ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

Review on Speech Production Model

Rupali V. Pawar, Dr. R.M. Jalnekar
Research Scholar, Vishwakarma Institute of Technology
Director, Vishwakarma Institute of Technology

Abstract*— Simplest and reliable way of communication amongst human beings is speech. Speech processing is one of the important applications of signal processing which involves techniques to process and analyse speech signals. It is important to understand the human speech production mechanism in order to model the same for various applications. The individual characteristics like pitch, fundamental frequency, formant frequency can be distinguishing components of human speech. This paper reviews the fundamentals of human speech production system, the characteristics of voiced and unvoiced sounds, and identification of voiced and unvoiced sounds from its spectrum.

***Index Terms*— Phonemes, voiced sound, unvoiced sound, formant frequency.**

I. INTRODUCTION

Speech conveys different forms of information to the listener. Along with the basic information about the language being spoken and the emotion, gender and the identity of the speaker also could be the part of information. This gives way to various applications in speech processing the major ones being Speech Recognition. Speech Recognition aims at recognizing the word spoken in speech, the goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity. [1] Quantitative models of human speech production and perception provide important insights into our speech production and perception mechanisms and lead to high-quality computer synthesis of speech, robust automatic speech recognition (ASR), and efficient speech and audio coders. These issues are of importance in the development of effective human-computer communications through the medium of human language. [2]

II. SPEECH PRODUCTION/PERCEPTION

A. Articulation

The movement of organs like tongue, lips, jaw to produce speech sounds is articulation. The accuracy of articulation can be affected by many factors physical contributed by factors such as health of muscles used to produce the sound or neurological damage.

B. Speech Generation

The process of speech production begins when the speaker formulates a message in his or her mind to communicate with the listener via speech, represented by first block (formulation of message). The next step would be Conversion of message into a language code, this involves converting the message into set of phonemes, comprising of correct sequence of words along with syntax, duration and loudness of sound. Figure 1 shows a different view of speech production and perception; the first four blocks comprising of message formulation, language code, neuro-muscular control and vocal tract system represent speech generation. The process is laid out along a line corresponding to the basic information rate of the signal at various stages of the process. The discrete symbol information rate is the raw message and is rather low of 50bps.After the language code conversion the information rate rises to about 200bps. The information becomes continuous with rate of 2kbps at neuromuscular control where the movement of articulators and its coordination with brain comes in picture and transfer of information takes place about 30-50kbps at the acoustic signal level. The transmission channel transmits acoustic waveform from the speaker to the listener [3].The speech Production process takes place inside the vocal tract, which begins at the glottis and ends at the lips.

C. Speech Recognition

The speech waveform generated from the vocal tract is analyzed by the basilar membrane (Ear), thus spectrum analysis is performed on the continuous signal. The features such as duration of vocal cord vibration, intensity of the sound, resonant frequency of the vocal cord are extracted at the neural transduction stage .Finally the language used for communication and the interpretation of message is done. The steps in speech perception mechanism can also be interpreted in terms of information rate in the signal and this follows a inverse pattern of production process. The higher level processing in brain converts the neural signals to a discrete representation which is ultimately decoded into a low bit rate [3].Figure 1 is shown in appendix.

III. GENERATION OF VOICED AND UNVOICED SOUNDS

Speech sounds are composed of sequence of sounds called phonemes produced as a result of acoustical excitation of vocal tract.

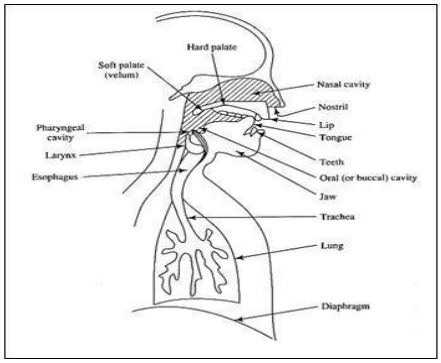

Fig 2: "Human Speech Production System"[5].

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

The vocal tract shape is determined from the position of the vocal organs, and speech is produced by controlling the speech production model using the vocal tract area [4]. Figure 2 show that the vocal tract is an acoustical tube which begins at the opening between the vocal cords and ends at the lips.The cross section area of the vocal tract varies from 0 to 20 square cm and is dependent on the position of the articulators. Nasal tract begins at the soft palate called velum and ends at the nostrils. Depending on the whether the vocal cord vibrates sound produced can be broadly classified as voiced and unvoiced sound. The vocal cords are tensed for sounds like a/e/i and vibrate to produce voiced sound. The vocal cords vibrate periodically and when air flows from the lungs resulting in a speech waveform which is quasi-periodic in nature. Air flows through vocal cords into vocal tract in discrete puffs. The vocal cords do not vibrate for sounds like s/f resulting in random speech waveform called unvoiced sound. The classification of the speech signal into voiced, unvoiced, and silence provides a preliminary acoustic segmentation of speech, which is important for speech analysis. [6] There are other sounds like nasal sound when vocal tract is coupled acoustically with nasal cavity through velar opening sound radiates from nostrils as well as lips. Plosive sounds are characterized by complete closure or constriction towards the front end of vocal tract, building up pressure behind the closure and sudden release. Utterance of p/t is examples of plosive sound while utterance of m/n is examples of nasal sound.

IV. RESSONANT FREQUENCY OF VOCAL TRACT

The spectrum of vocal tract has number of resonant frequencies called the formant frequencies. Speech signal has one formant frequency every 1 kHz hence there are 3 to 4 formant frequencies in 4 kHz of speech wave. For voiced sounds the amplitude of lower frequencies is larger than the amplitude of higher frequencies. While in case of unvoiced sounds the magnitude of lower frequencies is lower than that of higher frequencies [7]

V. SPEECH PRODUCTION MODEL

The basic assumption of speech processing system is that the source of excitation and the vocal tract system are independent [8] To develop a model for speech production it would be necessary to accurately represent Excitation mechanism for both voiced and unvoiced sounds, vocal tract model, lip or nasal radiation. The block schematic in figure 3 represents the model for the same. [7]

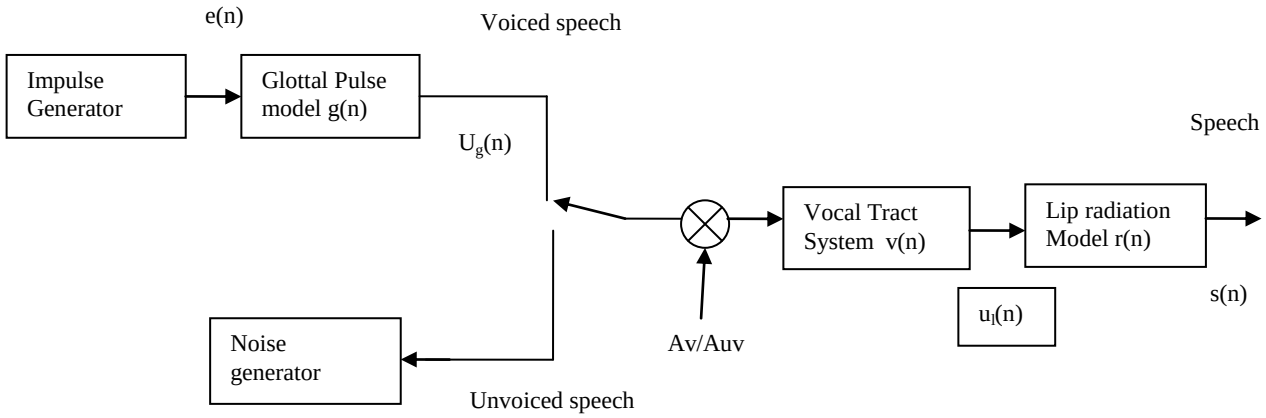

Fig 3: "Speech Production Model"

A. *Excitation Process*

The excitation process must take into account the voiced and unvoiced nature of speech, the operation of glottis and the energy of the speech signal in a frame of 10-30msec. For voiced speech excitation a train of Impulses spaced at intervals of pitch period are mathematically expressed as

$$e(n) = \partial(n - Pk) \quad \text{for k= 0,1,2….}$$

$$E(z) = \sum_{n=0}^{\infty} e(n)\, z^{-n}$$

$$= 1 + z^{-p} + z^{-2p} + \ldots$$

$$= \frac{1}{1 - z^{-p}}$$

For unvoiced speech

$$e(n) = random(n)$$

The glottal model for pulse shaping operation used only for voiced speech is expressed as

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2}$$

$CT << 1$, hence $e^{-cT} \approx 1$

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

$$G(z) \cong \frac{1}{(1-z^{-1})^2}$$ *for voiced speech*

G(z) = 1 for unvoiced speech

B. Vocal tract Model

The input for the vocal tract model coming as output of glottal pulse model or random generator is amplified to get A_v or A_{uv}. For voiced speech vocal tract can be represented by all pole filters. Typically two poles are required for each resonance or formant frequency.[7]Thus the transfer function will be

$$V(z) = \frac{U(z)}{Ug(z)} = \frac{1}{\prod_{k=1}^{N} 1 + b_k z^{-1} + c_k z^{-1}}$$

$$V(z) = \frac{1}{1+\sum_{k=1}^{P} a_k z^{-k}}$$

C. Lip radiation Model

It is a high pass filter

$$R(z) = 1 - 0.9z^{-1}$$

D. Overall Model for voiced and unvoiced speech

$$u_g(n) = Av.e(n) * g(n)$$

$$u_l(n) = u_g(n) * v(n)$$

$$s(n) = u_l(n) * r(n)$$

$$s(n) = A_v[e(n) * g(n) * v(n)] * r(n)$$

$$\frac{S(z)}{E(z)} = A_v G(z)V(z)R(z)$$

$$\frac{S(z)}{E(z)} = A_v \frac{1}{(1-z^{-1})^2} \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}} (1-z^{-1})$$

For unvoiced speech the overall transfer function can be represented as

$$\frac{S(z)}{E(z)} = \frac{A_{uv}(1-z^{-1})}{1+\sum_{k=1}^{P} a_k z^{-k}}$$

$$\frac{S(z)}{E(z)} = \frac{A_{uv}(1-z^{-1})}{1+\sum_{k=1}^{P+2L} a_k z^{-k}}$$

$$\frac{S(z)}{E(z)} = \frac{A_v}{1+\sum_{k=1}^{P+2L+2} a_k z^{-k}}$$

The zeros can be mapped as 2 poles Transfer function for voiced and unvoiced speech

$$\frac{S(z)}{E(z)} = \frac{G}{1+\sum_{k=1}^{q} a_k z^{-k}}$$

.

Choosing q equal to 12 sufficiently represents voiced and unvoiced speech.

VI. APPLICATION OF VOCAL TRACT MODEL

The vocal tract model can be used for applications like speech recognition, linear predictive coding. The overall all pole transfer function when represented in time domain gives LPC equation

$$\frac{S(z)}{E(z)} = \frac{G}{1+\sum_{k=1}^{q} a_k z^{-k}}$$

$$s(n) = \sum_{k=1}^{q} a_k s(n-k) + Gu(n)$$

The LPC equation represents speech signal which depends on previous samples and excitation. [7]

VII. CONCLUSION

The speech signal is generated by vocal cords and vocal tract. Understanding the human speech production, differentiating the voiced and unvoiced sounds helps generate model for speech production. This paper attempts to present the overall review of speech production model. The paper also gives the mathematical analysis at every stage of speech production model.

REFERENCES

[1] Douglas A. Reynolds Automatic Speaker Recognition: Current Approaches and Future Trends.

[2] A. Alwan, S. Narayanan, B. Strope, and A. Shen.Speech Production and perception models and their Applications to synthesis , Recognition and coding

[3] Lawrence Rabiner, Biing_Hwang Juang, B. Yegnanarayana. Fundamentals of Speech recognition, Pearson

[4] Honda M., "Human Speech Production Mechanisms", NTT Technical Review, Vol. 1 No. 2

[5] http://www.barcode.ro/tutorials/biometrics/voice.html

[6] Qi Y., Hunt R.B., "Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier", IEEE Transactions on Speech And Audio Processing, Vol. 1, No. 2, April 1993

[7] Prof Ambikairajah , speech and Audio Processing video lectures 1 to 6 UNSW learning

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

http://www.freevideolectures.com/Speech and Audio processing

[8] Dr. Shaila D. Apte. Speech and audio Processing, Wiley India Pvt.Ltd

APPENDIX

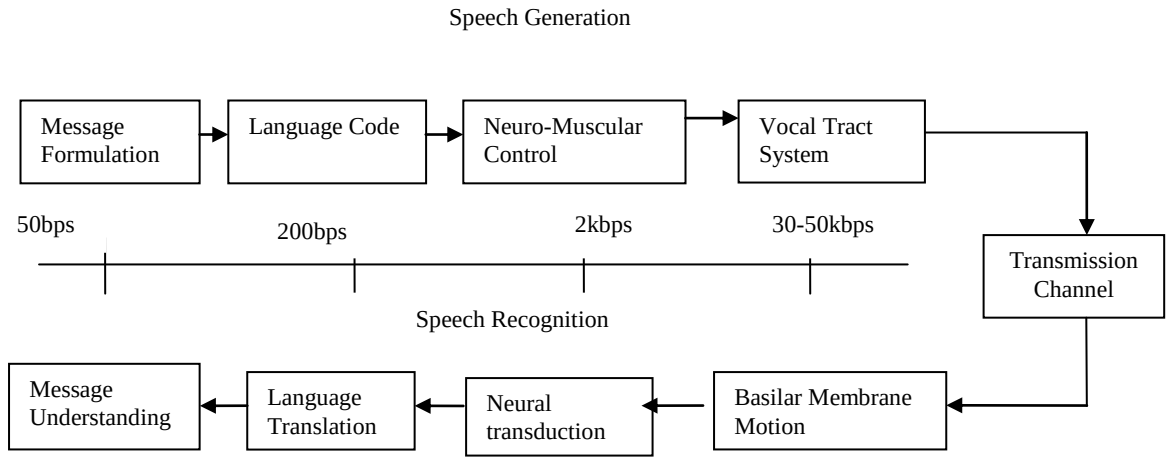

Fig 1: "Alternative view of speech production and perception" [3]