DIFFUSION-BASED SPECTRAL SUPER-RESOLUTION OF THIRD OCTAVE ACOUSTIC SENSOR DATA: IS PRIVACY AT RISK ?

Modan Tailleur[1*], *Chaymae Benaatia*[1], *Mathieu Lagrange*[1], *Pierre Aumond*[2], *Vincent Tourre*[3]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2] Université Gustave Eiffel, CEREMA, UMRAE, F-44344 Bouguenais, France
[3] Nantes Université, ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, F-44000 Nantes, France

ABSTRACT

Third octave spectral recording of acoustic sensor data is an effective way of measuring the environment. While there is strong evidence that slow (1s frame, 1 Hz rate) and fast (125ms frame, 8Hz rate) versions lead by-design to unintelligible speech if reconstructed, the advent of high quality reconstruction methods based on diffusion may pose a threat, as those approaches can embed a significant amount of *a priori* knowledge when learned over extensive speech datasets.

This paper aims to assess this risk at three levels of attacks with a growing level of *a priori* knowledge considered at the learning of the diffusion model, a) none, b) multi-speaker data excluding the target speaker and c) target speaker. Without any prior regarding the speech profile of the speaker (levels a and b), our results suggest a rather low risk as the word-error-rate both for humans and automatic recognition remains higher than 89%.

Index Terms— speech privacy, generative audio, acoustic sensor networks, audio encoding

1. INTRODUCTION

In recent years, the use of acoustic sensors for audio data collection has extended across diverse applications, encompassing domains such as smart homes [1], [2] and urban sensor networks [3], [4]. Ensuring the privacy of speech information is a key aspect of the deployment of such sensors, be it deployed on public or private places. A promising approach that emerged from previous studies involves the encoding of audio as fast third-octave spectrograms (FTOS), which are third-octave spectro-temporal data computed with 125ms windows and 20-29 frequency bands [5]. By considering a low sampling rate, this method proved effective in preserving speech privacy as phoneme average duration in spoken English is typically below 100ms [6]. With longer windows, the co-articulation of phonemes is lost, leading to an almost complete loss of intelligibility. This loss ensures ”speech privacy”, simply termed in this paper ”privacy”. Speaker privacy by means of its identity is another important issue that is not considered in this paper.

Considering an at the time state-of-the-art reconstruction approach combining a) the Moore-Penrose pseudo-inverse (PINV) for frequency retrieval and b) the Griffin-Lim algorithm [7] for phase retrieval, Gontier at al. empirically demonstrated that the recovered speech was unintelligible [8]. Based purely on signal processing techniques, this reconstruction method do not consider any *a priori* knowledge on spectro-temporal properties of spoken English. However, to our knowledge, no attempts have been made to recover speech information from FTOS data using deep learning methods. With the recent advancements in generative audio models [9], which leverage *a priori* knowledge from large amounts of speech data, we believe that there is a need to re-evaluate the aforementioned claim that is: FTOS encoding is *by-design* preserving speech privacy.

Particularly, the emergence of diffusion models [10] may pose a threat. These models are easier to train than Generative Adversarial Networks (GANs) and can thus be applied to a broader range of fields. Indeed, diffusion models have demonstrated super-resolution abilities [11], [12], and have notably shown good performances in enhancing the quality of speech [13], [14].

In order to evaluate potential privacy threats induced by training such algorithms, we define three distinct Attack Levels (AL) on FTOS data, based on training set selection:

- **AL0** denotes an unintentional attack, occurring when a model is trained on general urban data or without any specialized training. The aim of the attacker in this case is not to recover specifically speech information but rather to reconstruct general audio for analysis purposes.
- **AL1** denotes an attack resulting from training a model on general speech data. In this scenario, the attacker aims to recover speech information from an unknown speaker.
- **AL2** denotes an attack executed by training a model specifically on the voice of a target individual. In this

*Research partially funded by ANR AIBY4 (ANR-20-THIA-0011), and IRSTV (Institut de Recherche en Sciences et Techniques de la Ville).

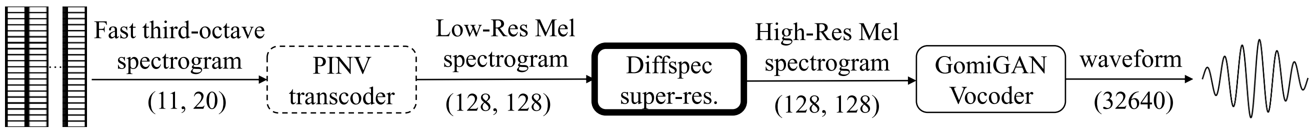

Fig. 1. Pipeline for audio super-resolution. Module outlined with dashes does not require any training, module with a plain outline is only pre-trained, and module outlined in bold is specifically trained for the task.

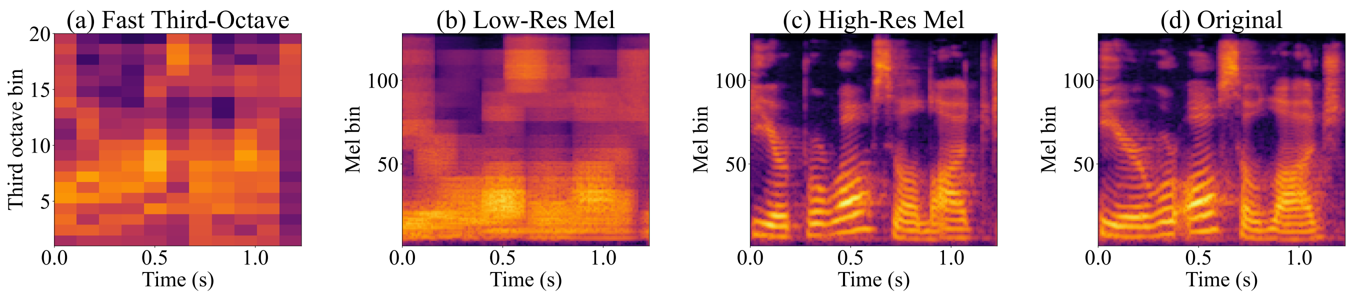

Fig. 2. Log spectrograms of the LJ002-0068.wav audio file from the LJSpeech evaluation set. The Diffspec model, trained on the LJSpeech training set, reconstructs a High-Res spectrogram (c) that closely matches the original (d).

scenario, it is assumed that the attacker has access to sufficient clear speech data from his target.

To evaluate the impact of large-scale attacks on sensor data, we anticipate that attackers will utilize Automatic Speech Recognition (ASR) systems on FTOS data. Given the uncertainty surrounding the effectiveness of ASRs in processing severely downsampled acoustic signals such as FTOS, we perform a subjective assessment through Human Speech Recognition (HSR) evaluation. This subjective evaluation provides a reference for interpreting ASRs performance under those kind of adverse conditions.

In Section 2, we present a super resolution technique based on diffusion [12] to recover speech from FTOS encoded speech. In Section 3, we describe the experimental protocol used to assess the privacy threat posed by this method. Sections 4 and 5, detail our main findings based on perceptual and computational assessment of intelligibility under scenarios simulating the 3 levels of attack.[1].

2. METHOD

Our proposed approach, which we refer to as the **Diffspec** method, transform FTOS into Mel spectrograms using a super-resolution algorithm, and use a vocoder for phase reconstruction. As audio is recovered, off-the-shelf ASRs or HSR can be used to recover speech information. We believe this approach is effective because it minimizes the need for extensive training. By relying on a pre-trained vocoder for the vocoding stage, the model only needs to focus on spectrogram reconstruction, thereby simplifying the overall process.

[1]Code operating solely on public data and audio examples are available at: `https://modantailleur.github.io/paperThirdOctavePrivacy/`

	FTOS	**GomiGAN Mel**
sample rate	32kHz	24kHz
window size	4096 (128ms)	1024 (43ms)
hop size	4000 (125ms)	320 (11ms)
window	Tukey	Hann
frequency bins	20	128
min frequency	125Hz	23Hz
max frequency	10kHz	12kHz

Table 1. Differences between fast third-octave spectrograms (FTOS) and GomiGAN Mel inputs.

As shown in Figure 1, we choose to use a Diffspec pipeline considering Saharia et al. [12] super-resolution algorithm applied on spectrograms, a conditioning method notably utilized in the NU-Wave 2 algorithm [15]. For the vocoder, we select the pre-trained GomiGAN model [16]. Consequently, the pipeline must align with the input dimensions of the GomiGAN model, as shown in Table 1. Going from FTOS to GomiGAN mels requires a 74x upscaling factor. After the vocoding stage, audio segments of 1.36s are computed and then concatenated with hops of 1.23s and cross-fades to match the initial audio length. An example of the spectrograms generated in the different Diffspec stages is shown in Figure 2.

First, we create an initial low-resolution approximation of a GomiGAN Mel spectrogram from the FTOS. We use a Moore-Penrose pseudo-inverse followed by Mel filtering to align with the target Mel frequency bins, and linear interpolation to match the number of time frames. This spectrogram will be referred to as **Low-Res Mel spectrogram** in the following sections. Detailed methodology for obtaining this pseudo-inverted Mel spectrogram can be found in Tailleur et al. [17].

The Low-Res Mel spectrogram is then refined using a

Diffspec based on the model proposed by Saharia et al. [12]. This refined spectrogram will be referred to as **High-Res Mel spectrogram**. The U-net model used for the diffusion includes two ResNet layers per block, organized into six blocks with 64, 64, 128, 128, 256, and 256 output channels, respectively, and contains a total of 28 million parameters. It features one attention block in both the downstream and upstream stages.

We then apply the GominGAN [16] vocoder on the output of the Diffspec model. GomiGAN is a general-purpose vocoder designed to convert any Mel spectrogram with characteristics shown in Table 1 to waveform audio. It is trained on a diverse range of audio datasets, including speech signals, music stems, animal sound recordings, and foley sounds. The GomiGAN model is based on BigVGAN [18], enhanced with Feature-wise Linear Modulation (FiLM) [19]. While the Griffin-Lim algorithm is also a potential vocoder alternative, GomiGAN offers several advantages. It runs approximately 15 times faster than Griffin-Lim with 32 iterations due to its GPU compatibility, and informal listening by the authors indicated superior audio quality. Objective and subjective evaluation have been considered for both vocoders. Only the Diffspecs using GomiGAN as vocoder are reported in Section 5, as no significant differences between the two vocoders are found.

3. EXPERIMENTAL PROTOCOL

Data

For each Attack Level (AL) defined in Section 1, we select a specific audio dataset to compute FTOS data to train our model. The chosen datasets are:

- **AL0** dataset: TAU Urban Acoustic Scenes 2020 Mobile dataset [20]. This dataset contains 10-second audio clips from 10 different acoustic scenes, including indoor public spaces, public transports, streets, and parks. While it includes some distant voice samples, it primarily focuses on ambient urban sounds, and totals 64 hours of audio.
- **AL1** dataset: Librispeech [21]. It consist of read audiobooks with more than 2,000 speakers. Specifically we use the "train-clean-100" subset, which includes 100 hours of audio data.
- **AL2** dataset: LJSpeech [2]. It comprises 13,000 audio clips from a single speaker reading seven non-fiction books, totalling 24h of audio. Like Librispeech, readings are available through the LibriVox project. The training dataset we considered comprises 12,900 audio clips, as 100 are kept for evaluation.

For evaluation, we randomly select 100 audio samples from the LJSpeech dataset. The relatively small size of the evaluation subset is due to the high computational cost of the inference of diffusion models.

Baseline

We compare our model against a simple pseudo-inverse approach using a PINV transcoder, as described in section 2. Compared to the proposed approach, this baseline simply bypasses the Diffspec step and applies the Vocoder directly to the Low-Res Mel spectrogram.

Learning procedure

The Diffspec model is trained with a learning rate of 10^{-4}, a batch size of 200, for 40,000 iterations.

Metric

To assess the privacy threat potentially induced by the generated audio samples, we measure the Word Error Rate (WER). The WER is a measure of the discrepancy between the reference transcriptions and those produced by HSR or ASR systems on the reconstructed speech. It is calculated as the sum of the number of substitutions, deletions, and insertions required to convert the inferred text into the reference text, divided by the total number of words in the reference text.

4. SUBJECTIVE EVALUATION

A subjective evaluation is performed on audio data reconstructed from FTOS, using WER on transcriptions from fluent english speakers who have reported normal hearing. Before the final analysis, the first author manually performs obvious grammar and typos corrections on all participants transcriptions.

8 audio samples are selected from the 100 audio samples of our LJSpeech evaluation subset, and are transformed into FTOS. These samples are chosen to be at least 6-s long and to contain content understandable without extensive cultural knowledge, avoiding names and slangs.

From informal listening done by the authors, some of the settings obviously lead to either full unintelligibility (AL0) or full intelligibility (original mel processed through GomiGAN). To keep the final perceptual test tractable and avoid cluttering the evaluation with settings that show highly contrasting WERs, we decide to evaluate those highly contrasted settings on a initial test conducted with only 3 participants.

20 other participants transcribe audios generated from the remaining two settings, which are the Diffspec models trained on Librispeech and LJSpeech (AL1 and AL2). Each participant transcribes a total of 16 audio samples: 4 samples from each of the two systems (AL1 and AL2) and 8 from the original audios. Participants whose transcriptions of the original audios lead to a WER exceeding 10% are excluded from the analysis. As a result, 3 participants are removed, leaving data of 17 participants for analysis.

As shown in Table 2, our reference human speech recognition (HSR) leads to a minimum of 92% WER on AL0, 90% of WER on AL1 and 64% of WER on AL2.

[2]LJSpeech dataset available at: https://keithito.com/LJ-Speech-Dataset/

Attack Level	Training Set	Method	HSR	FairseqS2T	W2V2	CRDNN	Whisper
-	-	Original (mel)	01 (±02)	10 (±16)	10 (±14)	09 (±13)	02 (±06)
-	-	White Noise	-	97 (±04)	100 (±00)	99 (±03)	95 (±04)
AL0	-	PINV transc.	98 (±04)	97 (±04)	95 (±06)	98 (±04)	82 (±20)
	TAU	Diffspec	92 (±09)	94 (±08)	93 (±09)	95 (±08)	92 (±13)
AL1	Librispeech	Diffspec	90 (±10)	91 (±13)	89 (±11)	91 (±11)	85 (±17)
AL2	LJSpeech	Diffspec	64 (±15)	53 (±21)	53 (±18)	46 (±23)	35 (±20)

Table 2. Word Error Rate (in %) on LJSpeech for the different combinations of methods and training sets. HSR represents the Human Speech Recognition evaluated with perceptual experiment. The "Original (mel)" method designates the original audio transformed into a GomiGAN Mel spectrogram and ran through GomiGAN vocoder. The confidence interval is based on standard deviation calculation.

5. OBJECTIVE EVALUATION

An objective evaluation is then performed using the WER computed for different state-of-the-art automatic speech recognition (ASR) models: Wav2Vec2 [22], the "large-v3" Whisper model [23], Fairseq S2T [24], as well as the CRDNN model from the Speechbrain library [25] called "asr-crdnn-rnnlm-librispeech". Table 2 presents the results of this evaluation.

The results indicate that all ASR systems yield WERs comparable to HSR, with the exception of Whisper. As they even outperform human evaluations at the AL2 attack level, this suggests that ASR systems are generally robust and effective for processing reconstructed audio. The Whisper ASR system exhibits a rather peculiar behavior, with overall low WERs accross all atack levels. This is especially worrying in the case of the PINV transcoder, showing a surprisingly low 82% WER, where human listening demonstrate almost complete unintelligibility (audio examples are provided on the companion page). We suspect that this behavior may be due to some sort of overfitting of this specific ASR to the LJSpeech dataset. Due to a lack of full understanding, the performance of Whisper is not discussed further.

For an unintentional attack resulting from training on a non-speech dataset or using an non-learned algorithms (attack level AL0), the results from the diverse ASR systems show that speech is nearly unintelligible. In this scenario, the WER is only a few percentage points lower than the one obtained with white noise, which is consistent with the results of Gontier et al. [8].

When targeting specifically speech information (attack level AL1), the results are slightly more concerning. W2V2 notably achieves an 89% WER. Although this might not seem alarming, understanding 10% of words could pose significant privacy risks in certain contexts, particularly when deploying systems in sensitive or private environments.

Targeting not only speech information but specifically the voice of the LJSpeech speaker (attack level AL2), the WER on W2V2 reaches up to 46%. While this level of WER indicates that a significant portion of the speech remains unintelligible, it is important to recognize that comprehending more than 50% of the words in a conversation might allow to grasp the overall meaning of the sentences. However, this scenario remains extreme, as it only occurs when a model is specifically trained on the target speaker's voice.

6. CONCLUSION

Using the Diffspec method across the different Attack Levels (AL) we have established, our model demonstrates a minimum Word Error Rate (WER) of 93% for AL0, 89% for AL1, and 46% for AL2 for several ASR systems on our LJSpeech test set.

Those results indicate a very low risk of extracting intelligible speech information, as an AL2 requires clean speech data from the target speaker. However, even with the seemingly high WERs for AL1 and AL2, the risk associated with these attacks is highly context-dependent. Future work could therefore consider perceptual assessment of acceptable risk thresholds for WER in specific application settings from public spaces, to offices or bedrooms.

We have shown in this paper that with the rise of generative models, fast third-octave spectrograms are no longer inherently privacy-aware as initially suggested by Gontier et al. [8]. Future research could involve training and testing models on multiple individual voices to further validate and strengthen these findings.

We believe that the proposed Diffspec approach nicely balance audio quality and training requirements in terms of data and power, but more complex approaches could be considered. For example, one could train an end-to-end automatic speech recognition (ASR) algorithm that uses fast third-octave spectrograms (FTOS) as input instead of Mel spectrograms [26]–[29], though this approach would require intensive training to reach the performance of off-the-shelf ASRs. One could also consider training vocoders to convert third-octave spectrograms directly to waveforms [30]–[33]. This latter approach could prove to be useful, but preliminary attempts by the authors demonstrated that the size of training dataset and computational power needed for training is notably larger than the ones required by the Diffspec method.

References

[1] S. Krstulović, "Audio Event Recognition in the Smart Home," in *Computational Analysis of Sound Scenes and Events*, 2018.

[2] M. Vacher *et al.*, "The sweet-home project: Audio technology in smart homes to improve well-being and reliance," in *EMBC*, 2011.

[3] A. Can *et al.*, "CENSE Project: General overview," in *Euronoise*, 2021.

[4] M. Cartwright *et al.*, "SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network," in *DCASE*, 2019.

[5] J. Ardouin *et al.*, "An innovative low cost sensor for urban sound monitoring," in *Internoise*, 2018.

[6] S. W. McKnight, A. O. Hogg, and P. A. Naylor, "Analysis of phonetic dependence of segmentation errors in speaker diarization," in *EUSIPCO*, 2021.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE TASLP*, 1984.

[8] F. Gontier *et al.*, "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," *MDPI Sensors*, 2017.

[9] S. Bengesi *et al.*, "Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers.," *IEEE Access*, 2024.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[11] B. Kawar *et al.*, "Denoising diffusion restoration models," *NeurIPS*, 2022.

[12] C. Saharia *et al.*, "Image super-resolution via iterative refinement," *IEEE TPAMI*, 2022.

[13] J. Serrà *et al.*, *Universal speech enhancement with score-based diffusion*, 2022. arXiv: `2206.03065 [cs.SD]`.

[14] R. Sawata *et al.*, "Diffiner: A Versatile Diffusion-based Generative Refiner for Speech Enhancement," in *Interspeech*, 2023.

[15] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, 2022.

[16] M. Kang *et al.*, "FALL-E: A Foley Sound Synthesis Model and Strategies," in *DCASE*, 2023.

[17] M. Tailleur *et al.*, "Spectral trancoder: Using pretrained urban sound classifiers on undersampled spectral representations," in *DCASE*, 2023.

[18] S.-g. Lee *et al.*, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," in *ICLR*, 2022.

[19] E. Perez *et al.*, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *DCASE*, 2018.

[21] V. Panayotov *et al.*, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

[22] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[23] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[24] C. Wang *et al.*, "Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq," in *AACL-IJCNLP*, 2020.

[25] M. Ravanelli *et al.*, *Speechbrain: A general-purpose speech toolkit*, 2021. arXiv: `2106.04624 [eess.AS]`.

[26] S. Kim, H. Kim, and S. Yoon, *Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data*, 2022. arXiv: `2205.15370 [cs.SD]`.

[27] D. Yang *et al.*, "Diffsound: Discrete Diffusion Model for Text-to-Sound Generation," *IEEE TASLP*, 2023.

[28] J. Tae, H. Kim, and T. Kim, "EdiTTS: Score-based Editing for Controllable Text-to-Speech," in *Interspeech*, 2021.

[29] R. Huang *et al.*, "ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech," in *ACM MM*, 2022.

[30] Z. Kong *et al.*, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *ICLR*, 2021.

[31] T. D. Nguyen *et al.*, "Fregrad: Lightweight and Fast Frequency-Aware Diffusion Vocoder," in *ICASSP*, 2024.

[32] S.-g. Lee *et al.*, "PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior," in *ICLR*, 2022.

[33] N. Chen *et al.*, "WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis," in *Interspeech*, 2021.