일반논문 (Regular Paper)
방송공학회논문지 제31권 제4호, 2026년 7월 (JBE Vol.31, No.4, July 2026)
https://doi.org/10.5909/JBE.2026.31.4.782
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

자기 조건부 확산을 이용한 수동 소나 데이터 생성의 시간-주파수 특성 향상

김 용 록[a)], 김 용 국[b)], 이 예 지[b)], 고 현 석[a)‡]

Enhancing Time–Frequency Characteristics in Passive Sonar Data Generation via Self-Conditioned Diffusion

Yongrok Kim[a)], Yongguk Kim[b)], Yejee Lee[b)], and Hyunsuk Ko[a)‡]

요 약

해양 환경에서 수동 소나 데이터는 표적의 위치 및 특성 분석에 활용되는 중요한 수중 음향 신호이다. 최근 딥러닝 기반 분석 기법이 활발히 적용되고 있으나 다양한 데이터를 충분히 확보하기에는 시간적·경제적 제약이 존재하며, 이는 모델의 일반화 성능 저하로 이어질 수 있다. 이를 해결하기 위한 대안으로 데이터 생성 기반 접근 방식이 주목받고 있으나, 1차원 수동 소나 신호 생성 연구는 아직 제한적인 상황이다. 본 연구에서는 자기 조건부 확산 기반 소나 데이터 생성 모델을 제안한다. 제안 방법은 이전 단계에서 추정된 신호를 조건으로 활용하는 자기 조건화 구조를 통해 생성 과정의 일관성과 안정성을 향상시키고, 시간-주파수 특성을 효과적으로 반영한다. 이를 통해 신호의 세부 구조를 점진적으로 정제하여 보다 현실적인 특성을 갖는 데이터를 생성한다. 실험 결과, 제안 방법은 생성된 소나 신호의 시간-주파수 표현을 보다 잘 보존하며, 기존 방법 대비 향상된 생성 품질을 보임을 확인하였다.

Abstract

Passive sonar data are important underwater acoustic signals used to analyze the location and characteristics of targets in marine environments. Recently, deep learning–based methods have been widely applied, but collecting sufficiently diverse data is limited by time and cost. This limitation can degrade model generalization. Data generation approaches have emerged as a promising solution, but research on 1D passive sonar signal generation remains limited. In this paper, we propose a self-conditioned diffusion model for passive sonar data generation. The proposed method uses signals estimated at previous steps as conditions. This improves the consistency and stability of the generation process and effectively reflects time–frequency characteristics. The model progressively refines the signal structure to produce more realistic data. Experimental results show that the proposed method better preserves the time–frequency representations of generated sonar signals and achieves improved generation quality compared to existing methods.

Keyword : Data generation, Diffusion model, Self-condition, Sonar

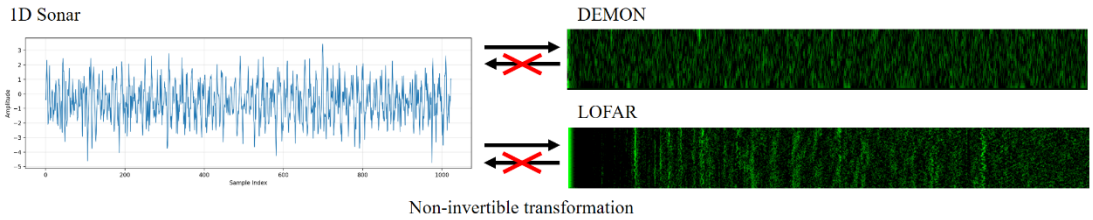

그림 1. 1D 소나 신호와 비가역 시간-주파수 변환 결과 (DEMON, LOFAR) 예시
Fig. 1. Example of 1D sonar signal and its non-invertible time–frequency representations (DEMON, LOFAR)

Ⅰ. 서 론

소나(Sonar; Sound Navigation and Ranging) 시스템은 수중 음향 신호를 활용하여 표적의 위치와 특성을 분석하는 기술로, 해양 감시 및 탐지 분야에서 핵심적인 역할을 수행한다. 최근에는 소나 신호 분석에 딥러닝 기반 기법[1-3]이 활발히 적용되고 있으나, 다양한 해양 환경을 반영한 대규모 학습 데이터를 확보하는 데에는 여전히 현실적인 제약이 존재한다. 실제 환경에서 다양한 소나 데이터를 수집하는 것은 높은 비용과 시간이 요구되며, 군사 및 보안적 특성으로 인해 공개 데이터 또한 제한적이다. 데이터 부족 문제는 딥러닝 모델의 일반화 성능 저하로 이어질 수 있으며, 생성 모델 기반 데이터 증강을 통해 완화할 수 있다[4-6].

특히 수동 소나(passive sonar)는 표적이 방출하는 음향 신호를 수신하여 분석하는 방식으로, 선박 소음과 같이 추진기 및 기계적 구조 특성을 반영하는 신호가 주요 분석 대상이 된다. 이러한 신호는 일반적으로 1차원(1D) 시계열 형태로 수집되며, 신호의 특성을 효과적으로 분석하기 위해 시간-주파수 영역에서의 해석이 필수적이다. 이에 따라 LOFAR(Low Frequency Analysis and Recording) 및 DEMON(Detection of Envelope Modulation on Noise)[7]과 같은 시간-주파수 변환 기법이 널리 활용되고 있으며, 이를 통해 신호를 2차원(2D) 스펙트럼 형태로 표현할 수 있다.

그림 1은 1D 소나 신호와 이를 LOFAR 및 DEMON으로 변환한 결과를 나타낸다. LOFAR 표현에서는 수직 방향으로 길게 이어진 강한 신호인 토널(tonal)과 배경 노이즈가 주로 나타나며, 이는 선박의 회전 기계 및 추진기에서 발생하는 주기적 성분을 반영한다. 반면 DEMON 표현은 신호의 진폭(envelope)을 추출하여 변조 성분을 강조한 것으로, 전체적으로 배경 노이즈가 지배적인 가운데 주기적인 변조 패턴이 나타나는 특징을 가진다.

그러나 이러한 변환 과정은 STFT(Short-Time Fourier Transform) 수행 후 magnitude 성분만을 사용하기 때문에 phase 정보가 소실되는 비가역적(non-invertible) 특성을 갖는다. 따라서 LOFAR 및 DEMON과 같은 스펙트럼 표현으로부터 원래의 1D 신호를 완전히 복원하는 데에는 근본적인 한계가 존재한다.

따라서, 원신호인 1D 데이터 차원에서 직접 생성하는 기법에 대한 필요성이 커지고 있다. 최근 1D 신호 생성 모델은 주로 오디오 분야를 중심으로 빠르게 발전해 왔으나, 대부분 지각적 음질 향상에 초점을 맞춰 설계되어 물리 기반 센서 신호의 고유한 스펙트럼 구조를 고려한 연구는 아직 제한적이다.

이에 본 연구에서는 1D 수동 소나 데이터를 생성하기 위한 자기 조건부 확산 기반 생성 프레임워크를 제안한다. 제안 방법은 확산 과정에서 예측된 노이즈로부터 복원된 신호를 다음 단계의 조건으로 활용하는 자기 조건화 구조를 도입하여, 생성 과정 전반의 일관성과 안정성을 향상시킨다. 또한 자기 조건 신호로부터 STFT 기반 시간-주파수 표현을 추출하고, 이를 1D 신호 특징과 교차 어텐션 방식으로 결합하는 시간-주파수 융합 모듈을 설계하였다. 이를 통해

a) 한양대학교 전자공학과(Department of Electrical and Electronic Engineering, Hanyang University)
b) LIG 디펜스&에어로스페이스 해양연구소(Maritime R&D Center, LIG Defense&Aerospace(LIG D&A))
‡ Corresponding Author : 고현석(Hyunsuk Ko)
E-mail: hyunsuk@hanyang.ac.kr
Tel: +82-31-400-5162
ORCID: https://orcid.org/0000-0002-7015-8351
※ 본 연구는 LIG D&A의 산학협력 연구과제(LIG D&A-2024-0689(00))의 지원을 받아 수행되었습니다.
· Manuscript received June 5, 2026; Revised June 15, 2026; Accepted June 15, 2026.

생성 과정에서 스펙트럼 구조를 명시적으로 반영하고, 신호의 세부 구조를 점진적으로 정제한다. 실험 결과, 제안 방법은 생성된 소나 신호의 시간-주파수 표현을 보다 잘 보존하며, 기존 방법 대비 향상된 생성 품질을 보임을 확인하였다.

Ⅱ. 1D 신호 생성 모델

1. 자기회귀 기반 생성 모델

1차원(1D) 신호에 대한 생성 모델 연구는 주로 오디오 영역을 중심으로 발전해 왔다. 초기에는 WaveNet[8]과 같은 자기회귀(autoregressive) 모델이 제안되어, 과거 샘플에 대한 조건부 확률 분포를 순차적으로 모델링하는 방식으로 파형을 생성하였다. 이러한 접근법은 높은 품질의 신호를 생성할 수 있었으나, 샘플 단위의 순차 생성 구조로 인해 계산 비용이 매우 크고 생성 속도가 느리다는 한계를 가진다.

2. GAN 기반 생성 모델

이후 생성적 적대 신경망(GAN)을 활용한 1D 생성 모델이 등장하면서 병렬 생성이 가능해졌고, WaveGAN[9], MelGAN[10], HiFi-GAN[11] 등은 생성 속도와 실용성을 크게 향상시켰다. 그러나 GAN 기반 모델은 학습 불안정성과 모드 붕괴 문제를 가지며, 장기 시간 의존성을 안정적으로 모델링하는 데에는 여전히 어려움이 있다.

3. 확산 기반 생성 모델

최근에는 확산 모델(Diffusion Model)이 1D 신호 생성 분야에서 효과적인 생성 기법으로 주목받고 있다. DiffWave[12], WaveGrad[13] 등은 점진적인 노이즈 제거 과정을 통해 안정적인 학습과 고품질 파형 생성을 동시에 달성하였다. 이후 연구에서는 긴 시계열 데이터를 보다 안정적으로 생성하기 위해 확산 모델과 자기회귀 구조를 결합한 DiffAR[14]와 같은 접근 방식이 제안되었다.

4. 1D 소나 데이터 생성의 한계

1D 생성 모델은 오디오 영역을 중심으로 발전해 왔으며, 주로 지각적 음질 향상이나 신호 변환에 초점을 두고 설계되었다. 그러나 소나와 같은 물리 기반 센서 신호는 전파 특성과 반사 구조에 의해 형성되는 고유한 시간-주파수 패턴을 가지며, 이러한 스펙트럼 구조의 물리적 일관성이 핵심적인 분석 요소가 된다. 기존 오디오 기반 생성 모델은 이러한 특성을 명시적으로 고려하지 않으며, 센서 도메인에 특화된 1D 신호 생성 연구는 제한적이다.

Ⅲ. 제안 방법

본 연구에서는 소나 신호의 스펙트럼 구조와 통계적 특성을 동시에 보존할 수 있는 1D 확산 모델 기반 생성 프레임워크를 제안한다. 제안 모델은 DiffAR의 Unconditional 구조를 기반으로 하며, 이전 단계에서 복원된 신호를 조건으로 활용하는 자기 조건화(self-conditioning) 기법[15]을 적용하였다.

특히, 복원된 신호로부터 시간-주파수 특성을 추출하고, 이를 1D 신호 특성과 효과적으로 결합하도록 하는 cross-attention 기반의 시간-주파수 융합 모듈인 STFF(Self-conditioned Time-Frequency Fusion)를 새롭게 제안한다. STFF 모듈은 주파수 구조 정보를 명시적으로 반영함으로써, 기존 1D 기반 생성 모델에서 제한적이었던 주파수 특성 표현을 효과적으로 보완한다. 전체 모델은 확산 과정에서 노이즈를 점진적으로 제거하는 방식으로 학습되며, 각 단계에서 자기 조건화와 STFF 모듈을 통해 보강된 시간-주파수 정보를 함께 활용하여 신호를 생성한다.

1. 확산 모델

확산 모델은 데이터에 점진적으로 가우시안 노이즈를 추가하는 확산 과정(forward process)과, 역으로 제거하는 복원 과정(reverse process)으로 구성된다. 확산 과정은 원본 신호 x_0에 시간 단계(time step) t에 따라 노이즈를 점진적으로 추가하여 노이즈가 포함된 신호인 x_t를 생성하는 과

정이며, 다음과 같이 정의된다.

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (1)$$

여기서 x_t와 x_{t-1}은 각각 시간 단계 t와 $t-1$에서의 노이즈가 포함된 신호를 의미한다. β_t는 시간 단계 t에 따른 노이즈 스케줄로, 각 단계에서 추가되는 노이즈의 크기를 조절한다. I는 단위 행렬을 의미하며, $N(\cdot)$은 평균이 $\sqrt{1-\beta_t}x_{t-1}$, 분산이 $\beta_t I$인 가우시안 분포를 나타낸다. 즉 식 (1)은 이전 단계의 신호 x_{t-1}에 가우시안 노이즈를 추가하여 다음 단계의 신호 x_t를 생성하는 과정을 의미한다.

복원 과정에서는 신경망 ϵ_θ를 통해 x_t에 포함된 노이즈를 예측하고, 이를 기반으로 원본 신호를 점진적으로 복원한다. 이를 위해 다음과 같이 diffusion loss를 최소화하도록 학습한다.

$$L_{diff} = E_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t)\|_1] \quad (2)$$

여기서 ϵ은 확산 과정에서 추가된 실제 가우시안 노이즈를 의미하며, $\epsilon_\theta(x_t, t)$는 시간 단계 t의 노이즈 신호 x_t를 입력으로 받아 신경망이 예측한 노이즈를 의미한다. θ는 신경망의 학습 가능한 파라미터이다. $E_{x_0,\epsilon,t}[\cdot]$는 원본 신호 x_0, 가우시안 노이즈 ϵ, 시간 단계 t에 대해 계산한 손실의 기대값을 의미하며, 실제 학습에서는 미니 배치 내 샘플들의 평균으로 계산된다. $\|\cdot\|_1$은 실제 노이즈와 예측 노이즈 간의 차이를 계산하는 Mean Absolute Error(MAE)를 의미한다. 따라서 식 (2)는 모델이 확산 과정에서 추가된 노이즈를 정확히 예측하도록 학습하는 손실 함수이다.

2. 네트워크 구조

제안 모델은 1D 신호 생성을 위해 DiffWave 구조를 기반으로 설계되었으며, 이를 그림 2에 나타냈다. DiffWave는 dilated convolution을 활용한 Residual Block 구조를 통해 긴 시간 의존성을 효과적으로 모델링한다.

각 Residual Block은 입력 신호와 time step을 embedding한 뒤 함께 활용하여 feature를 점진적으로 정제하며, skip connection을 통해 안정적인 학습을 수행한다.

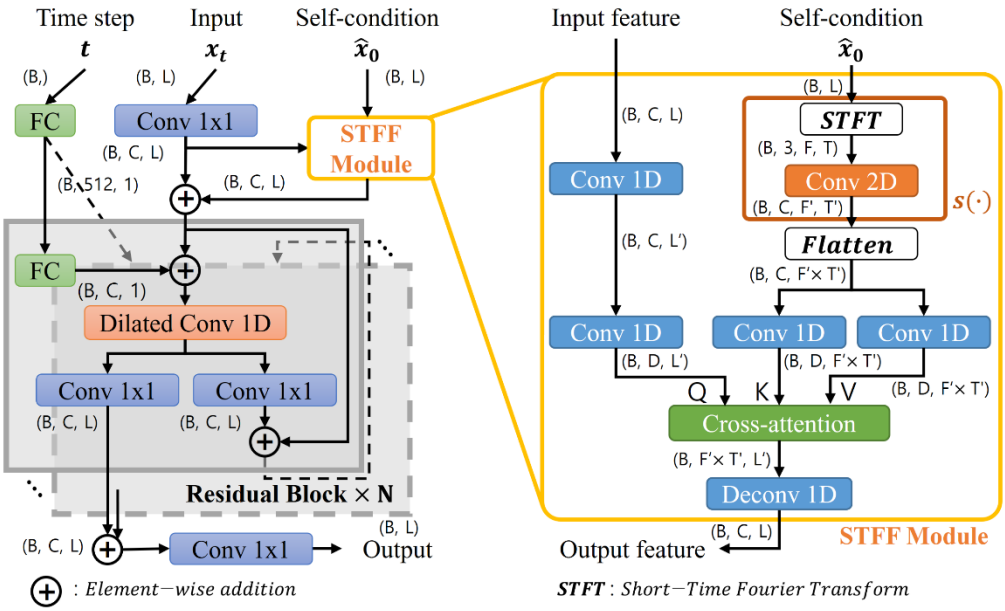

그림 2. 제안 네트워크 구조. B, C, D는 각각 배치 크기, 채널 수, 그리고 Query (Q), Key (K), Value (V)의 임베딩 차원을 의미한다.
Fig. 2. Proposed network architecture. B, C and D denote the batch size, the number of the channels, and the embedding dimension of Query (Q), Key (K), and Value (V), respectively.

3. 자기 조건화 (Self-Conditioning)

자기 조건화는 이전 단계에서 추정된 신호를 현재 단계의 조건으로 활용하는 기법이다. 제안 모델에서는 이전 step에서 추정한 노이즈를 이용해 복원된 신호 $\hat{x}_0$를 모델의 입력과 함께 제공하여, 노이즈 제거 과정에서 추가적인 조건 정보를 활용하도록 한다.

그러나 학습 과정에서는 실제 이전 step의 결과를 사용할 수 없기 때문에, 현재 단계의 노이즈 신호로부터 $\hat{x}_0$를 먼저 추정한다. 이때, 해당 추정 과정은 gradient를 차단한 상태에서 수행된다. 이후, 이렇게 얻은 $\hat{x}_0$를 self-condition으로 사용하여 모델 입력에 함께 제공하고 학습을 진행한다. 이를 통해 각 단계에서 보다 풍부한 정보를 기반으로 디노이징이 이루어지며, 생성 과정에서의 정보 전달을 효과적으로 보조한다.

자기 조건화가 적용됨에 따라 모델 입력은 x_t에서 $(x_t, \hat{x}_0)$로 확장되며, 이에 따라 식 (2)는 다음과 같이 수정된다.

$$L_{diff} = E_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, \hat{x}_0)\|_1] \quad (3)$$

4. STFF (Self-conditioned Time-Frequency Fusion) 모듈

시간-주파수 특성을 효과적으로 반영하기 위해 STFF 모듈을 제안한다. 해당 모듈에서는 먼저, 자기 조건화를 통해 얻어진 신호에 대해 STFT를 적용하여 시간-주파수 표현을 생성한다. STFT 결과는 magnitude와 phase로 구성되며, 제안 모듈에서는 magnitude의 동적 범위를 줄이기 위해 log-magnitude를 사용한다. 또한 phase 값은 $-\pi$에서 π 범위로 표현되는데, 위상 경계에서 발생할 수 있는 불연속성을 완화하고, 위상 정보를 연속적인 형태로 표현하기 위해 cos(phase)와 sin(phase) 성분으로 분해하여 사용한다. 따라서 STFT 변환 후의 결과는 log-magnitude, cos(phase), sin(phase)를 결합한 3채널 형태인 (B, 3, F, T)로 구성된다. 여기서 B는 배치 크기, 3은 채널 수, F는 주파수 빈의 개수, T는 시간 프레임의 개수를 의미한다.

다음으로, 해당 시간-주파수 표현에 2D convolution을 적용하여 주파수 및 시간 축의 구조적 특성을 추출한다. 이후, 추출된 시간-주파수 특성과 1D 신호 특성을 cross-attention 구조를 통해 결합한다. 이때 1D 신호 특성은 query로, 시간-주파수 특성은 key와 value로 사용된다.

이를 통해 시간 영역의 파형 정보와 주파수 영역의 구조적 정보를 순차적으로 반영할 수 있으며, 기존 1D 기반 생성 모델에서 부족했던 주파수 특성 표현을 효과적으로 보완한다.

학습 시, STFF 모듈이 자기 조건화를 통해 보강된 시간-주파수 특성을 효과적으로 반영하도록 하기 위해, 원본 신호와 복원 신호 간의 시간-주파수 표현 차이를 최소화하는 spectral loss를 적용한다. 그림 2와 같이 STFT 및 2D convolution 연산을 $S(\cdot)$으로 정의할 때, 해당 loss는 다음과 같이 표현된다.

$$L_{spec} = \|s(x) - s(\hat{x}_0)\|_1 \quad (4)$$

따라서 최종 학습에 사용되는 손실 함수는 식 (5)와 같이 정의되며, 실험을 통해 λ를 100으로 설정하였다. 모델은 해당 손실 함수를 이용해 end-to-end로 학습된다.

$$L_{spec} = \|s(x) - s(\hat{x}_0)\|_1 \quad (5)$$

Ⅳ. 실험 결과

1. 데이터셋

본 연구에서는 실제 해양 환경에서 취득한 수동 소나 데이터 256개를 사용하였다. 각 데이터는 2048Hz의 샘플링 주파수를 가지며, 약 14분 길이의 시계열 신호로 구성된다. 해당 데이터는 선박이 운항하는 실제 해양 환경에서 취득된 신호로, 해양 환경 소음과 선박 소음이 포함되어 있다.

각 신호는 min-max 정규화를 적용한 후, -1.0에서 1.0 범위로 매핑하였다. 생성 모델이 제한된 실제 수동 소나 데이터의 분포를 최대한 학습할 수 있도록, 학습 시에는 256개 데이터를 모두 사용하였다. 이때, 전체 신호 중 32초 길이에 해당하는 구간을 랜덤 크롭하여 입력으로 사용했으며, 추

가적인 외부 정보는 활용하지 않았다. 또한, self-conditioning에 사용되는 신호 역시 동일 데이터로 추정된 결과를 사용하였다.

평가 시에는 시간-주파수 변환 결과의 시각적 특성을 확인하기 위해 100초 길이의 신호 512개를 생성하였다. 또한, 정량 평가를 위해 원본 데이터 각각에서 100초 길이의 구간을 2개씩 랜덤 샘플링하여 총 512개의 실제 데이터를 구성하였으며, 이를 생성 데이터와의 비교에 사용하였다. 이를 통해 제안 모델이 실제 수동 소나 신호의 통계적 특성과 시간-주파수 구조를 얼마나 유사하게 재현하는지 평가하였다.

본 연구에서 사용한 데이터는 시뮬레이션 신호가 아니라 실제 해양 환경에서 취득된 수동 소나 신호이므로, 실제 운용환경에서 나타나는 해양 배경 소음과 선박 소음의 시간적·주파수적 특성을 포함하고 있다. 따라서 제안 모델의 학습 및 평가에 실제 환경 기반의 신호 특성을 반영할 수 있다.

2. 평가 지표

생성한 1D 신호의 품질을 정량적으로 평가하기 위해 Fréchet Inception Distance(FID)[16]를 사용하였다. FID는 생성된 데이터와 실제 데이터의 feature 분포 간 차이를 측정하는 지표로, 값이 작을수록 생성 품질이 우수함을 의미한다. 먼저 VGGish[17]를 통해 1D 신호를 feature 공간으로 변환한 뒤, 해당 feature 분포 간의 FID를 계산하였다. 또한, 1D 신호를 시간-주파수 영역의 2D 표현인 LOFAR와 DEMON으로 변환한 후, Inception[18] 네트워크를 활용하여 feature 공간으로 변환한 뒤 FID를 각각 측정하였다. 이를 통해 시간-주파수 영역에서 생성 품질을 평가하였다.

3. 비교 모델 및 실험 구성

Unconditional 설정에서 가변 길이 입력을 처리할 수 있는 1D 생성 모델을 대상으로 성능을 비교하였다. 비교 모델로는 DiffWave[12]와 DiffAR[14], 그리고 최신 오픈소스 모델인 Tiny Audio Diffusion(TAD)[19]를 사용하였다. DiffAR는 DiffWave를 기반으로 ResBlock의 채널 구성과 dilated convolution 구조를 확장한 모델이다. 제안하는 STFF 모듈과 spectral loss(SL)의 효과를 검증하기 위해, DiffWave와 DiffAR에 이를 추가하여 비교하였다.

모든 비교 모델의 입력 데이터 길이는 32초에 해당하는 65,536개 샘플로 설정했으며, batch size는 8로 설정하였다. TAD는 공개 코드에서 제공하는 기본 구성을 사용하였다. DiffWave와 DiffAR 기반 모델은 모두 Adam optimizer를 사용하고 learning rate는 2×10^{-4}로 설정하였다. DiffWave 기반 모델에서는 residual layer 수 (N), residual channel 수 (C), dilation cycle length를 각각 30, 64, 10으로 설정하였다. DiffAR 기반 모델에서는 각각 36, 256, 11로 설정하였다.

STFF 모듈에서 STFT는 FFT size, window length, hop length를 각각 2048, 2048, 512로 설정하고 Hann window를 사용했으며, 이에 따른 overlap은 75%이다. STFT 결과는 1,025개의 주파수 빈과 129개의 시간 프레임으로 구성되며, log-magnitude, cos(phase), sin(phase)의 3채널 feature로 표현되어 (B, 3, F=1025, T=129) 형태를 갖는다. 2D convolution은 4개의 layer로 구성되며, kernel size는 순서대로 2×2, 3×3, 3×3, 3×3이다. 이를 통해 (B, C, F'=64, T'=32) 형태의 시간-주파수 특성이 추출된다. 이후 flatten하여 F'×T'=2,048개의 토큰으로 변환되며, 이를 기반으로 cross-attention의 key와 value를 생성한다. 1D 신호 특성은 1D convolution을 통해 query로 변환되며, 이때 토큰 수 L'은 4096으로 설정하였다. Cross-attention은 4개의 attention head를 사용했으며, query, key, value의 embedding dimension인 D는 128로 설정하였다.

4. 정량 평가 비교 분석

표 1은 각 모델에 대한 1D FID와 LOFAR 및 DEMON 기반 2D FID 결과를 나타낸다. STFF 모듈을 적용한 경우, DiffWave와 DiffAR 모두에서 1D 및 LOFAR FID가 전반적으로 감소하는 경향을 보였다. 이는 self-conditioning으로부터 얻은 신호의 시간-주파수 특성을 1D feature에 효과적으로 반영함으로써, 생성 신호의 구조적 표현이 향상되었음을 의미한다. 여기에 spectral loss를 추가한 경우, 모든 지표에서 추가적인 성능 개선이 확인되었다. 특히 LOFAR

표 1. 1D 신호 생성 결과에 대한 FID 비교
Table 1. FID comparison of generated 1D signals results

Method	1D	LOFAR	DEMON
TAD	1.83	21.44	1.75
DiffWave	1.43	13.22	1.12
DiffWave + STFF Module	1.27	10.73	1.12
DiffWave + STFF Module + Spectral Loss	0.78	9.88	1.09
DiffAR	0.56	12.52	1.11
DiffAR + STFF Module	0.44	10.36	1.09
DiffAR + STFF Module + Spectral Loss	0.37	9.32	1.08

에서 FID의 감소는 spectral loss가 시간-주파수 feature 관점에서 보정하도록 학습함으로써 STFF 모듈의 효과를 보완한 결과로 해석된다. DEMON에서는 FID는 모든 모델에서 낮은 값을 보이며 전반적인 성능 차이는 크지 않으나, 제안 방법 적용 시 일관된 감소 경향이 나타났다. 이는 DEMON 표현이 배경 노이즈 성분에 크게 영향을 받는 특성에도 불구하고, 제안한 방법이 이러한 통계적 특성까지 안정적으로 반영함을 시사한다. 또한 TAD 모델은 LOFAR 및 DEMON FID에서 상대적으로 높은 값을 보였으며, 이를 통해 제안 모델이 시간-주파수 구조를 보다 효과적으로 반영함을 확인할 수 있다.

5. 정성 평가 비교 분석

그림 3은 각 모델에서 생성된 1D 신호의 waveform을 원본과 비교하여 나타낸다. TAD와 DiffWave는 전체적인 분포는 유지되지만 진폭 변동이 제한적이며, 특히 DiffWave는 진폭 범위가 축소되는 경향을 보인다. DiffAR는 상대적으로 유사한 분포를 보이나, 일부 구간에서 불균일한 변동이 나타난다. STFF 모듈을 적용한 경우, DiffWave와 DiffAR 모두에서 신호의 진폭 분포와 변동성이 보다 안정적으로 개선되는 것을 확인할 수 있다. 또한 STFF 모듈에 spectral loss를 추가한 경우, 신호의 진폭 변화가 더욱 자연스럽고 연속적으로 나타나며, DiffAR에 적용한 경우 전체적인 waveform 구조가 원본과 가장 유사하게 형성된다. 이는 표 1에서 확인된 1D FID의 지속적인 감소와 일관되며, 제안한 방법이 신호의 통계적 특성과 동적 변화를 보다 효과적으로 반영함을 시사한다.

그림 4와 그림 5는 LOFAR 및 DEMON 변환 결과를 토널 성분의 많고 적음에 따라 Tonal-rich case와 Tonal-sparse case로 구분하여 나타낸다. 그림 4의 LOFAR 변환 결과에

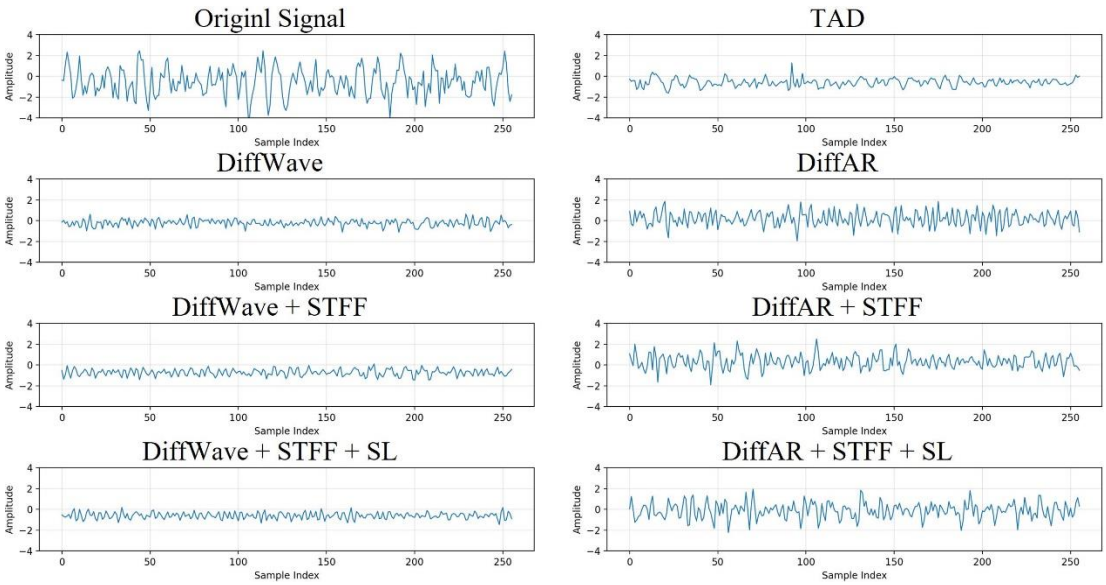

그림 3. 1D 신호 생성 결과에 대한 정성적 비교
Fig. 3. Qualitative comparison of generated 1D signals

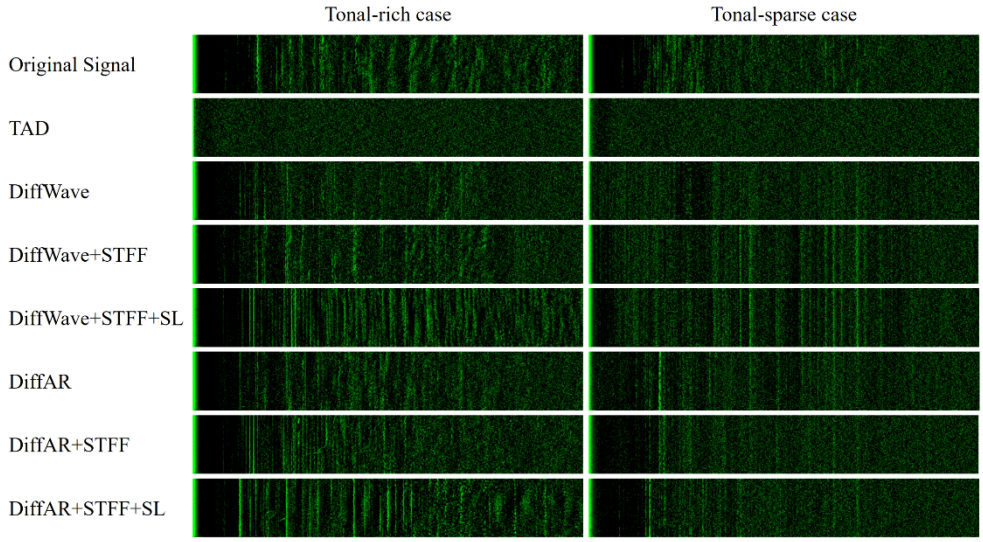

그림 4. LOFAR 변환 결과에 대한 정성적 비교
Fig. 4. Qualitative comparison of LOFAR transform results

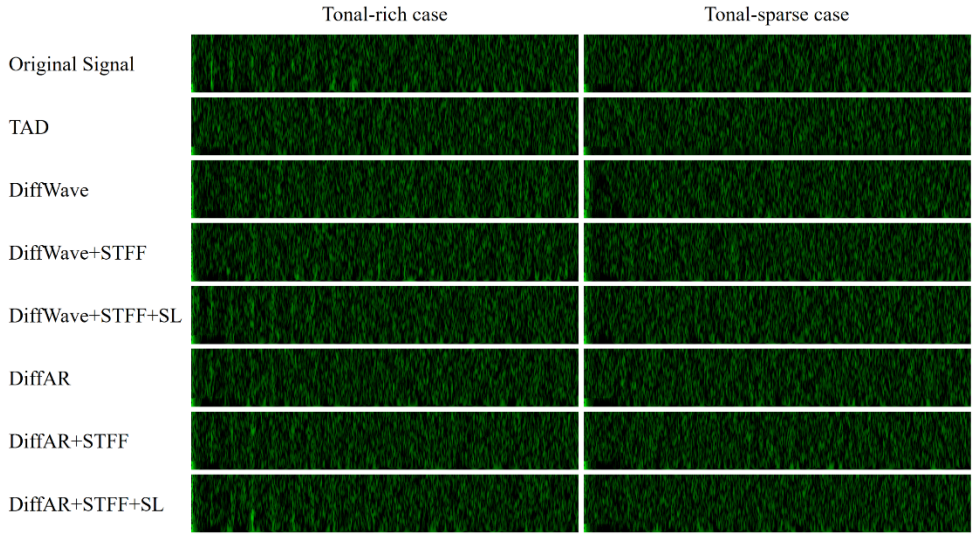

그림 5. DEMON 변환 결과에 대한 정성적 비교
Fig. 5. Qualitative comparison of DEMON transform results

서 Tonal-rich case는 시간 축을 따라 지속되는 다수의 수직 토널 성분을 보인다. TAD, DiffWave, DiffAR는 이러한 토널 구조를 충분히 재현하지 못하는 반면, STFF 모듈을 적용한 경우 토널 성분이 보다 뚜렷하고 연속적으로 나타나며, spectral loss를 함께 적용했을 때 이러한 경향이 더욱 강화된다. Tonal-sparse case에서는 원본 신호에 상대적으로 적은 토널 성분과 강한 배경 노이즈가 나타난다. TAD와 DiffWave는 원본 신호보다 토널 성분이 적게 나타나고, 저주파 영역의 배경 노이즈가 두드러지는 반면, 제안 방법을 적용한 모델은 주파수에 따른 에너지 분포를 더 안정적으로 반영한다. 그림 5의 DEMON 변환 결과에서는 두 경우 모두 배경 노이즈 성분이 지배적이며, 모델 간 시각적 차이

는 LOFAR 변환 결과에 비해 크지 않다. 그러나 tonal-rich case에서는 원본에 포함된 일부 토널 성분이 제안 방법을 적용한 모델에서도 부분적으로 재현되는 것을 확인할 수 있다. 종합적으로, STFF 모듈과 spectral loss를 적용한 경우 LOFAR와 DEMON 변환 결과에서 원본 신호의 시간-주파수 특성이 보다 효과적으로 반영된다.

6. LOFAR/DEMON 라인 검출 비교 분석

LOFAR/DEMON 도메인에서 원본과 유사한 토널 성분이 생성되었는지 확인하기 위해 라인 검출 방법인 DeepLSD[20]을 적용하여 결과를 비교했다. 그림 6은 원본 신호의 LOFAR/DEMON 변환 결과에 대한 라인 검출 예시이며, 표 2는 각 방법별 검출된 라인의 평균 개수와 평균 길이, 그리고 원본과의 차이를 비교한 결과를 나타낸다. 실험 결과 TAD는 LOFAR와 DEMON 모두에서 토널 성분을 충분히 생성하지 못하는 것으로 나타났다. 또한 DEMON은 LOFAR에 비해 토널 성분이 적게 나타나는 특성으로 인해 추출된 라인 수는 적었으나, 상대적으로 긴 라인이 검출되었다. DiffWave와 DiffAR 모두에서 STFF 모듈과 spectral loss를 적용한 경우 원본과의 평균 개수 및 평균 길이 차이가 감소하였으며, 이를 통해 제안 방법이 LOFAR/DEMON 도메인의 토널 구조를 보다 효과적으로 재현함을 확인하였다.

V. 결 론

본 연구에서는 해양 환경에서 취득되는 1D 수동 소나 신호의 데이터 부족 문제를 해결하기 위한 생성 프레임워크를 제안하였다. 제안 모델은 DiffWave 기반의 1D 확산 구조에 자기 조건화(self-conditioning)를 활용하고, 시간-주파수 융합 모듈(STFF)을 결합한 형태로 설계되었다.

제안한 STFF 모듈은 cross-attention 기반으로 시간-주파수 정보를 1D feature에 통합함으로써, 기존 1D 확산 모델에서 제한적이었던 주파수 구조 표현을 보완한다. 또한 spectral loss를 함께 적용하여, 예측된 신호를 시간-주파수 feature 관점에서 보정하도록 학습함으로써 생성 결과의 구조적 안정성을 향상시킨다.

실험 결과, STFF 모듈 적용 시 DiffWave 및 DiffAR 모

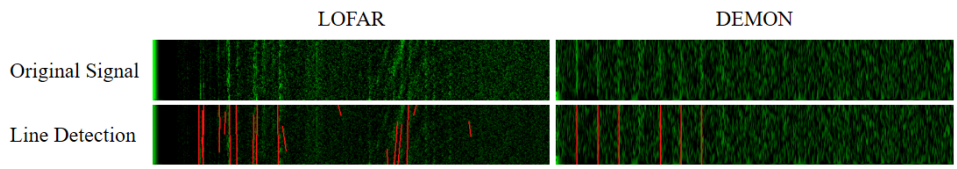

그림 6. 원본 신호의 LOFAR 및 DEMON 변환 결과에 대한 라인 검출 예시
Fig. 6. Example of line detection results for LOFAR and DEMON of the original signal

표 2. LOFAR 및 DEMON 변환 결과에 대한 라인 검출 예시
Table 2. Comparison of line detection results for LOFAR and DEMON

Method	LOFAR				DEMON			
	Number	Length	Number Diff.	Length Diff.	Number	Length	Number Diff.	Length Diff.
Original Signal	22.9	45.1	-	-	6.2	62.0	-	-
TAD	3.4	22.5	19.5	22.7	0.9	22.8	5.3	39.2
DiffWave	16.1	30.2	6.9	14.9	5.9	51.6	0.3	10.4
DiffWave + STFF	19.9	45.4	3.0	0.2	6.1	51.9	0.1	10.1
DiffWave + STFF + SL	21.6	49.4	1.3	4.3	6.1	52.8	0.1	9.2
DiffAR	15.4	33.6	7.6	11.6	6.3	51.5	0.2	10.5
DiffAR + STFF	17.3	34.8	5.7	10.3	6.1	51.7	0.0	10.3
DiffAR + STFF + SL	19.0	33.6	3.9	11.5	6.0	52.4	0.1	9.6

두에서 성능 개선이 확인되었으며, 여기에 spectral loss를 추가한 경우 추가적인 성능 향상이 나타났다. 특히 EnCodec 기반 1D FID와 LOFAR 및 DEMON 기반 2D FID 모두에서 일관된 개선을 보였으며, 이는 제안한 방법이 시간 영역뿐만 아니라 시간-주파수 영역의 구조적 및 통계적 특성을 효과적으로 반영함을 의미한다.

본 연구는 1D 신호 생성 문제에서 self-conditioning과 시간-주파수 표현을 결합한 확산 모델 설계의 유효성을 제시하며, 향후 다양한 센서 데이터 생성과 복원 문제로의 확장 가능성을 갖는다.

참 고 문 헌 (References)

[1] N. Hurtós, N. Palomeras, A. Carrera, and M. Carreras, “Autonomous Detection, Following and Mapping of an Underwater Chain Using Sonar,” *Ocean Engineering*, Vol.130, pp. 336-350, 2017.
doi: https://doi.org/10.1016/j.oceaneng.2016.11.072

[2] N. Palomeras, T. Furfaro, D. P. Williams, M. Carreras, and S. Dugelay, “Automatic Target Recognition for Mine Countermeasure Missions Using Forward-Looking Sonar Data,” *IEEE Journal of Oceanic Engineering*, Vol.47, No.1, pp.141–161, Jan. 2022.
doi: https://doi.org/10.1109/JOE.2021.3103269

[3] X. Zhang, H. Pan, Z. Jing, K. Ling, P. Peng, and B. Song, “UUVDNet: An Efficient Unmanned Underwater Vehicle Target Detection Network for Multibeam Forward-looking Sonar,” *Ocean Engineering*, Vol.315, pp.119820, 2025.
doi: https://doi.org/10.1016/j.oceaneng.2024.119820

[4] J. Lee, J. Im, and S. Bae, “A Diffusion Model-based Masking Data Augmentation Method for Improving Time Series Power Forecasting Performance,” *Journal of Broadcast Engineering*, Vol.30, No.3, pp.402-417, 2025.
doi: https://doi.org/10.5909/JBE.2025.30.3.402

[5] H. Seo, J. Lee, W. Park, H. Choi, K. Kim, and H. Jang, “Effective Data Preprocessing and Augmentation Methods of SAR Images for Optical Image Translation,” *Journal of Broadcast Engineering*, Vol.30, No.3, pp.480-489, 2025.
doi: https://doi.org/10.5909/JBE.2025.30.3.480

[6] H. Bong and M. Oh, “Conditional Variational Autoencoder-based Generative Model for Gene Expression Data Augmentation,” *Journal of Broadcast Engineering*, Vol.28, No.3, pp.275-284, 2023.
doi: https://doi.org/10.5909/JBE.2023.28.3.275

[7] D. K. Shin, and Y. D. Kim, “Automatic Frequency Line Detection Method for DEMON and LOFAR Gram using Transfer Learning,” *The Society of Convergence Knowledge Transactions*, Vol.11, No.3, pp.83-99, 2023.
doi: https://doi.org/10.22716/sckt.2023.11.3.028

[8] A. V. D. Oord et al., “WaveNet: A Generative Model for Raw Audio,” *arXiv*, 2016.
doi: https://doi.org/10.48550/arXiv.1609.03499

[9] C. Donahue, J. McAuley, and M. Puckette, “Adversarial Audio Synthesis,” *International Conference on Learning Representations (ICLR)*, 2019. url: https://openreview.net/forum?id=ByMVTsR5KQ

[10] K. Kumar et al., “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
doi: https://dl.acm.org/doi/10.5555/3454287.3455622

[11] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
doi: https://dl.acm.org/doi/10.5555/3495724.3497152

[12] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A Versatile Diffusion Model for Audio Synthesis,” *International Conference on Learning Representations (ICLR)*, 2021. url: https://openreview.net/forum?id=a-xFK8Ymz5J

[13] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “Wavegrad: Estimating Gradients for Waveform Generation,” *International Conference on Learning Representations (ICLR)*, 2021. url: https://openreview.net/forum?id=NsMLjcFaO8O

[14] R. Benita, M. Elad, and J. Keshet, “DiffAR: Denoising Diffusion Autoregressive Model for Raw Speech Waveform Generation,” *International Conference on Learning Representations (ICLR)*, 2024. url: https://openreview.net/forum?id=GTk0AdOYLq

[15] T. Chen, R. Zhang, and G. Hinton, “Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning,” *International Conference on Learning Representations (ICLR)*, 2021. url: https://openreview.net/forum?id=3itjR9QxFw

[16] M. Heusel et al., “GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
doi: https://dl.acm.org/doi/10.5555/3295222.3295408

[17] S. Hershey et al., “CNN Architectures for Large-Scale Audio Classification,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
doi: https://dl.acm.org/doi/10.1109/ICASSP.2017.7952132

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.2818-2826, 2016.
doi: https://doi.org/10.1109/CVPR.2016.308

[19] Tiny Audio Diffusion, https://github.com/crlandsc/tiny-audio-diffusion (accessed Mar. 2, 2026).

[20] R. Pautrat et al.,“DeepLSD: Line Segment Detection and Refinement with Deep Image Gradients,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp.17327-17336, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.01662

저 자 소 개

김 용 록

- 2020년 2월 : 한양대학교 ERICA 전자공학부 학사
- 2020년 3월 ~ 현재 : 한양대학교 전자공학과 박사과정
- ORCID : https://orcid.org/0009-0001-1148-4544
- 주관심분야 : 영상신호처리, 인공지능, 컴퓨터비전

김 용 국

- 2006년 2월 : 전남대학교 전자컴퓨터정보통신공학부 학사
- 2008년 8월 : 광주과학기술원 정보기전공학부 석사
- 2023년 2월 : 광주과학기술원 전기전자컴퓨터공학부 박사
- 2011년 1월 ~ 현재 : LIG 디펜스&에어로스페이스 수석연구원
- ORCID : https://orcid.org/0000-0001-6300-2538
- 주관심분야 : 음향신호처리, 소나 시스템 디자인

이 예 지

- 2022년 2월 : 명지대학교 전자공학과 학사
- 2024년 2월 : 명지대학교 전자공학과 석사
- 2024년 ~ 현재 : LIG 디펜스&에어로스페이스 선임연구원
- ORCID : https://orcid.org/0009-0005-2378-4703
- 주관심분야 : 소나 시스템, 데이터처리, 인공지능

고 현 석

- 2006년 8월 : 연세대학교 전기전자공학과 학사
- 2009년 2월 : 연세대학교 전기전자공학과 석사
- 2015년 5월 : University of Southern California 박사
- 2009년 2월 ~ 2010년 5월 : 삼성전자 연구원
- 2015년 6월 ~ 2020년 2월 : 한국전자통신연구원 선임연구원
- 2020년 ~ 현재 : 한양대학교 ERICA 전기공학부 부교수
- ORCID : https://orcid.org/0000-0002-7015-8351
- 주관심분야 : 영상압축, 영상 인지 화질 측정/개선, 컴퓨터비전, 딥러닝 기반 멀티미디어 신호처리