ISOLATED URDU WORD RECOGNITION: NATIVE AND NON-NATIVE SPEAKER PERSPECTIVES

Abstract

This study addresses isolated Urdu word recognition by encompassing both native speaker 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 speaker 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

Authors

Shalini Vijay Sathe
Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad, India.
shalinisathe55@gmail.com

Dr. R. R. Deshmukh
Department of Computer Science & Information Technology
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad, India.
rrdeshmukh.csit@bamu.ac.in

Copyright © 2024 Authors

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 text-to-speech andspeech-to-text conversion could enhance virtual assistants, transcription services, and accessibility tools, benefiting various users.

Copyright © 2024 Authors

I. INTRODUCTION

The most common and organic method of communication between people is speech. The world has many different spoken languages. Speaking is the primary mode of human communication, so it makes sense that people would expect speech interfaces for computers. Potentially, speech might be used to communicate with computers. Making a computer that can comprehend and communicate like a person has long been a goal of humankind. Researchers have worked to create a system for analyzing and categorizing voice signals in this direction (Shrishrimal P. P.,et. al. 2012).

Agriculture, healthcare, and government services can all greatly benefit from computer systems that can understand spoken language. Few people who can read or understand a given language have access to the majority of information in the digital world. Language technology can offer answers in the form of intuitive interfaces that will enable widespread distribution of digital content and promote communication among individuals who speak various languages. The present levels of accuracy reached by automatic speech recognition (ASR) systems can prove useful in numerous sectors where they can play a key role in facilitating daily operations.

Speech-based human-computer interaction (HCI) is one of these areas. This line of research holds great promise for applications where keyboards might not be suitable and natural language communication is preferred. This includes control applications where speaking commands can be used effectively because hands and eyes may be occupied at the same time. Additionally, people who have vision-related problems, poor motor control, or crippled devices might greatly benefit from such systems. This can give those who lack the ability to read and write, as well as those who may have a high literacy rate but lack computer skills, a way to access information in less developed nations with low literacy rates. Anyone who can speak and listen will be able to use computers thanks to speech-based HCI (Kumar, Y. et. al. 2019).=

1. **Native Speaker:** A natural speaker is someone who picked up a language throughout their early development. The mother tongue or native land of a native speaker.

2. **Non –Native speaker:** On the other hand, non-native speakers are those who have acquired this particular language as a 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.

Copyright © 2024 Authors

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 speaker and non-native speakers of Urdu are involved in this study project. Recognition methods have utilized MFCC and HMM.

II. AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR)

Speech recognition (SR) refers to the process of turning spoken words into written ones. I also go by the labels "ASR (automatic speech recognition)," "computer speech recognition," and "STT (speech to text)" Speech recognition, also known as automatic speech recognition (ASR) or computer speech recognition, is the act of transforming a speech signal into a string of words using computer software that applies the required algorithm. (Shaikh Naziya, et. al. 2016).

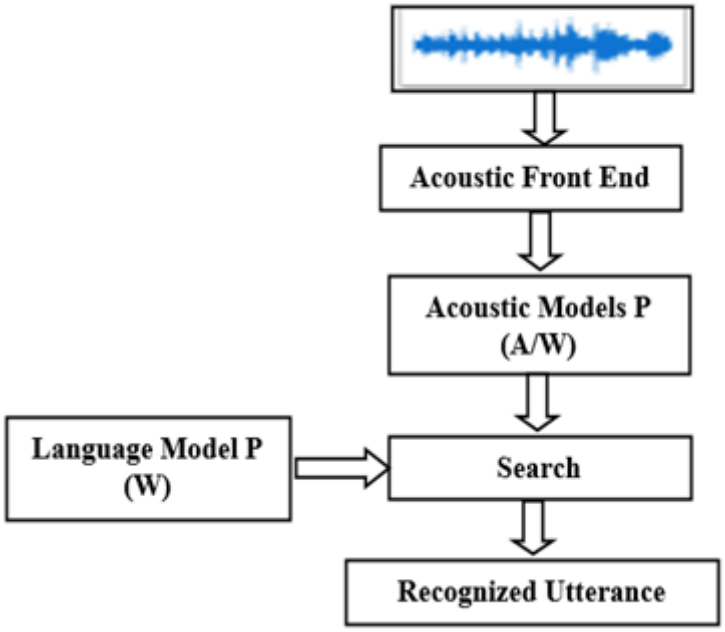

Figure 1: Basic Model of Speech Recognition

1. **Types of Speech Recognition**: The three speech patterns are disconnected, continuous, and isolated. Only words that are pronounced independently can be handled by isolated speech recognition systems. The most prevalent speech recognition technology now in use is this. Between each word or command spoken, the user must pause. Depending on the utterances they can recognize, speech recognition systems can be divided into a variety of groupings. These classes fall under the following categories:

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.

3. **Connected Words:** Similar to isolated words, a connected words system enables many

Copyright © 2024 Authors

assertions to be "run-together" with only a small pause in between. A computer will vocalise an utterance when it hears a word or group of words that all have the same meaning.

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.

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**.**

III. TYPES OF SPEAKER MODELS

Due to his individual physical characteristics and personality, each speaker has a different voice. The two primary classes of the system are developed based on these characteristics.

1. **Speaker Dependent Model:** Speaker-dependent, speaker-specific model. These models are less expensive to implement and simpler to utilise. For one speaker, it produces a more accurate result, however for other speakers, it produces a less accurate result. Systems designed for a certain speaker type are called speaker dependent systems. Although some speakers may find them less exact, they are typically more accurate for the particular speaker. These systems are frequently easier to build, more cost-effective, and more accurate. These systems, however, lack the flexibility of speaker-independent systems.

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).

IV. 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

Copyright © 2024 Authors

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).

V. 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%

Copyright © 2024 Authors

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%

VI. DESIGN AND DEVELOPMENT

1. Acquisition Environment of speaker and Instrumental setup:

- We collect the speech data from native people 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.

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.

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.

Copyright © 2024 Authors

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.

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.

Table 1: 0-9 Speech Corpus

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

Table 2: Days in a Week Speech Corpus

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

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.

Copyright © 2024 Authors

Table 3: Shows the Information About the Metadata.

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

VII. 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.

VIII. 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.

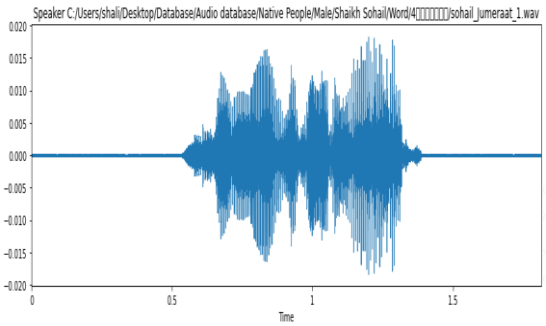

Figure 2: Native Speech wave form

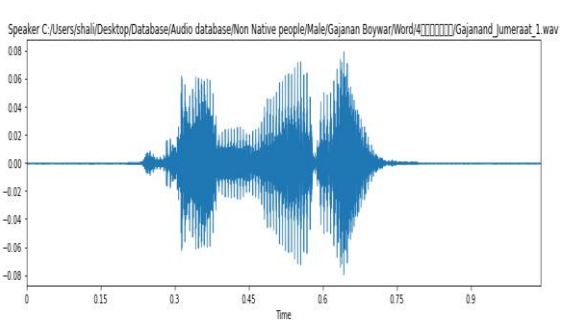

Figure 3: Non-Native Speech wave form

Copyright © 2024 Authors

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.

2. **Mel Frequency Cepstral Coefficients (MFCC):** A popular feature extraction method for voice processing and recognition is MFCC. The following steps are part of the MFCC technique:

    - **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):** To produce the magnitude spectrum, the windowed frames are transformed to the frequency domain using the FFT algorithm.

    - **Mel-frequency Wrapping:** This technique warps the magnitude spectrum onto the nonlinear Mel scale, which simulates the perception of frequency by the human auditory system.

    - **Cepstral Analysis:** The Discrete Cosine Transform (DCT) is used to convert the Mel-scaled spectrum to the Cepstral domain and produce the Mel-frequency Cepstral Coefficients (MFCCs). The first 12 MFCCs, which capture the most significant spectral features, are often employed.

    - **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)

Copyright © 2024 Authors

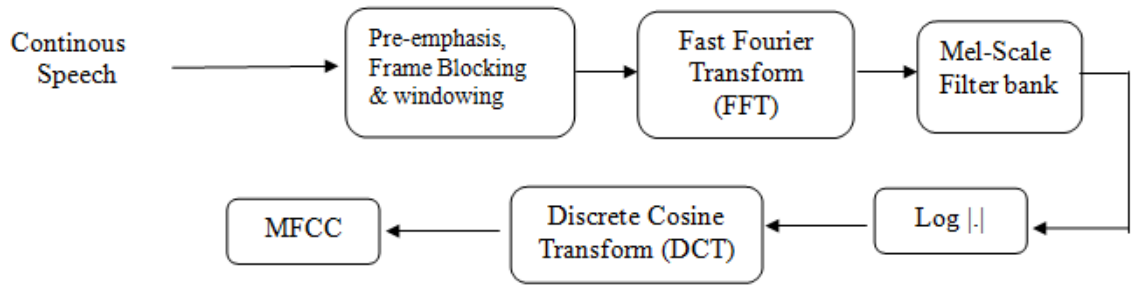

Figure 4: Steps involved in MFCC feature extraction

Two different types of filters—linearly spaced filters and logarithmically spaced filters—are used in the MFCC approach. Signal is conveyed using the Mel frequency scale to capture the phonetically significant aspects of speech. The Mel scale is mostly based on research on how humans perceive pitch or frequency. The scale is broken down into mel units. Ordinarily, the Mel scale is mapped linearly below 1000 Hz and logarithmically 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

$$Mel=2595 \log_{10}(1+f/700) \quad (1)$$

3. **HMM (Hidden Markov Model):** A phrase or a vocabulary is represented by a larger HMM that is created by concatenating the smaller HMMs for each unit of speech, such as a phoneme or a word. The goal of a hidden Markov model (HMM) is to identify the hidden parameters from the observable data by assuming that the system being modelled is a Markov process with unknown parameters. In a hidden Markov model, variables that are affected by the state but not the state itself are observable. A probability distribution over all potential output tokens exists for each state. Consequently, an HMM's series of tokens provides some information on the sequence of states. (Shaikh Naziya et.al.2017).

   The hidden variables that determine which mixture component will be chosen for each observation in a hidden Markov model are associated through a Markov process as opposed to being independent of one another. By using known utterances to build stochastic models, HMM assesses the likelihood that an unknown speech was produced by each model. Our feature vectors are (kind of) arranged into a Markov matrix (chains) that maintains probability of state transitions using statistical theory. That instance, if each of our code words were to stand for a particular state, the HMM would track the order in which those states changed and create a model that took those probabilities into account. HMMs are increasingly widely used since they can be automatically trained and are straightforward and computationally practical. HMM models these frames for recognition by treating the speech signal as quasi-static for short durations. (Kurzekar, P. K. et.al.2014).

4. **Prosodic Feature:** When we combine sounds in connected speech, prosodic elements take place. Prosodic elements should be taught to students since effective communication depends as much on rhythm, intonation, and emphasis as it does on the accurate pronunciation of sounds. It provides context, gives words meaning, and keeps listeners

Copyright © 2024 Authors

interested. Prosody entails accentuating the proper words, adjusting voice pitch and modulation, and employing the right 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.

IX. RESULTS AND DISCUSSION

In this study, in addition to the speech Hidden Markov Model (HMM) features, we used a variety of inputs, such as prosodic and speech spectral characteristics, pitch contour values, gender information, and syllable duration. We were able to successfully complete the experiment because there was enough data available for Urdu languages. Using our collected dataset of 30 male and 30 female speakers—both native and non-native speakers—the algorithm was trained using HMM. Based on the input features given, the trained network built using the HMM model was able to distinguish between native and non-native speakers of Urdu.

We conducted the experiment in two stages to examine the effects of spectral and prosodic characteristics on dialect digit and word recognition systems. The Mel-frequency cepstral coefficient (MFCC) characteristics of 13 features were initially used to train the HMM-based system. We then added syllable duration and pitch contour variables to the network's training set while maintaining the HMM structure.

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 80%. Notably, the system's performance exhibited a notable enhancement when trained with both spectral and prosodic information using the HMM model.

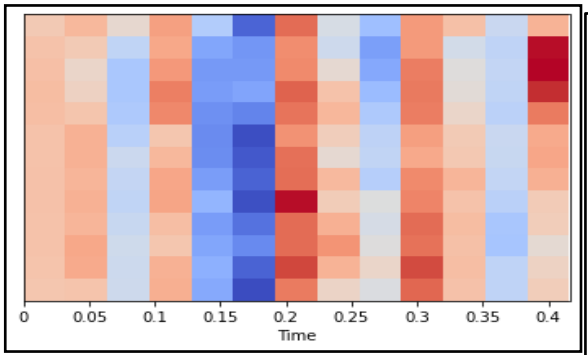

Figure 5: MFCC Spectrogram

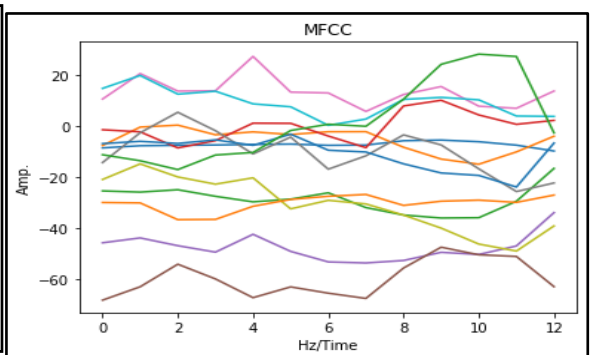

Figure 6: MFCC Feature

Copyright © 2024 Authors

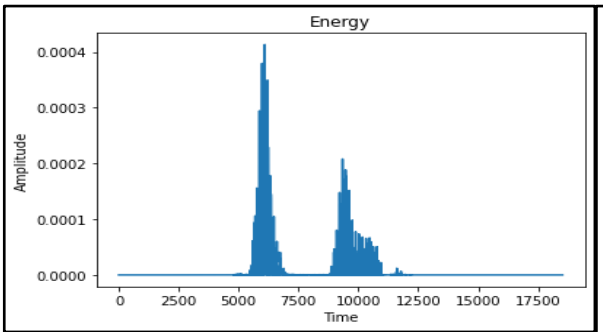

Figure 7: Energy

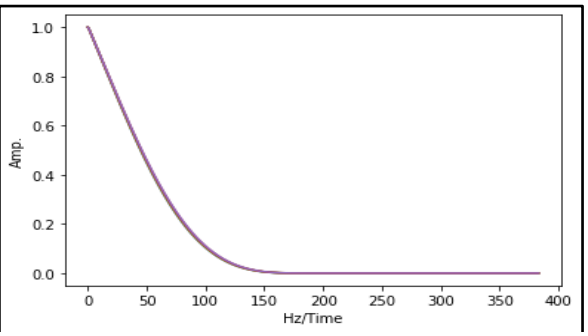

Figure 8: Tempo

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 सिफर (0) of Female Native Speaker

Table 4: Show MFCC Frame for सिफर (0) Female Native Speaker Sample with their Mean,Median and Standard Deviation

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									

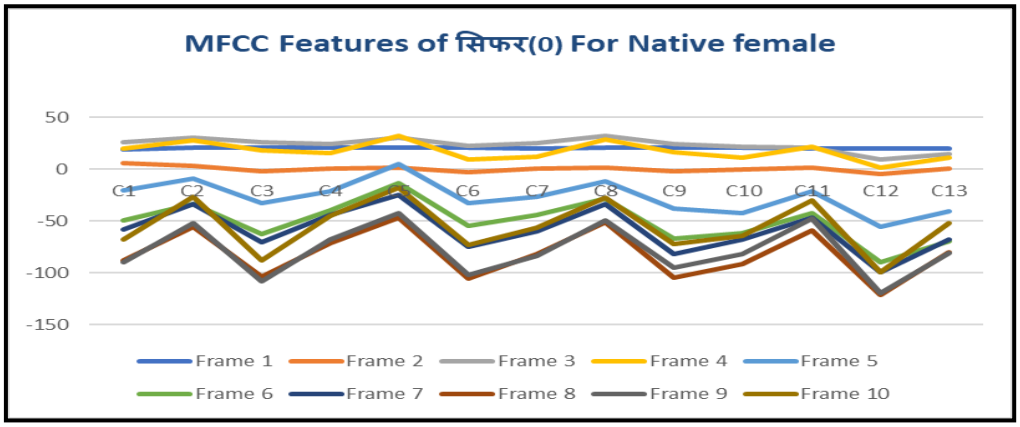

Figure 9: Shows the plot of MFCC features for सिफर (0) Female Native Speaker Sample

Copyright © 2024 Authors

The Table No. 4 consists of the 13 MFCC features and 10 frames: the numbers of the frames calculated varies according to the speech signal length of Non-Native female sifer(0) 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 सिफ़र (0) one person sample with Amplitude.

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

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

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									

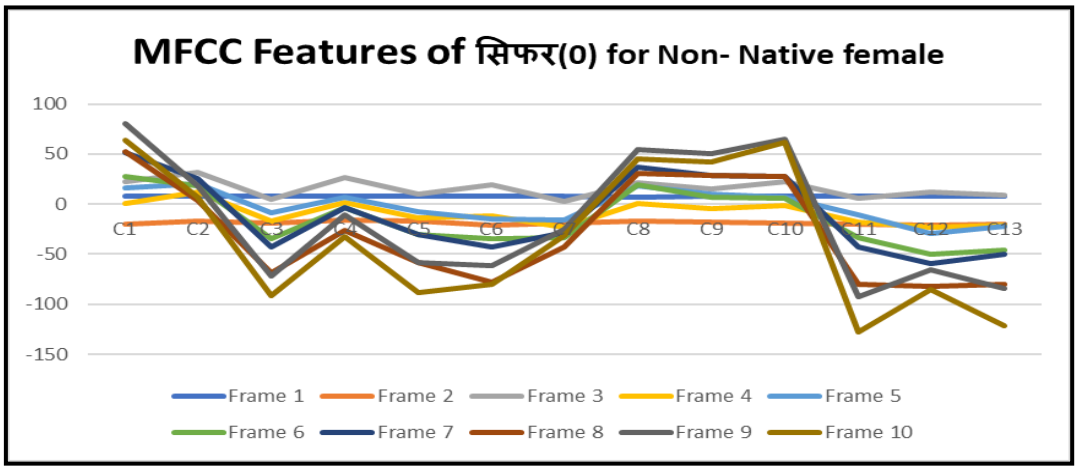

Figure 10: Shows the plot of MFCC features for सिफर (0) Female Non-Native Speaker sample

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, Medianand 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 सिफर(0) one person sample with Amplitude.

Copyright © 2024 Authors

X. 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 80% 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

[1] 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.
[2] 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.
[3] Shaikh Naziya, S., & Deshmukh, R. R. (2016). Speech recognition system—a review. IOSR J. Comput. Eng, 18(4), 3-8.
[4] 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).
[5] 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.
[6] 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.
[7] Saksamudre, S., & Deshmukh, R. (2015). Isolated word recognition system for Hindi Language. International Journal of Computer Sciences and Engineering, 3(7), 110-114.
[8] Oirere, A. M., Janvale, G. B., & Deshmukh, R. R. (2015). Automatic speech recognition and verification using lpc, mfcc and svm.
[9] 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.
[10] 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.
[11] 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.
[12] 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), 18006-180

Copyright © 2024 Authors