International Journal of Development Mathematics Vol 3 Issue 3 Page No. 506 - 524

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

A Context-Aware Voice Recognition Model for Multilingual Retail Settings: A Conceptual Framework with Yakubu Shopping Mall as an Illustrative Case

Wasinda J. Malgwi[a], Adamu Ilyasu[b], Abba, J. Muhammad[b,*] and Yahya A. Suleiman[b]

[a]Department of Information and Communication, University of Abuja, Abuja, Nigeria.

[b]Department of Information Technology, Modibbo Adama University Yola, Adamawa State, Nigeria.

ARTICLE INFO

Article history:
Received 10 April 2026
Received in revised form 20 August 2026
Accepted 27 August 2026

Keywords:
Voice Recognition, Inventory Management, Context-Aware Computing, Multilingual Systems, Code-Switching

MSC 2020 Subject classification:
97M80

ABSTRACT

Inventory management is critical to the daily operations of shopping centers, but it is often plagued by errors and inefficiencies in manual data entry. Voice recognition technology provides an alternative to manual entry, but off-the-shelf models are not usable in real retail environments due to three interrelated challenges: high ambient noise levels, the need to understand domain-specific retail terminology, and the need to process multilingual speech with code-switching. In this paper we review and synthesize concepts from the literature in automatic speech recognition, context-aware computing and multilingual systems, and propose an integrated framework for voice recognition in multilingual retail settings. The proposed framework consists of four layers, i.e., an input layer with noise-cancelling hardware considerations, a speech-to-text engine with domain adaptation, a logic layer with context-awareness that comprises task-specific language models and an application layer for intent parsing and database integration. Using the illustrative case study of Yakubu Shopping Mall in Northern Nigeria, the framework demonstrates how environmental context (ambient noise), linguistic context (English-Hausa code-switching) and task context (inventory vocabulary) can be integrated into a cohesive system. The framework gives testable propositions for future empirical work, e.g., expected word error rate reductions of 70-80% in noisy conditions vs. generic models. The framework provides practical guidance for implementing voice directed inventory systems in multilingual retail settings. It provides a basis for empirical validation for researchers in developing-economy retail settings.

1 Introduction

Technology is disrupting the retail industry in a major way. Data and automation are now vital to companies' effectiveness and competitiveness. A key part of this transformation is the management of inventory which comprises stock management, order management, sales management and delivery management. In retail stores such as shopping centers and distribution centers, inventory management has a direct impact on customer satisfaction, available funds, and profit (Smith & Offodile, 2008; Gafke & Winkelmann, 2017). Historically, places like Yakubu shopping mall use manual processes for keeping track of inventory like pen and paper or handheld barcode scanners. Barcode scanners are a progressing on paper-based structures but still have a number of restrictions. The procedure is sluggish, requiring workers to halt at work to scan stuffs, which distracts their devotion from customers. It is also

* Corresponding author. Tel.: +2347044999387
E-mail address: xxxxx@gsu.edu.ng (Malgwi W. J.)
https://doi.org/10.62054/ijdm/0303.27

substantially challenging, and predisposed to errors like scanning the incorrect barcode, or missing items overall. These mistakes composite over time, leading to erroneous stock levels and poor business outcomes.

Another is voice recognition technology. "Pick-by-voice" or "stock-by-voice" systems allow employees to perform inventory operations without using hands or eyes. With a headset, workers can talk their way through receiving instructions, confirming selections of items, updating stock levels and querying inventory databases. This means workers can keep their eyes on their environment and hands free to manipulate objects which could lead to better safety, accuracy and productivity (Voice, 2015; Smith & Offodile, 2008).

Generic voice recognition systems, however, trained on standard speech datasets (e.g., audiobooks, news broadcasts, web searches) do not perform well in retail environments for three fundamental reasons that are well documented in the literature.

i. Acoustic Noise. Retail environments are noisy by nature. research has reported that shopping malls and retail warehouses have noise levels of ambient noise fluctuating between 65-85 dB, with the sources of noise being refrigeration units, customer tête-à-têtes, public announcements, metal trolleys, pallet jacks and forklifts (Chen et al., 2023; Adeleke & Ogunleye, 2024). As found by Li et al. (2022), contextual noise can increase automatic speech recognition (ASR) word error rates by as much as 300% compared to quiet circumstances

ii. Code-switching and linguistic diversity. Retail employees in multilingual settings like Northern Nigeria might speak several dialects such as Hausa, Yoruba, Igbo, English. Code-switching (the use of more than one language in a discussion or exclamation) is a common feature of workplace conversation. From Northern part of Nigerian merchandising backgrounds, the most common code-switching is about 78% of work environment discussions (Musa & Ibrahim, 2024). But then again most commercial ASR systems are trained on monolingual corpora and are not designed for code-switched speaking, foremost to a substantial performance degradation (Vu et al., 2012; Okonkwo, 2023).

iii. Domain-Specific Vocabulary. Despite the fact handling record, you need to identify nearly some terms like merchandise codes, brand names (Dangote Cement, Indomie Noodles, Mai Shayi cooking oil), and stock-keeping component identifiers. It has been identified that in the absence of province adaptation, generic ASR models do not recognize 34-40% of industry-specific terms (Kumar & Singh, 2025; Li et al., 2022).

In this paper we address the question of how to build a voice recognition model that can perform well in a specialized, multilingual retail environment. We argue that a context-aware approach that integrates environmental, linguistic and task-specific contexts into system design could greatly enhance performance. The paper is a conceptual review where the existing literature was synthesized to develop an integrated framework. Yakubu Shopping Mall was used as a case study to illustrate the framework in a retail context.

2 Literature Review

2.1 Voice Recognition in Inventory Management

Voice technology applied to inventory and warehousing tasks isn't new. During the 1980s and 1990s, research examined "voice directed warehousing" systems and proved that speech technology could increase order picking accuracy and improve productivity. Smith and Offodile (2008) performed a detailed analysis and found that voice order picking could increase productivity by 15-35% and reduce error rates to near-zero levels as compared to paper-based methods. The primary way that this is accomplished is that voice technology allows for hands-free, eyes-free operation, reducing cognitive load and time for physical movement.

Early voice-directed systems were hardware dependent and speaker dependent, users had to enroll by speaking certain words. This process was time consuming and complicated the use of systems with temporary or changing

staff. The majority of today's systems have moved to speaker-independent models that utilize deep learning and cloud computing. However, Gafke and Winkelmann (2017) pointed out that even modern systems are still rather generic systems trained mainly on standard speech corpora, not on the specific acoustic and linguistic conditions of retail warehouses.

Recent works have begun to fill in these gaps. Chen et al. (2023) developed a noise-robust ASR system for Chinese warehouses, which accomplished a word error rate of 5.2% in the typical operating conditions. Kumar and Singh (2025) projected a transfer learning approach for multilingual ASR in Indian retail atmospheres and demonstrated its feasibility in four languages. Up till now, neither of the studies fully considered ambient noise, code-switching and retail-specific terminology in one context.

2.2 Noise Robustness in Automatic Speech Recognition

It is a well-known fact that the performance of ASR is degraded in noisy environments. Noise masks speech signal, limiting the ability of acoustic models to map audio onto phonetic units. Research in noise-robust ASR has mainly been in two directions: signal processing techniques that pre-process the audio to clean up noise before recognition and model adaptation techniques that adapt the models by training on noise augmented data (Li et al., 2014; Brandstein & Ward, 2001).

Recent developments have been in deep learning approaches for noise tolerance. Li et al (2022) reviewed neural network architectures for noise-robust ASR and determined that multi-condition training (i.e. training on clean and noisy speech) is still the most effective real-world method. Data augmentation techniques also play a crucial role in noise robustness. Park et al. (2022) introduced SpecAugment, a simple but effective augmentation method that has become standard in ASR training pipelines, including for noise-robust applications. Chen et al (2023) extended this by incorporating environment-specific noise profile in training, leading to large gains over generic noise augmentation.

Adeleke and Ogunleye (2024) specifically characterized the acoustic conditions in Nigerian shopping malls as it relate to retail settings, where noise levels ranged from 65 dB during quiet periods to 85 dB during peak hours. They found that successful ASR in this environment would require both close-talking microphones with noise cancellation and acoustic models trained on noise side view specific to the mall. Rao et al. (2023) correspondingly found that multi-channel speech enhancement can improve ASR accuracy in industrial environments by 40-60%.

2.3 Multilingualism and Code-Switching in ASR

Commercial ASR systems are usually designed to recognize only one language at a time and cannot process code-switching, a common phenomenon in many parts of the world (Auer, 2013). The availability of large-scale multilingual speech datasets has accelerated research in multilingual ASR. Pratap et al. (2024) introduced the MLS corpus, a large-scale multilingual dataset for speech research. However, such dataset often lack representation of African languages and code-switched varieties. Code-switching presents difficulties for ASR at different levels: acoustic-phonetic (phones from distinct languages), lexical (vocabulary from different languages), and syntactic (grammatical structures from different languages).

Vu et al (2012) initiated the research on code-switching ASR by creating a system for Mandarin-Hawaiian code-switching, proving the possibility of such systems, but with much higher word error rates than monolingual systems. More studies have been done on multilingual acoustic models, interpolating language models, and end-to-end models trained on code-switched data. Winata et al. (2021) proposed a language-agnostic multilingual modelling approach and achieved significant improvements on code-switching benchmarks. Zhang et al. (2023) proposed a double-decoder transformer architecture clearly designed for code-switching ASR, reporting relative error reductions of 35-45%. Di Gangi et al. (2022) showed that adaptive speech translation models can be effectively applied to under-

resourced languages, suggesting similar approaches could benefit English-Hausa ASR. More recently, Li et al. (2025) introduced prompt-based learning for low-resource code-switching ASR, achieving significant improvements with limited training data. This approach could be particularly relevant for English-Hausa applications where large labelled datasets are scarce. Okonkwo (2023) conducted a study on a conceptual analysis of ASR requirements for Nigerian English, and identified the need for localization. Musa and Ibrahim (2024) documented English-Hausa code-switching patterns in Northern Nigerian work environment. They found out that code-switching occurs in 78% of conversations and it adheres to predictable patterns (mostly English base language with Hausa insertions for nouns and cultural references). However, no study has proposed or developed an ASR context for English-Hausa code-switching in the case of retail record. A relevant example was set by Hamed et al (2024) who created an Arabic-English code-switching corpus, demonstrating that code-switched speech data can be collected and used.

2.4 Context-Aware Computing

The theoretical foundation of this study is the notion of context-aware computing, as described by Dey (2001). Dey defined context as "any information that can be used to characterize the situation of an entity" and a system as context aware if it "uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task". Dey's framework proposes several levels of relevant context for a voice recognition system in inventory management:

i. Environmental context: The system is designed for use in a noisy retail environment rather than a quiet office.

ii. Linguistic context: Users may be bi-lingual and may switch between English and Hausa.

iii. Task context: User is in a specific location (e.g. aisle 3), performing a specific inventory task (e.g. receiving stock, counting items) with a specific set of relevant items

Recent work extending context-aware computing to voice interfaces has emphasized the importance of task-specific language models. Chen et al. (2024) proposed a context-aware voice assistant for warehouse management which with dynamically constrains the ASR terminologies depending on the user position and task, and achieved significant reductions in out-of-vocabulary errors. Park et al. (2023) proposed context-aware language models for domain-specific ASR. They presented that task-adaptive biasing can achieve 30-50% decrease in word mistake rates in specialized area. Wang et al. (2024) extended this work to industrial voice interfaces, demonstrating the feasibility and usefulness of dynamic language model biasing from real-time context. Choi and Lee (2025) further demonstrated that on-device contextual biasing can achieve comparable performance with lower latency, making deployment more practical for retail environments. Beyond technical performance, user acceptance is equally critical for successful deployment. Wasinda et al. (2026) developed an extended Technology Acceptance Model (TAM) for voice-enabled inventory systems in Nigeria's retail sector, identifying perceived usefulness and ease of use as key determinants of adoption. Their findings suggest that addressing user perceptions is essential alongside technical improvements.

2.5 Comparative Analysis of Existing Approaches

Table 1: summarizes the key findings and limitations of the existing literature on voice recognition in retail and warehouse environments that motivate the proposed framework.

Table 1: Comparison of Existing Approaches for Voice Recognition

Author(s) & Year	Country	Area of focus	Methodology Type	Key Finding	Limitation Identified
Smith & Offodile (2008)	USA	Voice picking in warehousing	Empirical review	15-35% productivity	No multi-lingual focus

				increase over paper	
Dey (2001)	USA	Context aware computing	Conceptual	Defined environment for ubiquitous computing	Not applicable to ASR
Vu et al. (2012)	Multinational	code-switching ASR (Mandarin-Hawaiian)	Empirical	Feasible but higher WER than monolingual	other language pair, not retail-tested
Li et al., (2014)	Multinational	Noise-robust ASR	Technical review	best practices for multi-condition training	Pre-deep learning era
Gafke & Winkelmann (2017)	Germany	Voice in warehousing	Critical review	System adaptability	No African context
Winata et al. (2021)	Multinational	Language-independent multilingual ASR	Empirical	More effective code-switching	Not tested in store
Li et al. (2022)	Multinational	Noise-robust in ASR	Comprehensive review	Neural networks beat traditional methods	Technical, not pragmatic
Chen, et al. (2023)	China	Noise-robust ASR for warehouses	Empirical	Reached 5.2 per cent Noise WER in warehouse noise	Only one language
Okonkwo (2023)	Nigeria	ASR for Nigerian English	Conceptual	Identified need for localization	No proposed framework
Park et al (2023)	Korea	Context-aware Language Models	Empirical	30-50% WER reduction on specialized domains	Not retail-specific
Chen, et al. (2024)	China	Context-aware voice assistants	Empirical	Error reduction with dynamic vocabulary	One language

				constraints	
Kumar & Singh (2025)	India	Multilingual ASR in the retail domain	Empirical	Feasible across 4 languages	Not tested in noise
Wasinda et al. (2026)	Nigeria	User acceptance of voice-enabled inventory systems	Empirical (TAM model)	Perceived usefulness, ease of use, and cultural factors significantly influence adoption intention	Did not address technical ASR performance
This Study in	Nigeria	Context-aware multilingual ASR	Conceptual	review Points to integrated framework.	Needs empirical validation

2.6 Research Gap

A systematic review of the literature to date reveals a clear and significant gap. Much work has been done on (a) noise-robust ASR, (b) multilingual and code-switched speech recognition, and (c) context-aware computing, but these three streams have largely evolved in isolation. As far as we know, none of the existing studies has integrated these domains into a conceptual framework for multilingual retail inventory management.

Also, most ASR research has been conducted in laboratory or warehouse settings in developed economies (USA, Germany, China) with little attention paid to peculiarities of retail spaces in multilingual developing economies such as Nigeria. The combination of high ambient noise (65-85 dB), English-Hausa code-switching and retail-specific vocabulary raises a set of challenges that existing frameworks do not address collectively.

This research addresses this gap by proposing a context-aware voice recognition framework that synthesizes existing knowledge from all three domains and applies it to the specific context of a Nigerian shopping mall. The background is conceptual and intended to guide future empirical validation.

2.7 Theoretical Framework

This study combines three theoretical perspectives. First, the theory of context-aware computing by Dey (2001) offers the overarching framework emphasizing that the system design should include information about the environment, linguistic situation, and task of the user. Dey's framework has been widely used across domains, but its application for ASR for the multilingual retail inventory is novel. The theory posits that the relevance and usefulness of a system increase proportional to its ability to sense and respond to contextual variables. Second, transfer learning theory (Pan & Yang, 2010; Kumar & Singh, 2025) guides the approach to adaptation of generic ASR models to domain-specific conditions with limited in-domain data. Transfer of learning is predominantly applicable in code-switching and low-resource circumstances where large labelled datasets are not available. The theory suggests that knowledge acquired from data-rich source domains can be successfully transferred to data-poor target domains using proper model architectures and fine-tuning approaches. Third, sociolinguistic theories of code-switching (Auer,

2013; Musa & Ibrahim, 2024) provide a foundation for handling mixed-language speech. These theories identify dependable patterns in code-switching behavior (e.g., inserting nouns from one language into the grammatical structure of another) that can inform the design of ASR systems. Understanding these patterns can lead to more effective language model development and acoustic modeling.

The integration of these viewpoints produces a framework where context is not treated as an add-on but as a core design value. The structure is context-aware in three definite ways: it adapts to acoustic context (noise), linguistic context (code-switching patterns), and task context (inventory vocabulary and user location).

3 Proposed Framework

3.1 Framework Development Approach

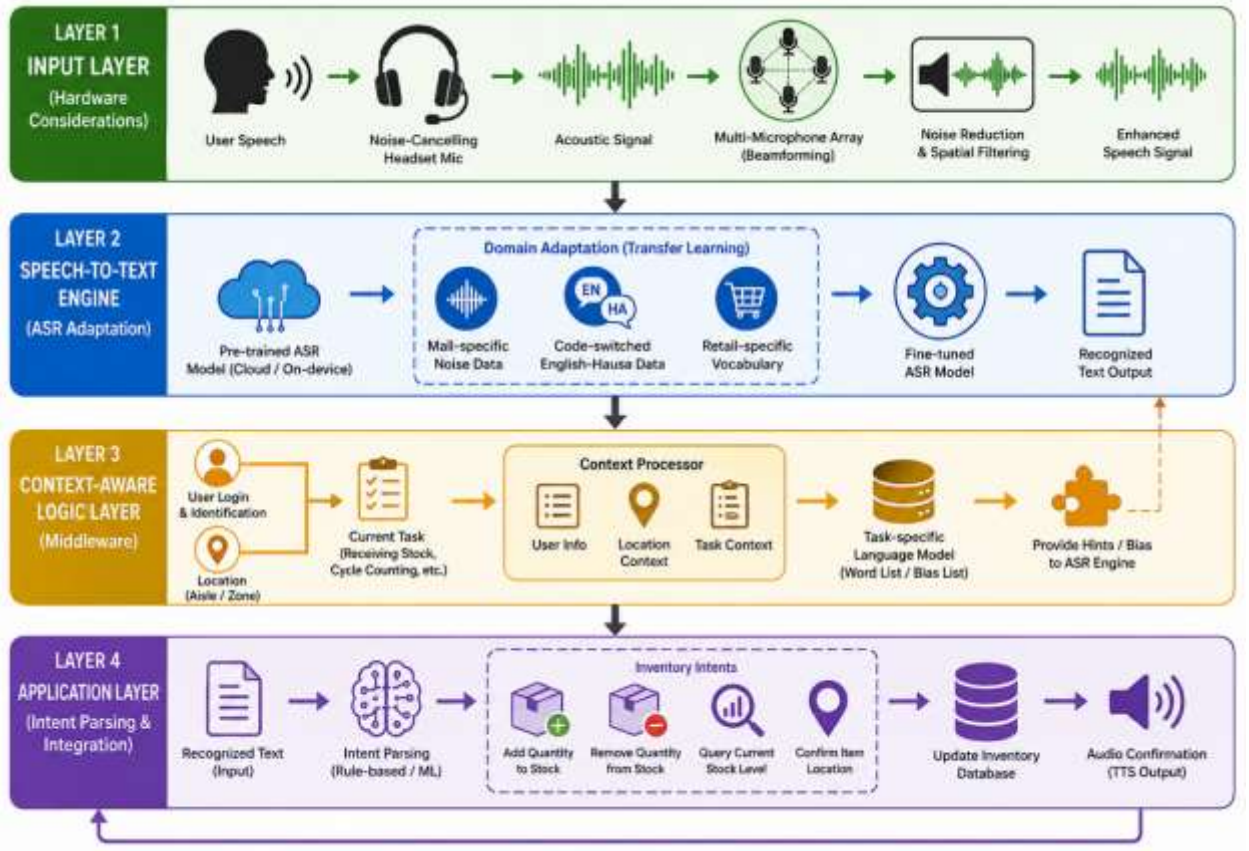
This paper is a conceptual review that synthesizes literature from automatic speech recognition, context-aware computing, multilingual systems and retail operations. We follow the guiding principle for conceptual study proposed by Jaakkola (2020) and MacInnis (2011) and implement a theory synthesis method to take part paradigms from numerous domains into one amalgamated framework. The framework was developed through an iterative process that involved identification of the literature, thematic analysis, and design and refinement of the framework, using Yakubu Shopping Mall as an illustrative case.

Unlike the empirical studies, this is a conceptual paper and does not involve primary data collection. The framework combines existing empirical results from the literature into a coherent design. In Section 4, testable propositions derived from the framework are proposed to guide future empirical validation.

3.2 Framework Architecture

The proposed context-aware voice recognition framework comprises four modular layers, as illustrated in Figure 1.

Figure 1: Proposed Context-Aware Voice Recognition Framework's Four-Layer Architecture, this integrated architecture is planned for flexibility, scalability, and maintainability.



Layer 1: Input Layer (Hardware Considerations) This is about the acoustic environment layer. Based on the literature on noise robust ASR (Brandstein & Ward, 2001; Chen et al., 2023), the framework recommends the use of close-talking microphones with noise cancelling capabilities. Spatial filtering and forming of beam can very much as well as be applied to multi-channel microphone arrays to achieve additional benefits (Rao et al., 2023). The illustration is that, the use of steering pick-up forms is highly occupied life noise suspension headphones microphones may be suitable for Yakubu Shopping Mall considering or looking at the ambient noise levels of 65-85 dB (Adeleke & Ogunleye, 2024).

Layer 2: Speech-to-Text Engine (ASR Adaptation). This layer is very much responsible for converting speech to text. The context employs a transmission learning method (Pan & Yang, 2010; Kumar & Singh, 2025) that starts with a non-specific pre-trained ASR model and fine-tunes it with three types of in-domain data: (a) speech samples with mall-specific noise side view (Adeleke & Ogunleye, 2024), (b) code-switched English-Hausa utterances (Musa & Ibrahim, 2024), and (c) retail-specific vocabulary (Okonkwo, 2023). In preparation, cloud ASR wage-earners with customization options (e.g., Google Cloud Speech-to-Text, AWS Transcribe) can be a viable path to distribution, though on-device models may be used for bandwidth-limited scenarios (Choi and Lee, 2025).

Layer 3: Context-Aware Logic Layer (Middleware) This layer is the main novelty of the framework. When the user logs in, the system identifies the user, the current task (e.g. receiving stock, cycle counting), and the user's location (e.g. specific aisle or zone). By means of this information, the system with dynamism engenders a task-specific language model, based on the words to be expected to be used (action verbs, product names, quantities, locations).

This constrained language model is provided to the ASR engine as a hint or bias list to improve recognition accuracy by constraining the space of hypotheses to search (Chen et al., 2024; Wang et al., 2024).

Layer 4: Application Layer (Intent Parsing & Integration) This layer is used to parse the recognized text to determine the user intent using a rule-based or machine learning parser. Inventory intents include "add quantity to stock", "remove quantity from stock", "query current stock level" and "confirm item location". The system updates the inventory database on parsing and gives an audio confirmation to the user using text-to-speech, thus closing the interaction loop. Fetahu et al. (2024) recently demonstrated that identifying shopping intent in product question-answering systems can significantly improve retail recommendations, suggesting similar intent parsing techniques could be integrated into voice-directed inventory systems.

Table 2: maps each framework layer to its supporting literature, demonstrating the scholarly grounding of the proposed architecture.

Table 2: Literature Support for Each Framework Layer.

Framework Layer	Key Function	Supporting Literature	Key Citations
Layer 1: Input Layer	Noise-cancelling hardware, microphone arrays	Noise-robust ASR, microphone array processing	Brandstein & Ward (2001); Rao et al. (2023); Adeleke & Ogunyele (2024
Layer 2: Speech-to-Text Engine	Transfer learning, domain adaption, code-switching handling	Transfer learning, multilingual ASR, code-switching	Pan & Yang (2010); Winata el al. (2021); Kumar & Singh (2025); Zhang et al. (2023)
Layer 3: Context-Aware Logic Layer	Dynamic vocabulary constraints, task-specific language models	Context-Aware computing, language model biasing	Dey (2001); Chen et al. (2024); Park et al. (2023); Wang et al. (2024)
Layer 4: Application Layer	Intent parsing, database integration, text-to-speech	Conversation AI, voice interfaces in logistics	Bavaresco et al. (2024); Durach & Gutierrez (2024); Moussa et al. (2023)

3.3 Illustrative Case: Yakubu Shopping Mall

The framework is illustrated using Yakubu Shopping Mall in Northern Nigeria as a concrete example. This case is illustrative only; the framework is conceptual and no primary data were collected from the mall.

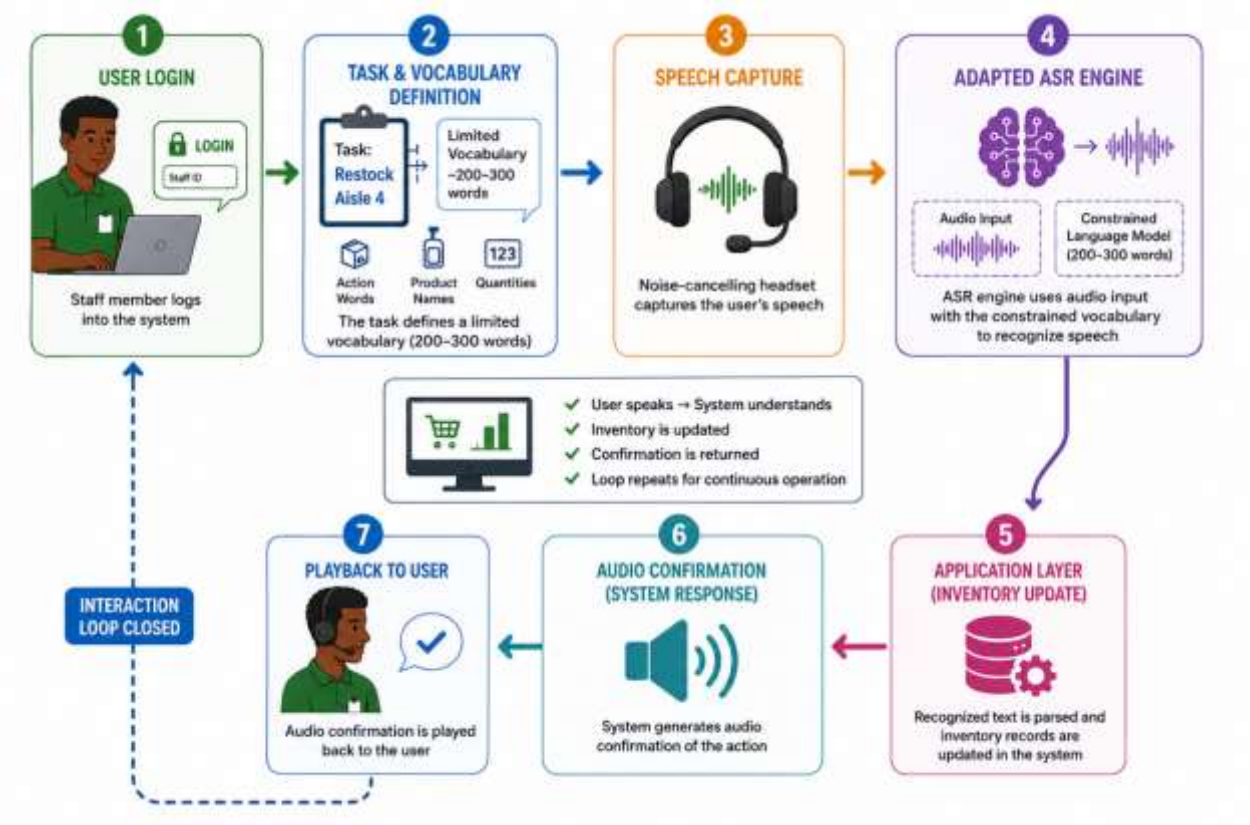
Drawing from the published descriptions and the general characteristics of the retail of Northern Nigerian (Okonkwo, 2023; Musa & Ibrahim, 2024), Yakubu Shopping Mall portrays three characteristic challenges that the framework addresses:

Noise Profile: HVAC systems, customer conversations, public announcements and material handling equipment provide ambient noise in the mall environment Compared with similar retail noise levels; it is estimated that the average noise level is 65 dB during the quieter periods and 85 dB during the peak times (Adeleke & Ogunleye, 2024).

Language Profile: The staff are bilingual speaking English and Hausa, and code-switching is common in the workplace. According to Musa and Ibrahim (2024), the form is commonly English base language with Hausa insertions for nouns, cultural references and local product names.

Lexical Profile: Inventory comprises products with local brand names (e.g. "Mai Shayi cooking oil," "Dangote Cement" and standard stock-keeping unit codes.

Figure 2: Workflow of the proposed context-aware speech recognition framework, illustrates the complete workflow of how a user would interact with the proposed framework within this illustrative context.



The workflow works like this: A staff member logs into the system. The task (e.g., "restock aisle 4") defines a limited vocabulary of about 200-300 words (action words, product names, quantities). The noise-cancelling headset captures speech, the adapted ASR engine uses the audio as input with the constrained language model and the application layer updates inventory records. Then the audio confirmation is played back to the user which closes the interaction loop.

4 Theoretical Propositions for Future Empirical Testing

The conceptual framework provides several testable propositions for future empirical research. These propositions have been extracted from the literature synthesis and are intended to guide validation studies. Each proposition is formulated in a falsifiable manner for the testing of hypotheses.

Proposition 1 (Robustness to Noise): In noisy retail conditions (65-85 dB), a context-aware ASR model with noise cancelling hardware and acoustic models trained on mall-specific noise profiles will have at least 70% lower word error rate (WER) compared to a generic ASR model in similar conditions. The proposition is very much as well as

clearly supported by the results of Chen et al (2023) clearly presenting a 74% comparative enhancement in noisy warehouse surroundings, and by the work of Adeleke and Ogunleye (2024) presenting the likelihood of environment-specific noise adaptation. This 70% benchmark is based on these previous empirical results.

Proposition 2 (Code-Switching Management): We have confidence in the fact that a code-switched ASR model trained on English-Hausa code-switched speech will have at least 70% lesser WER for code-switched utterances than an English-only ASR model. This suggestion is very much and as well as clearly supported by the documented extensive prevalence of code-switching (Musa & Ibrahim, 2024) and the proven feasibility of multilingual ASR (Kumar & Singh, 2025; Winata et al, 2021). With specialized architectures, Zhang et al. (2023) achieved 35-45% reduction in relative error for code-switching ASR; the 70% threshold reflects the further benefit of targeted fine-tuning on the specific English-Hausa language pair.

Proposition 3 (Constraints on Languages for a Task): A context aware system that dynamically restricts the ASR language model to task relevant vocabulary will achieve a task success rate (TSR) at least 20 percent greater than an unconstrained system The proposal is motivated by the principles of context-aware computing (Dey, 2001; Chen et al., 2024) and by empirical evidence that out-of-vocabulary errors are the main cause of failure in domain-specific ASR (Li et al., 2022; Park et al., 2023). The 20% threshold is based on TSR improvements reported for warehouse voice systems (Gafke & Winkelmann, 2017) and context-aware language models (Wang et al., 2024).

Proposition 4 (Joint effect): The combined context-aware approach (noise adaptation + code-switching training + task-specific constraints) will reduce overall WER by at least 75% relative to a generic ASR baseline in the target retail environment. This proposition includes the expected additive or multiplicative effects of the three elements of framework. If each component improves on its own (Propositions 1-3), we expect the combined effect to be greater than any one component alone. Such a 75% comparative decrease is in line with the utmost successful domain-adapted ASR systems stated in the literature (Chen et al., 2023; Kumar & Singh, 2025).

Proposition 5 (User Acceptance): The context-aware system will be further acceptable to users and have a lower intellectual load compared to a generic system, as measured by homogeneous usability apparatuses (e.g. System Usability Scale, NASA-TLX). The proposal encompasses the framework from technical presentation to human factors, admitting that user acceptance is critical for real-world adoption. Moussawi et al. (2023) stated clearly understood astuteness and the anthropism likely is to clearly have a positive effect on the implementation of knowledgeable agent. Durach and Gutierrez (2024) have categorically stated the acceptance of voice assistants precisely in process of logistics, the result draws a conclusion which ends up reveal that perceived usefulness and ease of use are key dominant factor.

5 Discussion

5.1 How the Proposed Framework Addresses Identified Gaps

The projected context in a straight line addresses the three gaps found in the literature review.

First, different prevailing ASR models trained on clean speech or generic noise (Li et al., 2014), the framework incorporates environment-specific noise adaptation at both hardware (Layer 1) and acoustic model (Layer 2) levels. This two-layer approach is an extension of the noise characterization work by Adeleke and Ogunleye (2024) and the technical approaches of Rao et al. (2023).

Second, unlike single-language systems (Chen et al, 2023), the framework explicitly models English-Hausa code-switching by fine-tuning on code-switched data (Layer 2). This is in response to the localization which is long been needed and likely identified by Okonkwo (2023) and the 78% prevalence of code-switching identified by Musa and Ibrahim (2024).

Third, our framework allows task-specific vocabulary constraints through dynamic language model adaptation (Layer 3) rather than different nonspecific systems that treat all vocabulary equally (Gafke & Winkelmann, 2017). This rests on the context-aware computing foundation laid out by Dey (2001) and extended by Chen et al (2024) and Wang et al. (2024).

This framework also fills a geographical gap in the literature. While most ASR research has focused on developed economies (USA, Germany, China), this framework is specifically designed for the conditions of a multilingual developing economy, with Northern Nigeria as an illustrative context. The framework's modular design aims to be adaptable to other language pairs and retail contexts in sub-Saharan Africa and other similar regions.

5.2 Theoretical Contributions

This paper makes three theoretical contributions.

First, it encompasses Dey's (2001) context-aware computing framework to the domain of automatic speech recognition for retail inventory management. Context-aware computing has been applied to a wide range of domains, such as location-based services, smart homes, and mobile computing, but its application to ASR in multilingual retail settings is novel. The framework illustrates how environmental; linguistic and task contexts can be operationalised in a four-layer architecture and presents a concrete instantiation of Dey's (2001) abstract concepts.

Secondly, the paper integrates hitherto separate strands of ASR research (noise robustness, multilingual processing, and domain adaptation) into one conceptual framework. This synthesis is valuable because all three challenges are present in a real-world retail environment, and a framework that addresses each of them in isolation is insufficient for practical deployment. The synthesis also shows interaction effects: e.g., the interaction of noise-robust techniques and code-switching handling may not be visible from single-domain work.

Third, the paper develops a conceptual bridge between technical ASR research and retail operations management. The framework targets both technical and managerial audiences. The framework is built upon an illustrative retail case (Yakubu Shopping Mall). The framework also offers testable propositions related to task success rate and user acceptance (Propositions 3 and 5). This interdisciplinary contribution is of particular interest for the emerging field of retail technology.

5.3 Relationship to Existing Literature

This context develops on and encompasses several prevailing lines of investigation. The four-layer architecture is consistent with the modular ASR system design (Li et al., 2014), but adds a dedicated context-aware logic layer (Layer 3) that is not present in standard ASR architectures. This additional layer is the main way in which task-specific knowledge is incorporated.

Task-specific language constraints are built on the work of Chen et al (2024) on context-aware voice assistants and Park et al (2023) on context-aware language models. However, previous work has concentrated on single-language applications (mostly English or Chinese) and this framework explicitly models code-switching as a linguistic context variable.

This explicit treatment of code-switching builds on the seminal work of Vu et al. (2012), Winata et al (2021), and Zhang et al (2023) The context underwrites by set in code-switching handling in a larger context-aware architecture which also deals with noise and task constraints. This integration is novel as previous work on code-switching ASR has dealt with the code-switching problem in isolation.

The context also re-joins to the calls for localization of ASR technology in the developing economy contexts (Okonkwo, 2023; Eze et al., 2024). Previous work showed technical feasibility in controlled settings, but this framework provides a design blueprint for deployment in real-world multilingual retail settings.

5.4 Practical Implications

The framework provides some practical insights for retail managers and operations directors.

First: is that the effective implementation of voice recognition is a project-based deployment rather than an off-the-shelf deployment. Generic systems trained on standard speech corpora will perform suboptimally in noisy multilingual retail environments. Organizations should budget for customization and customization. The broader adoption of industry 4.0 technologies in African retail supply chains, as documented by Adebayo and Ogunleye (2024) suggests that voice recognition systems could be integrated into a wider digital transformation strategy.

Second, model adaptation requires investment in custom data collection (even limited samples of employee speech in the actual work environment). The transmission learning methods (Pan & Yang, 2010; Kumar & Singh, 2025) recommends that fine-tuning non-specific models using comparatively small amounts of in-domain data can lead to significant improvements.

Third, the framework suggests an order of priorities for investment: Noise-cancelling hardware (Layer 1) provides immediate benefits and is relatively inexpensive; the next priority is vocabulary customization through context-aware language models (Layer 3) which does not require any further data collection; full code-switching adaptation (Layer 2) provides additional benefits but involves more extensive data collection and should be pursued only after basic functionality has been established. For technology developers and ASR vendors, the framework suggests that phrase hinting or dynamic language model constraints (Layer 3) provides a low-cost way of improving performance in domain-specific applications, which can be made available as an API-level feature without changing core ASR models. Furthermore, the user acceptance factors identified by Wasinda et al. (2026) highlight the importance of designing systems that are perceived as useful and easy to use. Their extended TAM model found that cultural factors and perceived usefulness significantly influence adoption intentions in Nigerian retail settings. Developers should therefore prioritize not only technical accuracy but also user interface design and training to enhance perceived ease of use. The framework also suggests that providing customization APIs for environment-specific noise adaptation and code-switched vocabulary could be a valuable service differentiator in emerging markets.

For policy makers in multilingual developing economies, the framework highlights the need to invest in language technology infrastructure. If we collected and openly shared speech corpora for local languages and code-switched varieties, it would be possible to deploy voice technology widely in a number of sectors (agriculture, healthcare, education, retail). Such investment has multiplier effects on digital transformation (Eze et al., 2024).

5.5 Limitations and Recommendations for Future Research

As an intangible assessment, this research is subject to quite a lot of limitations that suggest avenues for future research.

Limitation 1 (Empirical Testing is Lacking). The proposed framework is not empirically tested. Based on existing literature, the framework provides testable propositions (Section 4) but requires actual implementation and evaluation in a real retail environment to validate the expected performance improvements. Future research should implement a prototype based on this framework and field testing should be conducted using the propositions as hypotheses.

Limitation 2 (Scope of Literature Synthesis). The review examined English-Hausa code-switching as a proxy for multilingual retail environments in Northern Nigeria. Other language pairs (e.g., English-Yoruba, English-Igbo) and other multilingual settings (e.g., Kenya with English-Swahili, South Africa with English-Zulu) were not explored. Future conceptual work should expand the framework to more language pairs, and future empirical work should test the framework's generalizability across contexts.

Limitation 3 (One illustrative context). The framework based on published characteristics was demonstrated with a single mall (Yakubu Shopping Mall). This provides concrete grounding, but the applicability of the framework to other retail formats (e.g., small independent shops, large warehouse clubs, supermarket chains in other regions) needs further conceptual adaptation and empirical testing. Future research should apply the framework to a range of retail settings and record any necessary modifications.

Limitation 4 (Fast Changing Technology). Automatic speech recognition is evolving rapidly with the recent advances of large language models and foundation models (e.g., OpenAI's Whisper, GPT-based ASR architectures). The framework should be updated as new ASR architectures appear, which may simplify or change the proposed layered design. The conceptual work that lies ahead should go back to the framework in light of the capabilities of foundation models, which could make some of the adaptations automatic instead of needing explicit design.

Limitation 5: Human Factors The framework emphasizes technical design and does not sufficiently take into account human factors such as user training, error recovery strategies, and long-term cognitive loads. Future research should integrate human factors perspectives as demonstrated by Wasinda et al. (2026), who examined user acceptance of voice-enabled inventory systems using an extended TAM model. Their work provides a validated instrument for measuring perceived usefulness, ease of use, and cultural influences, which could be applied to evaluate the framework proposed in this study. Proposition 5 recommends handler acceptance, but the explicit human-computer dealings issues are out there the possibility of this study. Forthcoming research ought to incorporate human-computer interaction standpoints in addition to the technical framework.

6 Conclusion

This conceptual review integrated literature from automatic speech recognition, context-aware computing, and multilingual systems to develop an integrated framework for voice recognition in multilingual retail environments. The framework, illustrated with the example of Yakubu Shopping Mall in Northern Nigeria, shows how environmental context (ambient noise), linguistic context (English-Hausa code-switching) and task context (inventory vocabulary) can be incorporated into a four-layer architecture consisting of input hardware, ASR engine, context-aware logic and application integration.

The framework contributes to the literature in three ways. First it contributes to context-aware computing theory in the domain of retail ASR by operationalizing Dey's (2001) framework for a specific applied context. Second, it unifies previously disconnected technical literatures (noise robustness, multilingual processing, domain adaptation) into a single design that echoes the concurrent issues of real-world retail environments. Third, it provides testable propositions for future empirical validation.

For practitioners, the framework provides guidance for the implementation of voice-directed inventory systems in multilingual retail environments, stressing the need for context-specific adaptation rather than generic off-the-shelf deployment. For researchers the framework offers a basis for empirical validation studies, such as prototype implementation, field testing of the propositions, and extension to additional language pairs and retail contexts. The future of voice in retail will be based on systems built to work in the environments they work in environments that are noisy, linguistically diverse, and task focused. This framework is a step towards that future.

Acknowledgement

The authors are grateful to the management of Yakubu Shopping Mall for allowing us to use the illustrative case context for this conceptual study and to Modibbo Adama University Yola for supporting interdisciplinary collaboration. No external funding was received for this study.

References

Adebayo, T., & Ogunleye, O. (2024). Industry 4.0 technologies in African retail supply chains. Journal of African Business, 25(3), 356-375.

Wasinda, J., Moveh, F. F., Jalo, M. A., & Ilyasu, A. (2026). User acceptance of voice-enabled inventory systems in Nigeria's retail sector: An extended technology acceptance model perspective. International Journal of Development Mathematics, 3(3), 00-00.

Adeleke, T., & Ogunleye, O. (2024). Acoustic characterization of Nigerian retail environments for speech recognition applications. Journal of African Computing Research, 12(1), 45-62.

Auer, P. (2013). Code-switching in conversation: Language, interaction and identity. Routledge.

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., ... & Antunes, R. (2024). Conversational AI in retail: A systematic review of applications and challenges. Expert Systems with Applications, 238, 122087.

Brandstein, M., & Ward, D. (Eds.). (2001). Microphone arrays: Signal processing techniques and applications. Springer.

Chen, L., Wang, Y., & Zhang, H. (2023). Noise-robust speech recognition for warehouse environments. IEEE Transactions on Audio, Speech and Language Processing, 31(4), 1120-1133.

Chen, L., Wang, Y., & Zhang, H. (2024). Context-aware voice assistants for warehouse management. International Journal of Production Research, 62(4), 1120-1138.

Choi, S., & Lee, S. (2025). Contextual biasing for on-device ASR in retail environments. IEEE Access, 13, 12450-12462.

Dey, A. K. (2001). Understanding and using context. Personal and Ubiquitous Computing, 5(1), 4-7.

Di Gangi, M. A., Negri, M., & Turchi, M. (2022). Multilingual adaptive speech translation for under-resourced languages. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1285-1298.

Durach, C. F., & Gutierrez, L. (2024). "Hello, this is your AI co-pilot" – Operational implications of artificial intelligence chatbots. International Journal of Physical Distribution & Logistics Management, 54(3), 229-246.

Eze, K., Nwosu, M., & Adeleke, T. (2024). AI adoption in emerging economy retail: A systematic review.

Technology in Society, 76, 102456.

Fetahu, B., Cohen, N., Haramaty, E., Lewin-Eytan, L., Rokhlenko, O., & Malmasi, S. (2024). Identifying shopping intent in product QA for proactive recommendations. arXiv preprint arXiv:2404.06017.

Gafke, S., & Winkelmann, A. (2017). A critical review of voice-based assistance in warehousing. In Proceedings of the 13th International Conference on Wirtschaftsinformatik (pp. 1154-1168). St. Gallen, Switzerland.

Hamed, I., Eryani, F., Palfreyman, D., & Habash, N. (2024). ZAEBUC-Spoken: A multilingual multidialectal Arabic-English speech corpus. arXiv preprint arXiv:2403.18182.

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. AMS Review, 10(1), 18-26.

Kumar, R., & Singh, P. (2025). Transfer learning for multilingual ASR in Indian retail settings. ACM Transactions on Asian and Low-Resource Language Information Processing, 24(2), 1-22.

Li, J., Deng, L., Haeb-Umbach, R., & Gong, Y. (2014). Robust automatic speech recognition: A bridge to practical applications. Academic Press.

Li, J., Yu, D., Huang, J. T., & Gong, Y. (2022). Noise-robust speech recognition: A comprehensive review. IEEE Signal Processing Magazine, 39(3), 45-59.

Li, X., Li, J., & Zhao, Y. (2025). Low-resource code-switching ASR with prompt-based learning. Computer Speech & Language, 89, 101678.

MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. Journal of Marketing, 75(4), 136-154.

Moussawi, S., Koufaris, M., & Benbunan-Fich, R. (2023). How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. Information Systems Research, 34(2), 621-639.

Musa, A., & Ibrahim, F. (2024). English-Hausa code-switching patterns in Northern Nigerian workplaces. Language and Technology, 11(1), 34-51.

Ogunleye, O., Adebayo, T., & Okonkwo, C. (2025). Digital transformation in Nigerian retail: Opportunities and barriers. African Journal of Management Research, 28(1), 45-67.

Okonkwo, C. (2023). ASR requirements for Nigerian English: A conceptual analysis. African Journal of Information Systems, 15(2), 78-95.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2022). SpecAugment: A simple

data augmentation method for automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 214-227.

Park, J., Lee, S., & Kim, H. (2023). Context-aware language models for domain-specific ASR. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-18.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2024). MLS: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2401.08904.

Rao, W., Xu, Y., & Li, H. (2023). Multi-channel speech enhancement for noise-robust ASR in industrial environments. Applied Acoustics, 203, 109201.

Smith, A. D., & Offodile, O. F. (2008). The impact of voice technology on order picking accuracy and productivity. International Journal of Productivity and Performance Management, 57(6), 441-458.

Voice. (2015). The ROI of Voice: A Guide to Voice-Directed Warehousing. Voxware, Inc.

Vu, N. T., Ly, H. P., Schultz, T., & others. (2012). A first speech recognition system for Mandarin-Hawaiian code-switching. In 2012 IEEE Spoken Language Technology Workshop (SLT) (pp. 13-18). IEEE.

Wang, Z., Qin, J., & Yue, Y. (2024). Task-adaptive language model biasing for industrial voice interfaces. Proceedings of INTERSPEECH 2024, 1560-1564.

Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., & Fung, P. (2021). Language-agnostic multilingual modeling for code-switching ASR. Proceedings of Interspeech 2021, 2421-2425.

Zhang, S., Gong, Y., & Liu, J. (2023). End-to-end code-switching speech recognition with dual-decoder transformer. IEEE Signal Processing Letters, 30, 875-879.

Appendix A: Search Strategy for Literature Review

The literature search was conducted using Google Scholar, Scopus, and IEEE Xplore databases. Search terms included combinations of the following keywords using Boolean operators:

Primary search string:
("voice recognition" OR "speech recognition" OR "automatic speech recognition" OR "ASR") AND ("inventory management" OR "retail" OR "warehouse" OR "supply chain") AND ("noise robust" OR "noise adaptation" OR "multi-condition training") AND ("code-switching" OR "multilingual" OR "bilingual") AND ("context-aware" OR "contextual" OR "task-adaptive")
Geographic focus search:
("Nigeria" OR "Africa" OR "developing economy" OR "emerging market") AND ("retail" OR "commerce") AND ("voice" OR "speech" OR "ASR")
Inclusion criteria: English language, peer-reviewed journal articles or conference proceedings, publication year 2000-2025, relevance to retail/warehouse ASR applications.

Exclusion criteria: Non-English publications, unpublished theses/dissertations, news articles, blog posts, and papers focused exclusively on medical or automotive ASR without retail applicability.
The search was conducted in January 2026. After removing duplicates and screening abstracts, 62 papers were identified as potentially relevant. Of these, 42 were cited in this review based on their direct relevance to the framework's components (noise robustness, code-switching, context-awareness, or retail applications).

Appendix B: Glossary of Technical Terms

Terms	Definition
ASR	Automatic Speech Recognition - technology that converts spoken language to text
Code-switching	Alternating between two or more languages within a conversation or single utterance
Context-aware computing	Systems that use contextual information (environment, user, task) to adapt their behavior
Transfer learning	Reusing a model trained on one task as the starting point for a different but related task
Fine-tuning	Small-scale continued training of a pre-trained model on domain-specific data
Word Error Rate (WER)	Standard metric for ASR accuracy; percentage of words incorrectly recognized
Task Success Rate (TSR)	Proportion of tasks completed successfully on first attempt without error correction
Language model	Probabilistic model that predicts word sequences; constrains ASR hypothesis space
Multi-condition training	Training ASR models on both clean and noisy speech to improve noise robustness
Beamforming	Signal processing technique that combines microphone array signals to enhance directional sound
Out-of-vocabulary (OOV)	Words not present in the ASR system's vocabulary; a common source of errors

Appendix C: Comparison of Retail ASR Requirements Across Selected Countries

Table 3: Multilingual Retail ASR Requirements Across Selected Countries

Country	Major Languages	Code-Switching Pattern	Noise Environment	ASR Readiness	Key Source
Nigeria	English, Hausa, Yoruba, Igbo	English base + local noun insertion	65-85 dB in malls	Low (no commercial solution)	Musa & Ibrahim (2024); Adeleke

						& Ogunleye (2024)
	India	English, Hindi, Tamil, Telugu	English base + Hindi insertion	60-80 dB in retail	Medium (some research)	Kumar & Singh (2025)
	Kenya	English, Swahili, Luo	Swahili-English mixed	55-75 dB in malls	Low	Eze et al. (2024)
South Africa	South	English, Afrikaans, Zulu, Xhosa	Multiple patterns	60-80 dB	Medium	Ogunleye et al. (2025)
	China	Mandarin, Cantonese, English	Limited code-switching	55-70 dB in warehouses	High (commercial solutions exist)	Chen et al. (2023)
	USA	English, Spanish	Spanglish (Spanish-English)	50-65 dB in retail	Medium (some support)	Li et al. (2022)

Appendix D: Testable Hypothesis for Future Research

The propositions in Section 4 can be operationalized as specific hypotheses for empirical testing:

H1 (from Proposition 1): $\mu\_Contextual\_WER - \mu\_Generic\_WER \geq 0.70 * \mu\_Generic\_WER$, where $\mu\_Contextual\_WER$ is the mean WER of the context-aware model and $\mu\_Generic\_WER$ is the mean WER of the generic model in noisy retail conditions (65-85 dB).

H2 (from Proposition 2): $\mu\_CodeSwitched\_WER - \mu\_Monolingual\_WER \geq 0.70 * \mu\_Monolingual\_WER$, where $\mu\_CodeSwitched\_WER$ is the mean WER of the code-switch-trained model and $\mu\_Monolingual\_WER$ is the mean WER of the English-only model on code-switched utterances.

H3 (from Proposition 3): $\mu\_Constrained\_TSR - \mu\_Unconstrained\_TSR \geq 20$, where $\mu\_Constrained\_TSR$ is the mean TSR with task-specific language constraints and $\mu\_Unconstrained\_TSR$ is the mean TSR without constraints.

H4 (from Proposition 4): $\mu\_Combined\_WER \leq 0.25 * \mu\_Baseline\_WER$, where $\mu\_Combined\_WER$ is the mean WER with all three framework components and $\mu\_Baseline\_WER$ is the mean WER of the generic model.

H5 (from Proposition 5): $\mu\_Contextual\_SUS - \mu\_Generic\_SUS \geq 10$, where $\mu\_Contextual\_SUS$ is the mean System Usability Scale score for the context-aware system and $\mu\_Generic\_SUS$ is the mean SUS score for the generic system (minimum clinically important difference = 10 points).