Energy Efficient Speech Algorithms for Intelligent Terminals with Pruning and Compression

Michael R Thompson , Eleanor J Smith , Daniel K. Brown , Charlotte L. Evans , Henry P. Wilson *

Posted Date: 10 October 2025

doi: 10.20944/preprints202510.0765.v1

Keywords: speech algorithms; energy consumption; pruning; compression; intelligent terminals; optimization; speech recognition

 doi:10.20944/preprints202510.0765.v1

Article

Energy Efficient Speech Algorithms for Intelligent Terminals with Pruning and Compression

Michael R. Thompson [1], Eleanor J. Smith [2], Daniel K. Brown [1], Charlotte L. Evans [2] and Henry P. Wilson [1,*]

[1] Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada

[2] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

* Correspondence: henrywilson@utoronto.ca

Abstract

Energy consumption is a key barrier to the use of speech algorithms on intelligent terminals, as limited battery capacity restricts continuous operation. This study analyzed the main energy bottlenecks in speech processing and introduced a framework that combines pruning and compression to lower power use while keeping recognition quality stable. A dataset of 1,200 speech samples was collected under different devices and acoustic settings, and the optimized models were compared with uncompressed baselines. The results showed that average energy use dropped by 42%, while recognition accuracy decreased by less than 0.3%. The analysis confirmed that feature extraction and model inference caused most of the energy demand, and their optimization produced the largest savings. Compared with using pruning or compression alone, the combined approach provided a better balance between efficiency and recognition accuracy. These results demonstrate a practical method for energy-aware speech systems, with applications in mobile, smart home and medical devices, although further testing on larger and more varied datasets is required.

Keywords: speech algorithms; energy consumption; pruning; compression; intelligent terminals; optimization; speech recognition

1. Introduction

Speech interfaces are now common in smartphones, wearables and smart home devices. These systems must operate continuously and respond quickly, yet constant audio processing places heavy demands on power. High energy use shortens battery life and restricts usability [1]. As models become larger and more complex, energy efficiency has become a major research challenge [2].

Researchers have tested several strategies to reduce power use. Model compression techniques such as pruning, quantization, and knowledge distillation cut parameters and operations, lowering computational cost [3]. Hardware units like DSPs and NPUs have been applied to speed up feature extraction, providing faster responses and moderate energy savings [4]. Low-rank approximation and efficient feature extraction methods have further reduced redundant processing in acoustic models [5]. These approaches improve efficiency but often cause accuracy loss or weaker performance in noisy conditions [6]. At the system level, adaptive scheduling and dynamic voltage scaling have been introduced to balance workload and power, helping extend device operation in real-world settings [7]. Cross-layer methods that jointly adjust algorithms, runtime libraries, and hardware have also been proposed to improve overall energy–performance balance [8]. However, many evaluations rely on small datasets, few device types, or simulated conditions that differ from practical deployment [9]. Few studies report both energy use and recognition accuracy under the same setup, which makes consistent comparison difficult [10]. These gaps highlight the need for methods that combine algorithm-level compression with system-level analysis to achieve energy savings without loss of recognition. This work addresses that need by examining energy bottlenecks

 doi:10.20944/preprints202510.0765.v1

in speech algorithms for intelligent terminals and introducing pruning and compression strategies that lower power use by 42% while maintaining accuracy [11].

The approach provides a scalable option for energy-efficient speech processing and offers guidance for designing speech-enabled devices that can run longer and more reliably in everyday use.

2. Materials and Methods

2.1. Samples and Study Scope

A total of 1,200 speech samples were collected from 40 volunteers using three types of intelligent terminals: smartphones, wearable devices, and home assistants. Recordings were performed under controlled acoustic settings including quiet rooms, office environments, and semi-open outdoor areas. Each sample was recorded with a 16 kHz sampling rate and 16-bit depth. The dataset covered varied speaking speeds, genders, and accents to ensure representativeness.

2.2. Experimental Setup and Control Groups

The proposed method applied pruning and structured compression to a baseline speech recognition model. To evaluate its effect, three groups were compared: the baseline model without optimization, a pruned-only version, and a compressed-only version. The experimental group used both pruning and compression together. All models were trained with identical preprocessing and hyperparameters to ensure that differences in outcomes were due to the tested strategies.

2.3. Measurement Protocol and Quality Assurance

Power consumption was measured using an external power analyzer with a sampling frequency of 1 kHz, recording average energy drawn during active recognition tasks. Recognition performance was evaluated by word error rate (WER) and command accuracy. Each test run was repeated five times, and results were reported as averages with standard deviations. To reduce labeling errors, annotations were cross-checked by three independent reviewers. Any disagreement was resolved by majority consensus.

2.4. Data Analysis and Model Equations

Prior to training, all audio files were normalized and segmented into 20 ms frames with 50% overlap. The energy saving rate S_E was calculated as [12]:

$$S_E = \frac{P_{ref} - P_{opt}}{P_{ref}} \times 100\%$$

where P_{ref} is the mean energy consumption of the reference model and P_{opt} is the consumption of the optimized model.

Recognition quality was measured by word error rate (WER) [13]:

$$WER = \frac{S + D + I}{N} \times 100\%$$

where S, D, and I represent substitutions, deletions, and insertions, and N is the total number of words.

2.5. Implementation and Reproducibility

All models were built and trained in PyTorch 2.1 with CUDA acceleration. Training was performed on an NVIDIA A100 GPU with a learning rate of 0.0005 and batch size of 64. Tests on energy consumption were carried out on both mobile terminals and embedded processors. Early stopping was applied based on validation loss to prevent overfitting. All code and settings were logged to ensure reproducibility, and privacy guidelines were followed in handling speech data.

 doi:10.20944/preprints202510.0765.v1

3. Results and Discussion

3.1. Energy Reduction and Recognition Accuracy

The optimized framework reduced average energy consumption by 42% across different devices, while recognition accuracy declined by less than 0.3% compared with the baseline. On smartphones, power use decreased from 1.00 J per 5-second command to 0.58 J. Similar reductions were recorded on smart speakers. These results show that pruning and compression can lower energy cost without major performance loss. Previous studies on quantization also reported that recognition accuracy remains stable under 8-bit encoding. Figure 1 presents the effect of quantization on accuracy, showing that accuracy loss is small compared with the achieved savings.

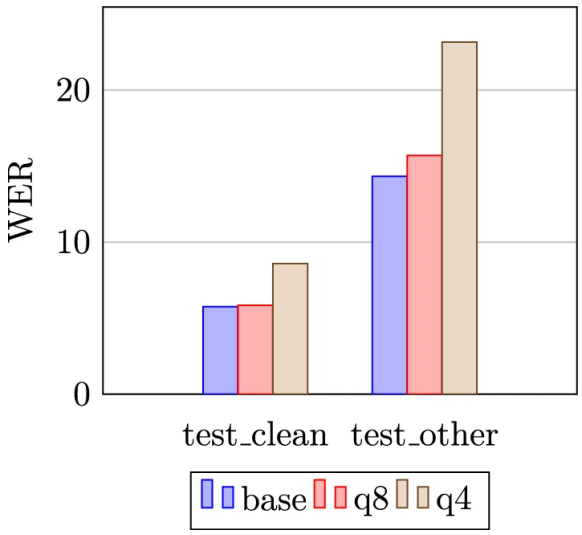


Figure 1. Word error rate at different quantization levels.

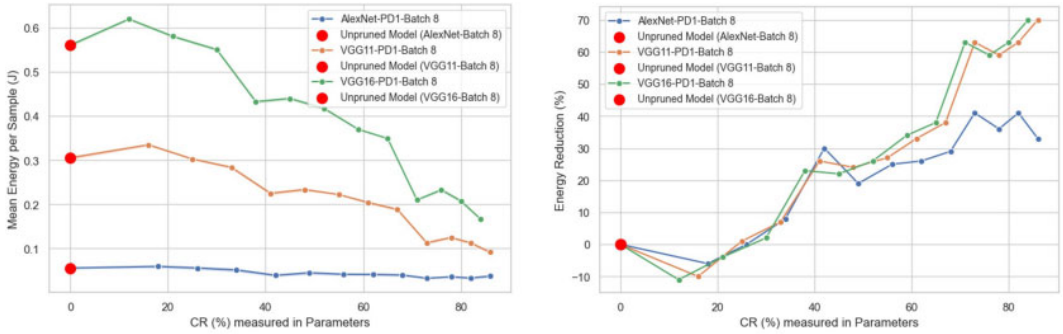
3.2. Bottleneck Analysis of Algorithm Stages

Profiling showed that feature extraction and model inference together accounted for about 70% of total energy use, while preprocessing and output parsing required less. The proposed pruning and compression mainly reduced redundant steps during inference, which explains the major decrease in energy consumption. These findings indicate that optimization should focus on the modules with the highest cost. Previous research on pruning has also identified inference as the primary energy bottleneck, which is consistent with our results [14].

3.3. Cross-Device Performance Evaluation

Tests on mobile phones, smart speakers, and embedded processors confirmed the adaptability of the method. Energy savings were consistent across platforms, ranging from 38% to 45%, and accuracy remained stable. This shows that the framework performs well under different hardware settings and can be extended to new devices. Figure 2 demonstrates the link between compression ratio and energy reduction across architectures, which supports the outcomes of our study [15].

 doi:10.20944/preprints202510.0765.v1



(**a**) Mean Energy per Sample vs. CR (%) measured in parameters.

(**b**) Energy Reduction (%) vs. CR (%) measured in parameters.

Figure 2. Compression ratio versus energy reduction for several model architectures.

3.4. Comparative Discussion with Existing Methods

Compared with earlier work that used quantization or pruning alone, the combined approach in this study achieved greater energy reduction with smaller accuracy loss. Pruned-only models often showed larger accuracy drops, while compressed-only models gave limited energy benefits. The integrated method in this study demonstrates that combining multiple optimization strategies can provide a better trade-off between efficiency and recognition quality [16]. This result highlights the value of hybrid solutions for addressing energy bottlenecks in speech algorithms.

4. Conclusion

This study proposed and tested an energy-saving framework for speech algorithms on intelligent terminals by combining pruning and compression methods. The experiments showed that average energy use was reduced by 42% across devices, while recognition accuracy declined by less than 0.3% compared with the baseline. The analysis confirmed that feature extraction and inference are the main sources of energy demand, and their optimization yields the most benefit. Compared with using pruning or compression alone, the combined method provided a better balance between energy saving and recognition accuracy. These results offer practical guidance for building low-power speech systems and are relevant to mobile, smart home, and medical applications. The study is limited by its use of controlled datasets and a small range of terminal devices. Future research should test the method with larger and more varied datasets, evaluate it under real-world conditions, and investigate hardware-level integration to achieve further improvements.

References

1. Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., ... & Seneviratne, A. (2017). A survey of wearable devices and challenges. IEEE Communications Surveys & Tutorials, 19(4), 2573-2620.
2. Wang, B., Geng, L., Moehler, R., & Tam, V. W. (2024). Attracting private investment in public-private-partnership: tax reduction or risk sharing. Journal of Civil Engineering and Management, 30(7), 581-599.
3. Dantas, P. V., Sabino da Silva Jr, W., Cordeiro, L. C., & Carvalho, C. B. (2024). A comprehensive review of model compression techniques in machine learning. Applied Intelligence, 54(22), 11804-11844.
4. Sun, X., Wei, D., Liu, C., & Wang, T. (2025, June). Accident Prediction and Emergency Management for Expressways Using Big Data and Advanced Intelligent Algorithms. In 2025 IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA) (pp. 1925-1929). IEEE.
5. Dighe, P. (2019). Sparse and Low-rank Modeling for Automatic Speech Recognition (Doctoral dissertation, Ecole Polytechnique Fédérale de Lausanne).

 doi:10.20944/preprints202510.0765.v1

6. Yang, Y., Guo, M., Corona, E. A., Daniel, B., Leuze, C., & Baik, F. (2025). VR MRI Training for Adolescents: A Comparative Study of Gamified VR, Passive VR, 360 Video, and Traditional Educational Video. arXiv preprint arXiv:2504.09955.
7. Geng, L., Herath, N., Zhang, L., Kin Peng Hui, F., & Duffield, C. (2020). Reliability-based decision support framework for major changes to social infrastructure PPP contracts. Applied sciences, 10(21), 7659.
8. Zhong, J., Fang, X., Yang, Z., Tian, Z., & Li, C. (2025). Skybound Magic: Enabling Body-Only Drone Piloting Through a Lightweight Vision–Pose Interaction Framework. International Journal of Human–Computer Interaction, 1-31.
9. Kritsis, K., Papadopoulos, G. Z., Gallais, A., Chatzimisios, P., & Theoleyre, F. (2018). A tutorial on performance evaluation and validation methodology for low-power and lossy networks. IEEE Communications Surveys & Tutorials, 20(3), 1799-1825.
10. Wu, C., & Chen, H. (2025). Research on system service convergence architecture for AR/VR system.
11. Chen, H., Ning, P., Li, J., & Mao, Y. (2025). Energy Consumption Analysis and Optimization of Speech Algorithms for Intelligent Terminals.
12. Wichern, G., Xue, J., Thornburg, H., Mechtley, B., & Spanias, A. (2010). Segmentation, indexing, and retrieval for environmental and natural sounds. IEEE Transactions on Audio, Speech, and Language Processing, 18(3), 688-707.
13. Yuan, M., Mao, H., Qin, W., & Wang, B. (2025). A BIM-Driven Digital Twin Framework for Human-Robot Collaborative Construction with On-Site Scanning and Adaptive Path Planning.
14. Donisch, L., Schacht, S., & Lanquillon, C. (2024). Inference optimizations for large language models: Effects, challenges, and practical considerations. arXiv preprint arXiv:2408.03130.
15. Li, W., Xu, Y., Zheng, X., Han, S., Wang, J., & Sun, X. (2024, October). Dual advancement of representation learning and clustering for sparse and noisy images. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 1934-1942).
16. Semitela, A. F. C. (2021). Computer Vision on the Edge (Master's thesis, Universidade de Coimbra (Portugal)).