CHAPTER 3

HUMAN SPEECH PRODUCTION MECHANISM

Chapter Overview:

- Background
- Mechanism of Speech Production
- Psychology of Speech Production
- Biology of Speech Production
- Speech Sound
- Difficulties in Speech Research
- Concepts of Mathematics used in Speech Processing
- Filter and Convolutions
- Sampling of the Speech Signal

CHAPTER 3

HUMAN SPEECH PRODUCTION MECHANISM

3.1: Background:

"India is the birthplace of human speech, the mother of history"

- Mark Twain

Speech is the secondary task of our mouth and is the most convincing way of communication among human and best known medium of the thought process exchange in the universe. Mechanizing this primary tool of human interaction by analyzing speech has attracted researchers for decades. Serious efforts are made by scientists' and researchers from various disciplines in developing voice synthesis system, followed by recognition, Identification and Authentication system using machines. This is considered to be the most obvious and spontaneous way for communication by us. Voice comes out in the form of series of sounds. Speech is extremely abstract signals by nature to manage and analyze. Lot of efforts have been put by researcher community all around the world in this area in recent times with a agreed upon philosophy that voice recognition possess the capability and potential to substitute others existing forms as the best ever interface.

The tremendous growth in information technology (IT) sector, have given rise to the projects targeting the human-machine communication, using natural speech, and interest growth is from both the academic and business communities. Speech researchers are challenged by the task of developing a system that can understand natural language and authenticate the speaker. In a spoken language inherent challenges are:

- Sentence vocabularies are infinite.

- Ambiguity exists within it. Several Words can have more than one meaning.
- Single utterances contain/sends context wise separate messages.

Characteristics mentioned above pose a huge challenge in the development of a SAS. To understand the message, any system should have capacity to absorb the linguistics of the natural language considered for processing. An efficient program must possess knowledge of the language structure (consists of words/phrases/sentences). Also one should have adequate information of the meaning along with their context. Classifications of the linguistic knowledge necessary to understand the language are described in brief below:

Table 3.1: Classifications of the linguistic knowledge

Phonology	**It relates sounds to the words we recognized.**
Morphological	Involves in morphemes forming meaningful words.
Syntactic	Indulge in formation of grammatically proper sentence.
Semantic	Involves in giving meaning to a sentences.
Pragmatic	Context difference is dealt with by pragmatism. At the same time indicates the contextual inferences in conveying the meaning within.
World	Consists of language information, understood and perceived by speakers, carries further in conversation. This lets us get the speaker's logic and senses.

3.2 Mechanism of Speech Production:

Remarkable progresses have been made in developing systems that can interact with human voice using natural language and IT devices. In near future human-machine communication will definitely be a ground level reality with the rapid progress made by scientists in natural language processing arena. In the forward stride to have a

computer to imitate human voice, there must be adequate knowledge of human speech mechanisms by which speech is produced and perceived. It will immensely help in developing speech information processing technologies that utilizes these functions. Although speaking is a natural phenomenon and mostly an effortless regular activity, perfect knowledge of the physiology and anatomy and speech production mechanism will increase clarity how the brain processes information. It will definitely help in marching ahead towards the development of better natural voicing instruments.

3.2.1 Psychological Aspects Related to Speech Production:

A series of sounds that involves the articulators of our body like mandible, tongue, lips, teeth etc are perceived as a speech wave. The shapes, sizes and altered positions over time to produce sounds bring about differences to utter different sounds for different meaning and context. The Psychological aspects leads us to divisions of the speech production process in three different stages [Figure 3.1] described below:

1. Conceptualization. 2. Formulation. 3. Articulation.

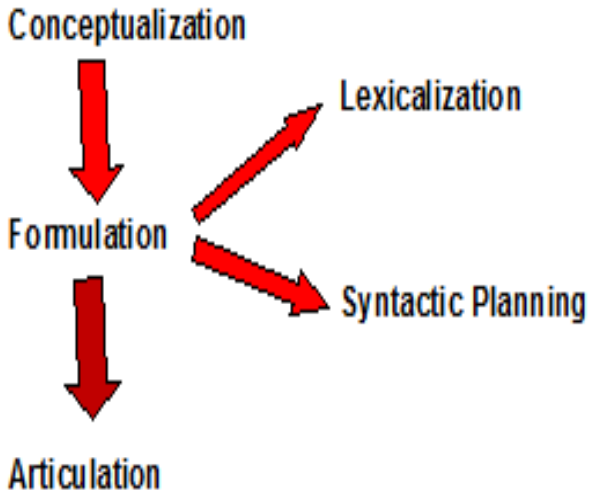

Figure 3.1: Different Stages of Speech Process

The process of speaking begins with central nervous system (brain) as a thinking procedure which is a pre-verbal message and the process is said to be conceptualization. Next phase is speech formulation where human thought (pre-verbal message) gets a makeover into linguistic formation and termed as speech formulation.

There are two sub-stages:

a) Lexical formation: Speaker's mind gets expressed in the format of sounds carrying meaning.
b) Syntactic formulation: proper placement of words to ascertain the required message.

Ultimate phase is Articulation of the voice wave applying appropriate body parts.

3.2.2 Biology (Physiology and Anatomy) of Speech Production:

The speech related neural capacity and mechanisms are required in order for a human to speak (Lieberman 2012). Initially, ability to conceptual formulation of the utterance to be spoken is essential, and then, various vocal apparatuses are to be stimulated in co-ordination properly to produce audible sound. Following discussion will give an insight to these aspects.

In our breathing process as soon as the airflow gets out of the lungs moves upwards towards mouth and nose, a pressure is created on air then speech is produced. While moving from lungs all along the vocal tract, air passes via the vocal tracks (chords) and generates vibration in that happens in a range of frequencies. Human vocal mechanism is depicted below through Figure 3.2. Subsequently we present sophisticated articulatory organs of our body along with their functions. Breathing is a function that is driven by the expansion and contraction of the diaphragm, and it

also has an important role in speech production (Heywood, Murphy et al. 1996). In order to speak, a person must gradually exhale, creating a steady flow of air from the lungs, through the trachea toward the next destination, the larynx.

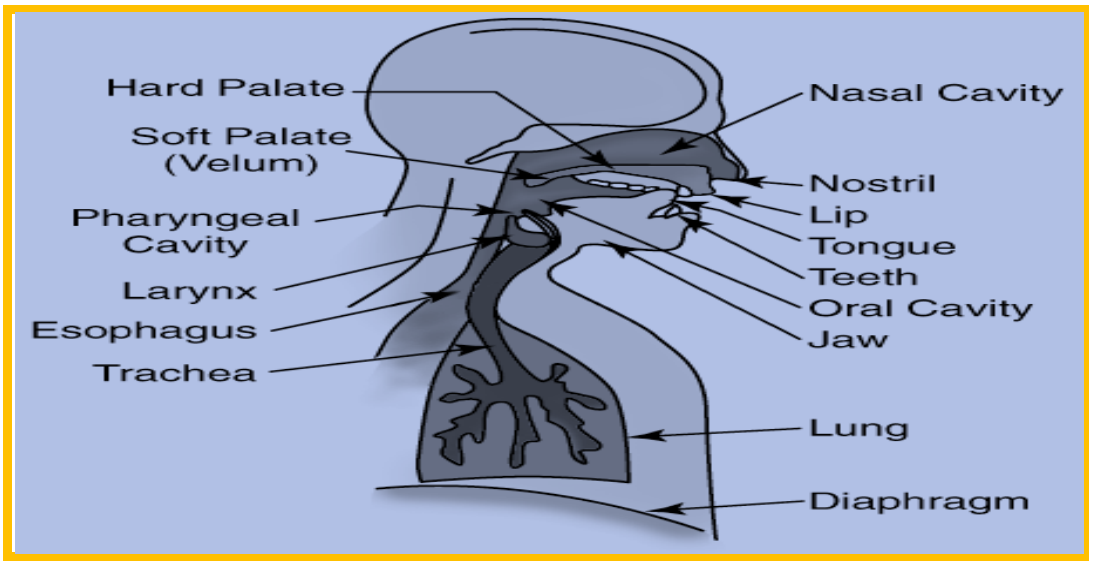

Figure 3.2: Human vocal system

3.2.2.1 The Vocal Tract:

As the acoustic signal (as generated in the larynx) passes through to the vocal tract, some frequencies are filtered by the surface area en route to the vocal tract (Ghazanfar, Rendall 2008). The process of filtering and exciting the acoustic signal, resulting from the vocal folds, is what ultimately converts it into speech that is released from the mouth, in conjunction with the signal that travels through the nasal cavity and is released from the nostrils (Hennebert 2009). After crossing the larynx lung sourced airflow transforms further while coming up and reaches the vocal tract. Combining together the first part the oral tract and the second part the nasal tract becomes 17 cm (on average). The resonating cavities consist of Upper pharynxial cavity, mouth and nostrils. Together with the vocal cord and other parts used to form sounds are named as articulators. They are further divided into active (movables like lips/tongue) and passive (fixed like teeth, soft/hard palate). In general it is the

combination of at least one active and passive articulator to generate most of the sounds. The interaction of articulator with one another is called articulation (upper lip meets the lower lip to say / ब /).

3.2.2.2 Pharynx Structure:

This muscular part of our body looks like a funnel (average length 8-12cm). The divider passage between air flow and food flow the pharynx and the stomach is normally kept closed by a muscle and open only at the time of swallowing food.

3.2.2.3 The Glottis:

The epiglottal fold normally called as the Glottis. This cartilage having leaf resembling shape is adjacent to the anterior part of the thyroid cartilage and to the root of the tongue. Glottis envelops the entrance to the larynx in the food swallowing process, and do not allow solids to go towards the trachea. Glottis helps in the production of some special sounds for instance, pharyngeal sounds in Arabic.

3.2.2.4. Velum Function:

Velum helps in separating the nasal cavity from the oral cavity. It acts to prevent air from coming through the nose (called velic closure) seen while producing oral sounds. When velum is lowered, air moves through both the mouth and nose helping in nasal sound production. Those sounds articulated with simultaneous oral and nasal articulations are known as nasal. A point to be marked here is that if we do not fully raise the velum while producing oral sounds, some amount of air will escape through the nasal cavity and those speakers doing it habitually produces nasalized speech and called as nasal twang. When someone has cold, nasalized speech is heard from him/her. Again when an effective closure is not possible for someone (for

defect in the soft palate) there will be an overall nasal effect in vowels and the failure to utter some selective sounds (like /b/, /g/, /d/). When the tongue is in contact with the lower side of the velum consonants are produced and called velar consonants (like /g/ or /k/).

3.2.2.5. The Tongue:

Primary function of the tongue is to taste and additionally performs the function of the bussiest articulator in speech production (movements of up to 9/sec). It may position almost in unlimited number of positions, both vertically and laterally and this versatility of the tongue helps in its primary function-eating and secondary function- talking. Primary physical function of the tongue is to push around the solid food in mouth and pharynx during chewing, swallowing and drinking. In formation of vowel sounds it is the principal agent. The tongue tip comes in contact behind the lower teeth for uttering vowels.

3.2.2.6. The Role of the Teeth and Lips:

Teeths (mainly upper part) plays vital role in utterance of many consonant sounds (e.g. that-/dat/ and think-/TINK/). Lips help in forming vowel and consonant sounds. For example, we need to know whether lips are to be rounded (/o/ in orange) or spread (/i/ in heed). To produce bilabial sounds (/b/, /p/) we need to bring the lips in contact. Again for articulating labiodental sounds (e.g. /f/, /v/), the upper teeth are touched by the lower lip. The mandibular attachment provides mobility for lower teeth and lip.

3.2.2.7. The Nose:

Our nose consists of two cavities (nostrils) joined by the central bone- septum. The floor of the nose is smooth and relatively wide compared to the roof of the nasal

cavity (too narrow). The surfaces are quit uneven. Back part goes to the nasopharynx. The nostrils provides the humidification and heating of the air in the respiration process and it acts as an air filter.

3.2.2.8 The Larynx:

Voice is generated when air travelling up through the trachea reaches and excites the vocal folds. Two types of excitation of the vocal folds are whispering and phonation (Campbell 1997, Hennebert 2009). The larynx is the source of sounds also called power house. The pitch and volume of a sound produced gets alternated in the larynx. Larynx accumulates power in expiration to enhance the voice loudness. The source sound is attached with a particular fundamental frequency (pitch). The flow moves towards vocal tract and voice is fine tuned. Now the voice gets configured with speaker specific attributes with individual difference according to on the articulatory adjustments. This vocal track filter mechanism results changes in a source sound. This will results in creation of different vowel and consonant sounds and different tones, and is a universal mechanism.

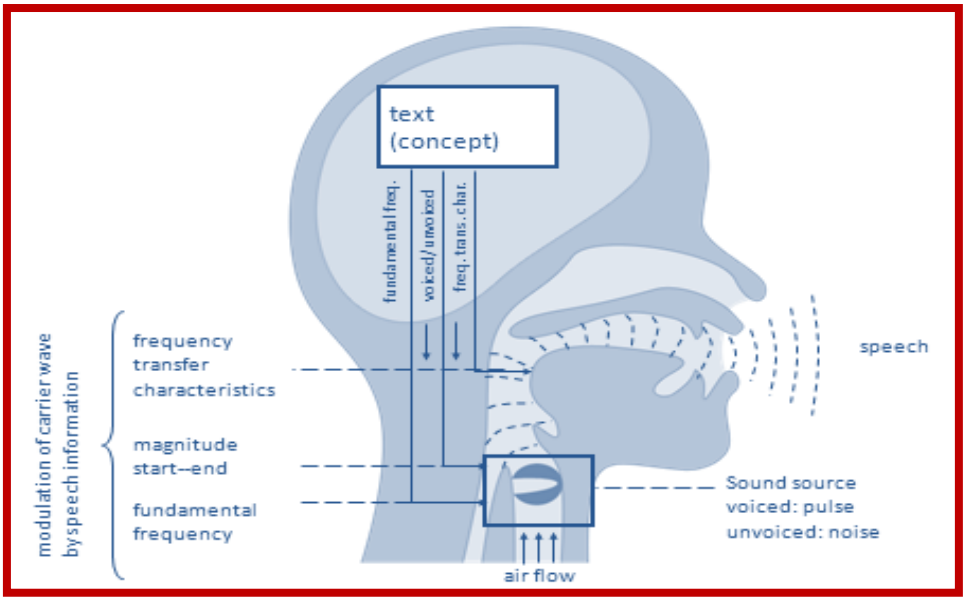

Figure 3.3: Human speech production mechanism and characteristics identification

3.3: Speech Sound:

The differentiation of vowels and consonants can be done depending on three primary aspects:-

1. Physiological: involved with airflow / constriction
2. Acoustic: Determined by prominence
3. Phonological: involves syllabic context

Normally consonants possess a higher degree of constriction than vowels. The oral and nasal stops, fricatives and affricates demonstrate this nicely. Again, consonants are less prominent than vowels. Occasionally, certain consonants shows greater total intensity than adjacent vowels still vowels are almost in every occasion more intense at low frequencies than adjacent consonants. It is an established fact that of the respiratory muscles movement at least creates a syllable. Syllables formed combining vowels usually surrounded by a number of consonants. Vowels are the **nucleus** of each syllable. Only a single peak of prominence per syllable is notable and it is predominantly a vowel. Opposite to this, consonants form the less prominent valleys between the vowel peaks. Speech is uttered in the form of sound sequence. A classification of events in speech is customized into three state oriented categories (shown below).

Voiced (V)	Periodical vibration (approximation and abduction) of vocal cords out of tension
Unvoiced (U)	No vibration in vocal cords
Silence (S)	No speech is produced in audible range

Speech organs movements forms sound, which is produced in air and the features perceived by listeners. Now a major activity is to relate a group of symbols for textual representation of the spoken utterances effectively.

3.3.1 Vowel Sounds:

Vowels are voiced phonetics defined with an open approximation of vocal cords without any obstruction, partial or complete, in the air passage.

Vowel sounds plays most crucial role in speaker recognition and authentication system development. They are produced by the vibration of the vocal cords, by exciting an essentially fixed vocal tract shape with quasi periodic pulses of air. The measurement of the vocal tract varies with uniqueness for each and every speaker, and defines the resonance frequencies (formants) of the tract. This rather defines sounds produced by individuals. Major factors which determine a vowel sound are: lips, jaw and the velum positions. The prolong duration of vowel utterances differentiates them from consonant sounds and posses stronger spectra. Thus vowels specification is easy, reliably recognized and immensely affects the recognition and authentication of speech and speaker. It holds well for both humans and for machines.

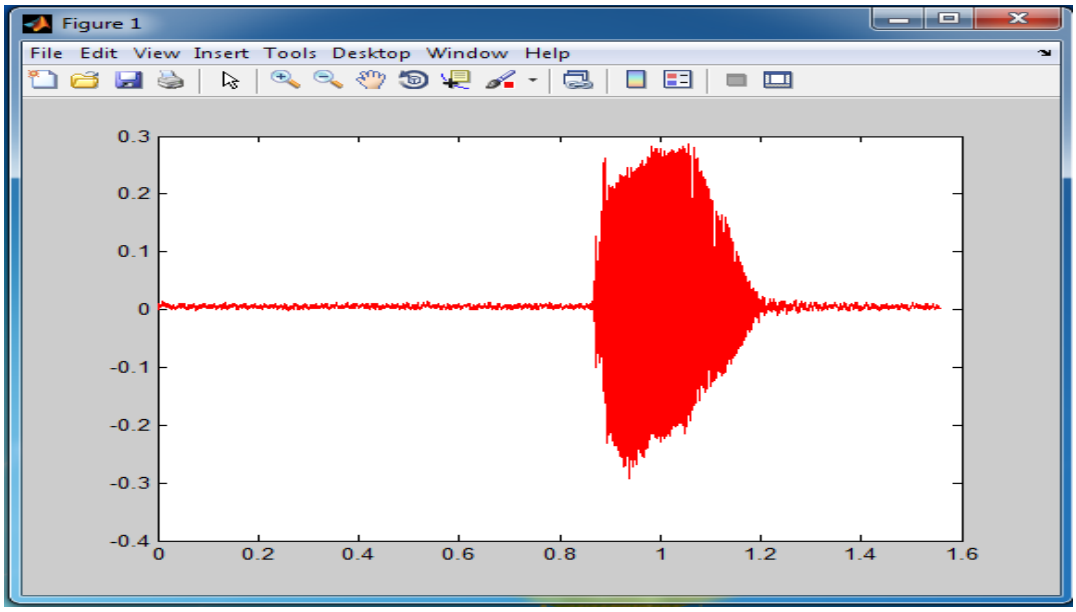

Figure 3.4: Snap shot of BODO vowel /अ/ /o/ spectrogram

3.3.2 Consonant Sounds:

The sounds that can't be uttered without the help of a vowel are consonants. In other words consonants are sound generated through creating a complete or partial stoppage of the air flow. It conveys that while uttering a consonant the lung generate air is fractionally or completely obstructed by narrowing down the air path or causing full closure of the same.

3.3.3 Syllables:

The elementary lexical units or fundamental parts without which words can't be formulated, while throwing an utterance are syllables. It is generally

i. A lone vowel.
ii. A vowel accompanied by one or more consonants and /or vowel.

Vowels getting the attention of a nucleus syllable, keeps conjugating with others to form themselves by positioning consonants either at the beginning or at the end in a word.

a. Arresting consonant positions itself at end of a syllable.
b. Releasing consonant positions itself at start of a syllable.

Again marginal elements occur either before the nucleus (vowel phoneme) or after it or some before and after it. In certain instances a group of two or three consonants before and/or after the vowel also exists. Continuous stream or at constant pressure is not seen with the air generated by power house lung in our body that initiates voice creation. Small puffs of air are the form in which we release it (average 5 times/second) that results in a syllabic utterance. Taking count of the syllable seen in the word it is assigned the tag of mono/di/……/penta/hexa/……/poly- syllabic word.

3.3.4 Phonemes:

One of the basic units of language is phoneme. Linguistically distinct speech sounds (Phonemes) of a particular language forms its characteristics and behavior of that particular language. It is linguistic lexical units in which if one phoneme is altered with another, the meaning of that word no more remain the same.

In 1870's the term phoneme was first introduced prominantly by Kruszewski. The other prominent phoneticians, whose contribution are noteworthy are, Sapir, Roman Jakobson, Louis Jhelmslev, Trubetzkoy, Bloomfield and Pike etc. The fundamental and basic unit of speech utterance that posses' inherent specificity among units can be termed as a phoneme.

3.4 Speech Research Hurdles:

Speech research fraternity often meets with the following inherent challenges while aiming speaker authentication

- Number of Speakers: Speech variability from one speaker to another becomes a major issue with increasing number of speakers, coping up with this is the prime concern for a Speaker Authentication system. To overcome this issue usually requirement is of using larger database in training phase.
- Speaking style: In case of isolated word level authentication is applied, it restricts the speakers to insert artificial pause between the successive utterances. To cope up with natural speech utterances deal with the words tied together or affects of co-articulation, robust spontaneous speaker authentication system is deployed.
- Volume of vocabulary: The authentication rate is inversely proportional to the vocabulary size increment in general.

- Language associated abstractness: Imposition of syntactic, semantic rules improvises authenticator system. It helps in achieving higher rate of authentication for continuous speech.
- Environmental Condition: There is a crucial negative effect of the following factor in the overall performance of the SAS- Background noise, Signal distortion and Transmission media difference.

Another critical hurdle to be dealt with cautiously in speech research projects is its interdisciplinary connection. Some of such associated disciplines that can create bottleneck are with one or more speech research problems are: *Forensic, Signal processing, Acoustics, Information and Communication Technology, Linguistics, Computer Science, Electronic and Instrumentation Science, Psychology, etc.*

Chapter summary: *In this chapter different aspect of human speech production including Anatomy and Physiology of human speech is elaborated for clear understanding.*