Al-Rafidain Journal of Computer Sciences and Mathematics (RJCSM)

www.csmj.uomosul.edu.iq

A Comprehensive Review of Offline Voice Control Systems for Smart Homes: Technological Advances, Existing Challenges, and Prospective Developments

Dalya Talal[1], Azam Al-Rawachy[2] and Abdalla Eblabla[3]

[1]Department of Medical Laboratories Techniques, Mosul Medical Technical Institute, Northern Technical University, Mosul, Iraq
[2]College of Computer Science and Mathematics, Software Department. University of Mosul, Mosul 41002, Iraq
[3]Center for High Frequency Engineering (CHFE), School of Engineering, Cardiff University, CF10 3AT Cardiff, U.K
Email: prog.dalya@ntu.edu.iq[1], azzam.esam@uomosul.edu.iq[2] and eblablaA@cardiff.ac.uk[3]

Article information

Article history:
Received 14 November, 2025
Revised 5 January, 2026
Accepted 15 February, 2026
Published 25 June, 2026

Keywords:
Offline voice control,
Smart homes,
TinyML,
Keyword spotting,
Embedded speech recognition,
Privacy,
Edge AI

Correspondence:
Dalya Talal
Email: prog.dalya@ntu.edu.iq

Abstract

The offline voice control systems are becoming more prominent in the smart home environment because of their ability to deliver quick feedback, preserve the privacy of the user, and be able to function even without network connectivity. However, their usage has not been mainstreamed, due to restrictions on vocabulary, resistance to noise, and hardware performance. In this paper, the authors attempt to provide a critical assessment of the recent developments in offline voice control technology in smart homes. On devices with limited resources, it evaluates a variety of datasets, frameworks, embedded platforms, and algorithmic enhancements. Moreover, the paper examines the issues linked to the growth of vocabulary, audio variability, and privacy, as well as interoperability and compares offline and online paradigms in terms of a more detailed analytical framework. Among the emerging trends highlighted include lightweight keyword spotting, embedded spoken language understanding, neural processing units, and on-device personalization, as being crucial in the evolution of future systems. This paper summarizes the current situation and outlines the future trends of the offline voice control technologies as it is applicable to smart home usage by synthesizing recent achievements and pinpointing the current limitations of the technology.

DOI: 10.33899/rjcsm.v20i1.60652,

1. Introduction

Voice control has become an essential part of smart homes which allows users to interact with the devices they are connected to in an intuitive way. The typical voice recognition systems are based predominantly on processing via the cloud, thus utilizing the comprehensive acoustic/linguistic models and, therefore, leading to high-quality recognition and large vocabularies [1, 2] . Even though this solution provides acceptable performance there is a need to have a good internet connection and, in most cases, data is sent remotely to servers, and this raises question of latency, privacy and reliability. By contrast, offline voice control systems process speech directly on the embedded devices, they are faster to respond to and operate in the face of lack of network connectivity as well they provide greater privacy [3, 4] . Offline solutions are particularly attractive in the setting where network connectivity can be limited or where security is a major priority [5, 6, 7].

Several disadvantages exist for offline systems. Limited memory and computing capacities of traditional hardware limit the sophistication of models [8, 9]. Vocabulary sizes and adaptability to noisy or adaptive acoustic learning are limited, thereby restricting resilience to these conditions [10, 11]. Additionally, unavailability of cloud service makes updating of the system difficult, expands language restrictions, and prevent connectivity to a variety of devices. To meet these trade-offs, work has been done to develop smaller algorithms, more efficient

frameworks, and dedicated hardware that can provide high-level functionality while using relatively little power and memory [12, 13, 14].

The paper aims to give a detailed and critical analysis of off-line voice-control technologies that would be used in smart homes. It identifies the key elements needed in the development of effective offline systems such as datasets, toolkits, embedded platforms, progress in algorithms, and performance tests. The paper compares the paradigms of offline and online approaches, highlights the ongoing problems and addresses the new directions in the field, including the key-word spotting, embedded language understanding in speech and on-device personalization. The aim is to synthesize the recent research to determine the present capabilities in this area as well as the prevailing gaps.

The paper is structured as follows: Section 2 discusses the technical background whereas in Section 3, the description of the main elements of the system is described. Part 4 to 7 discuss data sets, frameworks, technology innovations, and current limitations. Section 8 provides the comparative analysis of offline and online strategies, Section 9 discusses the examples of implementation, Section 10 considers the system performance, and Section 11 represents the future perspectives.

2. Review Methodology

In this paper, the review approach chosen is structured narrative in the analysis of recent developments in offline voice control systems in the context of smart home applications. A systematic search strategy and a qualitative synthesis were used to conduct the review which should offer an in-depth but concise evaluation of technologies, platforms, datasets, and performance trade-offs applicable to on-device voice processing.

This review had inclusion criteria as follows: (i) the study had to be offline or on-device voice processing without cloud based recognition, (ii) the study had to be in smart home or embedded system context and (iii) the study had to provide at least one measurable performance metric which could be recognition accuracy, latency, power consumption or memory footprint. The studies that were limited to cloud-based speech recognition, proprietary systems with undisclosed evaluation metrics or articles that were not related in any way to embedded or smart home application were all eliminated.

After the filtering procedure, the set of studies was interpreted and classified on the basis of main system components, design dimensions, such as datasets and benchmarks, algorithmic methods, in-built structures, hardware platforms, and performance attributes. Instead of a meta-analysis, the review focuses on qualitative comparison and trend analysis, demonstrating that trade-offs that are common to accuracy, latency, energy efficiency, and vocabulary size are common across various offline voice control systems.

The developed narrative method allows synthesizing uneven results that are presented in the literature without losing clarity as to assessment conditions and system limitations. The methodology will guarantee that the performance measurements and design data are understood within the context of their respective experiments to enable the reproducible and balanced comparison of the offline voice control technologies to be used in smart homes.

Peer-reviewed journal articles and reputable conference papers that are relevant to the requirements of offline voice control, keyword spotting, and on-device speech recognition in smart home and edge computing environments were prioritized in the selection. Thematic grouping, instead of statistical grouping, of system-level implementations, model architectures and evaluation metrics was done since the main goal was to synthesize design trends, performance trade-offs, and practical constraints rather than to conduct a quantitative meta-analysis.

3. Technical Background and Core Components

The proper analysis of the offline voice control systems is not possible without the knowledge of the basic technologies underpinning their functioning. What these systems have in common are a number of closely interconnected stages that make up an entire signal-processing chain which starts with the acquisition of acoustic signals and finishes with the execution of local commands. It starts with capturing of the sound by using a microphone which acts as the main point of contact between the user and control mechanism. The audio signal captured is then handled at feature extraction stage where different characteristics that are frequency- and amplitude-related are used to create an informative and succinct representation of the speech signal. This is a very crucial step as it prepares the data to follow through the further processing of data because it highlights the important features at the expense of the irrelevant features hence enhancing the recognition reliability as well as computational efficiency [15, 16, 23, 24].

After extracting the features, the system then go further to the recognition interval whereby key word spotting (KWS) algorithms are used to identify the existence of the extracted features in relation to the predefined commands or activation phrases. KWS is at the heart of the offline voice control because it provides the ability to detect wake-word and issue commands at low levels of latency and energy consumption. The current developments of lightweight KWS models have played a crucial role in securing recognition accuracy on small-sized embedded systems. Specifically, small-footprint convolutional neural

networks (CNNs) with phoneme-level or phoneme-scale decoding strategies have proven to have high potential of real-time embedded voice recognition. The phoneme-based decoders are however limited by the use of memory and computation burden particularly when subjected to low-resource microcontrollers [17, 18, 25, 26].

The development of the offline voice control systems in smart-home contexts also requires the choice and combination of appropriate hardware platforms. Simple voice command applications have been well taken up using hardware packages like Elechouse VR V3 using Arduino Uno boards that have been shown to be easy to integrate and to execute commands reliable. In higher-performance applications, dual-core processors can be used, like the ESP32, with built-in Wi-Fi and Bluetooth functionality available, making it easy to support larger vocabularies and more complex recognition pipelines [19, 20, 27, 28]. Although these platforms may have such benefits, they frequently demand more of the effort of the developers as complex signal-processing algorithms and numerous layers of software libraries are required to realize full utilization of their capabilities.

Wake-word-based architectures also provide additional benefits like energy efficiency and system responsiveness with the ability of an offline voice control system to hold an off-line listening state until a certain activation phrase is recognized. It is a very efficient method of saving power and at the same time sustaining speedy reaction of the system as soon as a valid command is detected, enhancing the total user-friendliness and compatibility in long-duration functioning in smart-home scenario interactions [25, 27].

The effectiveness of the offline voice control systems depends highly on the ambient noise conditions in real-world deployment. Noise-robustness methods are, therefore, crucial in giving credible functionality. Adaptive filtering schemes and noise suppression machine-learning algorithms are typically used in isolating speech commands under background interference to improve recognition accuracy and usability of a given system under aesthetically demanding conditions [21, 22, 29, 30]. The maintenance of good performance in various noise conditions is one of the major research problems with an active field in assessment.

An example of a regular setup of an offline voice control system that can be used in smart-home applications is shown in Figure 1. Acoustic signals are sampled by a microphone and processed by preprocessing and feature extraction to come up with compact representations that can be processed by embedded systems. Wake-word detection is done with a lightweight keyword spotting engine after which spoken commands are classified in a local recognition engine. The resulting control signals are sent directly to target devices so that fully offline operation is possible without reliance on cloud-based infrastructure.

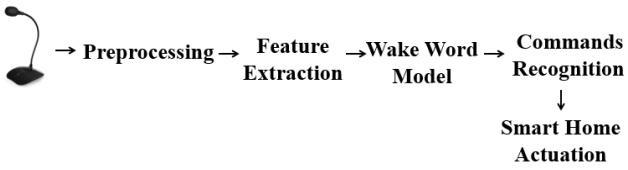

Figure 1. The standard architecture of an offline voice control system designed for smart home applications.

4. Datasets and Benchmarks

The massive selection of the datasets considerably affects the efficiency of the offline voice control systems. There is a difference in constraints between embedded models for public corpora versus private. Although the standard evaluation measures, including accuracy, false accept/ reject rates, and area under the curve (AUC), are imperative, they fail to evaluate the important trade-offs that are critical during embedded deployment. Other measures like energy use per inference, memory usage, and airtime at high frequencies of wake-word activation are becoming important items in embedded metrics which ought to be highlighted with other measure items [31].

The main reference of tiny keyword spotting (KWS) models is the Google Speech Commands dataset and plays a vital role in impacting the TinyML literature surrounding smart home devices. Nonetheless, this dataset is mainly made up of brief utterances with relatively clean-speech conditions, in fact, it is easy to find a model that performs well on such a baseline that reduces in accuracy by 10-25% when used in a far- field, reverberant, or multi-speaker condition much like someone talking at home [32]. This variation indicates that the models used in KWS should be subjected to training or modification in terms of real-world application as part of laboratory results to optimal by considering the additions of realistic noise and room impulse-response models to be effective in practice [33].

A number of such surveys have covered TinyML in varied perspectives. General ideas on TinyML, hardware limitations and target deployment issues are captured by broad overviews whereas more application oriented use case and benchmark issues are captured by the others. Unlike those, the given review is not expected to replicate a general TinyML survey instead it summarizes and puts into context previous results concerning offline voice control and keyword spotting systems in smart home settings with a focus on system-level trade-offs, privacy concerns, and operational constraints.

Speaker-dependent and private corpora may be highly accurate on-device when addressing narrow-purpose tasks but are poor in generalization and working outside the domain. Their pragmatic benefit is that vocabulary reduction and the ability to make the acoustic environment

controlled, over smaller form factors can enable smaller models to have a sub 1MB footprint and yet learn to run at acceptable latency. The negative thing is, though, that what the data based on such datasets can be compared to across systems cannot be simply achieved without unified data gathering and assessment measures [34] . A suggested solution to success in smart-home systems is a combination of such small public corpora to train them with their work as the basis and lightweight, on-the-board adaptation steps on sample collections within family settings to reduce the steps in deployment differences [35].

5. Frameworks, Toolkits, and Embedded Platforms

The supported frameworks and embedded platforms stand as remarkable to the real-life performance of the offline voice control systems. Embedded frameworks have to balance the entire model accuracy, latency, power consumption and memory utilization unlike cloud-based solutions where the model complexity is primarily limited by the server capabilities [36]. To achieve this balance, in most systems event-driven activation schemes are used, where the device is in an ultra-low-power mode until a wake word is heard after which the mainstream recognition processing is initiated. This architecture is good to contain the waste of energy and retain the immediate responsiveness, yet it has restrictions on the size of the models and also on the accuracy of the wake detector which may need careful handling on the part of the real-life applications [37].

Lightweight machine-learning frameworks have been specifically designed to alleviate these limitations. An example is TensorFlow Lite Micro which allows the use of small neural networks with as little as several tens of kilobytes of RAM since it does not use dynamic memory allocation but instead provides a fixed memory plan. In real-time voice interactions, predictable latency and simplified deployment are vital aspects. In addition, Edge Impulse features an upper-level tool chain providing data collection, model training and deployment over a wide range of hardware, reducing development effort at the sacrifice of slightly larger binaries and fewer controls over manual optimisation gain [38] . The two frameworks enable developers to avoid cloud reliance and still take advantage of modern model frameworks, even in resource-constrained surroundings [39].

The hardware choice also affects the limits of operating the system. Other modules like the Elechouse VR V3 are simple, packaged solutions which recognise isolated commands with the least development effort. Their command set is limited to about eighty entries and could be sufficient with simple applications to the smart home but might not be sufficient with more complex applications [40]. Microcontrollers such as the Arduino Uno offer a readily available platform upon which you can combine these modules but are hindered by the limited computing power making it useful when you want a fixed set of commands and use it in simple control. It also adds a significant boost in capability thanks to its dual-core processor and integrated connectivity, allowing for more sophisticated signal-processing and greater vocabulary support [41]. However, to make proper use of the ESP32, specific firmware development and better library choices are usually required to extract features, deal with buffers and inferences in the most efficient way possible [42].

It is evident from these differences that developers will have to make trade-offs between easy-to-install turnkey modules that require little power, but do not offer flexibility and programmable microcontroller platforms that are more flexible and easier to control, without complicating development. The choice of which framework/hardware combination to use depends on matching application needs, e.g. vocabulary size and latency goals, privacy requirements capabilities with the computational needs. With the ever-evolving nature of embedded frameworks and the continued increase in number and variety of hardware accelerators, these trade-offs are moving such that it is now possible to implement rather complex model implementations fully offline without running out of power and memory [43].

6. Technological Advances

Advances in technology have heavily redesigned the feasibility of the offline voice-controlled infrastructures in smart space. The move toward wholly local calculation rather than server-based processing has been inspired by the improvement in both hardware and algorithms techniques specially created to meet the high requirements of latency and power constraints of embedded devices. Time-constrained microcontrollers require the use of off-the-shelf and neurotically optimised processing pipelines due to the real-time nature of their use. In contrast to cloud-based systems when latency is caused by delivering the network transmission, offline systems are required to provide recognition and response in milliseconds ranges, but within very stringent power constraints often only limited to a few milliwatts. The achievement of such a fine balance will require detailed optimisation at every stage of the signal processing process as well as specialized architecture [44].

According to recent works on lightweight keyword spotting KWS show that even in a constrained embedded system high recognition performance can be attained when the system is evaluated under controlled conditions. A number of publications note the values of key word recognition accuracies of around 92 percent to 97 percent with train and test on the Google Speech Commands data set with clean or slightly noisy speech, using STM32- or ESP32-level microcontrollers and minimal convolutional neural network configurations. [45, 57, 60, 61, 69]. These are normally obtained by short command utterances, fixed

vocabulary, preset decision thresholds, and do not necessarily translate to the far-field, multi-speaker, and highly reverberant smart-home setting. As a result, these precise numbers must be put into the framework of the data features, the acoustic environment, and testing procedures presented by the respective research.

related to the introduction of a powerful spoken-language understanding (SLU) facility. Instead of viewing speech recognition and semantic analysis as independent processes that rely on high-performance servers, new embedded SLU pipelines can implement these two steps on limited resource-based devices. Integration allows the offline systems to read and follow commands without network connection hence allowing robust, personal, and real-time control. Embedded SLU is increasingly indicative of a larger tendency in low-resource automatic speech recognition to provide reference to a larger user expectation of natural interface without compromising privacy [46]. With the increased availability of microcontrollers with special neural accelerators, and advances in model-reduction algorithms, it is incredibly easy to defuse simple KWS systems into complex ASR-SLU pipes based on embedded hardware.

7. Challenges and Limitations

Offline voice-controlled systems have structural issues that are based on a tradeoff between vocabulary size, sonic integrity, hardware features, privacy, and cross-operability. The main limitation is vocabulary limitation. Most offline systems allow only a limited number of defined commands due to the significant scaling increase of the models and processing time related to the expansion of the vocabulary. The commonest implementation uses small sets of keywords to fit within sub-megabyte memory limits to ensure real-time responsiveness; however, this limits in effect the user interaction and versatility of the system [47]. Increasing the vocabulary usually requires advanced compression techniques or edge-cloud architectures, which may compromise the system's offline capability [48].

The issue of noisiness is one of the critical challenges of offline voice-controlled systems. Whereas the current and/or filtering methods alongside denoising algorithms have enhanced functionality in controlled laboratory scenarios, high-fidelity in real-life households habitually corrodes collecting and processing with a factor of several, owing to simultaneous recurring multiple speakers, reverberation, and uncontrolled noisy distractions [49]. Adaptive filtering and lightweight convolutional denoisers among many are effective against isolated sources of interference but not very reliable in multifaceted acoustic environments. The ongoing discrepancy indicates that currently, the offline systems are yet to achieve such acoustic resiliency as that offered by larger ACR models based on the cloud, as they use large training corpora and neuroadaptive acoustic models [50].

All the above challenges are impacted by the hardware shortages and processing resource constraints given by most microcontroller boards. Many microcontrollers can support more profound and complicated model architectures. This forces the designers to use small models with heavily quantized counterparts that are bound to result in a balance between accuracy and efficiency [51]. This can be particularly emphasized in situations where the system is under stress to deal with longer commands or ambiguous speech since the deprived representational depth does not allow the processing chain to continue. In spite of that, although some gains have been realized, via pruning, operator fusion and neural processing unit assisted acceleration, the resultant performance disparities are measurable [52].

Although offline systems have some privacy benefits, there is no audio data outside the device limiting some of these benefits. Despite local data storage and management, transparency of managing firmwares, and protection against unauthorized access, end users are worried [53]. Therefore, it is inherent that data processing processes, including that information stored in a local environment must be encrypted and users must have access to metadata regarding voice samples that were previously recorded, become essential, which will lead to shaping user trust [54]. Also, there is the matter of interoperability which is another complication. Modern smart-home systems are decentralized, and lots of appliances are based on high-level proprietary protocols. In offline systems, custom configurations or middleware are required therefore, to support the communication between the platforms further exacerbating the implementation and restricting scalability [55]. However, in contrary to cloud-based systems which provide integrated APIs that help integrate into the system without any difficulties, offline systems do not provide unified interfaces, which makes integration processes more tedious [56].

8. Comparative Analysis: Offline vs Online

The voice-controlled systems (online and offline) are based on significantly opposite principles of design, have different trade-offs, and performance features. The recognition on the internet is based on large-scale acoustic and linguistic models which are managed on the cloud. It can provide high vocabulary coverage and high noise performance by leveraging huge datasets and utilizing remote server processing power [57]. However, it reduces unavoidable dependencies on reliable internet connectivity, creates latency because of routing data to the network and remote processing, and gives recurring concerns over privacy since the data of audio must be relayed outside of the contextual area and environment of the user [58]. Also, operational expenses are higher due to the upkeep of the infrastructure of servers which could inhibit accessibility by some users [59].

On the other hand, offline systems do not require a connection to any network to process all audio and to therefore avoid network latency along with providing the capability to act immediately without a network connection as well as allowing an uninterrupted operation even when there is no connection to the Internet [60]. This is also able to enhance privacy because the information of the user is not transferred to third parties [61]. There are serious limitations encountered in offline systems however: they have lower vocabularies and are built from small models that are not always noise robust like their cloud-based counterparts [62]. Additional factors that make it imperative to minimize the complexity of the model to achieve a result that can work within the tight memory and power constraints of embedded platforms and hence restrict flexibility and adaptability [63]. Accordingly, offline systems prove to be especially beneficial when quick reaction, resistance to appear disconnected, and authority over user data is of primary consideration, or online systems are more applicable to the parts where operating in broad lexical scope, the alternative interaction open-domain, or continuous renewal of content [64].

Table 1. A Comparative Analysis of Offline and Online Voice Control Systems.

Feature	Offline Voice Control Systems	Online Voice Control Systems
Latency	Low, commands processed locally	Higher, depends on internet connection and server response
Privacy	Strong, no data leaves the device	Weak, user data transmitted to cloud servers
Vocabulary Size	Limited, often restricted to predefined commands	Extensive, dynamic and updated continuously
Noise Robustness	Still challenging in many acoustic environments	More robust, benefits from large-scale cloud models
Hardware Needs	Low-power MCUs/NPUs, resource-constrained devices	Requires stable internet and cloud resources
Reliability	Works even without connectivity	Fails if internet connection is lost
Energy Use	Optimized for local low-power operation	Higher due to data transmission and server computation
Cost	One-time device cost, no subscription needed	Ongoing costs for cloud infrastructure and services

Table 1 presents complete information on the design priorities, privacy level, latency, the number of vocabularies, and noise resilience to explain why offline or online algorithm is better in specific scenarios. The new technological trends gradually bridge the disparage between the two paradigms. With the use of low-power microcontrollers that have a neural accelerator, offline systems can use more advanced models without significantly increasing latency and energy consumption. All these hardware extensions, combined with better model-compression techniques and the combination of spokes speech language understanding (SLU) pipelines, indicate that the advantages of online systems obtained by existing means can decrease over time [65]. The choice between online and offline architecture will depend more on application-specific consideration than the difference of technical competence as this convergence proceeds further.

9. Case Studies and Implementations

Functional examples of application of offline voice control demonstrate how frameworks and technologies are implemented into practice. A lot of smart-home prototypes have already proved the features of providing effective voice recognition on embedded devices with no involvement in cloud services. Such implementations may also include low-power microcontrollers, radio responsive audio front ends, and thin recognition models to enable and support real time wake-word hardware residency and command execution [66].

Commonly, dedicated voice modules, such as the Elechouse VR v3, are integrated with microcontrollers running preprogrammed commands, such as the Arduino or the ESP32. These systems are also commonly applied in the field of access-control and environmental-automation applications such as smart locks, lighting control, and heat-ventilators where low latency and privacy is of primary importance [67]. However, the narrow vocabularies and preset sets of commands make them work great in structured interactions but less adaptive to the dynamic environment [68].

More complex versions make use of models like TensorFlow Micro and Edge Impulse, which enables the usage of small neural networks on gadgets offering constrained neural enablers. Such systems are capable of supporting an extended vocabulary and could be more accurate in recognition with respect to simple voice modules, still viable in entirely offline situations. They are considered as one of the most effective tools under security and assistance that might not consistently be in touch with each other and require a consistent real-time reaction [69]. The main disadvantage is that such systems are currently performed manually to maintain performance in the divergent setting of acoustic environment and heterogeneous hardware configurations [70].

The two key implementation strategies are therefore illustrated in the case studies. The former is more focused, simplified and with less development requirements, using existing readymade modules which offer a common command set. The second approach is more flexible but still technically challenging to implement employing embedded machine-learning structures to obtain excellent performance. Both arguments emphasize the importance of balancing the application need set with the computational and development resources available, and it shows that offline systems are being worked into a growing range of increasingly diverse uses in smart-home applications- both

simple control functions and intelligent conversation.

10. Performance Evaluation

Testing offline voice-controlled systems requires stringent testing on different performance aspects such as accuracy of recognition, latency, energy consumption and capability to perform against diverse acoustical environments. In contrast to cloud-based systems where the quality of models and network conditions are of primary importance in the performance analysis, offline systems have to find a perfect balance between these parameters under strict resource capability requirements. Accurate analysis is, therefore, required to achieve the appropriate choice of models and platforms that will be adapted to a given application of smart home [71].

Typically, recognition accuracy is measured by the word-error rate and command-recognition accuracy metrics using standardized data sets, such as Google Speech Commands or domain-specific corpora. Lightweight models that are implemented on microcontrollers may perform similarly (90 to 96 percent in quiet speakers in clean-speech environments); however, they fail to perform as well in noisy and reverberating conditions. This observation illuminates the limitations on capacity of the model used and insufficiency of training diversity [72]. Latency is also one of the crucial variables because the processing of commands must be in real-time so that the user experience can be enhanced. By pairing together an optimized version of the keyword spotting models with efficacious inference engines, it is possible to obtain response times of less than 200 milliseconds on platforms such as the ESP32 or STM32, which serves as evidence that low-latency operation can be achieved even with the most tightly strained resource limits [73].

Energy consumption is evaluated in relation to the amount of power that is needed in continuous listening and inference. This is because implementation of event-based activation and activation of on-chip preprocessing significantly consumes less energy, which makes devices have longer durations of time when they are powered by batteries. Indicatively, systems that couple low-power microphones to optimised inference pipelines have shown power consumption of less than 50-milliwatts in active recognition, which makes them useful as always-on systems. However, both energy requirements and power consumption increase dramatically when bigger models, or more feedstock controlled by continuous ASR, are enabled, especially with general-purpose microcontrollers [74].

The strength is tested through assessing performance in a variety of acoustic conditions, such as, multiple-speaker setting, silent background, and mobile microphone-speaker distance and tasks. Such circumstances often lead to performance variability of offline systems due to the insufficient size of the current model of acoustic features and the absence of noise-adaptation strategies of huge scale that cloud-based systems rely on [75].

Table 2 gives an overview of standards in the literature of various offline voice-control designs using diverse framework and hardware platforms.

Table 2. Typical Performance Metrics for Offline Voice Control Systems.

Platform / Framework	Accuracy (Clean)	Accuracy (Noisy)	Latency (ms)	Power (mW)	Vocabulary Size	Reference
Elechouse VR V3 + Arduino	90–92%	70–80%	100–150	<30	≤80 commands	48
ESP32 + TFLM	93–96%	78–85%	150–200	40–50	100–200 words	61, 69
STM32 + Edge Impulse	91–95%	75–83%	180–250	35–45	100–150 words	69, 70
Cloud Baseline (Online)	97–99%	95–98%	300–500+	N/A	Extensive	75

Note: Reported metrics are obtained out of the native studies and cannot be directly compared because of the differences in datasets and evaluation protocols. Values are also typical ranges obtained out of literature and can change based on the choice of a dataset, acoustic environment, the size of the vocabulary, and hardware settings. Public datasets that include Google Speech Commands under controlled settings will tend to give clean accuracy values whereas noisy accuracy represents the phenomenon of performance under noise-enhanced or real-world conditions that is reported in the referenced studies.

Such findings suggest that, despite the general low noise sensitivity and smaller vocabulary selection on offline voice control systems as opposed to cloud-based ones, their latency and energy consumption levels are very competitive. This is why offline methods should be considered especially appropriate to the smart home setting where providing privacy and continuous service is essential. The given variation in the performance of the different platforms and architectures further indicates that the computational architecture, model optimization techniques, and processing constraints are more determinants of the system performance than speech recognition model.

11. Future Directions

A coordinated combination of building technologies, computer acceleration, and intelligent education techniques is functioning towards the predicted future development of the offline voice-control frameworks. Wake detection, which directly affects responsiveness and usage power, has stayed as one of the greatest issues. New innovations have shown consistent recognition with few training samples up to 25 in each category, and eliminating the need for cloud-based retraining in the cloud [76]. Such techniques as transfer learning with small-sized models, and aggressive

quantization, helpful in this regard make the difference between the memory and base requirements and meet the accuracy criteria. As a result, extensive customization is fast and attended by these tools to suit a wide range of acoustic conditions, user groups or languages, which individualizes offline systems to be much more feasible at scale [77].

Hardware trends are also very relevant. Recent low-power microcontrollers featuring Neural Processing Unit (NPU) can be used to directly execute more advanced models in embedded space. Unlike digital signal distinguished DSP based algorithms, NPUs can operate the quantized neural-network operations on fixed volumes of data concurrently, allowing complex keyword -locating maps to operate under stringent performance requirements that include both latency and power [78]. This is made to give offline systems the ability to expand their vocabularies, to include more wake-word acts, as well as to defeat simple multilingual support without filtering cloud infrastructure [79]. All these enhancements significantly improve the performance difference compared to online systems making offline solutions a more appropriate solution in real life situations requiring both stability and freedom.

The field of embedded learning is one of focus. These systems monitor the voice, accent, and other changes of a certain person without sending any audio information to other servers, which contributes to increased privacy and scalability over time in the face of new circumstances of background noise and speaking styles [80]. However, feats like careful management of memory bandwidth, computational demands and stability are critical to on-device learning to alleviate difficulties like catastrophic forgetting. This is still a dynamic research area in which model architecture, learning algorithms on an incremental basis, and hardware potentials need to be balanced together.

It is based on these developments that the future can be suggested to be one where the offline systems not just) simplify their online analogs. Instead, they will become purpose designed intelligent agents that recreatively change user privacy, as well as being reliably operated even on resource limited environments. The intersection between effective model structures, dedicated hardware and custom enabling features will be the keystone to the realization of this vision and will collectively become the next generation of offline voice-control systems.

Conclusion

Smart-home systems have made offline voice control an important part because of its benefits of having high privacy, low latency and does not require network connectivity. In this review, the fundamental elements, architectures, and design of constraints of an offline voice-controlled smart home have been considered, focusing on the aspects of accuracy, latency, energy use, and noise resistance. The paper brings together the conclusion of synthesizing the results across models, hardware platforms and deployment strategies to point out how the performance of system is dominated by architecture decisions, and optimization methods rather than the complexity of the model selections. Analysis The analysis gives practical recommendations on how efficiency and privacy-conservative voice interfaces can be designed on constrained edge devices.

Conflict of interest

None.

References

[1] R. Martinek, J. Vanus, J. Nedoma, M. Fridrich, J. Frnda, and A. Kawala-Sterniuk, Voice communication in noisy environments in a smart house using hybrid LMS+ICA algorithm. Sensors (Switzerland), vol. 20, no. 21, 2020.

[2] A. M. Rostami, A. Karimi, and M. A. Akhaee, Keyword spotting in continuous speech using convolutional neural network. Speech Commun., vol. 142, 2022.

[3] S. Yang, B. Kim, I. Chung, and S. Chang, Personalized keyword spotting through multi-task learning. in Proc. Annual Conf. Int. Speech Commun. Assoc. (Interspeech), 2022.

[4] S. Drgaś, "A Survey on Low-Latency DNN-Based Speech Enhancement," Sensors, vol. 23, no. 3, p. 1380, 2023.

[5] J. Bushur and C. Chen, Neural network exploration for keyword spotting on edge devices. Future Internet, vol. 15, no. 6, 2023.

[6] Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki, and A. S. Hafid, A comprehensive survey on TinyML. IEEE Access, vol. 11, 2023.

[7] D. Bermuth, A. Poeppel, and W. Reif, Jaco: an offline running privacy-aware voice assistant. arXiv preprint arXiv:2209.07775, 2022.

[8] C. Oumard, J. Kreimeier, and T. Götzelmann, Pardon? an overview of the current state and requirements of voice user interfaces for blind and visually impaired users. in Lecture Notes in Computer Science, 2022.

[9] J. Mishra, T. Malche, and A. Hirawat, Embedded intelligence for smart home using TinyML approach to keyword spotting. Engineering Proceedings, vol. 82, no. 1, p. 30, 2024.

[10] C. Gao, Y. Gu, F. Caliva, and Y. Liu, Self-supervised speech representation learning for keyword spotting with light-weight transformers, arXiv:2303.04255, 2023.

[11] V. Rajapakse, I. Karunanayake, and N. Ahmed, Intelligence at the extreme edge: a survey on reformable

TinyML. ACM Comput. Surv., vol. 55, no. 13, 2023.

[12] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[14] A. Berg, M. O'Connor, and M. Tairum Cruz, Keyword Transformer: A self-attention model for keyword spotting, in Proc. INTERSPEECH 2021, pp. 4249–4253, 2021.

[15] N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, Voice recognition and voice comparison using machine learning techniques: a survey. In 2020 6th Int. Conf. on Advanced Computing and Communication Systems (ICACCS), 2020.

[16] A. S. Dhanjal and W. Singh, A comprehensive survey on automatic speech recognition using neural networks. Multimed. Tools Appl., vol. 83, no. 8, 2024.

[17] Irugalbandara C., Naseem A. S., Perera S., Kiruthikan S., and Logeeshan V., A secure and smart home automation system with speech recognition and power measurement capabilities, Sensors, vol. 23, no. 13, article 6109, 2023.

[18] M. Nalini, S. Suveka, and S. A. C. Bukhari, AI-based fingerprint and voice recognition systems, in AI based advancements in biometrics and its applications, CRC Press, 2024, pp. 101–117.

[19] S. Heydari and Q. H. Mahmoud, Tiny machine learning and on-device inference: a survey of applications, challenges, and future directions, Sensors, vol. 25, no. 10, p. 3191, 2025.

[20] H. Han and J. Siebert, TinyML: A systematic review and synthesis of existing research, in 4th Int. Conf. on Artificial Intelligence in Information and Communication (ICAIIC), 2022.

[21] Z. Xie., The BIGAI Offline Speech Translation Systems for the IWSLT 2023 Evaluation, in Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT), Association for Computational Linguistics, pp. 243–248, 2023.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, and Y. Wu, Conformer: Convolution-augmented transformer for speech recognition, Proc. Interspeech, 2020.

[23] L. Hernández Acosta and D. Reinhardt, A survey on privacy issues and solutions for voice-controlled digital assistants, Pers. Media Comput. J., 2022.

[24] S. Liao, C. Wilson, L. Cheng, H. Hu, and H. Deng, Measuring the effectiveness of privacy policies for voice assistant applications, in ACM Int. Conf. Proceeding Series, 2020.

[25] M. Imam and G. Gupta, Precision location keyword detection using offline speech recognition technique, J. Internet Technol., vol. 23, no. 2, pp. 125–138, 2023.

[26] C. Cioflan, L. Cavigelli, M. Rusci, M. de Prado, and L. Benini, "On-Device Domain Learning for Keyword spotting on Low-Power Extreme Edge Embedded Systems," in Proc. IEEE 6th Int. Conf. Artificial Intelligence Circuits and Systems (AICAS), 2024.

[27] R. Aloufi, H. Haddadi, and D. Boyle, On-device voice authentication with paralinguistic privacy, arXiv:2205.14026, 2022.

[28] H. M. S. Di Leo, L. De Cicco, and S. Mascolo, Real-time speech-to-text on edge: a prototype system for ultra-low latency communication with AI-powered NLP, Information, vol. 16, no. 8, p. 685, 2025.

[29] R. Kruger and B. Klug, Voice assistant technology: Alexa® in the sim lab, Can. J. Crit. Care Nurs., vol. 29, no. 2, 2018.

[30] Akhtar Z., Khursheed M. O., Du D., Liu Y., Small-footprint slimmable networks for keyword spotting, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2023.

[31] C. Banbury, A. Reddi, M. Lam, W. Fu, S. Han, and V. Chandra, Micronets: Neural network architectures for deploying TinyML applications, in Proc. Machine Learning and Systems (MLSys), 2021, pp. 1–15.

[32] Z. Yang, S. Sun, J. Li, X. Zhang, X. Wang, L. Ma, and L. Xie, CaTT-KWS: A Multi-stage customized Keyword spotting Framework based on Cascaded Transducer-Transformer, in Proc. Interspeech 2022, pp. 4245–4249, 2022.

[33] C. Banbury, V. J. Reddi, P. Torelli, J. Holleman, N. Jeffries, C. Kirsch, and V. Sze, MLPerf Tiny: Benchmarking TinyML Systems, in Proc. 35th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2021.

[34] T. Malche, A. Hirawat, and J. Mishra, Voice-activated home automation system for IoT edge devices using TinyML, Discover Internet of Things, 2025.

[35] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, Speech recognition using deep neural networks: a systematic review, IEEE Access, vol. 7, 2019.

[36] Pietro Bartoli, Tommaso Bondini, Christian Veronesi, Andrea Giudici & Franco Zappa, end-to-end Efficiency in Keyword spotting: A System-Level Approach for Embedded Microcontrollers, arXiv:2509.07051, 2025

[37] A. Kintz, A. G. Howard, M. Sandler, and A. Zhmoginov, EdgeSpeechNets: Highly efficient deep neural networks for speech recognition on the edge, arXiv preprint arXiv:1810.08559, 2018.

[38] N. Hoy, Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, Medical Reference Services Quarterly, vol. 37, no. 1, pp. 81–88, 2018.

[39] A. Pandey and D. Wang, A new framework for supervised speech enhancement in the time domain, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1179–1188, 2019.

[40] J. Wang and S. Li, Keyword spotting system and evaluation of pruning and quantization methods on low-power edge microcontrollers, arXiv:2208.02765, 2022.

[41] S. Kim, T. Hori, and S. Watanabe, Joint CTC–attention based end-to-end speech recognition: Advances and trends, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2742–2756, 2022.

[42] Y. Ma, Y. Zhang, M. Bachinski, and M. Fjeld, Emotion-aware voice assistants: design, implementation, and preliminary insights, in proceedings of the 11th International Symposium on Chinese CHI (Chinese CHI), pp. 527–532, 2023.

[43] Y. Iliev and G. Ilieva, A Framework for Smart Home System with Voice Control Using NLP Methods, Electronics, vol. 12, no. 1, article 116, 2023.

[44] J. Wang, S. Kim, and M. Sunwoo, Hardware-efficient Customized Keyword spotting with Spectral-Temporal Graph Attentive Pooling, arXiv preprint arXiv:2409.00099, 2024.

[45] C. Zonios and V. Tenentes, Energy efficient speech command recognition for private smart home iot applications, international conference on smart internet of Things, 2023.

[46] I. López-Espejo, Z. Tan, and J. Jensen, Deep spoken keyword spotting: An overview, IEEE Access, vol. 10, pp. 4169–4199, 2022.

[47] Md N. Miah, Voice command recognition with deep neural network on edge devices, M.S. thesis, Dept. Electrical & Computer Engineering, Purdue University, 2021.

[48] M. A. Torad, B. Bouallegue, and A. M. Ahmed, A voice controlled smart home automation system using artificial intelligent and internet of things, TELKOMNIKA (Telecommunication, Computing, Electronics and Control), vol. 20, no. 4, pp. 808–816, 2022.

[49] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in Proc. NeurIPS, 2020.

[50] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, The Microsoft 2017 conversational speech recognition system, in Proc. IEEE ICASSP, 2018, pp. 5934–5938.

[51] G. Menghani, Efficient Deep Learning: A survey on making deep learning models smaller, faster, and better, acm comput. surv., vol. 55, no. 12, 2023.

[52] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, efficient processing of deep neural networks: a tutorial and survey, proceedings of the IEEE, vol. 105, no. 12, pp. 2295–2329, 2017.

[53] J. Lau, B. Zimmerman, and F. Schaub, Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers, in Proc. ACM CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13.

[54] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, Speech Technology for Healthcare: Opportunities, challenges, and future directions, IEEE reviews in Biomedical Engineering, vol. 14, pp. 342–356, 2021.

[55] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, Security, privacy and interoperability in the Internet of Things: challenges and opportunities, Computer Networks, vol. 76, pp. 146–164, 2015.

[56] P. Aimtongkul and K. Janchitrapongvej, Development and Assessment of Internet of Things-Driven smart home security and automation with voice commands, sensors, vol. 24, no. 3, p. 896, 2024.

[57] S. Majumdar and B. Ginsburg, MatchboxNet: 1D time-channel separable convolutions for small-footprint keyword spotting, in Proc. INTERSPEECH, 2020, pp. 1977–1981.

[58] P. Drahoš, Edge container for speech recognition, Electronics, vol. 10, no. 19, p. 2420, 2021.

[59] H. Kim and J. S. Han, Smart home advancements for health care and beyond: systematic review of two decades of user-centric innovation, *Sensors*, vol. 24, no. 11, p. 3317, 2024.

[60] P. Warden, Speech commands: a dataset for limited-vocabulary speech recognition, *arXiv preprint arXiv:1804.03209*, 2018.

[61] R. David et al., TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems, in Proc. Machine Learning and Systems (MLSys), 2021.

[62] A. Pervaiz, J. M. Rabaey, and M. Tariq, Incorporating noise robustness in speech command recognition by noise augmentation of training data, sensors, vol. 20, no. 8, 2020.

[63] A. Howard et al., Searching for MobileNetV3, in Proc. IEEE/CVF international conference on computer vision (ICCV), 2019, pp. 1314–1324.

[64] J. Lin, W.-M. Chen, Y. Lin, and C. Gan, MCUNet: Tiny deep learning on IoT devices, in Proc. NeurIPS, 2020.

[65] K. Ding, M. Zong, J. Li, and B. Li, LETR: a lightweight and efficient transformer for keyword spotting, in Proc.

IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2022.

[66] G. Cámbara, J. Luque, and M. Farrús, recycle your Wav2Vec2 codebook: a speech perceiver for keyword spotting, in Proc. 29th Int. Conf. Comput. Linguistics (COLING), 2022, pp. 7166–7170.

[67] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, end-to-end multi-look keyword spotting, in Proc. Interspeech, 2020.

[68] A. Barovic and A. Moin, TinyML for speech recognition, arXiv preprint, Apr. 2025.

[69] J. Lin, W.-M. Chen, J. Gan, S. Han, and Y. Lin, MCUNetV2: Memory-Efficient Patch-based Inference for Tiny Deep Learning, in Proc. 35th Conference on Neural Information Processing Systems (NeurIPS), 2021, pp. 2874–2887.

[70] Edge Impulse, Edge Impulse Documentation, 2024. [Online]. Available: https://docs.edgeimpulse.com. [Accessed: Jan. 15, 2026].

[71] V. J. Reddi, B. Griffith, P. Warden, A. Faust, and G. Janapa Reddi, Widening access to applied machine learning with TinyML, Communications of the ACM, vol. 65, no. 4, pp. 34–40, 2022.

[72] A. L. Georgescu, A. Pappalardo, H. Cucu, and M. Blott, Performance vs. hardware requirements in state-of-the-art automatic speech recognition, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, article 24, 2021.

[73] Yi Luo and Nima Mesgarani, Conv-TasNet: Surpassing ideal time-frequency masking for real-time end-to-end monaural speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, 2019.

[74] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, Direct modelling of speech representations for context-aware emotion recognition, IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1184–1197, 2023.

[75] Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Wayne Xiong, and Zhong Meng, Recent advances in end-to-end automatic speech recognition, Nanoscale Research Letters, vol. 15, no. 1, article 5, 2021.

[76] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers, in Proc. INTERSPEECH, 2023.

[77] M. Suresh, M. S. Roopa, and K. G. Srinivasa, IoT-based smart security and home automation system, in Intelligent Technologies for Sensors: Applications, Design, and Optimization of a Smart World, 2023.

[78] T. Higuchi, A. Gupta, and C. Dhir, Multi-task learning with cross attention for keyword spotting, in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021.

[79] S. Ghangam, D. Whitenack, and J. Nemecek, Dyn-ASR: compact, multilingual speech recognition via spoken language and accent identification, arXiv preprint arXiv:2108.02034, 2021.

[80] A. Diwan, C.-F. Yeh, W.-N. Hsu, P. Tomasello, E. Choi, D. Harwath, and A. Mohamed, Continual learning for on-device speech recognition using disentangled conformers, in proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.