

Chapter 1

Basis for Sound-Based Assistive Technology

Abstract In this chapter, the author discusses the methodology as well as the basic concepts of assistive technology based on cybernetics that regards the human body as a feedback system comprising the senses, the brain, and the motor functions. In the latter part of this chapter, the author explains some mechanisms of the auditory sense and speech production of human beings that are needed to understand the contents of the book. The author introduces a national project called “The Creation of Sciences, Technologies and Systems to Enrich the Lives of the Super-aged Society” that he has promoted since 2010.

1.1 Background of Assistive Technology

In the field of medical engineering, medical doctors and engineers have eagerly combined their research efforts with the aim of developing “life support technology” such as artificial organs and regeneration medicine in order to prolong life. However, in the case of people suffering from disorders that cannot be cured by modern medical treatment, there is hope in particular that their disabilities may be compensated for by making use of cutting-edge technology. “Bodily support technology” such as artificial sensors and artificial limbs is a typical example for the compensation of the disabilities. Furthermore, the support technologies will give them the ability to take part in numerous “social participation and activities” for the sake of QOL improvement and job assistance system [1, 2].

Figure 1.1 shows a comparison between the roles of assistive technology and medical engineering, and also shows examples of assistive tools that should be studied and developed [3]. However, the disabilities experienced by people are so diverse and complex that it has been difficult to construct a research approach to develop this assistive technology. At this point, the author would like to explain his personal view regarding the methodology for conducting research in assistive technology. The concept of assistive technology originated from “Cybernetics,” published by Norbert Wiener (1894–1964) in 1948 [4]. As indicated in the subtitle “Or Control and Communication in the Animal and the Machine,” he described the

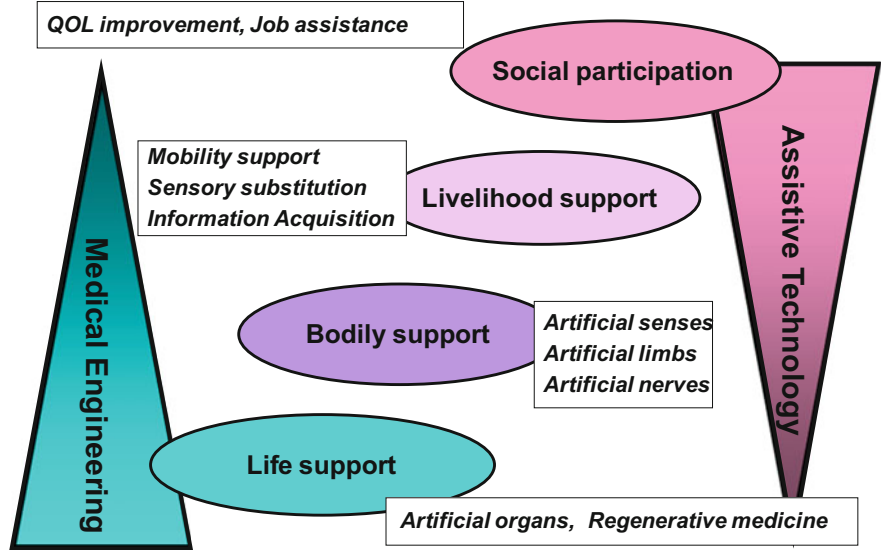


Fig. 1.1 Medical engineering and assistive technology

human functions of senses, brain and limbs—which are essential for living an intellectual life—as having similarities to the functions of the sensors, computers and actuators of automatic machines such as robots. Furthermore, he pointed out that “homeostasis” in both human beings and machines is maintained by feedback control systems. Thus, it can be said that assistive technology research lies in the undercurrent of cybernetics.

As shown in Fig. 1.2, the first point of departure in our research is to investigate and compare the mechanisms of the sensory systems, the brain functions, and motor ability in both healthy and disabled persons based on physiology and

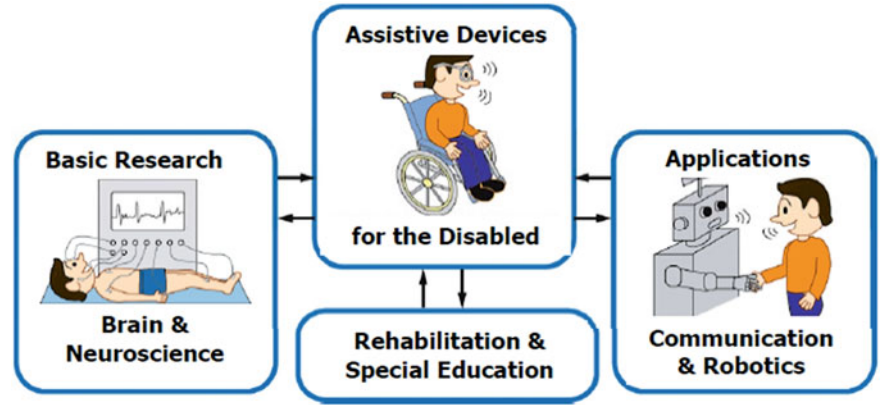


Fig. 1.2 Three stages of assistive technology research as presented in this study

psychophysics, as shown in the left side of the figure. After obtaining knowledge through this fundamental research, the second stage is to design tools that will enable substitution or compensation of a person's senses or limbs, as shown in the center of the figure. If the artificial senses or limbs created for the disabled function well, they can serve as excellent replacements for their original receptors or limbs. Conversely, if the devices are defective or inadequate in some way, it is necessary to come full circle by returning to the original starting point of the first stage. This methodology can thus be characterized by conducting continuous cycles of research that eventually lead to the desired goal.

Furthermore, the artificial senses or limbs thus created can be applied to computer pattern recognition as well as to robot actuators, as shown in the right side of the figure. This final stage could well result in healthy competition among manufacturers that would, in turn, lead to further improvements in this technology, creating a big market and making the price of the tools affordable. The author himself has realized significant progress by utilizing this approach in developing assistive technology over a period of more than 45 years.

On the other hand, as a highly advanced IT society is rapidly approaching, people have become surrounded by all sorts of information tools. As a consequence, more and more people are experiencing distress in their attempts to handle the complexity of both computers and the internet. In particular, a phenomena known as “technostress” is occurring on a wide scale, especially among the less technically inclined. This digital divide resulting from a knowledge gap is, in fact, becoming a major social issue. As information tools are especially difficult for the elderly and disabled to use, machines that can be operated with little or no conscious effort are in high demand. The role of assistive technology is therefore to achieve a state of “universal accessibility” that will empower anyone in society to easily reap the full benefits of the highly advanced IT.

1.1.1 Towards a Super-Aged Society

Currently, more than 25% of the Japanese population are over the age of 65, representing nearly 30 million people, as shown in the left side of Fig. 1.3. As the elderly population increases, so does the number of disabled people. In fact, the ratio of the number of elderly people to the number of disabled people increased from about 31% in 1970 to about 70% in 2011 in Japan, as shown in the right side of the figure [5]. Almost the same situation as that currently prevailing in Japan will extend all over the world, especially in European and East Asian countries. These disabled elderly people all have some form of disability in terms of hearing, speaking, reading, thinking, or moving.

Assistive technology for elderly disabled people is referred to as “geron-technology,” which is distinct from “barrier-free design” technology that includes supporting young disabled people. As shown in Fig. 1.4, the barrier-free design for the young uses the “plasticity” of the human body, especially the brain,

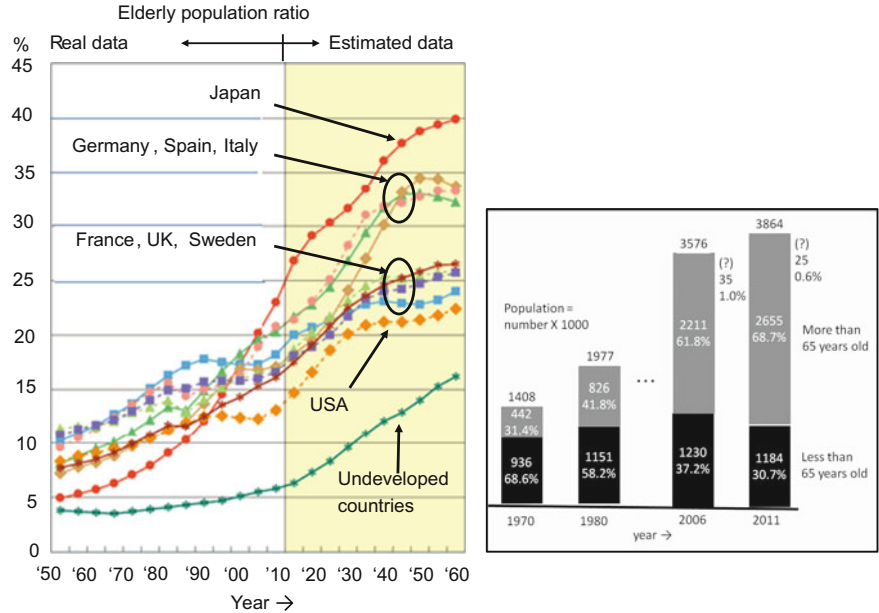


Fig. 1.3 Left Elderly population ratio [5], right elderly disabled/total disabled in Japan

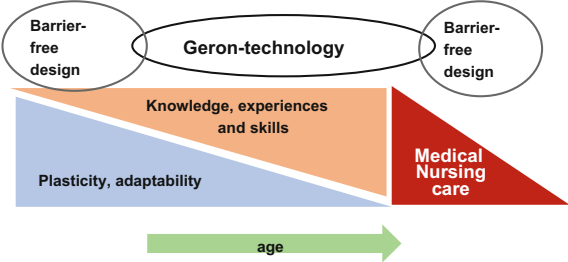


Fig. 1.4 Barrier-free design and geron-technology

because with the help of brain plasticity, residual functions work to compensate for other disordered functions. However, in general, this plasticity function decreases in the elderly, so that they will need to acquire abilities using their “experience” [6]. Overcoming this aging of society is already becoming an increasing burden for the Japanese government.

As a result, many projects have been launched with the Ministry of Economy, Trade and Industry and the Ministry of Labor, Health and Welfare leading the way. The author was asked to promote a new project called “The Creation of Sciences, Technologies and Systems to Enrich Society for the Aged in Japan,” which is one of the strategic innovation projects for a super-aged society organized by the Japan Organization for Japan Science and Technology (JST) in 2010. The project started in 2010 and will continue until 2019. It should be possible to apply almost the same

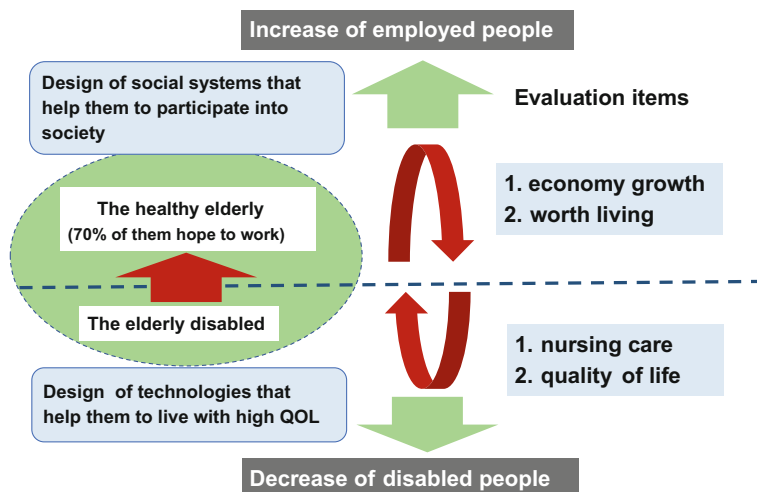


Fig. 1.5 Purpose and evaluation items of national project to enrich both healthy elderly and elderly disabled people

design approach for assistive tools for the elderly disabled as for the young disabled. A secondary purpose of the project is to create new assistive systems that will promote the social participation of healthy elderly persons as well as improve the quality of life (QOL) of elderly disabled persons.

Based on investigations into the functional ability of the elderly, the author divided elderly people into two groups: a healthy elderly group, and a disabled elderly group. In the project, for the healthy elderly, we strongly promote the design of assistive tools and systems so that the healthy elderly can participate in society. For the disabled elderly, we promote the design of rehabilitation technologies and nursing care networks so that the disabled elderly can live with a high QOL. These assistive technologies and systems will decrease social security expenses and may also increase employment and promote economic growth (Fig. 1.5).

Furthermore, five themes were set up concerning information communication technology (ICT) and information robotics technology (IRT) applications in order to realize the aim of our project, as shown in Fig. 1.6. These are:

- “Wearable ICT” to assist sensory functions and to detect the physical state for both groups,
- “Infra-structured ICT” to design job-matching systems for employment of the healthy elderly,
- “Mobility-assistive IRT” to enable free and safe movement within the community as well as at home for both groups,
- “Labor-assistive IRT” to help with heavy work and nursing care in daily life for both groups,
- “Brain-assistive ICT and IRT” to assist understanding, memory, and information production mainly for the disabled elderly.

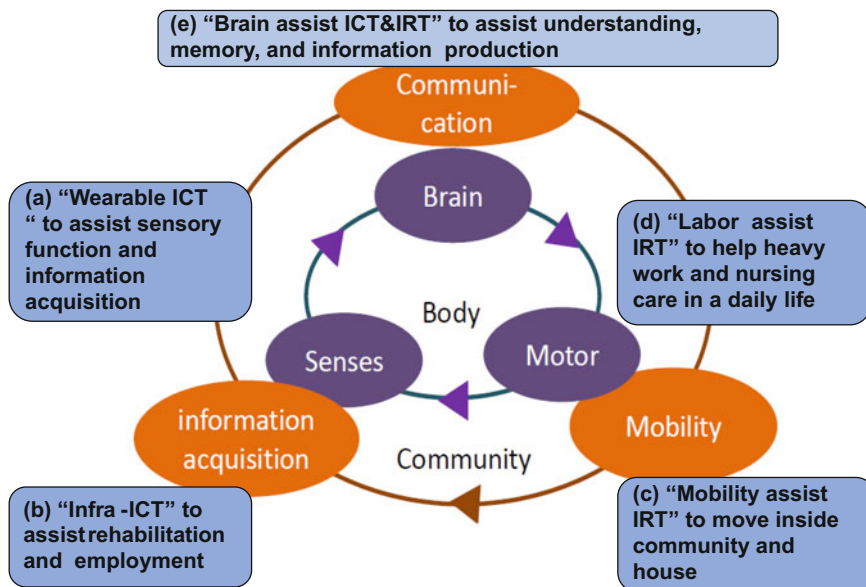


Fig. 1.6 Five themes to assist elderly using ICT and IRT

In the following chapters, the author would like to mention how sounds play an important role in the design of assistive tools in order to support the elderly as well as young people with hearing, speaking and seeing disabilities.

The project is advanced under three stages as shown in Fig. 1.7. The first stage (2011–2013) was to analyze the cognitive and/or behavior functions of the elderly and to investigate needs of the aged society as well as the elderly themselves. The second stage (2014–2016) was to design assistive tools for the elderly and to create systems needed in the aged society. Third stage (2017–2019) is to apply the tools and the systems in a real society and to evaluate the project from a viewpoint of social security, economy growth and QOL of the elderly. These three stages correspond “science”, “technology” and “systems” as shown in the figure. The third stage becomes very important to open a big market for mobile phones, automatic driving cars, virtual reality and care robotics that are capable of supporting the elderly as well as the disabled. A project “Labor-assistive IRT” carried out by a joint research of Mitsubishi Electric Engineering (Co., LTD) and Hokkaido University completed in 2014. The author will briefly introduce two projects related “Mobility-assistive IRT” and “Infra-structured ICT”. A project related to “Brain-assistive ICT and IRT” will be mentioned in Chap. 5.

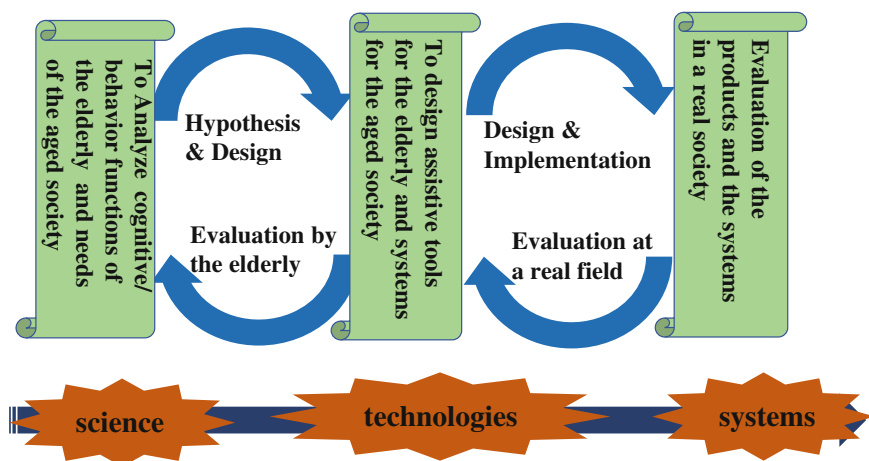


Fig. 1.7 Three stages of a project “The Creation of Sciences, Technologies and Systems to Enrich Society for the Aged in Japan”

1.1.1.1 Autonomous Driving Intelligent Systems to Assist Elderly Drivers

The research project titled “Autonomous driving intelligent systems to assist elderly drivers” was adopted as the mobility-assistive IRT, which has been carried out by a joint research of TOYOTA (Co., LTD) and Tokyo University of Agriculture and Technology. The project has been conducted by Inoue from TOYOTA and his co-researchers. It is reported that traffic accidents dramatically increase in the case of the elderly drivers, as shown in the left of Fig. 1.8. Major accidents are classified as collision accidents between cars, collisions to pedestrians and bicycles, deviation from road lanes and encounter accidents at crossroads as shown in the right of Fig. 1.8. The purpose of the research project is to construct the autonomous driving intelligent system that helps the elderly drivers to avoid these accidents by using an expert driver’s model. The expert driver’s model works only when there is a risk that an accident will occur. The model was constructed using big data of drive recorders that more than 60,000 professional drivers took just before near misses of car accidents occurred. The risks are predicted based on two kinds of information [7, 8].

As shown in Fig. 1.9, one of them is surrounding information around the car detected by various sensors such as a laser ranging device, several cameras and etc. The other geographic-traffic information obtained by map data, weather reports, GPS and etc. When the probability of the predicted risk is greater than the threshold, the expert driver’s model controls a break and a steering wheel instead of the elderly drivers.

They have been also investigating acceptability of the aged society as well as the elderly themselves in order to make the regulations to accept the new technologies.

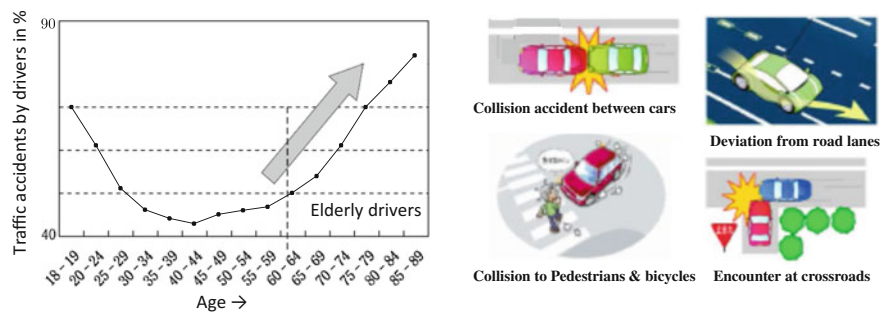


Fig. 1.8 Left Traffic accidents versus age of drivers, right four major traffic accidents

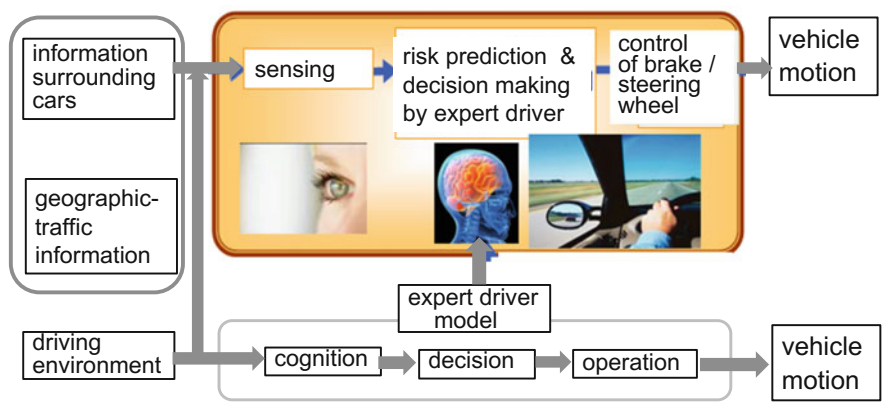


Fig. 1.9 Autonomous driving intelligent system consisting of information detected sensors, geographic-traffic information and expert driver’s model

It is expected that the autonomous driving intelligent systems may lead to expand their actions range and prompt social participation.

1.1.1.2 “Senior Cloud” Using Knowledge, Experiences and Skills of the Elderly

The research project titled “Senior cloud using knowledge, experiences and skills of the elderly” was adopted as the infra-ICT, which has been carried out by a joint research of the university of Tokyo and Japan IBM. The project has been conducted by Hirose at the university of Tokyo and his co-researchers. As shown in the left of Fig. 1.10, recently diversify of working style of the elderly has rapidly increased. For example, their working time has shifted from full-time to any-time, and their working place is from wide area to their home. Furthermore, their acquired knowledge, experiences and skills also have diversified [9, 10].

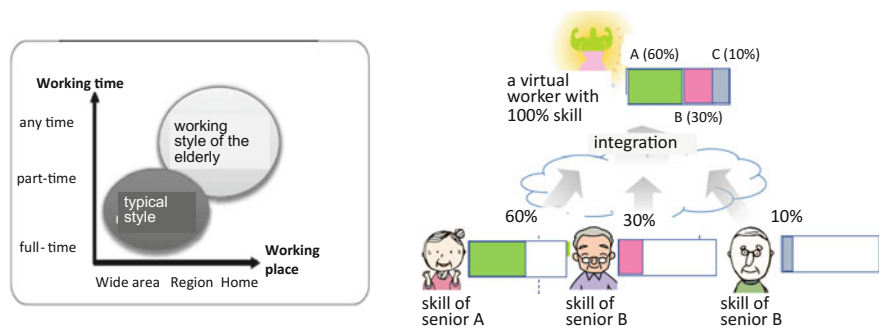


Fig. 1.10 *Left* Working styles of the elderly, *right* an example of mosaic model (skill mosaic)

The purpose of the research project is to promote working and participations in society by constructing “MOSAIC model” and job-matching platform that positively utilize the diversity and integrate their acquired abilities. The concept of the mosaic model is shown in the right of Fig. 1.10. For example, if a senior A has 80% of skills of averaged labor in the working field X, a senior B has 10%, a senior C has 10%, then one virtual worker with 100% skill is expected by integrating these three persons’ skills. As well as the skill mosaic model, it can be constructed that a working time mosaic model, a working place mosaic model, an experience mosaic model and so on.

Two interfaces and one platform were designed in order to realize the concept of the senior cloud; knowledge acquisition interface, knowledge structuring platform and knowledge transfer interface as show in Fig. 1.11. The knowledge acquisition interface includes “Question first method” that may elicit the various ability and characteristics from interviews and communication to them. As a result, the interface can estimate who, when, what, where, how the elderly can do. In the knowledge structuring platform, various mosaic models are constructed based on the data obtained by the knowledge acquisition interface and job matching for the elderly are performed. In the knowledge transfer interface, ICT and IRT such as virtual reality technologies and tele-existence robots are used so that the elderly may transfer their knowledges, experiences and skills to the people who work together in the desired location and favorite time. It is expected that the senior cloud systems may lead to sustain their purpose in life and also maintain their healthy as well as increasing their QOL.

1.1.1.3 Future Prediction Expected from Assistive Technologies

Furthermore, our research approach based assistive technology using the ICT and IRT may contribute to increase worker population, to decrease social security expenses and to open new markets. As population rate of the elderly is increasing in the other countries, it is expected that the assistive technology as well as the

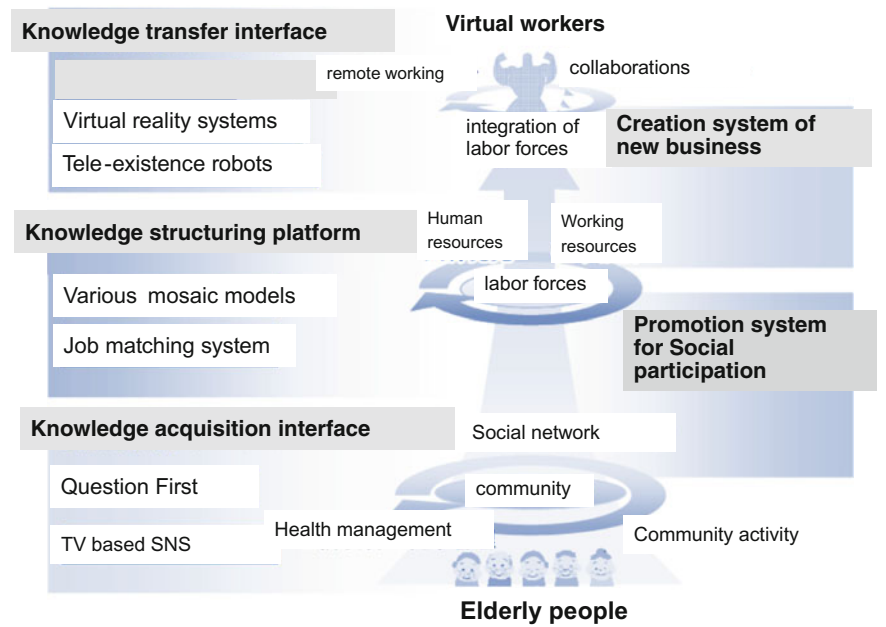
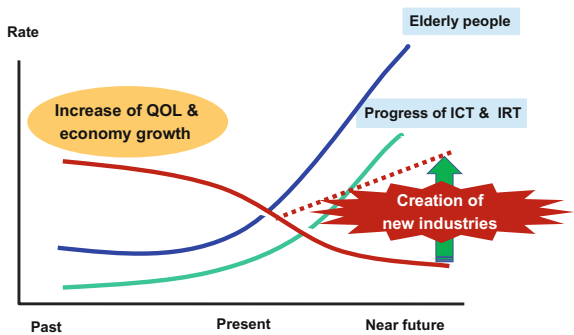


Fig. 1.11 A concept of “Senior cloud” consisting of knowledge acquisition, knowledge structure and knowledge transfer

geron-technology will take an important role for almost all countries. As a result, our national project may contribute to create new export industries and also to increase the economy growth in the near future as shown in Fig. 1.12.

Under this background, the assistive technology to support hearing, speaking and seeing will occupy an increasingly important position, especially for prompting a social participation of the elderly as well as the disabled. The author would like to focus on the sound-based science and technology because they are essential for communication and information acquisition in highly-sophisticated information society.

Fig. 1.12 Assistive technology for the elderly as well as the disabled promotes both economy growth and increase of QOL, making new industries



1.2 Roles of Sound in Assistive Technology

Sound became a vehicle to transmit information over a great distance ever since the Earth began to have an atmosphere. In particular, while both animals and humans have been using sound for communication, the invention of speech has enabled mankind to create civilizations and cultures. In this regard, it can be said that speech serves as a social tool for the transmission of civilizations and cultures. In fact, although some tribes do not possess any written language, there is no civilization that exists that lacks a spoken language.

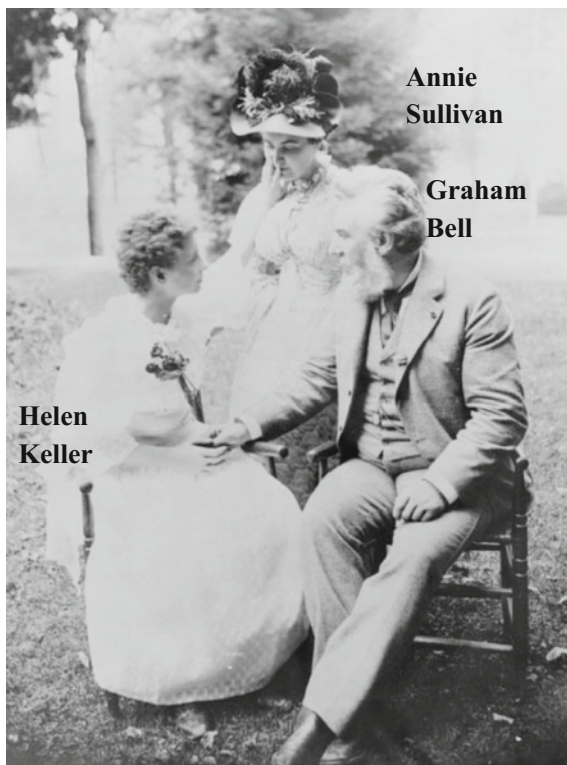
On a different note, when speech is considered on an individual basis, it could be categorized as a tool for thought. Specifically, the acquisition of speech has enabled humanity to make full use of word combinations that give rise to abstract thought, which, in turn, leads to creativity. Such silent speech, which takes place in the head, is known as “internal speech”. Thanks to it, the cerebral cortex of human beings has acquired increasingly complex capabilities. It gives one pause to consider that this extraordinary process arose out of the mere existence of air and its ability to transmit information through the vehicle of sound.

The well-known Helen Keller (1880–1968) (see Photograph 1.1) was once asked whether it was a greater disability to be blind or deaf. She answered: “To be deaf is to lose the tool of sound—a tool that greatly enhances one’s ability to be socially active as well as to think. In that sense, being deaf is more inconvenient than being blind.” Moreover, as speech, derived from sound, provides the building blocks that lead to the creation of abstract thought in conjunction with cerebral development, speech concepts could still be formed without the use of sight. In contrast, for those with congenital deafness, the lack of speech will undoubtedly retard development of the cerebrum. In severe cases, this could result in little or no cerebral development beyond a mental age of around nine.

It goes without saying that the transmission of speech is an important object of study in the field of engineering. In this regard, when two or more people are in close proximity, communication can be accomplished through the vehicle of air; however, when separated by a great distance, other media become necessary if the message is to be received. This was the essential reason the telephone was invented. In particular, it is worth noting that Alexander Graham Bell (1847–1922), the inventor of the telephone, was a linguist whose wife was hearing impaired. His invention was therefore motivated by a fervent desire to communicate with his wife by any means possible. With this in mind, it is no exaggeration to say that an extraordinary effort by one man to assist his wife is what eventually led to the development of today’s enormous telephone system.

When information is conveyed over a great distance, as in the case with telephones, radios and television, one must consider both the transmission of sound as well as its reproduction on the receiving end. Specifically, in order to understand how transmission takes place, matters such as information processing, compression, and restoration must be analyzed from an engineering perspective.

Photograph 1.1 Alexander Graham Bell with Helen Keller and Annie Sullivan at the meeting of the American Association to Promote the Teaching of Speech to the Deaf, July 1894, in Chautauqua, N.Y.



Another application of engineering is technology allowing communication between human beings and machines. There was an article in a newspaper written in 1901 [11] that attempted to predict future technology over the next century. In particular, the authors foresaw great technological progress in the field of communication. However, they believed that technology would eventually make it possible for humans to communicate with animals. Despite this prophecy's obvious inaccuracy, it was not entirely off the mark if one were to replace the word "animals" with "machines." Devices that can understand as well as respond to human speech are being applied to automatic recognition and sound synthesis. If the quality of such devices is sufficiently improved, they could provide adequate substitutes for the hearing and speech impaired.

Although speech synthesizers have been successfully produced, sound-recognition systems require further research. More specifically, a more in-depth investigation must be conducted regarding neurological functions that pertain to the understanding and production of speech. Due to the lack of such neurological data, it is difficult to determine the best way to equip computers with speech capability. As various biomedical measurements have recently become available for brain analysis, it might soon be possible to finally understand the brain—by far the most mysterious part of human beings. Furthermore, as our

understanding deepens, we will be in a better position to assist the hearing and speech impaired. In any case, no real progress in sound-recognition technology will be possible without such a clearer understanding of the brain.

This chapter will first outline the process of how sound is received and how this information is conveyed to the brain through the auditory nervous system. In doing so, some consideration will also be made as to why speech disorders occur.

1.3 Structure and Functions of the Auditory System

1.3.1 External Ears

Figure 1.13 shows the structure of the hearing organs from the auricle to the internal ear. The external auditory canal is a winding tube 10 mm in diameter and 35 mm in length with a tympanic membrane obliquely placed at the bottom. Acoustically speaking, the external auditory canal serves as a resonance tube and has the effect of increasing the decibel level from 10 to 20 dB for sound in the range of 2000–4000 Hz, as illustrated in Fig. 1.14.

However, the external ear, including the auricle, does not merely collect and filter sound. For example, when listened to through headphones, sound is heard as if it were originating from inside the head; whereas, in the case of the external ear, the feeling is that the sound is coming from outside. Another important function of the auricle is to provide the listener with a sense of whether the sound originates from above or below. As will be discussed in Chap. 7, when the auricle is covered with a substance such as silicone, a sound coming from below the listener's ear is actually heard as if coming from above.

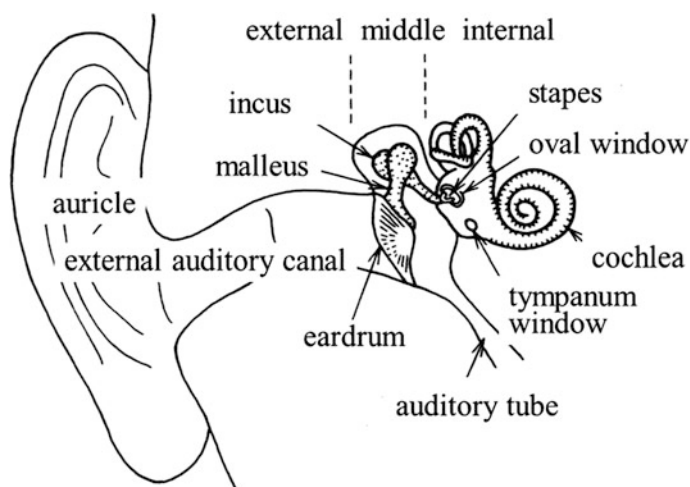
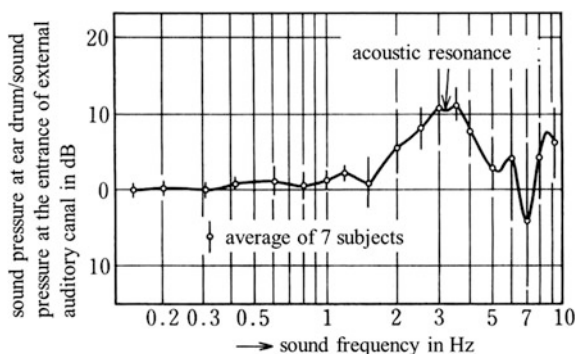


Fig. 1.13 Schematic representation of auditory organs

Fig. 1.14 Acoustic frequency characteristics of external auditory canal



When listening through headphones, a sound recorded through a microphone that has been placed on a tympanic membrane is experienced as originating from outside of the head—a phenomenon known as external sound localization. More specifically, this artificially produced sound is made possible by first using a computer to calculate the head-related transfer function (HRTF) between the origin of the outside sound and the tympanic membrane. In the field of virtual reality (VR)—also called artificial reality—NASA has succeeded in creating external sound localization that offers a very realistic presence by calculating the HRTF using a digital signal processor in real time [12].

A further function of external ears is illustrated in the case of people who are completely blind, most of whom make use of their external ears to develop what is known as an “obstacle sense”—the ability to recognize obstacles by depending exclusively on sound.

1.3.2 Middle Ear

It is said that the origin of the hearing receptors of mammals was a sensor in fish found along its sides, called a “lateral organ”. It is a sensor that measures the speed of its own movement and also detects the speed of the seawater when a fish is stationary. Under magnification in a microscope, it is found that the lateral organ is lined with many “hair cell receptors” (see Fig. 1.20a). As will be explained later, these hair cells play an important role as auditory receptors.

As fish evolved into mammals and started living on land, the middle ear developed in such a way that the highly sensitive hair cells could be used to detect air movements. When a sound wave is emitted toward seawater, only about 0.1% of the sound’s energy enters the water from the air; the rest is reflected when striking the surface of the sea. It is due to this phenomenon that hair cells in the lymph fluid cannot be used efficiently. Therefore, in order for a sound wave to enter the lymph fluid effectively, acoustic impedance matching by the middle ear became necessary.

Gradually, through evolution, the large tympanic membrane became capable of receiving sound waves. Once this occurred, it was possible for this membrane to then transmit the vibration to a small bone called “the hammer” that is one of auditory ossicle, finally reaching another small bone known as “the stapes” that is located on the entrance of the cochlea called oval window (see Fig. 1.15). The ratio of the area of the tympanic membrane to that of the oval window is about 20 to 1. This great difference in area results in the compression of the sound energy received at the bottom of the stapes bone. This compression increases the sound pressure per unit area by about 26 dB. The auditory ossicle muscle also contributes to the acoustic impedance by functioning like a kind of seesaw. Therefore, it is the middle ear that converts the acoustic impedance.

Furthermore, the auditory ossicle muscle functions to prevent strong sounds from entering the internal ear by lowering the sensitivity of the middle ear through feedback. As this muscle is also quite delicate, certain problems such as middle ear inflammation can easily occur. Therefore, as will be explained in a later chapter, artificial middle ears have been developed to substitute for the functions of natural middle ears.

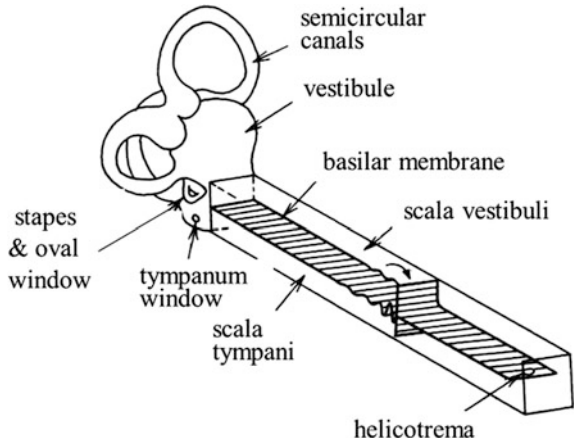
By making a rough estimate of the amplitude of the vibration of the tympanic and basilar membranes from the auditory threshold, it is found that the amplitude of the vibration of the tympanic membrane is roughly equal to the diameter of a hydrogen atom. In the case of the basilar membrane, the amplitude of the vibration is roughly the size of the nucleus of a hydrogen atom. 16,000 hair cells aligned in four rows at the basilar membrane detect subtle vibrations. If mammals had evolved on land from the beginning, they would have bypassed the rather roundabout, and largely impractical, evolutionary process just described. Instead, they would have quickly developed sharp sensors to detect air vibrations, much like microphones do today. Therefore, it can be said that the rather high incidence of hearing disorders in mammals arises in part as a result of a cumbersome evolutionary process.

1.3.3 Mechanism of Frequency Analysis in the Internal Ear

After mammals started living on land, hair cells were aligned on a thin “basilar membrane” inside a compact, snail-shaped cochlea. The basilar membrane is covered with lymph fluid that contains elements similar to seawater. The reason for the cochlea’s snail-like formation was to enable it to wind itself roughly $2^{3/4}$ times into a compact coil. If this coil were completely stretched out, its function would be like a rather long tube, reaching a length of 35 mm (Fig. 1.15). Additionally, while the cochlea gradually becomes narrower as it approaches the end tip, the basilar membrane, in contrast, gradually becomes wider towards the tip. In particular, the width of the basilar membrane that contains the hair cells is 0.1 mm at the entrance but 0.5 mm at the very end.

The pressure applied to the stapes and oval window, located at the entrance of the cochlea, reaches the cochlear tip via the lymph fluid. This pressure, in turn,

Fig. 1.15 Schematic representation of extended cochlea



bends the basilar membrane. This bending, appearing at both the oval window and the round window (tympanum window), reaches the tip within about 5 ms. Some hypotheses that frequency analysis occurs at the basilar membrane have existed for a long time.

Among these, Helmholtz’s (Hermann Ludwig Ferdinand von Helmholtz, 1821–1894) hypothesis gained widespread support for some time. He believed that the thin fiber on the basilar membrane played a role, similar to the keys on an automatic piano, in allocating sound according to its pitch. That is to say, when a certain sound enters the ear, only specific keys vibrate, meaning that an extremely sharp frequency analysis is already done at the basilar membrane. As will be explained more fully later, this hypothesis has been proved from a different point of view.

On the other hand, Békésy (Georg von Békésy, 1809–1972) doubted that only the thin fiber vibrated as the basilar membrane stretched itself out in a wide manner. Therefore, he used elephants’ ears, which are much larger than those of humans, to observe the vibration of the basilar membrane. As shown in the left side of Fig. 1.16,

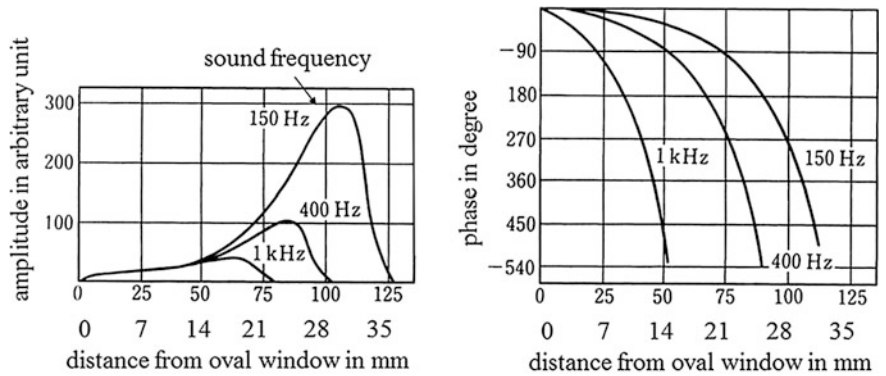


Fig. 1.16 Amplitude and phase of vibratory pattern of basilar membrane

the results indicated that the vibration of the basilar membrane extends over a very large area, and that it disappears at the end of the cochlea [13]. This is a sort of traveling wave, corresponding to the waves that disappear at the end of a long rope fixed to a wall after being shaken by a hand. For this reason, Békésy's theory is known as the "traveling wave theory."

This means that the part that vibrates most depends on the frequency of the sound. Specifically, a broad frequency analysis occurs because high-frequency sounds are analyzed at the entrance of the cochlea, whereas low-frequency sounds are analyzed at the end of the cochlea. The average speed of the traveling wave is 6–7 m/s. As the wave approaches the end of the basilar membrane, it slows down because the basilar membrane gradually softens towards the end. Békésy also performed detailed research on the relation between the intensity and the phase of the vibration, thereby clarifying that the phase moves in a negative direction from the part where the vibration is largest as shown in the right side figure. This means that there is a delay factor in the transmission function. Although this place theory was later revised, the fundamental idea is still in effect.

1.3.4 Hair Cells and the Auditory Nerve System

According to Békésy's study, the sharpness in the resonance of the basilar membrane can be represented by a Q-factor (center frequency/3-dB bandwidth) of 6. Since this value corresponds to a slope (sound intensity/frequency) of 6-dB/octave in a low-frequency region and -20 dB in the high-frequency region, the resonance characteristics of the basilar membrane do not show such a high resolution.

The human auditory system has an extraordinary high-frequency resolution for sounds, making it possible to differentiate 1003 Hz from 1000 Hz. This fact means that the differential limen of the sound frequency is only 0.3 $((1003-1000 \text{ Hz})/1000 \text{ Hz})$. Békésy hypothesized that this high-frequency resolution occurs in the auditory nervous system. Katsuki attempted to prove that the auditory nervous system might sharpen the low-frequency resolution. He did this by using a nervous lateral inhibition function predicted by his studies [14]. Those studies shall be explained in detail later.

Békésy indirectly proved the existence of such a sharpening mechanism by utilizing the human tactile sense. Specifically, he designed a membrane on a tube filled with liquid water, similar to a basilar membrane on a cochlea filled with lymph liquid inside an ear, as shown in Fig. 1.17. He then attached the membrane to a human forearm in order to correspond hair cells and the auditory nervous system to tactile receptors and the tactile nervous system, respectively.

Although the membrane on the tube vibrated widely by applying a vibration to the edge of the tube, the perceived vibratory patterns were much sharper than the vibrating membrane patterns, as schematically represented in Fig. 1.18. From this simulation of the cochlea using the tactile sense, Békésy indirectly proved that sharpening in the auditory sense might be realized by a function of the nervous system. He further hypothesized that the same function in the nervous system as the

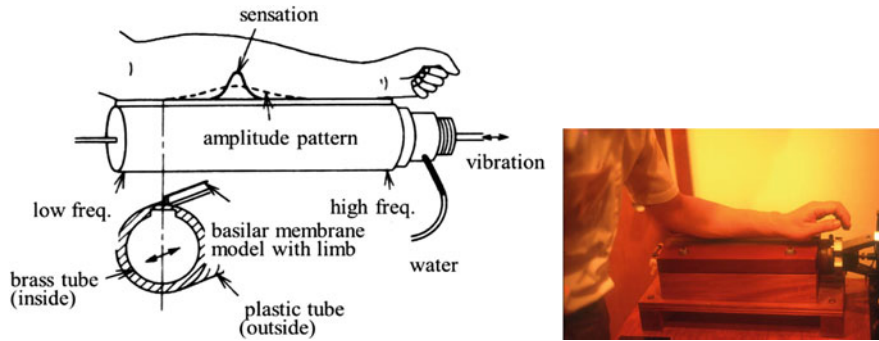


Fig. 1.17 Fluid tube model (left) of cochlea and the photograph (right)

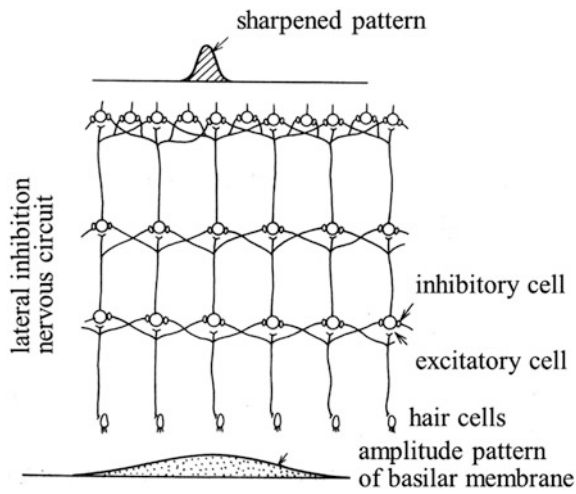


Fig. 1.18 Sharpening characteristics inside auditory nervous system

auditory sense also exists in the tactile nervous system. This idea was the origin of a “tactile aid” that utilizes the tactile sense of the hearing impaired as a substitute for the auditory system.

However, there were inherent limitations to Békésy’s experiments, as mentioned in the previous paragraph. Since he only used cochlea removed from elephant cadavers, measurements of basilar membrane movement could only be made under intense pressure. Specifically, it has long been a supposition that a higher frequency resolution might be observed in the cochlea membrane even for weak sounds if a human cochlea were used.

Since a new method of measurement, known as the Mössbauer method [15], was discovered in 1980 to detect living cochlea membrane movements, the frequency-analyzing mechanism in the cochlea has been revised by many auditory

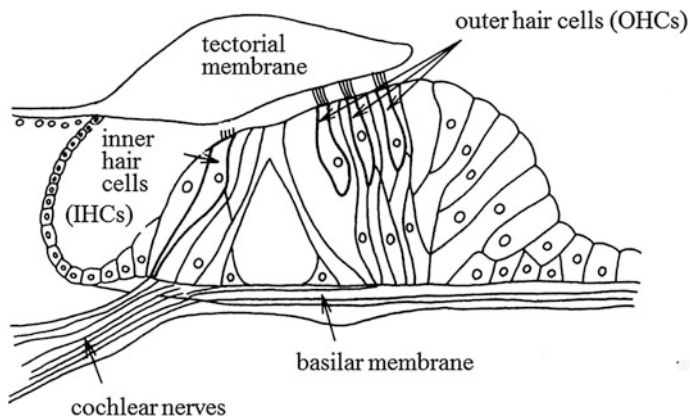


Fig. 1.19 Structure of Corti organ, and OHCs and IHCs

neuroscientists. According to the findings obtained through the Mössbauer method, it was learned that when the stimulating sound level is less than 30, the living cochlea has a very high resonance level (Q): approximately 13 in the low-frequency region and 100 in the high-frequency region.

Furthermore, it has been proven that the basilar membrane itself has such a high frequency resolution and the frequency analysis works adaptively according to sound level due to feedback control of the hair cells, as shown later. The basilar membrane works as a low-frequency analyzer for loud sounds and as a high-frequency analyzer for weak sounds.

As mentioned, the extraordinarily weak movements of the basilar membrane are received by 16,000 hair cells, arranged in four rows on the membrane. The hair cells inside the cochlea of human beings are divided into two types: outer hair cells (OHCs) and inner hair cells (IHCs) as shown in Fig. 1.19. The total number of OHCs, arranged in three rows, is around 12,000, whereas the total number of IHCs, arranged in one row, is about 3500. One OHC is about $8\text{ }\mu\text{m}$ in diameter and has around 140 hairs, whereas an IHC is about $12\text{ }\mu\text{m}$ in diameter and has around 40 hairs. The hair cells are so small and delicate that their function decays through natural aging and is sometimes lost due to the influence of certain antibiotic medicines. In fact, some forms of hearing impairment are caused by the loss of function of the hair cells.

Although all hair cells are inside the Corti organ (see Fig. 1.19), Corti discovered that only the hairs of the OHCs are connected to a tectorial membrane, causing them to bend and vibrate according to the basilar membrane's movements. Although the hairs of the IHCs are not connected to the tectorial membrane, the hairs are bent through the movement of lymph liquid inside the cochlea when the vibration of OHCs increases beyond a certain intensity. It is said that the frequency resolution in the basilar membrane is caused by a combination of OHCs and IHCs.

Through the bending of these hair cells, their mechanical energy is converted into electrical energy, thereby increasing the receptor potential of the hair cell. The

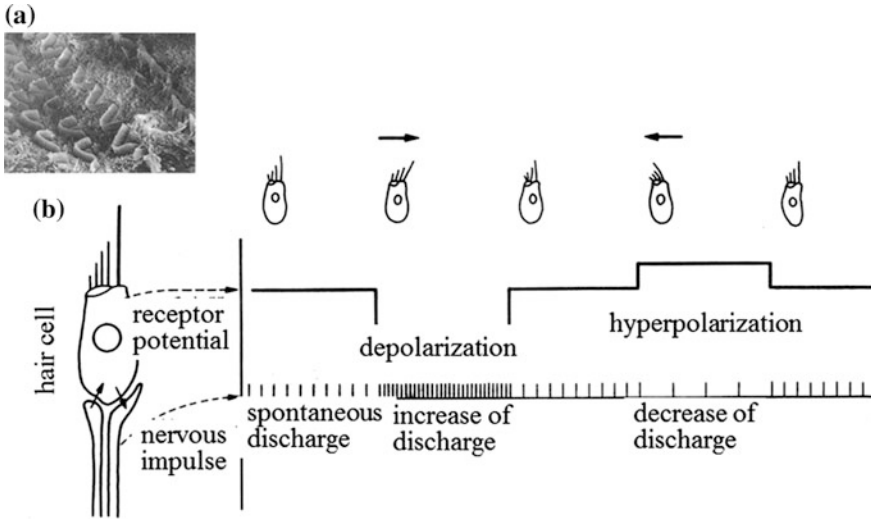


Fig. 1.20 **a** Micrograph of hair cells inside cochlear of guinea pig, **b** auditory nervous impulses produced by bending hairs of hair cells

increase in the potential enables neurotransmitters to travel to the cell membrane, where they are released outside the cell. The transmitters raise the synaptic membrane potential of auditory nerves, resulting in an increase in the membrane potential as shown in Fig. 1.20. When the membrane potential exceeds a certain threshold, impulses are produced that travel to the central nervous system (CNS) through auditory nerves.

The nerve fibers through which the impulses travel from the peripheral nervous system to the CNS are called “afferent nerves,” but when the impulses travel from the CNS to the peripheral nervous system, they are called “efferent nerves.” Since all receptors are connected by both kinds of nerves, the receptor potential undergoes complex changes through control by such nerves.

Before the roles of the OHCs and the IHCs were clarified, Katsuki had proved the existence of the sharpening mechanism of the auditory frequency resolution from the viewpoint of neuroscience. Specifically, he was able to measure the impulses of auditory neurons in cats by utilizing a microelectrode technique while producing sound stimulation of various intensities and frequencies. As the auditory nervous system of humans is basically the same as that of cats, as schematically shown in Fig. 1.21, it was believed that Katsuki’s experimental results could also be applied to understanding the human auditory system.

Auditory nerves that are synaptically connected to hair cells are called “auditory primary neurons.” In Katsuki’s study, after inserting a microelectrode into one of the auditory primary neurons, cats were subjected to sound stimulation of varying intensities and frequencies in order to measure the thresholds at which the neurons fired. By plotting the threshold as a function of the sound frequency, it became clear

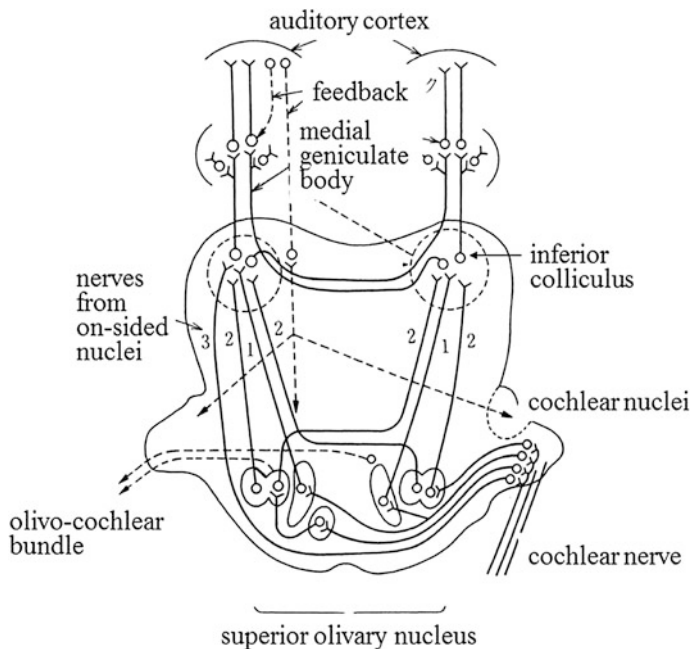


Fig. 1.21 Schematic representation of auditory nervous system

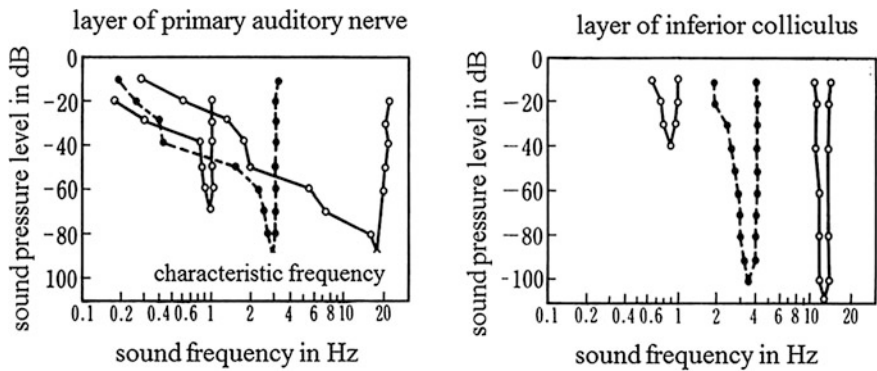


Fig. 1.22 Sharpening characteristics inside auditory nervous system

that, at a certain frequency, the threshold showed a minimum. This frequency is known as the “characteristic frequency” (CF). Katsuki’s study also indicated that the threshold gradually increases at lower frequencies than the CF, while it does not increase so gradually at higher frequencies than the CF, as shown in the left part of Fig. 1.22. It is assumed that the pattern showing the frequency-dependent threshold would faithfully reflect the spatial movement pattern of the basilar membrane. This

similarity implies that the frequency resolution at the level of the auditory primary system is also broad, like that for the basilar membrane.

Most nerve fibers connecting to the auditory primary neurons cross over at the medulla level, called the “olive nucleus,” and then reach the next layer, known as the “inferior colliculus” (IC). By inserting a microelectrode into one of the IC neurons, the thresholds for sound stimulation were obtained as a function of the sound frequency, using the same measurement method as for primary neurons. The results are shown in the right part of Fig. 1.22. Comparing the two graphs, it is apparent that the threshold-frequency pattern for the IC neurons is much sharper than that for the primary neurons. In support of Békésy’s hypothesis, this fact indicates that a sharpening mechanism does indeed exist between the primary neurons and the IC neurons.

It is hypothesized that the “lateral inhibition” neural circuit performs an important function in the sharpening process. Many nerve fibers that horizontally connect to surrounding neurons have been found in the IC layer. Furthermore, it is known that nerve cells include a number of inhibitory neurons that decrease the membrane potential of the surrounding neurons by means of the lateral inhibitory function. In this way, the weakly firing inhibitory neurons cause a slight decrease in the membrane potential of the surrounding neurons. Consequently, the frequently firing neurons fire at an even higher rate than the surrounding neurons. This process causes the broad threshold-frequency pattern in the primary neuron layer to be sharpened in the IC neuron layer.

It is assumed that the sharpening mechanism functions at least for a sound intensity greater than 30 dB, an intensity for which the sharpening mechanism of the basilar membrane ceases to function. Actually, recent studies of cochlear implants (see Chap. 3) still indicate that the auditory nervous system plays an important sharpening role, as patients with such implants retain this sharpening ability despite the failure of their basilar membranes to function.

Békésy hypothesized that the visual and tactile senses possess the same sharpening function as the auditory sense. Based on psychophysical experiments, he verified that the sharpening function is also present for the visual and tactile senses. He coined the term “sensory inhibition” to indicate the sharpening function that can be observed by other senses. He gave the name “neural unit” to the spatial band-pass filter. A method by which to estimate the neural unit will be discussed later in this chapter.

It has been found that the threshold-frequency pattern becomes sharper at a higher level in the auditory nervous system. This pattern becomes sharpest at the next layer of the IC, called the “medial geniculate body.” Furthermore, the sensation of loudness with regard to sound intensity is also sensed in the medial geniculate body, since the firing frequency of impulses is linearly proportional to the degree of loudness within the level of the medial geniculate body. It has been ascertained that the cerebral cortex processes the more complex sounds, such as their tone color and time sequence.

1.4 Mechanism of Speech Production

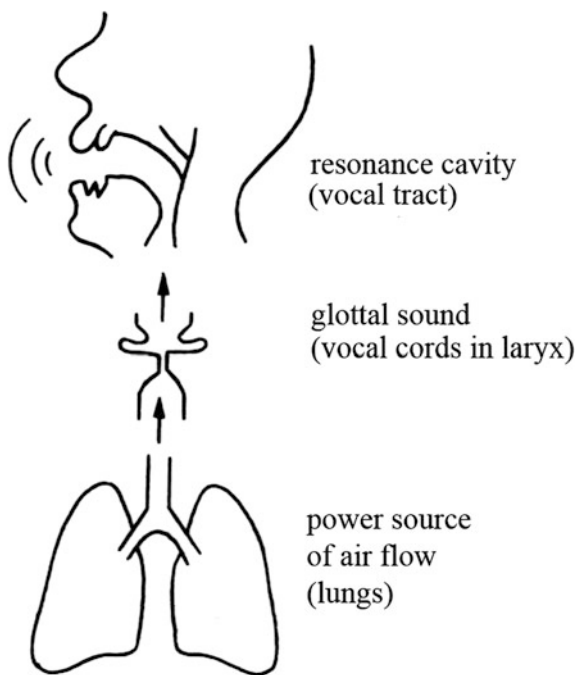
1.4.1 Structure of the Speech Organ

There are many animals that can produce vocal sounds. The majority of these animals can express their emotions or thoughts by changing the tone of the sounds they make. However, even in the case of chimpanzees, it is nonetheless impossible for them to produce sounds like human speech. This is because their speech cortex does not function like the “Broca’s area” in humans. Specifically, the Broca’s area functions in a way that is superior to the speech cortex of other animals. Thereby, the human speech organs—including the lips, tongue, and jaw—constitute the essential differences between humans and other animals.

As shown in Fig. 1.23, the speech-production process is divided into three stages. The first stage involves exhaling through the larynx; the second is characterized by the glottal sounds produced in the vocal folds, and in the third stage, the final sounds are formed through the use of the vocal tract—including the tongue, lips and jaw.

By inhaling during the first stage, as the lungs and chest expand, they produce the elastic force necessary for the exhalation pressure. As people exhale, they adjust the pressure to produce specific sounds by consciously controlling the exhalation muscle. By increasing the exhalation pressure, both the intensity and pitch

Fig. 1.23 Three stages of speech production



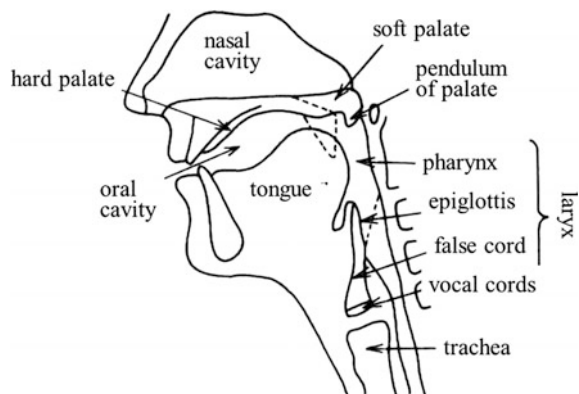
frequency of the sounds increase. This positive correlation between the intensity and the pitch frequency makes it possible to change the pitch frequency just by controlling the exhalation pressure. One example of this is in the case of people who have lost their larynx yet retain control of pitch frequency by the use of exhalation pressure, as shown in Chap. 6.

The exhalation pressure that is produced during the first stage becomes the power source for speech in preparation for the second stage. Although the exhalation pressure is held relatively steady at a frequency of around 1 Hz, it is converted into an alternative form of acoustic energy at a pitch frequency of around 200 Hz through the larynx. The name “larynx” can be traced to the Greek word “larugx ($\lambda\alpha\rho\nu\gamma\xi$),” meaning to shout or the Latin word “lurcare,” meaning to eat voraciously. As a result, the meaning of larynx in modern English has been used to imply either voice or swallowing.

Figure 1.24 schematically represents a structure of the speech organ. From the figure, it is found that the larynx is located under the pharynx, where the trachea and the esophagus separate. The original role of the larynx was to protect the trachea from food toxins. Its structure consists of some cartilages with attached muscles and vocal folds that produce sounds (see Fig. 1.25a). Muscle ligaments and mucous membranes are attached to the bones and vocal folds that produce sounds. The vocal folds’ elastic tissue is composed of muscle ligaments and mucous membranes, which are located inside the cartilage, as shown in Fig. 1.25b.

The tension and elasticity of the vocal folds are changeable, as is the length and width of the vocal folds themselves. Furthermore, the open space within the vocal folds is adjustable and the location of the vocal folds themselves can be moved vertically through movement of the larynx. While a person is speaking, the above changes are constantly occurring. Through the evolution of the larynx from a simple valve to a speech-production organ, the larynx has developed complex movements by controlling numerous muscles that work together. In the event of a disease such as cancer of the larynx, a surgeon will remove both those muscles as well as the larynx itself. Following such an operation, information to control the

Fig. 1.24 Schematic representation of speech organ



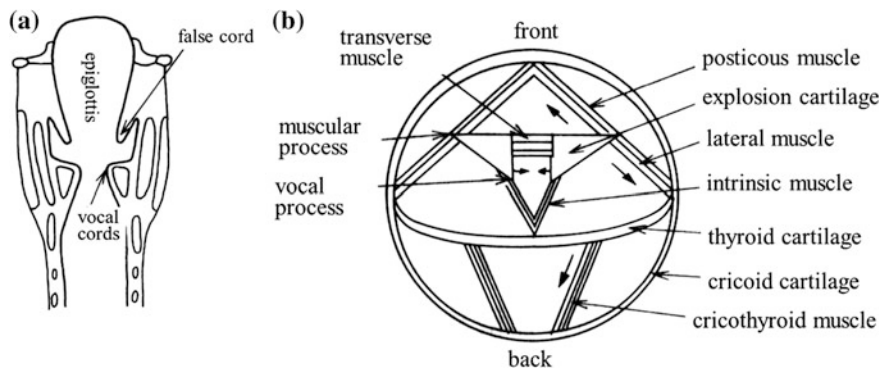


Fig. 1.25 a Profile of larynx, b schematic representation of laryngeal muscle

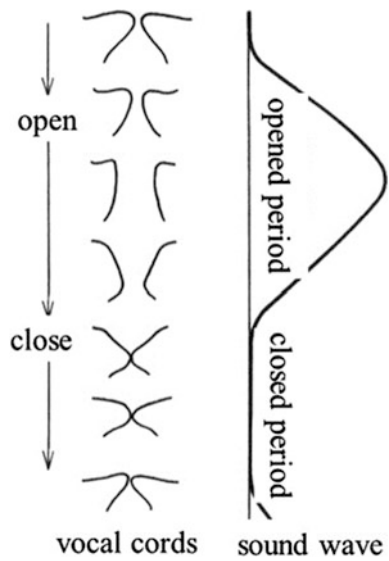


Fig. 1.26 Movement of vocal folds and glottal sound wave

larynx is lost. Therefore, it becomes difficult to produce natural speech, even by using such substitute speech mechanisms as an artificial larynx or through “eso-phageal speech.”

In the process of producing glottal sounds, first the vocal folds bend inwardly and then the glottis closes, as shown in the left figure of Fig. 1.26. Next, the pressure under the glottis increases, making the glottis open. When the glottis is open, air is able to flow through. The effect of this air flow causes the glottis to close because of both the effect of Bernoulli’s principle and the elasticity of the vocal folds. As this opening-closing process repeats, the vocal folds vibrate, thereby resulting in the production of glottal sounds. The vibratory frequency of the vocal

folds is determined by the elasticity, tension and mass of the folds. Each of these adjustments is performed by the muscles attached to the cartilage of the larynx.

The frequency of the glottal sound is mainly adjusted by controlling the tension and the mass of the vocal folds. The adjustments are performed by the vocal fold muscle and the cartilage muscle inside the larynx. Very low frequency speech sounds, known colloquially as “thick sounds” or “breath sounds,” are controlled by the above muscles. However, speech sounds in the very high frequency range, known as “falsettos” or “head sounds,” are controlled solely by the force of exhalation.

Prior to the final emission of the sounds from the mouth, the glottal sounds pass through the pharynx, the oral cavity and the nasal cavity, during which time the tone color of those sounds is altered by the acoustic resonance within those cavities called the “vocal tract”. This process constitutes the third stage of speech production. The sounds produced by the vibration of the vocal folds, called “puffing sounds,” are shown as sound waves in the right part of Fig. 1.26. From this figure, it can be seen that the sound waves are triangular in shape. The waveform becomes larger as the exhalation pressure increases. However, the shape of the waveform does not change much.

The fundamental glottal sound frequency increases as the tension in the vocal fold muscles increases. Although the waveforms of the glottal sounds do not change in the vocal folds, the sounds produced from the lips exhibit a variety of tones. The reason for this is that resonance frequencies can be changed by the shape of the vocal tract. For instance, the resonance frequencies can change significantly by altering the position of the tongue inside the vocal tract. The differences in these resonance components, called “formants,” characterize different vowels.

1.4.2 Resonances Inside the Vocal Tract

As shown in Fig. 1.27, which resonance frequencies are inside the vocal tract can be predicted by considering the vocal tract to be like a cylindrical tube with a length (L) of 17 cm that corresponds to the distance from the vocal folds to the lips. When the maximum amplitude of the sound wave is at the end of the tube, the sound

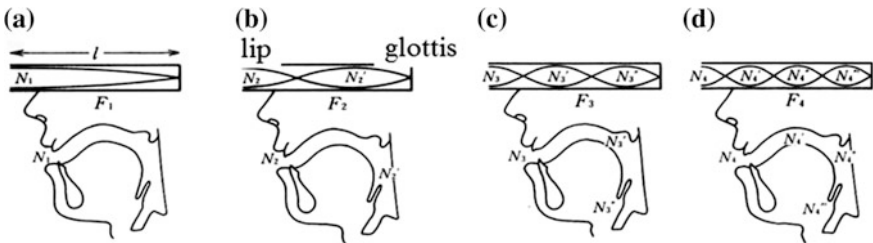


Fig. 1.27 Resonance in vocal tract

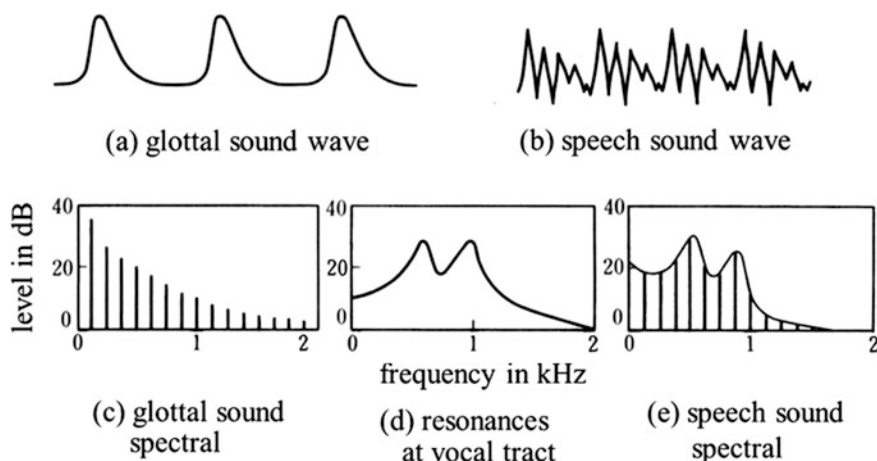


Fig. 1.28 Production process for speech power spectrum

becomes loudest. In other words, when $1/4$ th of the sound wavelength (λ) is equal to the length of the tube resonance occurs. In the case of a tube with a length of 17 cm, this corresponds to a resonance frequency (f) of 500 Hz when the sound velocity (C) is 340 m/s. It is calculated by the equation $f = C/\lambda = 34,000 \text{ cm}/(4 \times 17 \text{ cm})$, where $\lambda/4 = L$.

Furthermore, when $3/4$ ths and $5/4$ ths of the wavelength are equal to the tube length, the sound intensity reaches a maximum at the end of the tube. In the same way as above, resonances are created at frequencies of 1500 and 2500 Hz. Since the waveform at the end of the tube is triangular, its spectrum has a harmonic structure with multiple frequency components in addition to a fundamental frequency. Since the waveform at the beginning of the vocal tract is also nearly rectangular in shape, its spectral pattern decreases at 12 dB/oct, as shown in the left figure of Fig. 1.28. Under the influence of the resonance inside the vocal tract, the corresponding formants and their subsequent complex waveforms are created as shown in the middle of the figure. Actually, the spectral pattern produced at the position of the lips decreases at 6 dB/oct due to the influence of acoustic emissions.

Although the speech spectrum has an unlimited number of formants, all vowel sounds are determined by the combination of the first and second formants. As people can control the shape of their vocal tract, the formant frequencies can therefore be determined at will. Even in the case of substitute speech used by people who have undergone a laryngectomy, they can still control the formant frequencies by changing the shape of their vocal tract. Based on this principle, various speech substitutes have been proposed, as discussed in Chap. 6.

1.4.3 Speech Production Model

By using a cylindrical tube model, we can easily prove that the shape of the vocal tract and the waves produced through them have a one-to-one correspondence. It is known that we can estimate the shape of the vocal tract, according to the waves coming from the lips, by using a cylindrical tube model. The vocal tract has a continuous and smooth shape that, for the purpose of the model, can be divided into sections. As Fig. 1.29a shows, the tube model is made up of several disks of equal diameter that approximate the structure of the vocal tract. The left side of the figure corresponds to the glottis, while the right side corresponds to the lips. In forming this model, we usually use a total of 12 or 13 disk-shaped tubes. The parts are indicated by n , $n + 1$, $n + 2$, etc., and the corresponding areas are indicated by A_n , A_{n+1} , etc. Sound waves emitted from the left side are partially reflected at the points where the area changes. The coefficient of reflection is denoted by (1.1):

$$\alpha_n = (A_n - A_{n+1}) / (A_n + A_{n+1}) \quad (1.1)$$

$$\begin{aligned} F_{n+1} &= F_{n,p} + B_{n+1,r} = (1 - \alpha_n)F_n + \alpha_n B_{n+1} \\ B_{n+1} &= F_{n,r} + B_{n+1,p} = \alpha_n F_{n+1} + (1 - \alpha_n)B_{n+1} \end{aligned} \quad (1.2)$$

As Fig. 1.29b shows, the components that allow sound to penetrate $F_{n,p}$ and the components to reflect sound $F_{n,r}$ are produced at points n and $n + 1$. The components that allow reflected sound to penetrate from the right side ($B_{n+1,p}$) and the components to reflect the sound ($B_{n+1,r}$) are denoted by the coefficient of reflection α_n , thereby rearranging the formula as (1.2).

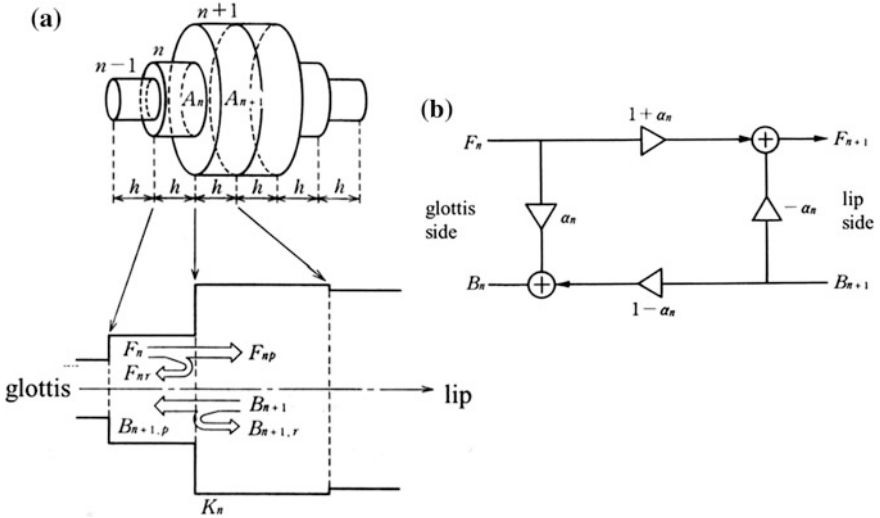


Fig. 1.29 a Acoustic tube model of vocal tract, b block diagram representation of acoustic tube model

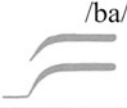
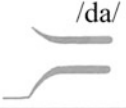
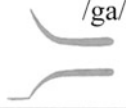
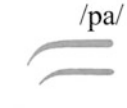
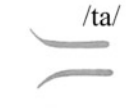

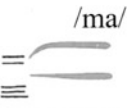
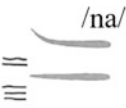
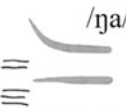
		place of articulation		
		lips	alveolar ridge	palate
method of articulation	voiced plosive	/ba/ 	/da/ 	/ga/ 
	unvoiced plosive	/pa/ 	/ta/ 	/ka/ 
	nasal	/ma/ 	/na/ 	/ŋa/ 

Fig. 1.30 Nine consonants classified by a place of articulation and a method of articulation. The time spectral patterns for the consonants are schematically indicated in the figure

The above formula can also be represented as shown in Fig. 1.29b.

Since we can calculate the area (A_n, A_{n+1}, \dots) of any cross section of the vocal tract by knowing α_n , it is also possible to calculate that area when we know the shape of the sound waves. In reverse fashion, by first knowing any cross-sectional area of the vocal tract, it is possible to synthesize any of the vowels. Lastly, when the vowels waveform is known, we can use that information to calculate any cross-sectional area of the vocal tract. This is the basic idea of speech analysis and synthesis methods: to make use of a few parameters to express sound information by calculating the cross-sectional area of the vocal tract from the shapes of vocal fold waves or vice versa. This principle is applied to synthesize speech sounds, especially vowel sounds, with a few parameters.

Most consonants are produced by rapidly changing the shape of the vocal tract in time. The duration from the starting point of each consonant to the following vowel is in the range from around 10 to 100 ms. Typical examples of consonants are shown in Fig. 1.30, where each consonant is classified by a “method of articulation” and a “place of articulation”. The place of articulation indicates the place of “narrowing point” of the vocal tract made by lips or tongue and it also means the starting point of the movement of the lips or the tongue. The movements are strongly related to the changes of the second formant frequency as represented in the time spectral patterns of the figure. The method of articulation is largely divided into the voiced and the unvoiced consonants. In the Fig. 1.30, the unvoiced consonants are the “plosives” /p/, /t/, /k/ and the voiced consonants are both “nasal” /m/, /n/, /ŋ/ and plosives /b/, /d/, /g/.

In addition to the examples described above, the “fricatives” /sh/, the “affricative” (/ts/ and /ch/), and the “semivowel” (/w/ and /y/) are used in various languages. Needless to say, there are many different consonants depending on the language. For example, the “liquid sound” /l/ and the “nasal sound” /ŋ/ are Japanese-specific consonants.

PARCOR (Partial Correlation) invented by Itakura [16] is the basis of all methods in use throughout the world. Furthermore, CELP (Code Excited Linear Prediction) method developed by Schroeder and Atal is widely used for recent mobile phones [17]. Since it is possible to synthesize various sounds with a few parameters and random noises, it is also used in the production of sound synthesizers for speech disorders. Furthermore, it has become possible to quantitatively express speech impediments caused by abnormal shapes of the human mouth.

1.5 Maximum Transmitted Information by Human Senses

Research in sensory substitutes occupies an important position in assistive engineering. However, the most important factor yet to be understood is how sound can be conceptualized by means of residual senses and nervous systems. At this point, the author will not deal with the conceptual aspects of sound but rather limit his discussion to a practical understanding of the design of devices that can serve as sensory substitutes.

Generally speaking, when we substitute one sense for another, it is important to determine quantitatively how much information the substituting sense is able to transmit. To illustrate this point, we will consider the sense of touch as an example. In particular, by applying communication theory and psychological experiments, the author will describe the method used to acquire the maximum quantity of information through vibratory stimulation to two points within a certain limited area of the tactile sense. It goes without saying that the concept explained below is likewise applicable to both visual and auditory senses.

1.5.1 Category Decision Test and Channel Capacity

As Miller has indicated, the maximum transmitted information (R_t) accrued through one point of stimulation can be evaluated by means of the following equation based on a category decision test [18].

$$I(X, Y) = H(X) - H_y(X) \quad (1.3)$$

$$R_t = \max\{I(X, Y)\}$$

$H(X)$ represents the entropy of the source signal, while $H_y(X)$ equals the dissipated information. Each can be found with the following equation.

$$H(X) = -\sum P(x_i) \log_2(1/P(x_i)) \quad i = 1, n$$

$$H_y(X) = -\sum \sum P(x_i, y_j) \log_2(1/P(x_i/y_j)) \quad i = 1, n, j = 1, n \quad (1.4)$$

$P(x_i)$ equals the occurrence probability of an intensity level (x_i). Here, the level (x_i) of stimulation intensity is divided into m degrees. Therefore, if all occurrence probabilities are same, $P(x_i)$ would equal $1/m$. The intensity level $P(x_i/y_j)$ represents the probability of a given intensity (x_i) when a subject replies that he/she received a stimulation of intensity (y_j). When we find the value of $I(X, Y)$ by varying the degree m , the largest $I(X, Y)$ should be designated as R_t . A conceptual diagram for the transmitted information is shown in Fig. 1.31. The author would like to explain how to determine R_t based on his experimental results [19].

In the experiment on the designation of R_t , the vibratory frequency (200 Hz)—having a sensation level of 14 dB (0 dB equals the threshold of sensation)—was equally divided logarithmically to m degree. The vibrations were applied 20 times at random to the fingertip of the normal subjects without any disabilities, at which time the subjects were asked to identify which degree of the stimulation intensity was perceived. The results of relationship between $I(X, Y)$ and m are shown in Fig. 1.32. The horizontal axis indicates m , and the vertical axis indicates the transmitted information $I(X, Y)$. It can be seen that $I(X, Y)$ reaches a saturation point when m is 5 or greater. Consequently, the maximum transmitted information (R_t) is around 1.75 bits.

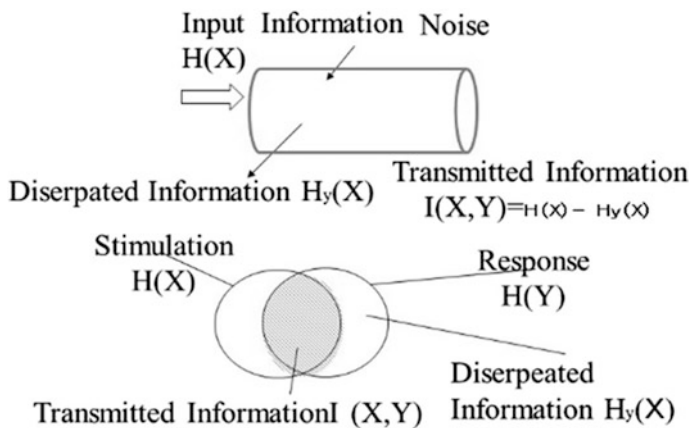
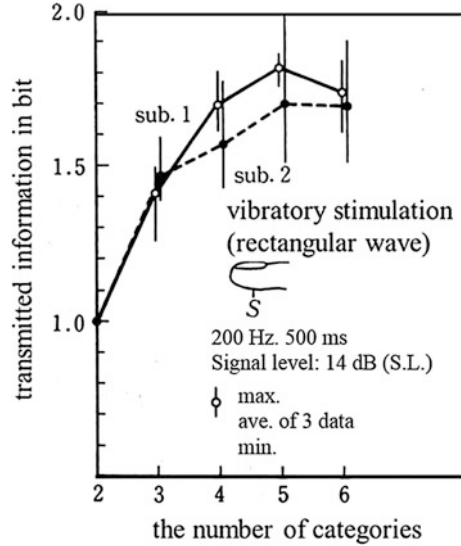


Fig. 1.31 Conceptual figure of transmitted information

Fig. 1.32 Transmitted information $I(X, Y)$ as function of the number of categories



1.5.2 Maximum Transmitted Information for Multiple Stimulation

In this manner, R_i can be ascertained rather easily in the case of a one-point stimulation. However, when a multiple stimulation is applied simultaneously on certain limited tactile surfaces, how does R_i change? To examine this, it is necessary to measure R_i when two points are stimulated according to a category decision test.

For the experiment, two vibrations were chosen from a vibrator array arranged in 16 rows, each 1 mm apart, with 3 mm between each vibrator. Next, with one of the two vibratory stimulations, a masker stimulation of the constant intensity level was generated while the signal vibratory stimulation was overlapped with the other vibrator. In the category decision test, the intensity level of the signal was only varied at a random rate. Moreover, the intensity level of the signal was equally divided into 5 degree. Both signals and maskers were applied every 2–3 s on the fingertip of the subjects' index fingers, and the subjects were asked to identify to which of the 5 degrees the signal belonged.

As shown in Fig. 1.33, it was found that the maximum transmitted information corresponding to R_i varied greatly depending on the distance x [mm] between two stimulation points and the intensity level M [dB] of the masker. When we consider the relation between the masker level (M) and R_i , by fixing the distance x at 3 mm, it was found that R_i tends to decrease as M increases. In the case of two simultaneous stimulations, there is mechanical interference combined with the inhibition of nervous systems. This likely results in a weaker sense perception of the signal compared to the case of a single point of stimulation. Accordingly, a decrease in subjective intensity level of the signals constitutes one of the factors that reduce R_i .

Fig. 1.33 Transmitted information as function of masker level

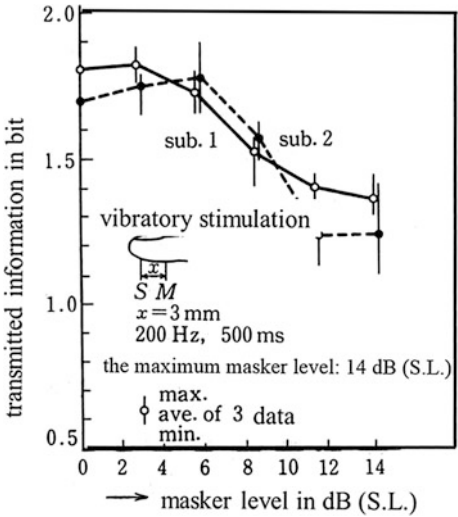
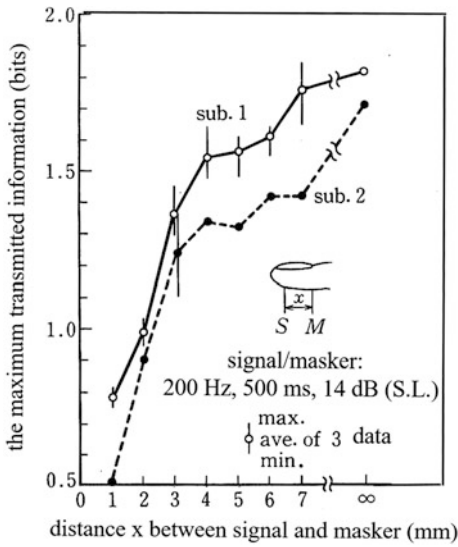


Fig. 1.34 Maximum transmitted information as function of distance between signal and masker



Next, as seen in Fig. 1.34, when we fixed the masker level (M) at 14 dB and determined the value of R_t by varying the distance x between the masker and the signal, it was found that R_t decreases in proportion to x . The degree of the decrease became particularly obvious when x is less than 3 mm. As stated previously, the cause of this decrease would also appear to result from the decrease of subjective intensity of the signals by the masker. Furthermore, when the distance between the

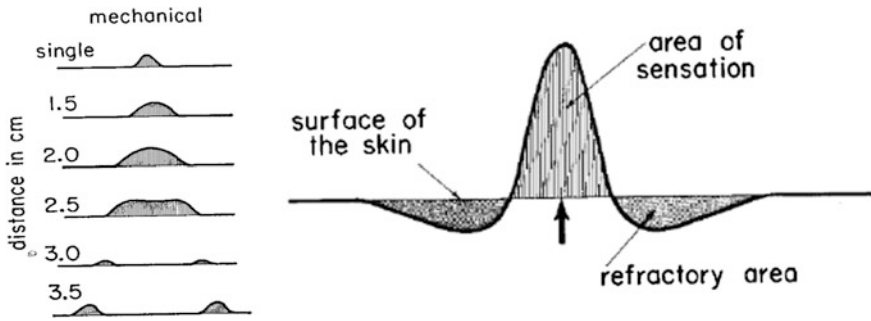


Fig. 1.35 Local distribution of sensory magnitude for mechanical stimulation with two points and its surrounding area of inhibition

two points of stimulation decreased, the overlap of the receptive field and mechanical interference became obvious. Therefore, the difficulty in perceiving the signals separately is also probably one of the causes of the decrease in R_r . The phenomenon is something that cannot be evaluated when the maximum transmitted information is derived from one point of stimulation.

The reason the maximum transmitted information decreased for two point stimulations may be explained by using the “neural unit” proposed by Békésy, as shown in the left side of Fig. 1.35. The neural unit shows the local distribution or receptive field of sensory magnitude for mechanical stimulation with two points and its surrounding area of inhibition. It is hypothesized that when two points of stimulation are added simultaneously to the tactile surface, the subjective intensity (S') and the subjective number (n') of the multiple stimulation greatly change due to the suppressive effect and the overlap of the receptive field as shown in the right side of Fig. 1.35.

This means that the maximum transmitted information may be obtained by using the estimated method mentioned above if S' and n' are determined. Therefore, if the inhibitory and receptive fields of the neural unit can be obtained by psychophysical experiments, the subjective intensity S' and the subjective number n' for the stimulation can also be estimated. One of approaches to determine S' and n' using the masking effect and the two-point threshold will be proposed in Chap. 4.

1.6 Recognition Time Difference Among the Tactile, Auditory and Visual Senses

Being different from letters or figures, sounds constitute information that changes moment by moment. As such, the question is whether the other senses can receive those sounds in real time. Stimulations such as light, sound, and vibrations require a certain time for processing in order for this information to be perceived by each sense organ via their respective sensory channels. It is preferable that the perceiving

time for both the auditory sense and substitute sense be simultaneous when vocal sounds are fed back through the visual or tactile senses. It is known that even people with normal hearing may begin to stutter when they are made to hear their own vocal sounds after a delay of roughly 0.2 s. With this in mind, it is easy to imagine that the delay in perception time is a crucial factor in the control of pronunciation. However, as we have not yet found an adequate method to directly measure this basic perception time, this factor constitutes the first challenge in designing a sensory aid.

Now we will refer to the results of measuring the differences in perception time among the visual, auditory, and tactile senses. This was accomplished by presenting two stimulations at different times. Specifically, when there was a slight time gap in the pulse-like presentation of light and sounds, or when there was a slight time difference in the presentation of sounds and vibrations to an index finger, those differing stimuli were perceived to have occurred simultaneously. Furthermore, the author would like to consider whether the visual sense or the tactile sense is more advantageous as a substitute sense of the auditory sense. In order to answer this question, the author and his colleague examined both the perception time difference and the reaction time—that is to say, the speed of the vocalization reflex—between a subject's receiving stimulations and vocalizing them.

For the measurement of the perception time differences, as our source of light stimulation, we used alternating colors whose brightness was constant and which continuously changed back and forth from green to red, as shown in Fig. 1.36. We

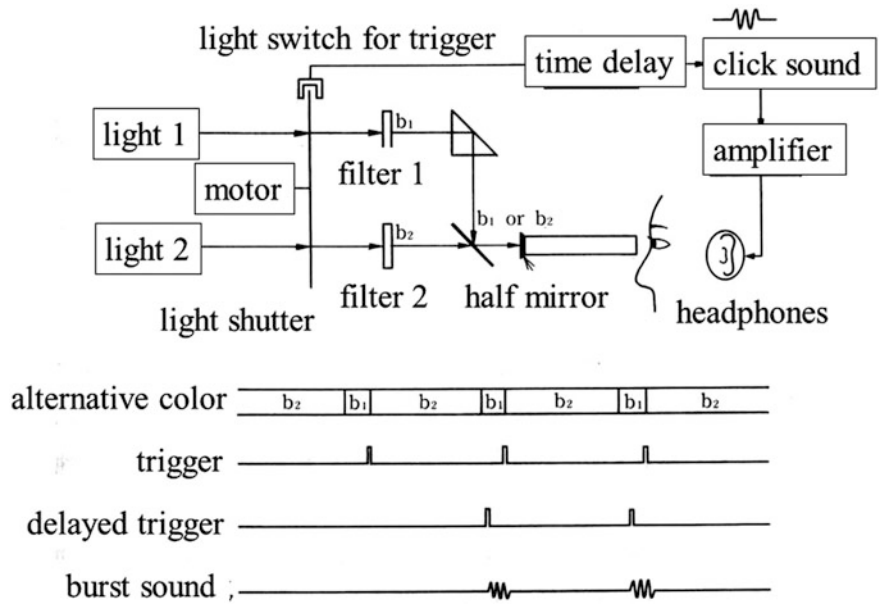


Fig. 1.36 Experimental system for measurement of recognition time difference between vision and hearing

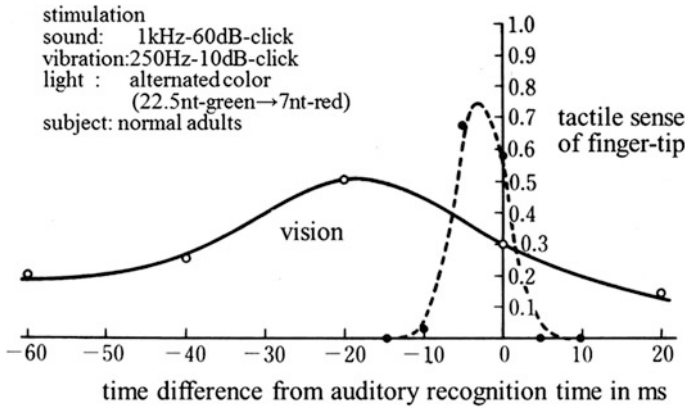


Fig. 1.37 Response probability that two stimulation presented to two different senses were simultaneously perceived. The data were obtained by constant method

set the starting time of the color change as the reference point. We used single sound waves [1 kHz, 60 dB (S.L.)] for our sound stimulation, and single-wave stimulation [250 Hz, 10 dB (S.L.)] for our vibration. The measurements were conducted using a “constant method” and the subjects included people with normal hearing from 10–30 years of age.

In Fig. 1.37, the vertical axis indicates to what degree the subjects perceived the sound and vibratory stimulations simultaneously, and the horizontal axis indicates the time difference (Δt) when subjects simultaneously perceived the two forms of stimulation. The solid line represents the auditory and visual senses, while the dotted line indicates the auditory and tactile senses. From the figure, it is clear that the time delay Δt from the auditory sense to the visual sense was approximately 30 ms, while the Δt from the auditory sense to the tactile sense was also about 5 ms [20]. This fact tells us that compared to the tactile sense, the visual sense experiences a delay from the auditory sense in perception time.

The difference in the time required for perception depends mainly on the number of synapses that a stimulation goes through from the time it is received until it is perceived. It is assumed that the number of synapses to perceive visual stimulation is larger for the visual sense than for the auditory sense. Furthermore, the perception times differed according to the intensity of the stimulation. As the intensity approached the threshold, the delay in perception time reached 100 ms in the case of the visual sense while there was hardly any delay for the tactile sense.

Moreover, it is assumed that these time perception characteristics also affect the reaction time, which is to say, the period of time between the presentation of a stimulus and the receiver’s vocalization. This reaction time corresponds to a time delay in the vocalization system, which is a basic factor that influences our evaluation of the combined intensity of the substitute sense and the vocal system. For the vocalization experiment, flashlight or pulse-like vibrations were presented to subjects with a random stimulation intensity. The subjects were instructed to

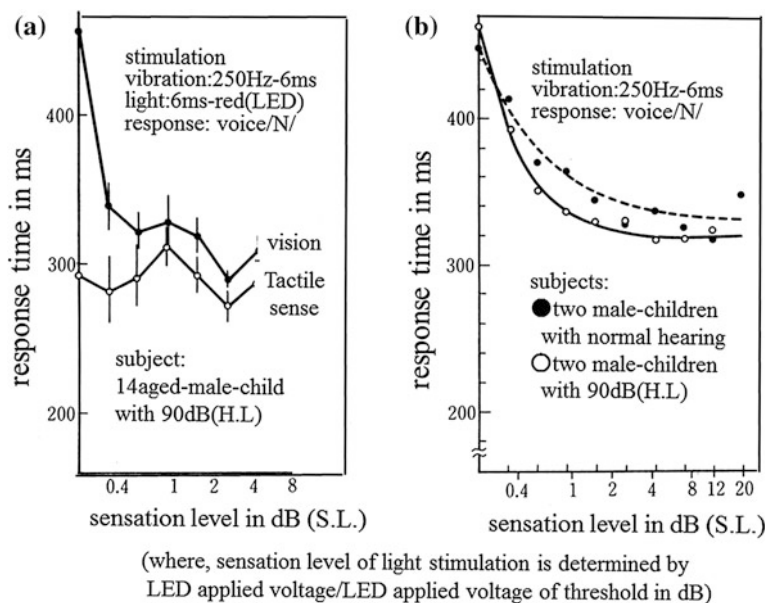


Fig. 1.38 Response time of voice /N/ as function of stimulation level. **a** Comparison of visual stimulation with tactile stimulation, **b** comparison of deaf children with normal hearing

pronounce /N/ as quickly as possible. Deaf students between 12–14 years of age participated in addition to students with normal hearing.

Figure 1.38 shows examples of experimental data of the relation between the time of the earliest sound vocalization(/N/) appearing after the sensory stimulation is introduced (vertical axis) and the stimulation's intensity (horizontal axis). Figure 1.38a is an example of a 14-year-old boy whose hearing level is around 90 dB. Figure 1.38b shows an example of two deaf boys, aged 12 and 13, and two boys with normal hearing of almost the same age. In this experiment, we measured the vocal reaction time resulting from only vibrating stimulation.

Figure 1.38a indicates that the vocal reaction time for visual stimulation is a few dozen milliseconds slower than that for tactile stimulation. In the case of deaf children, it was expected that the combined functions of the visual sense and vocalization would be strong because they have developed habits in their conversations that rely on the visual sense. This reaction time difference is an exact reflection of the perceptive time difference. As Fig. 1.38b shows, in the case of the tactile sense we did not recognize any difference between the children with normal hearing and deaf children of the same age.

Thus, from the point of view of perception time and vocalization reaction time, the tactile sense is more advantageous than the visual sense as a substitute sense for vocal sound perception and vocalization control. It is quite probable that a method could be found that would compensate for this disadvantage in perception time as the visual sense covers a much larger area of perceptive ability compared to the

tactile sense. Naturally, this factor should be taken into consideration in the design of sensory aids.

In the case of auditory substitutes, as mentioned before, the important question to consider is whether it is possible to form a concept of language with voice sound information that is conveyed through other senses, and whether the hearing sense can be combined with the language concepts one has already acquired. Taking the ideas of auditory substitutes into account, from Chaps. 3–5 the author will show how speech information passes through the inside of the human brain from a point of view of studies on “aphasia,” a concept known as a “speech chain” and recent progress in brain research using advanced biomedical measurements.

1.7 Language Area and the Speech Chain in the Human Brain

In 1861, Broca (Pierre Paul Broca, 1824–1880) described a patient who could understand language but could not speak. He had no motor deficits to account for his inability to speak. He could whistle, utter isolated words, and sing the lyrics of a melody. Postmortem examination of his brain showed a lesion in the posterior region of the frontal lobe. This region is now called “Broca’s area.” A few years later, in 1876, Wernicke (Karl Wernicke (1848–1905) described another type of aphasia. This aphasia, or language disorder, involved a failure to comprehend language rather than a failure to speak. The location of the lesion in Wernicke’s patient was different from that in Broca’s patient. It was at the junction of the temporal, parietal and occipital lobes-now called “Wernicke’s area.” The junction of temporal, parietal, and occipital lobes is an important area to remember [21].

Wernicke proposed that language involves separate motor and sensory programs located in different cortical regions. The motor program, located in Broca’s area was suitably situated in front of the motor area that controls the mouth, tongue, and vocal cords. The sensory program, located in Wernicke’s area, was suitably surrounded by the posterior association cortex that integrates auditory, visual, and somatic sensations.

Wernicke’s model is still referred to today. According to this model, initial processing of spoken or written words takes place in the primary and unimodal sensory areas for vision and audition. This information is then conveyed to the “angular gyrus” of the posterior association area. This was thought to be the area where either written or spoken words were transformed into a common neural representation. Then, these representations were thought to be transferred to Wernicke’s area where they were recognized as language and associated with meaning. Without meaning, there can be no language comprehension. These neural representations along with their associated meanings are then passed along, via the “arcuate fasciculus”, to Broca’s area where it is transformed into a motor representation that allows for speech.

As Denes and Pinson describe in their book “Speech Chain,” speech communication consists of a chain of events linking the speaker’s brain with the listener’s brain [22]. They call this chain of events the “*speech chain*.” As shown in Fig. 1.39, the transmission of a message begins with the selection and ordering of suitable words and sentences. This can be called the “*linguistic level*” of the speech chain. The speech event continues on the “*physiological level*”, with neural and muscular activity, and ends on the speaker’s side with the generation and transmission of sound waves, the “*physical (acoustic) level*” of the speech chain.

At the listener’s end of the chain, the process is reversed. Events start on the physical level, when the incoming sound waves activate the hearing mechanism. They continue on the physiological level with neural activity in the hearing and perceptual mechanisms. The speech chain is completed on the linguistic level when the listener recognizes the words and sentences transmitted by the speaker. The speech chain, therefore, involves activity on at least three levels—linguistic, physiological and physical—first on the speaker’s side and then at the listener’s end.

Due to recent progress in non-invasive measurement technologies such as functional MRI and positron emission tomography (PET) of human cerebral cortex, the mechanisms operating in the human auditory cortex have been gradually made clear. The locations of Broca’s area and Wernicke’s area on the cerebrum have been identified and also the speech chain model has been established although it was found that there are other brain areas responsible for different aspects of language (Fig. 1.40). The technologies and findings should be utilized to design assistive devices for people with diseases of the language cortex and the auditory cortex including the sensorineural hearing impaired and aphasia as described in Chap. 6.

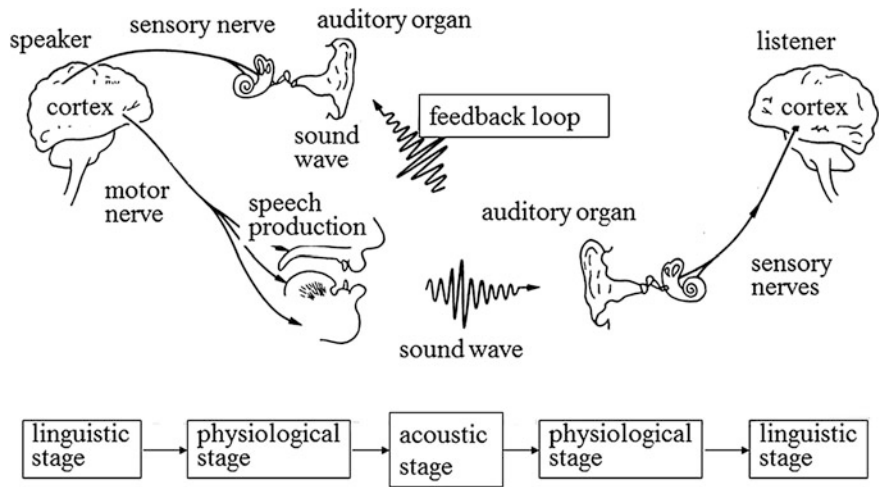


Fig. 1.39 Speech chain

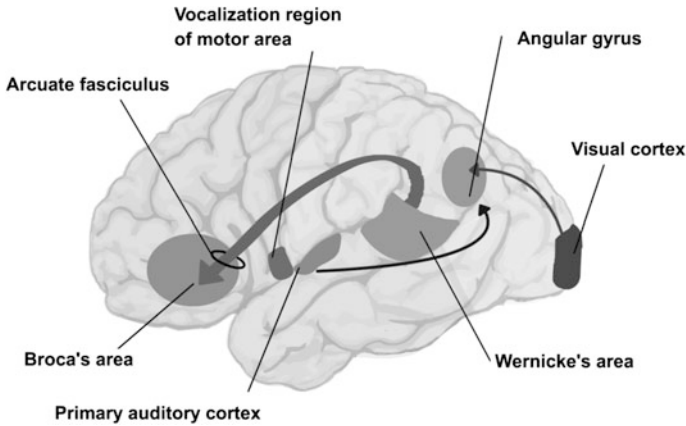


Fig. 1.40 Speech area based on cortex excision

References

1. T. Ifukube, Sound-based assistive technology supporting “seeing”, “hearing” and “speaking” for the disabled and the elderly. Key note speech, in *Proceedings of the INTERSPEECH2010:11–19* (2010)
2. T. Ifukube, A neuroscience-based design of intelligent tools for the elderly and disabled, in *Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly (WUAUC'01)* (ACM Press, 2001), pp. 31–36
3. T. Ifukube, *Challenge of Assistive Technology* (Chuokoron-Shinsha, 2004), p. 12 [in Japanese]
4. N. Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine* (Princeton University Press, 1948)
5. *Annual Report on the Aged Society* (Cabinet Office, 2012)
6. <https://www.jst.go.jp/s-innova/research/h22theme05.html> (2016)
7. R. Hayashi, J. Isogai, P. Raksincharoensak, M. Nagai, Autonomous collision avoidance system by combined control of steering and braking using geometrically-optimized vehicular trajectory. *Veh. Syst. Dyn.* **50**(Suppl), 151–168 (2012)
8. R. Raksincharoensak, Katsumi, M. Nagai, Reconstruction of pedestrian/cyclist crash-relevant scenario and assessment of collision avoidance system using driving simulator, in *Proceedings of 11th International Symposium on Advanced Vehicle Control (AVEC)* (2012)
9. M. Hirose, Koreisha cloud no kenkyukaihatsu (Research and development of senior cloud). *Trans. Virtual Reality Soc. Jpn.* **19**(3), 21–25 (2014) [in Japanese]
10. A. Hiyama, M. Kobayashi, H. Takagi, M. Hirose, Collaborative ways for older adults to use their expertise through information technologies. *ACM SIGACCESS Newslett.* **110**, 26–33 (2014)
11. Article in Newspaper (1901) 20th century prophecies. *Hochi Shimbun* (Houti Newspaper) [in Japanese]
12. E.M. Wenzel, Three-dimensional virtual acoustic displays, in *Multimedia Interface Design*, ed. by M.M. Blattner, R. B. Dannenberg (ACM Press, 1992), p. 257
13. G. Békésy, Neural funneling along the skin and between the inner and outer hair cells of the cochlea. *J. Acoust. Soc. Am.* **31**(9), 1236–1249 (1959). doi:[10.1121/1.1907851](https://doi.org/10.1121/1.1907851)
14. N. Suga, Sharpening of frequency tuning by inhibition in the central auditory system: tribute to Yasuji Katsuki's paper. *Neurosci. Res.* **21**(4), 287–299 (1995)

15. P. Gilad, S. Shtrikman, P. Hillman, M. Rubinstein, A. Eviatar, Application of the Mössbauer method to ear vibrations. *J. Acoust. Soc. Am.* **41**(5), 1232–1236 (1967)
16. F. Itakura, S. Saito, On the optimum quantization of feature parameters in the PARCOR speech synthesizer, in *Proceeding of the Conference on Speech Communication and Processing*, pp. 434–437 (1972)
17. M. Schroeder, B. Atal, Code-excited linear prediction(CELP): high-quality speech at very low bit rates. *Acoust. Speech Signal Process. IEEE Inter. Conf. ICASSP '85.* **10**, 937–940 (1985). doi:[10.1109/ICASSP.1985.1168147](https://doi.org/10.1109/ICASSP.1985.1168147)
18. G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81 (1956)
19. T. Ifukube, Maximum information transmission on a tactual vocoder: In the case of time invariant stimulation. *Jpn. J. Med. Electron. Biol. Eng.* **17**(3), 230–236 (1979). doi:[10.11239/jsmbe1963.17.230](https://doi.org/10.11239/jsmbe1963.17.230). [in Japanese]
20. T. Ifukube, Artificial reality based on biomedical engineering—as an example case of sensory substitute studies. *The J. Inst. Telev. Eng. Jpn.* **46**(6), 718–726 (1992). doi:[10.3169/itej1978.46.718](https://doi.org/10.3169/itej1978.46.718). [in Japanese]
21. P. Wilder, R. Lamar, *Speech Chain and Brain Mechanism* (Princeton Legacy Library, 1965)
22. P.B. Denes, E.N. Pinson, *The Chapter 4 Speech Chain: The Physics and Biology of Spoken Language* (1973)

Sound-Based Assistive Technology
Support to Hearing, Speaking and Seeing
Ifukube, T.

2017, XII, 239 p. 196 illus., 11 illus. in color., Hardcover
ISBN: 978-3-319-47996-5