Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automatic speech recognition system for court proceedings in Ethiopia

Tedros Kebede Bidada
Aksum University
Dawit Teklu Weldeslasie
dawit.tekulu@gmail.com

Aksum University https://orcid.org/0009-0000-5388-5197

Research Article

Keywords: Under-resource language, ASP, HMMs, MFCC, Ethiopia Court proceedings, Legal informatics

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10610158/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Automatic speech recognition (ASR) technologies have revolutionized specialized domains like healthcare and broadcasting, yet their application within legal and judicial proceedings remains unused for under-resourced languages. While foundational ASR research exists for Ethiopian language, a specialized system tailored to the unique linguistic and structural demands of the judiciary has been notably absent. This study addresses this gap by proposing a speaker-independent Tigrigna ASR system specifically designed for court proceedings in Ethiopia. Utilizing a Hidden Markov Model (HMM) framework, we developed a dedicated Tigrigna speech corpus featuring recordings from twenty-four native speakers, partitioned into training (90%) and testing (10%) subsets. The conventional ASR architecture integrates acoustic, language, and dictionary models. Both context-independent and context-dependent phone-level acoustic models were engineered using the CMU Sphinx toolkit, with speech waveforms parameterized into acoustic feature vectors via Mel-Frequency Cepstral Coefficients (MFCC). Furthermore, a phonetic dictionary was developed using a custom Java application, and a trigram language model was constructed via the SRI Language Modeling Toolkit. Experimental results indicate that the best recognition performance is obtained using a phoneme-based, context-dependent model configured with 11 Gaussian mixture components. Furthermore, comparative analysis shows that the context-dependent acoustic model substantially outperforms the context-independent variant, reaching an optimal baseline word accuracy of 73.84%. Accordingly, this study establishes a solid baseline for the digitalization of judicial workflows in Ethiopia.

1. Introduction and context

Speech recognition technology enables computational systems to capture human vocalizations and translate them into corresponding text or executable commands. Automatic Speech Recognition (ASR) has become a transformative tool in modern workflows, optimizing document preparation, data retrieval, and voice-driven command execution [1]. Consequently, ASR remains a foundational area of computational research with a broad spectrum of applications, including automated dictation, command recognition, data entry, and interactive voice response systems. In dictation tasks, ASR systems streamline documentation by replacing manual typing with real-time speech-to-text conversion [2]. Today, automated dictation represents one of the primary deployments of ASR, proving particularly vital in specialized, high-stakes environments such as medical transcription, corporate data processing, and legal proceedings. To support these applications, numerous open-source speech recognition frameworks have been developed, traditionally leveraging Hidden Markov Models (HMMs) as their core mathematical architecture.

A critical phase in any speech recognition pipeline is feature extraction, wherein raw acoustic waveforms are converted into parametric representations suitable for downstream processing and analysis [3]. Currently, Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) represent the two most prominent feature extraction techniques. The MFCC technique characterizes the acoustic "fingerprint" of audio files by mimicking the human ear’s critical bandwidth frequencies; it utilizes filters spaced linearly at lower frequencies and logarithmically at higher frequencies to capture critical speech characteristics [4][5].

Standard ASR architectures generally integrate three primary components; an acoustic model, a language model, and a lexicon (pronunciation dictionary). The acoustic model maps the extracted features of speech signals to the expected phonetics of a hypothesized sentence. Statistically, HMMs remain a robust and widely adopted approach for effective acoustic modeling [6]. Within this framework, HMMs are constructed for individual words or phonemes to capture their distinct acoustic profiles. These phonemes are further categorized into context-independent units (monophones) and context-dependent units (such as triphones), with Gaussian Mixture Models (GMMs) frequently employed to model the emission distributions of the hidden states [7].

Complementing the acoustic model, the language model applies linguistic constraints within a statistical framework to calculate the probability of various word sequences, thereby guiding the recognizer toward the most contextually

appropriate textual output. Bridging these models is the pronunciation dictionary, a critical training component that maps vocabulary words directly to their corresponding phonemic representations.

Significant progress has been made in developing ASR systems for Ethiopian languages. Notable benchmarks include the development of spontaneous speech recognition for Amharic using HMMs [8][9], as well as large-vocabulary, speaker-independent continuous systems for the Tigrigna language [10]. Globally, domain-specific ASR systems have also gained traction, as evidenced by the automatic speech recognition system for the judicial domain [11][12].

Within the judicial domain, transcribers face the highly demanding task of manually translating courtroom audio into verbatim legal text. Developing automated dictation systems capable of converting speech into accurate text holds substantial parametric and operational significance for the efficiency of legal systems. To address this need in an under-resourced linguistic context, this research aims to design and evaluate a speaker-independent, Tigrigna-based automatic speech recognition system tailored specifically for court proceedings of Ethiopia using a Hidden Markov Model framework.

2. Problem statement

Speech serves as the most primary and intuitive medium of communication among human beings. Within the evolution of human-computer interaction (HCI) [13], voice-driven interfaces offer a powerful mechanism to reduce physical reliance on traditional computing hardware, shifting workflows away from continuous manual and visual engagement. Furthermore, ASR functions as a critical pillar for assistive technologies, enabling individuals with visual impairments or motor disabilities to seamlessly interact with digital systems [14][15]. Consequently, the core objective of ASR research lies in perfecting the conversion of human vocalizations into precise, actionable texts or executable commands.

While ASR technologies have achieved high levels of maturity and widespread deployment for dominant, high-resource global languages, a profound disparity remains in low income countries. In the context of Ethiopia, research in language technologies remains severely limited, particularly for the Tigrigna language. Tigrigna, an official working language in Ethiopia, belongs to the Ethiopian Semitic branch of the Semitic language family. Boasting 6 million speakers, it stands as one of the largest Semitic languages, surpassed only by Arabic and Amharic [16]. Despite its regional and cultural prominence, Tigrigna remains a digitally under-resourced language, lacking the robust computational frameworks necessary for modern automated workflows.

This linguistic digital divide is particularly evident in the Ethiopian legal ecosystem. In the judicial domain, court proceedings rely on precise, manual transcription, which is inherently time-consuming, prone to human fatigue, and susceptible to administrative backlogs. While some foundational ASR research has emerged for Tigrigna language [17][18], a specialized system tailored to the specific vocabulary, acoustic constraints, and structural demands of courtroom environments has yet to be realized.

Driven by the need to advance regional HCI capabilities and bridge the digital gap in legal informatics, this study focuses on the development of a speaker-independent, Tigrigna-based ASR system designed specifically for court proceedings. By automating transcription workflows within the judiciary, this research aims to preserve linguistic equity, optimize legal administration, and establish a foundational baseline for automated legal documentation in the Tigrigna language. The following research questions were addressed based on the technical, legal and linguistic aspects of building, training, and testing the developed ASR system's accuracy.

1. Which acoustic and language model configurations optimize the performance of the Tigrigna ASR system when processing domain-specific legal vocabulary?
2. How does the system handle real-world courtroom audio challenges, such as background noise and varying recording qualities?
3. How does the automated transcription workflow compare to traditional manual methods in terms of time efficiency and processing costs within Tigrigna-speaking court jurisdictions?

3. State of the art

This section outlines the principles of Statistical Pattern Recognition (SPR), examines the acoustic and phonetic properties characterizing the phonology of the Tigrigna language, and reviews relevant global and domestic advancements in domain-specific ASR.

3.1 Statistical pattern recognition approach

SPR is a predominant paradigm in contemporary ASR, offering a well-formulated mathematical framework that establishes consistent, robust representations of speech patterns [19]. The SPR methodology fundamentally operates across two distinct phases: pattern training and pattern comparison. During the training phase, statistical models are optimized using a substantial volume of representative acoustic data to model targeted speech units, such as phonemes, within a specific task domain. In the pattern comparison phase, unseen speech signals are evaluated against these trained models to identify and select the sequence that yields the maximum likelihood match.

The dominance of pattern-matching techniques in modern speech recognition is largely driven by the efficiency of Hidden Markov Models (HMMs) in capturing the dynamic, time-varying properties of speech signals [20]. HMMs facilitate computationally efficient model training and robust decoding via the Viterbi algorithm. Despite these advantages, a primary challenge in SPR-based architectures remains system robustness against signal variations, such as environmental noise and channel distortion. HMMs are highly effective for acoustic modeling due to their capacity to represent sequences of discrete or continuous observable symbols. Structurally, the underlying sequence of unobserved states forms a Markov chain, where state transitions are governed by an explicit set of transition probabilities [21].

Each hidden state possesses an associated probability density function that models the generation of continuous acoustic feature vectors. In standard ASR architectures, this general framework is typically simplified into a first-order, left-to-right topology. In a first-order HMM, the probability of transitioning to a subsequent state depends exclusively on the immediate preceding state. The left-to-right constraint dictates that state transitions are strictly unidirectional or self-looping, effectively capturing the temporal progression of spoken language. The emission probability distributions of these hidden states are frequently modeled using Gaussian Mixture Models [22][23]. A GMM is a parametric probability density function expressed as a weighted sum of Gaussian component densities:

$$\mathrm{p}\left(\mathrm{x}\right)=\sum_{m=1}^{M} c_m \mathrm{N}(\mathrm{x};\ \mu_m,\ \Sigma_m)$$

Where x is a continuous acoustic feature vector, c_m represents the mixture weights, and $N(x; \mu_m, \Sigma_m)$ denotes the component Gaussian densities characterized by mean vectors μ_m and covariance matrices Σ_m. GMM parameters are typically estimated from training data using the iterative Expectation-Maximization (EM) algorithm or via Maximum a Posteriori (MAP) adaptation from well-trained prior models. In small-vocabulary tasks, distinct HMMs can be constructed for entire words. However, large-vocabulary continuous speech recognition (LVCSR) tasks encompass thousands of unique lexical entries, rendering word-level modeling computationally intractable.

To resolve this, sub-word units, specifically phonemes, are utilized as the foundational building blocks of the ASR system, as most languages can be decomposed into several dozen phonemes. These phonetic units are further classified into context-independent models (monophones) and context-dependent models (such as triphones) to account for coarticulation effects.

3.2 Phonology of Tigrigna language

Tigrigna is a Semitic language belonging to the Afroasiatic language phylum. It serves as the official working language of the Tigray Regional State of Ethiopia and is recognized as a national official language in Eritrea, where it is extensively spoken across central territories. Tigrinya has 32 consonant and 7 vowel phonemes. Distinctive are the glottalized

consonants and the labialized velars [16]. The phonetic inventory of Tigrigna consists of a distinct arrangement of consonants and vowels, which are systematically categorized as shown in Table 1.

Table 1
, Phonetic inventory of Tigrigna consonants [19]

Manner of articulation	Places of articulation									
	Bi-labial	Labio-dental	Alveolar	Palatal	Velar	Labio-velar	Uvular	Labio-uvular	Pharyngeal	Laryngeal
Stop - Voiceless	p		t		k	k□				□
Stop - Voiced	b		d		g	g□				
Stop - Ejective	p’		t’		k’	k□’				
Nasal - Voiced	m	m	n	□	ŋ	ŋ□				
Fricative - Voiceless		f	s	□			χ	χ□	ħ	h
Fricative - Voiced		v	z	□					□	
Fricative - Ejective			s’				χ’	χ□’		
Affricate - Voiceless				t□						
Affricate - Voiced				d□						
Affricate - Ejective				t□’						
Approximant - Voiced		□		y		w				
Trill - Voiced			r							
Lateral - Voiced			l							

In the Tigrigna orthographic system (Ge'ez script), each grapheme represents a syllabic unit composed of a consonant-vowel (CV) combination. This tight integration between consonantal structures and vocalic modifiers requires specialized acoustic and lexical mapping during dictionary development (see Fig. 1).

3.3 Related works in domain-specific and regional ASR

Developing ASR systems for low-resource languages remains challenging due to the scarcity of annotated speech data. Addressing this issue, Alemu et al. [24] proposed an ASR framework for Amharic using the Kaldi toolkit, demonstrating that effective speech recognition can be achieved by adapting existing tools to resource-constrained languages. Similarly, Ejigu et al. [26] developed a noise-robust end-to-end ASR model for Amharic using TensorFlow, showing its suitability for real-world environments with background noise, such as courtrooms.

ASR technology has significant potential in Ethiopia's legal sector by improving the documentation, accessibility, and transparency of court proceedings. Automated transcription can enhance legal record management and support broader

access to justice within Ethiopia's civil law system [25][29].

Recent studies have also advanced language technologies for Tigrigna. Temesgen [27] proposed a deep learning approach for identifying Ethiopian languages written in the Ge'ez script, while Gebretatios et al. [28] demonstrated the effectiveness of pre-trained models for Tigrinya text classification. Furthermore, linguistic studies on Tigrigna discourse provide insights that can support the development of more context-aware speech recognition systems [31]. These efforts, together with the growing adoption of AI applications in Ethiopia [30], highlight the increasing interest in language technologies for Ethiopian languages.

While these regional studies established important baselines for Tigrigna ASR, they focused on general-purpose or spontaneous speech environments without addressing the domain-specific linguistic and acoustic challenges of the legal sector. To bridge this critical gap, the present study introduces an optimized, HMM-based speaker-independent Tigrigna ASR framework tailored specifically to capture and transcribe the rigorous terminology of court proceedings.

4. Research methodology

This study adopts the Design Science Research (DSR) methodology [32] to guide the conceptualization, execution, implementation, and evaluation of the specialized automatic speech recognition (ASR) system. DSR provides a structured, iterative framework ideal for developing innovative technological artifacts that address distinct real-world operational problems, in this case, the automation of judicial transcription for an under-resourced language.

4.1 Corpus preparation and pre-processing

The empirical foundation of this research relies on the compilation of two distinct speech datasets: a training corpus to optimize model parameters and a testing corpus to evaluate recognition performance. The primary speech corpus comprises three hours of spontaneous audio recordings capturing real-world courtroom interactions and multi-speaker conversations during actual legal proceedings. To robustly handle the acoustic variability inherent in legal domains, data-cleaning procedures were implemented to isolate high-quality domain-specific speech while identifying and filtering out irrelevant acoustic artifacts, read speech anomalies, and excessive background environmental noise. This acoustic database was systematically partitioned into a 90% training subset and a 10% testing subset. To supplement the speech corpus and ensure the construction of a robust, highly predictive language model, a collection of plain Tigrigna text documents was integrated into the pipeline, expanding the lexical and syntactic scope of the domain vocabulary.

4.2 Speech recognizer development process model

The architectural framework governing the design and implementation of the proposed Tigrigna ASR system is modeled after the CMU Sphinx architecture, integrating multiple interconnected data-processing phases (see Fig. 2).

Hence, following initial data collection, an ethical speech selection and curation process [33][34] was applied to the raw audio files to prepare the datasets for downstream evaluation. The curated speech waveforms were segmented at word, phrasal, and sentence levels to generate manageable acoustic units for training and verification. Textual transcripts were subsequently generated for each segmented audio file, establishing the baseline truth for training.

Simultaneously, a feature extraction pipeline parameterized the raw segmented waveforms into low-dimensional acoustic feature vectors via Mel-Frequency Cepstral Coefficient (MFCC) analysis. To map these files to the recognizer, dedicated control files were formulated to define the explicit directory paths for the training and testing audio segments.

In parallel, a pronunciation lexicon was developed directly from the transcriptions, mapping every unique vocabulary word to its corresponding phonetic sequence. During the training phase, these structural files, lexicons, and acoustic features were used to build the context-independent and context-dependent HMM-based acoustic models.

Finally, during the evaluation phase, the test data and control configurations were routed through the decoder, which leveraged a trigram language model computed from the combined transcriptions and plain texts to estimate word sequence probabilities and output the final recognized text.

5. The implementation

5.1 System development tools and frameworks

To implement the speaker-independent Tigrigna automatic speech recognition (ASR) framework, a suite of specialized, open-source speech processing and language modeling toolkits was integrated into a unified development pipeline.

The initial acoustic preprocessing and waveform segmentation phases were executed using the PRAAT speech processing framework. PRAAT provides a highly flexible environment for continuous speech analysis, supporting both standard and non-standard digital signal processing procedures, such as spectrographic analysis, articulatory synthesis, and neural network modeling. This tool was vital in accurately partitioning the continuous courtroom audio recordings into discrete, manageable phonetic and phrasal units.

Statistical language modeling was performed utilizing the SRI Language Modeling (SRILM) toolkit. The SRILM toolkit provides an extensible architecture for generating, training, and evaluating statistical language models (N-grams). Within this study, it was leveraged to compute the trigram probabilities that govern the grammatical and syntactic constraints of the Tigrigna legal lexicon.

Acoustic model training was carried out via the SphinxTrain toolkit, an HMM-based acoustic optimization framework. For the Tigrigna linguistic profile, a phonetic inventory consisting of 51 distinct phones was defined, with individual Hidden Markov Models subsequently trained for each sub-word unit.

To convert the raw, segmented speech waveforms into parametric data structures suitable for HMM training, the sphinx_fe front-end tool was executed. This module extracts acoustic feature vectors using the Mel-Frequency Cepstral Coefficients (MFCC) technique, capturing the primary spectral characteristics of the target speech.

Finally, the decoding and evaluation phases were conducted using the Sphinx decoder. Developed as a highly modular, flexible and pluggable Java-based framework, Sphinx enables seamless interaction between the trained HMM acoustic models, the SRILM-generated language models, and the custom pronunciation dictionary, thereby ensuring efficient text transcription during courtroom simulation tests.

5.2 Data set configuration

The empirical evaluation of the proposed ASR system relies on a meticulously partitioned speech corpus to ensure unbiased performance metrics. The distribution of the audio datasets used for training and evaluating the Tigrigna speech recognizer is summarized in Table 2.

Table 2
, Dataset distribution for ASR training and testing

No	Data set	Gender	No. of Speaker	No. of Utterance
1	Training	Female	10	900
		Male	10	950
2	Testing	Female	2	120
		Male	2	120
Total			24	2090

As outlined in Table 2, the allocation of testing utterances strictly follows established empirical conventions in speech informatics, wherein approximately one-tenth (10%) of the total corpus is reserved exclusively for system validation. The recorded utterances maintain an average duration of approximately 5 seconds per segment. To guarantee the scientific validity of the evaluation and prevent data leakage, the testing partition consists entirely of acoustic data that remains completely unseen by the acoustic models during the training phase.

5.3 Preprocessing and acoustic feature extraction

The raw courtroom audio corpus underwent a rigorous segmentation and annotation process utilizing the PRAAT speech processing framework. To maintain the high fidelity required for legal proceedings, the audio data was transcribed verbatim. This included transcribing exact spoken forms, encompassing colloquial variations, dialectal nuances, and technical or mathematical expressions typical of judicial arguments. Furthermore, non-speech acoustic events (such as ambient noise, pauses, and speaker breath) were systematically integrated into the transcription corpus using standard annotation conventions.

Following annotation, the segmented audio data was routed through the sphinx_fe executable, the front-end feature extraction module of the CMU Sphinx package. This process parameterized the continuous time-domain waveforms into low-dimensional acoustic feature vectors via Mel-Frequency Cepstral Coefficient (MFCC) analysis, preparing the sequence tokens for downstream Hidden Markov Model (HMM) training.

5.4 Development of pronunciation dictionary model

The pronunciation lexicon serves as a mapping layer between the orthographic text and its corresponding phonetic units. It enables the ASR system to map acoustic parameters to phonemes during training and compute the most probable phonetic sequences during decoding. Given that minor transcription errors degrade acoustic modeling accuracy, a meticulous dictionary compilation process was executed.

A canonical phoneme-based dictionary was constructed using a vocabulary of 5,205 unique lexical items extracted directly from the courtroom transcripts. This vocabulary was mapped using a phonetic inventory optimized for the Tigrigna language. Additionally, to handle conversational legal speech effectively, non-speech acoustic phenomena were assigned specialized entries within a dedicated filler dictionary, allowing the decoder to treat environmental noises or pauses as distinct token boundaries rather than linguistic data.

5.5 Development of acoustic model

The speaker-independent acoustic model was developed using the SphinxTrain toolkit, processing a spontaneous judicial speech corpus contributed by 20 distinct native speakers (balanced equally with 10 male and 10 female participants). The toolkit iteratively computed the essential parameters governing the acoustic states, including mean vectors, variance matrices, state transition probabilities, and mixture weights.

Acoustic model training integrated the extracted MFCC feature files, the verified verbatim transcripts, the target Tigrigna phoneme inventory, and the compiled pronunciation dictionary. The dictionary comprised 5,205 unique terms mapped across 47 distinct phone units, which included 39 base Tigrigna linguistic phonemes and 8 non-speech or silence fillers.

Adhering to standard SphinxTrain procedures, both context-independent (monophone) and context-dependent (triphone) acoustic models were generated. A continuous HMM topology was selected for all training configurations, with each phonetic unit modeled using a five-state left-to-right (Bakis) architecture featuring a skip topology to accommodate temporal variations in speaking rates. Parameter optimization was achieved via the Baum-Welch (forward-backward) expectation-maximization algorithm.

The initial Gaussian mean and variance vectors for each state were initialized to the global mean and variance values computed across the entire training corpus. To optimize triphone mapping and resolve data sparsity issues, state tying was

applied to constrain the system to 1,000 tied states, matching the recommended architectural thresholds for medium-sized vocabularies.

5.6 Development of language model

In statistical ASR architectures, a language model is critical to establishing the probability distribution of word sequences, thereby guiding the search space toward syntactically valid sentences. Language models estimate the probability of language units based on textual training data. Due to the inherent structural constraints of natural language, specific word pairings exhibit high contextual probabilities, whereas others are structurally inadmissible. By applying these constraints, language modeling constrains the decoder's search space, reducing candidate sequence hypothesis errors.

Because estimating the unconstrained probability distribution across an entire vocabulary is computationally intractable, the search is bounded using an N-gram approximation framework, which calculates probabilities based on a sliding window of N consecutive words.

This study implemented a statistical trigram language model utilizing the SRI Language Modeling (SRILM) toolkit. To address data sparsity and prevent zero-probability assignments for unseen word sequences, a modified absolute discounting smoothing technique was applied. The language model was trained on a 12 MB normalized text corpus derived from a combination of the judicial transcriptions and an external Tigrigna plain-text database.

Prior to training, a rigorous text normalization pipeline was executed: excessively long sentences were segmented, numeric expressions were converted to their full text equivalents, typographical and grammatical anomalies were corrected, and all punctuation marks were removed to establish a clean, standardized training corpus.

5.7 Performance evaluation and discussion

The empirical evaluation of the developed Tigrigna ASR system was executed across two distinct phases: first, by training and testing context-independent (CI) monophone acoustic models, followed by the deployment of context-dependent (CD) triphone acoustic models. To ensure a controlled experimental environment, both modeling phases utilized identical configurations for the phonetic lexicons, verbatim transcriptions, and extracted Mel-Frequency Cepstral Coefficient (MFCC) feature parameters within the HMM framework.

To assess the system's transcription accuracy quantitatively, the standardized metric of Word Error Rate (WER) was employed. The WER measures the degree of divergence between the text sequence generated by the decoder (the hypothesized word string) and the corresponding verbatim courtroom transcript (the reference string).

Computing the WER requires calculating the minimum edit distance at the word level between the hypothesized string and the reference string. This dynamic programming approach determines the minimum number of single-word insertions (I), substitutions (S), and deletions (D) necessary to align the hypothesized sequence with the reference text. The Word Error Rate is formally expressed as follows:

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

Where, S represents the total number of word substitution errors, D represents the total number of word deletion errors, I represents the total number of word insertion errors, and N denotes the total number of words in the baseline reference transcript.

To illustrate the comparative performance of the system, as shown in Fig. 3, across the different acoustic modeling configurations, sample outputs containing individual sentence hypotheses generated by the decoder, alongside their

corresponding reference transcriptions, are analyzed in the following sections. Accordingly, the following performance was achieved by the recognizer using the testing data set.

Table 3
, the performance evaluation result of the spontaneous Tigrigna speech recognizer

HMM states	Acoustic Model	# Gaussian mixture	%WER	% Accuracy
Five states	CI	-	46.88	53.12
With skip topology	**CD**	1	33.68	66.32
		2	32.15	67.85
		4	28.75	71.25
		8	27.65	72.35
		10	27.08	72.92
		11	**26.16**	**73.84**
		12	32.96	67.04

Key: - **Bold** recorded data indicate that the CD acoustic model utilizing 11 Gaussian mixtures yielded optimal results

Based on the empirical results compiled in Table 3, three critical observations emerge regarding the system's behavioral performance. Initial increments in the number of Gaussian mixture components yielded a corresponding rise in overall word recognition accuracy. However, performance reached an asymptotic threshold and began degrading at twelve Gaussian mixtures. This performance drop is attributed to overfitting; a 12-component GMM over-parameterizes the continuous feature space, forcing the model to capture acoustic noise and irrelevant spectral variations instead of robust phonetic traits.

Spontaneous courtroom speech is highly susceptible to coarticulation effects, a physiological phenomenon where the physical movement of human articulators (such as the tongue, lips, and velum) is continuous and constrained by inertia. As a result, an articulator begins transitioning toward the target position of a subsequent sound before completing the current phone. While monophone models fail to capture these transitional variations, context-dependent (CD) triphone models successfully map these coarticulation effects by evaluating each phoneme within its distinct left and right phonetic contexts. Consequently, the CD acoustic model significantly outperformed the CI framework, achieving a substantial 20.72% gain in word recognition accuracy.

Across all experiments conducted on the speaker-independent recognizer, the CD acoustic model utilizing 11 Gaussian mixture components yielded optimal results when exposed to the four unseen speakers, establishing a baseline word accuracy of 73.84%. An analysis of the underlying error distributions reveals an average insertion rate of 14.45%, a deletion rate of 26.15%, and a high substitution rate of 81.00%. The high rate of substitutions represents the primary source of accuracy degradation. This trend stems from phonetic ambiguities within the vocabulary, where words containing acoustically similar phonemic sequences were mapped interchangeably by the decoder (see Fig. 4).

6. Conclusion

This study successfully designed and evaluated a speaker-independent, automatic speech recognition (ASR) system for spontaneous Tigrigna speech within the specialized domain of judicial proceedings. Adopting a robust statistical framework, the system leverages Hidden Markov Models (HMMs) engineered via the CMU Sphinx toolkit. The underlying empirical database comprised a three-hour courtroom speech corpus collected from 24 native speakers, systematically

partitioned into a training set consisting of 20 speakers (1,800 utterances) and a held-out testing set consisting of 4 speakers (300 utterances).

Adhering to standard SphinxTrain architectures, both context-independent (monophone) and context-dependent (triphone) acoustic models were developed, supported by a domain-specific phonetic lexicon and a statistical trigram language model. Comparative performance evaluations demonstrated that the context-dependent acoustic model significantly outperformed its context-independent counterpart, yielding an optimal baseline word accuracy of 73.84%.

Crucially, the acoustic analysis highlighted unique sociolinguistic challenges inherent to legal informatics. Courtroom speech involves diverse actors, including witnesses and defendants, who are often unfamiliar with formal judicial environments or recording constraints. The resulting speech patterns are frequently characterized by high acoustic variability, anxiety-induced prosodic shifts, and unpredictable spontaneous colloquialisms, which directly impact decoder precision.

While this research was conducted under low-resource constraints and bounded data availability, the empirical findings demonstrate the viability of automating transcription workflows for under-resourced languages in legal ecosystems. To build upon these promising baselines, future research will focus on expanding the acoustic corpus with a more diverse demographic of speakers to improve generalizability; evaluating alternative, noise-robust feature extraction techniques alongside modern deep learning frameworks; and implementing advanced linguistic smoothing and adaptation algorithms to better accommodate spontaneous, non-standard vocalizations in high-stakes legal settings.

Declarations

Ethical approval was obtained from ethics committee with Reference No. AKU-AIT:01/230/2026) for the study by Dr.Birhane Gebreyohannes, Mr.Hagos Hadgu and Ms.Eyerusalem Dagnew, Aksum University, Ethiopia.

Accordance

All procedures were conducted in accordance with the research ethics of Aksum University.

Competing interests

We declare that all the authors have no competing interests.

9.1 Ethical approval

The study involved human participants; therefore, informed consent was conducted in accordance with Ethiopian and international standards.

9.2 Consent to participate: Not Applicable

9.3 Consent to publication: All authors have read, approved and consented to its publication

Data availability

All data generated and analyzed during the research are included within this paper and Science Data Bank with a full name of the database “Automatic speech recognition system for court proceedings in Ethiopia”, and DOI https://doi.org/10.57760/sciencedb.42104 or https://cstr.cn/31253.11.sciencedb.42104

Funding: Not Applicable

Author contribution

Tedros Kebede Bidada and Dawit Teklu Weldeslasie drafted the study design, developed the system, and selected the analytical tools and methodology. Tedros Kebede Bidada conducted the experiments and gathered the data. Dawit Teklu Weldeslasie drafted the initial manuscript. All authors critically reviewed, edited, and provided final approval for the published version.

References

1. B G, N., G, T.Y. Automatic speech recognition: challenges, enhancements, and evaluation metrics. *Multimed Tools Appl* 84, 46627–46645 (2025). https://doi.org/10.1007/s11042-025-20998-6
2. Salifu, A., Mensah, H., Tchao, E. *et al.* Enhancing speech recognition through diverse shared features accent classification. *Int J Speech Technol* 28, 461–481 (2025). https://doi.org/10.1007/s10772-025-10198-w
3. Prem, D., Raimond, K. & Jeyabose, A. Investigating the Impact of Feature Extraction Techniques for Classification and Synthesis of Dysarthric Speech. *Circuits Syst Signal Process* 45, 3728–3761 (2026). https://doi.org/10.1007/s00034-025-03366-5
4. Sebastian, R., O' Keefe, S., Trefzer, M.A. (2026). Enhancing MFCC Feature Extraction Through Reservoir Computing. In: Formenti, E., Manzoni, L. (eds) Unconventional Computation and Natural Computation. UCNC 2025. Lecture Notes in Computer Science, vol 16364. Springer, Cham. https://doi.org/10.1007/978-3-032-15641-9_20
5. Beigi, H. (2011). Signal Processing of Speech and Feature Extraction. In: Fundamentals of Speaker Recognition. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-77592-0_5
6. Nyambo, D.G., Mdegela, L.N., Kihara, W. (2026). Assessment of Multiple ASR Architectures on Common Speech Datasets: Analyzing Training Duration, Framework, and Word Error Rate in the Development of a Swahili ASR Model. In: Marx Gómez, J., Gatera, A., Godfrey Nyambo , D. (eds) Advancement in Embedded and Mobile Systems. ICTA-EMOS 2024. Progress in IS. Springer, Cham. https://doi.org/10.1007/978-3-031-99219-3_26
7. Raja, D.N., Giri, K.J. AUDIRE: a comprehensive review of speech recognition technologies – methods, uses, and challenges. *Int J Speech Technol* 28, 903–929 (2025). https://doi.org/10.1007/s10772-025-10213-0
8. Tachbelie, M.Y. *et al.* (2024). DNN-Based Supervised Spontaneous Court Hearing Transcription for Amharic. In: Debelee, T.G., Ibenthal, A., Schwenker, F., Megersa Ayano, Y. (eds) Pan-African Conference on Artificial Intelligence. PanAfriConAI 2023. Communications in Computer and Information Science, vol 2068. Springer, Cham. https://doi.org/10.1007/978-3-031-57624-9_13
9. Ejigu, Y., Asfaw, T. & Tadesse, S. A noise-robust end-to-end framework for Amharic speech recognition. *Int J Speech Technol* 29, 14 (2026). https://doi.org/10.1007/s10772-025-10224-x
10. Alemu, A.A., Melese, M.D. & Salau, A.O. Ethio-Semitic language identification using convolutional neural networks with data augmentation. *Multimed Tools Appl* 83, 34499–34514 (2024). https://doi.org/10.1007/s11042-023-17094-y
11. Alameer, A., Jahromi, H.K., Afzal, Z., Malik, M., Morphy, S., Mansouri, T. (2025). A Comparative Analysis of State-of-the-Art Speech-to-Text Models for Court Applications. In: Palaiahnakote, S., Palit, R., Saraee, M., Atrey, P.K., Bai, X., Raman, B. (eds) Data Science, AI and Applications. ICDSAIA 2025. Communications in Computer and Information Science, vol 2683. Springer, Cham. https://doi.org/10.1007/978-3-032-11355-9_32
12. Trancoso, I., Mamede, N., Martins, B., Pinto, H.S., Ribeiro, R. (2024). The Impact of Language Technologies in the Legal Domain. In: Sousa Antunes, H., Freitas, P.M., Oliveira, A.L., Martins Pereira, C., Vaz de Sequeira, E., Barreto Xavier, L. (eds) Multidisciplinary Perspectives on Artificial Intelligence and the Law. Law, Governance and Technology Series, vol 58. Springer, Cham. https://doi.org/10.1007/978-3-031-41264-6_2
13. Koulu, R. In Search of Living Law: How Should We Design for (Digital) Legal Interaction?. *DISO* 3, 31 (2024). https://doi.org/10.1007/s44206-024-00117-y
14. McCrocklin, S. (2025). Automatic Speech Recognition. In: McCallum, L., Tafazoli, D. (eds) The Palgrave Encyclopedia of Computer-Assisted Language Learning. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_30-

15. Morzy, M. (2025). Automatic Speech Recognition. In: Spoken Language Processing. Studies in Computational Intelligence, vol 1205. Springer, Cham. https://doi.org/10.1007/978-3-031-88566-2_4
16. D. Appleyard, Tigrinya, Encyclopedia of Language & Linguistics (Second Edition), Elsevier, 2006, Pages 715-717, ISBN 9780080448541, https://doi.org/10.1016/B0-08-044854-2/02091-5
17. Mihreteab, M., Kidane, S., Fisshaye, M., Teklemariam, F. (2026). Towards End-to-End Speech Synthesis for Tigrinya Language. In: Chakravarthi, B.R., *et al.* Speech and Language Technologies for Low-Resource Languages. SPELLL 2024. Communications in Computer and Information Science, vol 2656. Springer, Cham. https://doi.org/10.1007/978-3-032-05855-3_2
18. Ghebregiorgis, B.D., Tekle, Y.Y., Kidane, M.F., Keleta, M.K., Ghebraeb, R.F., Gebretatios, D.T. (2024). Tigrinya End-to-End Speech Recognition: A Hybrid Connectionist Temporal Classification-Attention Approach. In: Debelee, T.G., Ibenthal, A., Schwenker, F., Megersa Ayano, Y. (eds) Pan-African Conference on Artificial Intelligence. PanAfriConAI 2023. Communications in Computer and Information Science, vol 2068. Springer, Cham. https://doi.org/10.1007/978-3-031-57624-9_12
19. Zhang, H., Shang, J. (2025). Speech Text Recognition. In: Natural Language Processing and Applications. Springer, Singapore. https://doi.org/10.1007/978-981-97-9739-4_7
20. Devi, T.C., Thaoroijam, K. & Nongmeikapam, K. Multilingual speech representation for the Manipuri automatic speech recognition system. *Lang Resources & Evaluation* 60, 25 (2026). https://doi.org/10.1007/s10579-026-09908-w
21. Nguyen, V.D., Do, A.T., Le, D.K., Nguyen, L.V. (2025). Speech Recognition-Based Human–Computer Interaction: A Survey. In: Lanka, S., Cabezuelo, A.S., Vuppalapati, C. (eds) Trends in Sustainable Computing and Machine Intelligence. ICTSM 2024. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-96-1452-3_21
22. Sharma, N., Jamwal, S.S. (2026). Spoken Keyword Recognition Using Gaussian Mixture Models and MFCC Features. In: Jaiswal, A., Anand, S., Hassanien, A.E., Azar, A.T. (eds) Proceedings of International Conference on Computing Systems and Intelligent Applications. ComSIA 2025. Lecture Notes in Networks and Systems, vol 1501. Springer, Singapore. https://doi.org/10.1007/978-981-96-8350-5_20
23. Mujtaba, D.F., Mahapatra, N.R. (2025). Automatic Speech Recognition APIs: Models, Features, and Challenges. In: Arabnia, H.R., Deligiannidis, L., Shenavarmasouleh, F., Amirian, S., Ghareh Mohammadi, F. (eds) Computational Science and Computational Intelligence. CSCI 2024. Communications in Computer and Information Science, vol 2503. Springer, Cham. https://doi.org/10.1007/978-3-031-94940-1_3
24. Nardos Alemu, Chelsea Hua, Phuc H. Le, Khoi P. N. Nguyen, Melat Ali, and Nanette Veilleux. 2023. A Framework to Develop Automatic Speech Recognition for Low Resource Languages. In Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE 2023). Association for Computing Machinery, New York, NY, USA, 1228. https://doi.org/10.1145/3545947.3573271
25. Elias N. Stebek. (2025). Access to Legal Information in Ethiopia: Overview of Attainments and Prospects. Mizan Law Review, 19(2), 175–216. https://doi.org/10.4314/mlr.v19i2.1
26. Ejigu, Y., Asfaw, T. & Tadesse, S. A noise-robust end-to-end framework for Amharic speech recognition. Int J Speech Technol 29, 14 (2026). https://doi.org/10.1007/s10772-025-10224-x
27. MF Temesgen "TEXT-BASED LANGUAGE IDENTIFICATION FOR TYPOLOGICALLY RELATED ETHIOPIAN LANGUAGES: A DEEP LEARNING APPROACH", 2025.
28. Daniel Tesfai Gebretatios, Huansheng Ning, Doreen Sebastian Sarwatt, et al. Advancing Tigrinya Multiclass Text Classification with Pre-trained Models: A Comparative Analysis. TechRxiv. 22 July 2025.DOI: https://doi.org/10.36227/techrxiv.175322819.98359474/v1
29. Augustyniak, Z. (2026). The Sources of Ethiopian Law. In: Marriage in the Ethiopian Empire. Studies in the History of Law and Justice, vol 33. Springer, Cham. https://doi.org/10.1007/978-3-032-26065-9_4
30. BS Goshu "Adaptive Learning with Artificial Intelligence in the Ethiopian Education System: A Systematic Review of Opportunities, Implementation Barriers, and Strategic …", International Journal of Business, 2026.

31. DM Asgede "Backchannels in Tigrinya", Ethiopian Journal of Business and Social Sciences, 2023.
32. Seckler, C., Mauer, R., vom Brocke, J. *et al.* Design Science Across Disciplines: Building Bridges for Advancing Impactful Business Research. *Schmalenbach J Bus Res* 77, 719–761 (2025). https://doi.org/10.1007/s41471-025-00233-6
33. Foong, C. Content Curation and Disseminators' Primary Copyright Liability in Access-Based Attention Markets. *IIC* 57, 33–66 (2026). https://doi.org/10.1007/s40319-025-01669-3
34. Sutherland, K.E. (2024). Ethical Content Curation. In: Strategic Social Media Management. Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-99-9496-0_10

Figures

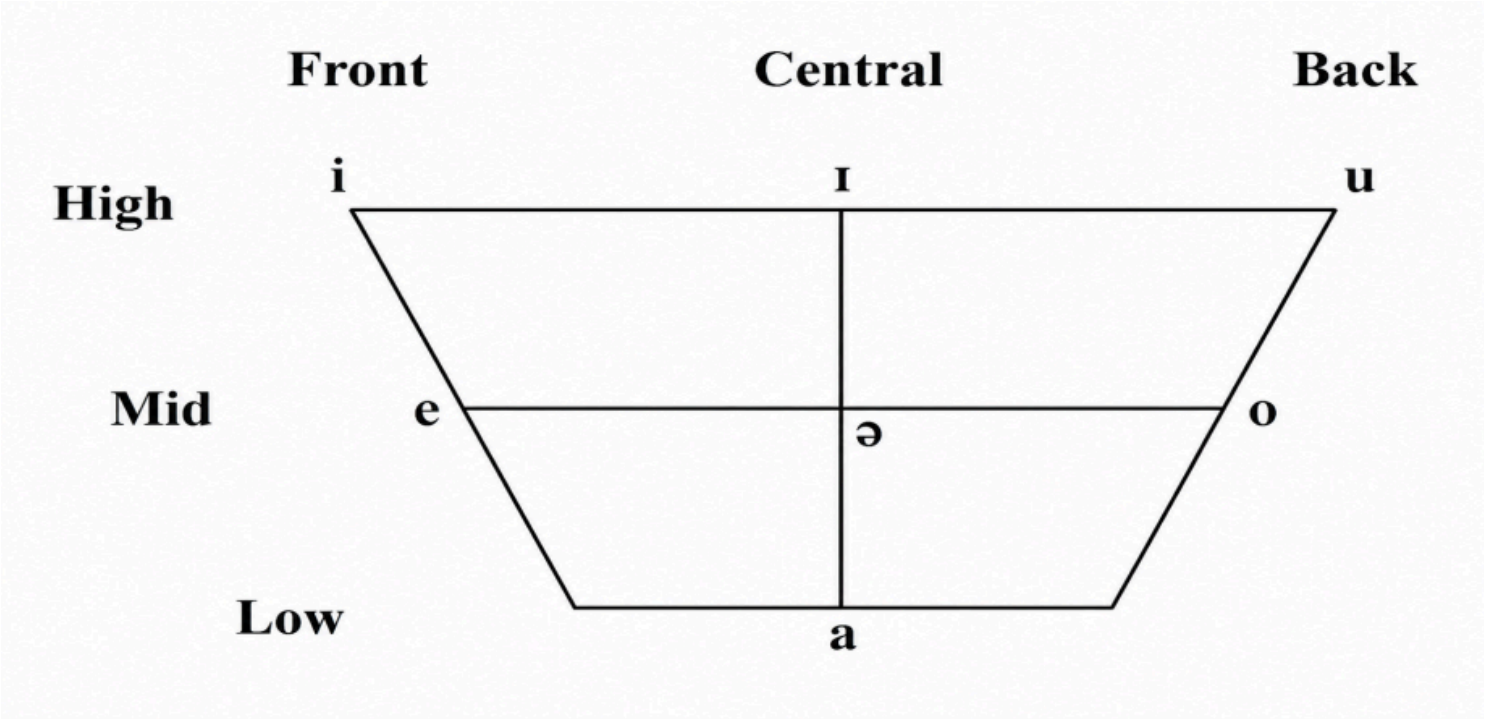


Figure 1

Vocalic inventory of Tigrigna vowels

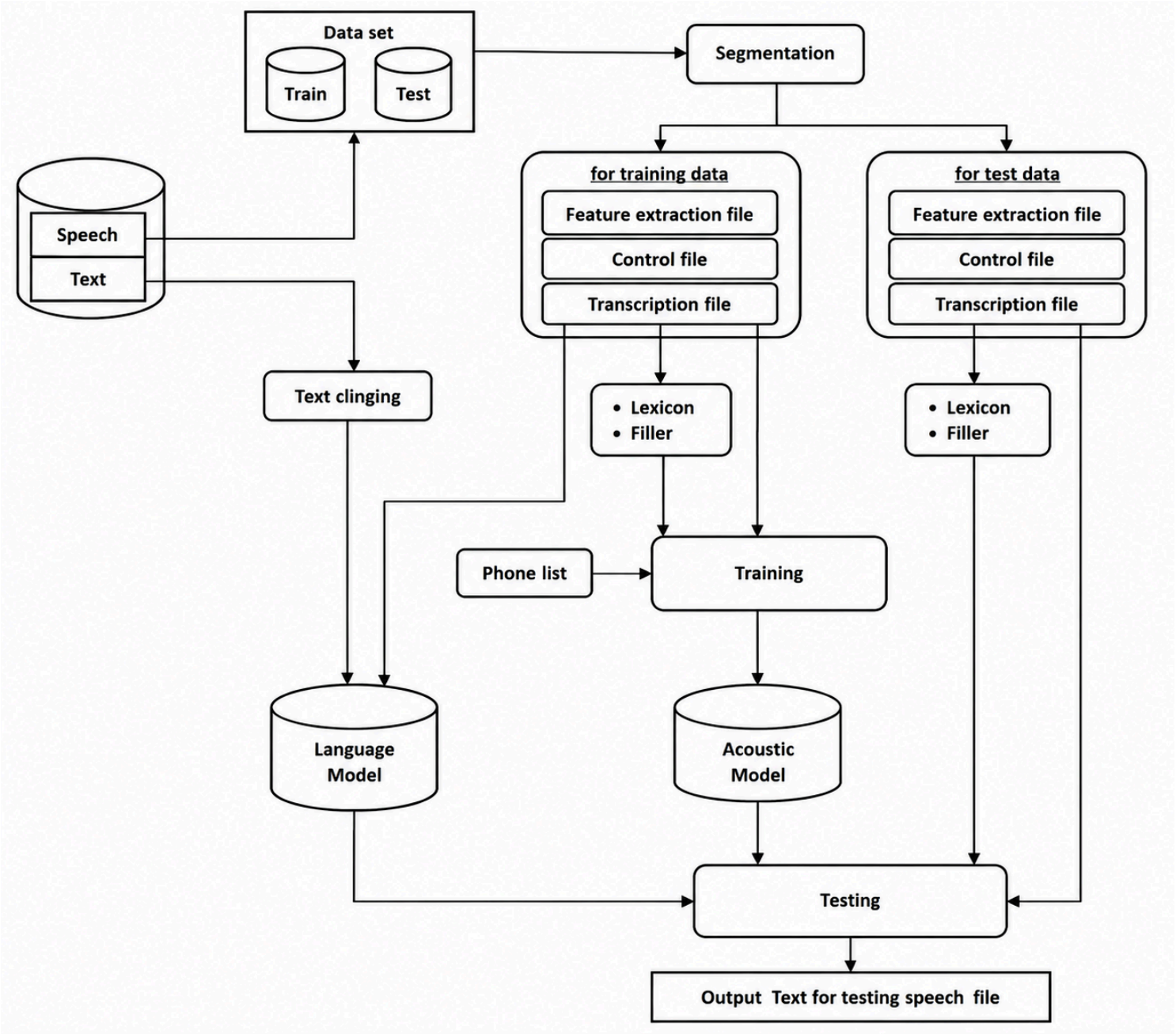


Figure 2

the process model of the automatic speech recognition system

መከላኸሊ ምስክር ተባሂለ እየ ተፀዊዐ (TRAIN-TSM11-008)
መከላኸሊ ምስክር ተባሂለ እየ ተፀዊዐ (TRAIN-TSM11-008)
Words: 5 Correct: 5 Errors: 0 Percent correct = 100.00% Error = 0.00% Accuracy = 100.00%
Insertions: 0 Deletions: 0 Substitutions: 0

እዚ ፅንዓት እዚ ከዓ ከይተማለአ ናብ መዘጋጃ ቤት መፅዩ እዩ (TRAIN-TSM28-005)
እዚ ፅንዓት እዚ ከዓ ከይተማለአ *** *** እናመዝገባ መፅዩ *** (TRAIN-TSM28-005)
Words: 10 Correct: 6 Errors: 4 Percent correct = 60.00% Error = 40.00% Accuracy = 60.00%
Insertions: 0 Deletions: 3 Substitutions: 1

ምናልባት ብኣኻትኩም ተጠየቕክዎ ጥያቄ ዝፈልጦ ተሃልዩ ክምልስ እየ (TRAIN-TSM11-011)
ምናልባት ብኣኻትኩም ተጠየቕክዎ ጥያቄ ዝፈልጦ ተሃልዩ ክምልስ ኢልና (TRAIN-TSM11-011)
Words: 8 Correct: 7 Errors: 1 Percent correct = 87.50% Error = 12.50% Accuracy = 87.50%
Insertions: 0 Deletions: 0 Substitutions: 1

ሲሒብና ምስ ናይ *** ካቢኔ ኣማኸርቲ ኣቃብቲ ሕጊ ተማያይጥና (TRAIN-TSM28-023)
ሲሒብና ምስ ናይ ካሊእ ድማ ካብታ ዓቃቢ ሕጊ ተማያይጥና (TRAIN-TSM28-023)
Words: 8 Correct: 5 Errors: 4 Percent correct = 62.50% Error = 50.00% Accuracy = 50.00%
Insertions: 1 Deletions: 0 Substitutions: 3

Figure 3

sample interface of the system

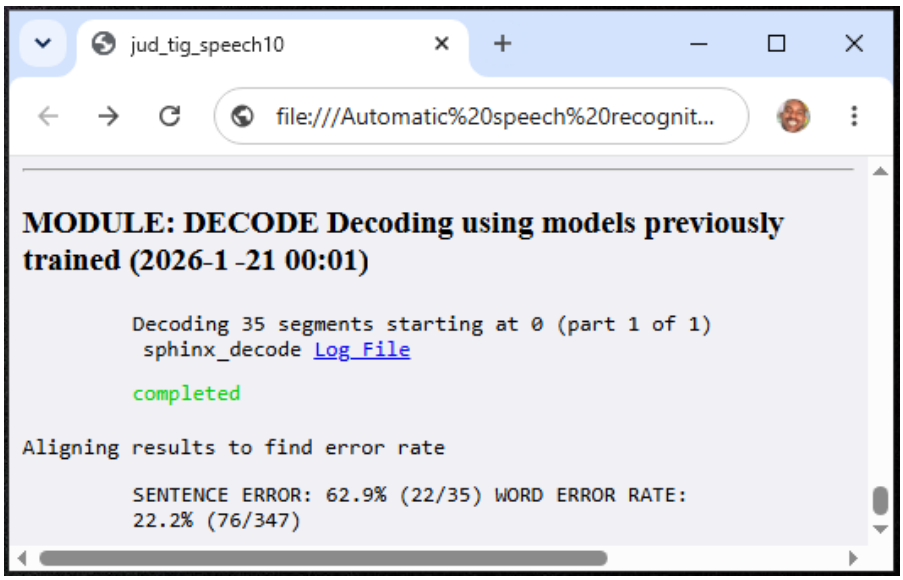


Figure 4

GUI showing system results