A Brief Overview of Speech Recognition: Hindi Language Perspective

Rajiv Kumar, Mridul
Shobhit Institute of Engineering and Technology (Deemed to be University), Meerut
Email Id- Rajiv.kumar@shobhituniversity.ac.in, mridul@shobhituniversity.ac.in

ABSTRACT: ***Information access in a convenient manner has become increasingly important in this age of information technology. People naturally expect to be able to have an outspoken conversation with a computer because speaking is the fundamental means of communication among humans. Ordinary people may talk to a computer to get information using a speech recognition system. A human-computer communication in the local language is desirable. Because Hindi is India's most commonly spoken language, it is the most obvious option for human-machine interaction. In Hindi, there are five pairs of vowels, one of which is longer than the other. The purpose of this study is to provide an overview of voice recognition systems. The qualities and characteristics of Hindi Phoneme, as well as how speech is formed.***

KEYWORDS: ***Acoustic modelling, Language model, MFCC, Mel-Frequency Cepstral Coefficients, Speech Recognition.***

1. INTRODUCTION

The ordinary man in India would be able to profit from information and communication technology if it is feasible to enable human-like contact with machines. The acceptability and usability of information technology by the general public will skyrocket in this situation. Furthermore, because 70% of the Indian population lives in rural regions, having a speech-enabled computer application designed in their native language becomes even more vital [1]. It's worth noting that, in recent decades, research has focused on continuous, large-vocabulary speech processing systems for English and other European languages, whereas Indian languages such as Hindi and others have received less attention. India is in the midst of a technological revolution. As a result, it is past time to create voice recognition technology for Indian languages. Speech recognition is the capacity to listen to spoken words (input in audio format) and recognize various sounds contained in them as words of a recognized language [1]. Speech recognition in the computer realm entails a number of phases, each with its own set of problems. Voice recording, word boundary detection, feature extraction, and recognition with knowledge models are the processes necessary to make computers do speech recognition [2].

The method of determining the start and end of a spoken word in a given sound stream is known as word boundary detection. It might be difficult to detect the word border while studying the sound stream. This may be ascribed to a variety of accents, as well as the length of the gap between words when speaking. Feature extraction is the process of converting a sound signal into a format that may be used in subsequent stages [3]. Extracting characteristics such as the signal's amplitude, frequency energy, and so on are examples of feature extraction. The process of sound recognition entails mapping the input (in the form of different characteristics) to one of the recognized sounds. For exact identification and ambiguity elimination, this may include the usage of multiple knowledge models. Knowledge models are models that aid the recognition system, such as phone acoustic models, language models, and so on [4]. The system must be trained in order to create the knowledge model. During the training stage, the system must be shown a set of inputs and the outputs to which they should be mapped. We explain the creation of a Hindi language voice recognition system in this work.

2. DISCUSSION

2.1. Existing System:

Although there are several promising voice synthesis and recognition solutions available, the most of them are English-centric. These systems' acoustic and linguistic models are for the English language. Before they can be utilized, the majority of them require several settings. There have also been attempts to convert it to Hindi and other Indian languages [5]. Explains how an acoustic model for English may be created from an existing acoustic model. SIP and Sphinx are two well-known open-source speech recognition programs. A comparison of voice recognition software available in the public domain. There is also commercial software available, such as IBM's Via Voice.

2.2. How Speech is produced?

The vocal organs shown in Figure 1 create human speech, with the lungs and diaphragm as the primary energy source. The vocal tract begins in the glottis, or entrance of the vocal cords, and ends at the lips [6]. The pharynx (the link between the esophagus and the mouth) and the mouth, or oral cavity, make up the vocal tract. The cross-sectional size of the vocal tract varies from zero to roughly 20 cm^2 depending on the location of the tongue, lips, jaw, and velum. The nasal tract runs from the velum to the nostrils. The nasal tract is acoustically linked to the vocal tract when the velum is lowered, resulting in nasal speaking sounds.

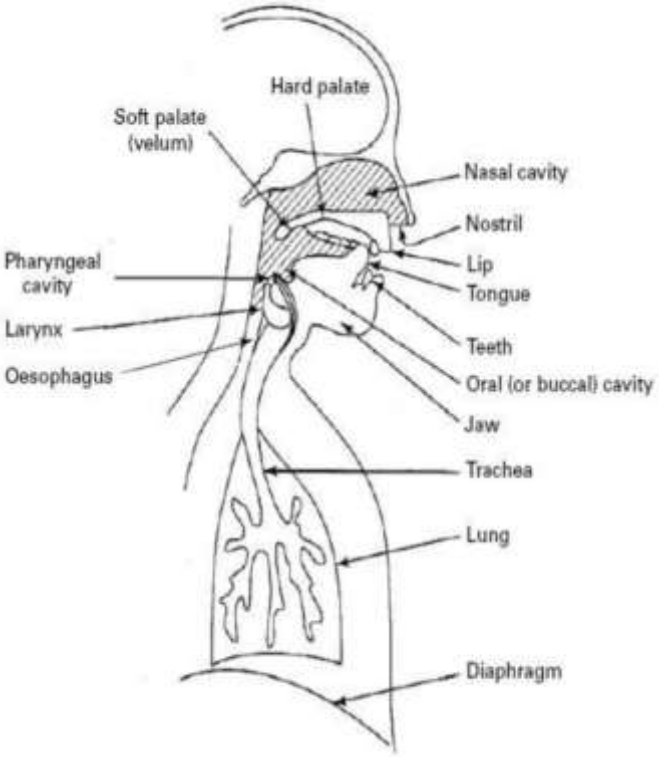

Figure 1: Diagram of the human speech production mechanism [7]

The regular breathing process allows air to enter the lungs. The air flow causes the tensed voice cords within the larynx to vibrate when air is evacuated from the lungs through the trachea. The air flow is split into quasi-periodic pulses that are modified in frequency as they pass through the throat, mouth cavity, and potentially nasal cavity. Different sounds are generated depending on the location of the various articulators (jaw, tongue, velum, lips, and mouth).

2.3. Architecture of Speech Recognition System:

Pattern recognition is a subset of speech recognition. Figure 2 depicts the stages of voice recognition processing. Training and testing are the two steps. In both phases, the method of extracting characteristics that are useful for categorization is the same. The parameters of the classification model are calculated using a large number of class exemplars during the training phase. During the testing step, each class's trained model is matched with the feature of the test pattern. The test pattern is assigned to the class whose model most closely matches the test pattern.

Speech recognition aims to create the best possible word sequence while keeping linguistic limitations in mind. The phrase is made up of linguistic elements such words, syllables, and phonemes. A sentencing model is believed to be a collection of smaller models in speech recognition. To hypothesis the sentences, the acoustic data supplied by the acoustic models of such units is coupled with the criteria for creating valid and understandable phrases in the language. As a result, the pattern matching system in voice recognition may be divided into two domains: acoustic and symbolic. A featured vector corresponding to a short segment to test speech is matched with the acoustic model of each class in the acoustic domain. The label of the class with the greatest matching score is applied to the section. For each feature vector in the feature vector sequence generated from the test data, the label assignment procedure is repeated. The recognized sentence is produced by combining the generated sequence of labels with the language model.

Figure 2 depicts the fundamental framework of a voice recognition system.

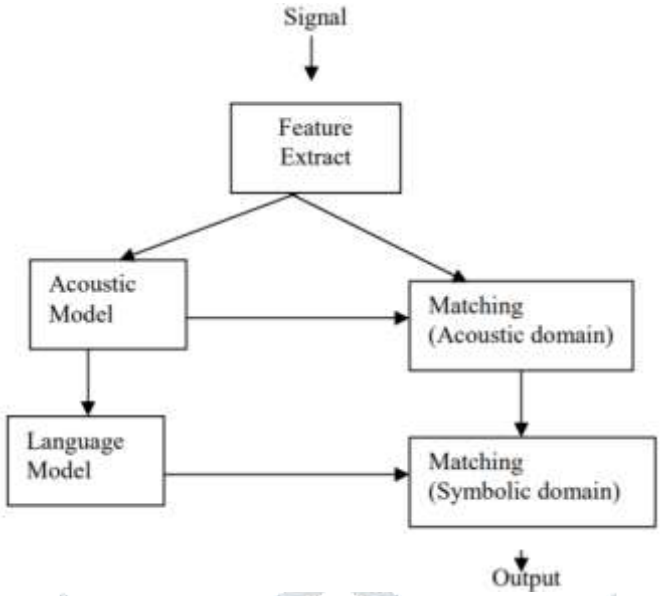

Figure 2. Illustrates a Speech Recognition System block diagram [8]**.**

2.4. Speech Signal:

Speech travels across a medium like air or water as a longitudinal wave. The speed of propagation is determined by the medium's density. A speech pressure waveform, or simply a speech waveform, is a graphic that shows the amplitude of air pressure fluctuation corresponding to a voice signal as a function of time. At the moment of recording, the speech signal is an analog signal that changes with time. To process the signal digitally, the continuous-time signal must first be sampled into a discrete-time signal, and then the discrete-time continuous valued signal must be converted into a discrete-time, discrete valued, or digital signal. We may break speech into a sequence of uncorrelated segments, or frames, and analyze the sequence as if each frame has fixed attributes since the properties of a signal change very slowly with time. We can extract the characteristics of each frame based on the sample inside the frame only if we make this assumption. And, in most cases, the feature vector will replace the original signal in further processing, implying that the speech signal is transformed from a time-varying representation of events in probability space, a process known as Signal Modeling.

2.5. Feature Extraction:

Frequency or spectrum analysis is a frequent stage in feature extraction. The goal of signal processing techniques is to extract features that are connected to the characteristics in order to identify them. The objective of feature extraction is to identify a collection of utterance qualities that have acoustic correlations in the spoken signal, i.e. parameters that can be approximated in some way using signal waveform processing. Features are the names given to these factors.

There are several distinct feature extraction algorithms, like as [9]

- Linear Predictive Cepstral Coefficients(LPCC)
- Perceptual Linear Prediction(PLP) Cepstra
- Mel-Frequency Cepstral Coefficients (MFCC)

Before converting it to Cepstral coefficient, LPCC computes the Spectral envelop. The LPCC is a cepstral coefficient generated from LP. The PLP combines crucial bands, equal loudness pre-emphasis, and compressed loudness intensity. The Nonlinear Bark scale is used to create the PLP. The PLP is a voice recognition system that eliminates speaker-dependent features. ASR has a lot of MFCC. MFCC is based on signal decomposition with the assistance of a Mel scale-based filter bank. The Mel frequency scale is based on the Discrete Cosine Transform (DCT) of a real logarithm of a short-time energy. The majority of study is focused on 12MFCC. The cepstral coefficients are a set of features that have been shown to be reliable in a variety of pattern recognition tasks using human speech. The human voice is well-suited to ear sensitivity. Twelve coefficients are preserved in voice recognition tasks, which indicate the slow fluctuation of the signal spectrum, which determines the vocal tract form of the spoken syllables. Mel-Frequency (Mel-Frequency) is a frequency that The Cepstrum Coefficient method is frequently used to produce sound file fingerprints. The MFCC is based on the known fluctuation of the crucial bandwidth

frequencies of the human ear, with filters separated exponentially at low frequencies and nonlinearly at high frequencies to capture significant speech features. Human perception of the frequency contents of sound for speech signals does not follow a linear scale, according to studies. As a result, a subjective pitch is assessed on a scale known as the Mel-scale for each shop with an actual frequency measured in Hz. The Mel frequency scale has a logarithmic spacing above 100Hz and is below 1000 Hz. The pitch of 1 KHz, tone, 40db just above sensory threshold of hearing is specified as 1000 Mels as a reference point. The Mel for a given frequency are calculated using the following formula:

$$\textbf{Mel (f) = 2595*log10(1+f/700)}$$

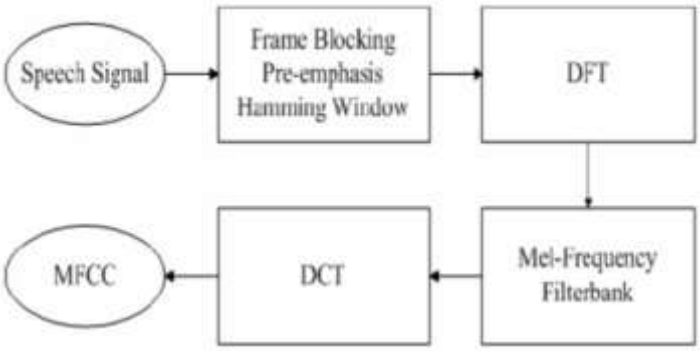

Figure 3: Illustrates the block diagram of the MFCC Processes.

Blocking the speech waveform is clipped in frame to eliminate any quiet or acoustical interference that may exist at the start or finish of the sound stream. By tapering the beginning and end of each frame to zero, the windowing block reduces signal discontinuities. Each frame is converted from the time domain to the frequency domain using the Fast Fourier Transform (FFT) block. The signal is filtered in the Mel Scale Filter block using a band-pass filter whose bandwidths and spacing are approximately equivalent to crucial bands and whose center frequency range includes the most significant frequencies for speech perception (300- 5000Hz). Cepstrum is created in the Inverse Discrete Fourier Transform (IDFT) block. Where Cepstrum is the log of the spectrum's spectrum. The MFCC characteristic is taken into account for both speaker-independent speech recognition and the speaker recognition job. The Table 1 below shows a comparison between LPCC and MFCC.

Table 1: Comparison between MFCC and LPCC

Categorization	MFCC	LPCC
Low Bandwidth	Higher result	Less Effective
Noisy	Effective	Less Effective
Vocal Tract	Yes	No
Human Ear	Good	Bad

2.6. Acoustic Modeling:

The retrieved features from the Feature Extraction module are compared to a model to determine the sound that was generated as the spoken word. Acoustic model is the name given to this model. The acoustic information and phonetics are established in this. Model as a speech signal is used to map a speech unit to its acoustic equivalent. The Hidden Markov Model is the most often used acoustic model (HMM) [10]. Word Model and Phone Model are the two kinds of acoustic models. Small vocabulary systems often utilize word models. The words are modeled as a whole in this model. As a result, each word must be modeled individually. We'll have to teach the system to identify a new term if we need to add support for it. The sound is compared against each of the models in the recognition process to get the best match. The uttered word is considered to be the best match. Instead of modeling the whole word, we model just portions of the word typically phones in Phone Model. The word is modeled as a phonetic sequence. The components are now identified once the heard sound is matched against them. The pieces that have been identified are combined to form a word. The word ak, for example, is formed by combining the letters a and k. This is helpful when a big vocabulary system is required. Adding a new word to the lexicon is simple since the sounds of phones are already known; all that is required is the potential phone sequence for the word and its likelihood.

2.7. Language Modeling:

The language model's aim is to generate an accurate probability value for the word W, Pr(w). The structural restrictions present in the language are used to create the probability in a language model [11]. Although there are words with identical sounding phonemes, people usually have no trouble recognizing them. This is due to the fact that they are familiar with the context and have a decent notion of what words or phrases may appear in it. The aim of a language model is to provide this context to a voice recognition system. The language model defines which words are valid in the language and in what order they may appear. Small vocabulary restricted activities, such as phone dialing, may often be represented using a grammar-based method, while big applications, such as broadcast news transcription, need a stochastic approach.

2.8. Acoustic-Phonetic Feature of Hindi:

Hindi's acoustic-phonetic system is distinct from that of European languages. The Hindi alphabet has ten vowels (two of which are diphthongs), four semivowels, four fricatives, and twenty-five stop consonants (including 5 nasals). In most Indian languages, the stop consonants are arranged in a systematic way, and this order may provide ideas for creating a recognition system. We discussed the characteristics of Hindi vowels and consonants in this paper. Table 2 shows the results. It is divided into three sections: The first part comprises of vowels, whereas the second section consists of phonemes that need full closure of the oral tract to produce. Semivowels and fricatives make speech sounds by moving the mouth from one position to another in the third segment.

Table 2: Illustrates the Hindi alphabet. The matching IPA symbol and the ASCII representation used in the database are displayed in the cell's second and third rows, respectively, for each phoneme.

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
a	*aː*	*i*	*iː*	*u*	*uː*	*e*	*eː*	*o*	*oː*
a	A	i	I	u	U	e	E	o	O

क	ख	ग	घ	ङ
k	k^h	*g*	g^h	ŋ
k	kh	g	gh	gñ
च	छ	ज	झ	ञ
tʃ	$tʃ^h$	*ʤ*	$ʤ^h$	ɲ
c	ch	j	jh	jñ
ट	ठ	ड	ढ	ण
ʈ	$ʈ^h$	ɖ	$ɖ^h$	ɳ
T	Th	D	Dh	N
त	थ	द	ध	न
t	t^h	*d*	d^h	*n*
t	th	d	dh	n
प	फ	ब	भ	म
p	p^h	*b*	b^h	*m*
p	ph	b	bh	m

य	र	ल	व	श	ष	स	ह
j	*r*	*l*	*ʋ*	*ʃ*	*ʂ*	*s*	*h*
y	r	l	w	s~	S	s	h

The Hindi alphabet has ten vowels (two of which are diphthongs), four semivowels, four fricatives, and twenty-five stop consonants (including 5 nasals). In most Indian languages, the stop consonants are arranged in a systematic way, and this order may provide ideas for creating a recognition system. We discussed the characteristics of Hindi vowels and consonants in this paper. Table 2 shows how vowels, consonants, and semivowels are arranged. There are three kinds of vowels. There's a front, a middle, and a rear. The three fundamental short vowels are (A), (E), and (U). The vowels (a), I and (u) are all long vowels. Diphthongs (ai) are a combination of two vowels (au).

Stop consonants are made by shutting the mouth cavity fully and then releasing the built-up air pressure. Various consonants have different sites of closure in the oral cavity: the glottis is the innermost point of closure, and the back, middle, and front portions of the tongue push against the corresponding areas of the upper palate are the other points of closure. Finally, beyond the teeth, the lips must be closed. As a result, the five rows in Table 2 indicate five distinct consonant classes. These relate to the five articulation points.

Non-nasalized consonants are the first four consonants in each row in Table 2. The first two consonants are nasal consonants, the following two are voiced, and the fifth is unvoiced. The un-aspirated kinds are shown in the first and third columns, while the aspirated types are represented in the second and fourth columns. The pattern of classifications according to the location of articulation and mode of production is shown in table 2 with the consonants arranged rows-wise and columns-wise. Each column in Table 2 represents a phoneme and contains three rows: the Devanagari script on the first row, the IPA symbol on the second row, and the roman character used to identify the phoneme in a spoken Hindi phrase on the third row. Each consonant has been represented using English letters. For example, the letter ‘H' stands for the sound ‘th,' whereas ‘W' stands for the vocal version ‘dh.'

3. CONCLUSION

The development of a human-oriented computer interface is urgently needed. Humans continue to utilize spoken languages as a method of communication. In a multilingual nation like India, the ability to communicate with computers in one's own language is critical. This paper explains the fundamentals of voice recognition systems and the characteristics of Hindi acoustic-phonetics.

REFERENCES

[1] “About 70 per cent Indians live in rural areas: Census report,” 2016. https://www.thehindu.com/news/national/About-70-per-cent-Indians-live-in-rural-areas-Census-report/article13744351.ece (accessed Sep. 15, 2018).

[2] S. Sinha, S. Sharan, and S. S. Agrawal, “O-MARC: A multilingual online speech data acquisition for Indian languages,” in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Nov. 2017, pp. 1–6, doi: 10.1109/ICSDA.2017.8384464.

[3] S. Narang and D. Gupta, “Speech Feature Extraction Techniques: A Review,” *Int. J. Comput. Sci. Mob. Comput.*, 2015.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, “Fundamentals of speech recognition,” in *Robust Automatic Speech Recognition*, 2016.

[5] R. K. Aggarwal and M. Dave, “Integration of multiple acoustic and language models for improved Hindi speech recognition system,” *Int. J. Speech Technol.*, 2012, doi: 10.1007/s10772-012-9131-y.

[6] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *J. Acoust. Soc. Am.*, 2007, doi: 10.1121/1.2404622.

[7] H. C. Mahendru, “Quick review of human speech production mechanism,” *Int. J. Eng. Res.*, 2014.

[8] S. Naziya S. and R. R. Deshmukh, “Speech Recognition System – A Review,” *IOSR J. Comput. Eng.*, 2016, doi: 10.9790/0661-1804020109.

[9] A. Madan and D. Gupta, “Speech Feature Extraction and Classification: A Comparative Review,” *Int. J. Comput. Appl.*, 2014, doi: 10.5120/15603-4392.

[10] W. W. Liu, M. Cai, H. Yuan, X. B. Shi, W. Q. Zhang, and J. Liu, “Phonotactic language recognition based on DNN-HMM acoustic model,” 2014, doi: 10.1109/ISCSLP.2014.6936704.

[11] R. Rasipuram and M. Magimai-Doss, “Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model,” *Speech Commun.*, 2015, doi: 10.1016/j.specom.2014.12.006.