

# High-Efficiency Neural Beamforming for Real-Time Speech Enhancement on Smart Low-Power Hearable Devices

Luca Bompani\*, Giovanni Oltrecolli\*, Marco Fariselli†, Francesco Conti\*

\* Department of Electrical, Electronic and Information Engineering, University of Bologna, Italy

† GreenWaves Technologies, Grenoble, France

**Abstract**—Accurate, low-latency spatial beamforming is a crucial component in emerging smart hearable devices, enabling speech enhancement while suppressing noise and interference. In this work we present an optimized methodology for real-time execution of a neural network–based minimum variance distortionless response (MVDR) beamformer on resource-constrained microcontroller units (MCUs). With mixed-precision quantization, beamforming weights can be estimated at an energy cost of 13.2 mJ, which, along a fixed-interval scheduling strategy enables end-to-end real-time operation under a 20 ms latency constraint while achieving a short-time objective intelligibility (STOI) score of 88.4. Efficiency is further improved by integrating a speech activity detection network, which bypasses speech enhancement during silence, resulting in a reduction in energy consumption to 0.62 mJ per execution with 98.5% accuracy. In realistic deployment conditions, we estimate a lifetime of 16 h on a 100 mAh battery.

**Index Terms**—beamforming, speech enhancement, mcu, real-time

## I. INTRODUCTION

Speech Enhancement (SE) is a core capability for hearable devices, directly impacting speech intelligibility, listening comfort, and the robustness of voice-driven AI assistants in noisy environments. Although deep learning has improved SE, most deployed systems remain single-channel, operating on audio from a single microphone. Multi-channel methods leverage spatial information from multiple microphones to achieve stronger noise suppression [1], [2] while preserving spatial cues and directionality essential for natural listening and sound localization [3]. Yet, they are often considered impractical for hearables due to stringent real-time latency requirements (below 20 ms) [4] and limited compute, memory, and energy budgets on MCU-class hardware.

In this work, we show that real-time multi-channel neural beamforming is feasible on the GAP9 ultra-low-power SoC. We deploy a complete SE pipeline combining a convolutional neural network (CNN) for mask estimation with a minimum variance distortionless response (MVDR) beamformer, enabled by an interleaved schedule that overlaps beamformer weight estimation and application, and by mixed-precision execution (`int8` neural inference, `float32` beamforming). Our system achieves a short-time objective intelligibility (STOI) score of 0.88 while consuming 17.28 mJ per beamforming inference. Under realistic conditions, 70% speech absence [5], six microphones, and a 100 mAh, 3.7 V battery, this translates to approximately 16 hours of continuous operation, demonstrating

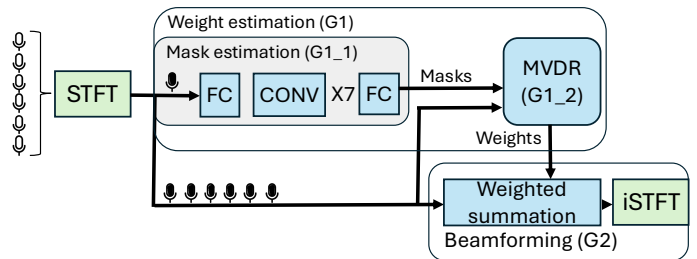


Fig. 1. Block diagram view of the elements in our neural beamforming application.

the practicality of advanced multi-channel SE for hearable devices.

## II. METHODS

We adopt the beamforming problem formulation presented by Ceolini et al. [6], and refer the reader to that work for a more comprehensive description. In brief, beamforming aims to enhance a target speech signal by linearly combining multiple microphone channels with frequency-dependent weights, chosen to preserve the desired source while suppressing interference and noise. In our work, the frequency-dependent weights are estimated through block **G1** (Fig. 1), which is organized into two stages: **G1\_1**, the mask estimation network (approximately 20k parameters), which is the most computationally demanding component of the pipeline, and **G1\_2**, which applies the masks to the spectrogram to compute the speech/noise spatial covariance matrices and derives the MVDR weights. Beamforming application, i.e., the weighted summation of the microphones' spectrograms, and inverse STFT are grouped into block **G2** (see Fig. 1), producing enhanced output frames at each hop  $t_{hop}$ .

After an initial buffering period to compute the first beamforming weights, we switch to an interleaved schedule: at each hop, **G2** runs using the most recent weights, while **G1\_1+G1\_2** advance incrementally in the remaining compute budget. This is effective because MVDR weights evolve slowly over time, allowing reuse across multiple hops without perceptible degradation.

The pipeline is deployed on the GAP9, which integrates a RISC-V host core, 1.5 MB of on-chip L2 memory, and a compute cluster of nine RISC-V cores operating at 370 MHz, supported by a neural engine accelerator (NE16) and four

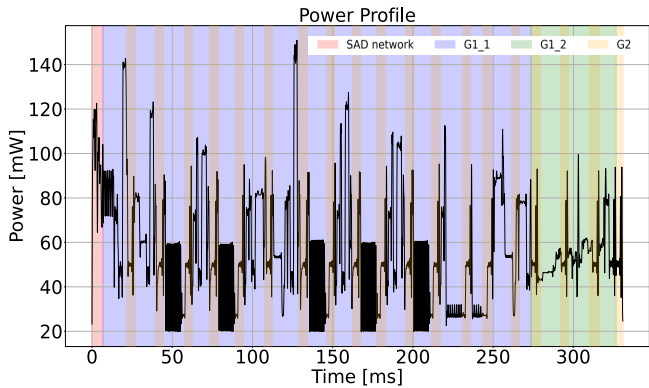


Fig. 2. Power profile of our application.

floating-point units (FPUs). Latency-critical buffers required by G2 are kept in L2 memory, while spectrogram data associated with G1 are stored in external memory and streamed as needed, enabling real-time multi-channel neural beamforming on a microcontroller-class platform.

### III. EXPERIMENTAL RESULTS

We evaluate the proposed neural beamforming pipeline on the GAP9, focusing on the pipeline real-time performance, and energy efficiency. The training of our network follows the setup of Ceolini *et al.* [6] using the Voice Bank + DEMAND (VBD) dataset [7], with multi-microphone acoustic scenes simulated using gpurir [8]. All signals are sampled at 8 kHz, with an STFT window of  $t_{\text{stft}} = 64$  ms and hop size  $t_{\text{hop}} = 15$  ms. Enhancement quality is evaluated using the STOI metric [9].

We found that a mixed-precision configuration, executing the mask estimation network (G1\_1) in `int8` while retaining MVDR weight computation (G1\_2) and beamforming (G2) in `float32`, preserves enhancement quality and is therefore adopted for deployment. This configuration achieves a STOI score of 88.4, improving over the noisy baseline (86.1).

The full pipeline is deployed end-to-end on GAP9 using the GAPflow toolchain. The mask estimation network (G1\_1), executed in `int8` and accelerated by the NE16 neural engine, achieves an efficiency of 3.1 MACs/cycle, allowing the complete network to execute within 203 ms. Including MVDR weight computation and STFT processing, a full set of beamforming weights is computed in approximately 244 ms. Combined with the interleaved execution strategy described in Section II, this results in an effective end-to-end latency of 19 ms, below the 20 ms real-time requirement [4]. Introducing a delay of 300 ms between weight estimation and beamforming leads to only a 0.4% STOI degradation, confirming the robustness of weight reuse [10].

Figure 2 reports the measured power profile of the deployed pipeline on the GAP9 evaluation kit, acquired using Nordic’s Power Profiler Kit. beamforming weight estimation (G1) consumes 13.2 mJ, while the complete enhancement pipeline (G1+G2+STFT and iSTFT) consumes 17.28 mJ per inference. To enable system-level gating during speech absence,

Model	Mpar	Quantization	ms/inf	MAC/cyc	GMAC/J
TinyLSTM [12]	0.33	INT8	4.26	0.36	0.14
TinyLSTM [12]	0.46	INT8	2.39	0.36	0.14
RNNNoise [13]	0.21	INT8	3.28	0.45	1.84
LSTM256 [14]	1.24	Mixed FP16-INT8	2.50	2.11	17.78
GRU256 [14]	0.98	Mixed FP16-INT8	1.70	2.41	17.46
<b>Ours</b>	<b>0.20</b>	Mixed FP32-INT8	<b>3.8</b>	<b>3.10</b>	<b>20.20</b>

TABLE I  
COMPARISON WITH OTHER MCU-DEPLOYED SE PIPELINES.

we employ a SAD network as the decision module for the gating mechanism. The SAD is a lightweight 1-D ResNet-8 [11] operating directly on the waveform, executed in `int8` and consuming 0.62 mJ per inference. It achieves 98.5% accuracy, with 97.8% precision and 99.9% recall. Assuming a 100 mAh, 3.7 V battery and six microphones consuming 0.9 mW each, continuous beamforming yields approximately 6.8 hours of operation; under realistic conditions where speech is present about 30% of the time [5], SAD-based gating extends battery life to approximately 16 hours.

### IV. STATE-OF-THE-ART COMPARISON AND CONCLUSIONS

We demonstrate that a real-time *multi-channel* neural beamforming speech enhancement pipeline can be deployed and executed on a microcontroller-class platform. By combining a compact CNN mask estimator with MVDR beamforming and enabling continuous streaming through interleaved scheduling and mixed-precision execution, our system makes six-microphone spatial speech enhancement practical under the tight memory and compute constraints of an MCU.

In Table I we compare with prior MCU-deployed *single-channel* speech enhancement systems. Our implementation achieves higher computational efficiency (MACs/cycle), despite operating in a more demanding multi-microphone setting and retaining part of the pipeline in `float32`. Relative to *RNNNoise* [13] and *TinyLSTM* [12], we achieve up to 6.9 $\times$  and 8.6 $\times$  higher MACs/cycle, respectively. Compared to the GAP9-based GRU256 and LSTM256 baselines introduced by Rusci *et al.* [14], we improve MACs/cycle by 1.29 $\times$  and 1.46 $\times$ , respectively. These gains are enabled by our mixed-precision quantization strategy and effective utilization of the GAP9 NE16 accelerator, without requiring quantization-aware training.

### REFERENCES

- [1] F.-J. Chang, M. Radfar, A. Mouchtaris, and M. Omologo, “Multi-channel transformer transducer for speech recognition,” 08 2021, pp. 296–300.
- [2] M. Ohlenbusch, C. Rollwage, and S. Doclo, “Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments,” 04 2024, pp. 416–420.
- [3] J. Pan, P. Shen, H. Zhang, and X. Zhang, “Efficient multi-channel speech enhancement with spherical harmonics injection for directional encoding,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9561–9565.

- [4] M. Stone and B. Moore, "Tolerable hearing aid delays. ii. estimation of limits imposed during speech production," *Ear and hearing*, vol. 23, pp. 325–38, 09 2002.
- [5] K. Smeds, S. Gotowiec, F. Wolters, P. Herrlin, J. Larsson, and M. Dahlquist, "Selecting scenarios for hearing-related laboratory testing," *Ear and Hearing*, vol. 41, 2020. [Online]. Available: [\url{https://journals.lww.com/ear-hearing/fulltext/2020/11001/selecting\\_scenarios\\_for\\_hearing\\_related\\_laboratory.3.aspx}](https://journals.lww.com/ear-hearing/fulltext/2020/11001/selecting_scenarios_for_hearing_related_laboratory.3.aspx)
- [6] E. Ceolini and S.-C. Liu, "Combining deep neural networks and beamforming for real-time multi-channel speech enhancement using a wireless acoustic sensor network," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.
- [7] H. Yen, F. Germain, G. Wichern, and J. Le Roux, "Cold diffusion for speech enhancement," 06 2023, pp. 1–5.
- [8] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools Appl.*, vol. 80, no. 4, p. 5653–5671, Feb. 2021. [Online]. Available: <https://doi.org/10.1007/s11042-020-09905-3>
- [9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.
- [10] L. Cheng, A. Pandey, B. Xu, T. Delbruck, V. Ithapu, and S.-C. Liu, "Modulating state space model with slowfast framework for compute-efficient ultra low-latency speech enhancement," 04 2025, pp. 1–5.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '16. IEEE, Jun. 2016, pp. 770–778. [Online]. Available: <http://ieeexplore.ieee.org/document/7780459>
- [12] I. Fedorov, M. Stamenovic, C. R. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "TinyLSTMs: Efficient neural speech enhancement for hearing aids," in *Interspeech*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:218862718>
- [13] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:21722177>
- [14] M. Rusci, M. Fariselli, M. Croome, F. Paci, and E. Flamand, "Accelerating RNN-based speech enhancement on a multi-core MCU with mixed FP16-INT8 post-training quantization," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Cham: Springer Nature Switzerland, 2023, pp. 606–617.