

A Dual-Resolution Binary Neural Network Inference Engine for Fully Implantable Cochlear Implant Speech Processing

Mert Doğan¹, Hasan Uluşan^{1,2}, Haluk Külâh^{1,2}

¹Department of Electrical and Electronics Engineering, METU, Ankara, Turkey

²ODTÜ MEMS CENTER, Ankara, Turkey

mertd@metu.edu.tr

Summary:

A dual-resolution Binary Neural Network inference engine replaces conventional filterbanks in cochlear implant speech processing. A 1 kHz split applies a 256-sample window (75% overlap) to low-frequency channels and a 128-sample window (50% overlap) to high-frequency channels, with a shared 64-sample hop. The architecture matches fixed N=256 intelligibility (STOI 0.826 vs 0.830) while halving high-frequency channel latency and reducing L1 inference cost and size by ~25%, at <2 mW inference power. The approach can be applied to edge AI sensor SoCs requiring low-power spectral analysis.

Keywords: cochlear implant, binary neural network, dual-resolution, embedded ML, implantable SoC

Background

Cochlear Implant (CI) users struggle with speech-in-noise due to two main limitations: the limited feature extraction capability of current processors [1], and suboptimal stimulation patterns. In sound processing, long windows resolve pitch and vowel formants for low frequencies, while short windows capture consonant transients and no single length serves both [2]. Additionally, conventional IIR/FFT filterbanks used in CIs are too power-hungry for fully implantable SoCs [3], and recent deep learning CI approaches [4, 5] remain offline-only. BNNs can bridge this gap via low-cost XNOR-popcount operations [6], enabling real-time embedded inference at implantable power levels. To the best of the authors' knowledge, this work presents the first implantable-scale neural CI speech processor SoC with embedded ML inference.

Method

Unlike single-window CI processors, a 1 kHz frequency split assigns a longer window to low-frequency channels (200–1000 Hz, 256 samples, 10.7 ms, 75% overlap) where wider spectral resolution benefits pitch and vowel encoding, and a shorter window to high-frequency channels (1–11 kHz, 128 samples, 5.3 ms, 50% overlap) where reduced latency and inference cost are prioritized. A shared 64-sample hop ensures both paths produce electrode envelopes at identical update rates, maintaining synchronisation without additional hardware complexity.

Envelope extraction replaces conventional multiply-accumulate-heavy IIR/FFT operations with

a BNN that operates directly on the raw I²S bitstream from the MEMS microphone, eliminating preprocessing entirely. XNOR-popcount reduces inference to single-cycle bitwise logic ($\mathbf{w}^T \mathbf{a} = 2 \text{ popcount}(\overline{\mathbf{w} \oplus \mathbf{a}}) - N$). A single pipelined 2048-bit engine is time-multiplexed across all layers of both paths (4-cycle, 48 MHz), completing inference in 0.111 ms (Fig. 1(a)). Bitwise output encoding maps 128 binary activations directly to 8-bit electrode envelopes, enabling positionally-weighted per-bit loss supervision during training. Binary weights are stored in on-chip BRAM; BatchNorm parameters are absorbed offline into fixed popcount thresholds, adding no runtime overhead. Reprogrammable weights enable post-implantation patient-specific adaptation: retraining on ECAP-derived cochlear maps [7] emphasizes channels with intact spiral ganglion populations without hardware modification.

Training uses synthetic windows with log-normal frequency distribution (centred at 2500 Hz, $\sigma = 0.8$ decades) and Beta(1,5) amplitudes across up to 32 sinusoidal components, biasing coverage toward the perceptually critical mid-speech range. Windows are normalized to unit peak before 8-bit signed quantization, matching the hardware MEMS pipeline exactly. The combined loss (1) sums peak-weighted MSE (2) and per-bit BCE (3). The MSE term applies peak-emphasis ($\alpha = 10$) with Gaussian band-weighting over the formant region (400–3000 Hz); BCE is weighted by bit significance ($w_b \propto 2^{-b}$, MSB weighted 128× over LSB), with separate loss instances per path.

$$\mathcal{L} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} \quad (1)$$

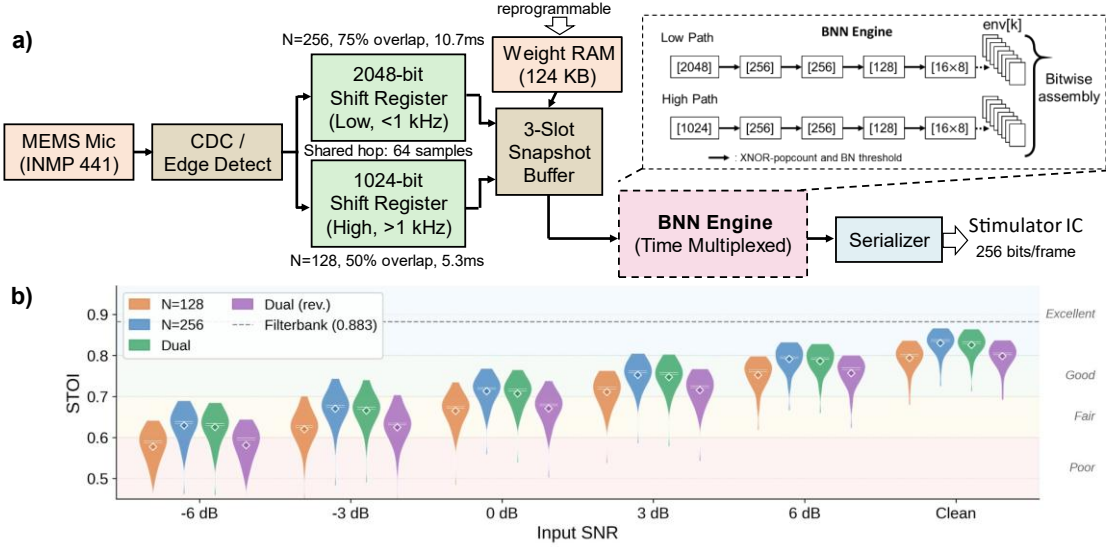


Fig. 1. (a) Top-level system architecture: dual-path dataflow from MEMS microphone through asymmetric shift registers and time-multiplexed BNN engine to serial electrode output. (b) STOI distributions across SNR conditions for fixed $N=128$, $N=256$, dual-res. and reverse dual-res. BNN, dashed line indicates clean filterbank ceiling.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{NB} \sum_{n,k} (1 + \alpha y_{nk}) (\hat{e}_{nk} - y_{nk})^2 \cdot w_k^{\text{band}} \quad (2)$$

$$\mathcal{L}_{\text{BCE}} = \frac{1}{8NB} \sum_{n,k,b} w_b^{\text{bit}} \cdot w_k^{\text{band}} \cdot \text{BCE}(l_{nkb}, y_{nkb}^{\text{bit}}) \quad (3)$$

Results

Table 1 and Fig. 1(b) summarize the four-way STOI [8] comparison across input SNR conditions. The reversed split ($N=128$ low / $N=256$ high) performs comparably to uniform $N=128$, while the proposed split ($N=256$ low / $N=128$ high) consistently outperforms both across all SNR conditions. This confirms that the direction of the frequency-window assignment is the critical design choice. The dual-resolution architecture achieves equivalent intelligibility to fixed $N=256$ (clean STOI 0.826 vs 0.830, difference within variance) while halving high-frequency channel latency from 10.7 ms to 5.4 ms and reducing weight memory from 156 KB to 124 KB. Both satisfy the 6–20 ms clinical latency target [1]. The design occupies 63 BRAM36 tiles (17.3%) and 757 LUTs (<1%) on the XC7A200T with 2 mW BNN dynamic power, supporting a

Tab. 1: Performance and hardware results

Metric	128	256	Dual (rev)	Dual
Clean STOI	0.794 ± 0.034	0.830 ± 0.030	0.798 ± 0.033	0.826 ± 0.032
High-path latency	5.4 ms	10.7 ms	10.7 ms	5.4 ms
Infer.Lat.	0.111 ms			
Memory	92 KB	156 KB	124 KB	124 KB
Usage	BRAM: 63 (17.3%) / LUT: 757 (<1%)			

<500 μW ASIC projection at 65 nm CMOS. The XNOR-popcount engine further generalizes to edge AI sensor SoCs where FFT or IIR filtering is computationally prohibitive.

Acknowledgements

This work was supported by TUBITAK under grant no.123C554 and partially by no. 112C340.

References

- [1] F.-G. Zeng et al., "Cochlear implants: System design, integration, and evaluation," IEEE Rev. Biomed. Eng., vol. 1, pp. 115–142, 2008.
- [2] J. H. Won et al., "Maximizing the spectral and temporal benefits of two clinically used sound processing strategies for cochlear implants," Trends Amplif., vol. 16, no. 4, pp. 201–210, 2012.
- [3] H. Uluşan et al., "A full-custom fully implantable cochlear implant system validated in vivo with an animal model," Commun. Eng., vol. 3, p. 132, 2024; doi: 10.1038/s44172-024-00275-4
- [4] E. H.-H. Huang et al., "ElectroNet: A deep-learning-based sound coding strategy for cochlear implants," IEEE Trans. Cogn. Dev. Syst., vol. 16, no. 1, pp. 346–357, 2024.
- [5] T. Gajeci et al., "A deep denoising sound coding strategy for cochlear implants," IEEE Trans. Biomed. Eng., vol. 70, no. 9, pp. 2700–2709, 2023.
- [6] F. Conti et al., "XNOR Neural Engine: A hardware accelerator IP for 21.6-fJ/op binary neural network inference," IEEE Trans. CAD, vol. 37, no. 11, 2018
- [7] C. Garcia et al., "The panoramic ECAP method:," J. Assoc. Res. Otolaryngol., vol. 22, no. 5, 2021.
- [8] C. H. Taal et al., "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 7, pp. 2125–2136, 2011.