

Specific Loudness Features for Sound Event Classification and Localization in Machine Learning

Adam Loper

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Committee:

Kaibao Nie, PhD

Tadesse Ghirmai, PhD

Pierre D. Mourad, PhD

Javier Herrero, PhD

Program Authorized to Offer Degree:
Electrical & Computer Engineering

©Copyright 2026

Adam Loper

University of Washington

Abstract

Specific Loudness Features for
Sound Event Classification and
Localization in Machine Learning

Adam Loper

Chair of the Supervisory Committee:
Associate Teaching Professor Kaibao Nie, PhD
Electrical Engineering, University of Washington-Bothell

This thesis investigates specific loudness representations as input features for machine learning systems that classify and localize sound events. Mel spectrograms are the de facto standard in audio machine learning, yet they are an *incomplete* perceptual transform: they warp the frequency axis to match human hearing but leave the magnitude axis in raw log-power, omitting the equal-loudness weighting, spectral masking, temporal integration, and compressive loudness growth of the auditory system. This work completes the transform by employing the ISO 532-1 (Zwicker) and ISO 532-3 (Moore-Glasberg-Schlittenlacher) specific loudness models, which map both the frequency and magnitude axes into perceptual units.

The thesis is primarily a study of sound event classification. On the ESC-50 benchmark, specific loudness features significantly outperform both linear STFT and mel spectrograms under an identical CNN, with accuracies of 39.7% for STFT, 53.1% for mel, 59.5% for ISO 532-1, and 63.0% for ISO 532-3. A magnitude-axis ablation localizes the cause: applying a perceptual transform to the magnitude of an ordinary mel spectrogram—a simple power-law compression, optionally combined with a fixed equal-loudness weighting, on the *same* mel filterbank—recovers most of the gain at essentially mel-level cost and without computing the full ISO 532 loudness model. Because these variants hold the mel frequency scale fixed and vary only the magnitude mapping, they isolate the perceptual transform of

the *magnitude* axis, rather than the auditory frequency scale, as the primary driver. The advantage replicates across two further datasets: UrbanSound8K, a second environmental-sound-event benchmark, and the TAU Urban Acoustic Scenes corpus, where it holds at roughly 14–16 percentage points over mel and extends even to acoustic scene classification, a task outside the sound-event framing of the thesis statement.

A second study extends the investigation to spatial audio as an exploratory probe. In a Sound Event Localization and Detection (SELD) experiment on STARSS23, 2-channel binaural specific loudness does *not* match the published 4-channel audio-only first-order ambisonics baseline ($ER = 0.97$). Instead, the ISO 532-3 model behaves as a more conservative detector: it obtains a lower error rate than our own FOA reproduction, but with lower F-score and localization recall. A multi-threshold analysis indicates that this lower error rate reflects conservative detection—fewer false alarms—rather than better localization, and that the F-score deficit persists even when direction of arrival is ignored. Classification, not localization, is the decisive result of the thesis.

Together, these results support the thesis that for sound event classes defined by human perception, perceptual input representations that include level-dependent spectral weighting align more strongly with classification targets than representations based solely on physical signal properties or partial perceptual transforms.

TABLE OF CONTENTS

	Page
List of Figures	iii
Glossary	v
Chapter 1: Introduction	1
1.1 Motivation	1
1.2 The Missing Dimension	2
1.3 Thesis Statement	4
1.4 Contributions	4
1.5 Thesis Organization	5
Chapter 2: Background and Literature Review	6
2.1 Psychoacoustic Frequency Scales	6
2.2 Loudness Models	7
2.3 Audio Features in Machine Learning	14
2.4 Sound Event Detection and Classification	16
2.5 Research Gap	16
Chapter 3: Classification Experiment: ESC-50	18
3.1 Experimental Design	18
3.2 CNN Architecture	19
3.3 Results	20
3.4 Per-Class Analysis	26
3.5 Cross-Architecture Generalization	37
3.6 Cross-Dataset Generalization	44
3.7 Discussion of ESC-50 Results	46

Chapter 4: SELD Experiment: STARSS23	48
4.1 Sound Event Localization and Detection	48
4.2 Binaural Hearing and Spatial Cues	53
4.3 Experimental Design	54
4.4 Dataset and Classes	55
4.5 Binaural Conversion Pipeline	55
4.6 Model Architecture and Training	62
4.7 Results: Standard 20° Threshold	64
4.8 Multi-Threshold Analysis	66
4.9 Detection-Only Class-Wise Analysis	70
Chapter 5: Discussion	72
5.1 The Loudness Advantage: The Magnitude-Axis Transform	72
5.2 Which Classes Benefit	73
5.3 Detection Versus Localization	74
5.4 Pipeline Alignment	75
5.5 The Role of Binaural Inhibition	77
5.6 Mel Adoption as Path of Least Resistance	77
5.7 Annotation Methodology	79
Chapter 6: Conclusion and Future Work	81
6.1 Summary of Findings	81
6.2 Future Work	82
Appendix A: Software and Reproducibility	84
A.1 Software Environment	84
A.2 Hardware	86
A.3 Validation	86
A.4 Code Availability	86
Bibliography	87

LIST OF FIGURES

Figure Number	Page
1.1 The perception-classification loop. Sound events are perceived and categorized by human listeners; these categories become training labels. The choice of input features determines how closely the ML system’s representation matches the perceptual transform that generated the labels.	3
2.1 Comparison of perceptual frequency scales	7
2.2 Block diagrams of the ISO 532-1 and ISO 532-3 loudness models	8
2.3 Equal-loudness contours	11
2.4 Perceptual alignment of audio features	13
2.5 Perceptual processing hierarchy	15
3.1 Feature representation comparison	19
3.2 ESC-50 classification accuracy comparison	21
3.3 Cross-fold variance comparison	24
3.4 Spectral comparison: crickets	28
3.5 Spectral comparison: thunderstorm	29
3.6 Spectral comparison: church bells	30
3.7 Spectral comparison: clock tick	31
3.8 Spectral comparison: washing machine	32
3.9 Spectral comparison: clock alarm	33
3.10 Per-class accuracy heatmap	34
3.11 Confusion matrix: mel spectrogram	35
3.12 Confusion matrix: ISO 532-3	36
3.13 How each architecture parses the input	42
3.14 Cross-architecture accuracy comparison	43
4.1 Ambisonics sound field capture and reconstruction	52
4.2 FOA to binaural conversion pipeline	56
4.3 Multi-threshold localization concept	68

4.4	Multi-threshold SELD analysis	69
4.5	Detection-only error rate comparison	71
5.1	Feature extraction computation cost	79

GLOSSARY

ADPIT: Auxiliary Duplicating Permutation Invariant Training, a loss function for multi-source SELD.

BARK: A psychoacoustic frequency scale proposed by Zwicker (1961), used in ISO 532-1. Divides the audible range into 24 critical bands.

BINAURAL: Relating to or involving both ears. Binaural audio preserves interaural level and time differences that encode spatial information.

DCASE: Detection and Classification of Acoustic Scenes and Events, an annual challenge and workshop series for audio classification research.

DOA: Direction of Arrival, the spatial angle from which a sound originates relative to the listener or microphone array.

ERB: Equivalent Rectangular Bandwidth, a psychoacoustic frequency scale proposed by Glasberg and Moore (1990). Used in ISO 532-3 with 150 bands from 1.75 to 39.0 Cam.

ESC-50: Environmental Sound Classification dataset containing 2,000 five-second recordings across 50 classes.

FOA: First-Order Ambisonics, a 4-channel spatial audio format encoding sound field pressure and particle velocity.

HRTF: Head-Related Transfer Function, describes how sound is filtered by the head, pinnae, and torso before reaching the eardrums.

ILD: Interaural Level Difference, the difference in sound pressure level between the two ears, a primary cue for lateral sound localization.

ISO 532-1: International standard specifying the Zwicker method for calculating loudness using the Bark scale.

ISO 532-3: International standard specifying the Moore-Glasberg-Schlittenlacher method for calculating loudness using the ERB scale, including binaural inhibition.

ITD: Interaural Time Difference, the difference in arrival time of a sound between the two ears.

MEL: A perceptual frequency scale proposed by Stevens et al. (1937), mapping perceived pitch to measured frequency. Widely used in audio machine learning via mel spectrograms.

SELD: Sound Event Localization and Detection, the joint task of identifying what sound events are present, when they occur, and where they are located in space.

PHON: The unit of loudness level. The loudness level of a sound, in phons, is the sound pressure level of a frontally incident 1 kHz sinusoidal plane wave, presented binaurally, that is judged by otologically normal persons as being equally loud as the given sound.

SONE: The unit of perceived loudness. Loudness is the perceived magnitude of a sound as estimated by otologically normal persons, depending primarily on sound pressure level but also on frequency, waveform, bandwidth, and duration. One sone is the loudness of a sound with a loudness level of 40 phon. A sound judged twice as loud corresponds to twice the number of sones.

SPECIFIC LOUDNESS: Loudness per unit bandwidth (sone/Bark or sone/Can), representing how loudness is distributed across frequency.

STARSS23: Sony-TAU Realistic Spatial Soundscapes 2023, a dataset for audio-visual SELD containing multichannel audio with spatiotemporal annotations.

STFT: Short-Time Fourier Transform, a spectral analysis method that decomposes a signal into frequency components over time.

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to the University of Washington Bothell, and in particular to Dr. Kaibao Nie for his guidance throughout this research. The author also thanks the reading committee—Dr. Tadesse Ghirmai, Dr. Pierre D. Mourad, and Dr. Javier Herrero—for their time and feedback. Special thanks to the acoustics community for the open-source tools that made this work possible.

Chapter 1

INTRODUCTION

1.1 Motivation

Sound event classification and localization are fundamentally perceptual tasks. The classes that machine learning systems are trained to recognize, *laughter*, *footsteps*, *door slam*, *music*, are not defined by their physical waveform properties but by how human listeners perceive and categorize them. A door closing produces a broadband transient; a laugh is quasi-periodic with formant structure. Yet the labels *door* and *laughter* are products of auditory cognition, not of acoustics. They are what cognitive science calls *folk categories*: natural groupings shaped by perceptual experience rather than by measurement [1].

Despite this perceptual origin, the dominant input representation in audio machine learning is the mel spectrogram. The mel scale, introduced by Stevens, Volkmann, and Newman in 1937 [2], maps measured frequency to perceived pitch by warping the frequency axis. Implemented in widely used libraries such as `librosa` [3], mel spectrograms have become the de facto standard for tasks ranging from speech recognition to environmental sound classification to sound event localization and detection (SELD).

However, the mel spectrogram applies only a *perceptual transform of the frequency axis*. It does not apply one to the *magnitude axis*. The relationship between physical sound pressure level and perceived loudness is highly nonlinear: it depends on frequency (via equal-loudness contours), on temporal context (via temporal integration and forward masking), and on spectral context (via simultaneous masking and excitation pattern spreading). These phenomena are well characterized by standardized loudness models, in particular ISO 532-1 [4] (the Zwicker method) and ISO 532-3 [5] (the Moore-Glasberg-Schlittenlacher method).

The specific loudness representation produced by these models, loudness per critical band

as a function of time, captures both the frequency and the level dimensions of human perception. It is, in effect, a more complete perceptual transform than the mel spectrogram.

1.2 *The Missing Dimension*

The mel spectrogram can be understood as a partial perceptual transform: it addresses the frequency axis but ignores the level axis. Figure 1.1 illustrates the perception-classification loop that motivates this work. Sound events originate in the physical world, are transformed by human perception into internal representations, and are then categorized into folk classes. If machine learning systems are trained to reproduce these perceptual categories, their input representations should approximate the perceptual transform as closely as possible.

The mel scale itself has known limitations. The original 1937 measurements were based on a small number of subjects using magnitude estimation, a psychophysical method with well-documented biases. Two alternative psychoacoustic frequency scales have since been developed with more rigorous methodologies:

- The **Bark scale** (Zwicker, 1961 [6]), based on critical-band measurements, divides the audible range into 24 critical bands; a discrete filterbank may sample this range more finely (here at 0.1 Bark, giving 240 bands).
- The **ERB scale** (Glasberg & Moore, 1990 [7]), based on notched-noise masking experiments, provides a continuous mapping spanning approximately 39 Cam—the ERB-number unit—across the audible range, sampled here at 0.25 Cam to give 150 bands.

Both are considered more accurate representations of auditory frequency selectivity than the mel scale. The dominance of mel in audio ML is attributable to software availability, particularly `librosa` [3], rather than to any systematic evaluation demonstrating its superiority over Bark or ERB representations. A review of DCASE 2023 Challenge submissions carried out for this thesis found that the overwhelming majority employ mel spectrograms as their primary audio representation, typically without comparison against alternatives.

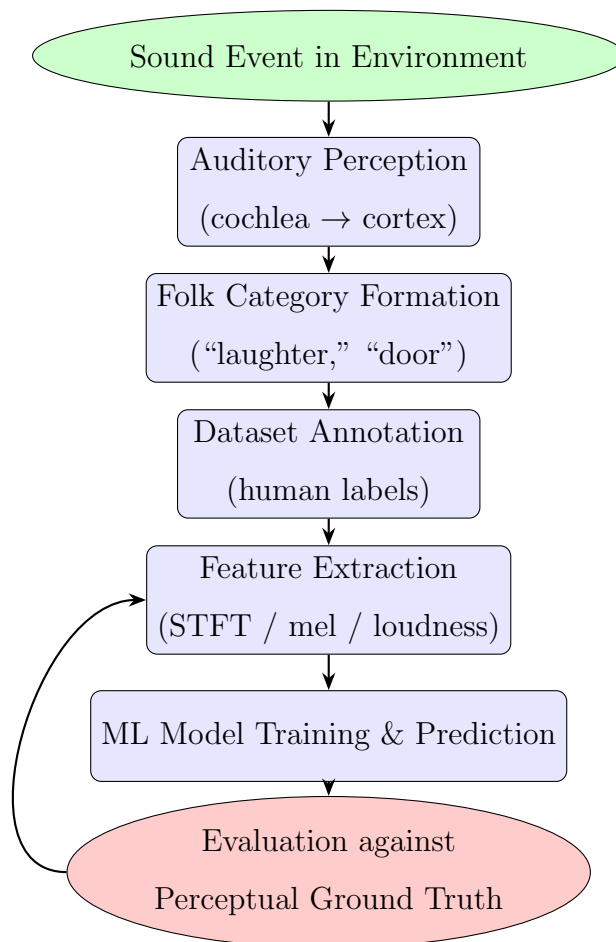


Figure 1.1: The perception-classification loop. Sound events are perceived and categorized by human listeners; these categories become training labels. The choice of input features determines how closely the ML system’s representation matches the perceptual transform that generated the labels.

1.3 Thesis Statement

For sound event classes defined by human auditory perception, perceptual input representations that include level-dependent spectral weighting align more strongly with classification targets than representations based solely on physical signal properties or partial perceptual transforms.

1.4 Contributions

This thesis makes the following contributions:

1. A controlled comparison of STFT, mel spectrogram, and ISO 532-1 and ISO 532-3 specific loudness features on the ESC-50 environmental sound classification benchmark, using the same CNN architecture and training procedure, and replicated on UrbanSound8K and TAU Urban Acoustic Scenes and across four model architectures (Chapter 3).
2. A magnitude-axis ablation isolating the perceptual transform of the *magnitude* axis—rather than the auditory frequency scale—as the primary driver of the gain, and showing that lightweight power-law and equal-loudness variants recover most of it at essentially mel-level cost (Chapter 3).
3. A SELD study on the STARSS23 dataset comparing 4-channel First-Order Ambisonics (FOA) features with 2-channel binaural specific loudness from ISO 532-1 and ISO 532-3 (Chapter 4).
4. A multi-threshold analysis that separates detection from localization, showing that the loudness model’s lower error rate reflects more conservative detection—fewer false alarms—rather than better localization (Chapter 4).
5. Per-class analysis identifying the acoustic and perceptual properties that predict loudness feature advantage (Chapters 4 and 3).

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews the psychoacoustic foundations of frequency scales and loudness models, surveys the evolution of audio features in machine learning, and identifies the gap that this work addresses. Chapter 3 presents the classification experiment on ESC-50, establishing the magnitude-axis benefit of specific loudness in a single-channel setting. Chapter 4 extends the investigation to spatial audio with a SELD experiment on STARSS23, introducing the binaural rendering pipeline and a multi-threshold analysis that separates detection from localization. Chapter 5 synthesizes the findings across both experiments. Chapter 6 summarizes the contributions and outlines future work.

Chapter 2

BACKGROUND AND LITERATURE REVIEW

2.1 *Psychoacoustic Frequency Scales*

Three psychoacoustic frequency scales are relevant to this work, each mapping physical frequency to a perceptual correlate of auditory frequency selectivity (Figure 2.1).

The **mel scale** [2], the oldest and most widely adopted in audio ML, approximates perceived pitch as $m = 2595 \log_{10}(1 + f/700)$. It is approximately linear below 1 kHz and logarithmic above. Its dominance in audio ML is attributable to software availability (**librosa** [3]) rather than demonstrated perceptual superiority.

The **Bark scale** [6] divides the audible range into 24 critical bands based on the ear’s frequency selectivity. It underlies the Zwicker loudness model (ISO 532-1).

The **ERB scale** [7], derived from notched-noise masking experiments, provides finer resolution with approximately 150 bands across the audible range. The Equivalent Rectangular Bandwidth is $ERB(f) = 24.7(4.37f/1000 + 1)$ Hz. The ERB scale underlies the Moore-Glasberg-Schlittenlacher loudness model (ISO 532-3).

All three scales compress high frequencies relative to low frequencies, reflecting the ear’s broader bandwidth at higher frequencies. Both Bark and ERB are considered more accurate representations of auditory frequency selectivity than mel, being derived from direct masking measurements rather than pitch-matching experiments.

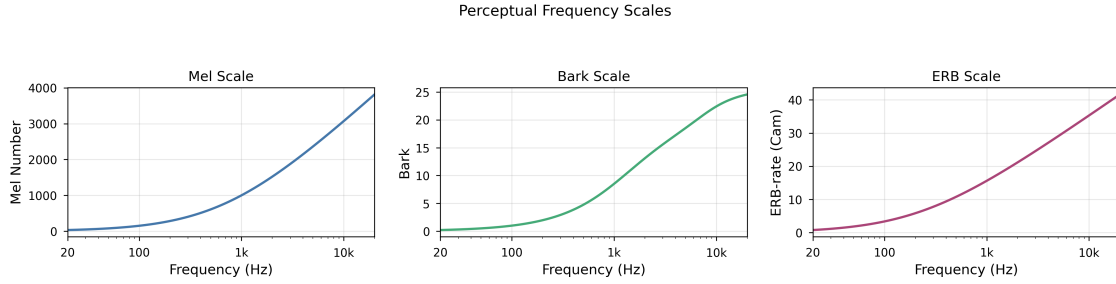


Figure 2.1: Comparison of mel, Bark, and ERB frequency scales showing how each warps the linear frequency axis differently. All three compress high frequencies relative to low frequencies, but differ in their specific mappings.

2.2 Loudness Models

Both standardized loudness models share a common processing chain—a peripheral filter-bank, an excitation stage that models masking, a compressive transformation to specific loudness, and integration to total loudness—and differ mainly in their frequency resolution, filter shapes, temporal dynamics, and, for ISO 532-3, an explicit binaural stage. Figure 2.2 places the two pipelines side by side; the subsections below describe each in turn.

2.2.1 ISO 532-1: The Zwicker Method

The Zwicker loudness model, standardized as ISO 532-1 [4], computes loudness in the following stages:

1. **Spectral analysis:** The input signal is decomposed into third-octave bands.
2. **Outer and middle ear transfer:** Each band level is adjusted by the transmission characteristics of the outer and middle ear.
3. **Excitation pattern:** The adjusted band levels are transformed into an excitation pattern on the Bark scale, incorporating upward and downward spread of masking.



Figure 2.2: High-level block diagrams of the two loudness models. Both share a perceptual-transform backbone (dashed box): an outer/middle-ear filter, an auditory filterbank, an excitation pattern that models spread of masking, a compressive *specific-loudness* transform (highlighted), and temporal integration. ISO 532-1 (Zwicker) uses third-octave analysis mapped to 24 rectangular Bark bands; ISO 532-3 (Moore–Glasberg–Schlittenlacher) uses ~150 ERB roex filters and adds a binaural-inhibition stage in which the louder ear masks the quieter, encoding the interaural level difference. The highlighted specific-loudness block is the magnitude-axis transform central to this thesis.

4. **Specific loudness:** The excitation pattern is converted to specific loudness (sone/Bark) using a compressive power law that implements the equal-loudness contour weighting.
5. **Total loudness:** Integration of specific loudness across all 24 Bark bands yields total loudness in sone.
6. **Temporal integration:** For time-varying sounds, a temporal integration stage with attack and release time constants produces a time-varying loudness trace.

The specific-loudness stage (stage 4) is the compressive, level-dependent transformation at the heart of the model. Zwicker maps the excitation E in each critical band to specific loudness

$$N'(z) = 0.08 \left(\frac{E_{\text{TQ}}}{E_0} \right)^{0.23} \left[\left(0.5 + 0.5 \frac{E}{E_{\text{TQ}}} \right)^{0.23} - 1 \right] \quad (\text{sone/Bark}), \quad (2.1)$$

where E_{TQ} is the excitation at the threshold in quiet and E_0 a reference excitation [4]. The exponent 0.23 makes the mapping strongly compressive: a tenfold (10 dB) rise in excitation produces only about a $10^{0.23} \approx 1.7$ -fold rise in specific loudness. This compression, together with the frequency-dependent threshold term E_{TQ} and the outer- and middle-ear transfer applied at stage 2, produces the level- and frequency-dependent weighting that distinguishes specific loudness from a log-power spectrogram. Total loudness follows by integrating specific loudness across the tonotopic axis,

$$N = \int_0^{24 \text{ Bark}} N'(z) \, dz \quad (\text{sone}), \quad (2.2)$$

so that each band contributes in proportion to its perceptual rather than its physical magnitude.

The specific loudness representation at stage 4 is a two-dimensional time-frequency representation analogous to a spectrogram but expressed in perceptual units (sone/Bark vs. time). Critically, it encodes both a perceptual transform of the frequency axis (via the Bark scale) and of the magnitude axis (via equal-loudness contour weighting and the compressive transformation).

2.2.2 ISO 532-3: The Moore-Glasberg-Schlittenlacher Method

The Moore-Glasberg-Schlittenlacher loudness model, standardized as ISO 532-3 [5], represents the state of the art in loudness computation. Its key differences from the Zwicker method are:

1. **Frequency resolution:** Uses the ERB scale with approximately 150 bands (1.75 to 39.0 Cam), providing substantially finer frequency resolution than the 24 Bark bands of ISO 532-1.
2. **Auditory filter model:** Employs the roex (rounded exponential) filter shape derived from notched-noise experiments, rather than the rectangular critical band assumption of Zwicker.
3. **Binaural inhibition:** Includes a binaural inhibition stage that models the interaction between left and right ear signals. When the loudness differs between ears, the louder ear partially inhibits the contribution of the quieter ear to overall loudness. This mechanism encodes spatial information via interaural level differences (ILD).
4. **Temporal processing:** Uses a temporal integration stage with separate attack and release time constants, producing short-term and long-term loudness.

The specific-loudness transformation takes an analogous but refined form. Between the threshold in quiet and roughly 100 dB SPL, the Moore-Glasberg model computes

$$N'(z) = C [(G E + A)^\alpha - A^\alpha], \quad (2.3)$$

where $\alpha \approx 0.2$ is the compression exponent, C a scaling constant, and G and A frequency-dependent gain and offset parameters that implement the equal-loudness weighting [5]. As in the Zwicker model, the sub-unity exponent compresses the wide physical dynamic range into the far narrower perceptual range presented to the network.

The binaural inhibition stage is particularly relevant for SELD, as it provides a physiologically motivated encoding of spatial information that complements the interaural level difference cues preserved by separate left/right channel processing.

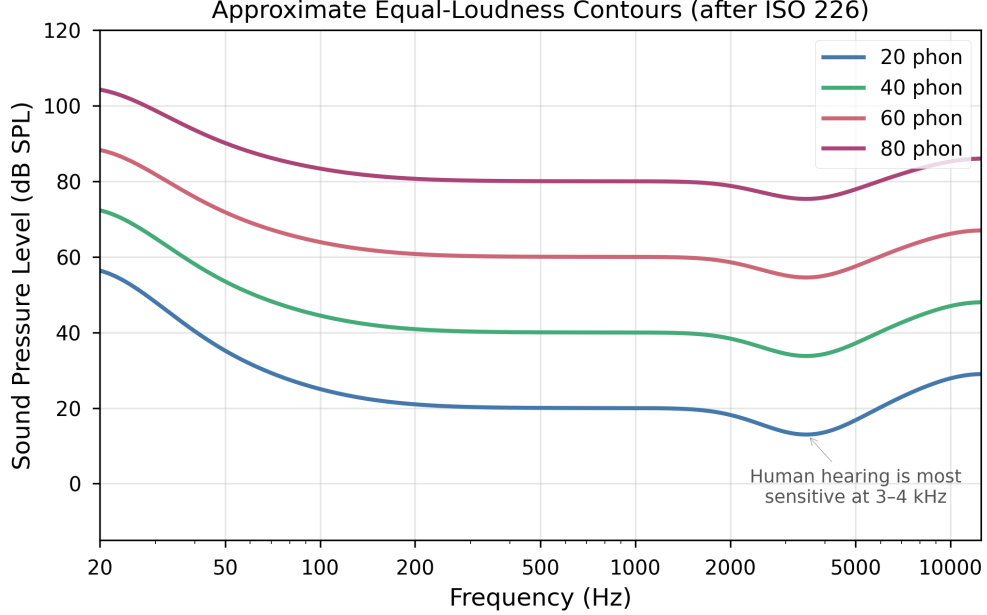


Figure 2.3: Approximate equal-loudness contours [8] illustrating the level-dependent frequency sensitivity of human hearing. At low levels (20 phon), perceived loudness drops dramatically at low and high frequencies. This level-dependent spectral reshaping is captured by loudness models but not by mel spectrograms.

2.2.3 The Missing Dimension Argument

Sound-event classes are defined by human perception, and their labels are judged by perception, so perception brackets the whole task. The true perceptual transform Φ maps the acoustic signal x_t to the filtered experience from which the dataset labels y are formed. A model instead applies an *estimate* $\hat{\Phi}$ to a measured signal M ; the STFT, mel, and specific-loudness front ends are estimates of increasing perceptual alignment (Figure 2.4). A representation that enters the perceptual domain (N') shares the domain in which the labels were

made; a physical one (STFT) is a misaligned proxy.

The single perceptual step the mel spectrogram omits is the transform of the magnitude axis—the “missing dimension” that this thesis investigates.

Two caveats bound this argument. It holds only for perceptually defined tasks: where the target is defined at the signal level—ultrasonic detection, say, whose content lies outside the audible range a loudness model preserves—a perceptual front end instead discards the information the labels depend on. And alignment does not cap a model at human accuracy; it may use the preserved cues more reliably than a listener, though it cannot recover what the transform has already removed.

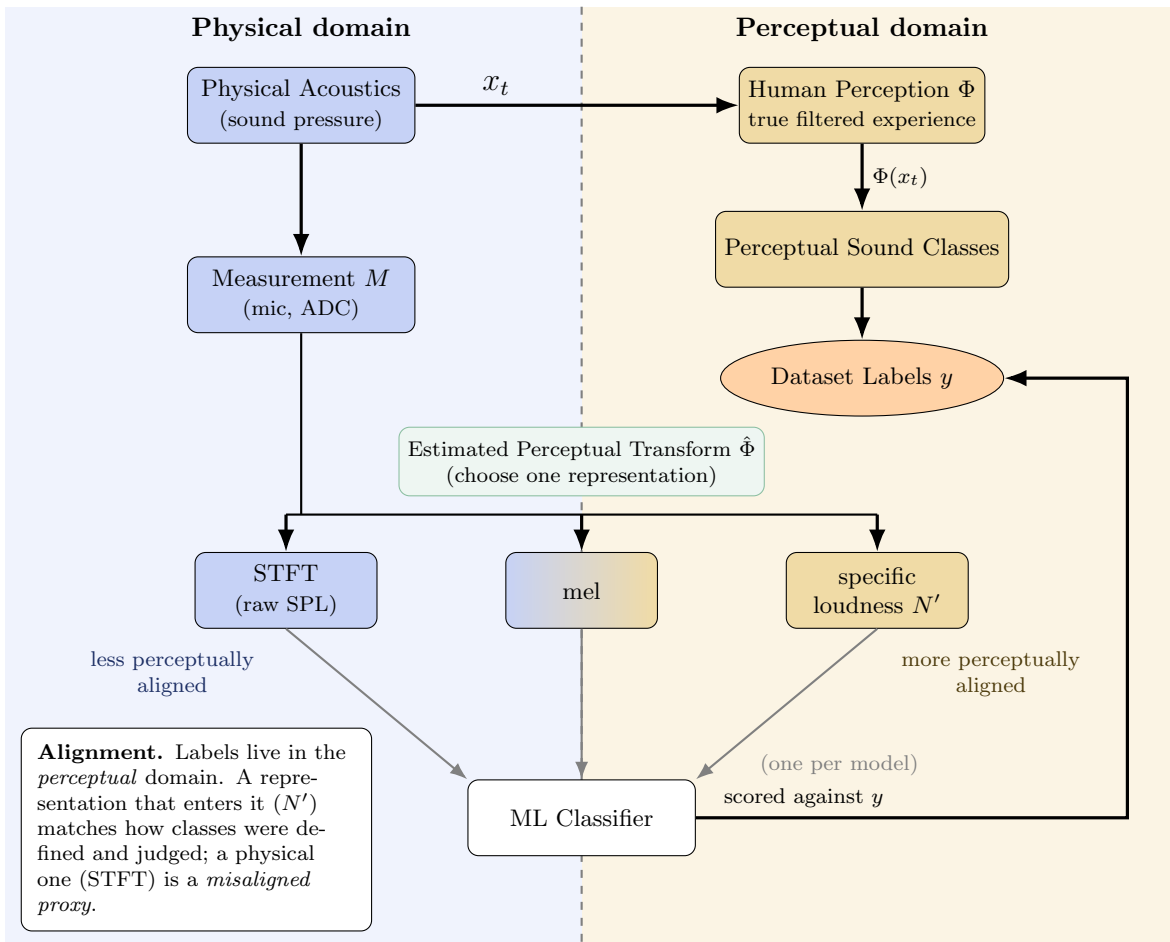


Figure 2.4: Audio front ends as estimates $\hat{\Phi}$ of the perceptual transform Φ that generates the labels, arranged from physical (STFT) to perceptual (specific loudness N').

2.3 Audio Features in Machine Learning

2.3.1 Mel Spectrogram Dominance

The mel spectrogram has become the near-universal front end for audio machine learning. On the ESC-50 benchmark, mel-based systems have achieved progressively higher accuracy through architectural advances: from the original 64.5% reported by Piczak [9] using a two-layer CNN, to over 90% with large-scale pretrained systems such as Pretrained Audio Neural Networks (PANNs) [10] and the Audio Spectrogram Transformer (AST) [11]. However, these improvements come from architectural scaling, pretraining on AudioSet (2 million clips), and data augmentation, not from changing the input representation. The mel spectrogram itself has remained essentially unchanged since its adoption.

In the DCASE challenge community, mel spectrograms dominate submissions across all tasks. A review of DCASE 2023 Task 3 (SELD) submissions carried out for this thesis found that the overwhelming majority use mel or log-mel spectrograms, with no submissions using loudness-based representations. Alternative features such as gammatone filterbanks [12] have shown benefits for specific tasks [13] but have not displaced mel as the default.

2.3.2 Feature Representation Hierarchy

The evolution of audio features in machine learning follows a trajectory from raw waveforms toward increasingly perceptual representations:

1. **Raw waveform:** Direct input of time-domain samples. Used in some end-to-end architectures [14] but requires the model to learn spectral decomposition.
2. **STFT magnitude:** The magnitude of the short-time Fourier transform provides a linear-frequency time-frequency representation, here compressed with $\log(1+x)$ rather than mapped to perceptual units.
3. **Mel spectrogram:** Applies a perceptual transform of the frequency axis via the mel

filterbank, typically with log-power compression. The current dominant representation in audio ML [3].

4. **Power-law-compressed features:** Representations such as perceptual linear prediction (PLP) [15] and power-normalized cepstral coefficients (PNCC) [16] replace logarithmic magnitude compression with a compressive power law motivated by Stevens' law of loudness, transforming the magnitude rather than the frequency axis.
5. **Gammatone features:** Uses a gammatone filterbank that more closely models peripheral auditory filtering than the mel filterbank [12].
6. **Specific loudness:** The representation investigated in this thesis. Applies a perceptual transform to both the frequency and magnitude axes via standardized loudness models [4, 5].

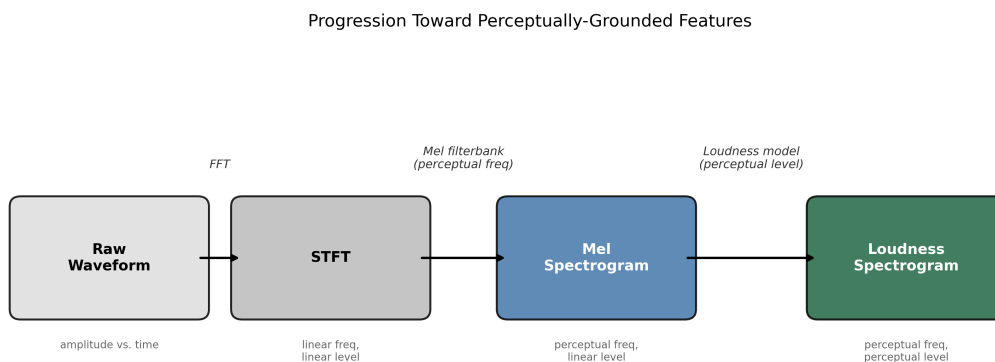


Figure 2.5: The progression of audio feature representations from raw waveform to specific loudness. Each step adds perceptual processing, with loudness models completing the transform by adding level-dependent spectral weighting.

2.4 Sound Event Detection and Classification

2.4.1 DCASE Challenges

The Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series [17] has been instrumental in advancing audio classification research. Beginning in 2013, the annual challenges have driven development of increasingly sophisticated architectures and feature representations for tasks including acoustic scene classification, sound event detection, and, since 2019, sound event localization and detection (SELD) [18].

2.4.2 ESC-50

The Environmental Sound Classification dataset (ESC-50) [9] contains 2,000 five-second recordings organized into 50 classes across five major categories: animals, natural soundscapes, human non-speech sounds, interior/domestic sounds, and exterior/urban noises. The dataset uses a predefined 5-fold cross-validation scheme. ESC-50 is widely used as a benchmark for environmental sound classification and provides a controlled setting for comparing feature representations.

2.4.3 Architectures

Convolutional neural networks (CNNs) have become the standard architecture for audio classification [19]. Typical architectures consist of stacked convolutional layers with batch normalization and max-pooling, followed by global average pooling and fully connected classification layers.

2.5 Research Gap

Despite the theoretical appeal of specific loudness as a more complete perceptual representation, no controlled comparison of loudness features against STFT and mel spectrogram features has been published for sound event classification. Prior work on loudness in audio

ML has been limited to loudness-based augmentation strategies [20] or loudness normalization [21], rather than using specific loudness as a primary input representation.

This thesis addresses two gaps in two stages. Chapter 3 evaluates specific loudness as a drop-in replacement for mel spectrograms in classification on the ESC-50 benchmark. Chapter 4 then extends the investigation to spatial audio, where a separate gap exists in the SELD literature: that gap is introduced in the opening of Chapter 4, alongside the binaural-hearing background it depends on.

Chapter 3

CLASSIFICATION EXPERIMENT: ESC-50

3.1 *Experimental Design*

This chapter presents a controlled classification experiment comparing four input representations on the ESC-50 dataset [9]; two further lightweight variants are introduced in the magnitude-axis ablation of Section 3.3.2:

1. **STFT**: Linear-frequency magnitude spectrogram, 257 frequency bins from a 512-point transform, compressed as $\log(1 + |X|)$ ($1 \times 257 \times 128$).
2. **Mel spectrogram**: 240-band mel spectrogram computed from a 2048-point transform, compressed as $\log(1 + \text{mel power})$ ($1 \times 240 \times 128$).
3. **ISO 532-1 specific loudness**: 240-band specific loudness computed using the Zwicker method on the Bark scale ($1 \times 240 \times 128$).
4. **ISO 532-3 specific loudness**: 150-band specific loudness computed using the Moore-Glasberg-Schlittenlacher method on the ERB scale ($1 \times 150 \times 128$).

All four representations were computed from the same source audio files. The mel and ISO 532-1 representations share the same dimensionality (240×128) to ensure that any performance difference between them is attributable to the representation itself, not to dimensionality. ISO 532-3 uses 150 ERB bands as specified by the standard, providing a direct comparison between the Bark and ERB frequency scales while both include a perceptual transform of the magnitude axis. The STFT and mel front ends, by contrast, differ in transform size (512 versus 2048 points) and in whether magnitude or power is compressed, so

the STFT-to-mel step is not a single-factor contrast. The magnitude-axis ablation of Section 3.3.2, which holds the mel front end fixed and varies only the magnitude mapping, is the controlled comparison on which the thesis statement rests.

Feature Representations of a Dog Bark (ESC-50)

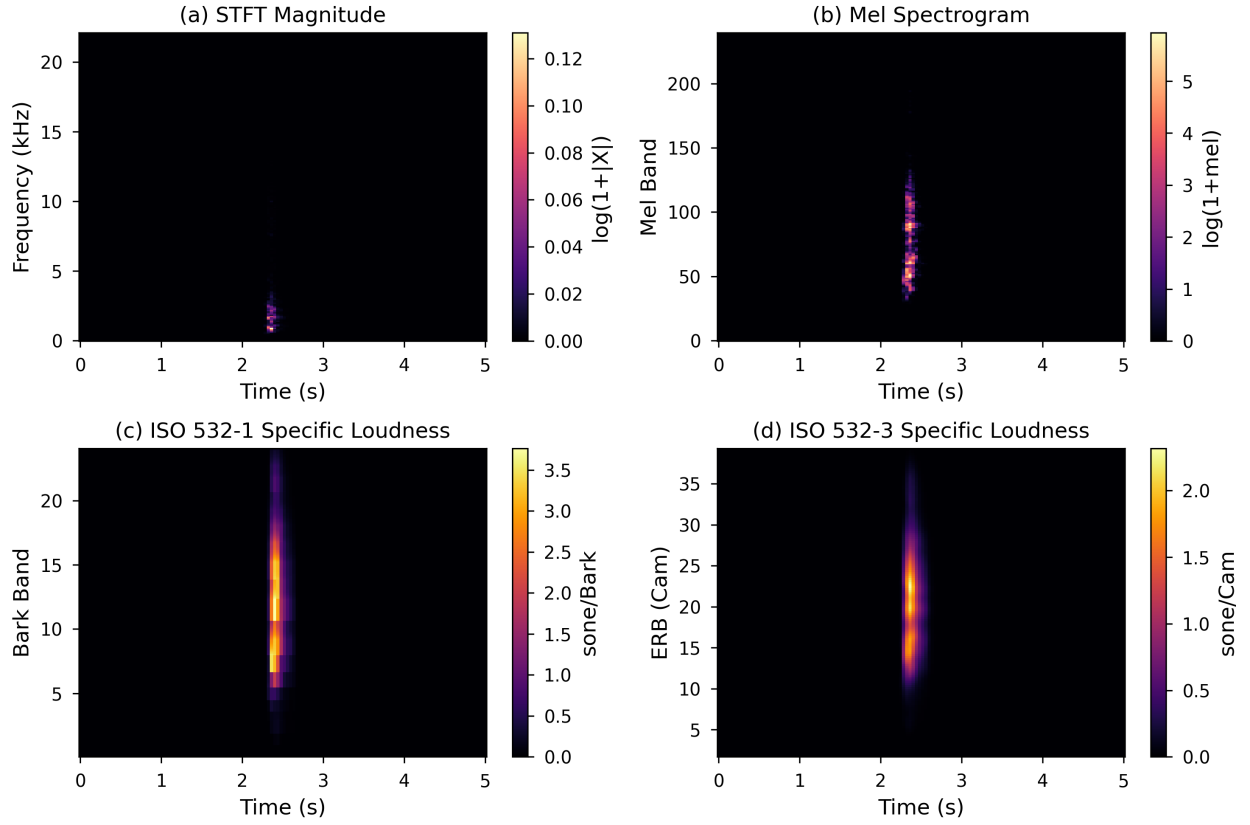


Figure 3.1: The same audio clip (dog bark from ESC-50) represented as four different features. Top-left: STFT (linear frequency). Top-right: mel spectrogram. Bottom-left: ISO 532-1 specific loudness (Bark scale). Bottom-right: ISO 532-3 specific loudness (ERB scale). Note the perceptual reshaping of the spectral content in the loudness representations.

3.2 CNN Architecture

A uniform CNN architecture was used for all feature types:

1. Three convolutional blocks, each consisting of:

- Conv2d layer with batch normalization
 - ReLU activation
 - Max pooling (2×2)
2. Adaptive average pooling to a fixed spatial size
 3. Fully connected layer with dropout
 4. Output layer with 50 units (one per class)

Feature normalization (zero mean, unit variance) was applied to all inputs. Training used early stopping with a patience of 15 epochs after a minimum of 30 training epochs, ensuring convergence without overfitting.

3.3 Results

Table 3.1 presents the 5-fold cross-validation results.

Table 3.1: ESC-50 classification accuracy (baseline CNN; 3 seeds $\times$ 5-fold cross-validation). Mel, “+ power law,” and “+ equal-loudness” share the same 240-band mel frequency scale and differ only in the magnitude-axis transform.

Feature	Freq. Scale	Accuracy (%)	Std (%)
STFT	Linear	39.7	± 9.7
Mel (240 bands)	Mel	53.1	± 3.2
Mel + power law	Mel	59.2	± 3.8
Mel + equal-loudness	Mel	61.9	± 4.0
ISO 532-1 Loudness	Bark	59.5	± 1.9
ISO 532-3 Loudness	ERB	63.0	± 1.3

The results show a clear progression driven by two distinct perceptual transforms. Moving from linear frequency (STFT, 39.7%) to a perceptual transform of the frequency axis (mel,

53.1%) yields a 13.4 percentage point improvement. Transforming the magnitude axis then adds a second, largely independent gain: a simple compressive power law (mel + power law, 59.2%) provides a further 6.1 points, and adding equal-loudness contour weighting (mel + equal-loudness, 61.9%) a further 2.7. Notably, these two lightweight magnitude transforms—which reuse the mel filterbank and add negligible cost—already match or exceed the full Zwicker loudness model (ISO 532-1, 59.5%). Only the state-of-the-art ISO 532-3 auditory model, with ERB-scale frequency resolution and refined temporal integration, does better still (63.0%). Section 3.3.2 dissects this magnitude-axis contribution in detail.

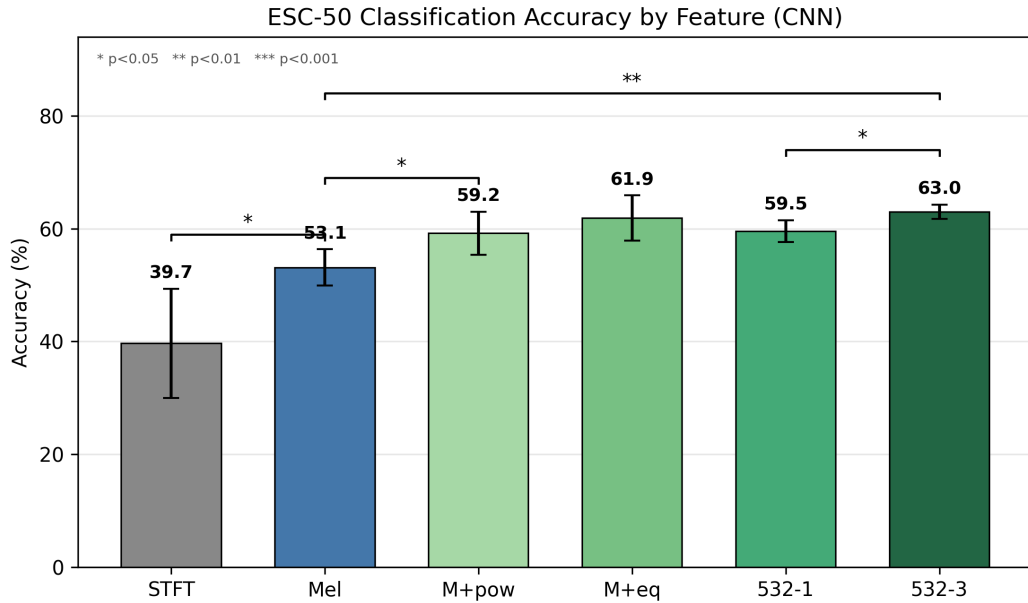


Figure 3.2: Overall classification accuracy on ESC-50 for six feature representations (baseline CNN). Error bars show standard deviation across 5 folds. Significance brackets indicate paired t -test results.

The full loudness features also exhibit lower variance across folds ($\pm 1.9\%$ and $\pm 1.3\%$ for 532-1 and 532-3) compared to STFT ($\pm 9.7\%$) and mel ($\pm 3.2\%$), indicating more consistent and robust feature quality.

3.3.1 Statistical Significance Analysis

To verify that the observed accuracy differences are not due to chance, paired t -tests were conducted across the five cross-validation folds. For each pair of features, the test compares the per-fold accuracy differences against the null hypothesis that the true difference is zero. Effect size is reported as Cohen’s d , computed as the mean difference divided by the standard deviation of the differences. By convention, $|d| > 0.8$ is considered a large effect.

Table 3.3 presents the full pairwise results. The per-fold accuracy values used in these tests are reported in Table 3.2 for reproducibility.

Table 3.2: Per-fold accuracy values used in statistical tests (5-fold CV; each value averaged over 3 seeds).

Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
STFT	0.424	0.381	0.268	0.536	0.377	0.397
Mel	0.496	0.554	0.496	0.557	0.552	0.531
Mel + power law	0.588	0.558	0.567	0.656	0.591	0.592
Mel + equal-loud.	0.588	0.569	0.662	0.652	0.625	0.619
532-1	0.573	0.576	0.608	0.610	0.610	0.595
532-3	0.622	0.625	0.628	0.652	0.623	0.630

Several observations emerge from this analysis:

1. **Both loudness models significantly outperform mel** (ISO 532-1: +6.4 pp, $p = 0.013$, $d = 1.92$; ISO 532-3: +9.9 pp, $p = 0.002$, $d = 3.37$), with large effect sizes comparable to the benefit of mel over STFT (+13.4 pp, $d = 1.58$).
2. **A cheap power law captures most of the magnitude-axis gain.** Replacing log-mel compression with a compressive power law (mel + power law) improves accuracy by +6.1 pp ($p = 0.026$, $d = 1.54$), and adding equal-loudness weighting (mel + equal-loudness) reaches 61.9%—which matches or exceeds the full Zwicker model (ISO 532-1,

Table 3.3: Pairwise paired t -tests across the 5 folds (each fold value averaged over 3 seeds). Δ Mean is the mean accuracy difference (first minus second), d is Cohen’s effect size. Significance level $\alpha = 0.05$.

Comparison	Δ Mean	t	p	Cohen’s d	Sig?
STFT vs Mel	−0.134	−3.54	0.024	−1.58 (large)	Yes
Mel vs Mel+power	−0.061	−3.45	0.026	−1.54 (large)	Yes
Mel+power vs Mel+eq. loud.	−0.027	−1.48	0.214	−0.66 (medium)	No
Mel+eq. loud. vs 532-1	+0.024	+2.22	0.091	+0.99 (large)	No
Mel vs 532-1	−0.064	−4.29	0.013	−1.92 (large)	Yes
Mel vs 532-3	−0.099	−7.54	0.002	−3.37 (large)	Yes
532-1 vs 532-3	−0.035	−4.59	0.010	−2.05 (large)	Yes

59.5%; the difference is not significant, $p = 0.09$). Most of the loudness benefit on mono audio is therefore available without the full ISO 532 computation; Section 3.3.2 develops this point.

- ISO 532-3 is the single best representation** and significantly exceeds ISO 532-1 (+3.5 pp, $p = 0.010$, $d = 2.05$). The refined ERB-scale resolution, roex auditory filters, and temporal integration of the Moore–Glasberg model thus provide a real, if modest, gain over the Zwicker method—correcting an earlier impression, from single-seed data, that the two were indistinguishable.
- The full loudness features exhibit the lowest variance.** Across folds, ISO 532-1 ($\pm 1.9\%$) and ISO 532-3 ($\pm 1.3\%$) are the most stable representations, versus mel ($\pm 3.2\%$) and STFT ($\pm 9.7\%$), indicating that the full auditory models produce feature representations that generalize most consistently across data splits.

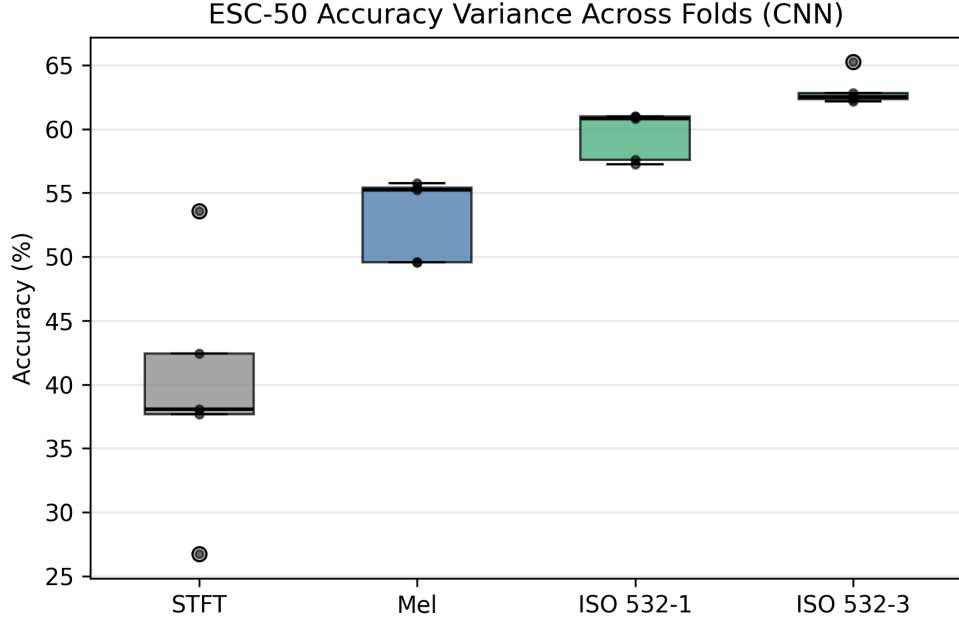


Figure 3.3: Distribution of per-fold accuracy for the four primary feature types. The full loudness features (ISO 532-1 and ISO 532-3) exhibit tighter distributions (lower variance) than STFT and mel, indicating more consistent representations across data splits.

3.3.2 The Magnitude-Axis Transform: An Ablation

The results above separate two perceptual transforms that are usually conflated: a transform of the frequency axis (mel warping) and a transform of the magnitude axis (the compressive, level-dependent mapping from physical level to specific loudness). Because mel already applies the frequency transform, the further gain from mel to the loudness features points to the magnitude axis. The full loudness models also warp the frequency scale (to Bark or ERB), however, so a clean attribution requires holding that scale fixed. Table 3.4 does exactly this on the baseline CNN, inserting two lightweight features that apply a magnitude transform on top of the *same* mel filterbank, without the full ISO 532 machinery: “mel + power law” raises the mel power spectrogram P to a compressive exponent ($P^{0.3}$, Stevens’ loudness exponent) in place of logarithmic compression, and “mel + equal-loudness” first weights P

by the equal-loudness-contour (ELC) weighting w_{ELC} , a fixed per-band gain derived from a 60-phon equal-loudness contour, before the same power law $((P \cdot w_{\text{ELC}})^{0.3})$.

Table 3.4: Magnitude-axis ablation on ESC-50 (baseline CNN, 3 seeds $\times$ 5-fold CV). Mel, mel + power law, and mel + equal-loudness share the same 240-band mel frequency scale and differ only in how the magnitude axis is treated; the two ISO rows are included as full-model endpoints and use their own auditory scales (ISO 532-1: 240 Bark bands; ISO 532-3: 150 ERB bands). Δmel is the gain over the log-mel baseline.

Feature	Magnitude transform	Acc. (%)	Δmel
STFT	linear magnitude, linear freq.	39.7	—
Mel	$\log(1 + P)$	53.1	—
Mel + power law	$P^{0.3}$	59.2	+6.1
Mel + equal-loud.	$(P \cdot w_{\text{ELC}})^{0.3}$	61.9	+8.8
ISO 532-1	Zwicker specific loudness	59.5	+6.4
ISO 532-3	Moore–Glasberg loudness	63.0	+9.9

Three findings stand out. First, a single compressive power law recovers +6.1 of the +9.9 point mel-to-532-3 gain—roughly 60% of the full benefit—at essentially the cost of a mel spectrogram (Table 5.2). Second, adding equal-loudness contour weighting reaches 61.9%, which *matches or exceeds* the full Zwicker model (ISO 532-1, 59.5%; the difference is not statistically significant, $p = 0.09$). Most of what ISO 532-1 provides on mono audio is thus a frequency- and level-dependent reweighting that a two-line modification of the mel front end reproduces. Third, only the state-of-the-art ISO 532-3 model exceeds the cheap variants, and by a modest margin (+1.1 over mel + equal-loudness).

This decomposition also generalizes across model families of increasing capacity, from the baseline CNN to ResNet, Transformer, and MLP-Mixer; Section 3.5 presents that comparison in full.

Compressive power-law magnitude scaling is not itself novel: it underlies perceptual linear prediction (PLP) [15], which applies a cube-root intensity compression motivated by Stevens’

law, and power-normalized cepstral coefficients (PNCC) [16]. The contribution here is not the transform but the systematic demonstration—across three datasets spanning sound events and acoustic scenes, and four architectures—that the magnitude axis, rather than the more commonly studied frequency axis, carries the bulk of the perceptual advantage, and that a low-cost approximation captures most of it.

3.4 *Per-Class Analysis*

Of the 50 ESC-50 classes, the two specific-loudness features (ISO 532-1 and ISO 532-3) together achieve the highest accuracy on 38, mel on 9, and STFT on only 3. The per-class values quoted in this section are drawn from a separate single-seed, four-feature run whose overall accuracies (38.4%, 52.0%, 59.3%, and 60.6% for STFT, mel, ISO 532-1, and ISO 532-3) differ slightly from the three-seed means reported earlier in this chapter; the relative ordering of the features is unchanged. The classes where the loudness advantage is largest share common acoustic properties:

- **Tonal and periodic sounds:** chirping birds (532-3: 82.5% vs. mel: 27.5%), church bells (92.5% vs. 75.0%), clock tick (82.5% vs. 47.5%), rooster (85.0% vs. 80.0%).
- **Spectrally complex sounds:** thunderstorm (532-1: 87.5% vs. mel: 65.0%), crackling fire (62.5% vs. 27.5%), where masking interactions between frequency bands are significant.
- **Environmental sounds:** footsteps (532-3: 72.5% vs. mel: 45.0%), toilet flush (90.0% vs. 62.5%), pouring water (532-1: 77.5% vs. mel: 42.5%).

The classes where mel matches or exceeds both loudness models tend to involve sounds whose raw spectral shape is highly distinctive without requiring level-dependent reweighting: cow (mel: 82.5%) and chainsaw (82.5%), where mel clearly leads, and sneezing (82.5%) and snoring (77.5%), where mel ties ISO 532-1.

3.4.1 Spectral Visualization of Class-Dependent Performance

To understand *why* loudness features outperform on certain classes but not others, Figures 3.4 through 3.9 present side-by-side spectral representations of representative clips from classes where the performance gap is largest.

A consistent set of magnitude-axis effects links these visual differences to the accuracy gaps: equal-loudness weighting raises sensitivity around the 3–4 kHz hearing peak and lowers it at low frequencies; level-dependent masking sharpens strong spectral peaks by suppressing weaker neighbours; and the critical-band (Bark/ERB) warping concentrates narrowband energy. Where a class is defined by structure these effects sharpen or isolate, loudness helps: the high-frequency chirps of crickets are concentrated into a clean band and freed of surrounding noise (Figure 3.4; ISO 532-1 80.0% vs. STFT 25.0%), the thunderstorm’s broadband balance is reshaped toward salient mid frequencies (Figure 3.5), and the church-bell harmonic series is sharpened by masking that suppresses inter-harmonic noise (Figure 3.6).

The same effects explain where loudness features *lose* (Figures 3.8 and 3.9). The washing machine is defined by a low-frequency hum below roughly 500 Hz—exactly the region equal-loudness weighting attenuates—while its weaker 2–4 kHz broadband content sits near the hearing peak and compression raises the noise floor. Its distinctive low-frequency signature is thus flattened to parity with broadband content rather than silenced (the loudness panels remain bright and busy), leaving it easily confused with other broadband sounds and dropping to 20.0% against STFT’s 47.5%. A clock alarm’s sharp tonal onset likewise relies on fine frequency detail that the broader critical bands and temporal integration smooth over (Figure 3.9). Perceptual weighting thus helps classes whose identity survives—or is sharpened by—the auditory transform, and hurts those whose defining cues it flattens or discards.

Spectral Comparison: Crickets

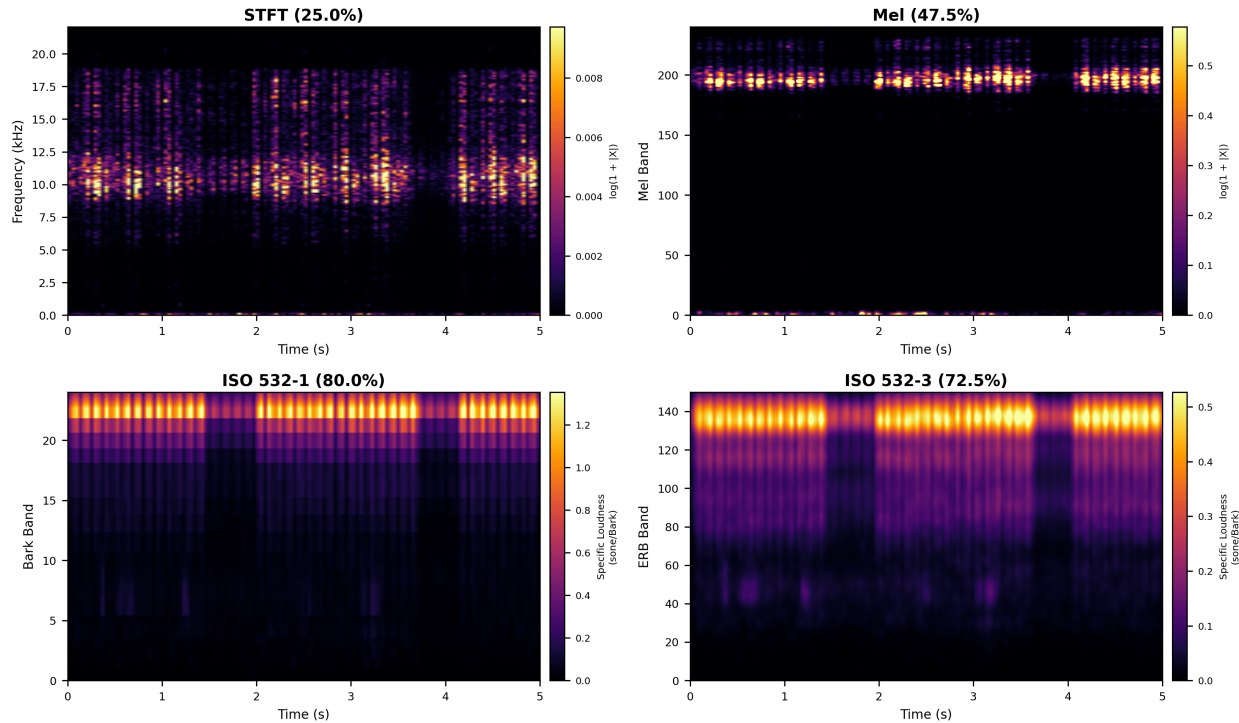


Figure 3.4: Feature representations of a crickets recording, whose energy is concentrated in the high frequencies (around 9–12 kHz). The loudness transform renders this periodic structure as a clean, isolated band—the wide high-frequency critical bands concentrate its narrowband energy and level-dependent masking suppresses surrounding noise—whereas the STFT spreads it across many bins and under-represents the highest frequencies. STFT accuracy: 25.0%, ISO 532-1: 80.0%.

3.4.2 Confusion Matrices

Figures 3.11 and 3.12 present full confusion matrices for mel and ISO 532-3 (the best-performing representation). The 532-3 matrix shows a stronger and more concentrated diagonal, indicating more confident and correct predictions across classes. Off-diagonal confusion patterns reveal that mel frequently confuses acoustically similar classes (e.g., animal sounds with each other, domestic sounds with each other), while the loudness representation reduces these confusions through its perceptual transform of the magnitude axis.

Spectral Comparison: Thunderstorm

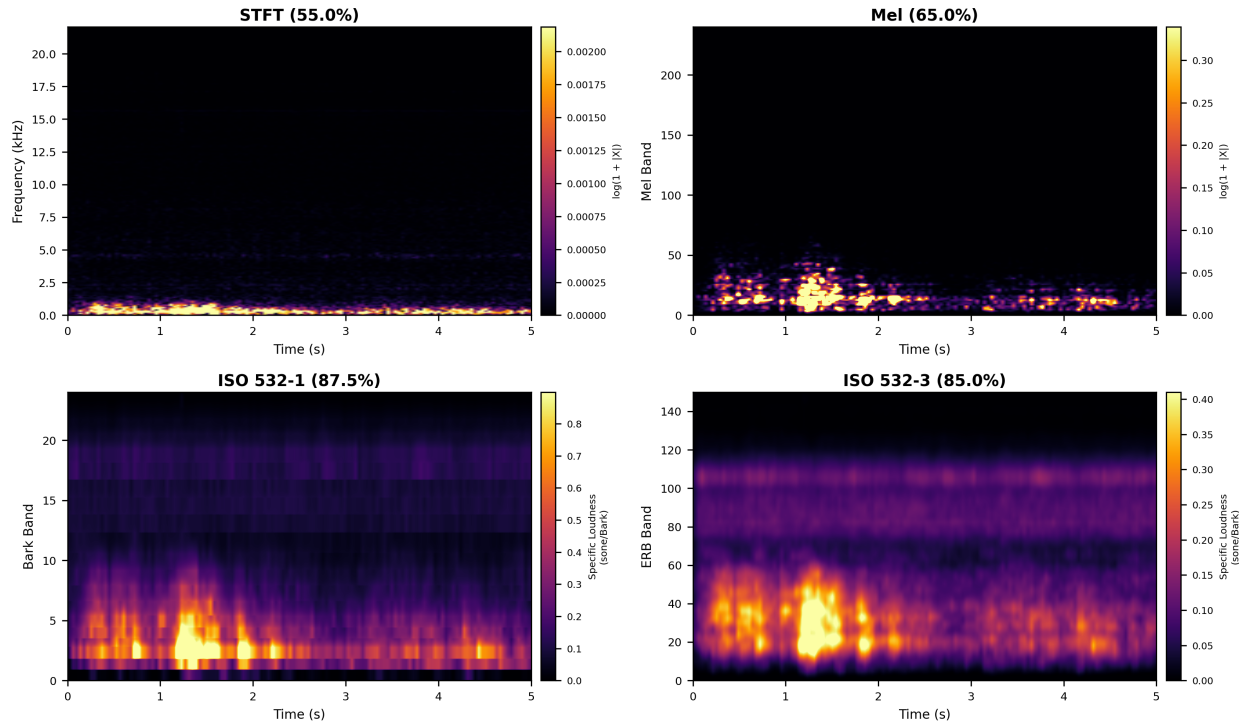


Figure 3.5: Feature representations of a thunderstorm recording. The equal-loudness weighting in the loudness models reshapes the broadband spectral balance, enhancing perceptually salient mid-frequency content. ISO 532-1 accuracy: 87.5% vs. mel: 65.0%.

Confusion matrices for all four representations (STFT, mel, ISO 532-1, and ISO 532-3) are available; the STFT and ISO 532-1 matrices show intermediate patterns consistent with their respective accuracies.

Spectral Comparison: Church Bells

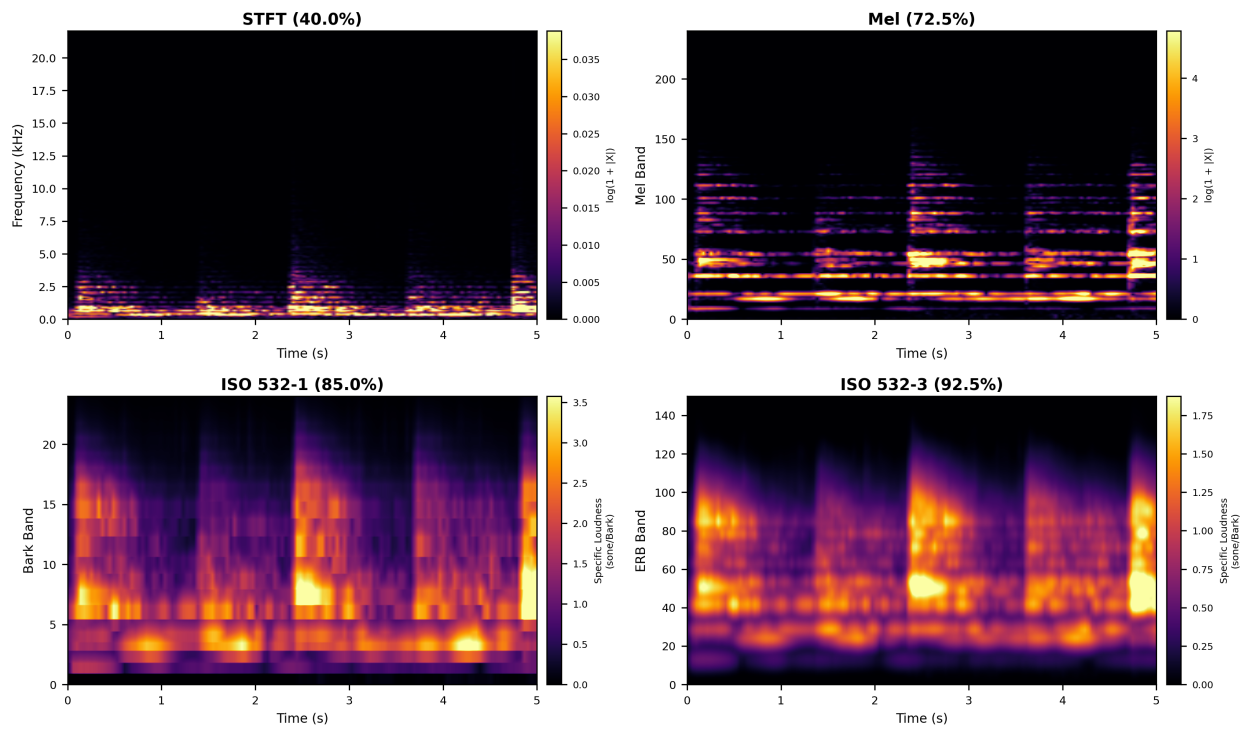


Figure 3.6: Feature representations of church bells. The harmonic series is preserved and sharpened in the loudness representations through spectral masking, which suppresses inter-harmonic noise. ISO 532-3 accuracy: 92.5% vs. STFT: 40.0%.

Spectral Comparison: Clock Tick

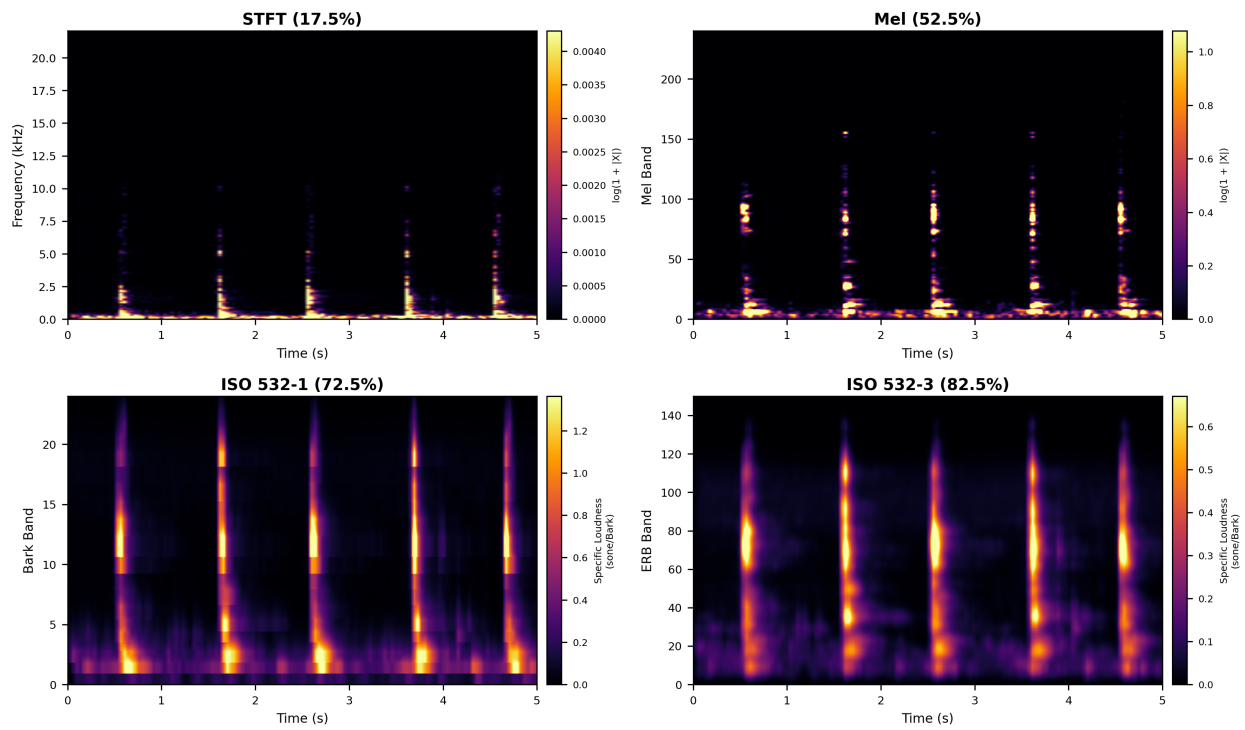


Figure 3.7: Feature representations of clock ticks. The periodic impulsive structure is captured differently across representations. Loudness temporal integration smooths the transients but preserves the rhythmic pattern that defines the class. ISO 532-3 accuracy: 82.5% vs. STFT: 17.5%.

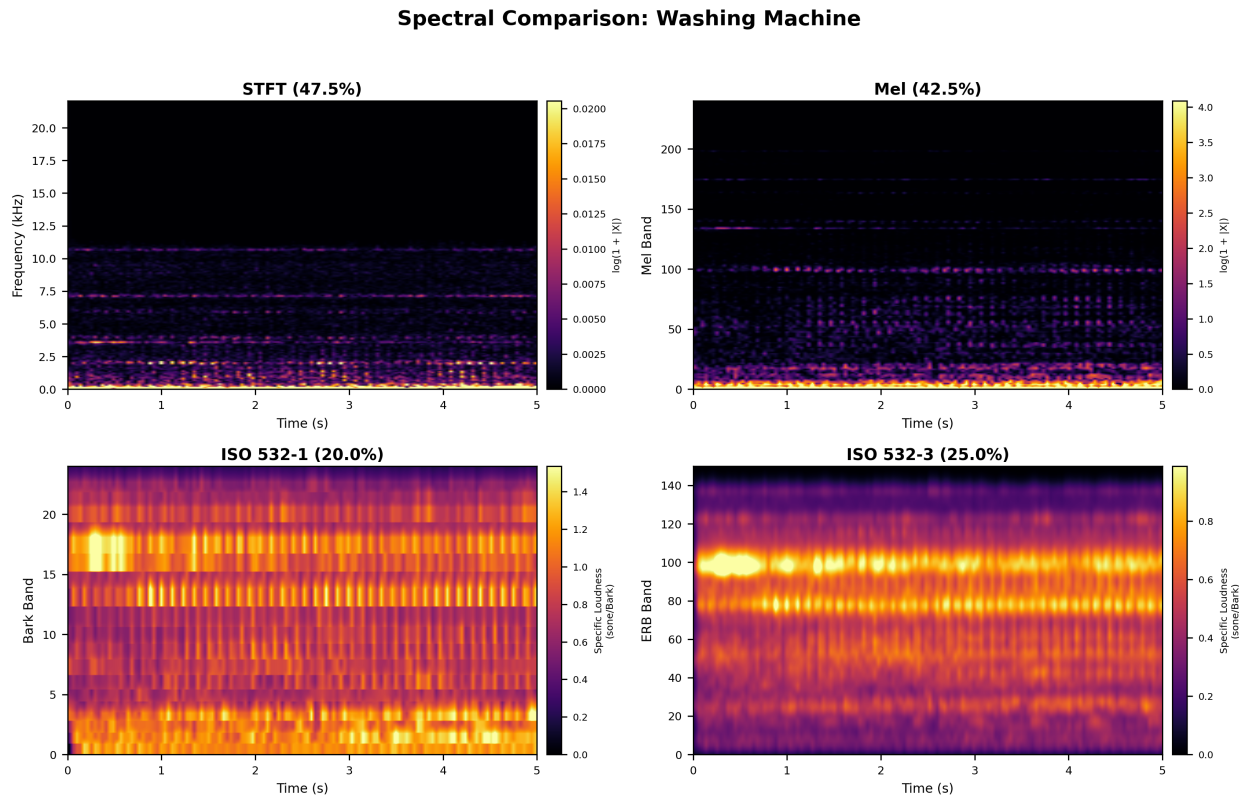


Figure 3.8: Feature representations of a washing machine. The equal-loudness weighting attenuates its defining low-frequency hum (<500 Hz), while the weaker 2–4 kHz broadband content sits near the ear’s sensitivity peak and compression raises the noise floor. The distinctive low-frequency signature is thus flattened to parity with broadband content—the loudness panels are bright and busy rather than silent—leaving the class hard to distinguish from other broadband sounds. STFT accuracy: 47.5% vs. ISO 532-1: 20.0%.

Spectral Comparison: Clock Alarm

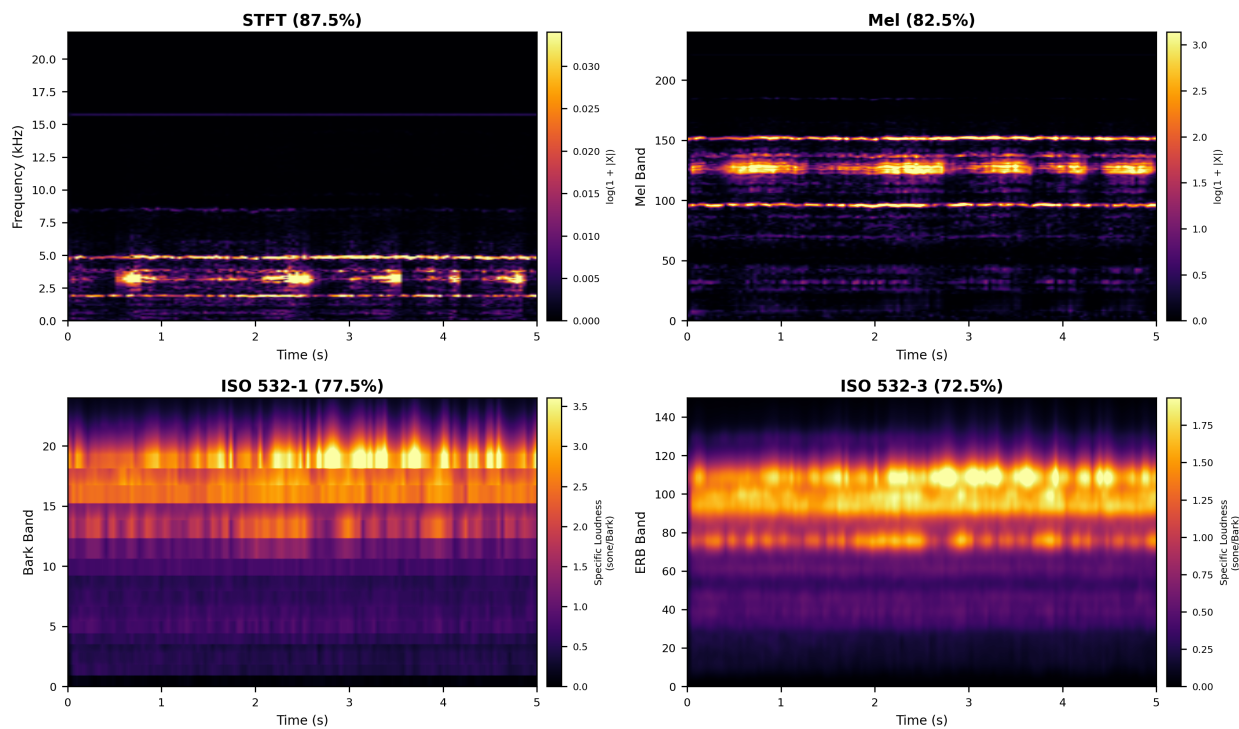


Figure 3.9: Feature representations of a clock alarm. The sharp tonal alarm is well-represented in all features, but the STFT's finer frequency resolution may preserve subtle harmonic details that the loudness model's broader critical bands smooth over. STFT accuracy: 87.5% vs. ISO 532-3: 72.5%.

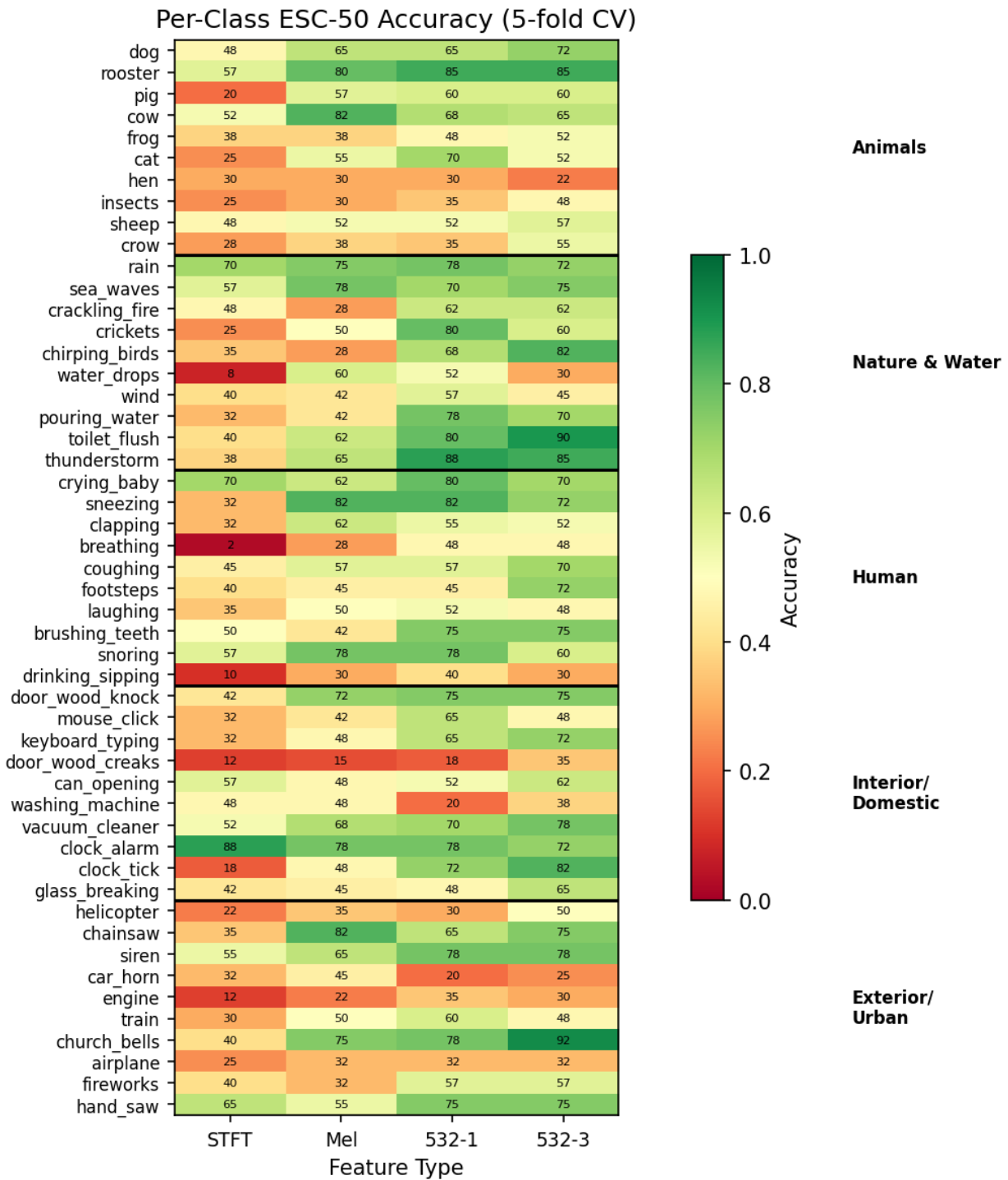


Figure 3.10: Classification accuracy per ESC-50 class for all four features. Classes are grouped by category. Loudness features (rightmost columns) show consistently higher accuracy across most classes, with particular strength in tonal, periodic, and spectrally complex sounds.

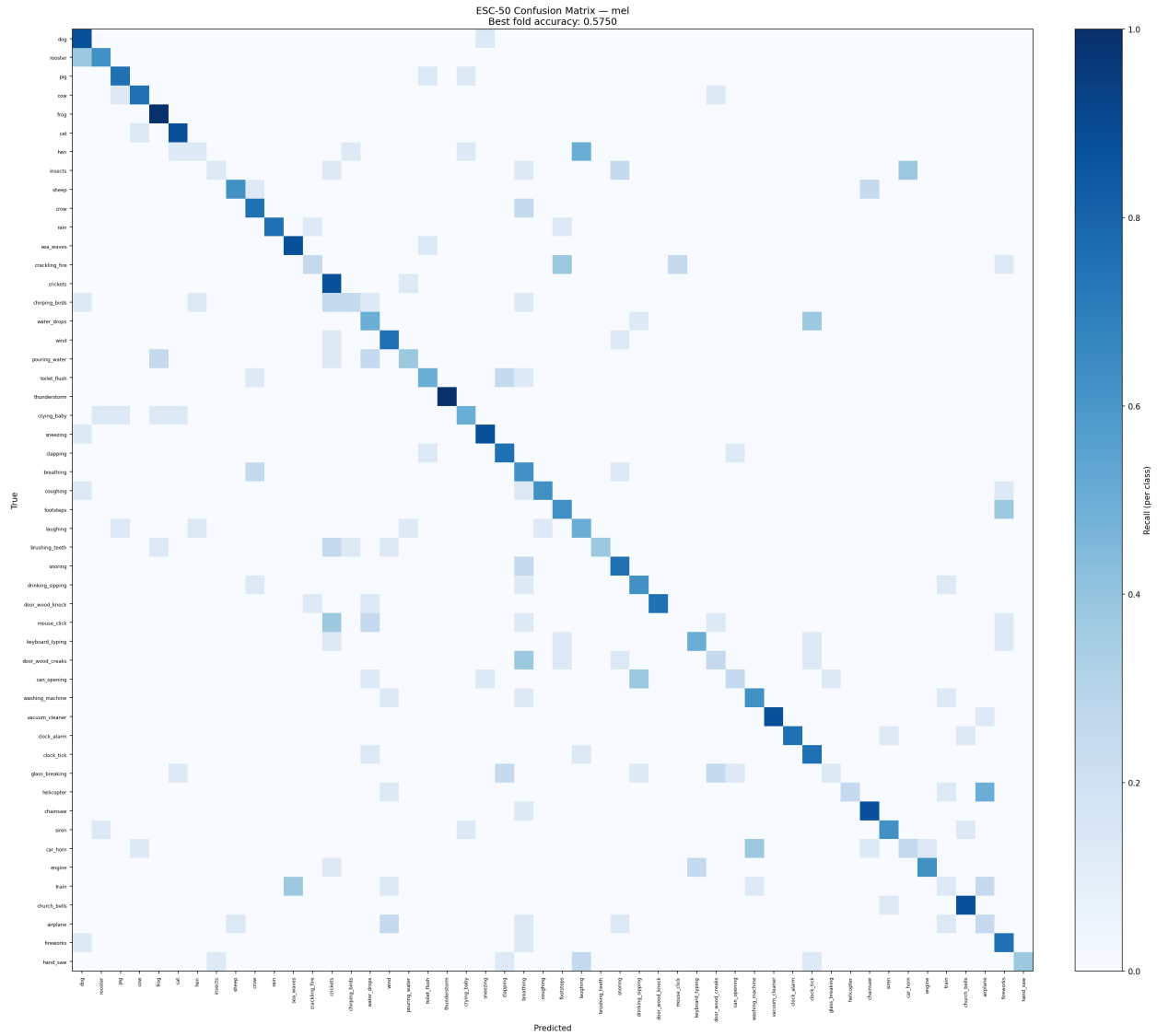


Figure 3.11: ESC-50 confusion matrix for mel spectrogram features (best fold accuracy: 57.5%). Off-diagonal entries indicate frequent misclassifications, particularly among acoustically similar classes.

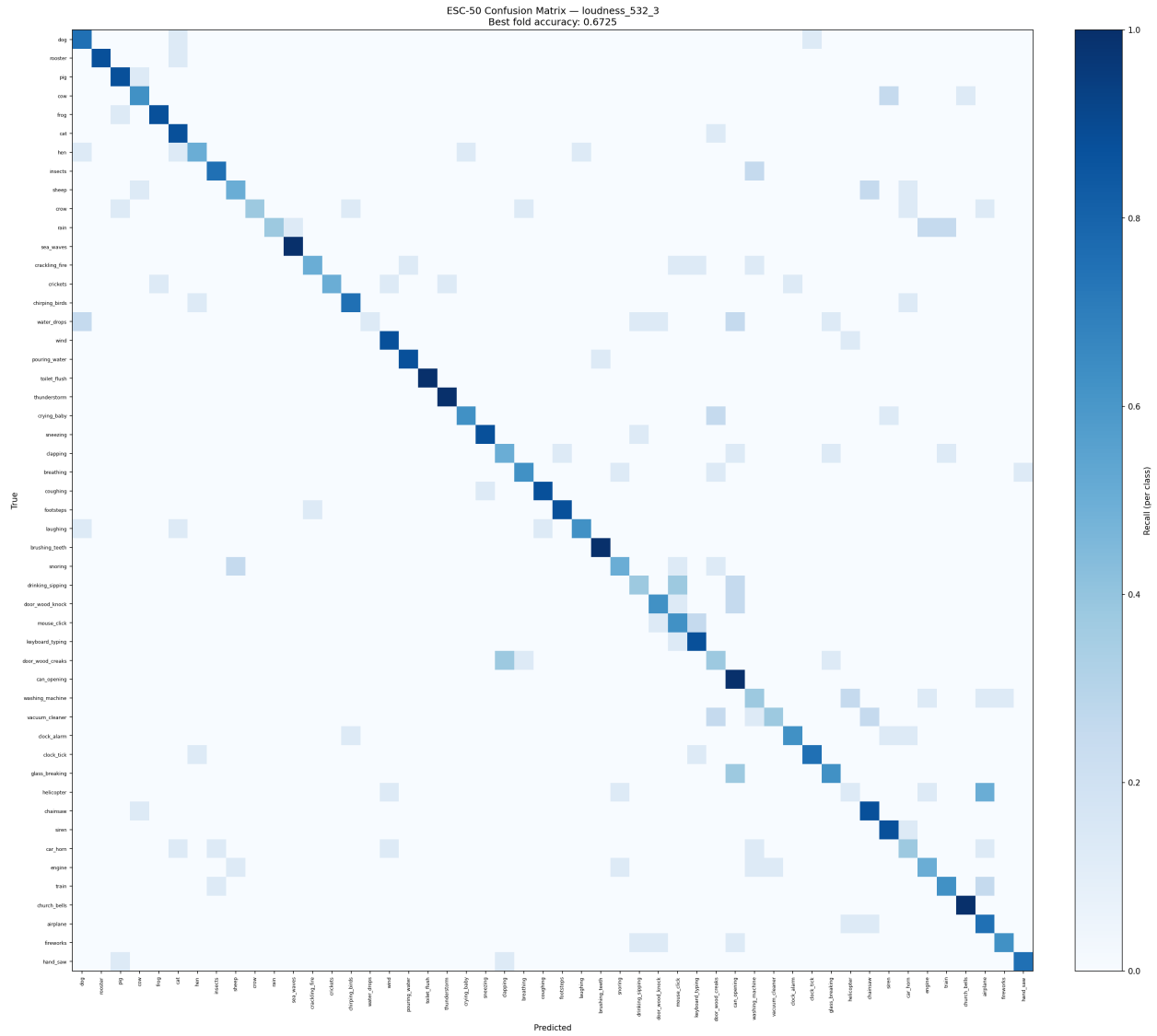


Figure 3.12: ESC-50 confusion matrix for ISO 532-3 specific loudness features (best fold accuracy: 67.3%). The stronger diagonal and reduced off-diagonal activity indicate that the perceptual transform of the magnitude axis helps distinguish classes that mel confuses.

3.5 Cross-Architecture Generalization

A key question is whether the loudness advantage is specific to the CNN used in the primary experiments or generalizes across model families. The ESC-50 experiment was therefore repeated with four architectures spanning a range of *inductive bias*—the built-in assumptions each makes about how structure is arranged in the input time–frequency plane—from strong locality (convolutional models) to weak (attention- and MLP-based models that must learn spatial structure from data). Figure 3.13 shows how each parses the same input.

1. **CNN:** The 3-layer network from the primary experiments (Section 3.2). Small 3×3 kernels slide over the time–frequency plane, so each unit responds to a *local* neighborhood; three convolution/batch-norm/max-pool blocks build a shallow hierarchy of local features before pooling and a linear classifier.
2. **ResNet:** A deeper convolutional network (ResNet-18 style): a 7×7 stride-2 stem and four residual blocks (two at 64 channels, two at 128 with downsampling). Skip connections let gradients bypass layers, enabling greater depth—larger receptive fields, more abstract spectro-temporal features—while keeping the same convolutional locality bias.
3. **Transformer:** A Vision Transformer (ViT)-style encoder without convolution. The input is split into non-overlapping 8×8 patches, each embedded into a 128-dimensional token; with a classification (CLS) token and learnable positional embeddings, four self-attention layers (four heads, feed-forward dimension 256) let *every* patch attend to every other. Receptive fields are global from the first layer, but the spatial prior is weak: structure must be learned from data.
4. **MLP-Mixer:** A patch model using the same 8×8 embedding but no attention. Four blocks each alternate a *token-mixing* MLP (across patch positions) and a *channel-*

mixing MLP (across the 128 channels). Mixing is global like attention, but through fixed learned weights rather than input-dependent ones.

The models thus parse the same representation in progressively less constrained ways: the convolutional pair (CNN, ResNet) assume discriminative cues are *local* and translation-equivariant; the Transformer assumes no locality and relates patches through data-dependent attention; the MLP-Mixer mixes all patches through fixed projections. Weaker inductive bias raises capacity but shifts the representational burden onto the input features themselves—which is why front-end quality interacts with architecture in the results below.

All four models used the same pre-extracted features and training protocol. Optimization used Adam [22]—which adapts a per-parameter learning rate from running estimates of the gradient’s first and second moments—with cosine annealing of the base rate (10^{-3}), weight decay 10^{-5} , batch size 32, and early stopping (up to 100 epochs, minimum 30). Table 3.5 presents the results.

Table 3.5: ESC-50 accuracy (%) across four architectures and six feature types (3 seeds $\times$ 5-fold CV). M+pow = mel + power law; M+eq = mel + equal-loudness. Δ is the best feature minus mel; the best per row is in bold.

Architecture	STFT	Mel	M+pow	M+eq	532-1	532-3	Δ
CNN	39.7	53.1	59.2	61.9	59.5	63.0	+9.9
ResNet	39.9	36.4	70.1	71.7	68.4	74.8	+38.4
Transformer	35.3	34.2	41.6	42.2	46.0	46.7	+12.5
MLP-Mixer	23.4	22.5	28.3	27.6	25.6	26.9	+5.8

The loudness advantage is consistent across all four architectures: the best perceptual-magnitude feature outperforms mel in every case, by margins ranging from +5.8 points (MLP-Mixer) to +38.4 points (ResNet). ISO 532-3 is the single best representation for the three stronger architectures (CNN, ResNet, Transformer); for the weak MLP-Mixer baseline the cheap mel + power law is marginally best. Crucially, the lightweight variants

are competitive with the full ISO models on three of the four architectures—mel + power law and mel + equal-loudness are on par with or exceed ISO 532-1 for CNN, ResNet, and MLP-Mixer—and only the Transformer clearly favors the full loudness computation. Section 3.3.2 examines this trade-off.

3.5.1 *Architecture-Dependent Advantage*

The variation in the loudness advantage across architectures reveals an important interaction between feature quality and model capacity.

ResNet shows the largest advantage (38.4pp). Residual connections enable the network to preserve and refine subtle patterns through deeper layers. Loudness features, which have been perceptually smoothed and structured by the auditory model, provide cleaner spectro-temporal patterns that ResNet’s skip connections are well-suited to exploit. Notably, mel features *degraded* with ResNet (36.4% vs. 53.1% with CNN), suggesting that ResNet’s greater capacity leads to overfitting on mel’s noisier representation, while loudness features provide a more robust learning signal.

CNN and Transformer show moderate advantages (9.9pp and 12.5pp). The CNN’s aggressive max pooling discards information indiscriminately, limiting its ability to fully exploit the richer loudness representation. The Transformer underperforms overall on this dataset (likely due to data hunger with only 1,600 training clips per fold), but the loudness advantage persists, confirming that the benefit is not specific to convolutional architectures.

MLP-Mixer shows the smallest advantage (5.8pp) and the lowest overall accuracy. Without any spatial inductive bias (no convolutions, no attention), the MLP-Mixer cannot effectively exploit the spectro-temporal structure of any representation. The small but consistent loudness advantage suggests that even when the architecture cannot leverage spatial patterns, loudness features provide inherently cleaner statistical distributions that marginally improve classification.

The key insight is that **loudness features and capable architectures are synergistic**: a more powerful architecture amplifies the loudness advantage rather than diminishing

it. This contradicts the argument that “the network can learn its own perceptual weighting,” since if that were the case, the advantage of pre-computed loudness features should decrease with more capable architectures. Instead, the opposite occurs, because perceptual preprocessing removes noise that even powerful architectures struggle to overcome with limited training data.

3.5.2 Cross-Architecture Statistical Significance

To verify that the loudness advantage is statistically significant across architectures (not only for the CNN in Chapter 3), paired t -tests were conducted for each architecture comparing ISO 532-3 accuracy against mel accuracy across the five cross-validation folds (each fold averaged over three seeds).

Table 3.6: Statistical significance of the loudness advantage (ISO 532-3 vs. mel) per architecture. Paired t -tests across 5 folds, $\alpha = 0.05$.

Architecture	ΔMean	p	Cohen’s d	Sig?
CNN	+9.9pp	0.002	3.37 (large)	Yes
ResNet	+38.4pp	0.0008	4.08 (large)	Yes
Transformer	+12.5pp	0.0003	5.08 (large)	Yes
MLP-Mixer	+4.4pp	0.025	1.57 (large)	Yes

The loudness advantage is statistically significant ($p < 0.05$) for all four architectures, with large effect sizes in every case. The ResNet result is particularly noteworthy: $p < 0.001$ with $d = 4.08$ represents an exceptionally strong effect, reinforcing the finding that capable architectures amplify the loudness advantage.

Notably, ISO 532-3 significantly exceeds ISO 532-1 on the two convolutional architectures (CNN: +3.5pp, $p = 0.010$, $d = 2.05$; ResNet: +6.4pp, $p = 0.003$, $d = 2.95$) but not on the Transformer (+0.7pp, $p = 0.59$) or MLP-Mixer (+1.2pp, $p = 0.18$). This suggests that the ISO 532-3 representation—which applies ERB-scale (roex) auditory filters together with

explicit temporal integration, rather than the Bark-band excitation pattern of ISO 532-1—yields additional discriminative structure that the convolutional inductive bias is best positioned to exploit. The difference is not one of raw dimensionality: as extracted here, ISO 532-1 provides 240 Bark bands against ISO 532-3’s 150 ERB bands.

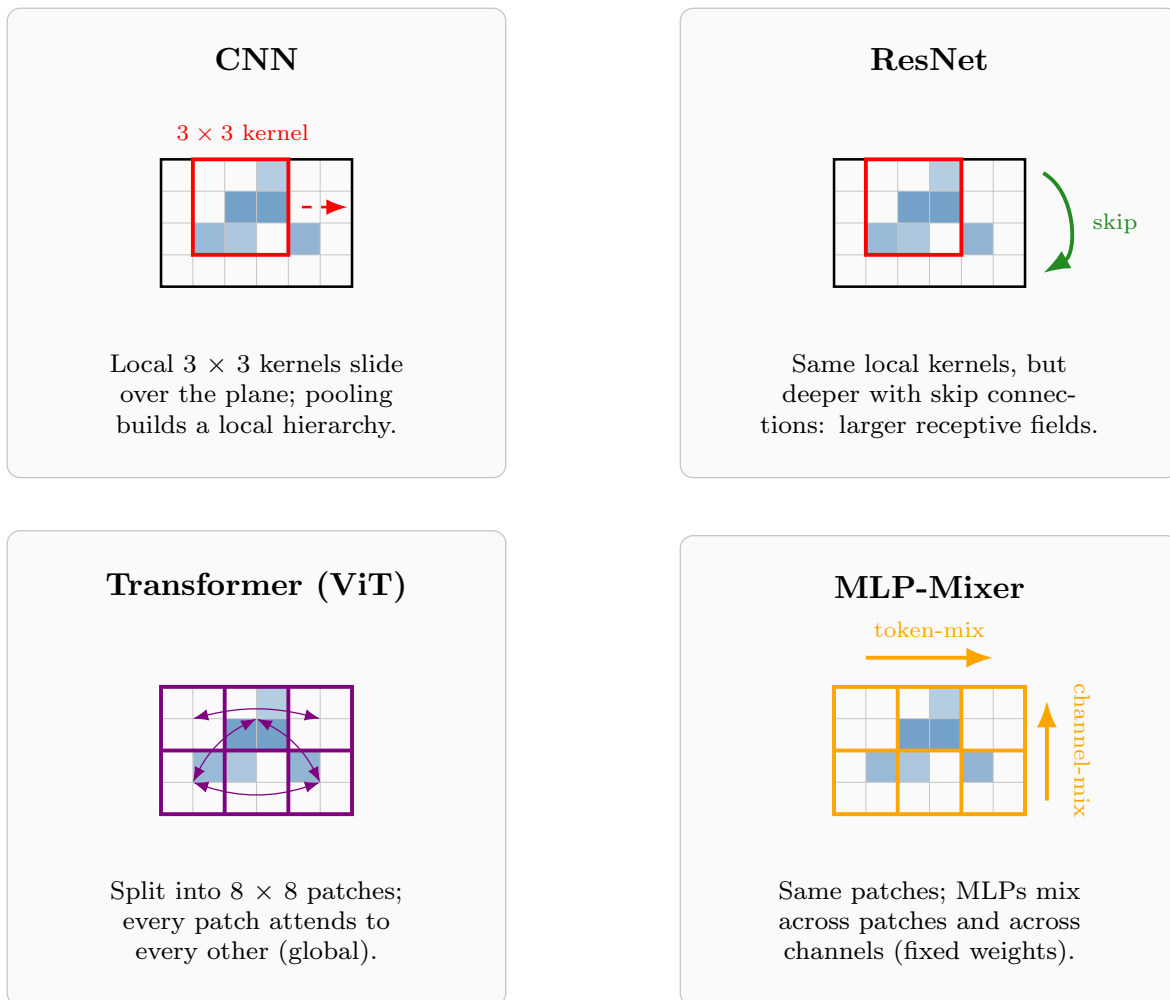


Figure 3.13: How the four architectures parse the same input time–frequency representation: the convolutional models (CNN, ResNet) apply local sliding kernels, while the patch-based models (Transformer, MLP-Mixer) split the plane into patches and relate them by attention or by MLP mixing.

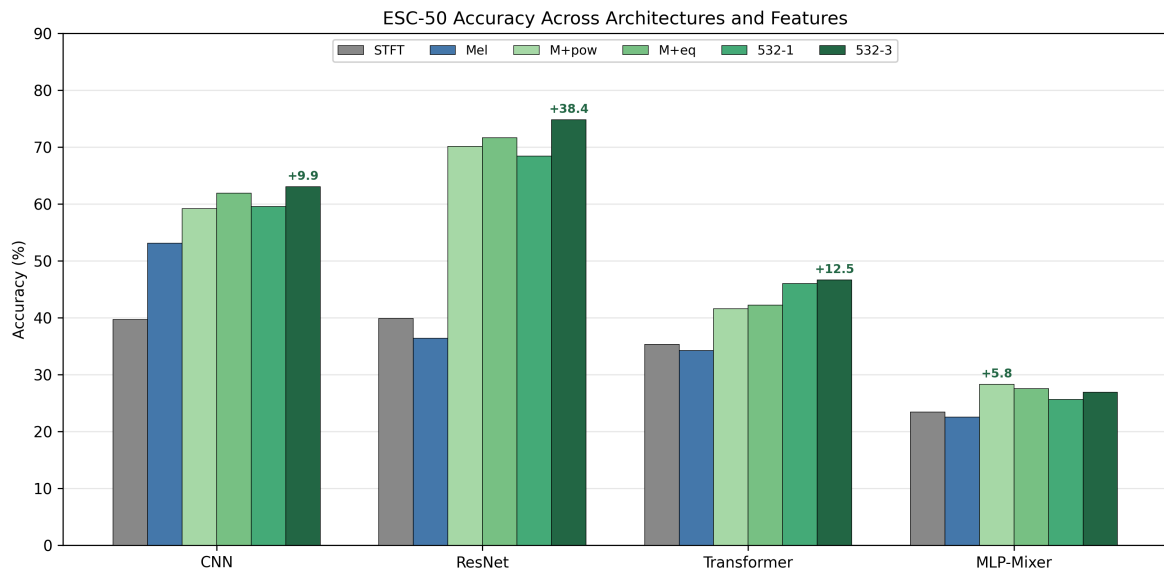


Figure 3.14: ESC-50 accuracy across four architectures and six features. The perceptual magnitude-axis features (green bars) consistently outperform STFT and mel across all architectures. The advantage is largest with ResNet (+38.4pp over mel) and smallest with MLP-Mixer (+5.8pp). Annotations show the improvement of the best feature over mel for each architecture.

3.6 Cross-Dataset Generalization

The preceding sections establish the magnitude-axis advantage on ESC-50 and show that it holds across model architectures. A separate question is whether the advantage generalizes across *datasets*, and whether it extends beyond isolated sound *events* to the related task of acoustic *scene* classification. To test this, the classification experiment was replicated on two additional datasets using the same baseline CNN and the same training protocol. UrbanSound8K was evaluated on all six feature variants; TAU on five, omitting ISO 532-3 for the reason given below:

- **UrbanSound8K** [23]: 8,732 urban sound-event clips across 10 classes with the prescribed 10-fold cross-validation — a second, independent environmental-sound-event benchmark. Results are averaged over three random seeds.
- **TAU Urban Acoustic Scenes 2019** [24]: 14,400 ten-second recordings across 10 acoustic scenes. Folds were assigned by stratified group k -fold ($k = 5$) grouped by recording location, so no location appears in both the training and test partitions. Scene classification is a distinct task from event classification, testing whether the advantage extends *beyond* the sound-event classes framed in the thesis statement.

3.6.1 UrbanSound8K (Sound Events)

Table 3.7 reports mean 10-fold accuracy (over three seeds) for each feature.

Table 3.7: UrbanSound8K accuracy (%) by feature (baseline CNN, 10-fold CV, mean over 3 seeds). M+pow = mel + power law; M+eq = mel + equal-loudness. Δ is the best feature minus mel.

Dataset	STFT	Mel	M+pow	M+eq	532-1	532-3	$\Delta(\text{Best-Mel})$
UrbanSound8K	50.9	55.5	69.3	70.7	71.6	72.1	+16.6

On UrbanSound8K, transforming the magnitude axis raises accuracy from 55.5% (mel) to 69.3% (mel + power law), 70.7% (mel + equal-loudness), 71.6% (ISO 532-1), and 72.1% (ISO 532-3), a gain of up to 16.6 percentage points (ISO 532-3 vs. mel: paired t -test, $p < 10^{-10}$, $d_z = 1.82$), replicating the ESC-50 finding on an independent environmental-sound-event benchmark. As on ESC-50, the lightweight power-law and equal-loudness features capture most of the gain—within about three points of the full ISO models. Here, however, ISO 532-3 and ISO 532-1 are statistically indistinguishable (+0.48 pp, $p = 0.58$): unlike ESC-50, the additional auditory detail of the Moore-Glasberg-Schlittenlacher model yields no measurable gain over the Zwicker model on this dataset. The frequency-axis transform alone (STFT $\rightarrow$ mel) yields only +4.6 pp ($p = 0.012$), so the magnitude axis again carries the advantage.

3.6.2 TAU Urban Acoustic Scenes (Acoustic Scenes)

Table 3.8 reports mean 5-fold accuracy for each feature.

Table 3.8: TAU Urban Acoustic Scenes 2019 accuracy (%) by feature (baseline CNN, stratified group 5-fold by location, single seed). M+pow = mel + power law; M+eq = mel + equal-loudness. Δ is the best feature minus mel. ISO 532-3 (—) was not computed for TAU (see text).

Dataset	STFT	Mel	M+pow	M+eq	532-1	532-3	$\Delta(\text{Best-Mel})$
TAU Scenes 2019	46.0	44.1	58.6	58.0	58.4	—	+14.5

Even for acoustic scenes — a task outside the sound-event framing of the thesis statement — transforming the magnitude axis lifts accuracy from 44.1% (mel) to 58.0–58.6% for the lightweight power-law and equal-loudness features and 58.4% for ISO 532-1, a gain of roughly 14 percentage points over mel (all $p < 0.04$, with large effect sizes, $d_z > 1.3$). As on the other datasets the cheap variants are statistically indistinguishable from one another ($p = 0.73$). Notably, mel does *not* improve on STFT here (−2.0 pp, n.s.), so on acoustic scenes the *entire*

advantage comes from the magnitude axis. Full ISO 532-3 loudness was not computed for TAU: at 14,400 ten-second recordings, its extraction was computationally prohibitive (on the order of several days of CPU processing for the full corpus), whereas ISO 532-1 and the lightweight features already establish the magnitude-axis effect on this dataset.

3.6.3 *Summary*

Across ESC-50, UrbanSound8K, and TAU — three datasets spanning both sound events and acoustic scenes — the perceptual transform of the magnitude axis yields a consistent advantage over mel—roughly 10 percentage points on ESC-50 and 14–16 on the two urban corpora. The effect is therefore not an artifact of a single dataset or task: it holds for an independent set of environmental-sound events and generalizes to acoustic scene classification, strengthening the central claim that perceptually transforming the magnitude axis aligns the representation with human-defined sound categories.

3.7 *Discussion of ESC-50 Results*

The ESC-50 results demonstrate a clear hierarchy of perceptual processing benefit. The progression from STFT (39.7%) to mel (53.1%) to ISO 532-1 (59.5%) to ISO 532-3 (63.0%) maps directly onto the degree of perceptual alignment in each representation:

1. **STFT $\rightarrow$ Mel (+13.4pp)**: A perceptual transform of the frequency axis alone provides substantial improvement, confirming the well-known benefit of mel features.
2. **Mel $\rightarrow$ ISO 532-1 (+6.4pp)**: Adding a perceptual transform of the magnitude axis (equal-loudness weighting, temporal integration, masking) provides a further significant gain. This is the “missing dimension” that mel spectrograms do not capture; as Section 3.3.2 shows, a lightweight power-law compression of the mel magnitude axis alone recovers most of it.
3. **ISO 532-1 $\rightarrow$ ISO 532-3 (+3.5pp)**: The more sophisticated auditory model with

ERB-scale resolution and refined temporal integration provides a further significant improvement ($p = 0.010$), even on mono audio where its binaural capabilities are unused.

The lower variance of the loudness results across folds is also noteworthy. It suggests that the full auditory models produce more invariant features, features that are less sensitive to recording conditions and amplitude variations, consistent with the automatic gain control function that loudness perception serves in the auditory system. This property does not follow from the magnitude transform alone: the lightweight power-law and equal-loudness variants show fold variance comparable to or higher than mel (Section 3.3.2).

Chapter 4

SELD EXPERIMENT: STARSS23

This chapter presents the SELD experiment on STARSS23. It begins by reviewing the SELD task formulation and the psychoacoustic principles that motivate the binaural processing pipeline, before turning to the experimental design, results, and a multi-threshold analysis that separates detection from localization performance.

4.1 *Sound Event Localization and Detection*

Sound Event Localization and Detection (SELD) is the joint task of answering three questions about an audio recording: *what* sounds are present, *when* do they occur, and *where* are they located in 3D space. This is challenging because real-world sound scenes contain multiple overlapping sources in reverberant environments with background noise. The system must simultaneously track and identify several concurrent events, each potentially moving through space over time.

The baseline architecture for SELD is a Convolutional Recurrent Neural Network (CRNN). The convolutional layers extract spectro-temporal patterns from the input spectrogram (similar to how image CNNs extract edges and textures), the recurrent layers (bidirectional GRU) model temporal dependencies across time, and the output layer predicts Multi-ACCDOA vectors. Multi-ACCDOA encodes both event presence and direction: each output is a 3D Cartesian vector (x, y, z) where the vector magnitude indicates presence confidence and the direction indicates the estimated Direction of Arrival (DOA).

4.1.1 *STARSS23*

The Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [25] was developed for the DCASE 2023 Task 3 challenge on SELD. It contains multichannel audio recordings in two formats, both derived from recordings made with a 32-channel Eigenmike em32 spherical microphone array [25]:

- **First-Order Ambisonics (FOA):** Four channels (W, X, Y, Z) representing B-format ambisonics, where W captures omnidirectional sound pressure and X, Y, Z capture three orthogonal particle velocity components. This is a mathematical decomposition of the sound field, analogous to how spherical harmonics decompose a function on a sphere. Spatial information is encoded through phase relationships between these mathematical projections.
- **Microphone array (MIC):** Four channels extracted from four of the 32 capsules arranged in a tetrahedral configuration. Spatial information is encoded through inter-channel level and time differences between these physical microphone positions.

Both formats provide 4-channel input, but they encode spatial information through fundamentally different mechanisms: FOA uses phase relationships between mathematical projections of the sound field, while MIC uses physical differences between microphone positions.

Annotations provide both temporal activity labels and direction-of-arrival (DOA) coordinates obtained via motion capture, making STARSS23 one of the few SELD datasets with ground-truth spatial annotations derived from physical measurement rather than subjective labeling. The dataset contains 13 sound event classes, listed in Table 4.1.

4.1.2 *SELD Architecture*

The baseline SELD architecture for DCASE Task 3 is the AudioOnlyCRNN, a convolutional recurrent neural network that processes multi-channel spectrogram inputs [25]. The architecture consists of:

Table 4.1: STARSS23 sound event classes.

Index	Class
0	Female speech
1	Male speech
2	Clapping
3	Telephone
4	Laughter
5	Domestic sounds
6	Footsteps
7	Door, open or close
8	Music
9	Musical instrument
10	Water tap
11	Bell
12	Knock

1. Convolutional feature extraction layers with batch normalization.
2. Recurrent (GRU) layers for temporal sequence modeling.
3. A single joint output branch predicting Multi-ACCDOA vectors, which encode event activity and direction of arrival together.
4. ADPIT (Auxiliary Duplicating Permutation Invariant Training) loss for handling multiple simultaneous sources.

4.1.3 SELD Evaluation Metrics

SELD performance is evaluated using four joint metrics [26]:

- **Error Rate (ER)**: Ratio of insertion, deletion, and substitution errors to the total number of reference events. Lower is better.
- **F-score (F)**: Frame-level F1 score for joint detection and localization. Higher is better. A detection is considered correct only if the predicted class matches the reference and the predicted DOA is within a spatial threshold (typically 20°) of the reference.
- **Localization Error (LE)**: Mean angular error in degrees between predicted and reference DOA for correctly detected events. Lower is better.
- **Localization Recall (LR)**: Fraction of reference events for which the model produces a DOA prediction within the spatial threshold. Higher is better.

Following the STARSS23 baseline and the DCASE 2023 Task 3 evaluation, these four metrics are reported *separately* rather than being collapsed into a single aggregate “SELD score”; unlike earlier DCASE editions, the 2023 evaluation ranks systems by their cumulative rank across the four metrics, so this thesis likewise compares the metrics individually.

Ambisonics Sound Field Reconstruction

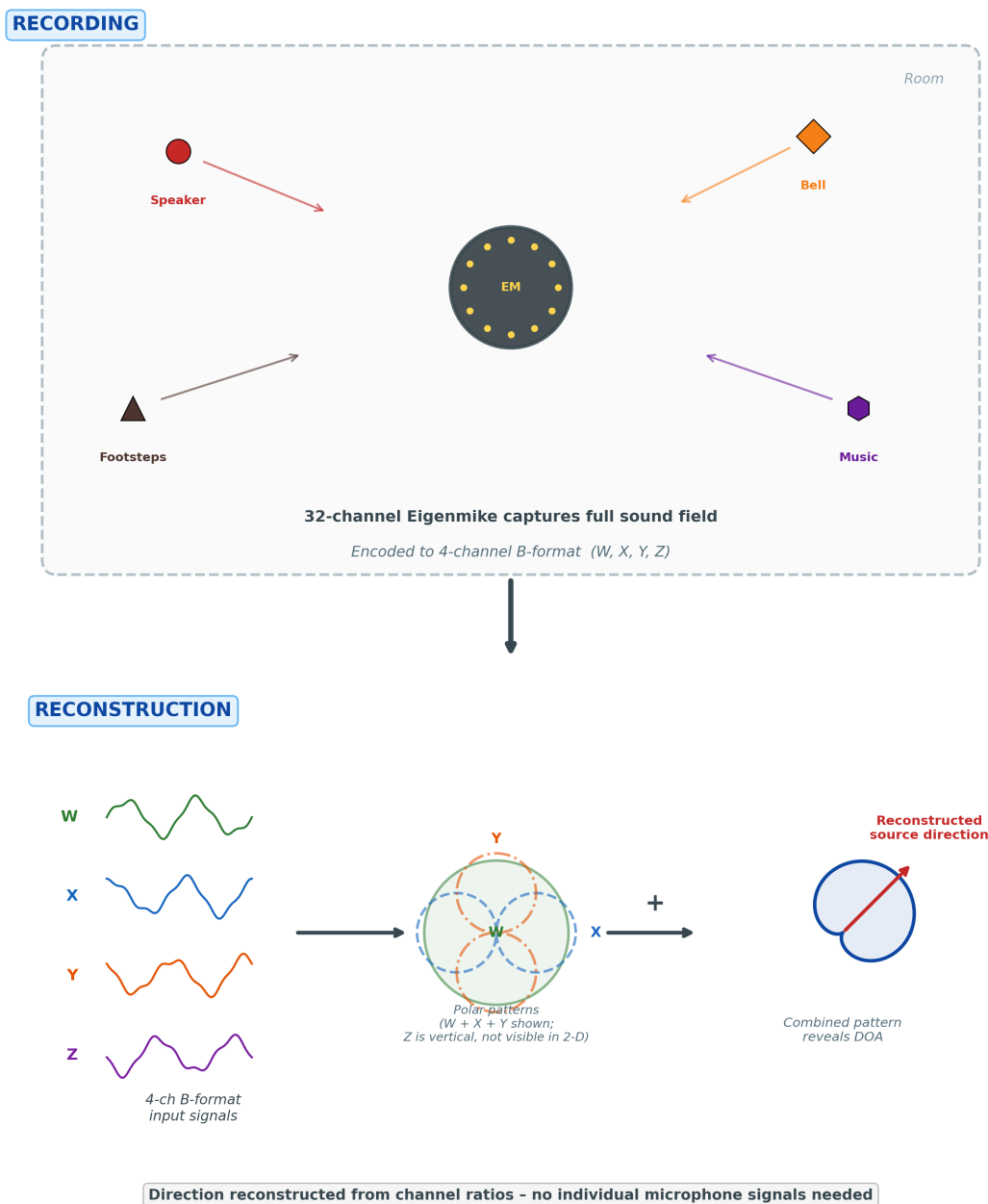


Figure 4.1: Sound field capture and reconstruction using First-Order Ambisonics. Top: a 32-channel Eigenmike captures the full sound field from multiple sources, encoded into 4-channel B-format (W, X, Y, Z). Bottom: the B-format signals combine omnidirectional and figure-8 directional patterns to reconstruct the spatial distribution of sound energy.

4.2 *Binaural Hearing and Spatial Cues*

Existing SELD systems process multi-channel ambisonics directly, without leveraging the spatial encoding that human binaural hearing provides through interaural differences. Binaural processing is a natural fit for SELD because it concentrates spatial information into interaural level and timing cues that are inherently coupled to spectral content, yet the combination of binaural rendering with specific loudness features has not been published. Closing that gap is the second research aim of this thesis.

The binaural processing pipeline introduced later in this chapter converts FOA recordings to binaural format before computing loudness features. This section reviews the psychoacoustic principles that motivate this approach.

4.2.1 *Interaural Cues*

Human sound localization relies primarily on two binaural cues that arise from the physical separation of the ears [27, 28]:

- **Interaural Time Difference (ITD):** The difference in arrival time between the two ears. For a source at azimuth θ , the ITD is approximately $\Delta t \approx (d/c) \sin \theta$, where d is the interaural distance (~ 17 cm) and c is the speed of sound. ITD is the dominant localization cue below approximately 1.5 kHz [29].
- **Interaural Level Difference (ILD):** The difference in sound level between the two ears, caused by head shadow. ILD increases with frequency as the head becomes a more effective acoustic obstacle, and is the dominant cue above approximately 1.5 kHz. At high frequencies, ILD can exceed 20 dB for lateral sources [27].

4.2.2 *Head-Related Transfer Functions*

The complete transformation imposed by the head, pinnae, and torso on incoming sound is described by the Head-Related Transfer Function (HRTF). An HRTF is a pair of filters

(one per ear) that captures direction-dependent spectral shaping, encoding both ILD and direction-dependent spectral cues (pinna notches, torso reflections) into the binaural signal. Databases such as CIPIC [30] provide measured HRTFs for many listener positions and head geometries.

For machine learning applications that require only the dominant binaural cues (ILD and gross spectral shaping) rather than individualized spatial rendering, a single representative HRTF is a practical and reproducible choice, obtained without listener-specific measurements. The downstream loudness integration attenuates fine-grained pinna cues but retains the broadband ILD patterns that encode horizontal source position.

4.2.3 Binaural Cues and Loudness Processing

The connection between binaural hearing and loudness processing is formalized in ISO 532-3, which includes a binaural inhibition stage (Section 2.2). When loudness differs between the two ears, the louder ear partially suppresses the quieter ear’s contribution to total loudness. This mechanism means that specific loudness computed from binaural input inherently encodes spatial information through ILD patterns.

For SELD, this creates an opportunity: by converting FOA recordings to binaural format and computing separate left and right loudness patterns, the spatial information encoded as ILD is preserved in the loudness representation. The ML model can then exploit both spectral content (from the loudness patterns) and spatial information (from the inter-channel level differences) without requiring explicit localization features.

4.3 Experimental Design

This chapter presents a SELD experiment comparing three input representations on the STARSS23 dataset:

1. **FOA baseline:** 4-channel First-Order Ambisonics (W , X , Y , Z), converted to 7 input feature channels—4 amplitude spectrograms and 3 inter-channel phase differ-

ences (IPDs), each with 257 frequency bins—reproducing the DCASE 2023 Task 3 baseline [25].

2. **ISO 532-3 binaural loudness:** 2-channel specific loudness computed using the Moore-Glasberg-Schlittenlacher method.
3. **ISO 532-1 binaural loudness:** 2-channel specific loudness computed using the Zwicker method.

4.4 *Dataset and Classes*

The STARSS23 dataset [25] contains 13 sound event classes (Table 4.1) with human-annotated temporal labels and motion-capture-derived DOA coordinates. This combination of perceptual temporal annotation and physical spatial annotation provides a uniquely rigorous ground truth for SELD evaluation.

4.5 *Binaural Conversion Pipeline*

Converting the STARSS23 FOA recordings to binaural loudness features required a multi-stage signal processing pipeline developed for this thesis. Each stage is described below.

4.5.1 *HRTF Construction from the ARI Database*

The binaural conversion requires Head-Related Impulse Responses (HRIRs) to map the ambisonics sound field to left- and right-ear signals. A single representative individual HRTF was used: subject NH10 from the Acoustics Research Institute (ARI) database, in SOFA format. The $\pm 45^\circ$ horizontal-plane measurements at 0° elevation were resampled to 24 kHz and convolved with the corresponding virtual-speaker signals. A single high-quality individual HRTF is preferable here to a cross-subject average for two reasons. First, averaging HRTFs across this database is not well-defined: the ARI subjects are measured on heterogeneous spatial grids, so their impulse responses cannot be combined element-wise, and naive

FOA to Binaural Conversion Pipeline

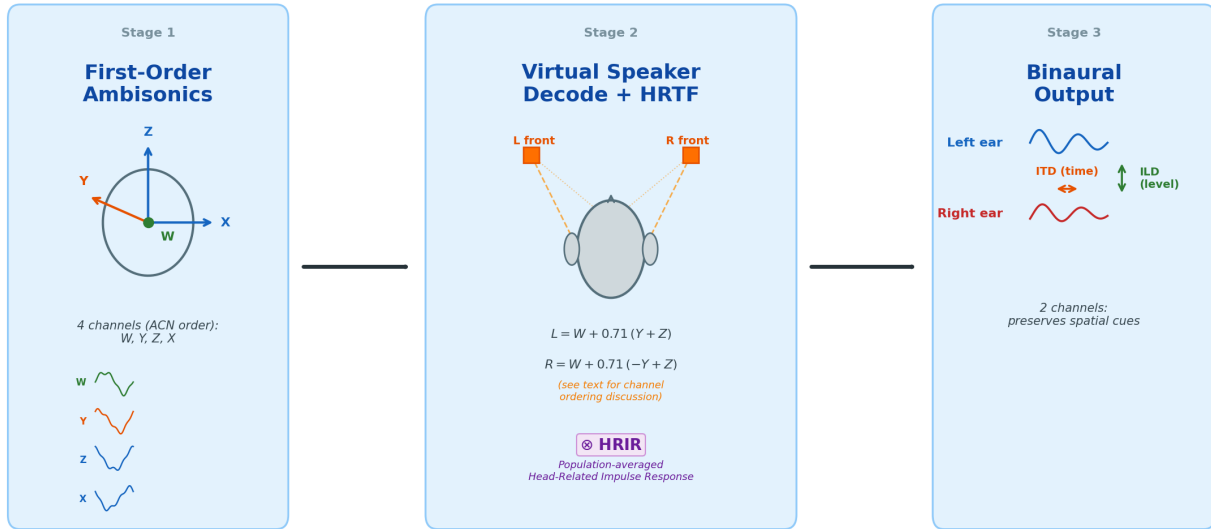


Figure 4.2: Three-stage binaural conversion pipeline. First-Order Ambisonics channels are decoded to virtual left and right speakers, convolved with HRIRs, and summed.

averaging across differing head geometries smears the interaural time delays, degrading the very spatial cues the rendering is meant to preserve. Second, and more fundamentally, the downstream ISO 532-3 loudness model integrates energy within ERB-scale (Cam) auditory filters, discarding the fine listener-specific spectral detail (pinna notches, torso reflections) that distinguishes one individual HRTF from another. The cue that survives to the loudness features is the broadband interaural level difference, which is set by the amplitude-panning decode geometry rather than by listener-specific HRTF fine structure. A single representative HRTF is therefore expected to be sufficient for this pipeline, although the sensitivity of the resulting features to the choice of subject was not tested directly here.

4.5.2 FOA to Binaural Decoding

The First-Order Ambisonics (FOA) recordings in STARSS23 contain four channels in the ambiX/SN3D convention: W (omnidirectional pressure), Y (left-right velocity), Z (up-down

velocity), and X (front-back velocity), stored as channels 0 to 3 respectively [25]. A binaural representation is required to feed the ISO 532-3 specific loudness model, which operates on left/right ear-input pairs.

The decode chosen for the main loudness pipeline is a two-virtual-speaker amplitude-panning configuration with speakers positioned at $\pm 45^\circ$ azimuth in the horizontal plane:

$$L = W + \alpha (X + Y) \quad (4.1)$$

$$R = W + \alpha (X - Y) \quad (4.2)$$

where $\alpha = 1/\sqrt{2} \approx 0.7071$. This gain is fixed by the geometry rather than chosen freely: for virtual speakers at azimuth $\pm\phi_s$, the first-order amplitude-panning weights are $\cos\phi_s$ on the front-back component X and $\sin\phi_s$ on the left-right component Y , so at $\phi_s = 45^\circ$ they coincide at $\cos 45^\circ = \sin 45^\circ = 1/\sqrt{2}$. The $\pm 45^\circ$ placement is a symmetric frontal pair that splits the frontal plane evenly and keeps both virtual speakers in the frontal hemisphere, consistent with the frontal left/right HRIRs; wider spacing would raise the ILD magnitude but move the speakers toward lateral positions, away from the symmetric frontal pair assumed here. Each speaker signal is then convolved with the corresponding ear's HRIR from the single representative HRTF described in the preceding subsection.

Pre-HRIR Interaural Level Difference

For a unit-amplitude FOA point source at azimuth ϕ and elevation θ in the ambiX convention, the spherical-harmonic projections are $W = 1$, $X = \cos\theta \cos\phi$, $Y = \cos\theta \sin\phi$, and $Z = \sin\theta$. Substituting into Equations 4.1–4.2 yields a pre-HRIR interaural amplitude difference—a linear signal difference, not a decibel-valued ILD—of

$$\text{ILD}_{\text{pre-HRIR}} = L - R = 2\alpha Y = \sqrt{2} \cos\theta \sin\phi. \quad (4.3)$$

The decode therefore produces a horizontal-azimuth gradient that varies as $\sin\phi$ for sources in the horizontal plane and attenuates as $\cos\theta$ with elevation. The X channel contributes equally to both ears and cancels in the difference, while Z does not enter the decode at all.

By construction, the binaural representation contains no front-back or up-down spatial cues prior to HRIR convolution.

Rationale for a Minimal Spatial Representation

The two-speaker configuration is below the $(N + 1)^2 = 4$ minimum required for a mathematically complete first-order ambisonics decode. It is therefore a deliberate sub-minimal choice rather than a standard playback configuration, motivated by two considerations specific to a loudness-feature pipeline.

Compatibility with the loudness feature pipeline. ISO 532-3 specific loudness integrates energy across approximately 150 equivalent rectangular bandwidth (ERB) auditory filters, with each filter output further smoothed in time. This band integration attenuates the fine spectral detail (pinna-notch zeros, head-shadow comb structures at specific frequencies) on which higher-order binaural rendering relies for elevation and front-back discrimination [27]. The spatial information that survives the loudness front-end is dominated by broadband interaural level differences. An amplitude-panning decode encodes horizontal source position directly as broadband ILD (Equation 4.3), aligning the decode’s spatial coding with what the loudness representation can carry through to the network.

Alignment with the asymmetry of human spatial hearing. Human spatial hearing is markedly asymmetric across spatial axes. The minimum audible angle (MAA) in the horizontal plane is approximately 0.97° at the frontal midline, while in the vertical (median) plane the MAA threshold rises to 3.65° under matched conditions [31] — a factor of approximately 3.75 in resolving power. Front-back disambiguation, which relies primarily on monaural pinna-spectrum cues rather than interaural difference cues [27, 28], exhibits a 6% confusion rate even in free-field conditions with broadband stimuli and ecologically-valid head pointing [32], and is further elevated when individualized HRTFs are replaced by averaged or generic ones [33]. This combination of high horizontal acuity, coarser vertical

resolution, and substantial front-back ambiguity — even under ideal listening conditions — characterises the spatial-information landscape the auditory system itself navigates.

A binaural decode that concentrates the spatial signal in the horizontal dimension is therefore aligned with the axis along which both the auditory system has the highest resolving power and the loudness representation is most likely to preserve the signal.

Implementation Details

Dynamic peak matching is applied between input and output to preserve the original FOA signal level, followed by clipping to $[-1, 1]$ for 16-bit PCM storage. The resulting binaural WAV files (2 channels, 24 kHz) serve as the input to the loudness feature pipeline.

4.5.3 Loudness Computation

Calibration Considerations

ISO 532 loudness models require input signals calibrated in physical units (Pascals), but the STARSS23 dataset provides only WAV files with no calibration metadata. The microphone sensitivity, preamplifier gain, and voltage-to-digital scaling used during recording are not documented in the dataset or accompanying paper. Since WAV format encodes audio relative to a digital full scale (± 1.0 for floating-point, or $\pm 2^{N-1}$ for N -bit integer), the absolute sound pressure level of the recordings is unknown.

This means an assumption must be made to map the WAV sample values to the sound pressure levels that the loudness models expect. Both the initial Artemis-based processing and the subsequent Python-based pipeline used an arbitrarily chosen calibration scale. In the Artemis workflow, an arbitrary sensitivity was assigned; in the Python implementation using the LoudnessModel library, the calibration parameter dB_{max} (the RMS sound pressure level of a full-scale sinusoid) was set to 92 dB SPL.

This arbitrary calibration is acceptable for the machine learning application because the CNN classifier operates on *relative* patterns in the specific loudness representation, not on

absolute loudness values. A shift in the assumed calibration level does not rescale specific loudness uniformly: both loudness models are compressive and level-dependent, so a change in input level alters bands by differing amounts, most strongly near threshold. What the fixed assumption does guarantee is internal consistency — every recording is processed under the same calibration, so the differences between sound events are computed on a common scale. The sensitivity of the reported results to the choice of calibration level was not tested here, and remains a limitation of this work.

To verify consistency between the Artemis and Python implementations, the Python pipeline’s output was validated against Artemis on the same input file, achieving a Pearson correlation of $r = 0.99$ for the specific loudness matrix. The magnitude ratio between the two implementations was approximately 1.1, consistent with slightly different calibration assumptions, and confirming that the spectral shape and temporal structure of the loudness representation are faithfully reproduced.

ISO 532-1 Processing

Each ear channel was processed independently using the `mosquito` Python library. The signal was resampled from 24 kHz to 48 kHz as required by the standard. The output is specific loudness in sone/Bark across 240 Bark bands at 2 ms temporal resolution, which was downsampled to 10 ms by averaging every 5 frames.

ISO 532-3 Processing

Both ear channels were processed jointly using the `LoudnessModel` library [34], a Python implementation of the Moore et al. (2016, 2018) auditory model [35, 36]. The signal was resampled from 24 kHz to 32 kHz as required by the model, and a free-field transfer function was applied. The output is per-ear specific loudness in sone/Cam across 150 ERB bands at 1 ms resolution, including binaural inhibition (cross-ear suppression), which was downsampled to 10 ms.

Sample Rate and Bandwidth Considerations

The STARSS23 recordings are natively sampled at 24 kHz, yielding a Nyquist frequency of 12 kHz. Both loudness models require upsampling: ISO 532-1 to 48 kHz and ISO 532-3 to 32 kHz. These minimum sample rates are required by the library implementations because the critical band filterbank coefficients, temporal integration step sizes, and other signal processing parameters are defined at these rates. The requirement does not imply that the input signal must contain wideband energy; a narrowband signal upsampled to the required rate is valid input, and frequency bands above the original Nyquist limit will correctly compute near-zero specific loudness. The upsampling is therefore a necessary technical step that does not distort the perceptual representation of the original signal content. Since 48 kHz also satisfies the ISO 532-3 requirement of ≥ 32 kHz, one might ask why both models were not run at a common 48 kHz rate. The answer is that upsampling cannot recover spectral content above the original 12 kHz Nyquist limit; the additional frequency bins would contain only interpolation artifacts, not genuine signal energy. Running ISO 532-3 at 48 kHz rather than 32 kHz would therefore increase computation time without adding meaningful information to the loudness representation. Both standards ultimately see identical underlying spectral content (0 to 12 kHz), and the performance differences reported in this thesis reflect genuine differences in the perceptual models rather than bandwidth artifacts from the resampling.

Artemis Workflow and Transition to Python

An initial version of the loudness processing was performed using HEAD Acoustics Artemis, a commercial psychoacoustic analysis tool. Due to Artemis’s limitation on the maximum number of data points per analysis window, long recordings (some exceeding 5 minutes) were processed in overlapping chunks that were stitched back together after computation, with careful alignment to remove onset ramp artifacts at chunk boundaries. The Python-based pipeline was subsequently developed to enable fully automated, reproducible processing without the manual overhead and chunk-size limitations of the GUI-based Artemis workflow.

4.5.4 Segmentation and Channel Handling

Loudness was computed on full-length recordings before segmentation. This ordering is critical: the temporal integration stage of both loudness models includes onset ramps (attack time constants of approximately 22 ms) that model the perceptual response to the onset of a sound. Computing loudness on pre-segmented 1.27-second clips would reintroduce this onset transient at every segment boundary, creating a comb-filter-like artifact in the time domain that does not reflect continuous auditory perception.

After full-length loudness computation, the specific loudness time series was segmented into 1.27-second windows with a 1.2-second stride, producing 128 time frames per segment at 10 ms resolution. This segmentation exactly matches the STARSS23 baseline framework, ensuring temporal alignment with the metadata labels.

Left and right ear specific loudness channels were kept separate (not summed) to preserve interaural level difference (ILD) spatial cues. While ISO 532-3 Section 7.9 specifies summing left and right long-term loudness to obtain total binaural loudness, this summation collapses the spatial information encoded in the L/R difference. For the machine learning application, preserving per-ear channels provides the network with the raw material to learn spatial patterns. The resulting input tensors have shape $(B, 2, F, 128)$ where F is 150 ERB bands (ISO 532-3) or 240 Bark bands (ISO 532-1), compared to $(B, 7, 257, 128)$ for the FOA baseline.

4.6 Model Architecture and Training

The AudioOnlyCRNN architecture from the DCASE 2023 Task 3 baseline was used for all experiments. The only modification was the number of input channels: 7 for the FOA baseline (4 amplitude spectrograms plus 3 inter-channel phase differences), 2 for the binaural loudness models. All other hyperparameters were held at the DCASE 2023 baseline values:

- Training iterations: 20,000

- Batch size: 16
- Optimizer, learning rate schedule, and loss function: identical to the DCASE 2023 baseline configuration.

All training was performed on the same hardware to ensure comparability:

- **GPU:** NVIDIA RTX 3090 (24 GB VRAM)
- **CPU:** AMD Ryzen 9 3900X (12 cores, 24 threads)
- **RAM:** 64 GB DDR4 at 2133 MT/s (Corsair)
- **OS:** Ubuntu 24.04.2 LTS

Table 4.2 summarizes the computation times for feature processing and model training across all configurations.

Table 4.2: Computation times for feature processing and model training.

Stage	Duration	Throughput
<i>Feature Processing (168 WAV files, total ~ 7.4 hrs audio)</i>		
ISO 532-3 loudness	~ 18 hours	$0.41 \times$ realtime
ISO 532-1 loudness	~ 24 hours	$0.31 \times$ realtime
<i>Model Training (20,000 iterations)</i>		
FOA baseline	~ 3 hr 28 min	~ 1.6 it/s
ISO 532-3, 2ch	~ 9 min	~ 37 it/s
ISO 532-1, 2ch	~ 9 min	~ 37 it/s

Feature-processing throughput is expressed as processed audio per unit wall-clock time, so values below unity indicate slower-than-realtime processing. These wall-clock totals are

not directly comparable with the single-clip benchmarks of Table 5.2: the ISO 532-3 run was distributed across 12 CPU cores, both models processed two-channel binaural audio rather than mono, and the long STARSS23 recordings incurred chunking and disk-I/O overhead. These pipeline differences, rather than the intrinsic per-clip cost of the two models, dominate the totals reported here.

The large difference in training speed between the FOA baseline and the loudness models is attributable to data loading, not model complexity. The FOA baseline computes STFT features on-the-fly from raw WAV files during each training iteration, while the loudness models load pre-computed feature arrays stored in NumPy’s compressed `.npz` format (NPZ). Both approaches perform the same number of gradient updates with the same batch size.

4.7 Results: Standard 20° Threshold

4.7.1 Understanding the SELD Metrics

The four SELD metrics may appear unusual compared to standard machine learning evaluation metrics, so a brief explanation is warranted before presenting results.

Error Rate (ER) is adapted from speech recognition’s word error rate. It is computed as $(\text{Substitutions} + \text{Insertions} + \text{Deletions}) / \text{Total Reference Events}$. Unlike a standard 0-to-1 error rate, ER can exceed 1.0 when the system produces more errors than there are actual events (for example, through false positive detections). Even state-of-the-art systems on this dataset achieve ER values around 1.0, reflecting the genuine difficulty of the task.

F-score (F) is location-sensitive: a detection counts as a true positive only if the predicted class matches the ground truth AND the predicted direction is within 20 degrees of the true direction. This is an extremely strict criterion. In standard classification, getting the class right suffices. Here, correctly identifying “female speech” but placing it 25 degrees from its true location counts as a miss. To put this in perspective, a system making random predictions achieves $F \approx 0.00077$, so even an F-score of 0.07 represents performance roughly 100 times better than chance [25].

Localization Error (LE) is the mean angular error in degrees between predicted and true directions, computed only for correctly detected events. Lower is better.

Localization Recall (LR) is the fraction of ground truth events that were correctly detected and localized. Higher is better.

SELD Error is a composite metric combining all four: $\text{SELD} = \text{mean}(\text{ER}, 1-\text{F}, \text{LE}/180, 1-\text{LR})$. Lower is better.

4.7.2 Standard Threshold Results

Table 4.3 presents the SELD evaluation metrics at the standard 20° spatial threshold.

Table 4.3: SELD results on STARSS23 at the standard 20° threshold. All systems are audio-only. The FOA paper baseline and our FOA reproduction use 4-channel first-order ambisonics; the loudness models use 2-channel binaural input. Lower ER and LE are better; higher F and LR are better.

Model	ER↓	F↑	LE (°)↓	LR↑
FOA paper, audio-only	0.97	0.148	55.9	0.351
FOA reproduction (ours)	1.18	0.110	54.9	0.440
ISO 532-3 (2 ch)	1.03	0.07	58.3	0.31
ISO 532-1 (2 ch)	1.17	0.09	56.1	0.31

At the standard 20° threshold, the 2-channel ISO 532-3 binaural loudness model does not reach the published 4-channel audio-only FOA baseline ($\text{ER} = 0.97$). Its error rate (1.03) is, however, lower than our own FOA reproduction (1.18): within an identical training and evaluation pipeline, the loudness front-end makes fewer total errors. This is not, by itself, a detection advantage. The absolute F-scores in Table 4.3 are all small because the metric is strict (correct class *and* direction within 20°): even the published FOA audio-only baseline reaches only 0.148, and a random system scores ≈ 0.00077 , so the loudness models’ 0.07–0.09 is of the same order—roughly 100 times chance—rather than a failure. Crucially, the loudness F-scores (0.07–0.09) and localization recall ($\text{LR} = 0.31$) fall below both the published

baseline and our reproduction ($F = 0.110$, $LR = 0.440$): the model detects and recalls fewer events. The result is best described as a *conservative detector*—fewer spurious detections lower the error rate, but fewer catches lower F and recall—rather than a system that matches first-order ambisonics. The following multi-threshold analysis tests whether this F deficit is an artifact of coarse localization.

4.8 Multi-Threshold Analysis

To disentangle detection performance from localization precision, the evaluation was repeated at multiple spatial thresholds: 20° , 45° , 90° , and 180° (the last being equivalent to detection-only evaluation, since all angular errors are $\leq 180^\circ$ —even a full left–right reversal falls within tolerance, so no localization error can cause a miss at this threshold). Results are shown in Table 4.4.

Table 4.4: Multi-threshold analysis: F-score and error rate at varying spatial thresholds. At 180° , the evaluation is detection-only.

Threshold	ISO 532-3	ISO 532-1	FOA
<i>F-score</i> $\uparrow$			
20°	0.07	0.09	0.11
45°	0.17	0.17	0.22
90°	0.27	0.24	0.31
180° (det.)	0.31	0.29	0.34
<i>Error rate</i> $\downarrow$			
20°	1.03	1.17	1.18
45°	0.88	1.01	0.99
90°	0.77	0.88	0.90
180° (det.)	0.71	0.79	0.85

The multi-threshold analysis does *not* support reading the F -score deficit as a localization

artifact. If the loudness model’s low F merely reflected coarse angular resolution, the gap to FOA would close once direction is ignored. It does not: at the 180° (detection-only) threshold, where even a full left–right reversal counts as a hit, the ISO 532-3 F -score (0.31) still trails our FOA reproduction (0.34), and the F gap stays roughly constant (≈ 0.03 – 0.05) across all four thresholds. The deficit is therefore in *detection recall*—the model catches fewer events—not in localization precision.

What the loudness model does gain at the detection-only threshold is a lower error rate (0.71 vs. 0.85 for our FOA reproduction), driven by fewer false alarms. This follows from the definition of ER: it counts substitutions, insertions and deletions against the reference total, so the loudness model’s lower localization recall implies *more* deletions, and a lower ER can then arise only from a larger reduction in insertions and substitutions. The evaluation used here does not report a per-error-type breakdown, so the split between those two terms is not resolved. Taken together, these observations describe a more conservative detector: it triggers less often, which reduces insertion errors (lowering ER) but also misses more events (lowering F and recall). Because this character is roughly constant across spatial thresholds, it is a property of detection sensitivity rather than of localization precision. The binaural loudness features do encode spatial information—via ILD and cross-ear inhibition—but with lower angular resolution than the phase-difference cues in 4-channel FOA, consistent with the slightly higher localization error ($LE = 58.3^\circ$ vs. 54.9°).

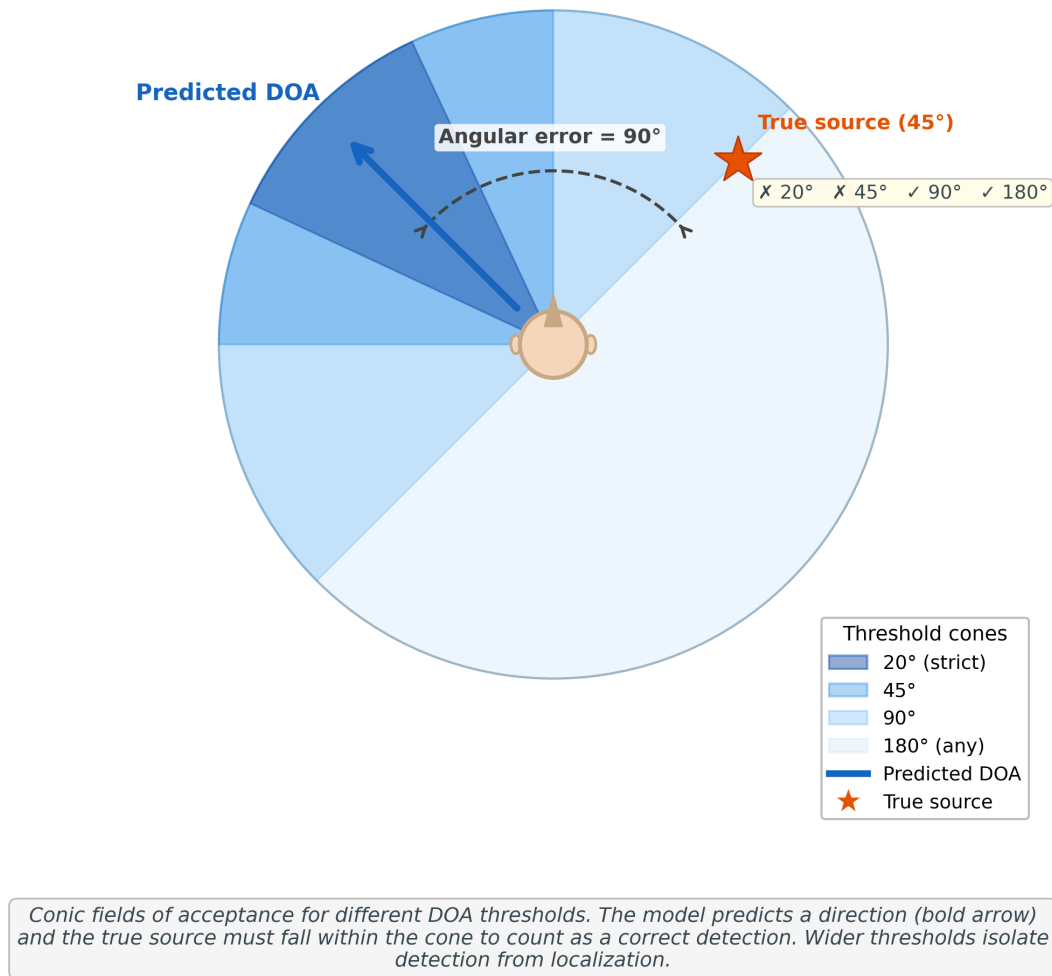


Figure 4.3: Visualization of the multi-threshold analysis. The model predicts a direction (bold arrow at 135°) and the true source is at 45°, yielding a 90° angular error. Concentric conic fields show acceptance regions for different thresholds. At strict thresholds (20°, 45°) this prediction counts as a miss; at relaxed thresholds (90°, 180°) it counts as a hit, isolating detection from localization performance.

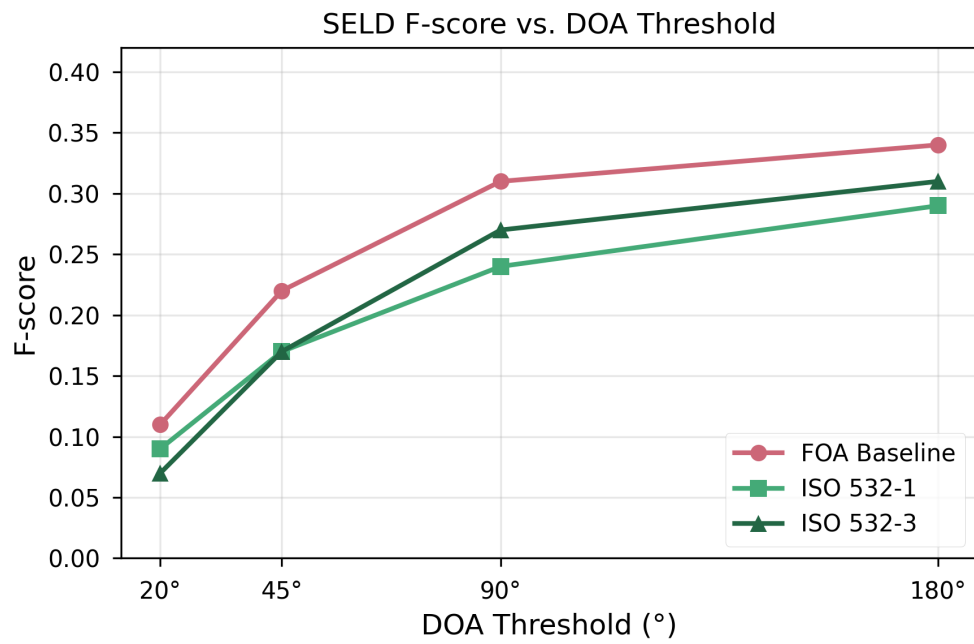


Figure 4.4: F-score as a function of DOA threshold for three feature representations. The loudness curves remain below the FOA reproduction at every threshold, and the gap does not close even at 180°, where direction is ignored (for ISO 532-3 it stays within 0.03–0.05)—indicating that the loudness model’s lower F reflects detection recall rather than localization precision.

4.9 Detection-Only Class-Wise Analysis

Table 4.5 lists per-class F-scores at the 180° (detection-only) threshold for the two binaural loudness models and the FOA reproduction, across all 13 STARSS23 classes.

Table 4.5: Detection-only (180°) per-class F-scores for all 13 STARSS23 classes. Bold indicates the highest value in each row.

Class	532-3 F	532-1 F	FOA F
Female speech	0.59	0.58	0.61
Male speech	0.66	0.68	0.69
Clapping	0.27	0.27	0.30
Telephone	0.07	0.09	0.09
Laughter	0.12	0.14	0.13
Domestic sounds	0.68	0.65	0.57
Footsteps	0.01	0.01	0.11
Door	0.13	0.26	0.21
Music	0.58	0.48	0.52
Musical instrument	0.45	0.37	0.38
Water tap	0.10	0.14	0.59
Bell	0.39	0.06	0.22
Knock	0.00	0.00	0.00

Even with localization removed, the FOA reproduction detects slightly more events overall (macro-F 0.34 vs. 0.31 for ISO 532-3) and wins the most individual classes (five of thirteen)—but the wins are distributed unevenly. ISO 532-3 leads on four classes, all sustained or tonal sounds: domestic sounds (+0.11 over FOA), bell (+0.17), musical instrument (+0.07), and music (+0.06). Bell is the single largest loudness advantage, consistent with the interpretation that the equal-loudness weighting and masking of the loudness model sharpen

tonal structure.

FOA’s own wins, by contrast, are mostly narrow—female and male speech by only +0.02 and +0.03, and clapping by +0.03—with two decisive exceptions: water tap (+0.49) and footsteps (+0.10), broadband and transient sounds that benefit from FOA’s finer spatial-spectral detail. The loudness model’s detection strength is thus confined to sustained and tonal classes, while its largest deficits are on broadband transients. This resembles, but does not simply repeat, the pattern seen in ESC-50 classification: footsteps is one of the larger ESC-50 loudness wins (Section 3.4) yet is a loudness deficit here, so the class-level pattern does not transfer intact between the two tasks.



Figure 4.5: Error rate at the 180° threshold (detection-only, no location constraint). The ISO 532-3 loudness model attains the lowest error rate—fewer total errors, consistent with fewer false alarms—using only 2 channels versus 4 for FOA. As the F-scores in Table 4.4 show, this reflects more conservative detection rather than higher recall.

Chapter 5

DISCUSSION

5.1 *The Loudness Advantage: The Magnitude-Axis Transform*

The results from both experiments converge on the same conclusion: the primary advantage of specific loudness features over mel spectrograms lies in a perceptual transform of the *magnitude* axis, not the frequency axis alone. In the ESC-50 experiment, adding a perceptual transform of the frequency axis (mel, 53.1%) over linear frequency (STFT, 39.7%) yields a 13.4 percentage point improvement. Adding a perceptual transform of the magnitude axis on top (specific loudness, 63.0%) yields a further 9.9 points. While frequency scaling contributes meaningfully, the level dimension provides substantial additional discriminative power that the mel spectrogram does not capture.

The level-dependent weighting in loudness models implements several perceptual mechanisms simultaneously:

- **Equal-loudness contour weighting:** Redistributes energy across frequency bands according to the ear’s frequency-dependent sensitivity, effectively emphasizing frequency regions where the human auditory system is most sensitive.
- **Compressive nonlinearity:** The conversion from excitation to specific loudness applies a compressive power law that mimics the basilar membrane’s response, providing automatic gain control.
- **Spectral masking:** The spread of excitation along the frequency axis models the masking interactions between simultaneous spectral components.

5.2 Which Classes Benefit

Whether the loudness transform helps a class depends not on its temporal character but on *what the transform modifies* relative to the cues that define the class. The specific loudness models reshape the input in a few ways: equal-loudness weighting raises sensitivity around the 3–4 kHz hearing peak and lowers it at low frequencies; masking and compression sharpen strong spectral peaks and compress the dynamic range; and critical-band integration with temporal smoothing (attack time constants of roughly 20–50 ms, release 50–100 ms) trades fine spectro-temporal detail for perceptual salience. A class benefits when its defining cue is preserved or enhanced by these operations, and suffers when that cue is attenuated or smoothed away.

This cuts against a simple sustained-versus-transient reading. Loudness helps the *transient* high-frequency chirps of crickets and impulsive ticks of a clock—their mid-to-high-frequency periodic structure survives the transform—as well as the sustained harmonics of bells and the broadband balance of thunder. It hurts the *sustained* washing machine, whose defining low-frequency hum is attenuated until it flattens into broadband noise, and the clock alarm, whose sharp tonal onset relies on fine detail the critical-band and temporal smoothing blur. Temporal integration removes fine *onset* detail but does not discard temporal structure wholesale: the time-varying specific loudness retains discriminative periodicity at its frame rate, as the clock-tick result shows.

The SELD experiment is consistent once its losers are read the same way. Speech is better served by the fine spatial detail of four-channel FOA phase features than by two-channel loudness—a spatial limitation, not a temporal one—while water taps lack the mid-frequency periodic or harmonic structure that loudness sharpens. Footsteps is the clearest exception to any single cross-task account: a substantial loudness win in ESC-50 classification but a loudness deficit in SELD detection, which suggests its SELD deficit arises from the spatial task rather than from the feature’s spectral character. The thread across both experiments is spectral rather than temporal: loudness features help where discriminative energy lies in

the mid-to-high-frequency, periodic, or harmonic structure the auditory transform preserves, and hurt where it lies in low-frequency energy or fine detail the transform reshapes away.

5.3 *Detection Versus Localization*

The multi-threshold analysis in the SELD experiment characterizes how the binaural loudness front-end behaves as a detector. At the detection-only threshold (180°), the ISO 532-3 model reaches a lower error rate than our FOA reproduction (0.71 vs. 0.85), consistent with its raising fewer false alarms. It does not, however, catch more events: its F-score and recall remain below FOA at every threshold. The loudness representation therefore trades detection sensitivity for a lower false-alarm rate—a conservative detector—while its angular localization is also slightly coarser than FOA’s.

This trade-off has a clear physical explanation. The FOA format encodes spatial information through phase differences between the W, X, Y, and Z channels of the ambisonics representation. These phase differences provide continuous, high-resolution directional information. Binaural loudness, in contrast, encodes spatial information primarily through interaural level differences (ILD), which are frequency-dependent and provide coarser angular resolution. The ISO 532-3 model’s binaural inhibition stage could in principle contribute an additional spatial cue, but it would operate on the overall loudness difference between ears rather than on fine-grained phase structure, and its contribution was not isolated in these experiments.

The multi-threshold analysis (Section 4.8) sharpens this picture. As the localization tolerance is relaxed, the error-rate advantage of the loudness model over our FOA reproduction persists ($ER = 0.71$ vs. 0.85 at detection-only), but the F-score gap does *not* close—so the two metrics are measuring different things. Detection and localization are served by different properties of the representation: the broadband level structure that survives the loudness front-end lowers false alarms, while the fine interchannel phase structure that FOA preserves supports resolving *where* a sound is. Reporting the four metrics separately, rather than as a single aggregate, is what makes this split visible: a composite score would average the

loudness model’s lower error rate against its lower recall and obscure both.

5.4 Pipeline Alignment

The binaural pipeline points to a broader design principle: each stage of a feature extraction pipeline should be aligned with what the *downstream* stage can exploit.

An apparent tension arises from the thesis results taken as a whole. A more sophisticated loudness model improves ML performance (Chapter 3), yet the two-channel binaural loudness pipeline does not match the four-channel FOA baseline for SELD (Chapter 4). This seems paradoxical: if perceptual processing helps, why does more of it not help here?

The resolution lies in the role each component plays. The loudness model *is* the feature extractor; it directly determines what the CRNN sees. Making it more perceptual improves the representation by compressing irrelevant variation (through equal-loudness weighting and masking) while preserving discriminative structure.

The binaural decode, in contrast, is *upstream* of the feature extractor. HRTF-based rendering encodes spatial information as frequency-dependent spectral notches and peaks, fine detail that the loudness model’s critical-band integration largely averages away. The HRTF cues are simply not the type of information that critical-band loudness preserves. Amplitude panning, by contrast, encodes spatial information as broadband ILD, which survives critical-band integration because every band sees the same left-right level difference.

This principle suggests that the binaural decode result does not undermine the loudness thesis; it *extends* it. The loudness features are effective precisely because they capture perceptually relevant level differences. An upstream stage that encodes spatial information as level differences (amplitude panning) is therefore naturally aligned with the feature extractor. An upstream stage that encodes spatial information as spectral shape differences (HRTFs) is misaligned, because the loudness model compresses spectral shape in favor of level.

A testable prediction follows: if STFT features (which preserve fine spectral detail) were used instead of loudness features, HRTF-based binaural rendering might improve localization, because the model could learn the HRTF spectral patterns directly. This prediction is

Table 5.1: Pipeline alignment: each stage transition is classified by whether the upstream stage produces information the downstream stage can exploit.

Stage Transition	Aligned?	Reason
Loudness features → CRNN	✓	Perceptual features match human-labeled task categories
HRTF binaural → loudness	×	HRTF spectral cues attenuated by critical-band integration
Amplitude panning → loudness	✓	Broadband ILD survives critical-band integration

left for future work.

5.5 The Role of Binaural Inhibition

In the SELD experiment the ISO 532-3 model achieves a lower error rate than ISO 532-1 at both the standard and detection-only thresholds ($ER = 1.03$ vs. 1.17 , and 0.71 vs. 0.79). It is tempting to attribute this to the one processing stage unique to the binaural setting: ISO 532-3 processes both ears jointly and applies binaural inhibition (cross-ear suppression), whereas ISO 532-1 processes each ear independently. That attribution, however, is not supported by these experiments. The same ordering—ISO 532-3 ahead of ISO 532-1—also holds on *mono* ESC-50 classification (63.0% vs. 59.5%), where there is only one channel and cross-ear inhibition cannot operate. The advantage is therefore not specifically a binaural effect: the non-binaural refinements of ISO 532-3 relative to ISO 532-1—finer ERB-scale resolution, roex auditory filters, and its 1 ms output resolution—are sufficient to account for a leading position that is already present without any binaural processing.

Binaural inhibition remains a *plausible additional* contributor in the spatial setting, because it introduces a cross-ear dependency that the per-ear ISO 532-1 pipeline lacks. Two mechanisms could in principle help: it could encode interaural level differences in a physiologically motivated way, supplementing the raw level differences between channels; and by suppressing the quieter ear it could sharpen the contrast between sources at different lateral positions—a primitive form of spatial source separation. Neither mechanism was isolated here—doing so would require an ablation that runs the binaural ISO 532-3 pipeline with cross-ear inhibition disabled—so both are offered as hypotheses for future work rather than as demonstrated effects.

5.6 Mel Adoption as Path of Least Resistance

The results of this thesis suggest that the dominance of mel spectrograms in audio ML is a product of software ecosystem dynamics rather than perceptual optimality. The mel spectrogram is computed in a single function call in `librosa` [3], while specific loudness computation requires specialized software implementing complex psychoacoustic models. The adoption of

mel is a path-of-least-resistance phenomenon: researchers use the representation that is most readily available, not the one that best matches the perceptual categories they are trying to classify.

This observation is reinforced by the computational cost analysis in Table 5.2. The mel spectrogram is computed at roughly $650\times$ realtime, making it practical for large-scale pipelines. ISO 532-1 specific loudness is about three orders of magnitude slower than mel (roughly $1,000\times$), running just below realtime at $0.64\times$. While this is dramatically slower, it provides a 6.4 percentage point accuracy improvement on ESC-50, a favorable cost-benefit ratio. ISO 532-3 is roughly four orders of magnitude slower than mel (about $25,000\times$) due to its per-millisecond sequential auditory model, yielding a further significant 3.5 percentage points. Most striking, however, is that a simple power-law compression of the mel magnitude axis recovers roughly +6 of these points at essentially the cost of mel (Section 3.3.2), the most favorable cost-benefit ratio of all.

Table 5.2: Computation time benchmarks for feature extraction (5-second mono audio clips at 44.1 kHz, single-threaded, AMD Ryzen 9 3900X; steady-state means over 5–30 clips per feature). ISO 532-3 timing varies substantially across clips (mean ~ 195 s, standard deviation ~ 100 s over 5 clips), so its cost is best read as an order-of-magnitude figure.

Feature	Per File	Realtime Factor	vs Mel
STFT	0.0034 s	$1,480\times$	$0.4\times$
Mel (240 bands)	0.0077 s	$650\times$	$1.0\times$
ISO 532-1 (Bark)	7.8 s	$0.64\times$	$\sim 1,000\times$
ISO 532-3 (ERB)	~ 195 s	$0.026\times$	$\sim 25,000\times$

These numbers highlight the practical motivation for developing optimized loudness computation libraries. On cost-benefit grounds the lightweight magnitude-axis variants—power-law compression and equal-loudness weighting—are the most attractive drop-in mel replacements, recovering most of the accuracy benefit at essentially the cost of mel. Among the full standardized models, ISO 532-1 is the more computationally practical of the two, but at

roughly three orders of magnitude above mel it is better suited to offline feature extraction than to large-scale or real-time pipelines.

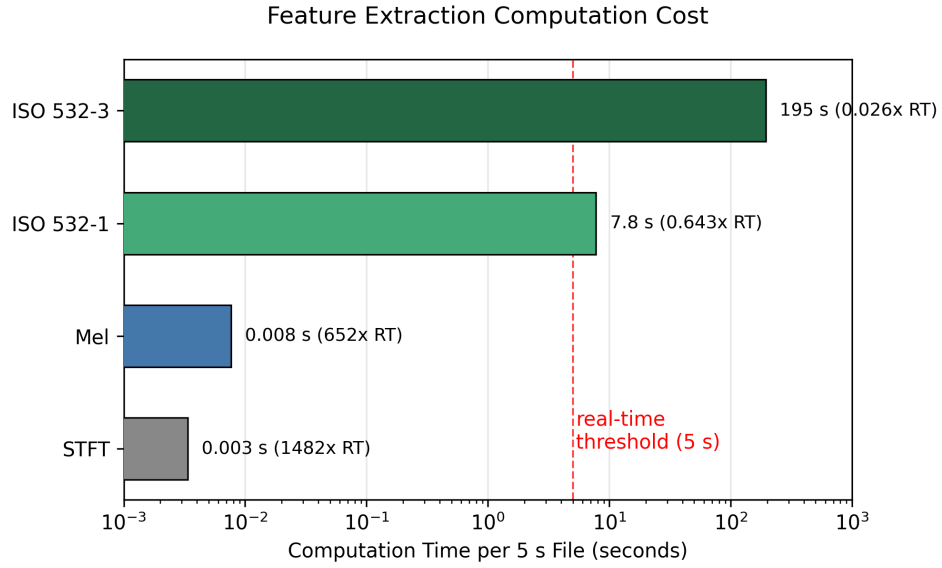


Figure 5.1: Wall-clock time for feature extraction per 5-second audio clip. Note the logarithmic scale. Realtime factors are annotated above each bar. Among the full ISO models, ISO 532-1 is the more practical of the two; the lightweight mel variants achieve most of the accuracy benefit at mel-level cost.

5.7 Annotation Methodology

The STARSS23 dataset’s annotation methodology deserves specific comment. Temporal annotations (onset and offset times) are produced by human listeners, a perceptual process, while spatial annotations (DOA coordinates) are derived from motion capture, a physical measurement. This hybrid annotation scheme means that the “ground truth” for detection is inherently perceptual, while the ground truth for localization is physical. A perceptually motivated representation might therefore be expected to align more naturally with the detection annotations than with the physically derived DOA annotations. The present results do

not test that prediction: the binaural loudness models trail the FOA reproduction on both F-score and localization error, and their relative advantage is confined to error rate. The distinction is nonetheless worth noting for future work that evaluates perceptual features against perceptually derived labels.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Summary of Findings

This thesis has investigated the use of specific loudness representations as input features for sound event classification and localization. The key findings are:

1. **Loudness features outperform mel and STFT for classification.** On ESC-50, ISO 532-1 specific loudness achieves 59.5% accuracy compared to 53.1% for mel and 39.7% for STFT (baseline CNN), a statistically significant improvement over mel (+6.4 pp, $p = 0.013$, $d_z = 1.92$).
2. **The advantage is due to the perceptual transform of the magnitude axis, not the frequency axis.** A magnitude-axis ablation on the shared mel filterbank recovers most of the loudness gain, and across architectures the magnitude transform helps universally whereas the frequency-axis step (STFT $\rightarrow$ mel) helps only the CNN and reverses on the other three architectures—isolating the level-dependent magnitude weighting as the robust source of the improvement.
3. **Binaural loudness does not match first-order ambisonics for SELD.** The 2-channel ISO 532-3 model does not reach the published 4-channel audio-only FOA baseline (ER=0.97); it obtains a lower error rate than our own FOA reproduction (1.03 vs. 1.18) but lower F-score and recall—a more conservative detector.
4. **The loudness model is a conservative detector, not a better localizer.** A multi-threshold analysis shows its lower error rate at detection-only (0.71 vs. 0.85 for our FOA reproduction) comes from fewer spurious detections, while its F-score stays

below FOA at every threshold—so the deficit is in detection recall ($LR = 0.31$ vs. 0.44), not localization precision.

5. **The loudness advantage is class-dependent.** Classes whose defining cue is mid-to-high-frequency, periodic, or harmonic structure show the largest benefit; those defined by low-frequency energy or by fine spectro-temporal detail the transform reshapes away show less benefit or a slight disadvantage.
6. **The full loudness models produce more consistent results.** The ISO 532-1 and ISO 532-3 features show markedly lower variance across ESC-50 folds than mel, suggesting more recording-condition-invariant features; the lightweight magnitude-axis variants do not share this property.

These findings support the thesis statement: for sound event classes defined by human perception, perceptual input representations that include level-dependent spectral weighting align more strongly with classification targets than representations based solely on physical signal properties.

6.2 *Future Work*

Several directions for future research emerge from this work:

1. **Controlled stimulus experiments.** A study using synthetic stimuli in an anechoic chamber would allow precise control over the acoustic variables (level, spectrum, spatial position) and enable direct comparison of model predictions with human perceptual judgments.
2. **Open-source loudness library.** A Python library implementing ISO 532-1 and ISO 532-3 with a `librosa`-compatible API would lower the barrier to adoption and enable systematic comparison across research groups.

3. **Faster loudness computation.** The main practical barrier to adopting specific loudness features is computational cost: the full ISO 532-3 model runs at roughly $0.026\times$ realtime (Table 5.2), about four orders of magnitude slower than mel. Because the models are largely a per-band, per-frame sequence of fixed operations, substantial speed-ups should be achievable through batched, vectorized, or GPU implementations, and through approximations that retain the compressive magnitude mapping. Reducing this cost—the single largest downside of the approach—would do the most to make perceptual loudness features a practical default.
4. **Larger-scale and pre-trained architectures.** The experiments here train every model from scratch. Extending to large pre-trained audio models (for example PANNs or AST) and to substantially larger corpora such as AudioSet would test whether the magnitude-axis advantage survives large-scale pretraining. The multi-architecture results (Section 3.5) suggest that more capable architectures amplify the loudness advantage.
5. **Hybrid representations.** Combining loudness features with FOA phase features in a multi-stream architecture could capture both the detection advantage of loudness and the localization precision of phase-based spatial encoding.
6. **State space models for SELD:** Recent state space models such as Mamba offer linear-complexity sequence modeling that may be particularly suited to longer audio recordings in SELD applications, where tracking sound events over minutes-long recordings requires capturing long-range temporal dependencies that current CRNN architectures handle less efficiently.

Appendix A

SOFTWARE AND REPRODUCIBILITY

A.1 Software Environment

Experiments were run in two separate Python environments: the DCASE 2023 audio-visual SELD baseline environment, whose versions are fixed by that project’s installation instructions, and a feature-extraction environment used for the loudness and spectrogram front ends. The two are listed separately because their dependency pins are mutually incompatible—the loudness stack requires a newer Python than the SELD baseline permits.

Table A.1: Software and library versions, by environment.

Component	Version	Purpose
<i>SELD training environment (DCASE 2023 baseline)</i>		
Python	3.8	Programming language
PyTorch	1.7.1	Deep learning framework
torchvision	0.8.2	Vision utilities
torchaudio	0.7.2	Audio utilities
culatoolkit	11.0	GPU acceleration
pandas	1.5.3	Data handling
<i>Feature-extraction environment (loudness front ends)</i>		
Python	≥ 3.9	Required by numpy 1.26 / scipy 1.11
LoudnessModel	Slaney impl.	ISO 532-3 loudness computation
mosquito	1.2	ISO 532-1 loudness computation
librosa	0.10	Audio I/O, mel spectrograms
numpy	1.26.3	Numerical computation
scipy	1.11.4	Signal processing, resampling
pysofaconventions	0.3	SOFA file I/O for HRTF data
scikit-learn	1.3	Evaluation metrics

A.2 Hardware

- GPU: NVIDIA RTX 3090 (24 GB VRAM)
- CPU: AMD Ryzen 9 3900X (12 cores)
- RAM: 64 GB DDR4

A.3 Validation

The ISO 532-3 implementation was validated against the HEAD acoustics Artemis software suite, which implements the same standard. On a common input file, the Pearson correlation between the Python implementation and Artemis was $r = 0.99$ for the specific loudness matrix; the magnitude ratio between them was approximately 1.1, attributable to differing calibration assumptions. The two implementations therefore agree closely in spectral shape and temporal structure rather than in absolute level.

A.4 Code Availability

The source code for all experiments, including feature extraction, model training, and evaluation scripts, is available at the thesis author’s GitHub repository.

BIBLIOGRAPHY

- [1] E. Rosch. Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, 1978.
- [2] S. S. Stevens, J. Volkman, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.
- [3] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–24, 2015.
- [4] International Organization for Standardization. Acoustics — methods for calculating loudness — Part 1: Zwicker method. Technical Report ISO 532-1:2017, ISO, 2017.
- [5] International Organization for Standardization. Acoustics — methods for calculating loudness — Part 3: Moore–Glasberg–Schlittenlacher method. Technical Report ISO 532-3:2023, ISO, 2023.
- [6] E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248, 1961.
- [7] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1–2):103–138, 1990.
- [8] International Organization for Standardization. Acoustics — normal equal-loudness-level contours. Technical Report ISO 226:2003, ISO, 2003.
- [9] K. J. Piczak. ESC: Dataset for environmental sound classification. *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, 2015.
- [10] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

- [11] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech*, 2021.
- [12] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *Auditory Physiology and Perception*, pages 429–446. Pergamon, 1992.
- [13] Kevin G. R. Jacome and Fabricio L. Grijalva. Sound events localization and detection using bio-inspired gammatone filters and temporal convolutional neural networks. 2023.
- [14] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals. Learning the speech front-end with raw waveform CLDNNs. In *Proceedings of Interspeech*, pages 1–5, 2015.
- [15] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
- [16] C. Kim and R. M. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1315–1329, 2016.
- [17] A. Mesaros, T. Heittola, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2017.
- [18] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 29, pages 684–698, 2020.
- [19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.
- [20] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. In *IEEE Signal Processing Letters*, volume 24, pages 279–283, 2017.
- [21] A. Ito and M. Tsuchiya. Effects of loudness normalization on audio event detection. *Applied Acoustics*, 176:107856, 2021.

- [22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [23] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1041–1044, 2014.
- [24] A. Mesaros, T. Heittola, and T. Virtanen. A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 9–13, 2018.
- [25] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Imoto, A. Mesaros, T. Virtanen, and Y. Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2024.
- [26] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen. Joint measurement of localization and detection of sound events. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 333–337, 2019.
- [27] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, revised edition, 1997.
- [28] John C. Middlebrooks and David M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42:135–159, 1991.
- [29] Frederic L. Wightman and Doris J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.
- [30] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. The CIPIC HRTF database. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, 2001.
- [31] David R. Perrott and Kourosh Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.
- [32] James C. Makous and John C. Middlebrooks. Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

- [33] Frederic L. Wightman and Doris J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, 85(2):868–878, 1989.
- [34] Malcolm Slaney and Jeremy JX Hsiao. Python code for calculation of the loudness of time-varying sounds. <https://github.com/MalcolmSlaney/LoudnessModel>, 2024.
- [35] Brian C. J. Moore, Brian R. Glasberg, Aarti Varathanathan, and Josef Schlittenlacher. A loudness model for time-varying sounds incorporating binaural inhibition. *Trends in Hearing*, 20:1–16, 2016.
- [36] Brian C. J. Moore, Michael Jervis, Lorna Harries, and Josef Schlittenlacher. Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition. *Journal of the Acoustical Society of America*, 143:1504–1513, 2018.