forum acusticum 2026
8-12 September · Graz, Austria

ON THE EFFECTS OF TIGHTNESS AND FREQUENCY RESOLUTION IN BIOACOUSTIC CLASSIFICATION

Clara Hollomey[1], Angela Stöger-Horwath[2], Daniel Haider[3], Matthias Zeppelzauer[1]

[1] *University of Applied Sciences St. Pölten, Austria*
[2] *University of Vienna, Austria, Vienna Zoo, Austria*
[3] *Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

We investigate whether the frame-theoretic properties of filterbanks can predict downstream neural network classification performance. Using a dataset of 3 080 African savanna elephant rumble vocalisations classified into four age groups, we train a deliberately small convolutional neural network on magnitude spectrograms derived from 15 filterbank configurations spanning equivalent rectangular bandwidth (ERB), constant-Q, gammatone, and mel families, including their tight-frame variants. Two findings emerge: First, the condition number κ is not a statistically significant predictor of CNN performance within the controlled ERB family ($r = 0.738$, $p = 0.155$, $n = 5$). Second, the considered invalid-frame configurations significantly outperform the valid-frame configurations ($p = 0.006$). An apparent overall correlation between condition number and F1 of $r = 0.954$ is identified as a filter-family confound. In summary, we find that in our setting tightness does not benefit CNN classification performance but that the filter family design—specifically frequency resolution and spacing—is the dominant factor.

Keywords: *frame theory, filterbanks, bioacoustics, convolutional neural networks, elephant vocalisations*

1. Introduction

Audio classification with neural networks often begins with a fixed time-frequency representation: a short-time Fourier magnitude spectrogram, mel-frequency cepstral coefficients, or a constant-Q spectrogram [1], [2]. These representations are chosen largely by convention, and their mathematical properties—how faithfully they preserve signal energy, whether inversion is stable—are seldom examined in the machine learning context. The implicit assumption is that the input representation is secondary to classifier design.

Frame theory offers a principled vocabulary for evaluating the mathematical properties of these representations [3], [4]. For finite dimensional input signals $f \in \mathbb{C}^L$, a filterbank with M subband channels constitutes a frame if there exist bounds $0 < A \leq B < \infty$ such that

$$A\|f\|^2 \;\leq\; \sum_{m=1}^{M} \|c_m\|^2 \;\leq\; B\|f\|^2 \tag{1}$$

for every signal f, where c_m are the (potentially downsampled) subband coefficients. In words, Eq. (1) states that the total energy of the subband coefficients is bounded both below and above by the signal energy, so the representation loses no information ($A > 0$) and cannot amplify it without bound ($B < \infty$). The bounds A and B govern reconstruction stability [5]: the canonical dual synthesis operator recovers the signal with a relative error that scales as $(B/A-1)/(B/A+1)$ per iteration of Landweber-type algorithms [3]. The condition number $\kappa = B/A$ thus quantifies how well-conditioned the inversion problem is: $\kappa = 1$ (a tight frame) yields perfect reconstruction in a single step, while large κ implies slow convergence and numerical sensitivity. When $A = 0$, the filterbank has spectral gaps and perfect reconstruction is fundamentally impossible [6].

From a pure frame-theoretic perspective, κ governs reconstruction stability, not discriminative power—a classifier never inverts the analysis operator, so there is no theoretical reason to expect κ to predict classification accuracy. However, $\kappa = 1$ also implies that signal energy is distributed uniformly across subbands, and this energy-balance property is widely invoked as a desirable inductive bias in applied audio machine learning: scattering networks [7], [8] use fixed tight wavelet frames to guarantee Lipschitz continuity, and learnable input representations such as LEAF [9] ini-

Corresponding author: *clara.hollomey@ustp.at.*

tialise with Gabor filterbanks near the tight-frame regime.

Recent work on learnable filterbanks [10] has shown that adapting the time-frequency representation to the task can improve performance, but these approaches do not analyse their learned representations in frame-theoretic terms. In [11], a trainable auditory-inspired filterbank operating on a frame theory-based learning objective is derived for encoder-decoder architectures, but the degree to which it is suitable for classification tasks is not investigated.

Here, we investigate the implicit assumption that generalisation benefits from energetically uniformly distributed discriminative information. We address this question in a use case focusing on elephant rumble vocalisations—near-infrasonic signals (fundamental frequency 8–80 Hz) [12] that place strong demands on low-frequency filterbank coverage. These vocalizations are harmonically rich, and occupy a frequency range where standard mel-frequency representations—designed for human hearing [13]—provide poor resolution. We ask: (i) do filterbanks with $A = 0$ perform worse than valid frames? and (ii) within the valid-frame regime, does lower κ lead to better performance?

Our contributions are: (1) a systematic study relating frame-theoretic filterbank properties to CNN classification performance across 15 configurations; (2) a null result showing that in our setting κ does not predict CNN performance within a filter family; and (3) evidence that filter family design outweighs frame conditioning.

2. Background and related work

2.1. Filterbanks as frames

A filterbank with M filters $\{G_m\}$ and hop sizes $\{a_m\}$ defines the analysis operator $T : \mathbb{C}^L \to \bigoplus_m \mathbb{C}^{L/a_m}$ by $(Tf)_m(n) = c_m(n) = (f * \tilde{g}_m)(a_m n)$ [14], where g_m is the impulse response of the m-th filter (with frequency response $G_m = \hat{g}_m$) and $\tilde{g}_m[k] = \overline{g_m[-k]}$ is its time-reversed conjugate. In the DFT domain, the frame operator $\mathbf{S} = T^*T$ has a main diagonal that is characterised by the frame response $s(\omega) = \sum_m (L/a_m)|G_m(\omega)|^2$. When each filter's frequency-domain support does not exceed L/a_m bins — the *painless case* [15], [16] — the frame operator is exactly diagonal, and the frame bounds equal the extrema of the frame response: $A = \min_\omega s(\omega)$, $B = \max_\omega s(\omega)$ [3]. All filterbank configurations in this study satisfy the painless condition. The filterbank with filters $\{s^{-1/2} \odot G_m\}$ yields the Euclidean-closest filterbank to $\{G_m\}$ that is *Parseval-tight*, i.e. a frame whose bounds satisfy $A = B = 1$. This is stronger than the general tight condition $A = B > 0$ and results in the analysis operator being its own pseudo-inverse, so perfect reconstruction reduces to applying the adjoint T^* without any rescaling. Fractional variants using $s^{-\alpha/2}$ for $\alpha \in [0, 1]$ trace a monotone path from the original frame ($\alpha = 0$, $\kappa = \kappa_0$) to the associated Parseval-tight frame ($\alpha = 1$, $\kappa = 1$).

2.2. Frame properties and classification

A filterbank with large κ might distribute energy unevenly: some frequency bands are over-represented while others are compressed. For a neural classifier, this means the gradient landscape is dominated by high-energy channels regardless of discriminative value. In the extreme case of $A = 0$, frequency gaps can make certain signal components entirely invisible to the network. Tight frames ($\kappa = 1$) ensure proportional contribution from every band. Hence, under the assumption that discriminative information is distributed across frequencies [17], tightness should benefit generalisation. However, whether CNNs are sensitive to κ in classification tasks in practice is ultimately empirical, since convolutional layers can in principle learn to compensate for energy imbalance [18]. The distinction between reconstruction quality (where frame bounds are essential) and classification performance (where they may not be) is central to this study.

2.3. Related work in bioacoustic front-end design

The dominant paradigm in bioacoustic deep learning is to use mel-spectrograms—magnitude-squared STFT coefficients weighted by triangular mel-frequency bands—and feed them to transfer-learned CNNs [1], [19], [20]. Although similar to a filterbank, this pipeline cannot be treated in a frame-theoretic sense. Alternative front-ends based on gammatone filterbanks [21] or constant-Q transforms [22] have been explored, but comparisons are typically made between complete systems rather than isolating the representation's mathematical properties. In this work we construct filterbanks directly as sets of frequency-domain windows at perceptually spaced centre frequencies, so that frame bounds and the condition number κ are well-defined and can be manipulated in a controlled manner.

3. Dataset

The dataset for this study consists of 3 080 elephant (*Loxodonta africana*) rumble vocalisations, labelled by age class: adult ($N = 2\,161$), calf ($N = 319$), infant ($N = 456$), juvenile ($N = 78$). The severe class imbalance (juvenile = 2.5%) motivates macro-averaged F1 as the primary metric, which weights each class equally regardless of prevalence. Recordings are resampled to 500 Hz, zero-padded or centre-cropped to $L = 2\,560$ samples (5.12 s), and normalised to unit RMS.

Elephant rumbles have fundamentals as low as 8 Hz, with harmonics extending to several hundred Hertz [12], [23]. Age differences are expressed in fundamental frequency, harmonic structure, call duration, and amplitude modulation patterns [24].

The rumbles in the dataset were mostly recorded in the wild, and have been annotated and preselected by expert listeners [25].

4. Methodology

4.1. Filterbank configurations

Filterbank features are extracted at hop size $a = 40$ samples, producing magnitude spectrograms of shape $(M, 64)$, with M the number of filterbank frequency channels. The filterbank ranges are set to $F_{\min} = 14\,\text{Hz}$, $F_{\max} = 240\,\text{Hz}$; the coverage gap below 14 Hz is discussed in Section 6. We evaluate 15 configurations organised into three groups designed to isolate specific design choices:

Group A (ERB family, κ variation): Starting from a standard ERB auditory filterbank ($M = 40$), we derive six configurations via fractional tightening ($\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$) and dual-frame inversion, yielding a controlled monotonic κ sequence from 6.08 to 1.00 while holding filter family, channel count, frequency range, and hop size constant. This group is the key controlled experiment.

Group B (channel count variation): Tight ERB ($\kappa = 1$) with $M \in \{20, 40, 60, 80\}$, isolating spectral resolution from frame conditioning.

Group C (family comparison, $M = 40$): Mel filterbank (C1), gammatone (C2), CQT $Q = 12$ (C3), CQT $Q = 24$ (C4), tight gammatone (C5). Each has different frequency spacing and coverage.

The frame bounds are computed as the extrema of the frame response over the full DFT grid using a Python port of the Large Time-Frequency Analysis Toolbox [26], [27]. Three Group C configurations fail the frame condition: the mel filterbank has $A \approx 2.1 \times 10^{-11}$ (effectively zero), while both CQT variants have $A = 0$ exactly. Table 1 summarises all configurations.

4.2. Classifier and training protocol

A deliberately small CNN (∼85K parameters) ensures that performance differences reflect representation quality rather than classifier capacity: two Conv2d layers (8 and 16 filters of size 3×3 with ReLU activation followed by 2×2 max-pooling), followed by a 32-unit fully connected layer and a 4-class softmax output. Training uses class-weighted cross-entropy (weights inversely proportional to class frequency), Adam optimiser (lr $= 10^{-3}$, weight decay 10^{-4}), cosine annealing over 100 epochs, and 5-fold stratified cross-validation with 5 random seeds (25 macro-F1 estimates per configuration). Features are per-channel normalised to zero mean and unit variance. No logarithmic compression is applied.

5. Results

5.1. Condition number vs. performance (Group A)

The controlled ERB experiment yields a null result. Despite the κ gradient ranging from 6.08 to 1.00, no monotonic performance trend is observed (Table 1, Fig. 1 right panel). The standard ERB (A1, $\kappa = 6.08$) marginally outperforms the tight ERB (A2, $\kappa = 1$): 0.537 ± 0.020 vs. 0.533 ± 0.019. Pearson correlation between $\log_{10} \kappa$ and macro-F1 within Group A is $r = 0.738$ ($p = 0.155$, $n = 5$)—not statistically significant. With $n = 5$, the critical threshold for $p < 0.05$ is $|r| > 0.878$. The trend direction is opposite to the hypothesis: larger κ is weakly associated with *higher* F1.

5.2. Channel count (Group B)

Varying M from 20 to 80 at fixed $\kappa = 1$ produces no meaningful change: F1 ranges from 0.532 to 0.535 (span 0.003), well within one standard deviation (∼0.022). Channel count is not a significant factor over this range.

5.3. Filter family comparison (Group C)

A comparison between filter families (Fig. 1, left panel) suggests that the three configurations CQT $Q = 12$, CQT $Q = 24$, and mel are the three best-performing configurations overall. Notably, all three configurations do not yield frames ($A = 0$). CQT $Q = 12$ achieves the highest macro-F1 of 0.577, outperforming the best ERB (A1, 0.537) by +0.040 F1 points. Welch's t-test comparing non-frame ($n = 3$) against valid-frame $M = 40$ configurations ($n = 8$) yields $t = -9.10$, $p = 0.006$, Cohen's $d = 8.2$. This is statistically significant.

5.4. Per-class analysis

The non-frame advantage is concentrated in rarer classes. The calf class shows the largest improvement: non-frame mean F1 $= 0.529$ versus 0.460 for valid frames ($\Delta = +0.069$). The infant class shows $\Delta = +0.038$. For the juvenile class—the rarest with only 78 training examples—CQT $Q = 24$ achieves F1 $= 0.100$ versus a valid-frame mean of 0.047, roughly doubling performance. The adult class shows minimal difference ($\Delta = +0.009$), consistent with adults being easily discriminated by any representation.

5.5. Confound analysis

Pooling all 15 configurations, Pearson correlation between $\log_{10} \kappa$ and macro-F1 is $r = 0.954$ ($p < 0.001$). This apparently strong relationship is a confound: all non-frame configurations are non-ERB families (mel, CQT), while all valid-frame configurations are ERB or gammatone. Partial correlation controlling for channel count M yields an unchanged $r = 0.954$, ruling out M as the confounding variable. Within the controlled Group A experiment—where filter family is fixed and only κ varies—the correlation drops to $r = 0.738$ ($p = 0.155$). This within-family estimate is the causally clean comparison; the overall $r = 0.954$ reflects the CQT/mel filter *family* advantage, not the effect of frame conditioning.

6. Discussion

6.1. The CNN compensation mechanism

The null result within Group A admits a mechanistic explanation. The CNN's first convolutional layer receives a spectrogram whose per-channel variance scales with the energy at each filterbank channel's center frequency. During training, the gradient signal may push the learned kernel weights toward down-

Table 1: Filterbank configurations, frame properties, and classification results. Macro-F1 is the mean over 25 estimates (5-fold × 5-seed cross-validation). *Configurations with $A \approx 0$ are not valid frames. Per-class F1 shown for the two most affected classes.

ID	Description	M	A	B	κ	F1 (mean)	(std)	F1-calf	F1-juv.
A1	Standard ERB	40	76.2	463	6.08	0.537	0.020	0.470	0.051
A2	Tight ERB ($\alpha = 1.0$)	40	1.00	1.00	1.00	0.533	0.019	0.458	0.049
A3	Dual ERB	40	0.002	0.013	6.08	0.528	0.029	0.448	0.042
A4	Partial ($\alpha = 0.25$)	40	25.8	99.9	3.87	0.534	0.021	0.458	0.050
A5	Partial ($\alpha = 0.50$)	40	8.73	21.5	2.47	0.532	0.018	0.464	0.036
A6	Partial ($\alpha = 0.75$)	40	2.96	4.64	1.57	0.532	0.025	0.458	0.048
B1	Tight ERB $M = 20$	20	1.00	1.00	1.00	0.533	0.026	0.468	0.040
B2	Tight ERB $M = 40$	40	1.00	1.00	1.00	0.532	0.022	0.454	0.053
B3	Tight ERB $M = 60$	60	1.00	1.00	1.00	0.535	0.027	0.454	0.061
B4	Tight ERB $M = 80$	80	1.00	1.00	1.00	0.533	0.022	0.448	0.061
C1*	Mel	40	$\sim$0	74.9	$\sim 10^{12}$	0.563	0.018	0.527	0.063
C2	Gammatone	40	76.2	463	6.08	0.539	0.026	0.464	0.061
C3*	CQT $Q = 12$	40	0	65.9	∞	**0.577**	0.025	**0.544**	0.080
C4*	CQT $Q = 24$	40	0	64.0	∞	0.572	0.025	0.516	**0.100**
C5	Tight gammatone	40	1.00	1.00	1.00	0.532	0.020	0.462	0.043

weighting high-energy channels — effectively approximating the inverse square root of that per-channel energy. The per-channel feature normalisation applied before training may further assist this implicit compensation.

6.2. Why non-frames outperform valid frames

The counterintuitive superiority of $A = 0$ configurations reflects a fundamental distinction: the frame condition guarantees stable inversion [3], [15]; classification does not require inversion. A CNN can extract discriminative features from an incomplete representation provided the relevant information is present.
As illustrated by Figure 2, the CQT and mel filterbanks' advantage may be better explained by frequency resolution. The filter quality factor Q (centre frequency divided by bandwidth) is constant for CQT filterbanks ($Q = 12$ or $Q = 24$) but varies for ERB filterbanks, dropping to $Q \approx 1.9$ at the geometric mean frequency of the elephant range (58 Hz). At $Q < 2$, even the second harmonic is unresolvable from the fundamental; at $Q = 12$, harmonics up to the 12th are cleanly separated. This interpretation aligns with the per-class analysis: the calf class, whose empirically observed fundamental frequency range differs most from that of adults, benefits most from high-Q representations.

6.3. Frequency coverage and future directions

The filterbank range begins at $F_{\min} = 14$ Hz, leaving the lowest fundamentals ($\sim$8 Hz) uncovered. A DC-centred Gaussian lowpass extension (41 channels) produced no improvement (ΔF1 < 0.001, $p = 0.975$), likely because a single broadband channel provides just energy without frequency resolution.
The absolute macro-F1 scores (0.53–0.58) reflect a deliberately small classifier and should be read as a controlled comparison of front-ends rather than a state-of-the-art age classifier — the persistent difficulty on the rare juvenile class underscores this. Whether a low condition number benefits a higher-capacity model, for instance by regularising against overfitting, cannot be resolved with the present architecture and is left to future work.

7. Conclusion

We presented a systematic study relating frame-theoretic filterbank properties to CNN classification performance. The key findings are: (1) the condition number κ is not a significant predictor of CNN performance within the ERB family ($r = 0.738$, $p = 0.155$); (2) the considered invalid-frame configurations significantly outperform valid frames ($p = 0.006$), likely driven by superior resolution at low frequencies. An apparent overall correlation of $r = 0.954$ is a filter-family confound. These results suggest that, for our setting, frame-theoretic reconstruction guarantees do not predict classification performance, and that filterbank design for discriminative tasks should prioritise frequency spacing and harmonic resolution over mathematical tightness. The observed independence of macro-F1 from κ should be read in light of our pipeline's per-channel normalisation, which already equalises energy across frequency bins and thereby subsumes the principal effect that frame tightening would otherwise provide.

8. Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) and by the State of Lower Austria [Grant ID: 10.47379/LS23024].

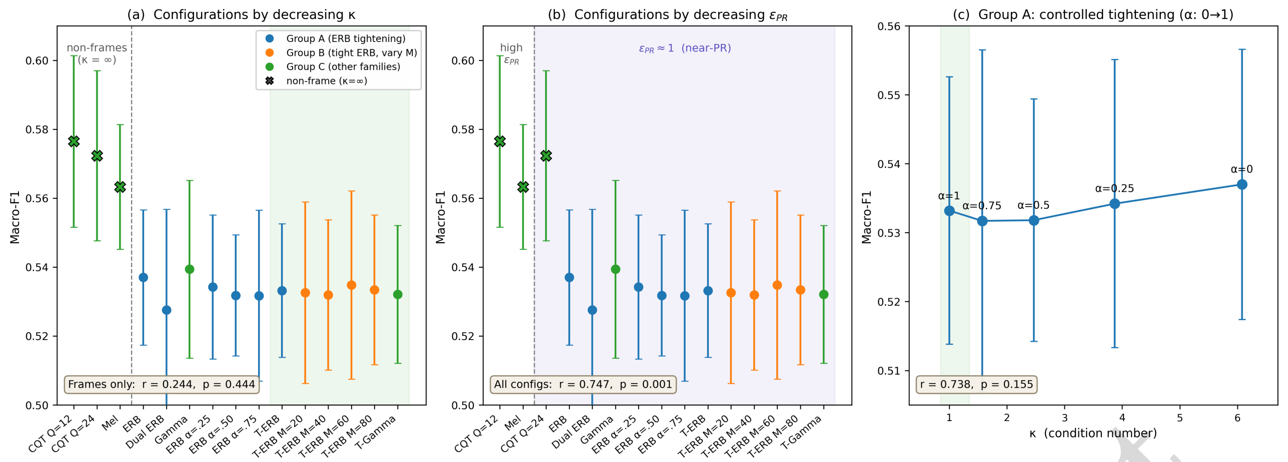

Figure 1: Frame properties versus classification performance across all 15 configurations (macro-F1; error bars ±1 SD over 25 runs; colour = filter group, × = non-frame configurations, $\kappa = \infty$; horizontal axes are ordered, not continuous). **(a)** Configurations ordered by decreasing condition number κ: the highest macro-F1 scores belong to the non-frame CQT and mel families, so the apparent overall association between κ and performance ($r = 0.954$) is a filter-family effect—within valid frames it vanishes ($r = 0.244$, $p = 0.444$). Green band: tight frames ($\kappa = 1$). **(b)** The same configurations ordered by decreasing perfect-reconstruction error $\varepsilon_{\mathrm{PR}}$. The positive association (all configs: $r = 0.747$, $p = 0.001$) is the same confound—the non-frame families combine high $\varepsilon_{\mathrm{PR}}$ with high macro-F1—so reconstruction error does not causally track discriminative performance. Lavender band: near-perfect reconstruction ($\varepsilon_{\mathrm{PR}} \approx 1$). **(c)** Controlled Group A experiment: within the fixed ERB family, macro-F1 against κ along the tightening path (α: $0 \to 1$). The non-significant $r = 0.738$ ($p = 0.155$, $n = 5$) and the opposite-to-predicted trend confirm that κ does not predict CNN performance.

References

[1] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, e13152, 2022.

[2] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP*, 2017, pp. 131–135.

[3] O. Christensen, *An Introduction to Frames and Riesz Bases*, 2nd. Cham: Birkhäuser, 2016, pp. 119–151.

[4] I. Daubechies, *Ten Lectures on Wavelets.* SIAM, 1992, pp. 53–105.

[5] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3256–3268, 1998.

[6] P. G. Casazza, "The art of frame theory," *Taiwanese Journal of Mathematics*, vol. 4, no. 2, pp. 129–201, 2000.

[7] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[8] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, 2014.

[9] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *Proc. ICLR*, 2021.

[10] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[11] D. Haider, F. Perfler, V. Lostanlen, M. Ehler, and P. Balazs, "Hold me tight: Stable encoder-decoder design for speech enhancement," in *Proc. Interspeech*, 2024.

[12] K. Payne, W. Langbauer Jr., and E. Thomas, "Infrasonic calls of the Asian elephant (*Elephas maximus*)," *Behavioral Ecology and Sociobiology*, vol. 18, pp. 297–301, 1986.

[13] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *J. Acoust. Soc. Am.*, vol. 33, no. 10, pp. 1344–1356, 1961.

[14] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Prentice Hall, 1993, pp. 42–99.

[15] K. Gröchenig, *Foundations of Time-Frequency Analysis.* Boston: Birkhäuser, 2001.

[16] T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor Analysis and Algorithms: Theory and Applications*, H. G. Feichtinger and T. Strohmer, Eds., Boston: Birkhäuser, 1998, pp. 267–294.

[17] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," *Proc. NeurIPS*, vol. 30, 2017.

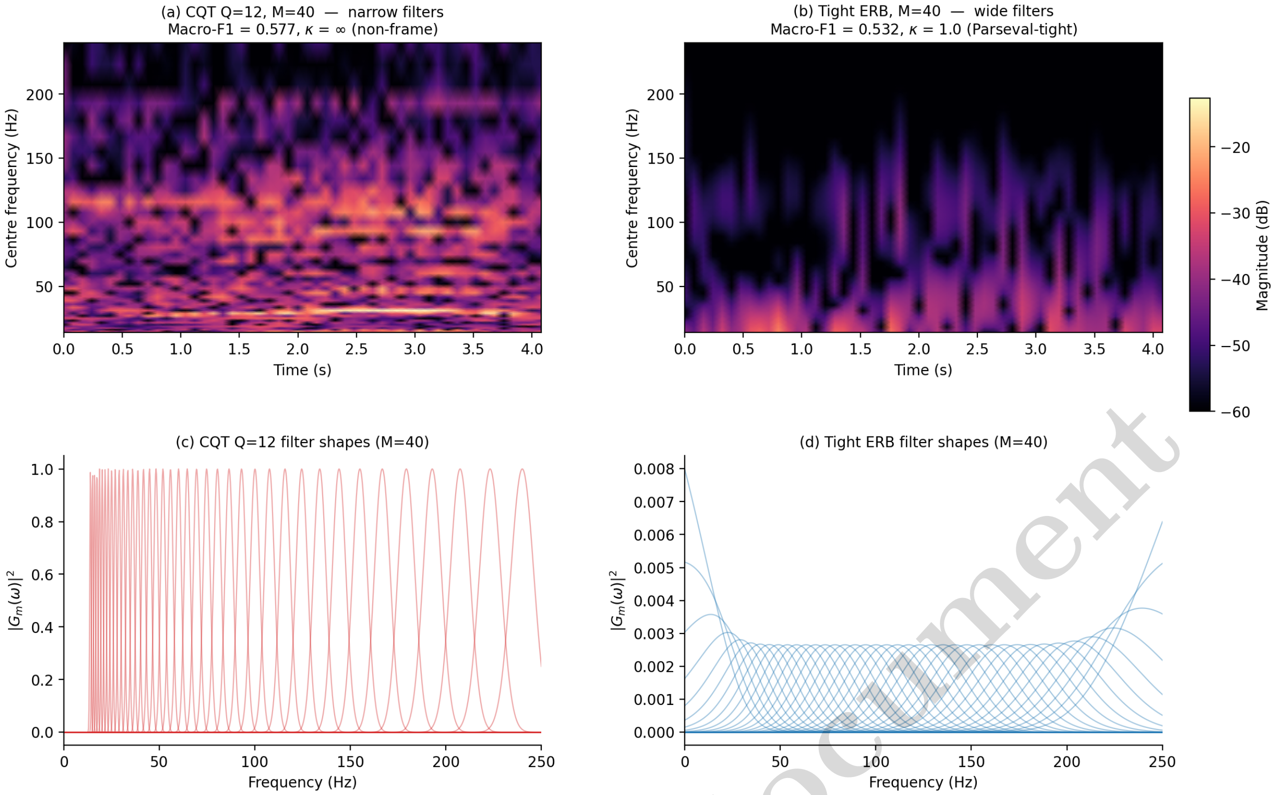

Figure 2: Filterbank spectrograms of an elephant rumble resampled to 500 Hz. (a) CQT filterbank ($Q = 12$, $M = 40$, $\kappa = \infty$): narrow constant-Q filters resolve individual harmonics of the fundamental ($F_0 \approx 20$ Hz). (b) Parseval-tight ERB filterbank ($M = 40$, $\kappa = 1$): broader auditory-scaled filters smear adjacent harmonics, reducing spectral detail visible to the classifier. (c, d) Corresponding squared frequency responses $|G_m(\omega)|^2$. Both filterbanks use 40 channels spanning 14–240 Hz; the difference in classification performance (macro-F1: 0.577 vs. 0.532) arises from frequency resolution, not frame conditioning.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[19] S. Kahl, C. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021.

[20] Y. Shiu et al., "Deep neural networks for automated detection of marine mammal species," *Scientific Reports*, vol. 10, p. 607, 2020.

[21] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proc. ISMIR*, 2015.

[22] N. Holighaus, M. Dörfler, G. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 775–785, 2013.

[23] K. McComb, D. Reby, L. Baker, C. Moss, and S. Sayialel, "Long-distance communication of acoustic cues to social identity in African elephants," *Animal Behaviour*, vol. 65, no. 2, pp. 317–329, 2003.

[24] A. S. Stoeger and P. Manger, "Vocal learning in elephants: Neural bases and adaptive context," *Current Opinion in Neurobiology*, vol. 28, pp. 101–107, 2014, SI: Communication and language, ISSN: 0959-4388. DOI: `https://doi.org/10.1016/j.conb.2014.07.001`.

[25] M. Zeppelzauer and A. S. Stoeger, "Establishing the fundamentals for an elephant early warning and monitoring system," *BMC research notes*, vol. 8, no. 1, p. 409, 2015.

[26] Z. Prusa, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The large time-frequency analysis toolbox 2.0," in *International Symposium on Computer Music Multidisciplinary Research*, Springer, 2013, pp. 419–442.

[27] C. Hollomey and N. Holighaus, "LTFATPY: Towards making a wide range of time-frequency representations available in Python," in *27th International Conference on Digital Audio Effects, DAFx 2024*, DAFx, 2024.