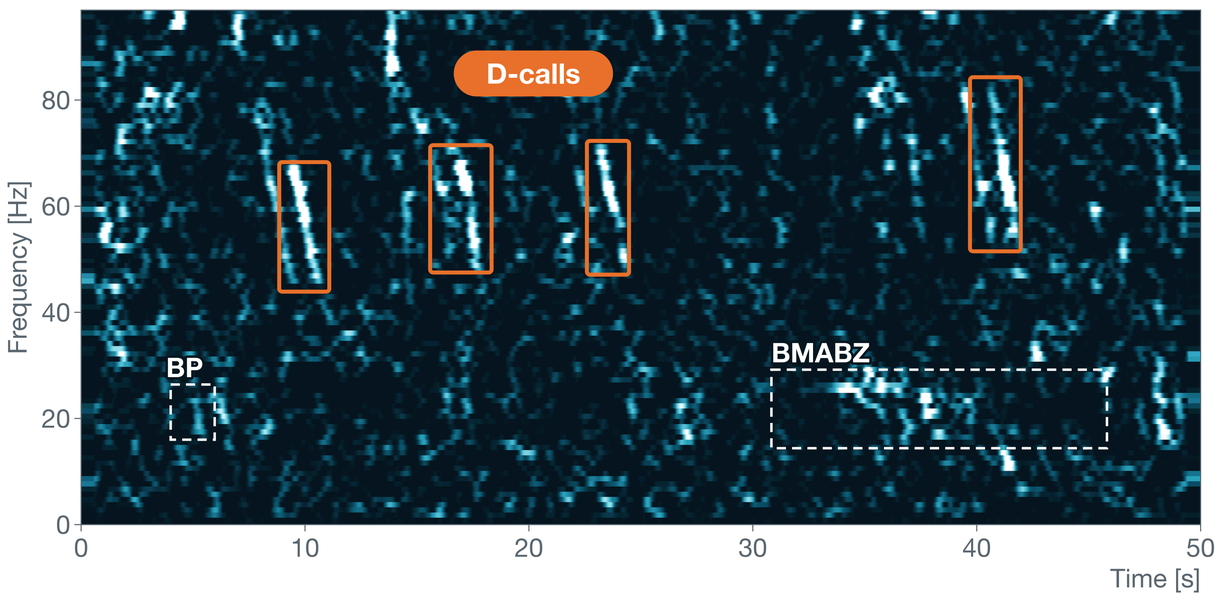
NAVE: A NORMALIZED, ADAPTIVE CONFORMER FOR ANTARCTIC WHALE VOCALIZATION DETECTION

Matthias Nagl[1], **Paul Primus**[1], **Gerhard Widmer**[1,2]

[1]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2]LIT Artificial Intelligence Lab, Linz, Austria

JOHANNES KEPLER UNIVERSITY LINZ
Institute of Computational Perception

D-calls are the blind spot of current detectors



D blue/fin-whale downsweeps **BMABZ** blue-whale A, B, Z **BP** fin-whale 20 Hz pulses

D-calls are faint downsweeps, often masked by noise. Prior detectors reach a D-call F1 of only **0.13** [1] and **0.34** [2].

OUR QUESTION

Can a better front-end and architecture improve D-call detection?

OUR ANSWER: NAVE

0.45
D-call F1
best prior: 0.34

0.51
Macro F1
best prior: 0.475

How: 1 normalized input, 2 call-length Conformer kernel.
Single model, only per-class thresholds tuned.

Leave-one-site-out (LOSO), 3 validation site-years, 5 seeds, fold-averaged as in [2].

NAVE: two improvements for D-call detection

Audio
250 Hz,
60 s tiles

→ **1 Normalized input**
mean-removed STFT + PCEN [4]
magnitude, phase (cos, sin), PCEN

→ **FDY stem [5]**
frequency-adaptive
convolutions

→ **2 Conformer [3]**
4 blocks, depthwise kernel
k = 153 frames (3.06 s)

→ **Events**
per-class
thresholds

Training: 8 site-years, 30 s segments, weighted BCE, 1:1 negative resampling, EMA weights.

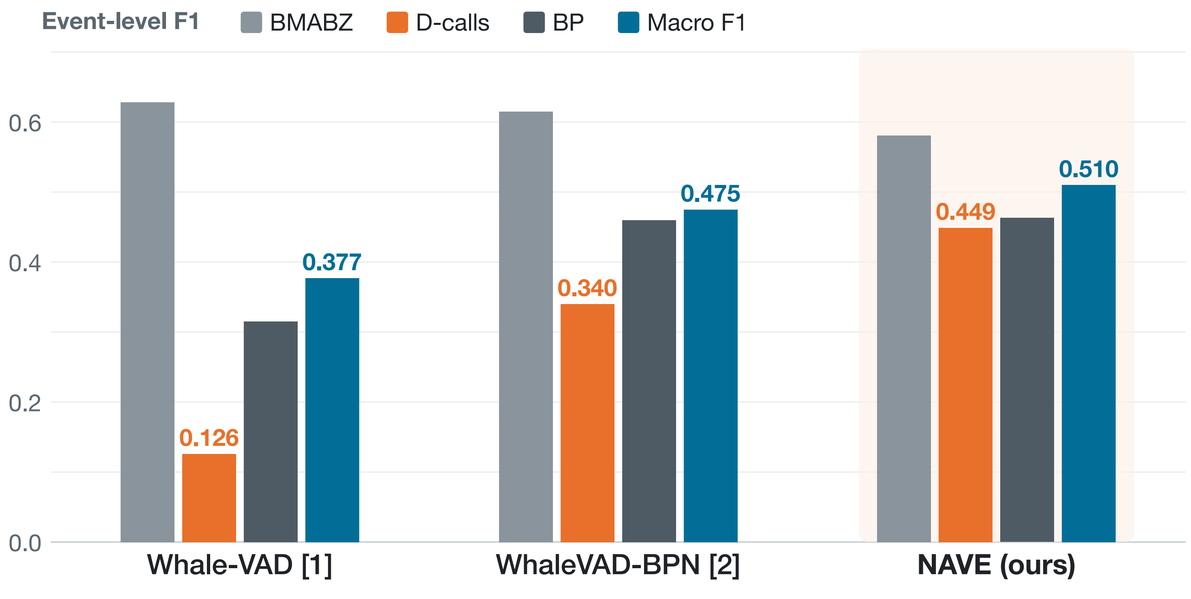
Makes D-calls more prominent
Mean subtraction removes constant site noise. PCEN tracks the noise floor over ≈ 0.8 s.

Mostly helps BMABZ
BMABZ F1 0.393 to 0.472. Filters adapt to each frequency band.

Spans ≈ 3 s, enough for a typical D-call
≈ 20k weights per block.

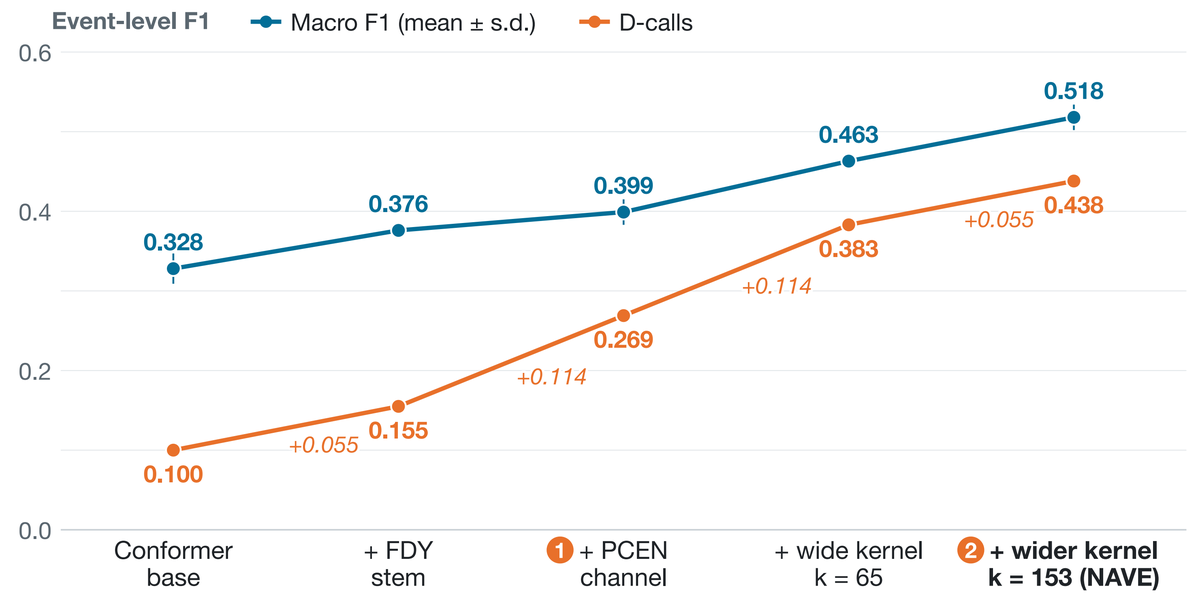
Fixed rules
from Whale-VAD [1], only thresholds tuned

NAVE lifts D-call F1 from 0.34 to 0.45



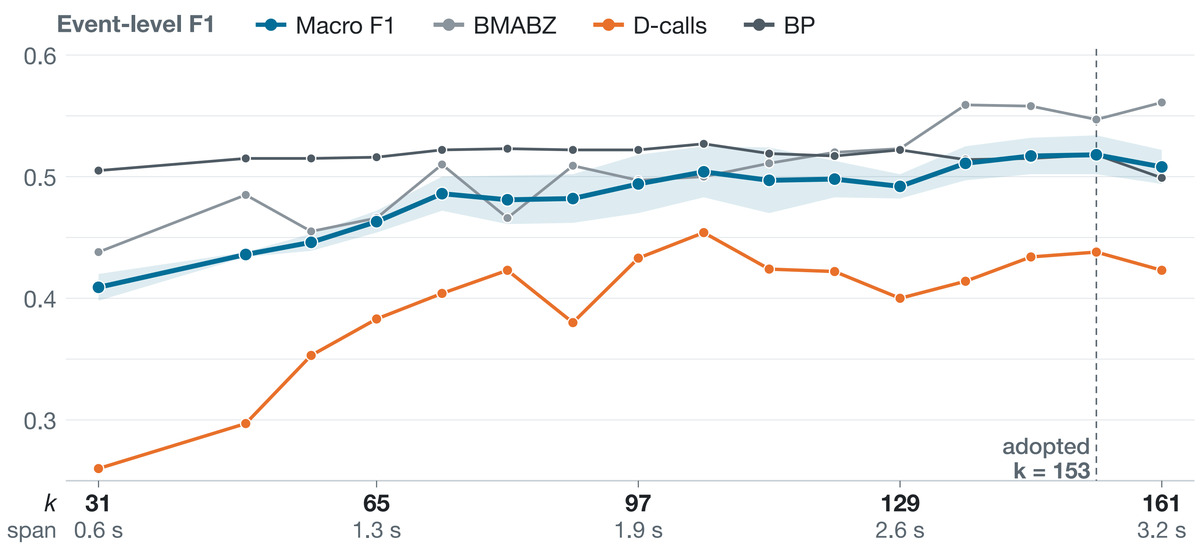
LOSO, 5-seed mean, fold-averaged as in [2]. Pooled (official scorer): **0.518 ± 0.016**.

PCEN and the wide kernel recover D-calls



Components added cumulatively. Pooled counts, 5 seeds. Orange: D-call gain per step.

Wider kernels help until they span a typical call



Full model, ≥ 3 seeds, band ±1 s.d. Plateau from k ≈ 65 (1.3 s). k = 153 beats k = 129 on all 5 paired seeds ($p = 0.036$).

Takeaways

1. **A better front-end and architecture** raise D-call F1 from 0.34 to 0.45.
2. **A single model outperforms WhaleVAD-BPN** (macro F1 0.510 vs. 0.475) without boundary proposals or post-processing search.
3. D-calls remain the hardest class. Future work: class-specific receptive fields and adaptation to unseen sites.

Code & pretrained model
github.com/TheFruitGuy/NAVE-SED

CONFORMER

[1] C. Geldenhuys et al., Whale-VAD, DCASE 2025. [2] C. M. Geldenhuys et al., WhaleVAD-BPN, arXiv:2510.21280, 2025. [3] A. Gulati et al., Conformer, Interspeech 2020. [4] Y. Wang et al., PCEN, ICASSP 2017. [5] H. Nam et al., Frequency dynamic convolution, Interspeech 2022. The LIT AI Lab is supported by the Federal State of Upper Austria.

Contact: Matthias Nagl, matthias.nagl@jku.at · Institute of Computational Perception, Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria