Isolated Urdu Word Recognition: Native and Non-Native Speaker Perspectives

Shalini V[1]. Sathe, Dr. R. R. Deshmukh2

[1,2]Department of Computer Science & Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

[1]shalinisathe55@gmail.com
[2]rrdeshmukh.csit@bamu.ac.in

Abstract

This study addresses isolated Urdu word recognition by encompassing both native and non-native speakers. Despite significant advancements in language technology, Urdu remains neglected, particularly in isolated word identification. This research aims to bridge this gap through a comprehensive analysis, creating a sophisticated recognition system tailored to Urdu's linguistic nuances. The primary goal is an efficient isolated word recognition system capable of accurately deciphering spoken Urdu words from both native and non-native speakers. The study dedicates significant effort to curating a diverse dataset, incorporating various accents, pronunciations, and speaking styles prevalent in Urdu. This dataset forms the cornerstone of the recognition system. The research navigates the challenges posed by both speaker groups, customizing the recognition system accordingly. By leveraging advanced machine learning techniques and signal processing, the aim is to achieve high accuracy and robustness in recognizing isolated Urdu words. Beyond language technology, the research has broad implications in education, accessibility, and communication. A robust recognition system could facilitate language learning, enhance human-computer interaction, and bridge linguistic barriers for non-native speakers. Ultimately, this study enriches Urdu language technology by focusing on isolated word recognition, combining linguistic insights, advanced technology, and diverse datasets to foster a more inclusive and effective interaction between individuals and technology in the Urdu language.

This study focuses on a regional language, benefiting non-native Urdu speakers through a speech interface system. Its main objectives include addressing issues in current speech devices, predicting speaker nativity, and recognizing spoken words. The technology's potential extends to diverse languages worldwide, enabling seamless cross-language communication. Refinements could improve analysis of prosodic features for accurate language identification and enhanced speech recognition. Progress in speech-to-text and text-to-speech conversion could enhance virtual assistants, transcription services, and accessibility tools, benefiting various users.

1. Introduction

Speech is the most prominent and natural form of communication between humans. There are various spoken languages throughout the world. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer. Speech has potential of being used as a mode of interaction with computer. Human beings have long been motivated to create computer that can understand and talk like human. In this direction, researchers have tried to develop system for analysis and classification of the speech signals (Shrishrimal P. P., et. al. 2012).

The computers System which can understand the spoken language can be very useful in domains like agriculture, health care and government services. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages.

There are many areas where Automatic Speech Recognition (ASR) systems can play a pivotal role in facilitating the daily activities, and where the current levels of accuracy that these systems have attained can prove useful. One of these areas is speech based Human Computer Interaction (HCI). This line of research promises to be of significant advantages in areas where keyboards may not be appropriate and natural language communication is desired. This includes control applications where hands and eyes may be busy at the same time and speech becomes a good means of issuing the commands. In addition to this, such systems can be of immense use for people with vision related disabilities, lack of motor control, crippled handset. In the under developed countries where literacy rate is poor, this can provide a mechanism of information access to people who are unable to read and write as well as people who may belite rate but not qualified in computing skills. Speech based HCI ideally brings computers within reach of anyone who can speak and listen (Kumar, Y. et. al. 2019).

1.1 Native Speaker: A native speaker is someone who learned to speak language as part of his or her childhood development. A native speaker's language is usually their parent or country.

1.2 Non –Native speaker: non-native speakers of language on the other hand are people who have learned this particular language as second or third language.

Automatic Speech Recognition (ASR) systems have achieved impressive accuracy levels, benefiting various sectors by optimizing daily operations. Notably, speech-based human-computer interaction (HCI) offers a promising avenue, particularly in situations where traditional keyboards are impractical, and natural language interaction is preferred. This is valuable for tasks requiring multitasking, as speech commands can be issued hands-free. Furthermore, individuals with motor impairments or visual disabilities can benefit greatly from this technology. In low-literacy regions, speech-based HCI offers access to computing for illiterate or computer-inexperienced individuals. Despite its potential, limited resources for indigenous languages remain a challenge. ASR systems, commonly used in telephones, can recognize numerals and simple instructions, making them accessible tools for diverse populations. This technology democratizes computer access through speech and hearing capabilities, overcoming barriers to digital inclusion.

The topic of speech recognition in Urdu is exciting yet understudied. There haven't been many efforts made to create frameworks for deciphering Urdu speech, though. Both native and non-native speakers of Urdu are involved in this study project. Recognition methods have utilized MFCC and HMM.

2 Automatic Speech Recognition System (ASR)

The act of converting audible words into written ones is called speech recognition (SR). Additionally, I go by the names "automatic speech recognition, or ASR," "computer speech recognition," and "speech to text, or STT." The process of converting a speech signal into a string of words using a computer software that applies the necessary algorithm is known as speech recognition, often known as automatic speech recognition (ASR) or computer speech recognition. (Shaikh Naziya, et. al. 2016).

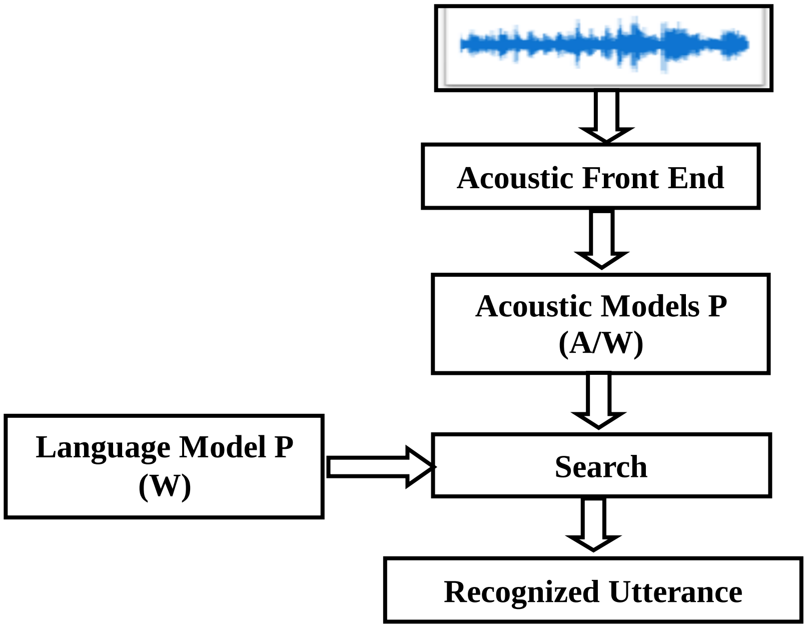


Fig. 1. Basic Model of Speech Recognition

2.1 Types of Speech Recognition

They are three styles of speech: isolated, connected and continuous. Isolated speech recognition systems can just handle words that are spoken separately. This is the most common speech recognition systems available today. The user must pause between each word or command spoken.

Speech recognition systems can be categorized into numerous groups based on the utterances they can recognize. The following categories apply to these classes:

2.2 Isolated Word

Isolated word recognition excels in single-word or utterance scenarios, enabling efficient user responses and command issuance. It's straightforward due to clear word boundaries and distinct pronunciations. Yet, its simplicity falters when handling multi-word demands. Altered boundaries can impact outcomes, representing a limitation of this approach.

2.3 Connected Words

A connected words system allows different assertions to be "run-together" with only a brief interval in between, similar to isolated words. A word or set of words that have a single meaning to a computer are vocalized as an utterance.

2.4 Continues Speech

Users of continuous voice recognition systems can speak almost naturally as the computer analyses the substance of their speech. It is basically digital dictation, to put it simply. Words flow into one another in this closest without any gaps or word breaks. The development of a system for continuous voice recognition is difficult.

2.5 Spontaneous Speech

Systems for spontaneous voice recognition can identify natural speech. It's typical for speech to come out of the mouth suddenly. One of the numerous aspects of genuine speech that an ASR system with spontaneous speech can manage is word runs. When speaking spontaneously, mistakes in pronunciation, false starts, and non-words can all happen.

3 Types of Speaker Models

Each speaker has a distinctive voice due to his unique physical attributes and personality. Based on these traits, the system's two main classes are established.

3.1 Speaker Dependent Model

Speaker-specific, the speaker-dependent model. These models are easier to use and less expensive to put into practice. It yields a more accurate result for one speaker while yielding a less accurate result for other speakers. Speaker dependent systems are those created for a certain type of speaker. While they might be less precise for some speakers, they are usually more accurate for the specific speaker. These systems are often simpler to construct, more affordable, and more precise. However, these systems are less flexible than speaker-independent systems.

3.2 Speaker Independent Model

Speaker Independent System is able to recognize a variety of speakers without any prior training. A system that is speaker-independent has been developed to function with any type of speaker. It is used in Interactive Voice Response Systems (IVRS), which need to receive input from a variety of users. The reduction in the quantity of words one can know is the sole drawback. The implementation of this system is the most difficult part. Additionally, it is more expensive and less accurate when compared to speaker independent systems (Saksamudre, S. K. et. al. 2015).

4 About Urdu Language

Urdu, a language steeped in a rich historical and cultural legacy, resonates with over 70 million individuals as their first language and another 100 million as a second language, primarily within the heartlands of Pakistan and India. This linguistic tapestry, woven through centuries, traces its roots back to the 12th century, unfurling in the northern expanse of the Indian subcontinent where it absorbed the intricate influences of Arabic, Persian, and Turkish. Sharing an ancestral lineage with Hindi, the two languages, while bound by a common history, diverge in the choice of script and vocabulary. Urdu's elegant script, a variant of the Persian Nastaliq, imbues it with a calligraphic allure, while its lexicon, enriched by the contributions of diverse cultures, testifies to its dynamic evolution. The pivotal year of 1947 witnessed the emergence of Pakistan as an independent nation, and it was in this crucible of change that Urdu ascended to the role of Pakistan's national language. Yet, its reach extended beyond political borders, spanning oceans and continents. Across the British Isles, Canada, the United States, and the Middle East, Urdu found resonance in diasporic communities, its lyrical cadence echoing through homes and gathering places.

However, Urdu's global presence surpasses the confines of Pakistan's borders. Remarkably, more speakers of Urdu reside in India than in Pakistan, a testament to the enduring legacy of this language in a region that transcends political demarcations. Yet, for non-native speakers, the intricacies of Urdu present a captivating challenge. Its phonetic subtleties, enigmatic idioms, and nuanced expressions demand a dedicated exploration, revealing layers of depth that connect individuals across linguistic frontiers. Urdu’s intricate evolution encapsulates the narrative of a language that has seamlessly woven itself into the fabric of diverse cultures. It symbolizes the bridge that spans temporal and spatial divides, uniting generations and regions. Through its mellifluous verses and resonant prose, Urdu not only preserves history but also serves as a conduit for the exchange of ideas, emotions, and aspirations. Its enduring significance underscores its role as a cultural treasure, a source of unity amid diversity, and a timeless testament to the power of language (Shaikh Naziya, et. al. 2017).

5 Urdu Language Related work

Urdu Language literature survey on different languages as follow:

Sr. No.	Title	Author and year	Techniques	Result
1	LPC and HMM Performance Analysis for Speech Recognition System for Urdu Digits	Shaikh Naziya S et.al. 2017	LPC, HMM	100%
2	Isolated English Words Recognition Spoken by Non-Native Speakers	Mr. V. K. Kale et.al. 2014	LPC, MFCC	95.75 % for MFCCs and 61.40 % for LPC.
3	Isolated Word Recognition System for Hindi Language	Suman K. Saksamudre et.al. 2015	MFCC, KNN CLASSIFIER	89%
4	Automatic Speech Recognition and Verification using LPC, MFCC and SVM	Aaron M. Oirere, et.al. 2015	MFCC, LPC SVM, LDA	For numeric data- MFCC-75%, LPC-72% Isolated word- MFCC-65.2%, LPC-66.67% Sentence data- MFCC-63.8% ,LPC-59.6%
5	Automatic Speech Recognition Of Urdu Digits With Optimal Classification Approach	Hazrat Ali, An Jianwei et.al. 2015	MFCC, SVM, LDA	MFCC-73% LDA-63%
6	Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling Approach	M. U. Akram et.al. 2004	MFCC, ANN,	54%
7	Marathi Digit Recognition System based on MFCC and LPC Features	Pukhraj P. Shrishrimal et.al. 2017	LPC, MFCC	MFCC-78.94%, LPC-66.17%

6 Design and Development:

6.1 Acquisition Environment of speaker and Instrumental setup:

- We collect the speech data from native and non- native people of Urdu language. All speakers are from Aurangabad district.
- The utterances were captured in mono sound at a sample rate of 16000Hz (.wav files), PRAAT software.
- We used Sennheiser HD450 microphones for audio recording.
- The speaker's mouth was around 5 cm away from the microphone. Each speaker was asked to utter a word from the produced text corpus. Each word is spoken five times.
- Selection of Native & Non-Native Speaker for Urdu Language Database

Both native and non-native Urdu speakers contributed speech data. Non-native speakers were those whose first language differed, while native speakers had Urdu as their first language. A balanced representation of genders and language backgrounds was ensured, including men and women, natives, and non-natives. Speaker ages (20-40) were randomly selected. Participants read phonetically balanced words to assess comfort and proficiency. Data was sourced from Marathwada speakers in Maharashtra's Aurangabad region.

6.2 Data Collection

The database consists of two parts: a numeric speech dataset covering digits zero to nine, and a days-of-the-week dataset covering Monday to Sunday. Additionally, 17 words from two categories were chosen for speech corpus development. Both native and non-native speakers recorded each word three times.

6.3 Corpus Text Selection

The suburban Urdu corpus integrates both read and spontaneous speech for effective speech recognition, benefiting from phonemic balance and coverage. This approach expedites corpus development, particularly for Urdu, a low-resource language, where spontaneous speech collection is challenging. Including word pronunciation translations aids non-native speakers' comprehension. The selected everyday terms for the corpus align with system design and development.

6.4 Numeric Corpus

The basic elements of any numbering system are its digits or numbers. Since Siffar (Zero) through Nau (Nine) play a key role in the number system, the most common numbers have been taken into consideration for the corpus creation. Ten digits in all have been taken into account for the corpus. In

Table No. 1, the corpus of numerical discourse is shown.

6.5 Days in a Week Corpus

The days from Pyir (Monday) to Itwaar (Sunday) have been taken into consideration for the corpus development because the terminology for the days of the week are ubiquitous. For corpus development, the complete seven-day workweek has been considered. Table No.2 displays the Days in a Week corpus selection.

Number	English	Urdu	Pronunciation	
0	Zero	۰	□□□□	Sifar
1	One	۱	□□	Aik
2	Two	۲	□□	Do
3	Three	۳	□□□	Teen
4	Four	۴	□□□	Chaar
5	Five	۵	□□□□	Paanch
6	Six	۶	□	Chha
7	Seven	۷	□□□	Saat
8	Eight	۸	□□□	Aanth
9	Nine	۹	□□	Nau

Table 1. Numeric 0-9 Speech Corpus

Sr. No.	English	Urdu	Pronunciation	
1	Monday	پیر	□□□	pyir
2	Tuesday	منگل	□□□□	Mangal
3	Wednesday	بدھ	□□□	Budh
4	Thursday	جمعرات	□□□□□□□	Jumeraat
5	Friday	جمعہ	□□□□	Jumaah
6	Saturday	ہفتہ	□□□□□	Sanichar
7	Sunday	اتوار	□□□□□	Itwaar

Table 2. Days in a Week Speech Corpus

6.6 Data Collection Statistics

Speech data was collected from 60 participants, evenly split between 30 native and 30 non-native speakers. Each speaker contributed 3 utterances, totaling 180 words per speaker. In total, there are

3060 recorded word utterances. This dataset includes 15 male and 15 female speakers from both native and non-native groups. The metadata details are outlined in Table 3.

Process	**Description**
Total No. of words selected	17
Utterances recorded	Three utterance of each word
Total utterance per speaker	51
Total speaker	60
Native Speaker	30
Non–native speaker	30
Native-male speaker	15
Native Female speaker	15
Non-Native male Speaker	15
Non-native female speaker	15
Total native speaker utterances	1530
Total non-native people's utterances	1530
Total utterances	3060
Tools	Sennheiser HD 450 Microphone
Recording Frequencies	16000HZ

Table 3. Shows the information about the Metadata.

7 Problem Faced During the Corpus Development

Obtaining accurate information posed a major challenge in the research. Creating the text corpus was time-consuming, and Urdu's phonetic nature required meticulous error checking. Teaching non-native speakers Urdu and ensuring clear enunciation were crucial. Recording data from non-native speakers was tough due to their distinct language use. Convincing speakers to allocate time for recording was challenging, leading to multiple sessions. Adjusting recording sessions to speakers' schedules was necessary. To ensure accurate samples, words were repeated three times during recording.

8. Lexical Tone: Articulation plays a crucial role in developing a unique dialect accent. Analyzing the variation in tone becomes essential for classifying different dialects. The pitch contour over the duration of pronunciation indicates the presence of tonal diversity, with low and high variability. Figure no. 2 and Figure no. 3 represent graphical representations of the tonal variation observed in native male and non-native male wave files for a specific sentence.

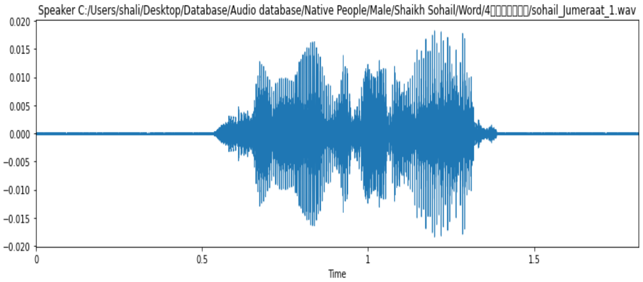


Fig.2 Native Speech wave form

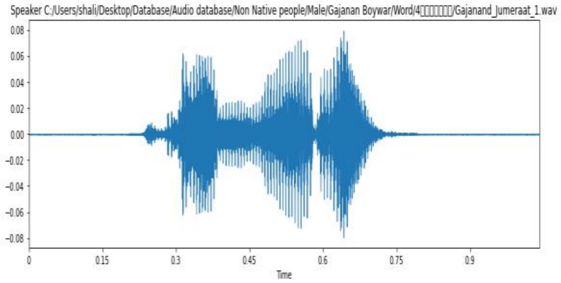


Fig.3. Non-Native Speech wave form

8.1 Feature Extraction

Phonetic qualities such as style, phonation, loudness dynamics, and flow of speech can be used to identify a person's speaking style. Speakers' linguistic qualities can be compared based on these characteristics. To examine the dialectal effect on individual speaking style, auditory phonetic analysis and spectrographic analysis of recorded samples for all dialects were performed. C1VC2 syllables were extracted to analyze the samples. Every study has been conducted with Urdu speakers in mind.

8.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a widely used feature extraction technique for speech processing and recognition. The MFCC technique involves the following steps:

- Pre-processing: The input speech signal is pre-processed by applying a pre-emphasis filter, which boosts the higher frequencies and reduces the lower frequencies. This helps in improving the signal-to-noise ratio.
- Frame blocking: The pre-processed speech signal is divided into small frames of typically 20-30 ms duration, with a 50% overlap between adjacent frames.
- Windowing: Each frame is multiplied with a window function, such as the Hamming window, to reduce spectral leakage and smooth the edges of the frame.

- Fast Fourier Transform (FFT): The windowed frames are converted to the frequency domain using the FFT algorithm to obtain the magnitude spectrum.
- Mel-frequency wrapping: The magnitude spectrum is then warped onto the Mel scale, which is a nonlinear scale that mimics the human auditory system's perception of frequency.
- Cepstral analysis: The Mel-scaled spectrum is transformed to the Cepstral domain using the Discrete Cosine Transform (DCT) to obtain the Mel-frequency Cepstral Coefficients (MFCCs). Typically, the first 12 MFCCs are used, which capture the most important spectral features.
- Energy coefficients: Finally, the log of the energy of each frame is computed, and the delta and double delta coefficients are also computed for the energy.
- Spectrographic Analysis: More minor speech sounds, such as vowels and consonants, combine to form syllables. Each word generates an utterance. Syllables are regarded as the fundamental processing unit for the Indian languages. Particular characteristics of the terms "superimposed" and "above" refer to the properties of spoken utterances in segments of Spectrographic examination of speech (Kurzekar, P. K. et.al.2014)

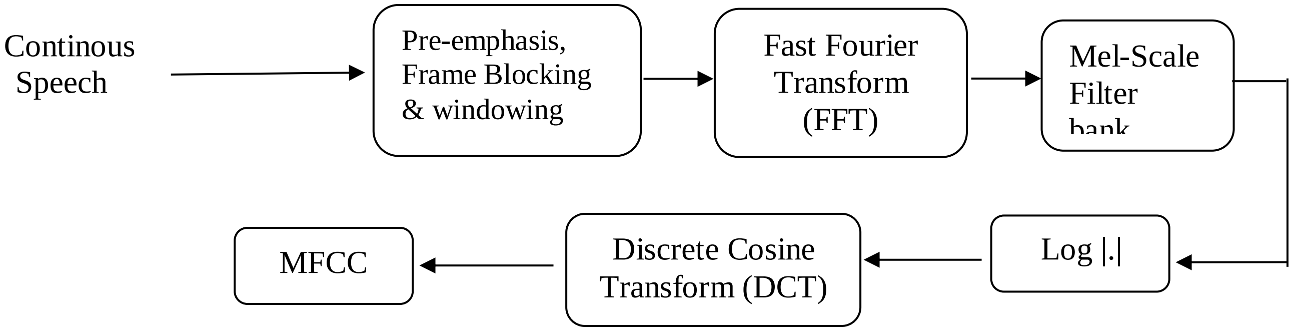

Continous Speech
Pre-emphasis, Frame Blocking & windowing
Fast Fourier Transform (FFT)
Mel-Scale Filter bank
Log |.|
Discrete Cosine Transform (DCT)
MFCC

Fig. 4. -Steps involved in MFCC feature extraction

The MFCC technique makes use of two types of filters, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel. The Mel scale is normally a linear mapping below 1000 Hz and logarithmically spaced above 1000 Hz (Shrishrimal, P. P. et.al. 2017).

Following Equation used to convert the normal frequency to the Mel scale the formula used is

$$Mel=2595\ log10\ (1+f/\ 700)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ (1)$$

8.3 HMM (Hidden Markov Model)

The basic idea is to define a HMM for each unit of speech, such as a phoneme or a word, and then concatenate them to form a larger HMM that represents a sentence or a vocabulary. A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters; the challenge is to determine the hidden parameters from the observable data. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states (Shaikh Naziya et.al.2017).A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state. HMMs are more popular because they can be trained automatically and are simple and computationally feasible to use HMM considers the speech signal as quasi- static for short durations and models these frames for recognition Kurzekar, P. K. et.al.2014).

8.4 Prosodic Feature

Prosodic features are features that appear when we put sounds together in connected speech. It is as important to teach learners prosodic features as successful communication depends as much on intonation, stress and rhythm as on the correct pronunciation of sounds. It provides context, gives meaning to words, and keeps listeners engaged. Prosody involves emphasizing the right words, using voice pitch and modulation, and taking appropriate pauses.

Example: accent, rhythm, Tempo, pitch, and intonation, Energy are prosodic features

We are working on Tempo and Energy Prosodic feature in our research work.

- **Tempo: -**Tempo is a measure of the number of speech units of a given type produced within a given amount of time.
- **Energy: -**Energy is most important features to extract value of energy in each Speech Frame.

9 Results and Discussion:

In this study, we employed various inputs, including prosodic and speech spectral characteristics, pitch contour values, gender information, syllable duration, in spoken utterance, along with the speech Hidden Markov Model (HMM) features. Sufficient data was available for Urdu languages, enabling us to conduct the experiment effectively. The algorithm was trained using HMM in our collected dataset consisting of 30 male and 30 female speakers, encompassing both native and non-native speakers. The trained network, based on the HMM model, was capable of classifying input utterances into Native and non-native Speaker of Urdu language based on the provided input features. To explore the impact of spectral and prosodic characteristics on dialect digit and word recognition systems, we conducted the experiment in two stages. Initially, the HMM-based system was trained using 13 features Mel-frequency cepstral coefficient (MFCC) characteristics. Subsequently, we further trained the network by incorporating syllable duration and pitch contour variables, still within the HMM framework.

The inclusion of MFCC features derived from prosodic characteristics significantly enhanced the digit recognition score, achieving an impressive 75%. Moreover, the utilization of prosodic characteristics alone contributed to an improved recognition score of 82%. Notably, the system's performance exhibited a notable enhancement when trained with both spectral and prosodic information using the HMM model.

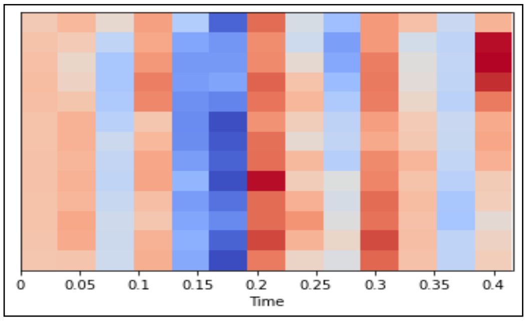


Fig.5.MFCC Spectrogram

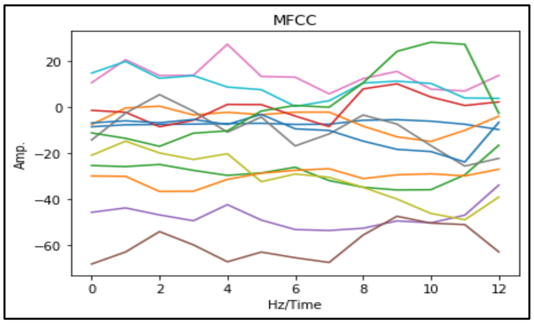


Fig.6. MFCC Feature

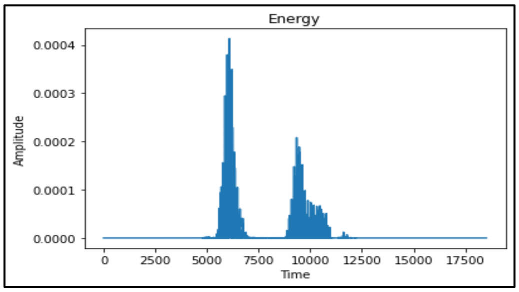


Fig.7. Energy

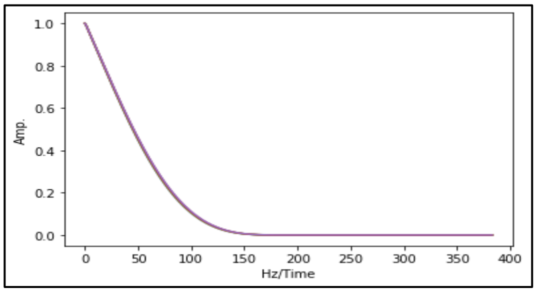
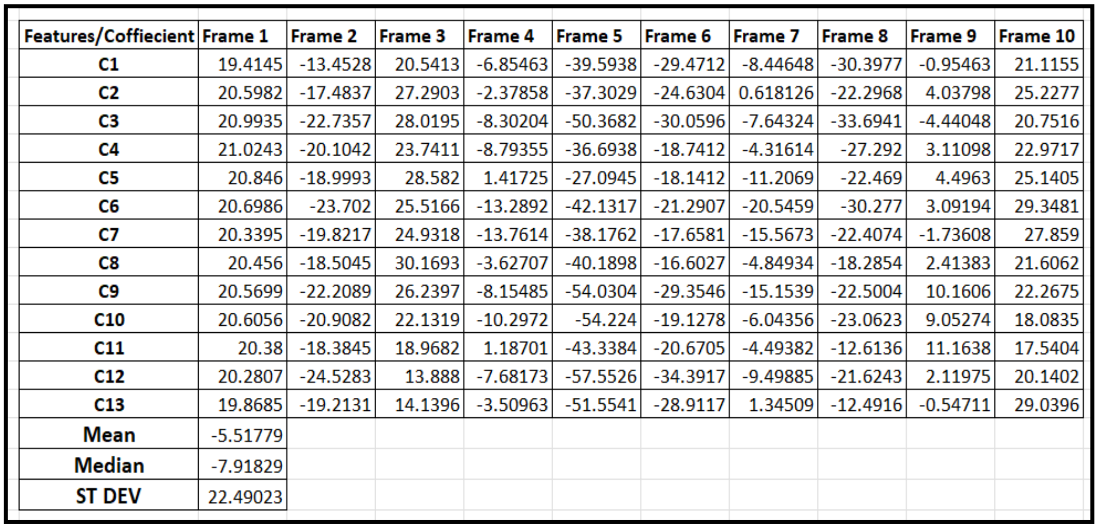


Fig.8. Tempo

From above figures, Fig. No.5 Shows the MFCC Spectrogram for Native female speaker sifer (सिफर) sample, Fig. No.7 Shows the Energy of that speech sample of Native female sifer (सिफर), 6. Shows the MFCC Feature and Fig No. 8 shows Tempo of that speech sample सिफर (0) of Native speaker.

MFCC Frame for □□□□ (0) of Female Native Speaker



Features/Coffiecient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
C1	19.4145	-13.4528	20.5413	-6.85463	-39.5938	-29.4712	-8.44648	-30.3977	-0.95463	21.1155
C2	20.5982	-17.4837	27.2903	-2.37858	-37.3029	-24.6304	0.618126	-22.2968	4.03798	25.2277
C3	20.9935	-22.7357	28.0195	-8.30204	-50.3682	-30.0596	-7.64324	-33.6941	-4.44048	20.7516
C4	21.0243	-20.1042	23.7411	-8.79355	-36.6938	-18.7412	-4.31614	-27.292	3.11098	22.9717
C5	20.846	-18.9993	28.582	1.41725	-27.0945	-18.1412	-11.2069	-22.469	4.4963	25.1405
C6	20.6986	-23.702	25.5166	-13.2892	-42.1317	-21.2907	-20.5459	-30.277	3.09194	29.3481
C7	20.3395	-19.8217	24.9318	-13.7614	-38.1762	-17.6581	-15.5673	-22.4074	-1.73608	27.859
C8	20.456	-18.5045	30.1693	-3.62707	-40.1898	-16.6027	-4.84934	-18.2854	2.41383	21.6062
C9	20.5699	-22.2089	26.2397	-8.15485	-54.0304	-29.3546	-15.1539	-22.5004	10.1606	22.2675
C10	20.6056	-20.9082	22.1319	-10.2972	-54.224	-19.1278	-6.04356	-23.0623	9.05274	18.0835
C11	20.38	-18.3845	18.9682	1.18701	-43.3384	-20.6705	-4.49382	-12.6136	11.1638	17.5404
C12	20.2807	-24.5283	13.888	-7.68173	-57.5526	-34.3917	-9.49885	-21.6243	2.11975	20.1402
C13	19.8685	-19.2131	14.1396	-3.50963	-51.5541	-28.9117	1.34509	-12.4916	-0.54711	29.0396
Mean	-5.51779									
Median	-7.91829									
ST DEV	22.49023									

Table No. 4 Show MFCC Frame for □□□□ (0) Female Native Speaker Sample with their Mean, Median and Standard Deviation

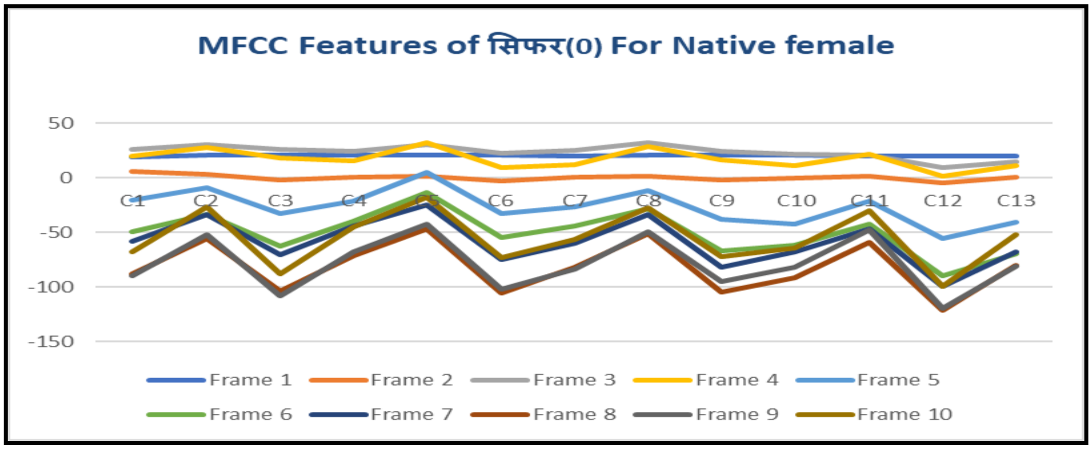
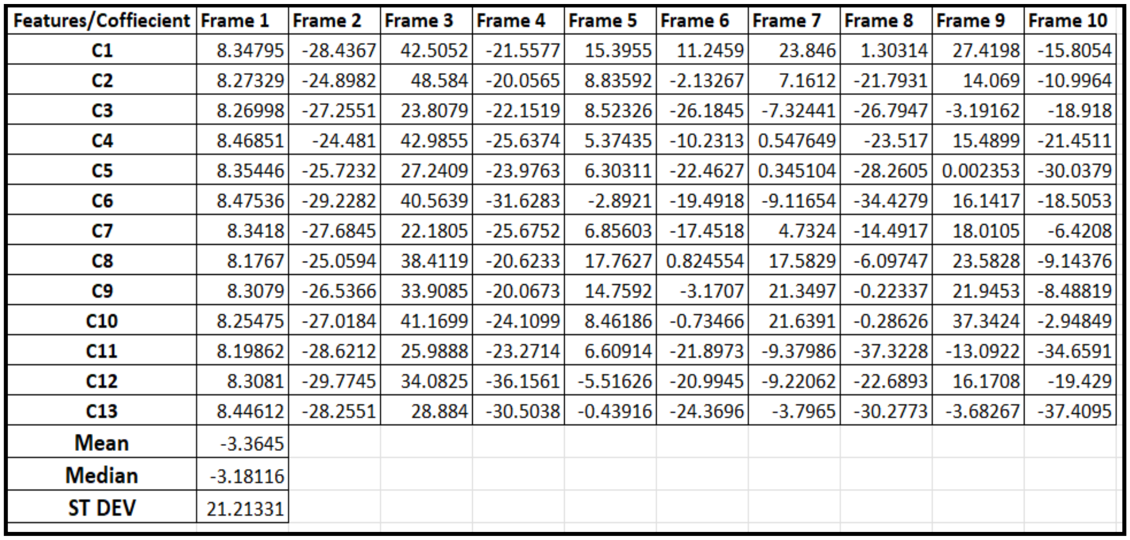


Figure. 9 Shows the plot of MFCC features for □□□□ (0) Female Native Speaker Sample

The Table No. 4 consists of the 13 features and 10 frames: the numbers of the frames calculated varies according to the speech signal length of Non-Native female sifer sample. The Mean, Median and the standard deviation for the complete MFCC were calculated for each utterance, we used them to analysis the performance of the features. The Figure No. 9 shows the plot of MFCC features for

10 frames of □□□□ (0) one person sample with Amplitude.

MFCC frame for □□□□ (0) Female Non –Native Speaker



Features/Coffiecient	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
C1	8.34795	-28.4367	42.5052	-21.5577	15.3955	11.2459	23.846	1.30314	27.4198	-15.8054
C2	8.27329	-24.8982	48.584	-20.0565	8.83592	-2.13267	7.1612	-21.7931	14.069	-10.9964
C3	8.26998	-27.2551	23.8079	-22.1519	8.52326	-26.1845	-7.32441	-26.7947	-3.19162	-18.918
C4	8.46851	-24.481	42.9855	-25.6374	5.37435	-10.2313	0.547649	-23.517	15.4899	-21.4511
C5	8.35446	-25.7232	27.2409	-23.9763	6.30311	-22.4627	0.345104	-28.2605	0.002353	-30.0379
C6	8.47536	-29.2282	40.5639	-31.6283	-2.8921	-19.4918	-9.11654	-34.4279	16.1417	-18.5053
C7	8.3418	-27.6845	22.1805	-25.6752	6.85603	-17.4518	4.7324	-14.4917	18.0105	-6.4208
C8	8.1767	-25.0594	38.4119	-20.6233	17.7627	0.824554	17.5829	-6.09747	23.5828	-9.14376
C9	8.3079	-26.5366	33.9085	-20.0673	14.7592	-3.1707	21.3497	-0.22337	21.9453	-8.48819
C10	8.25475	-27.0184	41.1699	-24.1099	8.46186	-0.73466	21.6391	-0.28626	37.3424	-2.94849
C11	8.19862	-28.6212	25.9888	-23.2714	6.60914	-21.8973	-9.37986	-37.3228	-13.0922	-34.6591
C12	8.3081	-29.7745	34.0825	-36.1561	-5.51626	-20.9945	-9.22062	-22.6893	16.1708	-19.429
C13	8.44612	-28.2551	28.884	-30.5038	-0.43916	-24.3696	-3.7965	-30.2773	-3.68267	-37.4095
Mean	-3.3645									
Median	-3.18116									
ST DEV	21.21331									

Table No. 5 Show MFCC Frame for □□□□ (0) Female Non-Native Sample with their Mean, Median and Standard Deviation

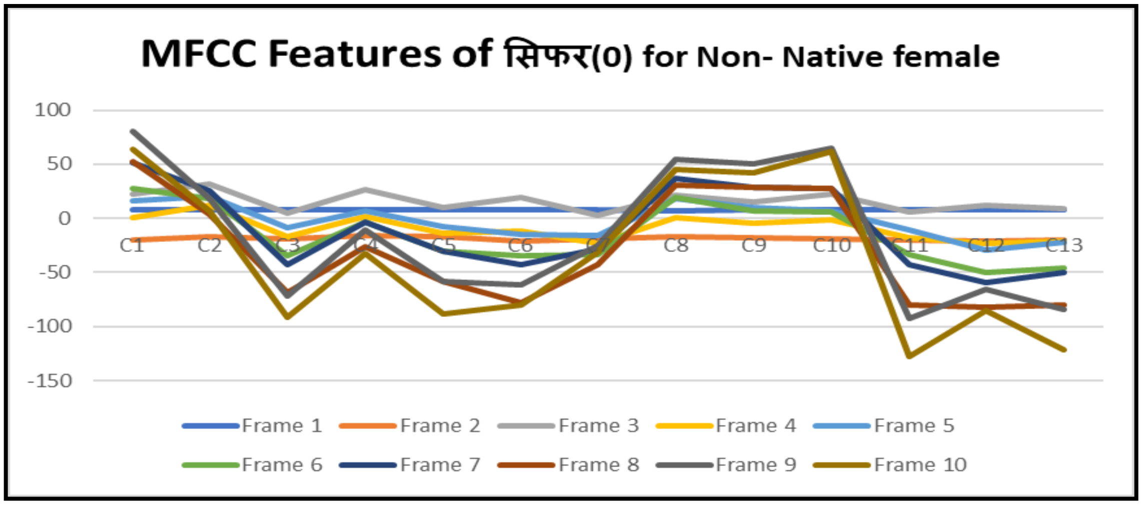


Figure No. 10 Shows the plot of MFCC features for □□□□ (0) Female Non-Native Speaker sample

The Table No.5 consists of the 13 features and 10 frames: the numbers of the frames calculated varies according to the speech signal length of Non-Native female sifer sample. The Mean, Median and the standard deviation for the complete MFCC were calculated for each utterance, we used them to analysis the performance of the features. The Figure No. 10 shows the plot of MFCC features for 10 frames of □□□□ (0) one person sample with Amplitude.

Conclusion

The research's main goal is to create an isolated word speech library for recognizing Urdu, catering to both native and non-native speakers. A database of isolated Urdu words was constructed, addressing the scarcity of such resources. The study uncovered the significance of language technologies for governance improvement. Challenges encompassed corpus evaluation, pronunciation for non-natives, and teaching Urdu to non-native speakers. Employing HMM for classification and MFCC for feature extraction, the voice database enabled a potent recognition system. An 82% accuracy was achieved using MFCC, shedding light on dialects and voice recognition in Urdu. The resulting library aids Urdu speech processing technology, benefiting applications like telecommunications, multimedia, customer care, and language learning.

The successful fusion of MFCC and HMM algorithms in Urdu speech recognition paves the way for exciting future avenues. Deep learning can unveil intricate patterns, while diverse linguistic and contextual cues enhance accuracy. Enriching the database with various dialects and demographics boosts real-world applicability. Real-time recognition, emotion detection, and cross-domain applications promise practical progress. User-centric interfaces, continuous speech recognition, and human-machine collaboration enhance experiences. This foundation sets the stage for a dynamic future in Urdu speech technology, spanning healthcare, education, and beyond.

References

Shrishrimal, P. P., Deshmukh, R. R., & Waghmare, V. B. (2012). Indian language speech database: A review. *International journal of Computer applications, 47(5), 17-21.*

Kumar, Y., & Singh, N. (2019, April). A comprehensive view of automatic speech recognition system-A systematic literature review. In 2019 *International conference on automation, computational and technology management (ICACTM) (pp. 168-173). IEEE.*

Shaikh Naziya, S., & Deshmukh, R. R. (2016). Speech recognition system—a review. IOSR J. *Comput. Eng, 18(4), 3-8.*

Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications, 115(22).*

Shaikh Naziya, S., & Deshmukh, R. R. (2017).LPC and HMM Performance Analysis for Speech Recognition Systemfor Urdu Digits. *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN, 2278-0661.*

Kale, M. V., Deshmukh, R. R., Janvale, G. B., Waghmare, M. V., & Shrishrimal, M. P. (2014). Isolated English Words Recognition Spoken by Non-Native Speakers.

Saksamudre, S., & Deshmukh, R. (2015). Isolated word recognition system for Hindi Language. *International Journal of Computer Sciences and Engineering, 3(7), 110-114.*

Oirere, A. M., Janvale, G. B., & Deshmukh, R. R. (2015). Automatic speech recognition and verification using lpc, mfcc and svm.

Ali, H., Jianwei, A., & Iqbal, K. (2015). Automatic speech recognition of Urdu digits with optimal classification approach. *International Journal of Computer Applications, 118(9), 1-5.*

Akram, M. U., & Arif, M. (2004, December). Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach. In 8th *International Multitopic Conference, 2004. Proceedings of INMIC 2004. (pp. 91-96). IEEE.*

Shrishrimal, P. P., Deshmukh, R. R., Janwale, G. B., & Kulkarni, D. S. (2017). Marathi Digit Recognition System based on MFCC and LPC. Reason, 2(67), 17-9.

Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology, 3(12), 18006-1801*