International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 9, Issue 10 (January 2014), PP. 48-54

Quick Review of Human Speech Production Mechanism

Harish Chander Mahendru
Department of Electronics & Communication Engg., Haryana Engineering College, Jagadhri, Haryana – 135003 (India)

Abstract:- This paper is presented to review the human speech production mechanism with an intention to portray different aspects involved in it, right from generation of thought in the mind to the production of sound wave from mouth. Speech production in human beings is a very common phenomenon which is experienced in their day to day life and is also part of speech communication between them. Although, speech production looks very simple from outside but its inside mechanism is very complex. Human being can generate many varieties of sounds whose frequency spectrum as well as the loudness changes very rapidly. This is possible due to the very sharp and precise articulatory movement control of the organs of speech production mechanism. The articulatory movement control is termed as Motor Control and is done by human brain through sensory nerve system connecting brain to the speech production organs such as lungs, vocal chords, tongue, jaw, lips, teeth, larynx etc. It is interesting to learn qualitative and quantitative aspects of the speech production organs for accurate speech analysis and synthesis through modeling. Through this review paper we will be describing in brief the complete speech production mechanism, the details of which are available in the referenced articles.

Keywords:- Articulatory Movement, Motor Control, Speech Generation, Speech Production Mechanism,Vocal Tract.

I. INTRODUCTION

Speech, being the natural form of communication, is the most basic and commonly used communication by all the human beings. For common people speech is just the sound waves coming out of the human mouth and perceived/listened through ears. But there is complex mechanism behind its production. The study of human speech production and perception mechanism is important and necessary for the development of devices for hearing aids, cochlear implant, speech recognition, speech enhancement, speech simulation, speech modeling etc. The complete speech production mechanism consists of mainly three functions which are represented through block diagram in Fig. 1 [1]. Motor control is the function driven by human brain which generates a thought of what to speak and accordingly it provides control signals through sensory nerves to the speech production organs. On receiving the control signals from the motor control unit, speech production organs move and take appropriate shape according to the words to speak or sound to be produced. The whole mechanism, termed as 'Ariculatory Motion' will be explained in the following paragraphs. Third function in the human speech production mechanism is the speech generation which consists of the air that comes out of the mouth and nasal cavity and is thrown in the open space in the form of acoustic wave. As far as speech perception is concerned, the acoustic wave generated through mouth and nasal cavity reaches the human ear and is perceived through sensory nerves connecting ear and the human brain. Reviewing speech perception is not the topic of this paper so we will concentrate on the speech production mechanism only.

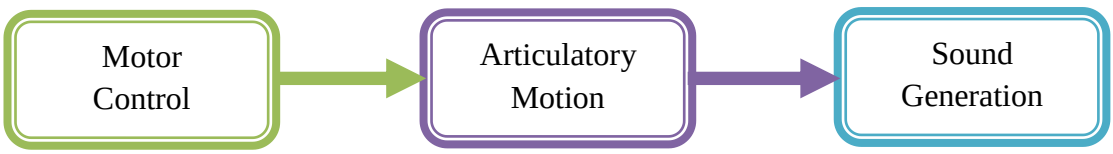

Fig. 1 Block diagram of human speech production mechanism.

Speech synthesis is required for many modern days' devices such as speech vocoders, speech recognition devices, hearing aids etc. Speech modeling, which is the first step of speech synthesis, is as complex as human speech production mechanism. Speech produced by different people differs in their basic parameters such as fundamental frequency, bandwidth and pitch. Producing exact replica of one's speech through machine simulation is quite difficult. However latest research and developments have made it possible to simulate human speech to large extant.

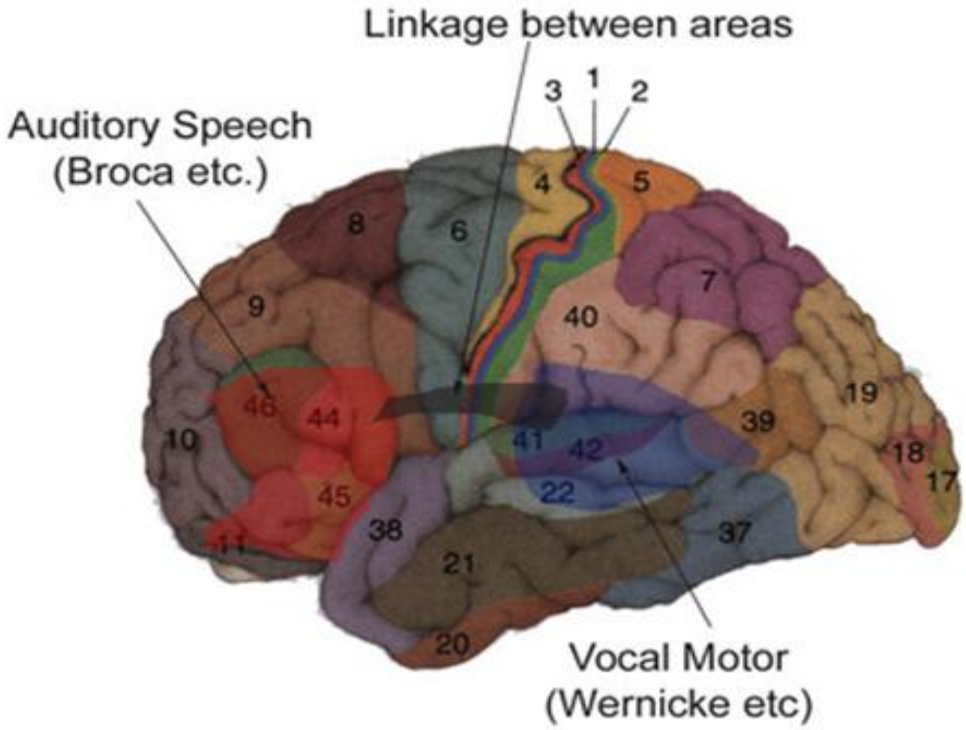

Fig. 2 Segmentation of the human brain showing Brodmann areas.

II. MOTOR CONTROL FUNCTION

We see people speaking to each other almost unconsciously without even knowing what is happening behind speech production mechanism. The production of speech starts with generation of idea in the mind about what to speak. This idea generation is passed onto the human vocal apparatus through sensory nerves. The whole process is termed as motor control function which is further divided into two parts, the language processing and the motor commands generation [2]. Researchers have evolved that our brain is divided into different segments. These segments are responsible to perform various control, think and memory functions [2]. Fig. 2 shows the segmentation of the human brain as suggested by Korbinian Brodmann [3]. The researcher in the field will be curious to know the answers to questions such as: How the brain extracts speech sounds from an acoustic? How the brain separates speech sounds from background noise? How the phonological system of a particular language is represented in the brain? How the brain stores and accesses words that a person knows? How the brain combines words into constituents and sentences? How structural and semantic information is used in understanding sentences? Phonetics, Phonology, Morphology, Lexicology, Syntax and Semantics are various subfields of Neurolinguistics through which researchers have discovered the answers to the above questions [6], [7]. In figure 2, the area, shaded red, numbered 44 is the auditory region and the area shaded blue, numbered 41 & 42 is the motor control region. The two areas are shown interconnected through dark shading indicating that these two regions work in tandem for accurate generation of sound signals. In fact the auditory region (Broca's region) gets input in the form of listening and visual gestures through other sensory organs for language processing, helping it to decide what to speak or what sound to produce [8]. Accordingly the motor control region (Wernicke's region) generates control signals to move the vocal tract apparatus and other speech production organs such as lungs, vocal cords, glottis, jaw, tongue, teeth, lips etc. The whole process is shown with the help of Fig. 3, below.

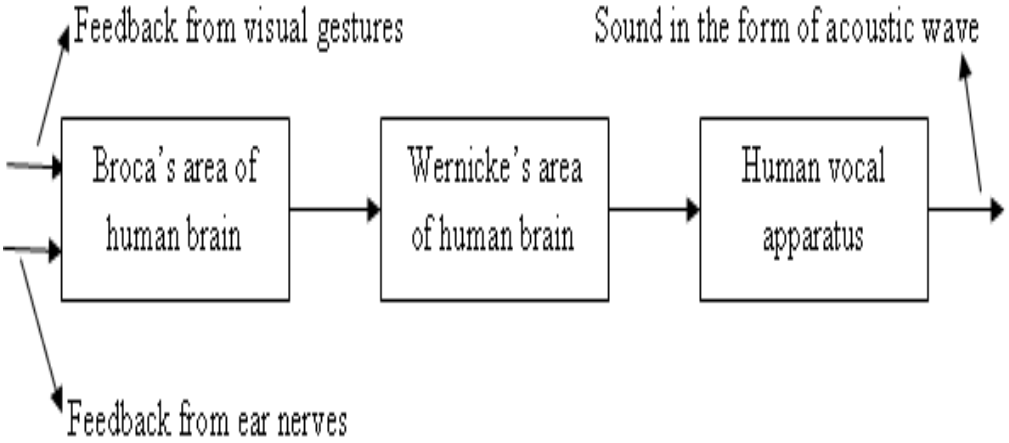

Fig. 3 Sound generation and production process.

III. ARTICULATORY MOTION

Various organs are involved in the production of speech & sound by the human beings. These organs are flexible in nature and their shape and size alters on the command of motor control signals received from the brain, as per the type of speech and sound to be produced. Lungs provide the necessary air force for the generation of sound in the form of acoustic wave. The air passes through the vocal tract (the pipe linking lungs and the throat), vocal cords, glottis, epiglottis, and other organs in the mouth and finally comes out through mouth and nasal cavities in the form of acoustic wave. Various organs through which the air passes during the process of generation of speech and sound are shown in the **Fig. 4 [9], [10].**

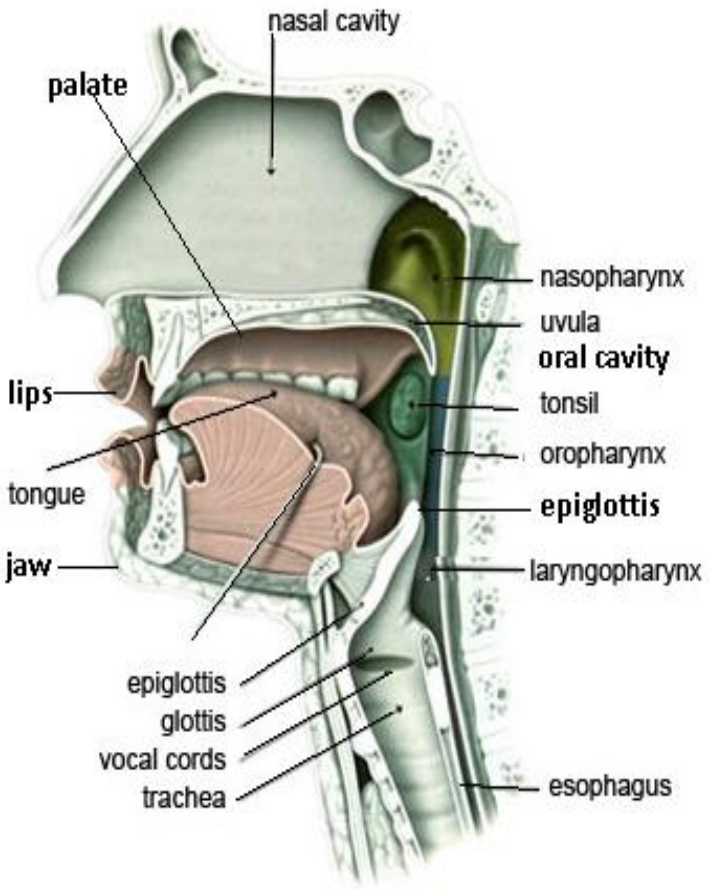

Fig. 4 Human vocal apparatus.

When we speak the air expelled from the lungs moves up through trachea and enters larynx. In the larynx the air is restricted by a pair of lip like tissues called vocal cords. These are very important membranes of vocal apparatus which decide the pitch of the speech produced. Vocal cords are pearly white in colour fixed at one end attached to the arytenoids cartilages at the back and to the thyroid cartilage at the front. Generally the males have low pitch voice with large larynx as compared to the high pitch voice for females with small larynx. The length of vocal folds for males varies from 17 to 25 mm and for females it varies from 12.5 to 17.5 mm. The vocal folds vibrate to produce voiced speech and provide momentarily restriction to produce unvoiced speech. In fact there are various other types of speeches called phonemes for which vocal cords open or close in different fashion to let the air pass through it and send it to the upper part of the vocal tract. Vocal tract is the tube like passage which runs from glottis at one end and with two openings, oral and nasal cavity, at the other end. It is of non uniform cross section whose approximate length for males is about 17 cm. It branches out at soft palate (velum), just at half way of the tract, and opens up at nostril as second branch. This part of the vocal tract is approximately 13 cm long. Air, after leaving the vocal cords enters into the pharyngeal, mouth and nasal cavities which provide necessary resonation to the sound as per word to speak by amplifying some of the frequencies and attenuating others. Other organs in the mouth such as soft palate, teeth, tongue, lips, jaw change their shape and move accordingly providing blockage or allowance to the air for the mouth and nasal exit and thus modulate the sound to give necessary shape and amplitude. Because of the difference in size and shape of different speech production organs, speech of an individual is unique. The beauty of the whole articulatory movement system is that even after having so many complexities it is able to react very fast catering to the fast changing speech parameters. Epiglottis and false vocal cords below the pharynx has an important role of preventing food to enter into the larynx and isolate the esophagus acoustically from the vocal tract.

IV. PHONEMES

Generally people speak in language according to the region they are brought up in. One does not need special training or knowledge to speak in their mother tongue. Children learn to speak at an early age of one year by understanding the audio and visual gestures. The language signs of any language can be pronounced with the help of symbols called phonemes. All the words with different tones of any language can be spoken using minimum set of phonemes. All the languages spoken in the world have 20 to 60 phonemes [11], [12]. Phonemes of any language include contextual effects, emotions and characteristics of the speaker to be pronounced which of course is not required for written text of the language. These phonemes are basically designed based on the articulatory movement of the vocal tract. In this paper we will be discussing about the phonemes of the most spoken language in the world i.e. English. All the possible words of English language can be described with 40 phonemes [13].

The phonetics of any language contains broadly two types of phonemes, the Vowels and the Consonants. Vowels are always voiced sounds whereas consonants can be voiced and non-voiced both. Voiced sounds are produced when the vocal cords vibrate frequently almost periodically when air passes through it having fundamental frequency about 110 Hz for men, 200 Hz for women and 300 Hz for children. Apart from the fundamental frequency the articulatory movement of the speech production organs generates resonance frequencies according to the phoneme. These 'N' number, F1, F2, ..…, Fn, of resonance frequencies are called Formant Frequencies. The normal range of the formant frequencies for adult males is F1 = 180-800 Hz, F2 = 600-2500 Hz, F3 = 1200-3500 Hz, and F4 = 2300-4000 Hz. The unvoiced sound on the other hand is totally random in nature. During the production of unvoiced sound the vocal cords are completely open, completely close or partially open. The most popular and widely used format of phonemes for the American English language is ASCII symbols known as ARPAbet [14], [15]. These sounds are represented by a set of 39 phonemes as given in Table 1. Vowel phonemes are produced due to frequent vibration of the vocal cords. As per the position of tongue in the mouth cavity, vowel phonemes are further divided into three types, namely 'Front' such as /IY/, /IH/, /EY/, & /EH/; 'Mid' such as /AA/ & /ER/ and 'Back' such as /AE/, /AO/, /UH/, /OW/,

Table I. North American English ARPAbet Phonetic

Class	Subclass	ARPAbet symbol	Example	Transcription
Vowels	Front	IY	beet	[B IY T]
		IH	bit	[B IH T]
		EY	bait	[B EY T]
		EH	bet	[B EH T]
	Mid	AA	bob	[B AA B]
		ER	bird	[B ER D]
	Back	AE	bat	[B AE T]
		AO	born	[B AO R N]
		UH	book	[B UH K]
		AH	but	[B AH T]
		OW	boat	[B OW T]
Diphthongs		UW	boot	[B UW T]
		AY	buy	[B AY]
		AW	down	[D AW N]
		OY	boy	[B OY]
Semivowels	Glides	Y	you	[Y UH]
		R	rent	[R EH N T]
	Liquids	W	wit	[W IH T]
		L	let	[L EH T]
Consonants	Nasals	M	met	[M EH T]
		N	net	[N EH T]
		NG	sing	[S IH NG]
	Plosives	P	pat	[P AE T]
		B	bet	[B EH T]
		T	ten	[T EH N]
		D	debt	[D EH T]
		K	kit	[K IH T]
		G	get	[G EH T]
	Fricatives	HH	hat	[HH AE T]
		F	fat	[F AE T]
		V	vat	[V AE T]
		TH	thing	[TH IH NG]
		DH	that	[DH AE T]
		S	sat	[S AE T]
		Z	zoo	[Z UW]
		SH	shut	[SH AH T]
		ZH	azure	[AE ZH ER]
	Affricates	CH	chase	[CH EY S]
		JH	judge	[JH AH JH]

& /AH/. Diphthong sounds are produced by transition between two vowels within a single syllable such as /AY/, /AW/, /OY/ & /UW/. Semivowels are produced when tongue or lips closes the vocal tract completely. Semivowels are of two types, 'Glides' such as /Y/ & /R/ and 'Liquids' such as /W/ & /L/. Consonants can either be voiced or unvoiced and are further classified into Nasal, Stop or Plosives, Fricative and Affricate. Nasal

sounds such as /M/, /N/ & /NG/ are produced when mouth cavity is closed and the air passes through nasal cavity via opened velum. Plosive sounds such as /P/, /B/, /T/, /K/, /D/ & /G/ are produced when pressure built up behind the vocal cords due to its momentarily closure is released suddenly. Here /B/, /D/& /G/ are voiced whereas /B/, /T/ & /K/ are unvoiced. When mouth cavity is not fully blocked and there is quasi-periodic flow of air due to vocal cord vibrations then the sound produced is named as fricative sounds such as /HH/, /F/, /V/, /TH/, /DH/, /S/, /Z/, /SH/ & /ZH/. Affricate sounds such as /CH/ & /JH/ are produced by dual action of plosives followed by fricative.

V. REPRESENTATION OF SPEECH SIGNALS

As mentioned earlier broadly the speech signal is divided into two categories, voiced and unvoiced. Voiced speech signals are periodic in nature whereas unvoiced look like random signals. These two types of speech signals are shown in Fig. 5. For the analysis, speech signals are represented in various forms such as time domain representation, frequency domain representation and spectrogram. Examples of these three types of representations are given in Fig. 6 (a, b, c). In the time domain representation signal is represented time versus amplitude.

In frequency domain representation signal is represented frequency versus amplitude. And in spectrogram representation the signal is represented time versus frequency. All of these representations are important from the point of view of speech synthesis and analysis. The analysis of speech signal is again very cumbersome due to its very fast changing (every 50 – 100 ms) parameters such as pitch and loudness. Due to this the speech analysis/processing is performed in blocks or frames. These blocks are formed by windowing the speech signal for short duration (of tens of millisecond) for which the speech parameters are expected to be constant. The blocks are formed on overlapping time frames which, after processing, are mixed again to retrieve the original speech signal. Various methods are used for short time analysis of speech signal. Some of these methods include, short time zero crossing rate; short time energy; short time auto correlation function, short time Fourier transform; short time z-transform, short time cepstrum, short time homographic filtering and short time spectrograph.

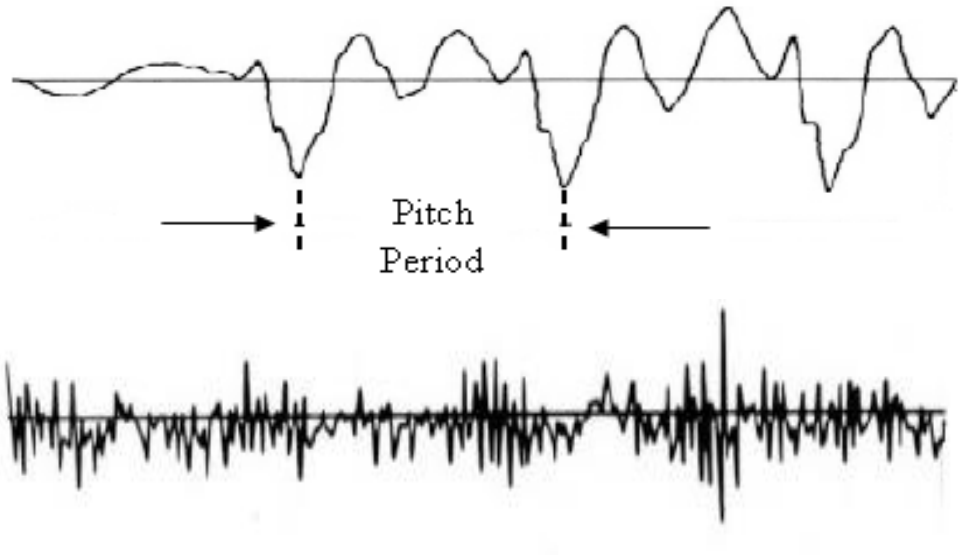

Fig. 5 Examples of Voiced (Top) and Unvoiced (Bottom) Speech Signal

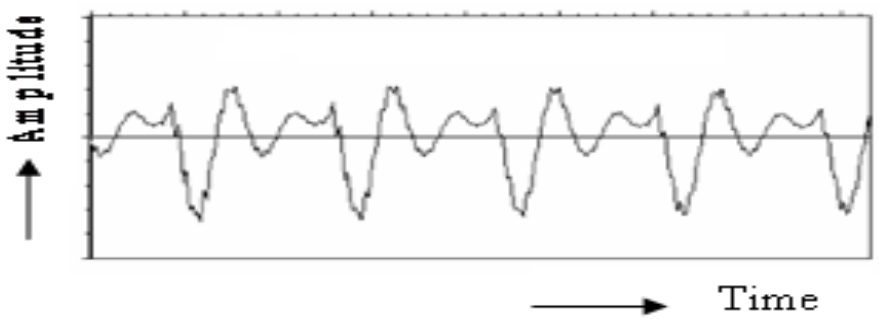

Fig. 6 (a) Amplitude versus Time representation of speech signal

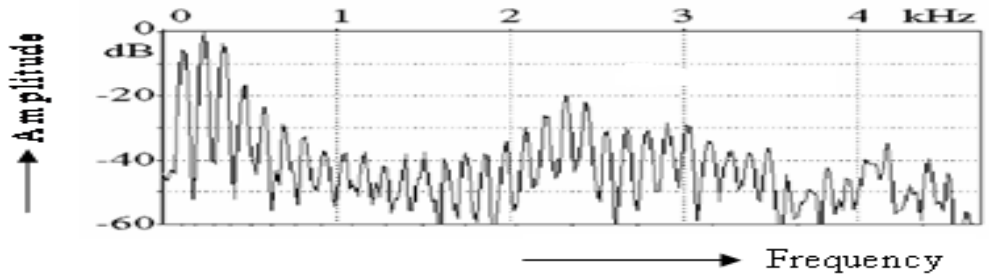

Fig. 6 (b) Frequency versus Amplitude representation of speech signal

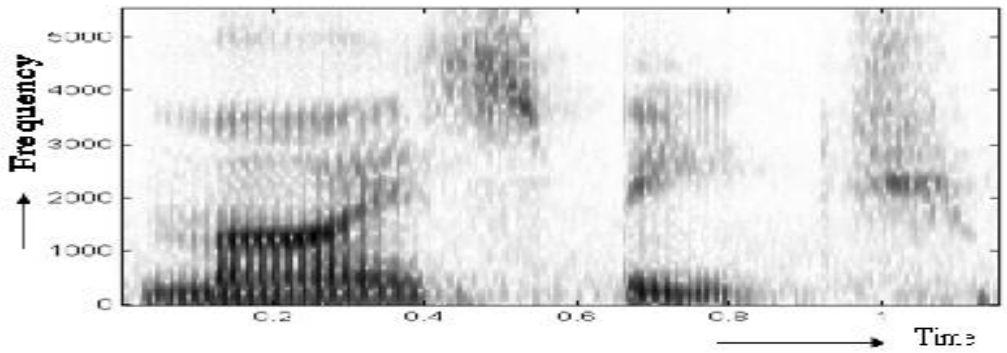

Fig. 6 (c) Time versus Frequency representation of speech signal

VI. CONCLUSION

The purpose of this paper is to present quick overview of the human speech production mechanism. The summarized research work of this paper will be use full for the researchers who are working on the synthesis and analysis of speech, such as, speech compression, speech enhancement, text to speech / speech to text conversion, cochlear implants and speech recognition. In brief it is concluded that it is very easy for a normal person to speak but there is big complexity involved behind its production.

REFERENCES

[1]. Masaki Honda, "Human Speech Production Mechanism", selected papers, NTT technical review, vol.1, No. 2, May 2003.
[2]. Edmund Blair Bolles, "Speech Circuitry", a blog on origins of speech posted at http://www.babelsdawn.com, Jul 2009.
[3]. Hockett, C. F., "The origin of speech" Scientific American, vol. 203, pp. 88-96, 1960.
[4]. Michael, C. Corballis, *From Hand to Mouth: The Origins of Language* (Princeton and Oxford: Princeton University Press, 2002).
[5]. Lieberman, P, *The Biology and Evolution of Language* (Cambridge, MA: Harvard University Press, 1984).
[6]. Goodall, J., *The Chimpanzees of Gombe. Patterns of behavior* (Cambridge, MA and London: Belknap Press of Harvard University Press, 1986).
[7]. Darwin, C., *The Expression of the Emotions in Man and Animals* (London: Murray, 1872).
[8]. Mc Caffrey, Patrick, *CMSD 620 Neuroanatomy of speech, Swallowing and language* (Neuroscience on the web, California State university, Chico, Feb 2009).
[9]. Tool module: The Human Vocal Apparatus, available at http://thebrain.mcgill.ca/flash/capsules/outil_bleu21.html
[10]. Fitch, W. T., "The evolution of speech: a comparative review", Trends in Cognitive Science 4, pp. 258-267, 2000.
[11]. O'Saughnessy D., *Speech Communication - Human and Machine* (Addison-Wesley, 1987).
[12]. Breen A., Bowers E., Welsh W., "An Investigation into the Generation of Mouth Shapes for a Talking Head", Proceedings of ICSLP 96 (4), 1996.
[13]. Rossing T., *The Science of Sound* (Addison-Wesley, 1990).
[14]. Rabiner, L. R. and Schafer, R. W., "Introduction to Digital Speech Processing", NOW, the essence of knowledge, Foundations and Trends in Signal Processing, vol. 1, No. 1-2, 2007.
[15]. [Online] available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.