

Chapter 2

Speech Processing Technology: Basic Consideration

Abstract In this chapter, we have described the basic considerations of speech processing technology. Starting with the fundamental description of speech recognition systems, we have described the speech communication chain, mechanisms of speech perception and the popular theories, and model of spoken word recognition.

Keywords Spoken word · Speech communication · Perception · Recognition · Hearing

2.1 Fundamentals of Speech Recognition

Speech recognition is special case of speech processing. It deals with the analysis of the linguistic contents of a speech signal. Speech recognition is a method that uses an audio input for data entry to a computer or a digital system in place of a keyboard. In simple terms, it can be a data entry process carried out by speaking into a microphone so that the system is able to interpret the meaning out of it for further modification, processing, or storage. Speech recognition research is interdisciplinary in nature, drawing upon work in fields as diverse as biology, computer science, electrical engineering, linguistics, mathematics, physics, and psychology. Within these disciplines, pertinent work is being done in the areas of acoustics, artificial intelligence, computer algorithms, information theory, linear algebra, linear system theory, pattern recognition, phonetics, physiology, probability theory, signal processing, and syntactic theory [1].

Speech is a natural mode of communication for human. Human start to learn linguistic information in their early childhood, without instruction, and they continue to develop a large vocabulary with the confines of the brain throughout their lives. Humans achieve this skill so naturally that it never facilitates anything specific enabling a person to realize how complex the process is. The human vocal tract and articulators are biological organs with non-linear properties, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. As a result, vocalizations can vary widely in terms

of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. Moreover, during transmission, our irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics. All these sources of variability make speech recognition, a very complex problem [2]. Speech recognition systems are generally classified as follows:

1. Continuous Speech Recognition and
2. Discrete Speech Recognition

Discrete systems maintain a separate acoustic model for each word, combination of words, or phrases and are referred to as isolated (word) speech recognition (ISR). Continuous speech recognition (CSR) systems, on the other hand, respond to a user who pronounces words, phrases, or sentences that are in a series or specific order and are dependent on each other, as if linked together.

Discrete speech was used by the early speech recognition researchers. The user has to make a short pause between words as they are dictated. Discrete speech systems are particularly useful for people having difficulties in forming complete phrases in one utterance. The focus on one word at a time is also useful for people with a learning difficulty.

Today, continuous speech recognition systems dominate [3]. However, discrete systems are useful than continuous speech recognition systems in the following situations:

1. For people with speech and language difficulties, it can be very difficult to produce continuous speech. Discrete speech systems are more effective at recognizing nonstandard speech patterns, such as dysarthric speech.
2. Some people with writing difficulties benefit from concentrating on a single word at a time that is implicit in the use of a discrete system. In the continuous speech, people need to be able to speak at least a few words, with no pause, to get a proper level of recognition. Problems can appear if one changes his/her mind about a word, in the middle of a sentence.
3. Children in school need to learn to distinguish between spoken English and written English. In a discrete system, it is easier to do, so since attention is focused on individual words.

However, discrete speech is an unnatural way of speaking, whereas continuous speech systems allow the user to speak with a natural flow at a normal conversational speed [3]. Both discrete and continuous systems may be speaker dependent, independent, or adaptive. Speaker-dependent software works by learning the unique characteristics of a single person's voice, in a way similar to voice recognition. New users must first train the software by speaking to it, so the computer can analyze how the person talks. This often means, users have to read a few pages of text to the computer before they can use the speech recognition software. Conversely, speaker-independent software is designed to recognize anyone's voice, so no training is involved. This means that, it is the only real option for applications such as interactive voice response systems, where businesses cannot ask callers to read pages of text

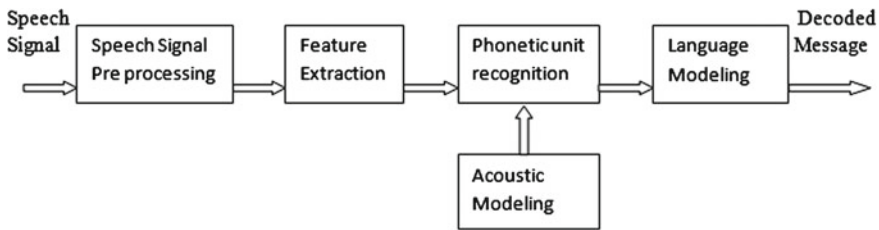


Fig. 2.1 Structure of a standard speech recognition system [2]

before using the system. The other side of the picture is that speaker-independent software is generally less accurate than speaker-dependent software [4].

Speech recognition engines that are speaker independent generally deal with this fact by limiting the grammars they use. By using a smaller list of recognized words, the speech engine is more likely to correctly recognize what a speaker said. Speaker-dependent software is commonly used for dictation software, while speaker-independent software is more commonly found in telephone applications [4].

Modern speech recognition systems also use some adaptive algorithm so that the system becomes adaptive to speaker-varying environment. Adaptation is based on the speaker-independent system with only limited adaptation data. A proper adaptation algorithm should be consistent with speaker-independent parameter estimation criterion and adapt those parameters that are less sensitive to the limited training data [5]. The structure of a standard speech recognition system is illustrated in Fig. 2.1 [2]. The elements of the system can be described as follows:

1. **Raw speech:** Speech is typically sampled at a high frequency, for example, 16 kHz over a microphone or 8 kHz over a telephone. This yields a sequence of amplitude values over time.
2. **Pre-processing:** Prior to applying the speech signal, preprocessing of speech signals is required. Although the speech signals contain all the necessary information within 4 kHz, raw speech signals naturally have very large bandwidth. High-frequency speech signals are always less informative and contains noise. Therefore, to make the signal noiseless, preprocessing involves a digital filtering operation. Various structures of digital filters can be used for these purposes.
3. **Signal analysis:** Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available that can extract useful features and compress the data by a factor of ten without losing any important information.
4. **Speech frames:** The result of signal analysis is a sequence of speech frames, typically at 10ms intervals, with about 16 coefficients per frame. These frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics. This typically leads to improved performance. The speech frames are used for acoustic analysis [2].

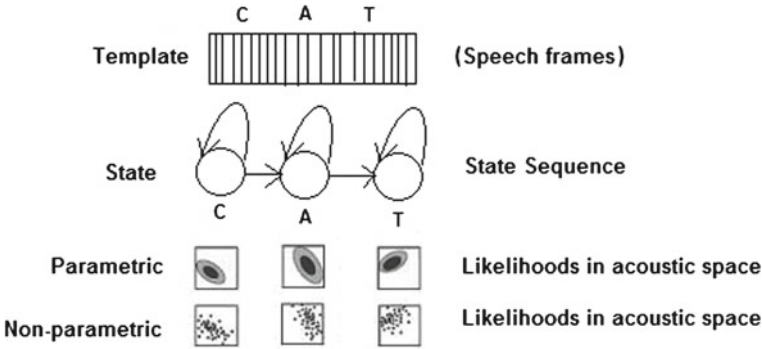


Fig. 2.2 Acoustic models: template and state representations for the word ‘cat’ [2]

5. **Acoustic models:** In order to analyze the speech frames for their acoustic content, we need a set of acoustic models. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties. Figure 2.2 shows two popular representations for acoustic models. The simplest is a template, which is just a stored sample of the unit of speech to be modeled, for example, a recording of a word. An unknown word can be recognized by simply comparing it against all known templates and finding the closest match. Templates have some major drawbacks. One is, they cannot model acoustic variabilities, except in a coarse way by assigning multiple templates to each word. Furthermore, they are limited to whole word models, because it is difficult to record or segment a sample shorter than a word. Hence, templates are useful only in small systems.

A more flexible representation, used in larger systems, is based on trained acoustic models, or states. In this approach, every word is modeled by a sequence of trainable states, and each state indicates the sounds that are likely to be heard in that segment of the word, using a probability distribution over the acoustic space. Probability distributions can be modeled parametrically, by assuming that they have a simple shape like Gaussian distribution and then trying to find the parameters that describe it. Another method of modeling probability distributions is nonparametric, that means, by representing the distribution directly with a histogram over a quantization of the acoustic space or as we shall see with an ANN.

2.2 Speech Communication Chain

The fundamental purpose of generation of speech is communication [6]. Speech communication involves a chain of physical and psychological events. The speaker initiates the chain by setting into motion the vocal apparatus, which propels a modulated airstream to the listener. The process culminates with the listener receiving the

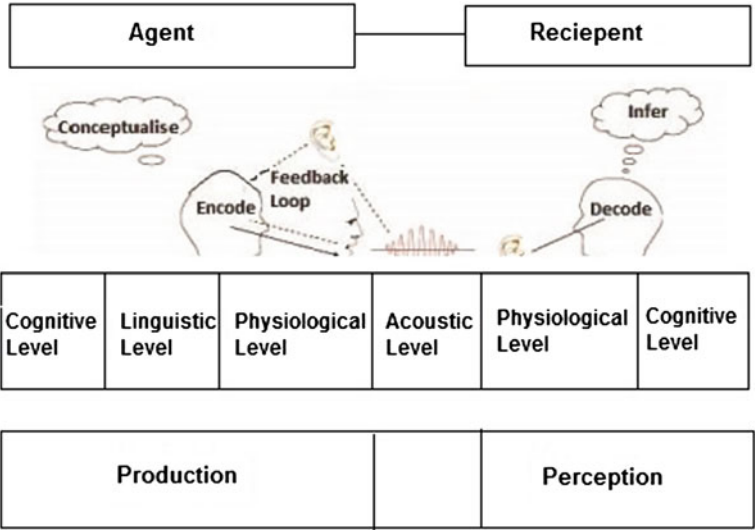


Fig. 2.3 Communication model [8]

fluctuations in air pressure through the auditory mechanisms and subsequently parsing the signal into higher-level linguistic units [7]. Any communication model is just a conceptual framework that simplifies our understanding of the complex communication process. This section discusses a speech transmission model, underpinned by Information Theory, known as the Speech Communication Chain. The three elements of the speech chain are production, transmission, and perception which depends upon the function of the cognitive, linguistic, physiological, and acoustic levels for both the speaker and recipient listener who are explained. Cognitive neuropsychological models that attempt to explain human communication, and language in particular, in terms of human cognition and the anatomy of the brain [8].

Figure 2.3 shows the communication chain underpinned by Shannon and Weaver’s Information Theory and highlights the linguistic, physiological, and acoustic mechanisms by which humans encode, transmit, and decode meanings. This model is an extension of the Speech Chain, originally outlined by Denes and Pinson (1973) [8], as shown in Fig. 2.4 [8].

- There are three main links in the communication chain:
1. Production is the process by which a human expresses himself or herself through first deciding what message is to be communicated (cognitive level). Next a plan is prepared and encoded with appropriate linguistic utterance to represent the concept (linguistic level) and, finally, produce this utterance through the suitable coordination of the vocal apparatus (physiological level) [8]. Production of a verbal utterance, therefore, takes place at three levels (Fig. 2.4):
 - a. **Cognitive:** When two people talk together, the speaker sends messages to the listener in the form of words. The speaker has first to decide what he or she

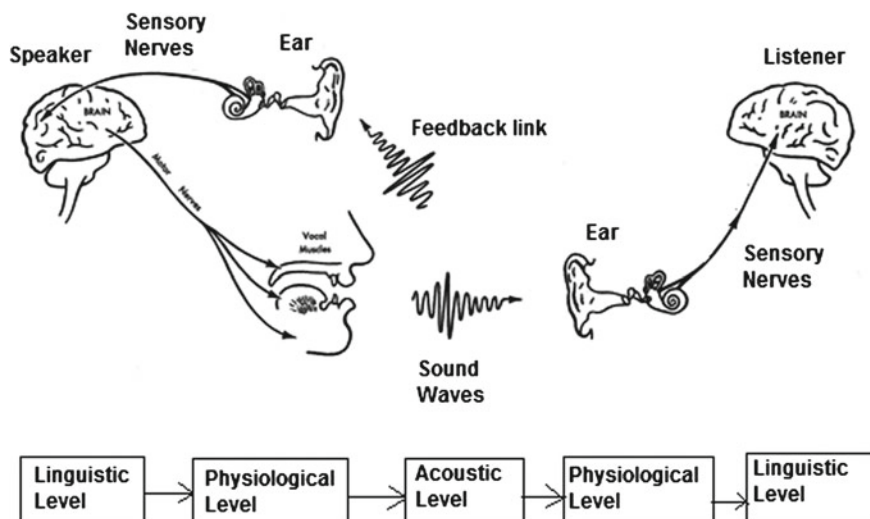


Fig. 2.4 The speech chain [8]

wants to say and then to choose the right words to put together in order to send the message. These decisions are taken in a higher level of the brain known as the cortex [8].

- b. **Linguistic:** According to connectionist model, there are four layers of processing at the linguistic level: semantic, syntactic, morphological, and phonological. These work in parallel and in series, with activation at each level. Interference and misactivation can occur at any of these stages. Production begins with concepts and continues down from there [9]. Human has a bank of words stored in brains known as lexicon. It is built up over time, and the items we store are different from person to person. The store is largely dependent upon what one have been exposed to, such as the job of work it does, where one have lived, and so on. Whenever human needs to encode a word, a search is made within this lexicon in order to determine whether or not it already contains a word for the idea which is to be conveyed [8].
- c. **Physiological:** Once the linguistic encoding has taken place, the brain sends small electrical signals along nerves from the cortex to the mouth, tongue, lips, vocal folds (vocal cords), and the muscles which control breathing to enable us to articulate the word or words needed to communicate our thoughts. This production of the sound sequence occurs at what is known as the physiological level, and it involves rapid, coordinated, sequential movements of the vocal apparatus.

It is clear that proper production depends upon normal perception. The prelingually deaf are clearly at a disadvantage in the learning of speech. Since normal hearers are both transmitters and perceivers of speech, they constantly perceive

their own speech as they utter it, and on the basis of this auditory feedback, they instantly correct any slips or errors. Infants learn speech articulation by constantly repeating their articulations and listening to the sounds produced [8]. This continuous feedback loop eventually results in the perfection of the articulation process [10].

2. Transmission is the sending of the linguistic utterance through some medium to the recipient. As we are only concerned with oral verbal communication here, there is only one medium of consequence, and that is air, i.e., the spoken utterance travels through the medium of air to the recipient's ear. There are, of course, other media through which a message could be transmitted. For example, written messages may be transmitted with ink and paper. However, because here, the only concern is the transmission of messages that use a so-called vocal-auditory channel, then transmission is said to occur at the acoustic level [8].
3. Reception is the process by which the recipient of a verbal utterance detects the utterance through the sense of hearing at physiological level and then decodes the linguistic expression at linguistic level. Then, the recipient infers what is meant by the linguistic expression at cognitive level [8]. Like the mirror image of production, reception also operates at the same three levels:
 - a. **Physiological:** When the speaker's utterance transmitted acoustically as a speech sound wave arrives at the listener's ear, it causes his or her eardrum to vibrate. This, in turn, causes the movement of three small bones within the middle ear. Their function is to amplify the vibration of the sound wave. One of these bones, the stapes, is connected to a membrane in the wall of nerve bundle called the cochlea. The cochlea is designed to convert the vibrations into electrical signals. These are subsequently transmitted along the 30,000 or so fibres that constitute the auditory nerve to the brain of the listener. Again, this takes place at the physiological level.
 - b. **Linguistic:** The listener subsequently decodes these electrical impulses in the cortex and reforms them into the word or words of the message, again at the linguistic level. The listener compares the decoded words with other words in its own lexicon. The listener is then able to determine that the word is a proper word. In order for a recipient to decode longer utterances and to interpret them as meaningful, it must also make use of the grammatical rules stored in the brain. These allow the recipient to decode an utterance such as "I have just seen a cat" as indicating that the event took place in the recent past, as opposed to an utterance such as "I will be seeing a cat" which grammatically indicates that the event has not yet happened but will happen some time in the future. Consequently, as with the agent, the recipient also needs access to a lexicon and a set of grammatical rules in order to comprehend verbal utterances.
 - c. **Cognitive:** In this level, the listener must infer the speaker's meaning. A linguistic utterance can never convey all of a speaker's intended meaning. A linguistic utterance is a sketch, and the listener must fill in the sketch by inferring the meaning from such things as body language, shared knowledge, tone of voice, and so on. Humans must, therefore, be able to infer meanings in order to communicate fully.

2.3 Mechanism of Speech Perception

Speech perception is the process by which the sounds of language are heard, interpreted, and understood. The study of speech perception is closely linked to the fields of phonetics and phonology in linguistics and cognitive psychology and perception in psychology. Generally speaking, speech perception proceeds through a series of stages in which acoustic cues are extracted and stored in sensory memory and then mapped onto linguistic information. When air from the lungs is pushed into the larynx across the vocal cords and into the mouth nose, different types of sounds are produced. The different qualities of the sounds are represented in formants, which can be pictured on a graph that has time on the x -axis and the pressure under which the air is pushed, on the y -axis. Perception of the sound will vary as the frequency with which the air vibrates across time varies. Because vocal tracts vary somewhat between people, one person's vocal cords may be shorter than another's, or the roof of someone's mouth may be higher than another's, and the end result is that there are individual differences in how various sounds are produced. The acoustic signal is in itself a very complex signal, possessing extreme interspeaker and intraspeaker variability even when the sounds being compared are finally recognized by the listener as the same phoneme and are found in the same phonemic environment. Furthermore, a phoneme's realization varies dramatically as its phonemic environment varies. Speech is a continuous unsegmented sequence and yet each phoneme appears to be perceived as a discrete segmented entity. A single phoneme in a constant phonemic environment may vary in the cues present in the acoustic signal from one sample to another. Also, one person's utterance of one phoneme may coincide with another person's utterance of another phoneme, and yet both are correctly perceived [10].

2.3.1 Physical Mechanism of Perception

The primary organ responsible for initiating the process of perception is the ear. The operation of the ear has two parts—the behavior of the mechanical apparatus and the neurological processing of the information acquired. Hearing, one of the five senses of human, is the ability to perceive sound by detecting vibrations via the organ ear. Mechanical process of the ear is the beginning of the speech perception process.

The mechanism of hearing comprises a highly specialized set of structures. The anatomy of ear structures and the physiology of hearing allow sound to move from the environment through the structures of the ear and to the brain. We rely on this series of structures to transmit information, so it can be processed to get information about the sound wave. The physical structure of the ear has three sections—the outer, middle ear, and the inner ear [11]. Figure 2.5 shows the parts of the ear. The outer ear consists of the lobe and ear canal, serve to protect the more delicate parts inside. The function of the outer ear is to trap and concentrate sound waves, so their specific messages can be communicated to the other ear structures. The middle ear begins to

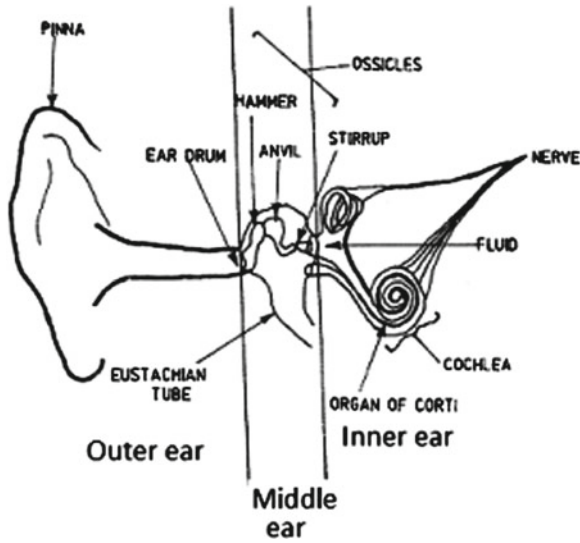


Fig. 2.5 Parts of the ear

process sounds specifically and reacts specifically to the types of sounds being heard. The outer boundary of the middle ear is the eardrum, a thin membrane that vibrates when it is struck by sound waves. The motion of the eardrum is transferred across the middle ear via three small bones named the hammer, anvil, and stirrup. These bones form a chain, which transmits sound energy from the eardrum to the cochlea. These bones are supported by muscles which normally allow free motion but can tighten up and inhibit the bones action when the sound gets too loud. The inner ear is made up of a series of bony structures, which house sensitive, specialized sound receptors. These bony structures are the cochlea, which resembles a snail shell, and the semicircular canals. The cochlea is filled with fluid which helps to transmit sound and is divided in two the long way by the basilar membrane. The cochlea contains microscopic hairs called cilia. When moved by sound waves traveling through the fluid, they facilitate nerve impulses along the auditory nerve. The boundary of the inner ear is the oval window, another thin membrane that is almost totally covered by the end of the stirrup. Sound introduced into the cochlea via the oval window flexes the basilar membrane and sets up traveling waves along its length. The taper of the membrane is such that these traveling waves are not of even amplitude the entire distance, but grow in amplitude to a certain point and then quickly fade out. The point of maximum amplitude depends on the frequency of the sound wave. The basilar membrane is covered with tiny hairs, and each hair follicle is connected to a bundle of nerves. Motion of the basilar membrane bends the hairs which in turn excites the associated nerve fibers. These fibers carry the sound information to the brain, which has two components. First, even though a single nerve cell cannot react fast enough to follow audio frequencies, enough cells are involved that the aggregate of all the

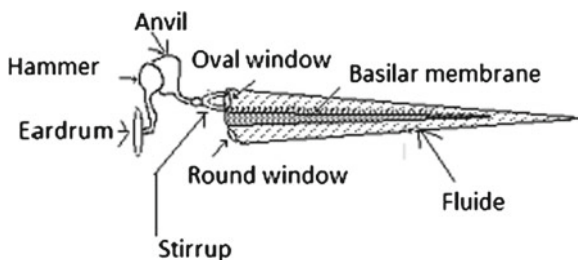


Fig. 2.6 Schematic of the ear

firing patterns is a fair replica of the waveform. Second, and most importantly, the location of the hair cells associated with the firing nerves is highly correlated with the frequency of the sound. A complex sound will produce a series of active loci along the basilar membrane that accurately matches the spectral plot of the sound. The amplitude of a sound determines how many nerves associated with the appropriate location fire, and to a slight extent, the rate of firing. The main effect is that a loud sound excites nerves along a fairly wide region of the basilar membrane, whereas a soft one excites only a few nerves at each locus [11, 12]. The schematic of the ear is shown in Fig. 2.6.

2.3.2 Perception of Sound

In the complex process of speech perception, a series of stages are involved in which acoustic cues are extracted and stored in sensory memory and then mapped onto linguistic information. Yet perception basically involves the neural operation of hearing. When sound of a particular waveform and frequency sets up a characteristic pattern of active locations on the basilar membranes, brain deals with these patterns in order to decide whether it is speech or non-speech sound and recognize that particular pattern. The brain's response may be quite similar with the way brain deals with visual patterns on the retina. If a pattern is repeated enough brain learns to recognize that pattern as belonging to a certain sound, as like as human learn a particular visual pattern belonging to a certain face. The absolute position of the pattern is not very important, it is the pattern itself that is learned. Human brain possess an ability to interpret the location of the pattern to some degree, but that ability is quite variable from one person to the next [11].

Although the perception process is not so simple as described above. The process involves lots of complexities and variations. Three main effects that are crucial for understanding how the acoustic speech signal is transformed into phonetic segments and mapped into linguistic patterns are—invariance and linearity, constancy, and perceptual units. All these are described below:

- **Variance and non-linearity:** The problems of linearity and invariance are the effects of coarticulation. Coarticulation in speech production is a phenomenon in which the articulator movements for a given speech sound vary systematically with the surrounding sounds and their associated movements [13]. Simply it is the influence of articulation of one phoneme on that of another phoneme. The acoustic feature of one sound will spread themselves across those of another sound, which will lead to the problem of nonlinearity, that is, for a particular phoneme in the speech, there should be single corresponding section in the speech waveform if speech were produced linearly. However, the speech is not linear. Therefore, it is difficult to decide where one phoneme ends and the other phoneme starts. For example, it is difficult to find the location of ‘th’ sound and the ‘uh’ sound in the word ‘the.’ Again invariance refers to a particular phoneme having one and only one waveform representation. For example, the phoneme /i/ (the “ee” sound in “me”) should have the identical amplitude and frequency as the same phoneme in “money.” But this is not the case, the two waveforms differ. The plosives or stop consonants, /b/, /d/, /g/, /k/, provide particular problems for the invariance assumption.
- **Inconstancy:** The perceptual system must cope with some additional sources of variability. When different talkers produce the same intended speech sound, the acoustic characteristics of their productions will differ. Women’s voices tend to be higher pitched than men’s, and even within a particular sex, there is a wide range of acoustic variability. Individual differences in talker’s voice are due to the size of the vocal tracts, which vary from person to person. This affects the fundamental frequency of a voice. Factors like flexibility of the tongue, the state of one’s vocal folds, missing teeth, etc., also will affect the acoustic characteristics of speech. In many cases, production of the same word is different by a single talker. Again in some cases, production of a particular word by one talker is acoustically similar with the production of a different word by another talker. Changes in speaking rate also affected the speech signal. Differences in voice and voiceless stops tend to decrease as speaking rate speeds. Furthermore, the nature of the acoustic changes is not predictable a priori.
- **Perceptual unit:** Phoneme and the syllable both have often been proposed as the ‘primary unit of speech perception’ [14, 15]. Phonemes are the unit of speech smaller than the words or syllables. Many early studies reveal that perception depends on the phonetic segmentation considering phonemes as the basic unit of speech. Phonemes contain the distinctive features that are combined into the one concurrent bundle called the phonetic segments. But it is not possible to separate formants of consonants and vowel syllable such as /di/. In contrast, many researchers argued that syllables are the basic unit of speech since listeners detect syllables faster than the phoneme targets [14].

2.3.3 Basic Unit of Speech Perception

The process of perceiving speech begins at the level of the sound signal and the process of audition. After processing the initial auditory signal, speech sounds are further processed to extract acoustic cues and phonetic information. This speech information can then be used for higher-level language processes, such as word recognition. The speech sound signal contains a number of acoustic cues that are used in speech perception [16]. The cues differentiate speech sounds belonging to different phonetic categories. One of the most studied cues in speech is voice onset time (VOT). VOT is a primary cue signaling the difference between voiced and voiceless stop consonants, such as ‘b’ and ‘p’. Other cues differentiate sounds that are produced at different places of articulation or manners of articulation. The speech system must also combine these cues to determine the category of a specific speech sound. This is often thought of in terms of abstract representations of phonemes. These representations can then be combined for use in word recognition and other language processes. If a specific aspect of the acoustic waveform indicated one linguistic unit, a series of tests using speech synthesizers would be sufficient to determine such a cue or cues [17]. However, there are two significant obstacles, that is,

1. One acoustic aspect of the speech signal may cue different linguistically relevant dimensions.
2. One linguistic unit can be cued by several acoustic properties.

Speech perception is best understood in terms of an acoustic signal, phonemes, lexicals, syntax, and semantics. The human ear receives sound waves called the acoustic signal. Phonemes are the basic units of human speech, smaller than words or syllables, which allow us to distinguish between the words. Lexical relates to the words or vocabulary of a language, and syntax refers to the combination of words (or grammar) forming language. Finally, semantics refers to the meaning of the spoken message. Experts agree that speech perception does not involve simply the reception of acoustic signals and its direct translation to a semantic understanding of the message. By applying the principle of recursive decomposition, analysis of speech perception can begin with the observation that sentences are composed of words composed of syllables that are arranged grammatically. But neither the syllable nor the phoneme is the basic unit of speech perception [18].

A single phoneme is pronounced differently based on its combination with other phonemes. Consider the use of the phoneme /r/ in the words red and run. If one were to record the sound of both spoken words and then isolate the /r/ phoneme, the vowel phonemes would be distinguishable. This effect is known as coarticulation which is used to refer to the phenomenon in which the mouth position for individual phonemes in the utterance of continuous sounds incorporates the effects of the mouth position for phonemes uttered immediately before and after [19]. The articulatory gestures required to produce these words blend the phonemes, making it impossible

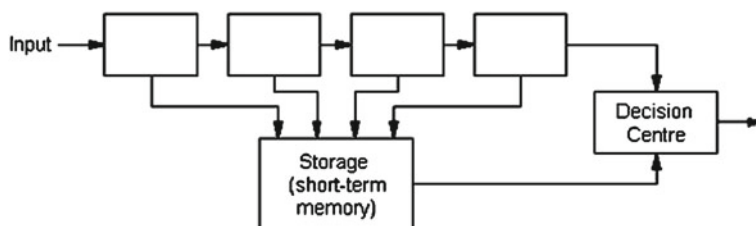


Fig. 2.7 Series processing

to distinguish them as separate units. Thus, the basic unit of speech perception is coarticulated phonemes [18].

2.4 Theories and Models of Spoken Word Recognition

The basic theories of speech perception provide a better explanation about how the process works. In series models, there is a sequential order to each subprocess. A contrast to series models is parallel models which feature several subprocesses acting simultaneously [18].

The perception of speech involves the recognition of patterns in the acoustic signal in both time and frequency domains. Such patterns are realized acoustically as changes in amplitude at each frequency over a period of time [10]. Most theories of pattern processing involve series, arrays, or networks of binary decisions. In other words, at each step in the recognition process, a yes/no decision is made as to whether the signal conforms to one of two perceptual categories. The decision thus made usually affects which of the following steps will be made in a series of decisions. Figure 2.7 shows the series processing scheme [10]. The series model of Bondarko and Christovich begins with the listener receiving speech input and then subjecting it to auditory analysis. Phonetic analysis then passes the signal to a morphological analysis block. Finally, a syntactic analysis of the signal results in a semantic analysis of the message [18]. Each block in the series is said to reduce/refine the signal and pass additional parameters to the next level for further processing. Liberman proposes a top-down, bottom-up series model which demonstrates both signal reception and comprehension with signal generation and transmission in the following format: acoustic structure sound, speech phonetic structure, phonology, surface structure, syntax, deep structure, semantics, and conceptual representation. Series models imply that decisions made at one level of processing affect the next block, but do not receive feedback. Lobacz suggests that speech perception is, in reality, more dynamic and that series models are not an adequate representation of the process [18].

The decision made in series processing usually affects which following steps will be made in a series of decisions. If the decision steps are all part of a serial processing

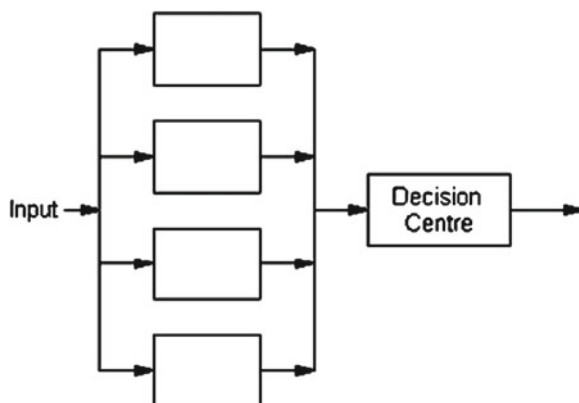


Fig. 2.8 Parallel processing

chain, then a wrong decision at an early stage in the pattern recognition process may cause the wrong questions to be asked in subsequent steps. It means that each step may be under the influence of previous steps. A serial processing system also requires a facility to store each decision so that all the decisions can be passed to the decision center when all the steps have been completed. Because of all these problems with serial processing strategies, most speech perception theorists prefer at least some sort of parallel processing as shown in Fig. 2.8. In parallel processing, all questions are asked simultaneously, that is, all cues or features are examined at the same time, and so processing time is very short no matter how many features are examined. Since all tests are processed at the same time, there is no need for the short-term memory facility, and further, there is also no effect of early steps on following steps, that is, no step is under the influence of a preceding step [10]. Thus, it can be said that the need to extract several phonemes from any syllable is taken as evidence that speech perception is a parallel process. Parallel models demonstrate simultaneous activity in the subprocesses of speech perception, showing that a process at one level may induce processes at any other level, without hierarchical information transmission. Lobacz cites Fants model as an example. The model consists of five successive stages of processing:

- Acoustic parameter extraction,
- Microsegment detection,
- Identification of phonetic elements (phonetic transcription),
- Identification of sentence structure (identification of words), and
- Semantic interpretation

Each stage includes stores of previously acquired knowledge. The stores hold inventories of units extracted at the given level with constraints on their occurrence in messages. The system also includes comparators that are placed between the successive stages of processing. The model provides for a possibility of direct connection between the lowest and the highest levels [18].

Further, speech perception theories can be considered to be of two types or a combination of both:

1. **Passive or Non-mediated theories:** These theories are based on the assumption that there is some sort of direct relationship between the acoustic signal and the perceived phoneme. In other words, perceptual constancy is in some way matched to a real acoustic constancy. These theories tend to concentrate on discovering the identity of such constant perceptual cues, and on the way, the auditory system might extract them from the acoustic signal. In one way or another, these theories are basically filtering theories and do not involve the mediation of higher cognitive processes in the extraction of these cues. These higher processes are restricted to making a decision based on the features or cues which have been detected or extracted closer to the periphery of the auditory system [10].
2. **Active or mediated theories:** These theories, on the other hand, suggest that there is no direct relationship between the acoustic signal and the perceived phoneme but rather that some higher-level mediation is involved in which the input pattern is compared with an internally generated pattern [10].

Spoken word recognition is a distinct subsystem providing the interface between low-level perception and cognitive processes of retrieval, parsing, and interpretation of speech. The process of recognizing a spoken word starts from a string of phonemes, establishes how these phonemes should be grouped to form words, and passes these words onto the next level of processing.

Research on the discrimination and categorization of phonetic segments was the key focus of the works on speech perception before 1970s. The processes and representations responsible for the perception of spoken words became a primary object of scientific inquiry with a curiosity of disclosing the cause and methods of how the listener perceives fluent speech. The following sections provides a brief description of some renown models and theory of spoken word recognition found in the literature.

2.4.1 Motor Theory of Speech Perception

Beginning in the early 1950s, Alvin Liberman, Franklin Cooper, Pierre Delattre, and other researchers at the Haskins Laboratories carried out a series of landmark studies on the perception of synthetic speech sounds. This work provided the foundation of what is known about acoustic cues for linguistic units such as phonemes and features and revealed that the mapping between speech signals and linguistic units is quite complex. In time, Liberman and his colleagues became convinced that perceived phonemes and features have a simpler (i.e., more nearly one-to-one) relationship to articulation than to acoustics, and this gave rise to the motor theory of speech perception. Every version has claimed that the objects of speech perception are articulatory events rather than acoustic or auditory events [20].

What a listener does, according to this theory, is to refer the incoming signal back to the articulatory instructions that the listener would give to the articulators in order

to produce the same sequence. The motor theory argues that the level of articulatory or motor commands is analogous to the perceptual process of phoneme perception and that a large part of both the descending pathway (phoneme to articulation) and the ascending pathway (acoustic signal to phoneme identification) overlaps. The two processes represent merely two-way traffic on the same neural pathways. The motor theory points out that there is a great deal of variance in the acoustic signal and that the most peripheral step in the speech chain which possesses a high degree of invariance is at the level of the motor commands to the articulatory organs. The encoding of this invariant linguistic information is articulatory and so the decoding process in the auditory system must at least be analogous. The motor theorists propose that 'there exists an overlapping activity of several neural networks—those that supply control signals to the articulators, and those that process incoming neural patterns from the ear and that information can be correlated by these networks and passed through them in either direction' [10].

This theory suggests that there exists a special speech code (or set of rules) which is specific to speech and which bridges the gap between the acoustic data and the highly abstract higher linguistic levels. Such a speech code is unnecessary in passive theories, where each speech segment would need to be represented by a discrete pattern, that is, a template somehow coded into the auditory system at some point. The advantage of a speech code or rule set is that there is no need for a vast storage of templates since the input signal is converted into a linguistic entity using those rules. The rules achieve their task by a drastic restructuring of the input signal. The acoustic signal does not in itself contain phonemes which can be extracted from the speech signal (as suggested by passive theories), but rather contains cues or features which can be used in conjunction with the rules to permit the recovery of the phoneme which last existed as a phonemic entity at some point in the neuromuscular events which led to its articulation. This, it is argued, is made evident by the fact the speech can be processed 10 times faster than can non-speech sounds. Speech perception is seen as a distinct form of perception quite separate from that of non-speech sound perception. Speech is perceived by means of categorical boundaries, while non-speech sounds are tracked continuously. Like the proponents of the neurological theories proposed by Abbs and Sussman in 1971 [10], the motor theorists believe that speech is perceived by means of a special mode, but they believe that this is not based directly on the recognition of phonemes embedded in the acoustic signal but rather on the gating of phonetically processed signals into specialized neural units. Before this gating, both phonetic and non-phonetic processings have been performed in parallel, and the non-phonetic processing is abandoned when the signal is identified as speech [10].

Speech is received serially by the ear, and yet must be processed in some sort of parallel fashion since not all cues for a single phoneme coexist at the same point in time, and the boundaries of the cues do not correspond to any phonemic boundaries. A voiced stop requires several cues to enable the listener to distinguish it. VOT, that is, voice onset time, is examined to enable a voiced or voiceless decision, but this is essentially a temporal cue and can only be measured relative to the position of the release burst. The occlusion itself is a necessary cue for stops in general in a

Vowel–Consonant–Vowel (VCV) environment, and yet it does not coexist in time with any other cue. The burst is another general cue for stops, while the following aspiration (if present) contains a certain amount of acoustic information to help in the identification of the stop’s place of articulation and further helps in the identification of positive VOT and thus of voiceless stops. The main cues to the stop’s place of articulation are the formant transitions into the vowel, and these in no way co-occur with the remainder of the stop’s cues. This folding over of information onto adjacent segments, which we know as coarticulation, far from making the process of speech perception more confusing, actually helps in the determination of the temporal order of the individual speech segments, as it permits the parallel transmission via the acoustic signal of more than one segment at a time [10].

2.4.2 Analysis-by-Synthesis Model

Stevens and Halle in 1967 have postulated that ‘the perception of speech involves the internal synthesis of patterns according to certain rules, and a matching of these internally generated patterns against the pattern under analysis...moreover, ...the generative rules in the perception of speech [are] in large measure identical to those utilized in speech production, and that fundamental to both processes [is] an abstract representation of the speech event.’ In this model, the incoming acoustic signal is subjected to an initial analysis at the periphery of the auditory system. This information is then passed upward to a master control unit and is there processed along with certain contextual constraints derived from preceding segments. This produces an hypothesized abstract representation defined in terms of a set of generative rules. This is then used to generate motor commands, but during speech perception, articulation is inhibited and instead the commands produce a hypothetical auditory pattern which is then passed to a comparator module. It compares this with the original signal which is held in a temporary store. If a mismatch occurs, the procedure is repeated until a suitable match is found [10]. Figure 2.9 shows the Analysis-by-Synthesis model.

2.4.3 Direct Realist Theory of Speech Perception

Starting in the 1980s, an alternative to Motor Theory, referred to as the Direct Realist Theory (DRT) of speech perception, was developed by Carol Fowler, also working at the Haskins Laboratories. Like Motor Theory, DRT claims that the objects of speech perception are articulatory rather than acoustic events. However, unlike Motor Theory, DRT asserts that the articulatory objects of perception are actual, phonetically structured, vocal tract movements, or gestures, and not events that are causally antecedent to these movements, such as neuromotor commands or intended gestures. DRT also contrasts sharply with Motor Theory in denying that specialized, that is,

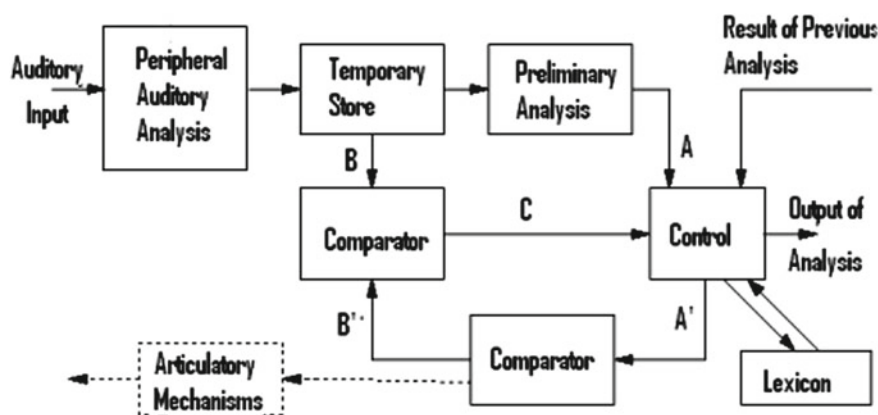


Fig. 2.9 Analysis-by-synthesis model (after Stevens 1972)

speech-specific or human-specific, mechanisms play a role in speech perception. This theory is elegantly summarized by Fowler in 1996 in the following passage: Perceptual systems have a universal function. They constitute the sole means by which animals can know their niches. Moreover, they appear to serve this function in one way. They use structure in the media that has been lawfully caused by events in the environment as information for the events. Even though it is the structure in media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure in those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure [20]. Thus, according to DRT, a talker gestures (e.g., the closing and opening of the lips during the production of /pa/) structure the acoustic signal, which then serves as the informational medium for the listener to recover the gestures. The term direct in direct realism is meant to imply that perception is not mediated by processes of inference or hypothesis testing, rather, the information in the acoustic signal is assumed to be rich enough to specify (i.e., determine uniquely) the gestures that structure the signal. To perceive the gestures, it is sufficient for the listener simply to detect the relevant information. The term realism is intended to mean that perceivers recover actual (physical) properties of their niche, including, in the case of speech perception, phonetic segments that are realized as sets of physical gestures. This realist perspective contrasts with a mentalistic view that phonetic segments are internally generated, the creature of some kind of perceptual-cognitive process [20].

2.4.4 Cohort Theory

The original Cohort theory was proposed by Marslen-Wilson and Tyler in 1980 [21]. According to Cohort theory, various language sources like lexical, syntactic,

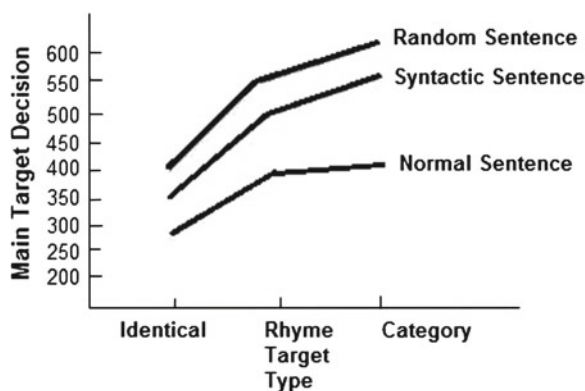


Fig. 2.10 Detection time in word sentence presented in sentences in the Cohort theory (Marslen-Wilson and Tyler 1980)

semantic, etc., interact with each other in complex ways to produce an efficient analysis of spoken language. It suggests that input in the form of a spoken word activates a set of similar items in the memory, which is called word initial cohort. The word initial cohort consisted of all words known to the listener that begin with the initial segment or segments of the input words. For example, the word elephant may activate in the word initial cohort members echo, enemy, elder, elevator, and so on. The processing of the words started with both bottom-up and top-down information and continuously eliminated those words if they are found to be not matching with the presented word. Finally, a single word remains from the word initial cohort. This is the ‘recognition point’ of the speech sound [14, 21].

Marslen-Wilson and Tyler performed a word-monitoring task, in which participants had to identify prespecified target words presented in spoken sentence as rapidly as possible. There were normal sentences, syntactic sentences, that is, grammatically correct but meaningless and random sentences, that is, unrelated words. The target was a member of given category, a word that rhymed with a given word, or a word identical to a given word. Figure 2.10 shows the graph of detection time versus target. According to the theory, sensory information from the target word and contextual information from the rest of the sentence is both used at the same time [21].

In 1994, Marslen-Wilson and Warren revised the Cohort theory. In the original version, words were either in or out of the word cohort. But in the revised version, words candidate vary in the activation level, and so membership of the word initial cohort plays an important role in the recognition. They suggested that word initial cohort contains words having similar initial phoneme [21].

The latest version of the cohort is the distributed Cohort model proposed by Gaskell and Marslen-Wilson in 1997, where the model has been rendered in a connectionist architecture as shown in Fig. 2.11. This model differs from the others in the fact that it does not have lexical representations. Rather, a hidden layer mediates between the phonetic representation on the one side and phonological and semantic

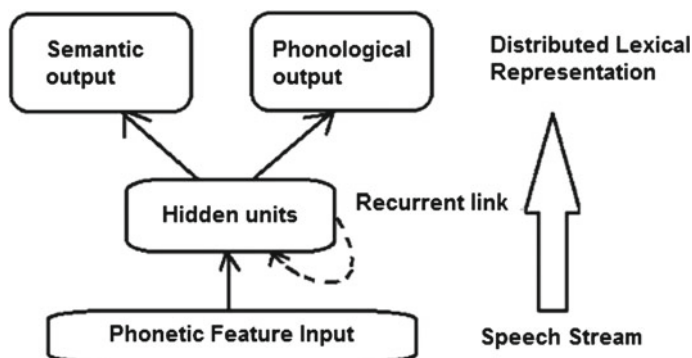


Fig. 2.11 The connectionist distributed Cohort model of Gaskell and Marslen-Wilson (1997)

representations on the other side. The representations at each level are distributed. In Cohort, an ambiguous phonetic input will yield interference at the semantic and phonological levels. Input to the recurrent network is passed sequentially, so given the beginning of a word, the output layers will have activation over the various lexical possibilities, which as the word progresses toward its uniqueness point, will subside [22].

In this model, lexical units are points in a multidimensional space, represented by vectors of phonological and semantic output nodes. The phonological nodes contain information about the phonemes in a word, whereas the semantic nodes contain information about the meaning of the words. The speech input maps directly and continuously onto this lexical knowledge. As more bottom-up information comes in, the network moves toward a point in lexical space corresponding to the presented word. Activation of a word candidate is thus inversely related to the distance between the output of the network and the word representation in lexical space. A constraining sentence context functions as a bias: The network shifts through the lexical space in the direction of the lexical hypotheses that fit the context. However, there is little advantage of a contextually appropriate word over its competitors early on in the processing of a word. Only later, when a small number of candidates still fits the sensory input, context starts to affect the activation levels of the remaining candidates more significantly. There is also a mechanism of bottom-up inhibition, which means that in case the incoming sensory information no longer fits that of the candidate, the effects of the sentence context are overridden [14].

2.4.5 Trace Model

The TRACE model of spoken word recognition, proposed by McClelland and Elman, in 1986, is an interactive activation network model. According to this, model speech perception can be described as a process, in which speech units are arranged into

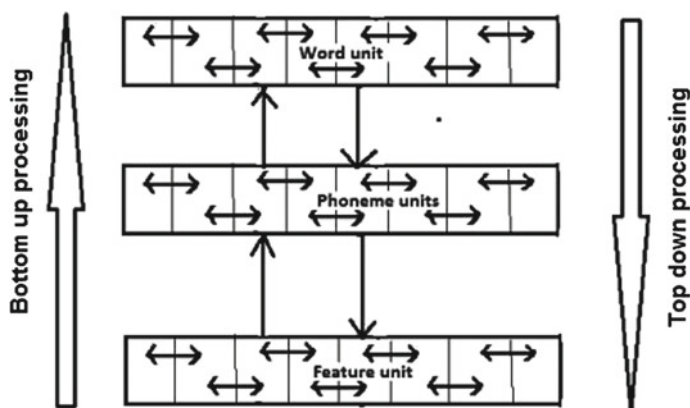


Fig. 2.12 Basic Trace model

levels which interact with each other. There are three levels: features, phonemes, and words. There are individual processing units or nodes at three different levels. For example, within the feature level, there are individual nodes that detect voicing. Feature nodes are connected to phoneme nodes, and phoneme nodes are connected to word nodes. Connection between levels operate in both directions. The processing units or nodes share excitatory connection between levels and share inhibitory links within levels [14, 23]. For example, to perceive a /k/ in ‘cake,’ the /k/ phoneme and corresponding featural units share excitatory connections. Again, /k/ would have an inhibitory connection with the vowel sound in ‘cake’ /eI/.

Each processing units corresponds to the possible occurrence of a particular linguistic unit (feature, phoneme, or word) at a particular time within a spoken input. Activation of a processing unit reflects the state of combined evidence within the system for the presence of that linguistic unit. When input is presented at the feature layer, it is propagated to the phoneme layer and then to the lexical layer. Processing proceeds incrementally with between-layer excitation and within-layer competitive inhibition. Excitation flow is bidirectional, that is, both bottom-up and top-down processing interact during perception. Bottom-up activation proceeds upward from the feature level to the phoneme and phoneme level to the word level. Whereas top-down activation proceeds in the opposite from the word level to the phoneme level and phoneme level to the feature level as shown in Fig. 2.12. As excitation and inhibition spread among nodes, a pattern of activation develops. The word that is recognized is determined by the activation level of the possible candidate words [23].

Trace is a model of accessing the lexicon with interlevel feedback. Feedback in the Trace model improves segmental identification under imperfect hearing conditions. In interactive activation networks, such as Trace, each word or lexical candidate is represented by a single node, which is assigned an activation value. The activation of a lexical node increases as the node receives more perceptual input and decreases when subject to inhibition from other words [21].

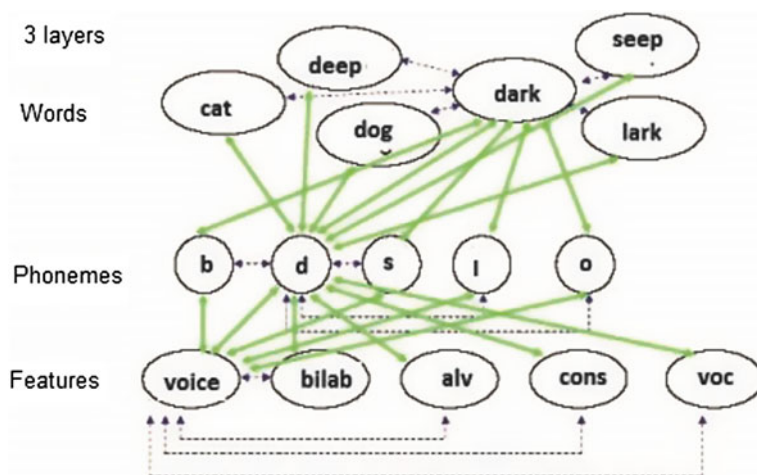


Fig. 2.13 Interaction in Trace model

In Trace model, activations are passed between levels, and thus, it confirms the evidence of a candidate to a given feature, phoneme, or words. For example, if the voicing speech is presented, that will activate the voiced feature at the lowest level of the model, which in turn pass its activation to all voiced phoneme in the next level of phonemes, and this activates the words containing those phonemes in the word level. The important fact is that through lateral inhibition among units, one candidate in a particular level may activate some similar units, in order to help the best candidate word to win the recognition. Such as, the word 'cat' at the lexical level will send inhibitory information to the word 'pat' in the lexical unit. Thus, the output may be one or more suggested candidate. Figure 2.13 shows the interaction in a trace model through inhibitory and excitatory connection. Each unit connects directly with every unit in its own layer, and in the adjacent layers. Connections between layer are excitatory that activates units in other layers. Connections within a layer are inhibitory that activates units in same layer.

2.4.6 Shortlist Model

Dennis Norris, in 1994, have proposed the Shortlist model, which is a connectionist model of spoken word recognition. According to Norris, a Shortlist of word candidates is derived at the first stage of the model. The list consists of lexical items that match the bottom-up speech input. This abbreviated list of lexical items enters into a network of word units in the later stage, where lexical units compete with one another via lateral inhibitory links [14]. Unlike trace flow of information in the Shortlist model is unidirectional and only bottom-up processing involves. The superiority

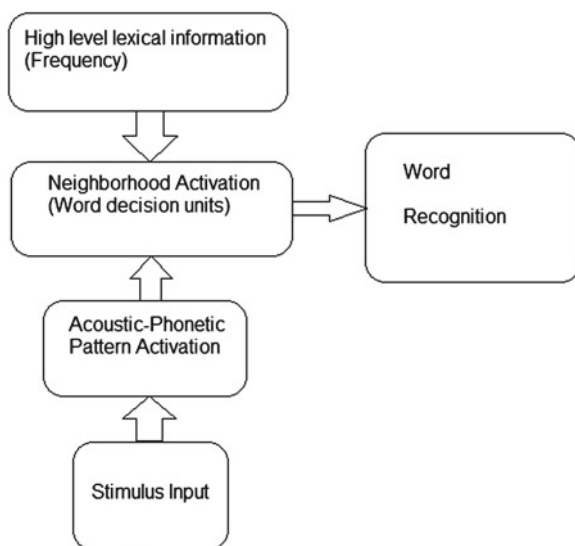


Fig. 2.14 Neighborhood activation model

of Shortlist comes from the fact that it provides an account for lexical segmentation in fluent speech.

2.4.7 Neighborhood Activation Model

According to the Neighborhood Activation Model, proposed by Luce, Pisoni and Goldinger in 1990, and revised by Luce and Pisoni in 1998, spoken word recognition is a special kind of pattern recognition task performed by human. Some similar acoustic-phonetic patterns are activated in the memory by the stimulus speech input. Once the patterns are activated, the word decision units interact with each of the patterns and tries to decide which pattern best matches the input pattern. The level of activation depends on the similarity of the pattern to the input speech. The more similar the pattern, the higher is its activation level [14, 24]. The word decision unit performs some probabilistic computation to measure the similarity of the patterns. It considers following facts:

1. The frequency of the word to which the pattern corresponds.
2. The activation level of the pattern, which depends on the patterns match to the input.
3. The activation levels and frequencies of all other words activated in the system.

Finally, the word decision unit decides the winning pattern which have the highest probability, and this is the point of recognition. Figure 2.14 shows the diagrammatic view of neighborhood activation model.

2.5 Conclusion

This chapter initially provides a fundamental description of generalized speech recognition systems and then provides a brief description of the speech communication chain and the flow of the various cognitive, linguistic, and physiological events involved in the process of the speech production and perception. Next, the speech perception mechanisms starting from the human auditory system and hearing mechanisms are described. Finally, spoken word recognition theories and model like Motor theory, Analysis-by-Synthesis model, DRT theory, Cohort theory, Trace model, Shortlist model, and Neighborhood activation model are described in a nutshell.

References

1. Automatic Speech Recognition. Available via www.kornai.com/MatLing/asr.doc
2. Speech Recognition. Available via www.learnartificialneuralnetworks.com/speechrecognition.html
3. Discrete Speech Recognition Systems. Available via <http://www.gloccal.com/discrete-speech-recognition-systems.html>
4. Types of Speech Recognition. Available via <http://www.lumenvox.com/resources/tips/types-of-speech-recognition.aspx>
5. Huang XD (1991) A study on speaker-adaptive speech recognition. In: Proceedings of the workshop on speech and natural language, pp 278–283
6. Rabiner LR, Schafer RW (2009) Digital processing of speech signals. Pearson Education, Dorling Kindersley (India) Pvt Ltd, Delhi
7. Narayan CR (2006) Acoustic-perceptual salience and developmental speech perception. Doctor of Philosophy Dissertation, University of Michigan
8. Speech Therapy-Information and Resources. Available via <http://www.Speech-therapy-information-and-resources.com/communication-model.html>
9. Levelt WJM (1989) Speaking, from intention to articulation. MIT Press, England
10. Mannell R (2013) Speech perception background and some classic theories. Department of Linguistics, Faculty of Human Sciences, Macquarie university, Available via http://clas.mq.edu.au/perception/speechperception/classic_models.html
11. Elsea P (1996) Hearing and perception. Available via http://artsites.ucsc.edu/EMS/Music/tech_background/TE-03/teces_03.html
12. The Hearing Mechanism. www.healthtree.com. 20 July 2010
13. Ostry DJ, Gribble PL, Gracco VL (1996) Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? *J Neurosci* 16:1570–1579
14. Jusczyk PW, Luce PA (2002) Speech perception and spoken word recognition: past and present. *Ear Hear* 23(1):2–40
15. Pallier C (1997) Phonemes and syllables in speech perception: size of attentional focus in French. In: Proceedings of EUROSPEECH, ISCA

16. Klatt DH (1976) Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J Acoust Soc Am* 59(5):1208–1221
17. Nygaard LC, Pisoni DB (1995) Speech perception: new directions in research and theory. *Handbook of perception and cognition: speech, language, and communication*. Elsevier, Academic Press, Amsterdam, pp 6396
18. Anderson S, Shirey E, Sosnovsky S (2010) Department of Information Science. University of Pittsburgh, Speech Perception, Downloaded on Sept 2010
19. Honda M (2003) Human speech production mechanisms. NTT CS Laboratories. *Selected Papers* (2):1
20. Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. *Annu Rev Psychol* 55:149–179
21. Eysenck MW (2004) *Psychology: an international perspective*. Psychology Press
22. Bergen B (2006) Speech perception. *Linguistics* 431/631: connectionist language modeling. Available via <http://www2.hawaii.edu/bergen/ling631/lecs/lec10.html>
23. McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? *Trends Cogn Sci* 10:8
24. Theories of speech perception, Handout 15, Phonetics Scarborough (2005) Available via <http://www.stanford.edu/class/linguist205/indexfiles/Handout%2015%20%20Theories%20of%20Speech%20Perception.pdf>

Phoneme-Based Speech Segmentation using Hybrid
Soft Computing Framework

Sarma, M.; Sarma, K.K.

2014, XXI, 187 p. 62 illus. in color., Hardcover

ISBN: 978-81-322-1861-6