International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V7I1P117
Eureka Vision Publication | Volume 7, Issue 1, 119-123, 2026

Original Article

Real-Time Speech and Audio Processing for Mobile Applications

Dheeraj Vaddepally
Independent Researcher, USA.

Received On: 20/12/2025 Revised On: 19/01/2026 Accepted On: 28/01/2026 Published On: 06/02/2026

Abstract - Real-time audio and speech processing are now a necessary component of modern mobile applications, enabling users to engage comfortably with voice-based inputs and speech recognition capabilities. Mobile environments do come with inherent challenges such as limited computation, battery constraints, and the need for low-latency noise-immune solutions. The present paper addresses the design and implementation of real-time speech recognition and command-detection systems optimized for mobile platforms. We focus on low-latency methods and algorithms for wake word spotting and command understanding in order to provide quick reaction to user feedback. We also address noise-robust modeling techniques to battle the various acoustic scenes of mobile scenarios, making the most out of deep neural networks architectures, noise-irrelevant feature learning, and data augmentation. For optimizing model performance on mobile, we discuss model compression, pruning, and the benefits of on-device inference for real-time processing. Next, we discuss case studies and identify future directions and challenges in deploying real-time speech systems on mobile, in terms of accuracy vs. latency vs. power consumption. The results discussed herein present a comprehensive framework for building mobile-based speech and audio processing technology.

Keywords - Real-Time Speech Processing, Mobile Applications, Low-Latency Recognition, Command Detection, Noise-Robust Modeling, Deep Learning Architectures, Feature Extraction, Model Compression, On-Device Inference, Real-Time Optimization.

1. Introduction

Real-time audio and speech processing have emerged as a significant feature of mobile applications in recent times, leading the way in the development of voice-based interfaces, virtual assistants, and hands-free functionality. Mobile applications, such as smartphones, wearables, and smart home devices, are increasingly employing speech recognition for various purposes like voice commands, voice search, and personal assistants. This shift towards voice-based interaction is extremely convenient, especially in hands-free or screen-limited scenarios. However, implementation of real-time speech recognition in mobile phones is confronted with a variety of technical issues primarily due to the need for low-latency output, noise resilience, and efficient use of available computation.[1]

Low-latency speech recognition is key to providing an interactive and natural user experience. Voice commands should be processed and results delivered back within mobile apps in real time with as little latency as possible to enhance user satisfaction. It can be done through optimized algorithms along with efficient hardware-software synergy in particular for mobile environments.[1] Besides, mobiles are commonly operated in noisy, unreliable acoustics such as busy roads, public transports, or in vehicles. Noise-robust modeling methods are therefore essential to speech recognition systems to maintain their accuracy under such situations.

This paper presents the techniques and strategies required to deploy real-time speech and audio processing on mobile phones. The paper focuses on two areas: low-latency speech recognition and command detection, and noise-robust modeling and implementation [2]. By using these solutions, the paper aims to realize the best possible performance on resource-constrained mobile hardware while guaranteeing the correctness and reliability of speech processing systems.

The following sections will describe in detail real-time speech and audio processing, low-latency recognition and noise-robustness approaches, and optimization techniques for deployment in mobile environments. The paper also reviews realistic case studies and conclude with a discussion of future directions in this topic.

2. Overview of Speech and Audio Processing

Speech and audio processing involve the conversion of acoustic signals into a computationally processable form that can be accessed and understood by computers so that decisions may be made based on it. These operations are the foundation of all real-time applications like voice control, transcription, and interactive voice response. In mobile phones, the most significant challenges are speech processing efficiently and accurately within real-time constraints due to limited computation and battery power.

2.1. Speech Processing Principles

Speech processing starts with digitization of audio sound waves of the speech to digital signals. The system extracts different steps on digitizing in order to yield meaningful information of the audio signal for recognition, classification, or command detection.[3]

Extraction: Feature extraction involves the isolation of the basic features of the speech signal that reflect meaningful features and elimination of unimportant or unnecessary details. Usual features include:

- Spectrogram: Graphical representation of the audio signal's frequency spectrum as a function of time. It assists in identifying frequency patterns, which are important for specifying the various phonemes or speech sounds.[2]
- Mel-Frequency Cepstral Coefficients (MFCCs): One of the most widely employed feature extraction methods employed in speech processing. MFCCs attempt to imitate the hearing process in humans by emphasizing frequency components of higher significance in speech recognition. MFCCs compress the dimensions of audio signals without eliminating informative data.
- Linear Predictive Coding (LPC): LPC is a linear predictive coding of the speech signal within a linear predictive model, where the speech envelope is modeled as vocal tract prediction. LPC is applied to compress speech data and extract features for effective speech recognition.

These features extracted are the foundation for further speech recognition and classification processes.

2.2. Real-Time Audio Processing

Real-time audio processing is the capability to record, process, and analyze audio data with little delay, such that applications can respond in real time to user input. In contrast to offline processing, which can provide more computationally costly options and greater latency, real-time processing places both latency and computational cost constraints.

Low Latency: In real-time systems, speech inputs should be done by the system and results or responses have to be given in terms of milliseconds to achieve an interactive and smooth user interface. Low latency would be required in voice-controlled assistant applications where prompt feedback is expected. Special hardware, i.e., Digital Signal Processors (DSPs), would most likely be used in mobile devices to obtain the required speed of processing.[4]

2.3. Differences Between Offline and Real-Time Processing

- Computational Limitations: Offline speech recognition can use more powerful and computationally intensive models (e.g., large transformer models, deep neural networks) since they can use high-end cloud servers. Real-time mobile speech recognition, however, is limited to light models that can be executed within the confines of limited hardware resources at the cost of accuracy versus performance.
- Latency Tolerance: Offline processing can withstand higher latency since there is no specific need for instant response. This permits increased model complexity and size flexibility. Real-time processing requirements must, however, be under strict latency limits at some point with compressed or optimized model use necessitated in order to fulfill such a requirement.[4]
- Power and Battery Limitations: Real-time audio processing, as applied to mobility, has to be sensitive to the power capacity of the battery on the system. In addition to speed optimization, models and algorithms need to optimize for power consumption so that the battery gets used up quickly.

3. Low-Latency Speech Recognition and Command Detection

Low-latency voice recognition and command interpretation are essential in real-time mobile use to give users immediate response. Low latency, when used on mobile phones, reacts in terms of milliseconds to make the operations interactive and fluid. Where processing is slow for operations like voice command interpretation or virtual assistant conversation, performance is unacceptable and makes the system seem unresponsive or slow. It is most difficult to prefer low-latency processing on mobiles, where the limitations are much more pronounced than the case of cloud-based solutions.[5]

Alternative algorithms that have to be computationally cheap and able to maintain accuracy are generally preferred by mobile speech recognition. Legacy models such as Hidden Markov Models (HMMs) have been conventionally applied in speech recognition since they can represent sequential data. But the deep learning architectures, Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have been trending in recent years due to their high performance in handling complex speech patterns. LSTMs are particularly well-suited to capture long-term dependencies of the speech sequence for accurate transcription and command identification.

Keyword spotting and wake word detection techniques are crucial in the interest of command recognition. These methods are employed to passively wait to hear specific words or phrases, such as "Hey Siri" or "Okay Google," and then activate more advanced recognition algorithms. Keyword spotting is typically optimized for low-latency detection with light models that run continuously in the background. Low power usage and low latency are facilitated by optimized algorithms combined with techniques such as model pruning and quantization.

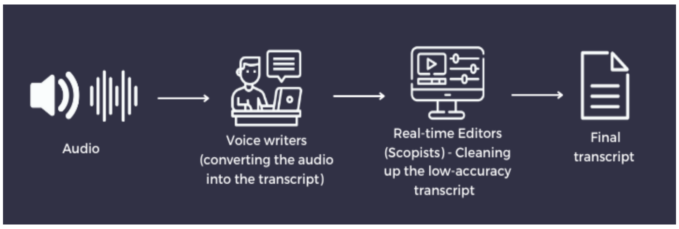
Audio
Voice writers (converting the audio into the transcript)
Real-time Editors (Scopists) - Cleaning up the low-accuracy transcript
Final transcript

Figure 1. Working with ASR

Mobile hardware is also tasked with providing low-latency performance. Accelerators and Digital Signal Processors (DSPs) are widely used to offload the processing load from the processor so that audio can be processed better. Software frameworks such as ONNX and TensorFlow Lite also enable developers with the optimization of machine learning models for mobiles, thus allowing real-time voice recognition and command identification. These frameworks provide tools to quantize models and map them to on-device inference-efficient representations, reducing latency and power consumption.[6]

4. Noise-Robust Modeling and Implementation Strategies

Noise is especially a major problem in mobile settings, where speech recognition models must work well under an incredibly broad range of acoustic conditions. Noises may arise from background talk, traffic noise, wind, and echoes, all of which can have the impact of decreasing speech recognition model performance significantly. This is especially challenging for mobile phones since users are most likely to utilize their phones in noisy, uncontrolled environments, and therefore it is essential to create models that can perform effectively in such environments.[7]

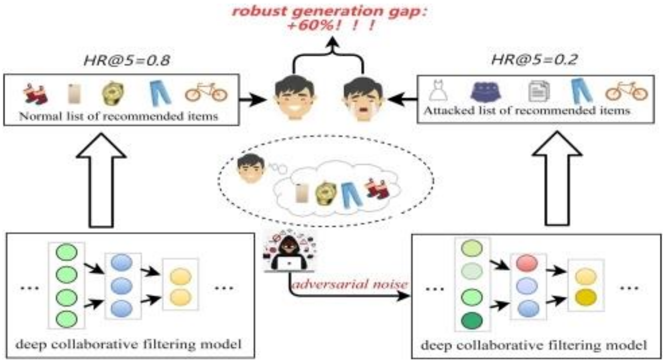
robust generation gap: +60%! ! !
HR@5=0.8
HR@5=0.2
Normal list of recommended items
Attacked list of recommended items
adversarial noise
deep collaborative filtering model
deep collaborative filtering model

Figure 2. Robust Voice Generation and Filtering Model Architecture

While creating noise-robust models, a number of methods can be employed. One of them is enhancing feature extraction by emphasizing noise-invariant features that are more resistant to environmental distortions. Methods such as spectral subtraction, in which the noise is estimated and subtracted from the speech input, have been shown to enhance speech intelligibility in noise. More recently, newer models such as Convolutional Neural Networks (CNNs) and attention-based models have also been shown to work well on noisy speech data. CNNs are suitable to acquire local patterns of speech of the speech signal, and transformer models and attention mechanisms are suitable to pay attention to areas of input that must be examined so there is higher recognition accuracy even under poor conditions.[7]

Another practice is to use data augmentation, where artificial noise is added to training data to train the model to generalize in noisy conditions. Training the model with numerous noisy conditions gives it more robustness when deployed in real-world conditions. Apart from using real noisy real-world data, synthetic data can be created to simulate different noise environments and offer more robustness.

Mobile-device applications of noise-resistance methods should achieve a compromise between computation cost and energy consumption. More sophisticated models and noise-reduction approaches provide better accuracy but come with the sometimes-additional processing load. Mobile devices must walk a line between such demands and the necessity to maintain in-place real-time operation and low energy consumption.[8] Optimization techniques like model pruning, quantization, and hardware accelerators are the reason that noise-resilient speech recognition systems remain responsive and efficient on mobile devices. Using these methods, developers can design speech recognition systems to perform well in a broad variety of acoustic conditions without violating the real-time and power efficiency limits of mobile devices.

5. Optimization Techniques for Mobile Applications

With increasing sophistication of speech and audio processing models, these need to be optimized for deployment on the mobile platform for delivering efficient runtime performance in resource-constrained environments. Model compression, deployment onto edge devices, and performance over battery consumption trade-offs are most essential optimization strategies.

5.1. Model Compression and Pruning

For mobile deployment of sophisticated speech recognition models, model compression with no negative impact on performance is essential. Methods such as quantization, pruning, and knowledge distillation are popular model compression techniques for this. Quantization is the reduction of the model's weights and activations precision from 32-bit to 16-bit or even 8-bit, reducing models in size and executing faster on mobiles. Pruning reduces redundant or unnecessary parameters, making the whole model simpler. This method improves the performance without reducing it but at the reduced cost of calculations. Knowledge distillation is also a potent method by which a light model (student) is trained from a large one (teacher) and can thereby utilize light, energy-efficient models and maintain slight loss of accuracy. Edge AI and On-Device Inference [9]

5.2. Edge AI and On Device Inference

Edge AI, or the implementation of machine learning models inside real devices, forms an inherent part of the entire process of audio and speech processing in real-time.

Inference within a device reduces the latency to almost zero, thus avoiding data communication over the internet for analysis on a cloud server. The resulting effect makes applications, including voice assistants and voice command recognition, immediately fast and responsive. Moreover, execution within a local device encourages enhanced security and confidentiality since no web communication is required to protect users' information. However, there are some trade-offs with cloud vs. on-device processing.

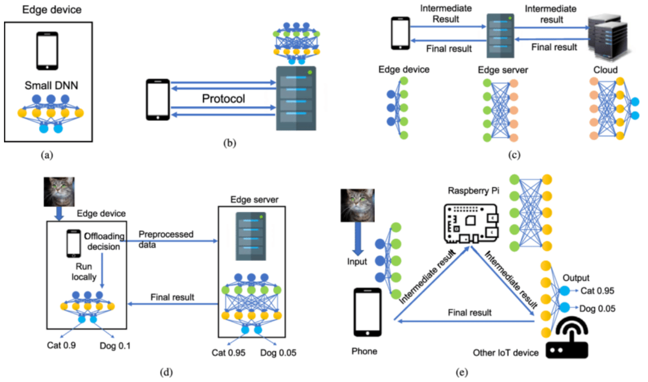

Figure 3. Architectures for Deep Learning Inference with Edge Computing

Cloud services, on the other hand, offer increased compute and can support more advanced operations well, but involve lag and require connectivity to a fast network. Inference on device, in contrast, gives zero response latency but must be properly optimized in devices with scarce memory and processing capabilities. Software packages such as TensorFlow Lite and ONNX support edge AI via their capability for deployment of model-optimized models onto mobile hardware. Battery and Performance Trade-offs [8]

5.3. Battery Performance Trade Off

One of the greatest challenges of mobile apps is achieving a balance between real-time ability and battery life. High-cap models, though accurate, drain a mobile device's battery quickly since they are computationally intensive. To compensate for this, developers must achieve a balance between real-time capability and power usage. Techniques such as adaptive model selection, where simpler models are used for less intense tasks, can be utilized in order to regulate power usage.

Further, utilizing special hardware, i.e., Digital Signal Processors (DSPs) and power-saving accelerators, intensive computations are performed by power-efficient devices. Power saving in speech recognition systems is also very important when the devices need to perform background processing of tasks such as keyword spotting. By reducing model size and making use of capabilities provided by hardware, developers can offer real-time audio processing without draining batteries too much. [10]

6. Case Study and Applications

Speech and audio processing are built into numerous mobile applications these days, and there are a few significant real-world applications and situations where these technologies have been used successfully.[6]

6.1. Well-known Mobile Speech Recognition Systems

There are a couple of well-known popular mobile speech recognition systems, including Google Assistant, Apple Siri, and Amazon Alexa. These applications are constructed upon industry-leading speech understanding algorithms to provide customers precise real-time feedback as an answer or direction. Every one of these platforms applies both cloud and local processing in mix for the benefit of precision as well as execution. As an example, detection of wake word (e.g., "Hey Siri" or "Alexa") is often done locally within the device, while more involved speech recognition capability can be cloud-offloaded so that it processes with higher precision and less system resource usage locally.

Google Assistant and Apple Siri are equally synchronized with most mobile and IoT devices so that customers can operate home smart conditions remotely, schedule events, send messages or play music by issuing a voice command. Amazon Alexa has also made contact in smart speakers, reaching out to home automation and voice control of various devices.

6.2. Use Cases

Real-time audio and speech processing capability makes possible a large number of new, innovative applications for enhancing the experience of performing ordinary mobile tasks by the users.

- Voice-Controlled Applications: Various applications from productivity suites to home appliances are now implementing voice control features to provide ease of accessibility and convenience even further enhanced. For instance, voice-controlled note-taking applications allow users to write reminders or lists remotely.
- Real-Time Language Translation: Mobile apps like Google Translate utilize speech-to-text recognition and translation algorithms to facilitate real-time communication between speakers of different languages. These apps usually combine automatic speech recognition (ASR) with machine translation algorithms to deliver accurate, near-simultaneous translations.
- Hands-Free Navigation: Voice-controlled navigation applications, for instance, those employed in vehicle infotainment, enable motorists to navigate and interact with their phones without having to take their hands off the wheel at all. Such applications rely substantially on low-latency speech recognition in order to provide real-time, safety-enhancing functionality in a car.

With mobile hardware optimization, these apps demonstrate how audio and speech processing in real time have been integrated into the user experience and transformed the way we engage with technology.

7. Future Directions and Challenges

The future of mobile speech recognition is packed with promising prospects. One of the areas where innovation may take place is via enhancement of transformer models and attention mechanisms, which have already transformed natural language processing. These models can be made even more optimized on mobile platforms to deliver faster and more accurate speech recognition systems. Hybrid approaches incorporating the conventional machine learning techniques, such as Hidden Markov Models (HMMs), with deep learning techniques, such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), will also be popular. They have the potential to provide the strengths of both worlds reaping the benefits of conventional reliability while taking advantage of the power to effectively deal with large amounts of data through deep learning.

Despite all these promising developments, there remain significant issues with the use of mobile speech recognition systems. One of the most important is privacy since cell phones will be processing extremely personal data. Speech recognition needs to keep on performing well and securely, and that will mean even more innovation in on-device processing to minimize data transmission to the cloud. The other problem will be to advance speech recognition with various accents and dialects. The models in existence for this at present are subpar, and it will be essential to develop more variable systems that can fit into an extremely broad spectrum of speech patterns to increase wider use. Further, advancing advanced models to resource-constrained mobile platforms such as less capable batteries and processing capacities used in some cell phones will also continue to be an issue. Edge computing, so full of promise in name, but yet to be developed to handle the requirements of low-latency and real-time applications without undermining the performance of devices.

8. Conclusion

We addressed here the most critical points of mobile real-time speech and audio processing, i.e., low-latency speech recognition, noise-robust modeling, and optimization feasible in mobile systems. We emphasized why low-latency is a blessing for enhancing the user experience, how methods for constructing noise-robust models are employed, and why efficiency in terms of utilization of mobile hardware to trade performance for battery consumption was needed.

In the future, better speech recognition models, particularly based on transformers and attention, will define the future of smartphone applications. However, challenges such as privacy, handling different accents, and applying sophisticated models on resource-constrained devices are still to be addressed. With constant research and development, real-time speech and audio processing within smartphones will definitely become more efficient, accurate, and omnipresent, paving the way for a wide range of new and advanced applications.

References

[1] Sehgal, A., & Kehtarnavaz, N. (2018). A convolutional neural network smartphone app for real-time voice activity detection. IEEE access, 6, 9017-9026.

[2] Omyonga, K., & Shibwabo, B. K. (2015). The application of real-time voice recognition to control critical mobile device operations.

[3] Bhat, G. S., Shankar, N., Reddy, C. K., & Panahi, I. M. (2019). A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone. IEEE Access, 7, 78421-78433.

[4] Gokul, G., Yan, Y., Dantu, K., Ko, S. Y., & Ziarek, L. (2016, August). Real time sound processing on android. In Proceedings of the 14th International Workshop on Java Technologies for Real-Time and Embedded Systems (pp. 1-10).

[5] Sehgal, A., & Kehtarnavaz, N. (2018, January). Utilization of two microphones for real-time low-latency audio smartphone apps. In 2018 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-6). IEEE.

[6] Deligne, S., Dharanipragada, S., Gopinath, R., Maison, B., Olsen, P., & Printz, H. (2002). A robust high accuracy speech recognition system for mobile applications. IEEE Transactions on Speech and Audio Processing, 10(8), 551-561.

[7] Drakopoulos, F., Baby, D., & Verhulst, S. (2019). Real-time audio processing on a Raspberry Pi using deep neural networks (pp. 2827-2834). Universitätsbibliothek der RWTH Aachen.

[8] Ghosh, R., Ali, H., & Hansen, J. H. (2021). CCi-MOBILE: A portable real time speech processing platform for cochlear implant and hearing research. IEEE Transactions on Biomedical Engineering, 69(3), 1251-1263.

[9] Iwaya, Y., & Katz, B. F. (2018). Distributed signal processing architecture for real-time convolution of 3d audio rendering for mobile applications. In Virtual Reality and Augmented Reality: 15th EuroVR International Conference, EuroVR 2018, London, UK, October 22–23, 2018, Proceedings 15 (pp. 148-157). Springer International Publishing.

[10] SM, U. S., & Katiravan, J. (2022). Mobile application based speech and voice analysis for COVID-19 detection using computational audit techniques. International Journal of Pervasive Computing and Communications, 18(5), 508-517.