3. Study and Analysis of Gujarati voice for Structural Feature Extraction and classification

3.1 Introduction to Gujarati Language

Gujarati serves as the official language of the Indian state of Gujarat and is one of the 23 official languages of India. The primary varieties are Standard Gujarati (between Ahmedabad and Vadodara), Surati (southeastern Gujarat), Kathiyawadi (Saurashtra peninsula), Charotari (central Gujarat), and Patani (northern Gujarat). The Gujarati spoken in Pakistan closely resembles Patani. Kutchi, also known as Kachhi, is a closely related language spoken in western Gujarat, influenced by the adjacent Sindhi language of Pakistan. The most notable variety outside South Asia is East African Gujarati[1].

3.2 Gujarati Speech Features

Gujarati is an Indo-Aryan language spoken globally in many states and regions. Gujarati possesses numerous characteristics, including accents, phonology, script, inflection, and dialects. The various characteristics render it intriguing to investigate. The Gujarati language possesses distinct phonetic, prosodic, and linguistic features that set it apart from other languages. These attributes are shaped by its profound historical and cultural legacy and significantly contribute to both communication and speaker identification. The following are the key attributes of Gujarati speech:

3.2.1 Voice Biometrics

3.2.1.1 Pitch (High vs. Low Pitch)

Pitch Variation: Like other languages, in Gujarati also, the pitch plays an important role for conveying meaning and distinguishing different types of statements. While uttering for a neutral or factual declaration of statement, the speaker uses a low pitch; to show the strings of seriousness, they use low pitch. This low pitch sets on the tone being informational in nature. A Gujarati speaker can easily indicate whether he or she is stating, interrogating, or showing wrong emotions by changing the pitch.

- "તમારા પ્રશ્નો બહુ મહત્વના છે." (Tamara prashno bahu mahtvna chhe) – "Your questions are very important."

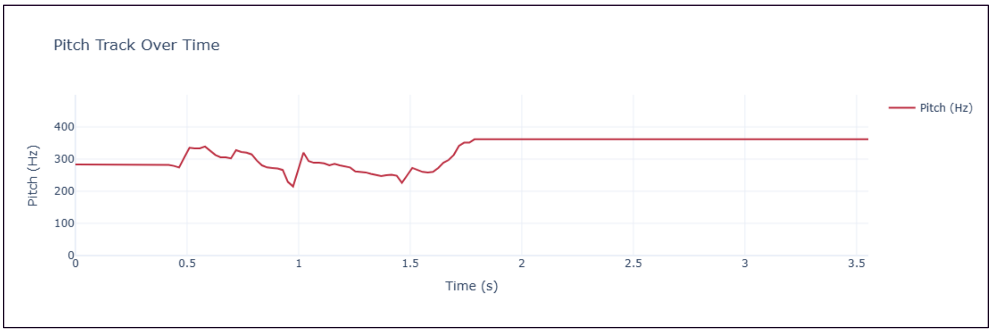


Figure 3- 1 Pitch Variation in Neutral Sentence

The above Figure 3-1 illustrate the Pitch Variation in spoken sentence type neutral as this type of sentences have the minimal variations.

Most Gujaratis use a rather high pitch quite more often to express intense oscillations of emotions like ecstatic ness, surprise, and, rather occasionally, even extreme anger. This adds this intensity to the speech at most cases, revealing the speaker's feeling evident: be it excitement for happy news, shock on watching something unexpected, or the anguish resultant from frustration with some unwarranted situation-the uptorn in pitch helps the catch the emotional weight of an utterance.

- "કેમ છો, દોસ્ત?" (Kem cho, dost?) – "How are you, friend?" (with a friendly, warm intonation).

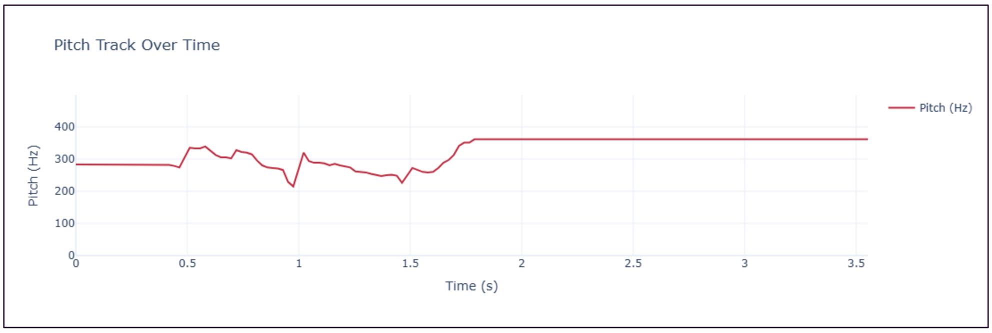


Figure 3- 2 Pitch Variation in Spoken Sentence Conveying Regards

In Gujarati, most statements or utterances related to certainty are said on a falling pitch. Thus, the usual intonation when speaking in assertive sentence mode is usually a fall to indicate finality as shown in Figure 3-3; a confident statement, at that. The drop in tone provides the speech with a sense of completeness, which reassures one of the speaker's positive beliefs in what he is saying.

- "આઈનાના પ્રોડક્ટ્સ ઘણી સરસ છે." (Ainana products ghani saras chhe) – "Aina's products are very good."

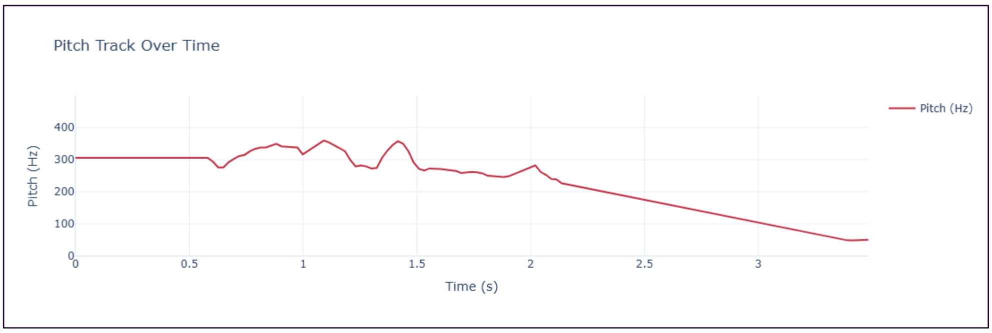


Figure 3- 3 Pitch Variation in Assertive Sentence

In Gujarati, the intonation rises in the case of a question or modal sentence, mostly towards the end of the sentence. The upward pitch shift indicates that the speaker is inquiring about something or doubts it. This is quite commonly found in interrogative sentences, which helps to distinguish them from statements.

- "તમે ક્યાં જઇ રહ્યા છો?" (Tame kyā jāi rahyā cho?) – "Where are you going?" The pitch rises at the end of the question.

3.2.1.2 Timbre

The quality that makes one voice distinct from another is called timbre. In the context of a speaker, timbre would form the basis of recognition or verification. Timbre concerns the acoustic features of a speaker's voice and is unrelated to the content of the speech, that is, what words are spoken. The specific particularities of the person result from the features that regard the shape of the vocal tract, resonance, or harmonic frequencies.

In a language like Gujarati, as indeed in any other language, timbre helps greatly in differentiating voices. When two different persons utter the same sentence in Gujarati, say, "હું ગુજરાતી બોલી રહ્યો છું" ("I am speaking Gujarati"), their voices will differ because of their different timbres. A deeper-voiced individual would therefore have a richer, fuller timbre, while a high-pitched voice would result in a lighter, sharper timbre. The difference in timbre depends not on the words spoken, but rather on the physical properties of their vocal apparatus.

Timbre is one major factor in speaker recognition systems on account of its uniqueness and stability, besides being reliable. Identification and authentication in a wide group of applications can be conducted properly with the timbre of one's voice-one's acoustic fingerprint-whether speaking Gujarati or even any other language.

3.2.1.3 Speech Rate (Tempo)

Normally Gujaratis speak at a moderate or slow rate in formal conditions or speeches, especially those on cultural or traditional ground, giving sufficient time to the audience. This allows them to gain more understanding of what's being told and also the words clear. For example, at the welcome speech, speech can be done slowly while being clear about things for clarity and effectiveness. On the other hand, in informal or day-to-day conversations, like among friends or family, speakers of Gujarati tend to hurry up, showing the relaxed nature of the setting.

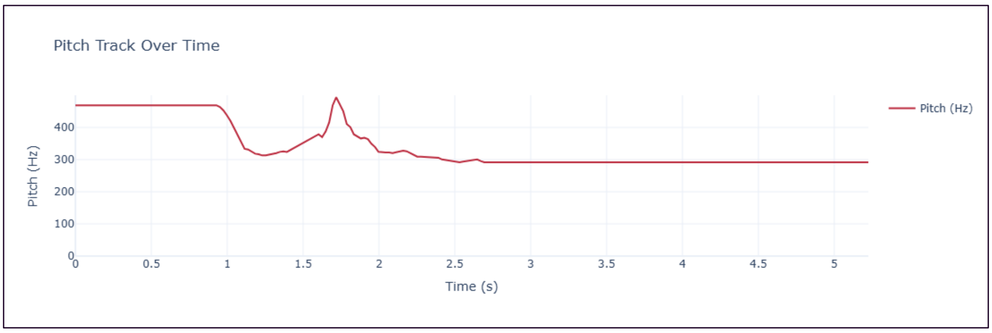


Figure 3- 4 Increasing Speech Rate with Blended Words

- "કેમ છો, શું ચાલી રહ્યું છે?" (Kem cho, shun chaali rahyo chhe?) – "How are you? What's going on?" In such contexts, the tempo might increase, and words might blend together more as shown in Figure 3-4.

3.2.1.4 Rhythm (Tempo and Flow)

Gujarati speech, very often, is on a regular rhythm, at an even pace, which becomes most marked in traditional storytelling or while reciting poetry, such as during "Garba" dances or Bhajans. This rhythmic flow not only gives the language a natural musicality but also enhances the emotional and cultural experience, making the speech feel more like a melody than mere words.

Rhythm is what plays the central role in every poetic performance, like 'Garba' or Bhajans. In every line, the syllable is set in a rhythm in such a way that it not only sounds musically appealing but also reveals the depth of culture in the speeches of Gujarati people. By adding meaning to this, their traditional practices are enriching both in sound and cultural expression.

3.2.1.5 Emotion in Speech (Expressiveness)

The intonation of the language holds a deep relationship with expressing emotions in Gujarati, like many languages. The art of injecting life into speech depends mostly on modulation of pitch, speed, and rhythm of speech. These factors introduce elements through which the speaker can put excitement, sorrow, surprise, or anger in his words to make a message come alive and relate well.

The Gujarati language is immensely expressive, and much of its emotion is conveyed through variations in intonation and pitch modulation. With a variation in pitch and tone, the speaker can show everything from joy and surprise to frustration and sadness. This subtlety in sound helps to convey the depth of emotion behind the words, adding richness and nuance to the communication.

Joy and Excitement can be Expressed with a higher pitch as shown in Figure 3 -5 and faster speech.

- "બધાને મજા આવી!" (Badhane maja aavi) – "Everyone had fun!" (with a joyful tone)

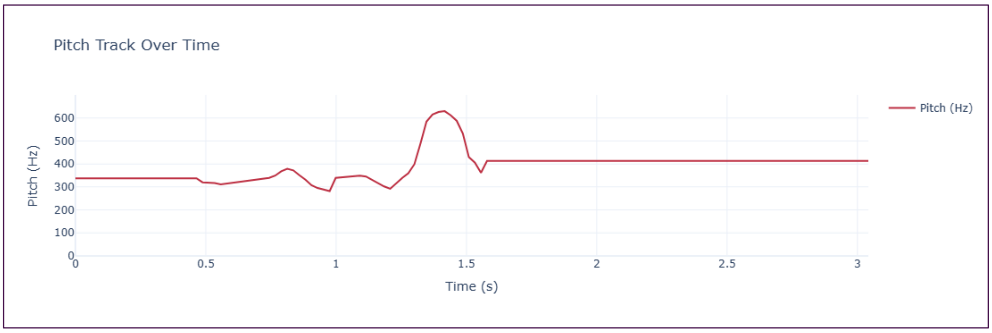


Figure 3- 5 Pitch Intonations of sentence Conveying Excitement

Anger or Displeasure Often signalled with a stronger, sharper tone as shown in Figure 3-6, especially in informal speech.

- "આ કરવાનું છે?" (Aa karvanu chhe?) – "Is this what you want to do?" (with frustration)

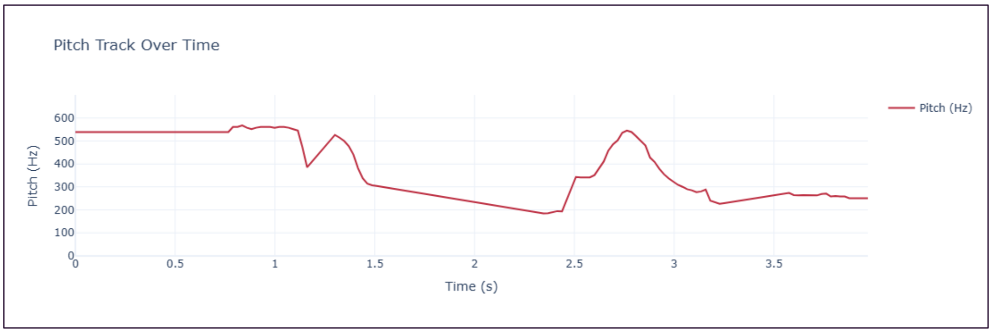


Figure 3- 6 Pitch Intonations of sentence Conveying Displeasure

Sadness or Contemplation are A lower pitch as shows the Figure 3-7 and slower speech are often used to reflect sadness or thoughtfulness.

- "આ દુનિયા ઘણી અન્યાયી છે." (A duniya ghani anyayi chhe) – "This world is so unfair." (said softly)

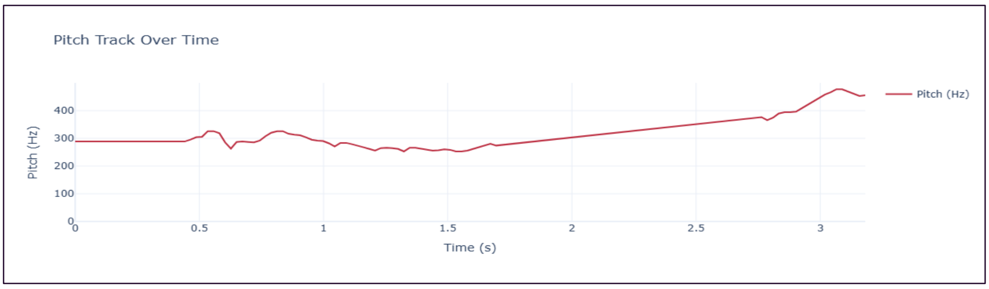


Figure 3- 7 Pitch Intonation of Sentence Conveying Sadness

For instance, when feeling delighted or excited, a Gujarati speaker always pitches high and speaks fast. Anything that depicts high energy would be something where one shows enthusiasm for something or shares very good news with another person:

- *"આજ ખૂબ મજા આવી છે!"* (Aaj khub maza aavi chhe!) meaning "Today is so much fun!" The excitement is evident in the fast pace and slightly higher pitch that shows the below Figure 3-8.

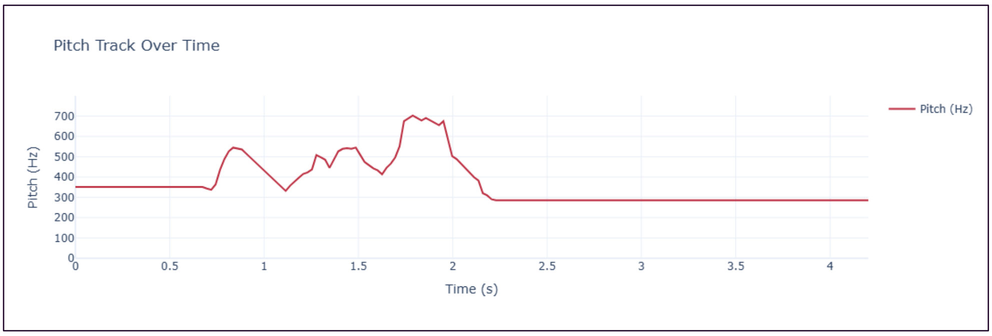


Figure 3- 8 Pitch Intonation of Sentence Conveying Emotions of Surprise

On the other hand, anger or frustration is usually conveyed with a lower pitch and much force in the voice that we can see in Figure 3-9. The speaker may also speak slowly or emphasize words to show the gravity or intensity of their emotions. For example,

- *"હવે હું સહન કરી શકતી નથી!"* (Havē huṁ sahana karī śakato nathī!) means "I can’t take this anymore!" Here, the tone is firm, the pace slow, and the pitch tends to be lower to communicate a sense of irritation.

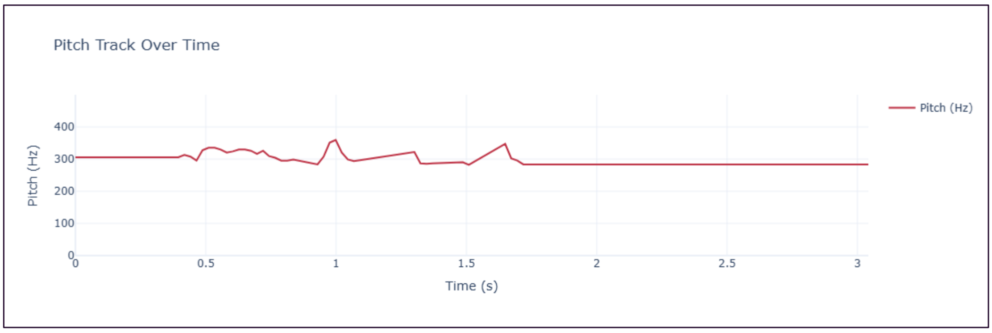


Figure 3- 9 Pitch Intonation of Sentence Conveying Anger

For emotions like surprise or shock, the pitch usually rises sharply and the speech rate slows down momentarily, as the speaker processes the unexpected. A phrase like

- *"તમે આ બધું કેવી રીતે કર્યું?"* (Tame ā badhuṁ kevi rite karyuṁ?) meaning "How did you do all this?" is spoken with a rising intonation as Figure 3-10 illustrates, conveying a sense of disbelief or astonishment.

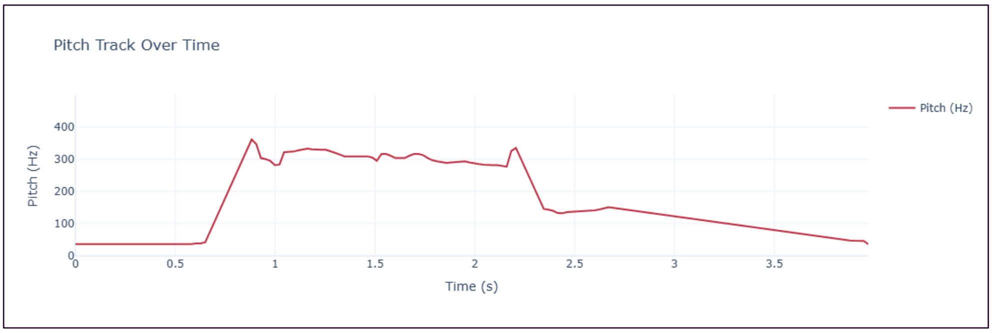


Figure 3- 10 Pitch Intonations of sentence Conveying Sadness

Sadness or melancholy is typically expressed through a slower rate of speech, often with a lower pitch and elongated vowels. A sentence like

- *"હું ખૂબ દુ:ખી છું."* (Hun khub dukhi chhu) meaning "I am very sad," is spoken more slowly and softly, with a softer, sadder tone indicating lower pitch intonations as shown in Figure 3-11.

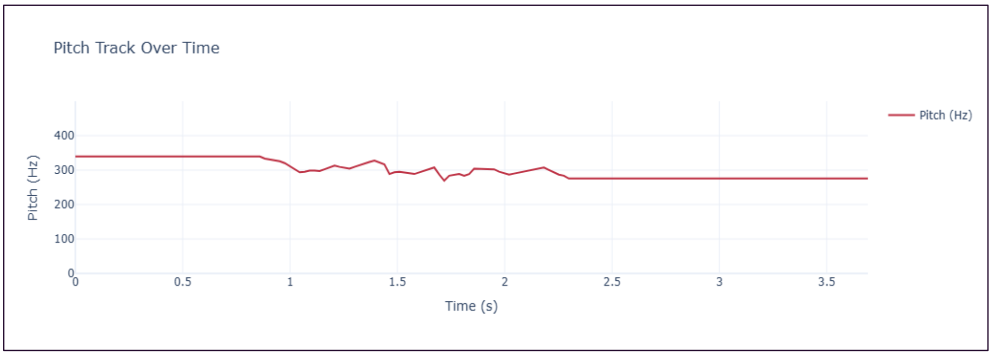


Figure 3- 11 Pitch Intonation of Sentence Conveying Melancholy

In Gujarati, the rhythm and melody of the language itself contribute greatly to how feelings are felt and perceived by the listener. For example, in storytelling or bhajans, there is often a rhythmic pattern that enhances emotional expression. Be it in a joyful tune or a sombre prayer, the rhythm of the language pulls the listener deeper into the emotional experience being shared.

By skilfully modulating the rate of speech, pitch, and rhythm, speakers of Gujarati are able to reveal a broad range of expression. It is such adaptability that allows the language to portray everything from absolute joy to deep sorrow, thereby bringing this emotional depth into the hearts of listeners.

3.2.2 Phonetic Features

The phonetic features of Gujarati speech delineate its unique sound system and affect the articulation, perception, and recognition of the language. The primary distinctions among several stages of dialects arise from phonetic variances in Gujarati. The primary phonetic characteristics of Gujarati are as follows:

3.2.2.1 Vowel System

Gujarati possesses both short and long vowels, exhibiting a clear phonemic distinction between the two. All vowels, with the exception of [e] and [o], appear in nasalised, murmured, and non-murmured variants. Gujarati possesses both short vowels shown in Table 1-1 and long vowels shown in Table 3-2; nevertheless, they are not contrastive. Long vowels are typically more tense than their short equivalents. The length of this vowel is essential for distinguishing words[2].

3.2.2.1.1 Short Vowels

Table 3- 1 Table showing the Examples of Short Vowels

Features	Description	Examples in Gujarati	IPA Transition
/a/	Short "a" sound	"કપ" (kap) – "cup"	[kəp]
/i/	Short "i" sound	"સિટ" (siṭ) – "sit"	[sɪʈ]
/u/	Short "u" sound	"પુટ" (puṭ) – "put"	[pʊʈ]
/e/	Short "e" sound	"બેડ" (beḍ) – "bed"	[beɽ]
/o/	Short "o" sound	"ડોગ" (ḍog) – "dog"	[ɖog]

3.2.2.1.2 Long Vowels

Table 3- 2 Table showing example of Long Vowels

Features	Description	Examples in Gujarati	IPA Transition
/aː/	Long "a" sound	"માતૃ" (mātṛ) – "mother"	[mɑːtr̥]
/iː/	Long "i" sound	"મશીન" (maśīn) – "machine"	[mɑːʃiːn]
/uː/	Long "u" sound	"બૂટ" (būṭ) – "boot"	[buːʈ]
/eː/	Long "e" sound	"કેક" (keḵ) – "cake"	[keːk]
/oː/	Long "o" sound	"ગો" (go) – "go"	[goː]

3.2.2.1.3 *Diphthongs*

Table 3- 3 Table Showing Diphthongs examples

Features	Description	Examples in Gujarati	IPA Transition
/ai/	"ai" as in "eye"	"દ્રષ્ટિ" (dṛṣṭi) – "sight"	[dɽɪʃʈi]
/au/	"au" as in "how"	"કાઉ" (kāu) – "cow"	[kɑːu]

3.2.2.1.4 *Nasalized Vowels*

Table 3- 4 Table Showing Nasalized Vowels

Features	Description	Examples in Gujarati	IPA Transition
/ŋ/	Nasalized "ng"	"સંગ" (saṅg) – "together"	[saŋg]

3.2.2.2 Consonant System

Gujarati comprises a total of 31 consonants, which include 20 stops, 3 fricatives, 3 nasals, and 5 liquids and glides. Stops and nasals are produced at five distinct locations, designated as labial, dental, retroflex, palatal, and velar. The palatal stops are, indeed, affricates. Each set of stops include voiceless and voiced consonants, as well as unaspirated and aspirated variants; this four-way distinction is exclusive to Indo-Aryan within the Indo-European language family. Proto-Indo-European exhibited a triadic contrast exclusively[2].

The table 3-5 shows the examples of consonant system used during the voice processing system.

Table 3- 5 Table showing Example of Consonant System

Features	Description	Examples in Gujarati	IPA Transition
/pʰ/	Aspirated bilabial	"ફોન" (phōn) – "phone"	[pʰoːn]
/bʰ/	Aspirated bilabial	"ભલાઈ" (bhalā'ī) – "goodness"	[bʰəlāːʌi]
/tʰ/	Aspirated dental	"થેલ" (ṭhēl) – "bag"	[ʈʰɛːl]
/dʰ/	Aspirated dental	"ધ્વનિ" (dhvani) – "sound"	[dʰʋəni]
/ʈʰ/	Aspirated retroflex	"ઠેકો" (ṭhēkō) – "correct"	[ʈʰeːkɔ]
/ɖʰ/	Aspirated retroflex	"ઢગલો" (ḍhagalō) – "pile"	[ɖʰəgəlɔː]
/kʰ/	Aspirated velar	"ખોટું" (khōṭu) – "wrong"	[kʰoːʈu]
/gʰ/	Aspirated velar	"ઘરો" (gharō) – "house"	[gʰəɾoː]
/pʰ/	Aspirated bilabial	"ફોન" (phōn) – "phone"	[pʰoːn]
/bʰ/	Aspirated bilabial	"ભલાઈ" (bhalā'ī) – "goodness"	[bʰəlāːʌi]
/tʰ/	Aspirated dental	"થેલ" (ṭhēl) – "bag"	[ʈʰɛːl]
/d/	Voiced dental	"દરજું" (darajũ) – "door"	[dəɽɪʈʃũ]
/ʈ/	Voiceless retroflex	"ટર્ન" (ṭarn) – "turn"	[ʈarn]
/ɖ/	Voiced retroflex	"ડ્રમ" (ḍram) – "drum"	[ɖram]

Gujarati possesses a distinct script. It is a syllabic alphabet (abugida), ultimately derived from Brāhmī, whereby each consonant possesses the inherent vowel [ə]. Its fundamentals resemble those of the Devanāgarī script. Diacritic vowel marks are affixed before, after, above, or below a consonant to denote non-initial vowels.

The Gujarati alphabet comprises 47 letters arranged according to phonetic principles, with conventional transliteration and International Phonetic Alphabet equivalents provided beneath each letter. Initially, the simple vowels are presented, succeeded by the syllabic vowels, and subsequently the diphthongs (e and o originate from old diphthongs and were regarded as such by the native grammarians). Above table shows the vowels are the stops and nasal consonants, categorised into five groups (each comprising five letters) based on their site of articulation (from posterior to anterior). In each group, the sequence is: voiceless unaspirated stop, voiceless aspirated stop, voiced unaspirated stop, voiced aspirated stop, nasal. Subsequent to these five groups, the semivowels (liquids and glides) are organised according to their point of articulation. Subsequently, the fricatives commencing with the sibilants[2].

3.2.3 Prosodic Features

Prosodic features encompass the aspects of speech that extend beyond mere phonemes, such as vowels and consonants. They encompass elements like as rhythm, stress, intonation, and pitch as shown in Table 1-6. These traits are essential for our interpretation of meaning, emotions, and emphasis in speech In Gujarati, just like in other languages, prosody helps us tell whether we're making a statement or asking a question, stressing a particular point, or even expressing different emotions.

Table 3- 6 Prosodic Features of Gujarati Speech

Features	Description	Examples in Gujarati	IPA Transition
Pitch	Highness or lowness of the voice	"તમારું નામ શું છે?" (What is your name?)	[təmɑːruṁ nɑːm ʃuṁ t͡ʃe?]
Stress	Emphasis on certain syllables in a word	"વિશ્વ" (viśv) – "world"	[vɪʃʋ]
Intonation	Rise and fall of pitch across phrases or sentences	"હું ઘર જાઈ રહ્યો છું." (I am going home.)	[huṁ gʰarē d͡ʒəɪ rəhjɔː t͡ʃuṁ]
Speech Rate	Speed at which speech is delivered	"તમે ક્યારે આવી રહ્યા છો?" (When are you coming?)	[tʰəmeː kjæɾe< aːvi< rəhjɔː ʧoː?]
Rhythm	The pattern of stressed and unstressed syllables	"તમારા બાળક ક્યાં છે?" (Where is your child?)	[təmɑːɾɑː bɑːɭək kəɳyɑː ʧe<]
Pauses	Pauses to create emphasis, uncertainty, or convey emotion	"હું... હું તમને સહયોગ કરવા માટે કહું છું." (I... I am asking you to cooperate.)	[huṁ... huṁ tʰəmɑːnə səʰo<g kəɾvāː maːṭe< kəɦuṁ tʃuṁ]
Response Tone	Tone used to respond to indicate agreement or disagreement	"હા, મારે તે કરવું છે." (Yes, I need to do that.)	[hɑː, mɑːɾe< te< kəɾvũ< tʃe<]

3.2.3.1 Stress Patterns (Stress and Accent)

Initial Stress: In Gujarati, emphasis predominantly occurs on the initial syllable of words, constituting a fundamental prosodic characteristic that influences the language's rhythm. This inclination imparts a uniform, nearly melodic cadence to Gujarati speech, as the opening syllables are generally accentuated more than the subsequent ones. This characteristic is ubiquitous throughout the majority of the language's lexicon, establishing a consistent and rhythmic auditory pattern. It aids in sustaining a specific rhythm in daily discourse, contributing to the language's vibrancy and dynamism.

Example:

- "ગુજરાત" (Gujarat) – stress on "ગુ" (gu)
- "સંગીત" (sangeet) – stress on "સં" (sang)

Unlike languages such as English, which exhibit unpredictable stress shifts between syllables in polysyllabic words, Gujarati adheres to a more consistent stress pattern. This indicates that, in the majority of instances, the stress is assigned to the initial syllable of a word, irrespective of its length. The constancy of Gujarati imparts a continuous rhythm, rendering its speech more consistent and predictable, in contrast to the diverse stress patterns seen in English, where the stressed syllable may vary based on the word or its context[1].

Syllable-Timed Rhythm: Gujarati is classified syllable-timed language. It generally indicates that every word used in the speech is take equivalent duration when spoken. The difference between stress-timed and syllable-times language is that in the stress-times language like English, stressed syllables are stretched out and relaxed ones are truncated, and because of that it results in more irregular rhythm. While in syllable timing languages like Gujrati communicates a smooth and uniform flow to the language and it interprets the discourse stable and rhythmic. The reliable rhythm improves the variability and harmony while speaking a language like Gujrati that can give pleasing experiences to hearing senses.

3.2.3.2 Intonation Patterns (Pitch and Tone)

Rising and Falling Intonation: Like number of other languages, Gujrati intonation is crucial for differentiating exploration and declarations. The pitch typically ascends at the conclusion with indication of curiosity or a question in interrogative sentences of

Gujarati language. While in declarative sentence, typically have a sinking tone which resulting in expressing certainly. The pattern of ascending and descending pitch in Gujarati shows its distinctive conversional rhythm similar to the intonation system seen in other languages.

Rising Intonation:

- "તમને કેમ છે?" (Tame kem chhe?) – "How are you?"

In the above sentence pitch rises at the end of the sentence as we can see in Figure 3-12 to indicate a question.

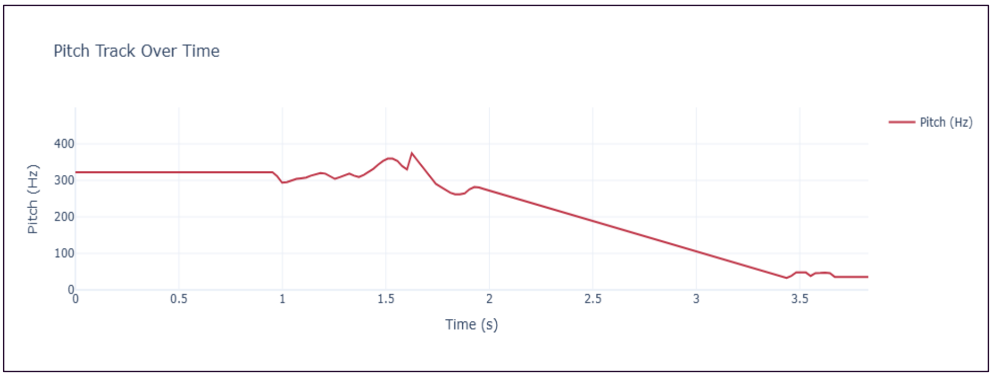


Figure 3- 8 Rising Intonation Pattern of Spoken Sentence

Falling Intonation:

- "આજે આપણી મીઠી સંસ્કૃતિને ઉજવણી છે." (Aaje aapni mithi sanskruti ne ujavani chhe.) – "Today we are celebrating our sweet culture."

In the above sentence pitch falls at the end of the sentence to indicate a statement shown in Figure 3-13.

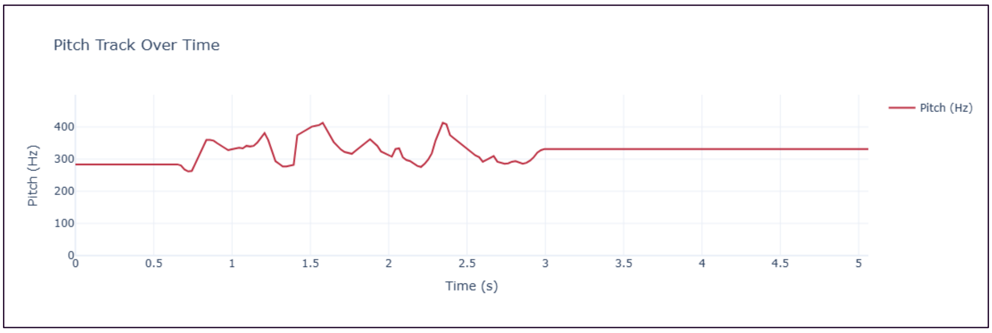


Figure 3- 9 Falling Intonation Pattern of Spoken Sentence

Neutral Intonation: In declarative statements or when expressing general observations, the intonation in Gujarati remains neutral and consistent. In Gujrati, the absence of notable pitch variations conveys a serene and harmonious tone to the voice. This intonation is distinguished when the speaker is merely conveying information or saying a fact, without any focus on inquiries or emotional nuances.

- "હું ઘરે જઈ રહ્યો છું." (Hun ghare jai rahyo chhu) – "I am going home." The sentence conveying neutral intonation shown in Figure 3-14.

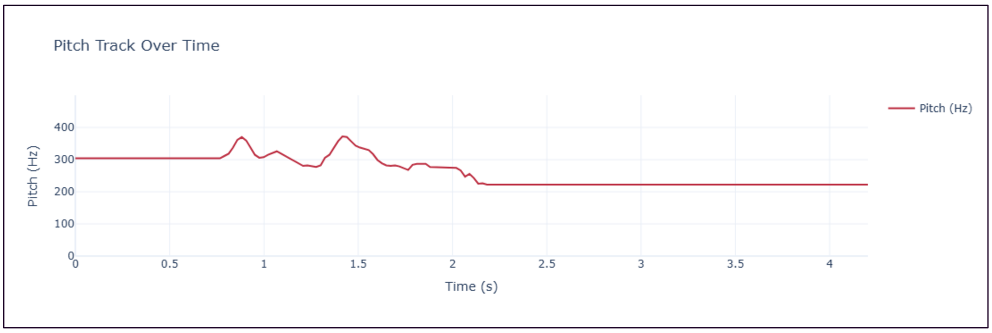


Figure 3- 10 Neutral Intonation Pattern of Spoken Sentence

Intonation for Emphasis: Gujarati Speakers frequently shows their pitch or execute unexpected tonal shifts to significant point or convey intense emotions such as enthusiasm, rage or surprise. These modifications highlight the emotional intensity of the words that renders the speech more vibrant and expressive. With the pitch modification like mentioned, the speaker expresses their emotional engagement with the topic and enhance the depth and importance of their message.

- "આજે તો આપણે મજા કરી લીધી!" (Aaje to aapne maja kari lidhi!) – "Today, we had so much fun!"

The word "મજા" (maja) is emphasized with a higher pitch that shows the excitement.

3.2.3.3 Pauses and Silences

The speakers pause occasionally in certain places in Gujarati, to give more depth to what is being conversed or to let a thing sink into the thinking of the listener. When these pauses are further indulged in, the lengthy sentences break down into smaller-size pieces that the listener can catch their breath with. This will make it certain that the main ideas come ahead and get enough space till they are digested well.

- "આજે... આ પ્રસંગ ઘણો વિશિષ્ટ છે." (Aaje... aa prasang ghanō vishisht chhe) – "Today... this event is very special." The Intonation of statement that has pause the pitch become neutral in between and then raise at the high point shows the Figure 3-15.

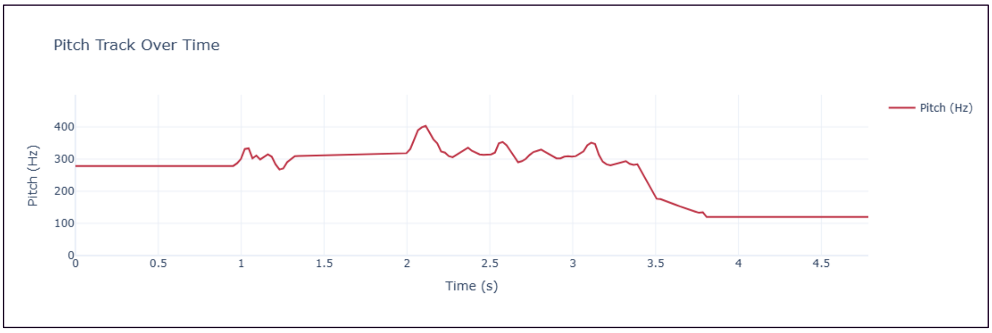


Figure 3- 11 Speech Rate of Sentence having Pause

Silence holds a great importance in speech, especially when one needs to be more formal or reflective. Silence helps point the concentration onto the words to come with higher strength. A timely pause has an effect of highlighting the statement to be grave or serious by giving weight and resonance to what the speaker is about to say.

3.2.4 Syntactic Features

Syntactic features are the structural features of a language, such as the rules that control how words, phrases, and larger sentences are combined to convey meaning in the language. In speaker recognition, syntactic features are less directly used compared to other features like voice biometric or prosodic features. However, they can be informative of a speaker's identity, especially in text-independent systems of speaker recognition, or any other analysis apart from the speaker's mark.

3.2.4.1 Word Order

Different languages and cultures use different patterns of word order. Within one language, the individual speaker has their ways of arranging the word pattern in the sentence. This feature could thus give an insight into the linguistic background of a speaker. A speaker's usual sentence structure or syntax, for example, may indicate his mother tongue, dialect, and sometimes even educational background.

Gujarati typically follows the word order of SOV. It might be said by a speaker as "હું તમારી સાથે જઈશ", following this structure. More linguistically, English would create a sentence, "I will go with you"; it follows SVO ordering. Differences in sentence structure may indicate which language you use a majority of the time or help specify someone's cultural background.

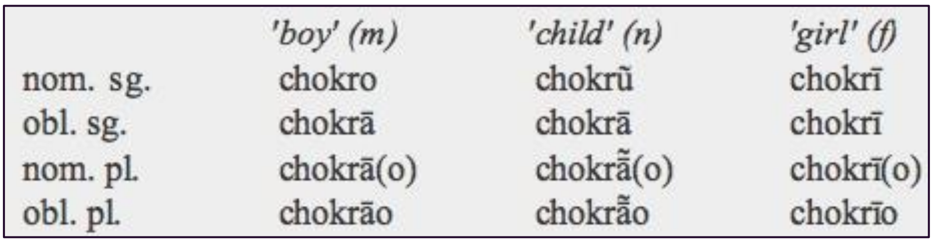
3.2.4.2 Postpositions

Languages like Gujarati make use of postpositions, for example, which can give away a lot about a speaker's linguistics; hence, postpositions depict syntax that always appear after a modified noun, as in this sentence: "સાથે" means "with" in "હું તમારી સાથે જઈશ.".

This might form one of the distinctive syntactic features which may distinguish speakers of languages like Gujarati, that use postpositions, from speakers of languages like English that use prepositions. Recognition systems analysing sentence structure may use such features to make out not just the language but sometimes even regional dialects or speech patterns.

3.2.5 Linguistic Features

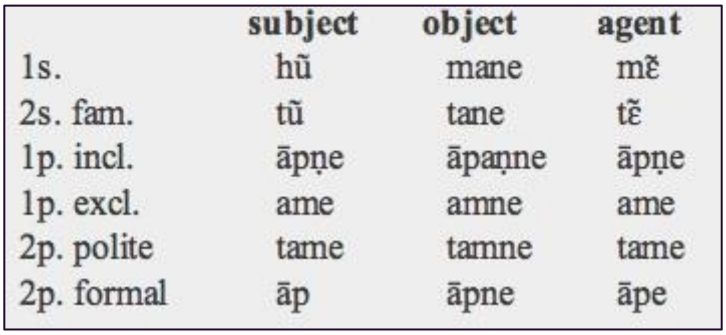
Gujarati boasts a rich and structured morphology, with nouns and adjectives that can either be inflected or remain invariable depending on their syntactic role. Inflected adjectives in Gujarati agree with the nouns they modify in terms of case, gender, and number. The language features three genders—masculine, neuter, and feminine—with masculine singular nouns typically ending in *-o*, neuter nouns in *-ũ*, and feminine nouns in *-ī*. As for number, Gujarati has both singular and plural forms, with the plural indicated by the suffix *-o*. In terms of case, Gujarati has three distinct forms: the nominative, used for the subject or direct object; the oblique, which appears when nouns are accompanied



	'boy' (m)	*'child' (n)*	*'girl' (f)*
nom. sg.	chokro	chokrũ	chokrī
obl. sg.	chokrā	chokrā	chokrī
nom. pl.	chokrā(o)	chokrā̃(o)	chokrī(o)
obl. pl.	chokrāo	chokrā̃o	chokrīo

Figure 3- 12 Linguistic Features of Gujarati for Different Gender by postpositions that define other syntactic functions; and the vocative, which is morphologically identical to the oblique and used when addressing someone directly[3].

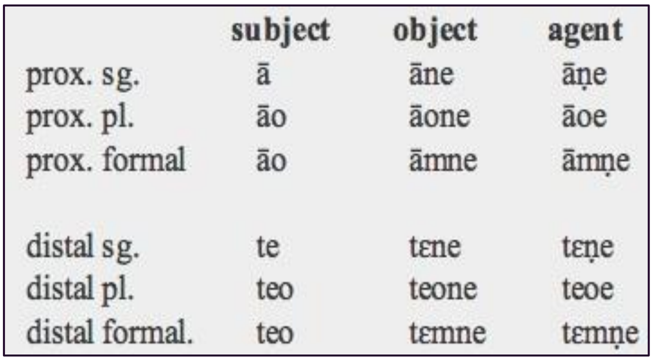
Gujarati pronouns come in various types, including personal, demonstrative, interrogative, relative, and indefinite pronouns. Personal pronouns distinguish between inclusive and exclusive first-person plural forms, and while they are gender-neutral, they feature three levels of formality in the second-person singular, ranging from familiar to highly formal. Demonstrative pronouns express proximity and distance and include a plural, polite form. Interrogative pronouns like *kɔṇ* (who?), *śũ* (what?), *kyāre* (when?), *kyā̃* (where?), and *kɛm* (why?) allow for asking questions, while relative pronouns inflect for both number and case (e.g., *je* for singular nominative and *jeo* for plural nominative), often linking with distal demonstrative pronouns in main clauses[2].



	subject	object	agent
1s.	hũ	mane	mɛ̃
2s. fam.	tũ	tane	tɛ̃
1p. incl.	āpṇe	āpaṇne	āpṇe
1p. excl.	ame	amne	ame
2p. polite	tame	tamne	tame
2p. formal	āp	āpne	āpe

Figure 3- 13 Pronouns in Gujarati Language

Gujarati also makes extensive use of compound words formed by combining adjectives and nouns in various patterns. In the adjective-noun type, the adjective modifies the noun (e.g., *black-bird*), while in the noun-adjective form, adjectives are used for comparisons (e.g., *ice-cold* = as cold as ice). [2].



	subject	object	agent
prox. sg.	ā	āne	āṇe
prox. pl.	āo	āone	āoe
prox. formal	āo	āmne	āmṇe
distal sg.	te	tɛne	tɛṇe
distal pl.	teo	teone	teoe
distal formal.	teo	tɛmne	tɛmṇe

Figure 3- 14 Compound Words

3.2.6 Speech Articulation

3.2.6.1 Dialects Variations

Like any other language, Gujarati, too, has its varieties or dialects, which range from the regional to communal levels. These dialects vary from one another on phonetic, lexical, and even some grammatical levels; however, they are generally comprehensible for speakers of standard Gujarati. Variations have been determined by historical, geographical, and cultural factors, not only across different parts of Gujarat but also in the world over in communities that use Gujarati.

Major dialects of Gujarati include Standard Gujarati, spoken between Ahmedabad and Vadodara; Surati, from southeastern Gujarat; Kathiyawadi, from the Saurashtra peninsula; Charotari, in central Gujarat; and Patani, from northern Gujarat. Curiously, Gujarati spoken in Pakistan is nearer to Patani. Kutchhi, also called Kutchhi, which is spoken in western Gujarat, is a language related to Sindhi but has been influenced mostly by Pakistan's Sindhi tongue. Exclusively outside South Asia, East African Gujarati must be one of the most salient varieties of this language.[3].

Here’s a closer look at some of the prominent Gujarati dialects. The table below presents examples of speech samples in various regional dialects. It demonstrates how different dialects are used to express the same sentences, highlighting the linguistic diversity across regions:

Table 3- 7 Region Wise Speech Accent Examples

Sentence (English)	Standard Gujarati	Kathiyawadi	Surti	Charotari	Kutchi
Where did you go?	તમે ક્યાં ગયા? (Tame kya gaya?)	ક્યાં પડ્યા તમણે? (Kya padya tamane?)	તમે ક્યા ગયા? (Tame kya gaya?)	તું ક્યા ગયો? (Tu kya gayo?)	ક્યા રવાય છે? (Kya ravaay chhe?)
It's scorching	આજ તપ્ત ગરમી છે. (Aaj tapt garmi chhe.)	આજ તાપ ચડે છે. (Aaj taap chade chhe.)	આજ ગરમ છે. (Aaj garam chhe.)	આજ બફાટ છે. (Aaj bafaat chhe.)	આજ તાપ ચડે છે. (Aaj

Sentence (English)	Standard Gujarati	Kathiyawadi	Surti	Charotari	Kutchi
hot today.					taap chade chhe.)
I need water to drink.	મારે પાણી પીવું છે. (Mare paani pivu chhe.)	મારે પાણું લાવવું છે. (Mare paanu lavvu chhe.)	મારે પાણી લેવું છે. (Mare paani levu chhe.)	મારે પાણુ જોઇએ. (Mare paanu joiye.)	મારે ઊરે પાણું લાવવું છે. (Mare ure paanu lavvu chhe.)
What did you eat?	તમે શું ખાધું? (Tame shu khadhu?)	તમે શું ખાધાં? (Tame shu khadha?)	તું શું ખાધું? (Tu shu khadhu?)	તું શું ખાધું? (Tu shu khadhu?)	તમે શું ખાધાં ભાઈ? (Tame shu khadha bhai?)
The whole village has come here.	આખું ગામ અહીં આવ્યું છે. (Aakhu gam ahi aavyu chhe.)	સારૂ ગામ આડે છે. (Saru gam aade chhe.)	ગામ ભેગું થયું છે. (Gam bhegu thayu chhe.)	ગામ બધું અહીં આવ્યું છે. (Gam badhu ahi aavyu chhe.)	સારૂ ગામ અહીં આવ્યું છે. (Saru gam ahi aavyu chhe.)
This thing is very good.	આ વસ્તુ બહુ સારી છે. (Aa vastu bahu saari chhe.)	આ વસ્તુ ઘણી સરસ છે. (Aa vastu ghani saras chhe.)	આ મજાની છે. (Aa majani chhe.)	આ સરસ છે. (Aa saras chhe.)	આ ખરું સારું છે. (Aa kharu saru chhe.)
Who brought all this?	આ બધું કોણ લાવ્યું? (Aa badhu kon lavyu?)	આ બધો કોણે ચડાવ્યું? (Aa badho kone chadavyu?)	આ બધું કોણ લાવ્યું? (Aa badhu kon lavyu?)	આ બધું કોણે લાવ્યું? (Aa badhu kone lavyu?)	આ બધો કોણ લાવ્યો? (Aa badho kon lavyo?)
Everyone in my house is caring.	મારા ઘરના બધા મમતા કરે છે. (Mara ghar na badha mamta kare chhe.)	મારા ઘરના માવડી ભલા છે. (Mara ghar na mavadi bhala chhe.)	મારા ઘરના લોકો સારાં છે. (Mara ghar na loko sara chhe.)	મારા ઘરના બધા મમતા કરે છે. (Mara ghar na badha mamta kare chhe.)	મારા ઘરના વંદા ભલા છે. (Mara ghar na vanda bhala chhe.)
This natural scenery is beautiful.	આ કુદરતી દ્રશ્ય સુંદર છે. (Aa kudrati drashya sundar chhe.)	આ કુદરતી દૃશ્ય સરસ છે. (Aa kudrati drashya saras chhe.)	આ દૃશ્ય મજાનું છે. (Aa drashya majanu chhe.)	આ કુદરતી દૃશ્ય ખૂબ સુંદર છે. (Aa kudrati drashya khub sundar chhe.)	આ કુદરતી દૃશ્ય વ્હાલું છે. (Aa kudrati drashya vhalu chhe.)
Where did your	તમારા ભાઈ ક્યાં ગયા છે? (Tamara bhai	તમારા ભાઈ ક્યાં દગ્યાં છે? (Tamara bhai	તમારા ભાઈ ક્યાં ગયા?	તારા ભાઈ ક્યાં ગયા?	તમારા ભાઈ ક્યાં ગયાં?

Sentence (English)	Standard Gujarati	Kathiyawadi	Surti	Charotari	Kutchi
brother go?	kya gaya chhe?)	kya dagya chhe?)	(Tamara bhai kya gaya?)	(Tara bhai kya gaya?)	(Tamara bhai kya gaya?)

From the Table 3-7 above, we can observe that different regions have distinct speech accents, even though the meaning of the sentences remains consistent. The people of Gujarat are known for their vibrant and colourful nature, which is also reflected in their speech. Gujarati, being the native language of Gujarat, deserves focused attention.

3.2.7 Voice Quality

3.2.7.1 Breathiness & Roughness

Breathiness and roughness are important features in terms of voice quality that represent a vital role in speakers' recognition. Mainly, these features reflect the tension of the vocal cords in the utterance and seriously influenced how the speaker's voice sounds to the listeners. Consequently, understanding these characteristics will significantly help in Speaker Recognition Systems to identify the speaker with some advanced applications in this regard, such as forensic or biometric voice analyses. Roughness is a voice quality characterized by irregularities in the vibration of the vocal cords, which produces a gravelly or raspy voice. It is often the result of tension and irregularities in the vibration of the vocal cords that cause a rough, uneven airflow.

Breathiness and roughness may, however, affect the sound of a sentence in Gujarati as in any other language.

- If someone says, "હું ગુજરાતી બોલી રહ્યો છું" ("I am speaking Gujarati") with a breathy voice, that would sound softer, more airy, almost whispered. This voice will have a gentle airy quality as if the speaker is uttering while exhaling.

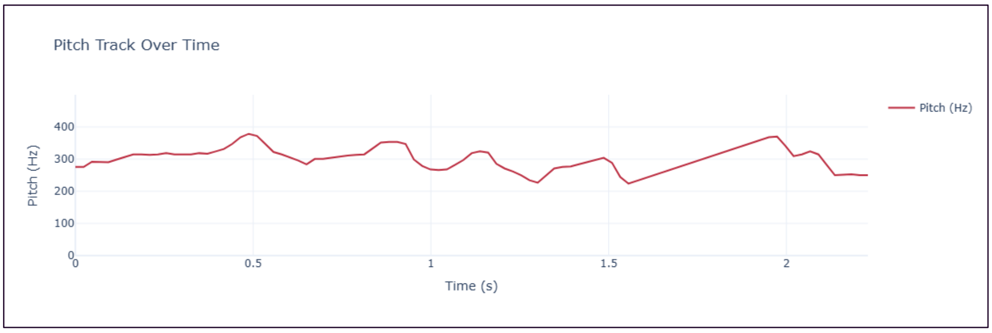


Figure 3- 15 Pitch Intonation of Sentence with Breathy Voice

The Pitch Intonation of sentence with breathy voice shows in Figure 3-15.

- If one speaks in a rough voice, the very same sentence would sound grating, hoarse, and with a raspy quality. He may sound as though he were straining his vocal cords or has a cold.

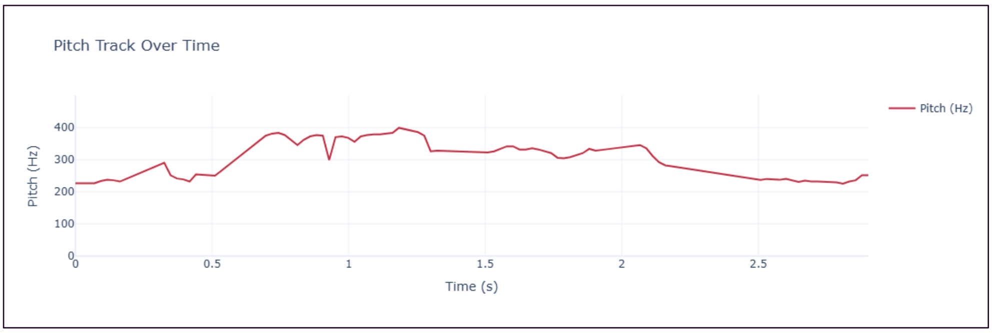


Figure 3- 16 Pitch Intonation of Sentence with Rough Voice

In both cases, the breathiness or roughness will change the unique acoustic signature of the voice shows in Figure 3-15 and 3-16, making it more distinctive from another speaker with a clear or non-breathy voice. Breathiness and roughness are two very critical aspects of voice quality and may contribute much to speaker recognition. They add unique acoustic patterns that might enable the systems to identify speakers even when this is hard to be done based on other features, such as pitch or speech rate.

3.2.7.2 Voice Clarity & Speech Rate

Rates and clarity make much difference to the conveyance in Gujarati. That is to say, the normal rate of delivery changes very fast depending upon emotional situations, as well as contexts wherein the utterances are going around. At casual, regular meetings, be it with close friends or family members, their speech velocity is moderate yet can rapidly increase when at periods the individual becomes overwhelmed with excitement while trying to have informal conversations with loved ones in everyday instances. For instance, phrases like below often see a faster delivery, reflecting a sense of urgency or enthusiasm.

- *"કેમ છો? તમે ક્યાં જઇ રહ્યા છો? હું તો તરત આવી જાઉં!"* (Kem cho? Tame kyan jai rahya cho? Hun to tarat aavi jao!), meaning "How are you? Where are you going? I'll be there right away!"

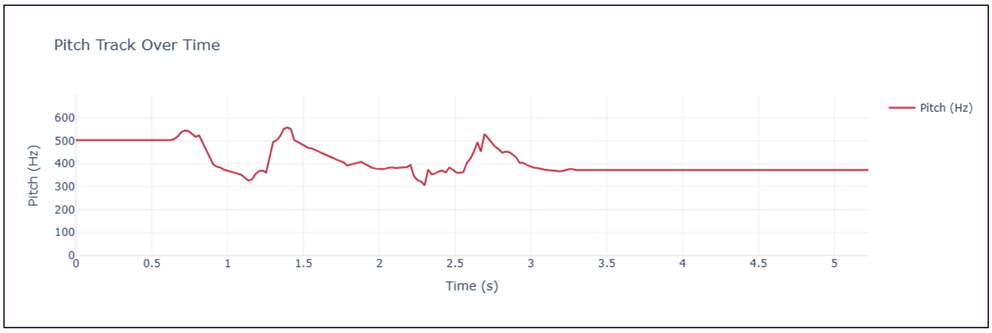


Figure 3- 17 Speech Rate of Informal Sentence

On the other hand, in more formal situations, like giving a presentation or participating in an interview, the pace slows down in order to ensure clarity, as seen in Figure 3-18.

- *"તમારા સૌનો આભાર. હું આ પ્રોજેક્ટના ફલસફાને પરિચય કરાવવાનો છું."* (Tamara sauno aabhar. Hun aa project-na philosophy-ne parichay karavva jai rahyo chhu), meaning "Thank you all. I am going to introduce the philosophy of this project."

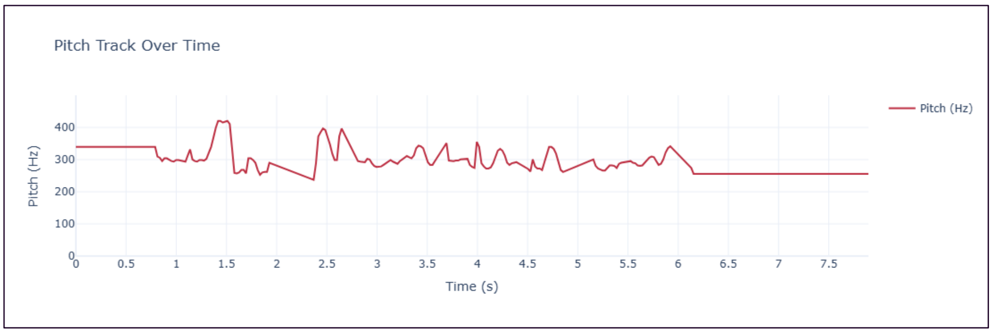


Figure 3- 18 Speech Rate of Formal Sentence

Clarity in Gujarati speech is heavily reliant on precise pronunciation, clear enunciation, and effective modulation of tone. A slight mispronunciation can completely alter the meaning of a word, as with

- *"સફળતા"* (Safalta - success) v/s *"સફલતા"* (Safalta - failure).

Clearness in Gujarati speech is much dependent upon correct pronunciation, articulation, and modulation of tone. A small mispronunciation can completely change the meaning of a word, as with Incorrect pronunciation of this subtle difference may result in confusion if not articulated well. Moreover, it is also important to put the right stress on a syllable. For example, stressing the right syllable may make all the difference between understanding a statement and misinterpreting it.

- *"તમે ખાવાનું આરંભ કરશો?"* (Tame khavanu aarambh karsho?), meaning "Will you begin eating?" is easily understood when the intonation rises at the end, signalling that it's a question. Using the right tone—rising for questions and falling for statements—is vital for accurate meaning.

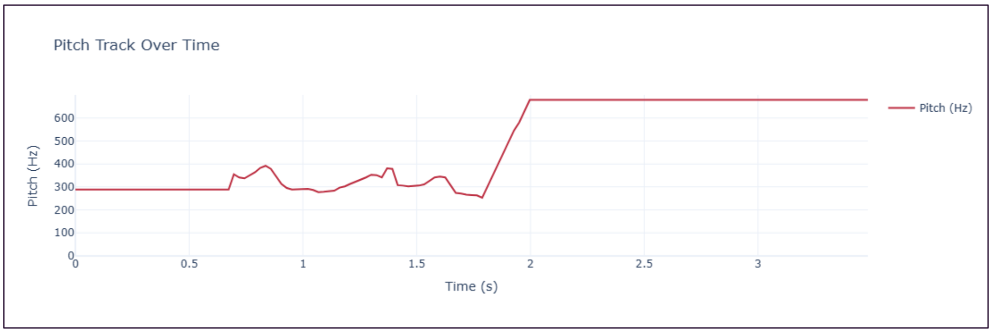


Figure 3- 19 Speech Rate of a Sentence Signalling Question

Regional variations also affect both speech rate and clarity. In urban areas, Gujarati speakers often speak fast, while in rural regions, the pace of speech is usually slow and more deliberate to ensure that each word is well enunciated. The context, whether formal or informal, also influences how fast one is speaking. Formal conversations, such as presentations or official dialogues, on the other hand, tend to be slower and more contemplative in their delivery because clarity in transmission necessitates it. By contrast, informal conversations are those that occur between two good friends or close family members. The speech can be a little faster, with many words shortened or merged in pronunciations for the convenience of communication.

Ultimately, speech rate and clarity are what would be required to master effectively in Gujarati. In adjusting the pace of speech and focusing on proper pronunciation and tone, speakers can ensure that their message is delivered with precision, making sure their audience understands them without any effort.[2].

3.3 Introduction to Feature Extraction Techniques

Feature extraction denotes the procedure of deriving certain information from a signal, including amplitude measurement, peak power, spectral density, and Hjorth parameters. The task is computationally demanding and entails mathematical analysis in univariate and multivariate contexts. Feature extraction can be conducted in laboratory settings utilizing local computers or cloud computing platforms, and it is essential for activities such as classification and data analysis in domains like Brain-Computer Interface (BCI)[4].

In speech processing, the procedure of extracting significant information from a voice signal while minimizing noise and extraneous data is termed feature extraction. The fundamental process of feature extraction includes spectral analysis, parametric transformation, and statistical modelling. The result is a parameter vector. Nonetheless, it is customary to forfeit valuable information when eliminating superfluous data. Feature extraction entails the conversion of the voice signal into a digital format[5]. The fundamental process of feature extraction is illustrated in below Figure 3-20.

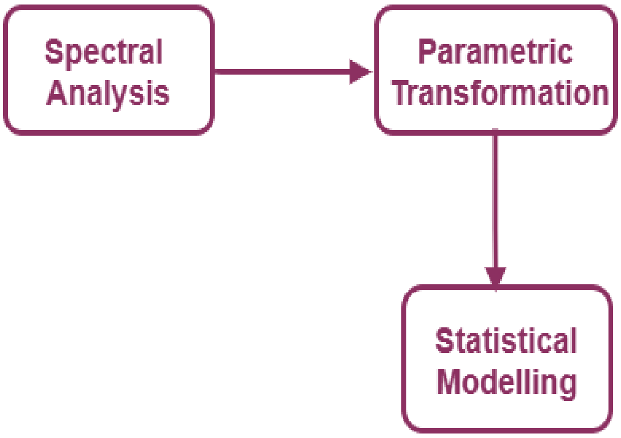


Figure 3- 20 Fundamental Process of Feature Extraction

Spectral Analysis constitutes the initial phase of speech analysis, encompassing the Spectro-temporal examination of the signal. In Parametric Transforms, two essential processes, differentiation and concatenation, are employed to generate signal parameters from signal observations. Signal parameters were derived from several underlying multivariate random processes during the Statistical Modelling phase[5].

The speech signal can be directly extracted from the digital waveform. Extensive speech signal data necessitates appropriate and dependable feature extraction methods. This can enhance performance and increase computational efficiency. It will exclude numerous sources of information, such as whether the sound is voiced or unvoiced, and whether speech is influenced by noise.

Speech processing involves with big amount of speech signal data. Consequently, data minimization is crucial for diminishing computing complexity and enhancing performance. However, data compression can result in the loss of significant speech signals[5]. The selection of a feature extraction technique is crucial for maintaining significant speech signals.

3.4 Different Feature Extraction Techniques for Gujarati Dialects

In voice recognition, accuracy and recognition rates diminish due to factors such as speaker variability; the utterances produced by the speaker may fluctuate with emotions and health conditions. Additionally, environmental unpredictability, speech signal noise due to the transmission channel, background noise, and reverberation compromise the integrity of the input voice signal during testing. The temporal function characteristic is inefficient as it varies considerably when the same speaker articulates the same utterance. The features that will yield accurate information and are resilient to noise should be computed in such instances. The features can be categorized as short-term spectral features, voice source features, Spectro-temporal features, prosodic features, and high-level feature. Feature extraction methods for voice recognition are typically classified as Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLPC), and Mel-Frequency Cepstral Coefficients (MFCC)[6].

3.4.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC[7] are a collection of features frequently employed in speech and voice recognition systems. They encapsulate the timbral attributes of sound, mirroring human auditory perception, especially with frequency and amplitude. This method is extensively utilized for Automatic Voice Recognition (ASR) because to its superior performance with clean data and minimal computing complexity. Performance deterioration is directly proportional to the signal-to-noise ratio (SNR); thus, performance significantly deteriorates when speech signals are contaminated by noise. In MFCC, the Mel-Spectrum is computed from a Fourier-transformed signal utilizing band-pass filters sometimes referred to as a Mel-filter bank. Here, Mel denotes the frequency recognized by human auditory perception. The Mel scale is linearly correlated with the physical frequency of speech tones below 1 kHz and logarithmically above 1 kHz[8] The Figure 3-21 shows the step-by-step process of extracting Features using MFCC techniques.

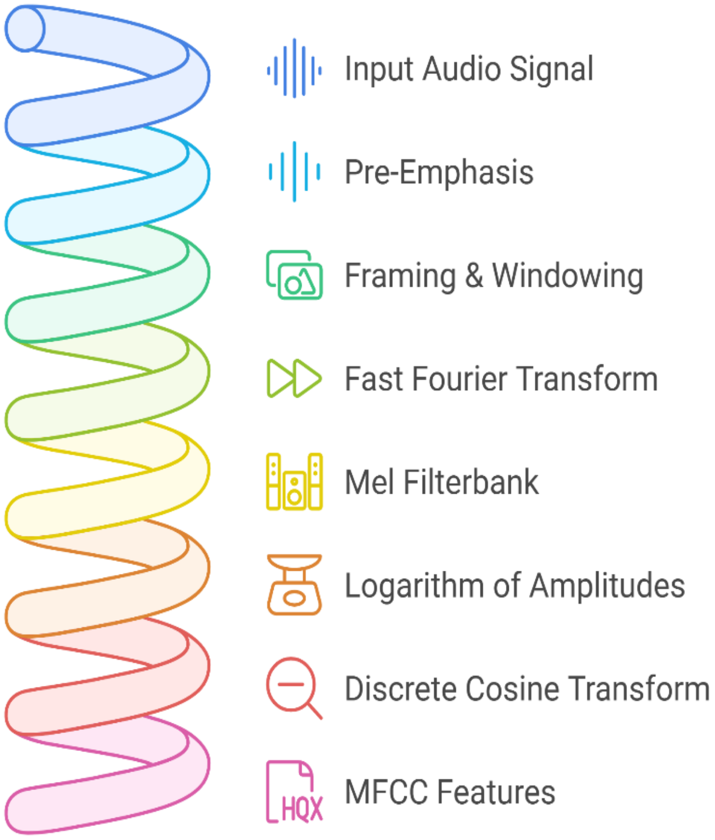


Figure 3- 23 MFCC Feature Extraction Process

3.4.1 Linear Prediction Cepstral Coefficients (LPCC)

LPC is used to estimate the spectrum of the signal. It predicts speech samples as a linear combination of past samples. This method minimizes the sum of squared error between past samples and linearly predicted samples over a defined window. The unique set of predictor coefficients can be determined by minimizing this error. The pre-emphasis of the speech signal is the first step in flattening the spectrum of the voice signal. Pre-emphasis amplifies the higher frequencies in the signal. Then, framing and multiplying by a window function is applied to diminish spectral leakage in the spoken frame. The vocal tract can be modelled as an all-pole model. It provides a set of autoregressive coefficients, known as Linear Prediction Coefficients (LPCs)[6].

One of the major useful methods in most of the speech processing aspects is. In most voice-communication systems, LPC is followed while encoding and decoding the speech to save bandwidth and maintain control over data bit rates. Yet this is a voice feature-extracting method, befitting for both speaker-dependent and speech recognition too. LPC exhibits poor performances, especially in the condition wherein the speech signal involves the presence of noise. The main aim of developing this technology is to develop a resonance structure exactly like the human vocal tract, generating the same sound[8].

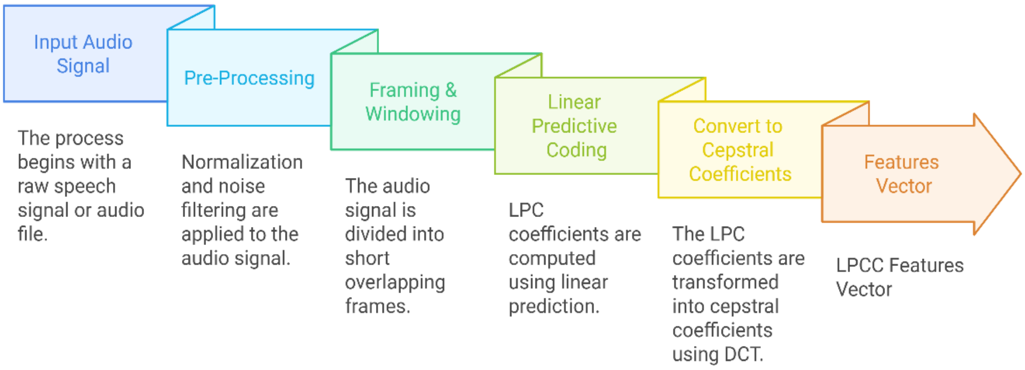

Input Audio Signal
The process begins with a raw speech signal or audio file.
Pre-Processing
Normalization and noise filtering are applied to the audio signal.
Framing & Windowing
The audio signal is divided into short overlapping frames.
Linear Predictive Coding
LPC coefficients are computed using linear prediction.
Convert to Cepstral Coefficients
The LPC coefficients are transformed into cepstral coefficients using DCT.
Features Vector
LPCC Features Vector

Figure 3- 26 LPCC Feature Extraction Process

The performance of LPCC is suboptimal in the presence of loud speech signals. To perform cestrum Linear Prediction (LP) analysis on a specified speech signal. The principle of employing LP analysis is to estimate the nth sample of a speech signal by incorporating a linear combination of the preceding p samples[8].The detailed process of extracting features using LPCC technique showed in Figure 3-22.

3.4.2 Perceptual Linear Prediction (PLP)

PLP is utilized to compute the power spectrum of the spoken signal. It alters the spectrum of the voice stream through various changes. The fundamental concept is to acquire the auditory spectrum and approximate it using an all-pole model. This method initially calculates a power spectrum estimate. The power spectrum is integrated via a Bark-scale filter bank, which simulates the crucial band frequency selectivity within the human cochlea[6].

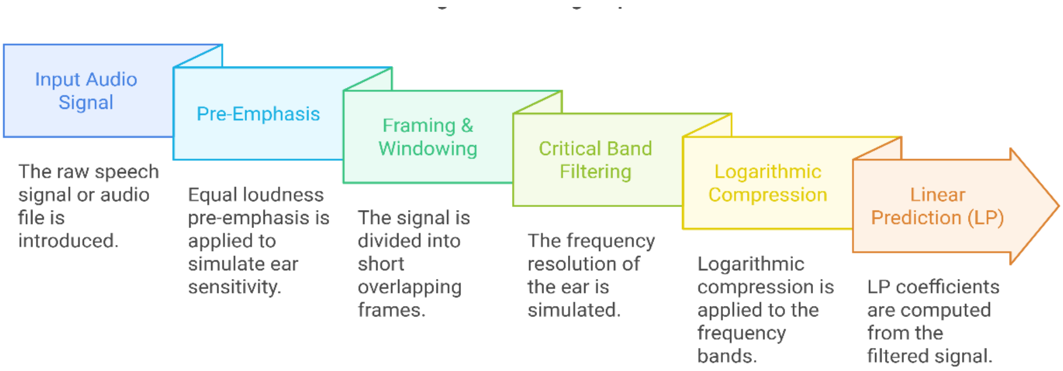

Input Audio Signal
The raw speech signal or audio file is introduced.
Pre-Emphasis
Equal loudness pre-emphasis is applied to simulate ear sensitivity.
Framing & Windowing
The signal is divided into short overlapping frames.
Critical Band Filtering
The frequency resolution of the ear is simulated.
Logarithmic Compression
Logarithmic compression is applied to the frequency bands.
Linear Prediction (LP)
LP coefficients are computed from the filtered signal.

Figure 3- 27 PLP Feature Extraction Process

The bark scale filters possess a trapezoidal configuration. The signal's pre-emphasis is executed via the equal loudness curve following frequency integration. The cube root of the power spectrum is utilized, as perceived loudness is approximately the cube root of intensity. The power spectrum is condensed in this phase. Subsequent to this stage, the inverse Fourier transform is applied to the filter outputs to derive the autocorrelation sequence, followed by the execution of Linear Predictive analysis to refine the spectrum. The ultimate features are derived through cepstral recursion from the LP coefficients. The PLP model is identical to the LPC model, except that in the PLP model, spectral features are modified to align with the attributes of the human auditory system as shown in Figure 3-23. The DFT and LP approaches are integrated within the PLP framework[6].

3.4.3 Discrete Wavelet Transform (DWT)

The primary goal of wavelet transform is to decompose a voice signal into a collection of functions referred to as wavelets. To analyse the frequency spectrum, WT employed a variable window to enhance the temporal resolution of speech analysis. It is advantageous for the analysis of non-stationary signals. Furthermore, it is a time-frequency transformation capable of multi-resolution analysis. DWT[9] is applied to an adaptable window size for the extraction of speech features. The objective of DWT is to decompose signals into sub-bands, facilitating the differentiation of features within each sub-band. The DWT parameters encapsulate the information of voice signals across many frequency scales. The wavelet transform technique efficiently represents non-stationary signals. This approach is sufficiently capable of extracting information in both the temporal and frequency domains from transient signals[8]. Figure 3-24 illustrates the procedures involved in the extraction of DWT features.

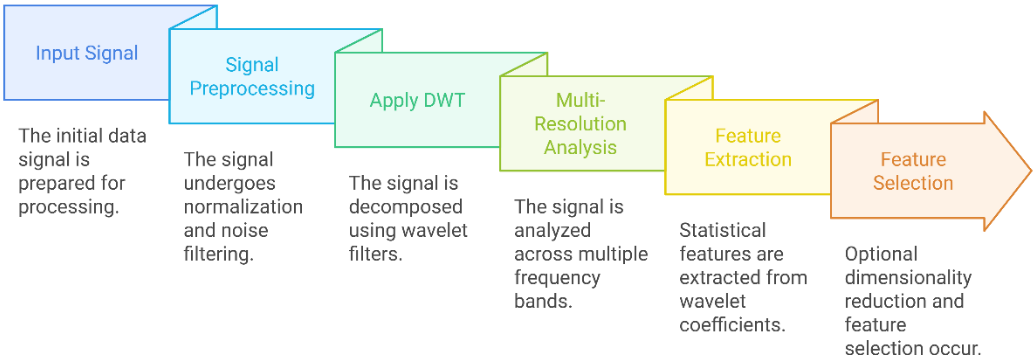

Input Signal
The initial data signal is prepared for processing.
Signal Preprocessing
The signal undergoes normalization and noise filtering.
Apply DWT
The signal is decomposed using wavelet filters.
Multi-Resolution Analysis
The signal is analyzed across multiple frequency bands.
Feature Extraction
Statistical features are extracted from wavelet coefficients.
Feature Selection
Optional dimensionality reduction and feature selection occur.

Figure 3- 28 DWT Feature Extraction Process

3.4.4 Fast Fourier Transform (FFT)

FFT is essential in Voice recognition systems as it helps in the conversion of raw audio signals from the time domain to the frequency domain. Voice recognition is about identifying or verifying a person based on his or her voice characteristics; hence, it depends highly on capturing the pattern of frequencies. In fact, it was here that the FFT plays the major role, changing the speech signal into its frequency components, whereby most of the important features representing pitch, harmonics, and formants can be brought out by the system. As a matter of fact, these elements are very particular and unique for every different speaker, which aids the system in distinguishing one voice from another[9], [10] Once the speech signal is converted with the FFT, the next most common steps involve the application of techniques such as MFCC and LPC. These methods further refine the frequency data, extracting more detailed features that are crucial for accurate voice recognition. The beauty of the FFT lies in its ability to work in the frequency domain, thus enabling voice recognition systems to model unique vocal traits-even in noisy environments. This makes the FFT a very important building block in voice biometrics or any state-of-the-art speech analysis where identification needs to be just correct. The below Figure 3-25 shows the process of FFT technique.

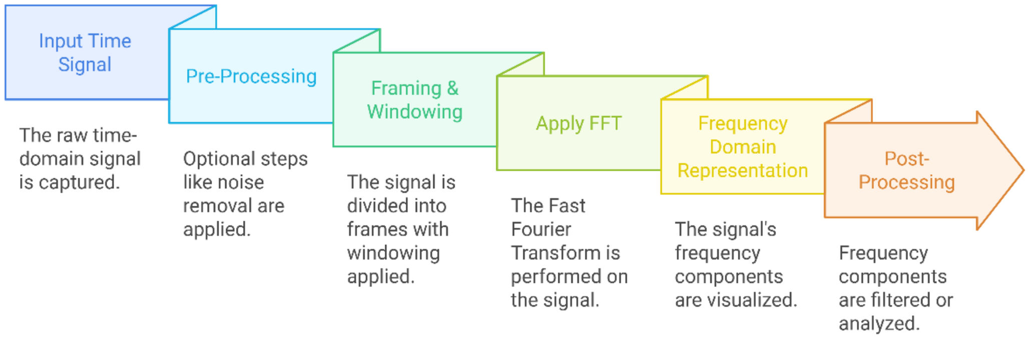

Input Time Signal
The raw time-domain signal is captured.
Pre-Processing
Optional steps like noise removal are applied.
Framing & Windowing
The signal is divided into frames with windowing applied.
Apply FFT
The Fast Fourier Transform is performed on the signal.
Frequency Domain Representation
The signal's frequency components are visualized.
Post-Processing
Frequency components are filtered or analyzed.

Figure 3- 29 FFT Feature Extraction Techniques

3.4.5 Relative Spectral (RASTA)

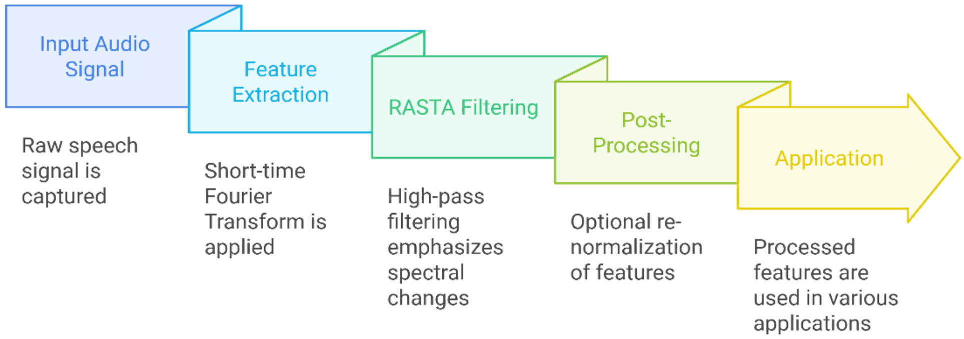

Input Audio Signal
Raw speech signal is captured
Feature Extraction
Short-time Fourier Transform is applied
RASTA Filtering
High-pass filtering emphasizes spectral changes
Post-Processing
Optional re-normalization of features
Application
Processed features are used in various applications

Figure 3- 30 RASTA Feature Extraction Process

RASTA processing shows in Figure 3-26 plays a crucial role in enhancing the reliability of speech features, especially in noisy or challenging environments[9]. It does this by filtering out the irrelevant distortions-for example, background noise or channel variations-and retains the most important dynamic speech features crucial in tasks such as voice recognition.

First, there is the extraction of speech-related basic features from raw audio signals, such as those generated by MFCC, which represent the spectral contents. Then, a filtering operation is performed through band-pass, with RASTA, in order to focus on temporal changes that are the most relevant in the frequency range from 1 to 20 Hz. This range is important for the capturing of unique vocal dynamics, including pitch and formants that make a voice unique. By removing both slow variations, which include noise, and fast fluctuations, probably caused by the communication channel, RASTA isolates the features.

Furthermore, RASTA cancels the effect of pre-emphasis, usually performed to amplify higher frequencies of speech. This step makes features even more robust and insensitive to the environment. Thus, this set of features generated becomes much more invariant, even against noisy or variable acoustic conditions, and therefore so ideal for voice identification and verification.

3.5 Evaluating Best Feature Extraction Technique for Gujarati Dialects

Feature extraction as a step is majorly important in any system designed for Voice Recognition as it enables the conversion of raw audio into a manageable dataset. There are several issues that arise from the different regions in the dialect of Gujarati: the size, the pronunciation, the intonation, and the pitch of the voice, all require proper consideration due to the diverse feature extraction that different regions present. The efficiency or otherwise of the feature extraction method compounded determines how well a system can distinguish between the different dialects. This section gives an assessment of the required feature extraction feature in speech recognition and determines the most suitable one for recognition of Gujarati dialects.

For Gujarati dialects, it is required that the linguistic diversity that best describes the phonetics, prosody and regional variations is well identified. We compared the above-mentioned features extraction techniques (section 3.4) to obtain the feature which classifies the different Gujarati dialects with better accuracy.

To evaluate the effectiveness of each feature extraction technique, the metrics that were considered are Accuracy, Precision, Recall, and F1 Score. The correctness of the entire system of predicting the correct dialect or phoneme for the input speech is based on accuracy. Precision, Recall, and F1 Score metrics allow their ability to correctly classify instances of different dialects while avoiding false positives and false negatives be assessed through the feature extraction method. Computation Time shows the real time performance of the recognition system also depends on the efficiency of the feature extraction process, and longer computation times can cause a noticeable deterioration in performance. The results, shown in the table below, demonstrate significant differences in performance across the methods:

Table 3- 8 Performance Measures Obtained from Various Feature Extraction Techniques

Sr. No	Metrics (%)	MFCC	LPC	FFT	DWT	PLP	RASTA
1	Accuracy	**96.84**	**97.00**	62.04	11.92	93.43	86.37
2	Precision	**95.82**	**96.90**	60.30	10.50	92.60	84.20
3	Recall	**97.50**	**97.10**	63.25	13.75	94.00	88.10
4	F1 Score	**96.65**	**97.00**	61.75	12.00	93.30	86.20

Our experiments showed that the MFCC technique achieved best overall performance over all the metrics, achieving 96.84% accuracy, 95.82% precision, 97.50% recall and 96.65% F1 score. They show that the most effective method of extracting these spectral and temporal features of speech for discriminating between Gujarati dialects is via extraction using MFCC. We show that the MFCC features adequately represent the short-term power spectrum of speech and are thus suitable to recognize dialectal differences in Gujarati.

While the LPC method performs well with accuracy of 97.00% it does not compare very much in precision and recall to MFCC. MFCC was slightly more precise (96.90%), but on par with its recall (97.11%) and F1 score (97.00%) with its. While LPC's performance was strong, it was unable to match MFCC's attention in capturing subtle regional differences.

The FFT method was working well and the accuracy of 62.04% and the precision of 60.30% are very poor. This means that FFT, which mostly focuses on processing frequency domain data, did not perform as well as was needed for accurate identification of Gujarati dialects. Even worse was the DWT method, which showed an accuracy of only 11.92 percent, which means it cannot be used for dialect classification.

Although neither FFT nor DWT reached the MFCC, the PLP technique showed an accuracy of 93.43%. Although it showed moderate precision (92.60%) and recall (94.00%) and can be useful in some noisy environments, the dialect classification performance was not as good in this study. Using this RASTA filtering approach, an accuracy of 86.37% was achieved on removing slow time variations in speech. At some improvement over FFT and DWT, it was not quite as effective as MFCC and LPC at capturing the relevant features of the English dialect.

It was found that the MFCC technique achieves highest recognition rate among all methods with respect to all of the key metrics. The greatest ability of dialect classification was to accurately capture the spectral and temporal characteristics of speech. Other methods like LPC and PLP respectively also did good, but did not hold MFCC overall performance. Future work could extend by combining MFCC with any of the other acoustic features, like pitch or formants improvement, in more challenging situations.

3.6 Evaluating Different Classifier for Voice Recognition

Classifiers are in charge of converting the features extracted from the speech signals to voice identity or speakers themselves. It was also revealed that the success of a Voice recognition system strongly depends on performance of the classifier used to differentiate the speaker’s voice in terms of the quality of the features extracted from the speech signal. The accuracy, robustness and real-time performance of the system strongly depends on the choice of the classifier, as is the choice classifier itself. An extensive literature survey of the different classification algorithms used in the voice recognition field is given, due to the different strengths and weaknesses of each. However, the classifiers vary in terms of complexity, computational cost, and their general performance under various acoustic conditions.

When working on Voice Recognition System, different classification algorithms will help a person identify the speakers and also recognize the dialects when working with Gujarati language. The biggest problem in dialect recognition generally involves phonetic, lexical, and acoustic differences among dialects within a single language. While dealing with voice recognition using Gujarati dialects, you would employ classifiers that can handle subtlety and variation due to dialects.

3.6.1 Hidden Markov Model (HMM)

Hidden Markov Models are a powerful tool for modelling speech, which is essentially sequential data. HMMs can model the sequential nature of speech with its intrinsic temporal characteristics, hence used for voice recognition and dialect classification. HMM is a rather rich mathematical structure used in modelling data in multiple applications like speech recognition, artificial intelligence, data compression, and pattern recognition. HMM stands for a non-stationary speech data modelling. HMM model represents speech as sequence of probable observation and defines in different states[11].

Because of its ability to model sequential data with temporal dependencies, Hidden Markov Models (HMMs) are a widely used statistical model in speech and voice recognition tasks. In HMMs, the system is arbitrarily divided into a finite number of states: at any particular time, the system is in one of these states and its sequence of states is a Markov process (i.e., the future state depends only on the current state and not on the history of the events leading up to the current time). This term means that the actual states are not directly observed but we have observations (including these extracted from speech in case of speech recognition) which the states emit, and use these observations to find the sequence of the state.

HMMs are used to model the temporal dynamics of speech in voice recognition, which allows a separating of different speakers. The set of HMM parameters unique for each speaker characterize the statistical properties of the speaker’s speech patterns with time. The model is trained on a sequence of features of speech samples and the task is to determine which speaker's HMM generated the speech we have observed.

Due to the fact that voice recognition is a challenging problem to solve as it must be able to accurately extract voice sounds from voices that change their pitch, speed, and intonation over time, HMMs are a particularly powerful technique for this task. They also are suitable to variable length sequences, because speech segments may be different than the length, but can be modelled with sufficiently high accuracy. But HMMs need a lot of data to estimate their parameters well and can be cumbersome if the non-linear relationship incurs more complex than the deep learning model such as CNN or RNN is capable of receiving.

3.6.2 Gaussian Mixture Model (GMM)

GMM is a voice recognition technology, which uses the segmented statistical features of the speech spectrum to identify the speaker. A probabilistic model that assumes data are generated by means of a mixture of several Gaussian distributions, each corresponding to a different component of the speaker's voice, the Gaussian Mixture Model (GMM). To characterize a speaker's unique speech properties, GMMs are used in voice recognition, to model statistics of a speaker's speech features, such as MFCC (Mel frequency coefficient) sets, that capture individual features of the speaker's speech. The system assumes that the input features come from a set of seconds (the GMM drawn from each second) based on each speaker’s speech samples, and calculates the likelihood of the input features under various speaker models identify or verify the speaker. The speech data is observed, and the Expectation–Maximization (EM) algorithm is used on the model to optimize the parameters of the Gaussian distributions to best match that observed speech data.

One major appeal of GMMs for speaker recognition is the flexibility of describing complex, multimodal distributions, and they are also probabilistic, offering them the ability to deal with uncertainty in speech data. However, they offer a natural way to calculate the likelihood of a speaker’s identity allowing towards voice identification and verification tasks. Nevertheless, GMMs suffer from high computational complexity at training time and a sensitivity to noise due to the speech signal.

3.6.3 K-Nearest Neighbors (KNN)

Classification tasks such as voice recognition often use a simple, instance-based learning method K-Nearest Neighbors (K-NN). K-NN classifies the input feature of an unknown speaker using the majority vote of K Nearest Neighbors from feature space, where K is a user defined constant. We call the closest labelled feature vectors in the training dataset 'Neighbors', usually use distance, for example, Euclidean distance. The problem of voice recognition involves extracting feature vectors from a set of speech produced by each speaker and then using K-NN to classify an unknown speech as the closest match with a known speaker feature set.

Despite its namesake, KNN is an insanely intuitive model that does not require any explicit model training. Its performance is based on the choice of K and the distance metric used. K-NN is one of the key strengths because it is simple, can handle complex, nonlinear decision boundaries. While it might be computationally expensive too, especially with big data, we have to calculate the distance to all points in the training set before classification. K-NN is less sensitive to noise and irrelevant features, but may affect the accuracy of the K-NN classifier in distinguishing between speakers.

3.6.4 Support Vector Machine (SVM)

Supervised learning algorithm used widely for classification tasks like voice recognition is called Support Vector Machine (SVM). Optimizing hyper plane in high dimensional feature space, SVM is used for finding the optimal hyper plane between data points of different classes. The features in voice recognition are MFCC or LPC and we need to find the words in the sentence that we think will give a feature which can distinguish the speakers. SVM is an algorithm that uses a hyperplane to build the one that maximize the gap between two examples of a class support vector. Increase in the margin improves the classifier’s generalization ability.

One of the major advantages of SVM in dealing with high dimensional feature space spaces makes it ideal for complex classification such as voice recognition. Also, SVMs are academically good to use if the data is not in a linear separable form and have a small-to-medium sized dataset. To this end, they can be used by use of Kernel functions like RBF kernel which maps the data in a high dimensional space. This flexibility allows for effective classification of speakers even when the underlying features have nonlinear relationships to each other. SVM, however, can be computationally expensive as C and gamma are tuning parameters, and SVM cannot be overfit or underfit if these parameters are not adjusted properly. While facing these problems, SVM has been used as one of the most successful and robust classifiers for voice recognition tasks in many aspects, such as distinguishing speakers from voice characteristics.

3.6.5 Convolution Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning models primarily used in image and spatial data processing but have also gained popularity in speech and voice recognition tasks. Unlike traditional machine learning models, CNNs automatically learn hierarchical feature representations directly from raw data (such as waveforms or spectrograms) without requiring explicit feature extraction. In voice recognition, CNNs are applied to spectrograms, mel-spectrograms, or other time-frequency representations of speech to learn features that are invariant to small variations in pitch, tone, or noise.

CNNs are particularly effective for voice recognition because they can capture local patterns in the input data, such as phonetic characteristics or speaker-specific traits, through convolutional layers. These layers apply filters to the input, detecting spatial hierarchies and features, such as edges or textures, that contribute to voice identity. The deeper layers of the CNN then learn increasingly complex patterns in the data, which enhances the model's ability to recognize speakers across different acoustic conditions. CNNs are known for their efficiency in handling large datasets and their ability to generalize well to unseen data.

However, CNNs require a substantial amount of training data to perform well, and their training can be computationally expensive, particularly with large datasets or deep architectures. Additionally, CNNs may not perform as well with very short speech samples or highly variable acoustic conditions, as they rely on the quality and quantity of the input data.

3.6.6 Long Short-Term Memory (LSTM)

Recurrent Neural Networks (RNN) is a type of neural network that takes an input sequence and produces an output sequence, similar to most machine learning tasks. Long Short-Term Memory (LSTM) networks are a particular kind of RNN and are great for problems with temporally dependent data (e.g slides containing data across time). LSTMs are unlike traditional RNNs in that they solve the problems with vanishing and exploding gradients the present in long sequence training. They accomplish this via their unorthodox architecture, which consists of memory cells and gates that determine the flow of information so the model can "remember" relevant information for a long period of time and "forget" irrelevant data.

In voice recognition LSTMs can take as input raw audio signals (for example, source signal itself) or more simple time-frequency encodings (such as MFCCs or spectrograms) and capture temporal interaction within the speech signal. More specifically, they are extremely effective for voice verification and identification tasks, because speech can be sequenced and voices can vary over time in pitch, tone, and rhythm, so this kind of representation has extra benefits. Being able to maintain context for longer with the LSTM makes it very good at separating speakers even in a noisy environment or when the speakers exhibited varying speech patterns.

The main nice properties of LSTMs in so voice recognition are their ability to model temporal dependencies and learn without any handcrafted feature extraction from unstructured sequential data. While LSTMs are computationally intensive and need quite a lot of data to train for deep architectures, however. Furthermore, they tend to overfit when the training data set is too small, or uniformly sampled.

3.6.7 Time Delay Neural Networks (TDNN)

Time Delay Neural Networks (TDNN) is a class of feed forward neural network used for sequential data; time delay is included in the model. There is some advantage to using TDNNs for speech and voice recognition tasks because they can capture time dependent patterns over different scales of time. This differs from traditional neural networks, which process data in a non-sequential, static way, but TDNNs allow the network to learn patterns in the input sequence through temporal delays, local and global context in the sequence.

TDNNs are used in voice recognition to feature such as MFCCs or filter bank coefficients that model the speech signal over time. Such relationships are learnt between consecutive frames of audio to identify speaker dependent characteristics (e.g. voice timbre, pitch and rhythm). Especially when the model needs to look for patterns across different Bins, TDDNNs are very powerful as they will allow you to keep those temporal relationships but without the complexity of the RNNs, which can be very computationally expensive.

One of the strengths of TDNNs is to model temporal dependencies over different time spans, which can facilitate modelling temporal dependencies even in a dynamic and fluctuating voice of a speaker in a voice recognition task. However, TDNNs also have their limitations: First, they are less efficient in training if they are given limited data, and second, they are quite sensitive to the size of the temporal window that we use to process.

3.6.8 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a family of neural networks that are created to work with sequential data (meaning they are a function of previous computations and output values at each time step). While traditional neural networks are feedforward, RNNs are cyclic (with loops in their architecture) that allow information to persist, and are especially well suited for tasks such as speech recognition, language modelling and voice identification. However, in a voice recognition context, RNNs, given MFCCs as input, are able to process such features over time to learn speaker specific temporal dependencies and patterns in speech.

Due to its ability to capture sequential relationships in speech signals, i.e.,.pitch, rhythm, and intonation, that are key for knowing a speaker, RNNs can learn these characteristics. RNNs are able to learn long term dependencies in speech and adapt to variable patterns of speech, such as speaking rate or speaking tone, because, as a voice is primarily consistent across time, it makes sense. This allows RNNs to work effectively in voice recognition tasks where temporal signal is critical for discriminating or confirming who a speaker was.

Nevertheless, RNNs suffer some limitations such as vanishing gradient (VGD) problem preventing the network from learning long range dependencies. Consequentially, RNNs can only learn over limited length of the sequence. When dealing with this problem we usually come up with some specialized variants such as Long Short-Term Memory (LSTM) networks or Gated Recurrent Units (GRUs) in order to keep track of long term dependencies while mitigating the vanishing gradient problem. RNNs are good for modelling temporal data, and are therefore easily applied to voice recognition, however they are computationally complex, and cannot handle long sequences with as much ease as other deep learning models like CNNs or TDNNs. But RNNs remain a powerful tool for speech processing applications where the sequence of data is important for this particular task.

References

[1] Manaorama Vyas, “PHONATION TYPES IN GUJARATI,” 1978.

[2] B. A. Alejandro Gutman, “The Language Gulper.” **https://www.languagesgulper.com/eng/Gujarati.html**

[3] Sakshi A. Patil, Gaurav A. Varade, and Vikram Hankare, “Tracing Gujarati Dialects Philogically and Sociolinguistically,” *International Journal of Modern Developments in Engineering and Science*, vol. 2, no. 5, May 2023, [online]. Available: https://www.ijmdes.com

[4] “Feature Extraction.” **https://www.sciencedirect.com/topics/computer-science/feature-extraction**

[5] M. A. Mazumder and R. A. Salam, “Feature extraction techniques for speech processing: A review,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.3 S1, pp. 285–292, 2019, doi: 10.30534/ijatcse/2019/5481.32019.

[6] S. B. Dhonde and S. M. Jagade, “International Journal on Recent Technologies in Mechanical and Electrical Engineering (IJRMEE) Feature Extraction Techniques in Speaker Recognition: A Review,” 2015, [Online]. Available: http://www.ijrmee.org

[7] S. T. Aung, H. Myo Tun, Z. M. Naing, and W. K. Moe, “Analysis of Speech Features Extraction using MFCCs and PLP Extraction Schemes.” [Online]. Available: www.ijsetr.com

[8] N. Singh, N. Parveen, P. Chandra, and B. Banarasi, “Feature Extraction Algorithms for Speaker Recognition System and Fuzzy Logic,” *International Journal of Advanced Science and Technology*, vol. 29, no. 7s, pp. 3068–3076, 2020, [Online]. Available: https://www.researchgate.net/publication/341576926

[9] S. Ajibola Alim and N. Khair Alang Rashid, “Some Commonly Used Speech Feature Extraction Algorithms,” in *From Natural to Artificial Intelligence - Algorithms and Applications*, IntechOpen, 2018. doi: 10.5772/intechopen.80419.

[10]Jasmeet Kaur Hundal and Dr. S. T. Hamde, “Some Feature Extraction Techniques for Voice based Authentication System,” May 2013, [Online]. Available: http://arxiv.org/abs/1305.1145

[11]J. H. Tailor and D. B. Shah, “HMM-Based Lightweight Speech Recognition System for Gujarati Language,” in *Lecture Notes in Networks and Systems*, vol. 10, Springer, 2018, pp. 451–461. doi: 10.1007/978-981-10-3920-1_46.