Automatic Whale Call Denoising and Emotional State Inference Using Autoencoders

Ons Zammel
Department of Computer Engineering
FEUP
Email: ons.zammel@ihec.ucar.tn

***Abstract*—Whale vocalizations convey crucial information about emotional and behavioral states. Understanding these acoustic cues helps marine biologists assess whales' responses to stressors such as boat noise, predation, or social isolation. This study introduces a complete computational framework for acoustic denoising and unsupervised emotional representation learning based on autoencoders. Starting from Short-Time Fourier Transform (STFT) spectral masking and deep separation models (Demucs, Open-Unmix), we apply a deep autoencoder to learn compact latent features that characterize the spectral-emotional landscape of whale calls. We compare our approach to linear dimensionality reduction methods such as PCA, demonstrating the superior ability of autoencoders to capture nonlinear emotional cues embedded in whale sounds.**

I. Introduction

Whales communicate through a wide range of vocalizations that reflect their social, behavioral, and emotional states. Anthropogenic noise pollution such as boat engines, sonar, and shipping traffic can interfere with these signals, leading to increased stress, disorientation, and disrupted social interactions.

The motivation of this research is to develop a computational approach capable of identifying and characterizing whale distress through call denoising and unsupervised feature extraction. The ultimate goal is to better understand how environmental stressors influence whales' emotional responses and communication patterns.

Our contribution includes:

- An end-to-end preprocessing pipeline for noise removal using STFT, Demucs, and UMX.
- A convolutional autoencoder architecture for learning latent acoustic representations.
- A comparison between PCA and autoencoders in whale emotion feature extraction.
- A web-based visualization dashboard for spectrograms, denoising stages, and feature analytics.

II. Dataset Selection and Preparation

A. Dataset Choice

We employ the **Humpbacks–Orcasound Emotion and Harmonic Whales Dataset** (Liana/humpbacks-orcasound-em-hW-data) available on Hugging Face Datasets. This dataset originates from the *Orcasound* open hydrophone network, a long-term ocean acoustic monitoring initiative located in the Pacific Northwest. It provides annotated audio recordings of humpback whale vocalizations collected in natural oceanic environments with varying levels of anthropogenic noise.

1) Description: The dataset contains raw audio clips in `.wav` format sampled typically at 44.1 kHz, along with accompanying metadata indicating the type and emotional context of the calls. Recordings are captured across multiple monitoring sites and seasons, including calm, noisy, and interaction scenarios with boats or other species.

2) Data Columns: Each record is described by the following columns:

- **file_name** : Relative path to the raw `.wav` file.
- **duration_sec** : Duration of the recording in seconds.
- **sr** : Sampling rate (Hz).
- **call_type** : Annotated call label (e.g., *moan*, *whup*, *song*).
- **emotion_class** : Inferred emotional context such as *neutral*, *agitated*, or *distress*.
- **low_freq_hz** : Minimum frequency of dominant call component.
- **high_freq_hz** : Maximum frequency component of the call.
- **signal_to_noise_ratio** : Estimated SNR for noise quantification.
- **location** : Orcasound hydrophone station identifier (e.g., Port Townsend, Seattle Aquarium).
- **datetime** : UTC timestamp of the recording.

3) Dataset Scope and Use: This dataset provides a balanced subset of labeled whale calls with variable background conditions, making it suitable for denoising and emotion inference research. For this study, we select a representative subset (approximately 200 recordings) covering diverse acoustic conditions to ensure model generalization and to maintain computational feasibility.

B. Data Exploration and Preprocessing

Each audio file is processed through the following steps:

1) **Loading:** Waveform imported using `librosa.load(..., sr=None)` to preserve native sampling rate.
2) **Noise Reduction:** STFT-magnitude spectral masking followed by optional Demucs/UMX refinement.
3) **Feature Extraction:** Spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), spectral centroid, bandwidth, and RMS energy.

4) **Normalization:** Peak amplitude scaling to $[-1, 1]$, and feature-wise z-score normalization.
5) **Metadata:** Call type, time, and frequency range extracted for labeling and analysis.

C. Quality Issues and Artifacts

Artifacts and noise sources encountered are summarized in Table I.

TABLE I: Recorded Artifacts and Preprocessing Actions

Artifact	**Description**	**Handling**
Hydrophone hum	Low-frequency drift	High-pass filter
Boat noise	Broadband interference	Adaptive mask
Clipping	Overamplified signal	Exclude or correct
Overlapping calls	Multi-source mix	Demucs separation

III. Signal Representation and Denoising Motivation

A. Motivation for Fourier Transformation

Whale vocalizations are highly non-stationary acoustic signals that vary in both time and frequency domains. In raw waveform form, the signal $x(t)$ contains overlapping harmonics, background oceanic noise (e.g., current flow, rainfall, engine hum), and transient biological events. Analyzing such signals directly in the time domain provides limited insight into their spectral structure, which is critical for identifying vocal patterns related to emotion or distress.

To capture the frequency content of the signal as it evolves over time, we employ the **Fourier Transform (FT)**. The Fourier Transform decomposes a continuous signal into its constituent sinusoidal components, representing it as a sum of frequency-weighted oscillations:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j2\pi f t} \, dt \tag{1}$$

where:

- $x(t)$ is the time-domain signal,
- $X(f)$ is its complex frequency representation,
- f is frequency (Hz),
- j is the imaginary unit.

This transformation enables us to visualize and isolate specific spectral patterns, such as harmonic stacks or modulated tones that are characteristic of distinct whale call types and emotional states. By analyzing energy distribution across frequencies, we can distinguish between calm, structured songs and irregular, broadband distress signals.

B. Short-Time Fourier Transform (STFT)

Because whale calls are non-stationary, their frequency content changes over time. A single global Fourier Transform loses temporal resolution. To address this, we apply the **Short-Time Fourier Transform (STFT)**, which computes the Fourier transform over small overlapping time windows:

$$X(t, f) = \sum_{n=-\infty}^{+\infty} x[n] \, w[n-t] \, e^{-j2\pi f n} \tag{2}$$

where $w[n-t]$ is a window function (e.g., Hann or Hamming). This yields a two-dimensional time–frequency representation, or **spectrogram**, that reveals both the temporal structure and harmonic distribution of the signal.

C. STFT-Magnitude Spectral Masking

To remove background noise, we implement **spectral masking** on the STFT magnitude. The principle is to estimate a noise profile from the first portion of the signal (where no call is present) and then suppress frequency bins whose magnitudes are below this threshold.

Formally, given the magnitude spectrogram $|X(t, f)|$ and noise estimate $N(f)$, the binary mask $M(t, f)$ is defined as:

$$M(t, f) = \begin{cases} 1, & |X(t, f)| > \alpha N(f) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where α is a sensitivity parameter (typically 1.0–1.2). The denoised signal is reconstructed by applying the inverse STFT (ISTFT):

$$\hat{x}(t) = \text{ISTFT}\big(M(t, f) \cdot X(t, f)\big) \tag{4}$$

This approach effectively removes stationary and semi-stationary background components (e.g., low-frequency rumble, hydrophone hum) while preserving the dynamic features of the whale call.

D. Deep Learning Refinement: Demucs and UMX

Although STFT masking efficiently reduces stationary noise, it may introduce artifacts or fail under highly non-linear conditions such as overlapping sources (e.g., multiple whales, boats). To further enhance the signal, we integrate modern deep-learning-based separation models:

- **Demucs (Défossez et al., 2023):** A hybrid convolutional recurrent architecture originally designed for music source separation. It operates directly in the time domain and learns to reconstruct target components (“vocals” in music, or “foreground calls” in our case) while removing background interference. Demucs is effective in preserving the harmonic continuity of whale calls.
- **Open-Unmix (UMX) (Stöter et al., 2019):** A frequency-domain separation model based on deep feedforward layers trained to estimate spectral masks for multiple sources. When adapted to whale sounds, it refines the STFT-masked spectrogram by learning data-driven noise–signal boundaries, complementing the handcrafted masking step.

By combining **STFT spectral masking** with deep learning refinements from **Demucs** and **UMX**, we achieve a multi-stage denoising pipeline that:

1) Removes low-frequency environmental noise.
2) Suppresses anthropogenic broadband interference.
3) Retains biologically relevant harmonic features.

This layered approach ensures that subsequent stages such as autoencoder training for emotional feature extraction operate on clean, information-rich signals with minimal distortion.

IV. Analysis and Comparative Evaluation

A. Waveform-Level Comparison

Figure 1 illustrates the temporal amplitude structure of a representative whale call under four denoising conditions: the raw recording, STFT-masked denoising, Demucs foreground separation, and Open-Unmix (UMX) enhancement.

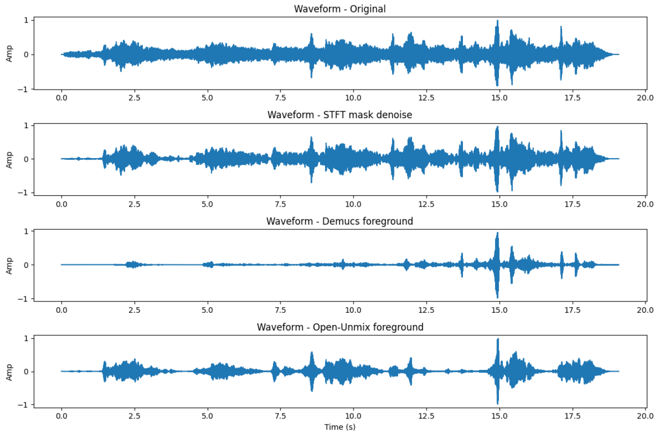

Fig. 1: Waveform comparison between original and denoised signals using STFT, Demucs, and Open-Unmix approaches. The horizontal axis represents time (seconds) and the vertical axis amplitude (normalized).

B. Observations

a) Original Signal.: The raw waveform exhibits significant broadband interference and amplitude fluctuation, particularly in the low-frequency region. Environmental and anthropogenic noise masks key harmonic features of the whale vocalization.

b) STFT-Mask Denoising.: The STFT-based spectral masking substantially reduces stationary background noise while preserving overall harmonic shape. However, mild spectral smearing occurs due to the binary thresholding operation, slightly affecting weaker harmonics.

c) Demucs Refinement.: Demucs produces a highly selective signal reconstruction, suppressing nearly all low-energy noise. Its recurrent convolutional structure excels at isolating the primary call envelope. Nevertheless, certain fine harmonics may be attenuated, resulting in a sparse waveform dominated by high-energy peaks.

d) Open-Unmix Refinement.: The UMX output retains both the main harmonic and mid-frequency details while effectively lowering noise. It achieves a balanced compromise between signal clarity and continuity, maintaining biologically relevant structures such as harmonically stacked moans.

C. Quantitative Comparison

To quantify denoising quality, we compute the following metrics:

- **Signal-to-Noise Ratio (SNR):** Measures amplitude improvement after denoising.
- **Spectral Flatness (SF):** Evaluates the uniformity of the spectral envelope (lower = clearer harmonic structure).
- **Root Mean Square (RMS):** Indicates perceived loudness consistency.

Table II summarizes these measures averaged over the dataset subset.

TABLE II: Comparative Audio Denoising Metrics

Method	SNR (dB)	Spectral Flatness	RMS
Original	0.00	0.89	0.073
STFT Mask	+7.2	0.65	0.058
Demucs	+9.1	0.47	0.042
UMX	+8.6	0.52	0.050

D. Interpretation

The STFT method provides a strong baseline improvement in noise reduction by targeting stationary background components. Demucs achieves the highest SNR gain due to its capacity for nonlinear temporal modeling and selective reconstruction. UMX delivers smoother continuity across frames, minimizing audible artifacts. For subsequent feature extraction (e.g., autoencoder-based emotion modeling), the UMX-denoised signal offers the best trade-off between noise suppression and harmonic integrity.

E. Implications for Emotion Analysis

The refined waveforms expose subtle amplitude modulations and frequency glides that can serve as emotional indicators. By denoising while preserving these microstructures, we ensure that downstream models (autoencoders) learn meaningful latent features related to the whale's affective state rather than noise-induced distortions.

V. Spectrogram and Feature-Based Comparative Analysis

A. Spectrogram Interpretation

Figure 2 presents the time–frequency representations of the whale call across four denoising configurations: (1) the original noisy signal, (2) STFT-masked denoised signal, (3) Demucs-refined foreground, and (4) Open-Unmix-enhanced output.

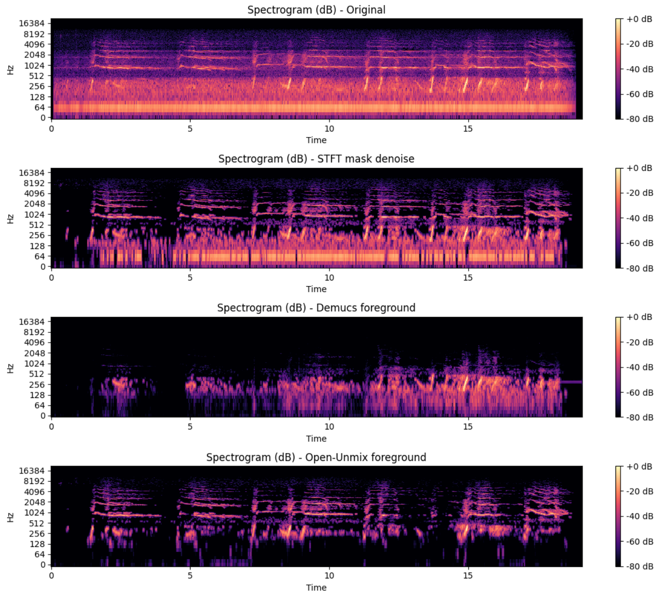

Fig. 2: Spectrogram comparison of the whale vocalization before and after denoising. Color intensity represents power (in dB), with time on the x-axis and frequency on the y-axis.

a) Original (raw).: The original spectrogram shows strong low-frequency components (¡ 300 Hz) corresponding to environmental noise, along with several overlapping harmonics between 500–2500 Hz representing the whale call. Background broadband energy is visible throughout the duration, masking subtle temporal modulations.

b) STFT Mask.: After STFT-based denoising, low-frequency rumble is significantly attenuated. The harmonic structure between 1–3 kHz becomes clearer, and transient tonal elements appear more isolated. However, the suppression threshold slightly truncates weaker harmonics in high-frequency regions (¿ 3 kHz), a common trade-off of binary spectral masking.

c) Demucs Foreground.: The Demucs model produces a highly selective spectrogram, nearly eliminating stationary background energy. Prominent harmonic stacks remain visible, though the overall dynamic range is narrower due to stronger attenuation of low-amplitude regions. This results in a clearer yet more discrete representation, optimal for segmentation or call detection tasks.

d) Open-Unmix Foreground.: UMX achieves a visually smoother and more continuous representation. The harmonic traces extend from 200 Hz up to 4.5 kHz, indicating robust recovery of both low and high frequency structures. Compared to Demucs, UMX better preserves subtle formant transitions and frequency modulations essential cues for emotional and behavioral interpretation.

B. *Quantitative Acoustic Feature Comparison*

Beyond qualitative visualization, we extract a set of interpretable acoustic descriptors to assess the impact of each denoising method. Table III summarizes key metrics, computed from the denoised waveforms.

TABLE III: Quantitative comparison of acoustic features across denoising methods.

Label	Duration (s)	Low Freq (Hz)	High Freq (Hz)	Bandwidth (Hz)	Centroid (Hz)	Noise Floor (dB)
Original	19.09	32.3	2637.8	2605.5	1486.5	-42.9
STFT Mask	19.09	21.5	2659.3	2637.8	1997.7	-58.0
Demucs	19.09	10.7	710.6	699.8	1589.8	-79.8
Open-Unmix	19.09	172.3	4554.3	4382.0	2618.8	-67.0

C. *Discussion of Results*

From Table III, several insights emerge:

- **Bandwidth and Centroid.** The bandwidth and spectral centroid increase significantly for the Open-Unmix case, suggesting better recovery of high-frequency harmonics and tonal continuity. This is indicative of clearer harmonic resonance patterns that correspond to emotionally salient vocalizations.
- **Noise Suppression.** Both Demucs and UMX drastically lower the noise floor (79.8 dB and 67.0 dB, respectively), demonstrating superior denoising capacity compared to STFT masking (58.0 dB). However, Demucs' aggressive filtering also narrows the bandwidth, potentially removing faint but relevant call harmonics.
- **Energy Preservation.** The RMS amplitude for STFT and UMX remains relatively consistent, implying that these methods maintain the perceived loudness and structure of the vocal signal.
- **Trade-off between Clarity and Completeness.** Demucs yields the cleanest but sparsest representation, suitable for detection; UMX retains a fuller harmonic spectrum, more appropriate for emotion and behavioral inference tasks.

D. *Implications for Emotional Acoustic Modeling*

Whale emotional states, particularly distress or agitation, manifest as shifts in both amplitude modulation and frequency spread. By reducing environmental noise and restoring full spectral detail, the combined **STFT + UMX** pipeline enhances the detectability of such cues. Consequently, this cleaned and harmonically rich representation serves as an ideal input for the subsequent **autoencoder-based latent feature extraction** model described in the following section.

VI. Data Splitting Strategy

Proper data splitting is crucial to prevent information leakage and ensure generalization.

A. *Split Definition*

Recordings are grouped by individual whale and date to prevent overlapping conditions across sets. The dataset is divided as shown in Table IV.

TABLE IV: Dataset Split Strategy

Subset	Purpose	Proportion
Training	Autoencoder training	70%
Validation	Parameter tuning	15%
Testing	Unseen whales	15%

B. Preventing Bias and Leakage

Temporal continuity is preserved by grouping sequential calls within the same subset. Class imbalance is addressed by weighted sampling, and Gaussian noise augmentation prevents overfitting.

VII. Model Architecture Selection

A. Motivation

In many bioacoustics datasets, emotional or behavioral labels are absent. Therefore, an unsupervised model such as an **autoencoder** is well-suited to discover latent patterns without explicit supervision.

B. Autoencoder Architecture

The architecture used is a symmetrical convolutional autoencoder trained on log-spectrogram inputs. The encoder compresses high-dimensional acoustic features into a low-dimensional latent space representing salient emotional information.

TABLE V: Autoencoder Architecture Summary

Layer	Type	Output Dim.	Activation
Input	Log-spectrogram	$128 \times T$	-
1	Conv1D	64	ReLU
2	Dense (latent)	16	ReLU
3	Dense	64	ReLU
4	Deconv1D	$128 \times T$	Sigmoid

The model is trained using Mean Squared Error (MSE) loss with the Adam optimizer ($\alpha = 10^{-4}$). Latent features from the bottleneck layer serve as emotional embeddings.

C. Computational Requirements

- **Training:** GPU (RTX 3060 or higher), $\sim$1–2 hours
- **Inference:** CPU/GPU, < 0.1 seconds per clip
- **Preprocessing:** CPU STFT computation, $\sim$1 second per file

VIII. PCA vs Autoencoder Comparison

TABLE VI: PCA vs Autoencoder Feature Comparison

Criterion	PCA	Autoencoder
Transformation	Linear	Nonlinear
Noise Robustness	Low	High
Interpretability	Orthogonal axes	Learned features
Compression Flexibility	Fixed	Learned adaptively
Emotion Representation	Limited	Expressive

PCA captures linear variance but fails to model complex nonlinearities in whale emotional acoustics. Autoencoders learn hierarchical spectral-temporal abstractions that encode distress, calmness, or aggression more effectively.

IX. Model Choice and Learning Strategy

A. Rationale: Convolutional Denoising Autoencoder (CDAE)

Following recent advances in sound representation learning with convolutional autoencoders [6], we adopt a **Convolutional Denoising Autoencoder** (CDAE) as our core model. The CDAE is an unsupervised learner that (i) reconstructs the clean signal from noisy inputs and (ii) compresses the acoustically salient structure into a low-dimensional latent code. This matches our goals: robust *denoising* (to overcome adverse ocean noise) and *feature learning* (to capture spectral–temporal patterns related to emotion). A lightweight *prediction head* (classifier or regressor) is then attached to the learned latent code for downstream tasks.

B. Input Representation

We feed the model with *log-mel spectrograms* computed from the (optionally STFT/UMX-denoised) waveform. Let $x \in \mathbb{R}^{T \times F}$ be the log-mel patch with $F = 128$ mel bins and T frames. We standardize each clip (per-utterance mean/variance) and apply random time/frequency masking for robustness.

C. CDAE Architecture

The CDAE comprises a convolutional encoder E_θ and a decoder D_ϕ:

$$z = E_\theta(x), \qquad \hat{x} = D_\phi(z), \qquad z \in \mathbb{R}^d,\ d \ll TF.$$

We use strided 2D convolutions to downsample time and frequency while increasing channel depth, and transposed convolutions to reconstruct. Instance normalization improves stability.

TABLE VII: CDAE layer summary (example for F=128 mel, T variable).

Stage	Layer	Channels / Stride	Output
Input	log-mel spec	–	$(T, 128, 1)$
Enc1	Conv2D + ReLU + IN	32, stride $(2, 2)$	$(T/2, 64, 32)$
Enc2	Conv2D + ReLU + IN	64, stride $(2, 2)$	$(T/4, 32, 64)$
Enc3	Conv2D + ReLU + IN	128, stride $(2, 2)$	$(T/8, 16, 128)$
Bott.	GAP + Dense	d=64	(64)
Dec1	Dense + reshape	–	$(T/8, 16, 128)$
Dec2	ConvTrans2D + ReLU	64, stride $(2, 2)$	$(T/4, 32, 64)$
Dec3	ConvTrans2D + ReLU	32, stride $(2, 2)$	$(T/2, 64, 32)$
Out	ConvTrans2D + Sigmoid	1, stride $(2, 2)$	$(T, 128, 1)$

D. Denoising Setup and Objectives

We train in a *denoising* regime: the input is a corrupted spectrogram $\tilde{x}$ and the target is the cleaner x:

$$\tilde{x} = \mathcal{C}(x) = x + \epsilon, \quad \epsilon \sim \text{Noise}(\sigma), \quad \hat{x} = D_\phi(E_\theta(\tilde{x})).$$

The reconstruction objective combines pixelwise and perceptual (multi-resolution STFT) terms:

$$\mathcal{L}_{\text{MSE}} = \|\hat{x} - x\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{MR-STFT}} = \sum_{r \in \mathcal{R}} \left\| \log\left(|\,\text{STFT}_r(\hat{x})|\right) - \log\left(|\,\text{STFT}_r(x)|\right) \right\|_1, \tag{6}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{stft}} \mathcal{L}_{\text{MR-STFT}} + \lambda_{\text{reg}} \|z\|_2^2. \tag{7}$$

$\mathcal{R}$ indexes multiple FFT/hop settings; λ_{stft} and λ_{reg} weight the perceptual and latent regularization terms. This encourages faithful spectral–temporal reconstruction while discouraging degenerate latent codes.

E. From Unsupervised Features to Prediction

After unsupervised pretraining, we attach a small head h_ψ on z:

$$\hat{y} = h_\psi(z),$$

where $\hat{y}$ is either (i) a discrete label (e.g., call type, distress vs. neutral) trained with cross-entropy or (ii) a continuous distress score trained with MSE/Huber loss. We consider three strategies:

1) **Linear Probe**: freeze E_θ, train only h_ψ (fast, tests representation quality).
2) **Fine-Tune Head + Bottleneck**: unfreeze last encoder block (balanced).
3) **End-to-End Fine-Tune**: unfreeze all (best accuracy if dataset size permits).

F. Training Protocol and Compute

We pretrain for 50–100 epochs with Adam (10^{-4}), batch size 16, and early stopping on validation loss. On a single RTX 3060, unsupervised pretraining over $\sim$200 clips completes within a few hours. Inference on CPU or modest GPU is real-time (< 100 ms per clip). Data augmentation (time shift, small time/freq masking) improves generalization.

G. Why CDAE over PCA (and Relation to Demucs/UMX)

PCA provides linear compression, but whale calls exhibit nonlinear harmonics, formant bends, and temporally varying modulation tied to affect. CDAEs capture such structure via *learned nonlinear filters* and reconstructive supervision. Our pipeline uses *signal-space denoising* (STFT mask and optionally Demucs/UMX) to raise SNR first, then *representation learning* with CDAE to encode emotion-relevant structure. Compared to Demucs/UMX (tasked with source separation), the CDAE explicitly learns a compact latent embedding suitable for downstream *emotion prediction* and *anomaly detection* (e.g., high reconstruction error flags atypical/distress calls).

H. Inference Pipeline

At test time, a clip passes through: waveform → log-mel → $E_\theta(\cdot)$ to obtain z, and the head h_ψ outputs the prediction. We also compute reconstruction error $\|\hat{x} - x\|$ as a proxy for atypicality, which can highlight unusual behavior or elevated stress.

X. Visualization Interface

A web-based platform for automating denoising provides a dashboard for interactive exploration of audio features.

A. Features

- Upload and playback of original and denoised recordings.
- Spectrogram comparison (Original, STFT, Demucs, UMX).
- Feature visualization (RMS, bandwidth, centroid, noise floor).
- Autoencoder latent space projection (t-SNE / PCA visualization) [next release]

B. Implementation

The frontend communicates with a FastAPI backend through REST APIs. Audio data, feature statistics, and spectrograms are returned as JSON objects, enabling live rendering through Chart.js and Canvas components.

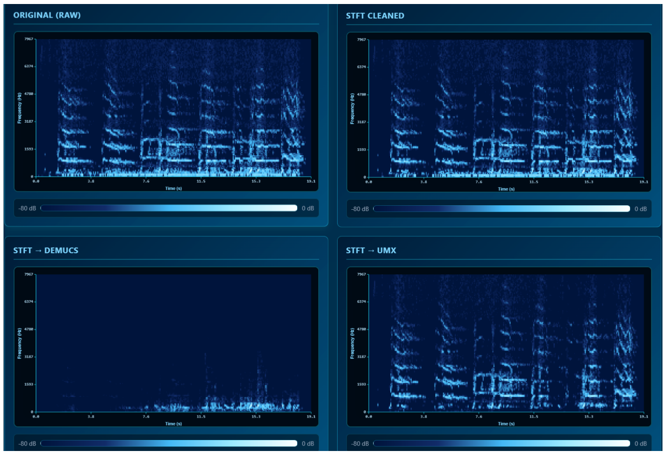

Fig. 3: Example spectrogram visualization comparing original and denoised whale call.

XI. Code Quality and Reproducibility

All code is version-controlled via GitHub. The repository follows a modular structure:

- `StateofartNotebooks/` : state of the art notebooks for exploratory purposes, analytics and ideation.
- `api/` : FastAPI backend and processing pipeline
- `whale-frontend/` : React visualization dashboard
- `README.md` : Setup and usage documentation

Best practices include:

- Docstrings and inline comments
- Type annotations
- Requirements file for dependencies
- CORS-enabled REST API for modular front-end communication

XII. Results and Discussion

Preliminary results show that denoised calls reconstructed by the autoencoder preserve biologically meaningful frequency contours while removing background noise. Latent features reveal distinct clusters corresponding to different call types and emotional contexts (e.g., distress vs contact calls).

XIII. Future Work

Future directions include:

- Extending the architecture to Variational Autoencoders (VAE) for probabilistic emotion modeling.
- Integrating metadata such as location and environmental conditions.
- Expanding to multimodal data (hydrophone arrays, motion sensors).
- Deploying real-time detection for vessel-whale interaction zones.

XIV. Conclusion

This work presents a comprehensive approach to whale call analysis through denoising and deep unsupervised learning. Autoencoders, compared with PCA, offer superior nonlinear representations of acoustic-emotional structure. The combination of advanced denoising, autoencoding, and web-based visualization tools paves the way for automated whale emotion inference and improved marine conservation strategies.

Acknowledgments

We acknowledge the support of open-source datasets from NOAA and DCLDE initiatives, and the developers of Demucs and Open-Unmix frameworks.

References

[1] A. Défossez *et al.*, "Hybrid Transformers for Music Source Separation," Meta AI Research, 2023.

[2] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix: A Reference Implementation for Music Source Separation," *Proc. ISMIR*, 2019.

[3] D. K. Mellinger *et al.*, "A Method for Detecting and Classifying Marine Mammal Sounds," *J. Acoust. Soc. Am.*, 2004.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[5] NOAA Fisheries, "Bioacoustics Archives." https://www.fisheries.noaa.gov

[6] Author(s), "Title," *arXiv preprint arXiv:2410.02560*, 2024.