ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 9, November 2012

THE PROCESS OF SPEECH RECOGNITION, PERCEPTION, SPEECH SIGNALS AND SPEECH PRODUCTION IN HUMAN BEINGS

DR. Sunita Rana
Associate Prof. & H.O.D App.Sc.
Hindustan Inst. Of Tech. & Mgt.

ABSTRACT

Automatic recognition of speech by machine has been a goal of research for more than five decades. In spite of glamour of designing an intelligent machine that can recognize the spoken word and recognize its meaning, and in spite of the enormous research efforts spent in trying to create such a machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments. Thus, an important question in this paper is, what do we mean by "speech recognition by machine"? Another important question is what are the issue in developing conversational speech recognition and understanding systems by machine? Because we do not know how to solve the ultimate challenge of speech recognition. One of the most difficult aspects of performing research in speech recognition by machine is its interdisciplinary nature, and the tendency of most researchers to apply a monolithic approach to individual problems. My aim in this paper is to give a series of presentations on the fundamental principles of most modern, successful speech- recognition systems so as to provide a framework to understand the conception of speech recognition.[1]

Key Words- Speech mechanism, speech perception, speech production, speech recognition Speech signal, Speech Recognition, Speech Process ,

INTRODUCTION

In this paper we discuss the mechanics of producing and perceiving speech in human beings, and we will see how an understanding of these processes leads naturally to several different approaches to speech recognition by machine. We begin by showing how the different classes of speech sounds, or phonetics, can each be characterized in terms of broad acoustic features whose properties are relatively invariant across words and speakers. The idea of acoustic- phonetic characterization of sounds lead naturally to straightforward implementation of a speech recognition algorithm based on sequential detection of sounds and sound classes. The strengths and weaknesses of such an approach are discussed. An alternative approach to speech recognition is to use standard pattern-recognition techniques in a framework in which all speech knowledge is "learned" via a training phase. Such a "blind" approach has some natural advantages for a wide range of speech recognition systems. Finally we show how aspects of both the acoustic-phonetic approach and the pattern-recognition approach can be integrated into a hybrid method that includes techniques from artificial intelligence as well as neural network methods.[2] There are some contents can consider the disciplines that have been applied to one or more speech recognition problems:-

Signal Processing-the process of extracting relevant information from the speech signal in an efficient robust manner. Included in signal processing is the form of spectral analysis used to characterize the time-varing properties of the speech signal as well as various types of signal preprocessing to make the speech signal robust to the recording environment.

Physics(acoustics)-the science of understanding the relationship between the physical speech signal and the psychological mechanisms that produced the speech and with which the speech is perceived.[3]

Pattern recognition-the set of algorithms used to cluster data to create one or more prototypical patterns of a data ensemble, and to match a pare of patterns on the basis of future measurements patterns.

Communication and Information theory-the processors for estimating parameters of statistical models; the methods for detecting the presence of particular speech patterns, the set of modern coding and decoding algorithms used to search a large but finite grid for a best path corresponding to a "best" recognized sequence of words.[4]

Linguistic-the relationship between sounds, words in a language, meaning of spoken words and sense derived from meaning. Included within this discipline are the methodology of grammar and language parsing.

Physiology-understanding of the higher order mechanisms within the human central nervous system that account for speech production and perception in human beings. Many modern techniques try to embed this type of knowledge within the framework of artificial neural networks.

Computer science-the study of efficient algorithms for implementing, in software and hardware, the various methods used in a practical speech recognition system.

Psychology-the science of understanding the factors that enable a technology to be used by human beings in practical tasks. [5]

THE PROCESS OF SPEECH PRODUCTION AND PERCEPTION IN HUMAN BEINGS

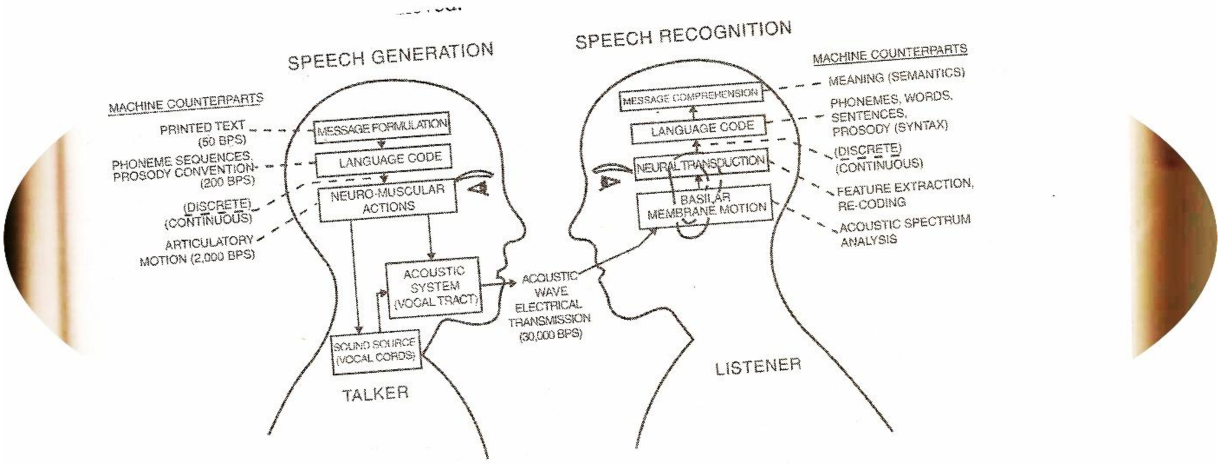


Figure 2.1 shows a schematic diagram of the speech-production/speech-perception process in human beings. The production (speech-generation) process begins when the talker formulates a message (in his mind0 that he wants to transmit to the listener via speech. The machine counterpart to the process of message formulation is the creation of printing text expressing the words of the message. The next step in the process is the conversion of the message into a language code. This roughly corresponds to converting the printing text of the message into a set of phoneme sequence corresponding to the sounds that make up the words, along with prosody markers denoting duration of sounds, loudness of sounds, and pitch accent associated with the sounds. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output. The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the lips, jaws, tongue, and velum (a "trapdoor" controlling the acoustic flow to the nasal mechanism).

Once the speech signal is generated and propagated to the listener, the speech-perception (or speech-recognition) process begins. First the listener processes the acoustic signal along the basilar membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process. In a manner that is not well understood, the neural activity along the auditory nerve is converted into a language code at the higher centres of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

A slightly different view of the speech-production/ speech-perception process is shown in **Figure 2.2.** Here we see the step in the process laid out along a line corresponding to the basic information rate of the signal 9or control) at various stages of the process. The discrete symbol information rate in the raw message text is rather low (about 50 bps [bits per second] corresponding to about 8 sounds per second, where each sound is one of about 50 distinct symbols). After the language code conversion, with the inclusion of prosody information, the information rate rises to about 200 bps. Somewhere in the next stage the representation of the information in the signal (or the control) becomes continuous with an equivalent rate of about 2000 bps at the neuromuscular control level, and about 30,000-50,000 bps at the acoustic signal level.

A transmission channel is shown in **Figure 2.2** [6], indicating that any of several well-known coding techniques could be used to transmit the acoustic waveform from the talker to the listener. The steps in the speech-perception mechanism can also be interpreted in terms of information rate in the signal or its control, and follow the inverse pattern of the production process. Thus the continuous information rate at the basilar membrane is in the range of 30,000-50,000 bps. The higher level processing within the brain converts the neural signals to a discrete representation, which ultimately is decoded into a low-bit-rate message.

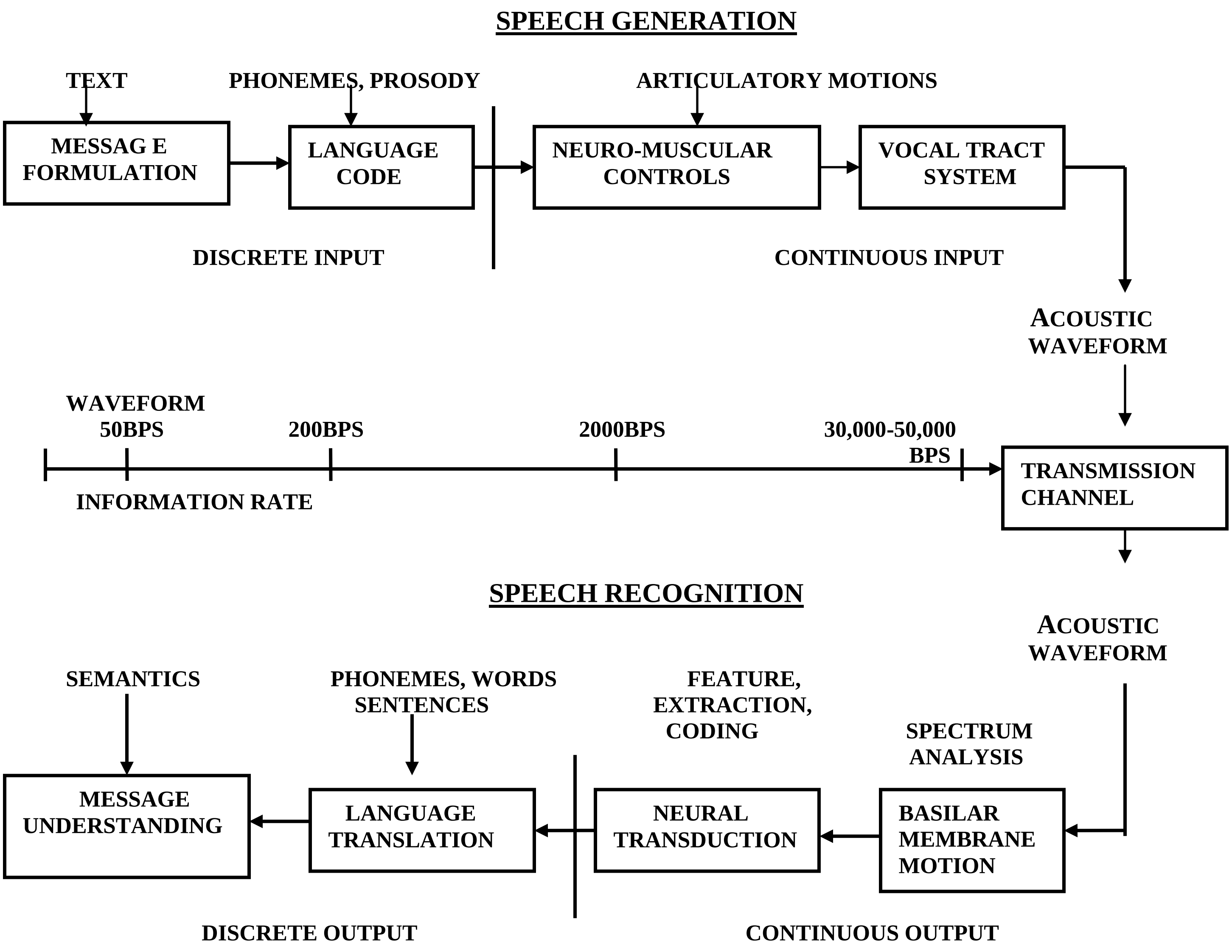
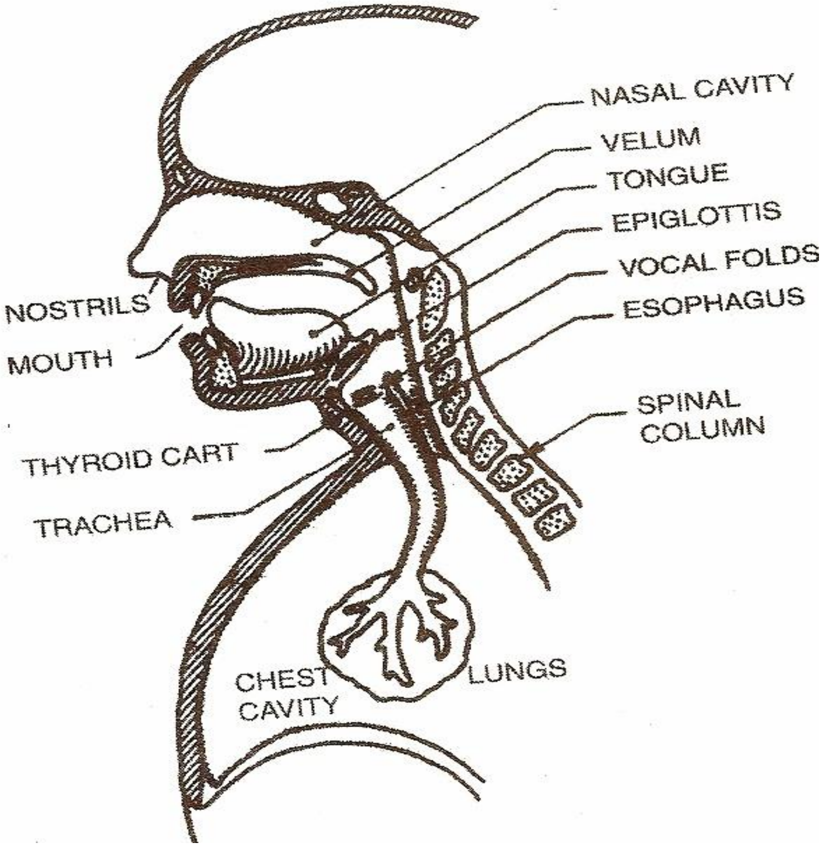


Figure 2.2 Alternative view of speech-production/speech perception process

THE SPEECH PRODUCTION PROCESS

A schematic diagram of the human vocal mechanism is shown in **Figure 2.3** [7]. The vocal tract begins at the opening of the vocal cords, or glottis, and ends at the lips. The vocal tract consists of the pharynx (the connection from the esophagus to the mouth) and the mouth, or oral cavity. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract, determined by the positions of the tongue, lips, jaw, velum, varies from zero (complete closure) to about 20 $cm^{2.}$ The nasal tract begins at the velum and end at the nostrils. When then velum, (a trapdoor-like mechanics at the back of the mouth cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.



Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the trachea (or windpipe), the tensed vocal cords within the larynx are caused to vibrate (in the mode of a relaxation oscillator) by the air flow. The air flow is chopped into quasi-periodic pulse which is then modulated in frequency in passing through the pharynx (the throat cavity), the mouth cavity, and possibly the nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, mouth), different sounds are produced.

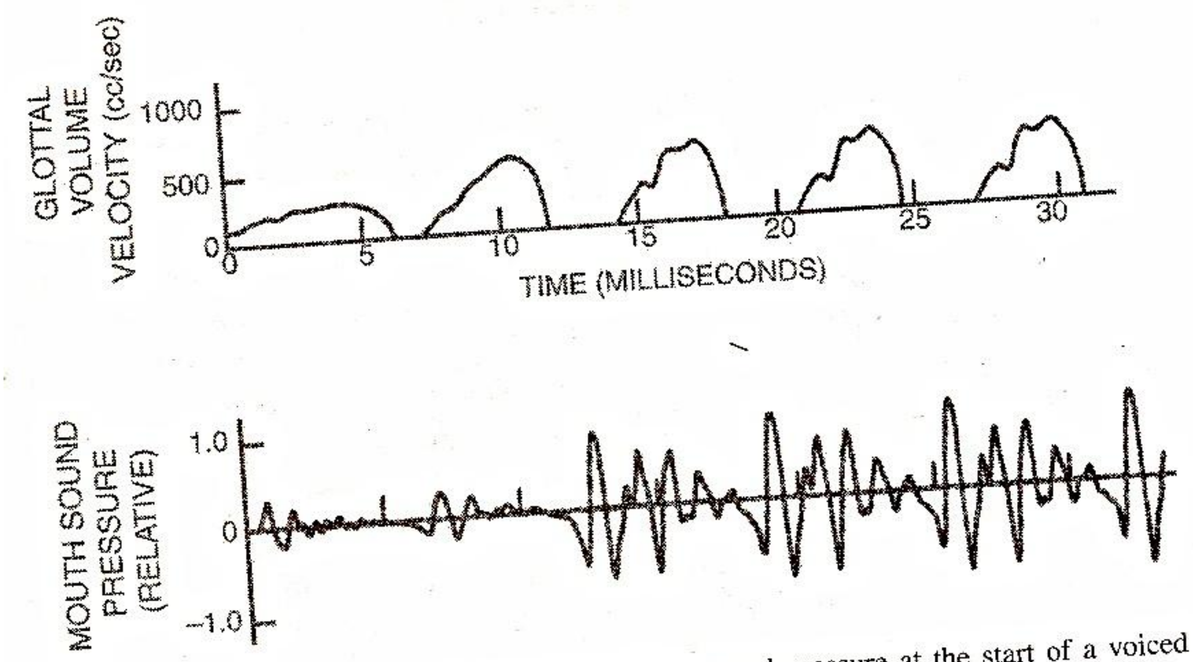


Figure 2.4 Glottal volume velocity and resulting sound pressure at the start of a voiced sound (after Ishizaka and Flanagan [3]).

Figure 2.4 shows plots of the glottal air flow (volume velocity waveform) and the resulting sound pressure at the mouth for a typical vowel sound [8].

The glottal waveform shows a gradual build- up to a quasi periodic pulse train of air, taking about 15 msec to reach steady state. This build up is also reflected in the acoustic waveform shown at the bottom of the figure.

A simplified representation of the complete physiological mechanism for creating speech is shown in **Figure 2.5[**9].

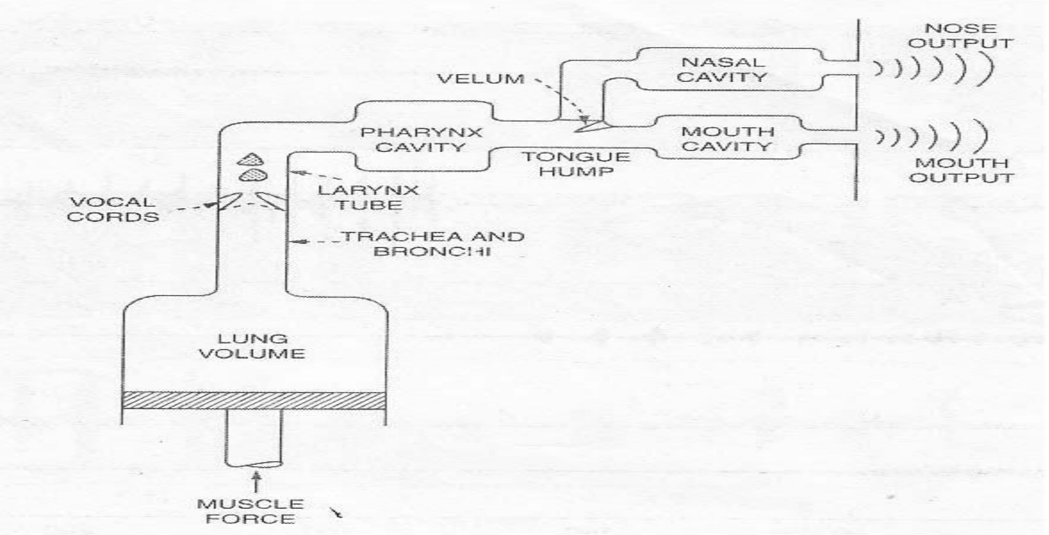


Figure 2.5 Schematic representation of the complete psysiological mechanism of speech production.

The lungs and the associated muscles act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the bronchi and trachea. When the vocal cords are tensed, the air flow causes them to vibrate, producing so- called voiced speech sound, the air flow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sounds, or it can build up pressure behind a point of total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and adruptly released, causing a brief transient sound.

Speech is produced as a sequence of sounds. Hence the state of the vocal cords, as well as the positions, shapes, and sizes of the various articulators, changes over time to reflect the sound being produced.

REPRESENTING SPEECH IN THE TIME AND FREQUENCY DOMAINS

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. An illustration of this effect is given in **figure 2.6**, which shows the time waveform corresponding to the utterance of the word “speech” as spoken by a male speaker. Each line of the waveform corresponds to 250 msec (1/4 second) of signal; hence the entire plot encompasses about 1 sec.

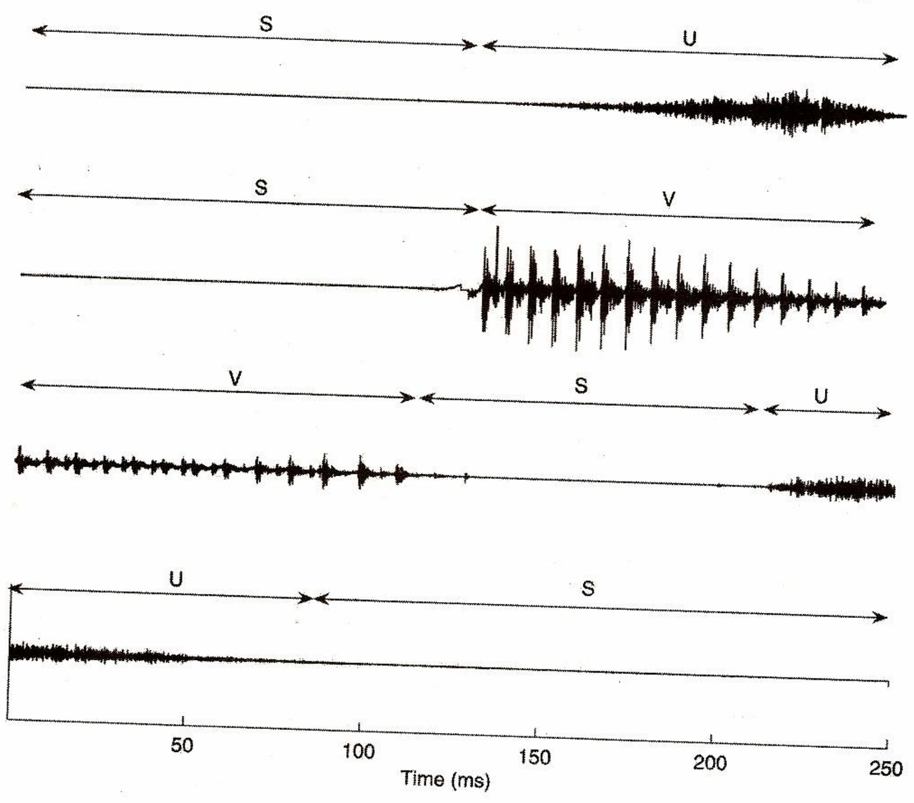


Figure 2.6 Waveform of the utterance “speech.”

There are several ways of classifying (labeling) events in speech. Perhaps the simplest and most straightforward is via the state of the source of speech production-the vocal cords. It is an accepted convention to use a three-state rep-resentation in which the states are (1) silence (S), where no speech is produced; (2) unvoiced (U), in which the vocal cords are not vibrating, so the resulting speech waveform is aperiodic or random in nature; and (3)voiced (V), in which the vocal waveform is aperiodic or random in nature;and(3) voiced (V),in which the vocal cords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting speech waveform is quasi-periodic. The result of applying this type of classification to the waveform of Figure **2.6** is shown in the figure.

It should be clear that the segmentation of the waveform into well defined regions of silence, unvoiced and voiced signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or a weak , voiced sound (like /v/ or /m/) from unvoiced sounds or even silence. However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications.[10]

Conclusion

In this paper I have presented a brief discussion of the basic speech-production/perception mechanism in human beings, and illustrated how we can exploit the so- called acoustic-phonetic properties of speech to identify basic sounds. Acoustic phonetics is the broad underpinning of all speech recognition work. Differences in approach lie in the degree of reliance on how much acoustic phonetics can be used in the recognition process. At one extreme is the class of pattern-recognition approaches that do not make a priori assumptions on the phonetic characteristics and instead choose to “relearn” the appropriate acoustic-phonetic mapping for specific word vocabularies and tasks via an appropriately designed training set. Finally, there is the hybrid class of artificial intelligence approaches that exploit, in various degrees, aspects of both extreme views of the speech-recognition process.

References

[1] K.H Davis, R. Biddulph, and S. Balashek, “Automatic Recognition of Spoken Digits”, J. Acoust. Soc.Am.,24 (6): 637-642,1952.

[2] D.B. Fry, Theoretical Aspects of Mechanical Speech Recognition”; and P.Denes, “The Design and Operation of The Mechanical Speech Recognizer At University College London,” J. British inst. Radio Engr., 19: 4, 211-229, 1959.

[3] J.W. Forgie and C.D. Forgie, Result Obtained From a Vowel Recognition Computer Programme, J. Acoust .Soc.Am.,31 (11):1480-1489, 1959.

[4] J. Suzuki and K. Nakata, “Recognition of Japanese Vowels-Preliminary to the Recognition of Speech,” J. Radio Res. Lab, 37 (8): 193-212, 1961.

[5] T. B. Martin,A.L.Nelson, and H. J. Zadell, “Speech Recognition by Feature Abstraction Techniques,”Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[6] D.R. Reddy, “An Approach to Computer Speech Recognition By Direct Analysis of the Speech Wave,”Tech. Report No. C549, Computer Science Dept., Stanford Univ.,September 1966.

[7] L.R.Rabiner and S.E Levinson, “Isolated and connected Word Recognition-Theory and Selected Applications,”IEEE Trans. Communications, COM-29(5):621-659, May 1981.

[8] K.Ishizaka and J.L. Flanagan,” Synthesis of Voiced Sounds from a Two- Mass Model of the Vocal Cords, “Bell System Tech.J.,50(6):1233-1268, July-Aug.,1972.

[9] J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed., Springer-Verlag, New York,1972.

[10] J.E. Shoup,”Phonological Aspects of Speech Recognition,”125-138, Ch.6 in Trends in Speech Recognition, W.A.Lea, E.d., Prentice-Hall, Englewood Cliffs, N.J.1980.