Sound-based Assistive Technology Supporting "Seeing", "Hearing" and "Speaking" for the Disabled and the Elderly

Tohru Ifukube[1]

[1] Research Institute for Advanced Science and Technology, University of Tokyo, Japan

ifukube@rcast.u-tokyo.ac.jp

Abstract

With a rapid increase of a population rate of the elderly, disabled people also have been increasing in Japan. Over a period of 40 years, author has developed a basic research approach of assistive technology, especially for people with seeing, hearing, and speaking disorders. Although some of the required tools have been practically used for the disabled in Japan, the author has experienced how insufficient a function of the tools is for supporting them. Moreover, the author has been impressed by amazingly potential ability of the human brain that compensates for the disorders.

In this report, the author shows some compensation abilities formed by "brain plasticity", and also shows extraordinary ability such as voice imitation of mynah bird and obstacle sense of the blind. Furthermore, the author introduces six assistive tools borne by solving mystery of the compensation function and the extraordinary animal. Finally, the author emphasizes that these assistive tools will contribute to design a new human interface that may support the elderly as well as the disabled.

Index Terms: assistive technology, the blind, the deaf, speech disorders

1. Introduction

As shown in figure 1, nowadays, around 1/4 of population are elderly in Japan, so actually near 30 millions people are more than 65 years old. With the rapid increase of the elderly, disabled people who need care are also increasing. Actually, in 2005, the number of the disabled was beyond 6 millions. Most of them have disability regarding hearing, speaking, reading, thinking and moving functions.

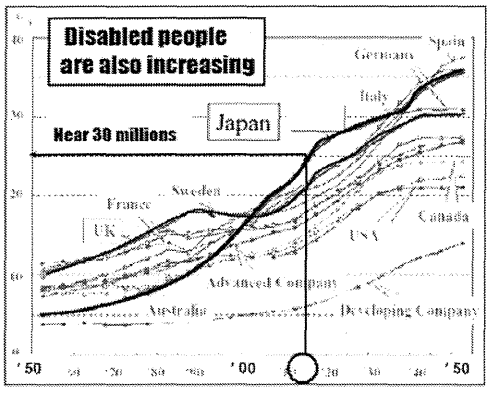

Figure 1: *Population rate of the elderly from 1950 to 2050.*

In this situation, assistive technologies should be urgently designed for supporting these people in addition to medical side technologies. However, the disability is so diverse and complex that it has been difficult to construct a research approach of the assistive technology.

Our research approach has three steps as shown in figure 2. The first step is to analyze the human function such as sensory, brain and motor functions based on neuroscience and cognitive science. The second step is to design various assistive tools based on the findings obtained by the basic research. If the assistive tools are insufficient for supporting the disabled, the research should go back to the first step. The third is to apply the basic findings and the assistive tools to human interface systems such as robotics and virtual reality systems. The third step is important to open a big market and to make a price of the tools cheap [1], [2].

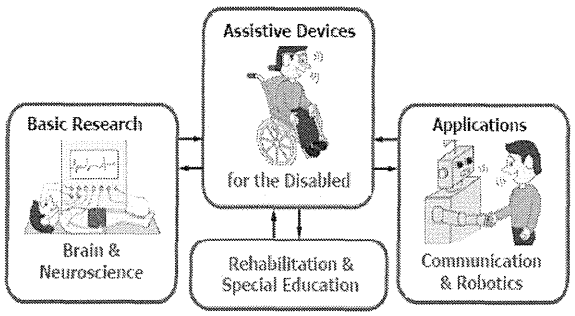

Figure 2: *Three steps of our assistive technology research.*

The assistive technology for elderly disabled people is mainly called "Geron-technology" that is somewhat different from "Barrier-free design" that includes supporting young disabled people. As shown in figure 3, the barrier-free design for the young should positively use "plasticity" of the human body, especially the brain because residual functions work to compensate the disordered function by the help of brain plasticity.

On the other hand, in general, the plasticity function decreases in the elderly, whereas, they will acquire abilities using their "experience". The acquired experiences should positively be used in the geron-technology. The following examples of six assistive tools, which we have designed, were mainly designed for the young disabled.

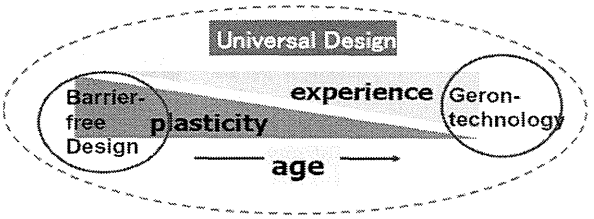

Figure 3: *Barrier-free design and Geron-technology.*

2. Assistive tools supporting "Hearing"

To assist "hearing" for auditory disorders, there are three approaches as shown in figure 4. One is to convert sound signals into tactile stimulations, the second is to electrically stimulate the surviving auditory nerves, and the third is to convert speech signals into letters. Two examples of assistive tools for hearing impairments are shown in this section.

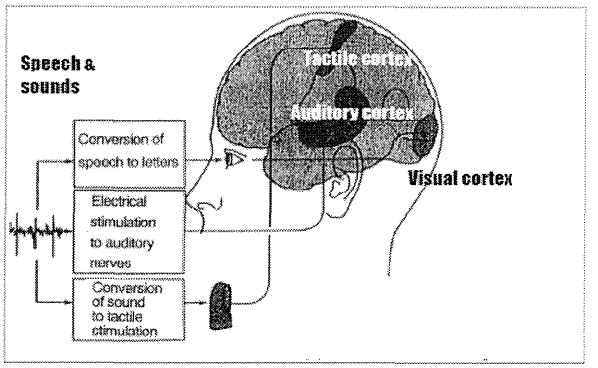

Figure 4: *Three research approaches of assistive tool design for the hearing impaired.*

2.1. Tactile communication aids for the deaf and/or the blind

The first topic is an assistive tool called "tactile aid" that is the author's first research started from 1972. Tactile information is conbineant for suppoting hearing as well as seeing because it can be used together with the visual and/or hearing sense. Through fundamental research regarding similarities between auditory and tactile information processing, we developed a prototype tactile aid named "tactile vocoder", as shown in figure 5, in 1975.

The tactile vocoder is a device which produces sound spectral patterns of which frequency analysis method is modeled after cochlear mechanism. In the tactile vocoder, spectral patterns of speech sounds were divided into 16 frequency components and each component was transmitted to lateral inhibition circuit in order to sharpen the spectral patterns. Finally, the 16 components were converted into 16 vibratory stimuli of which level of the is corresponding to each intensity of the 16 components.

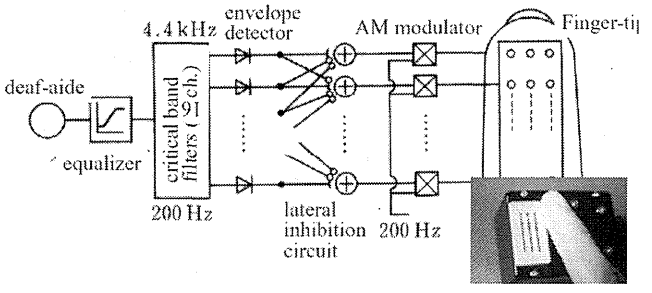

Figure 5: *Block-diagram of tactile vocoder and a vibrator array (3x16).*

When an index fingertip touches a vibrator array that was composed of 16 rows and 3 columns, the device made it possible to discriminate the first and the second formants of vowels as well as some consonants such as fricatives, semi-vowels and nasals [3].

The device was used for a telephone communication aid at a deaf school after deaf boys learnt tactile patterns corrresponding to several words such as "hello", "yes", "no", and etc. Although the device was manufactured in Japan about 30 years ago, it has not been widely used because nobody knows whether or not the tactile information is associated with speech understanding area in the brain.

After 25 years ago of the tactile aid research, the author moved to the University of Tokyo to attend at barrier-free project. The project has been conducted by a professor Satoru Fukushima who became blind at the age of 9 years old and deaf at 18 years old. From a viewpoint of the deaf-blind, he has suggested how we should perform the barrier-free research. He ordinarily communicates with us using both the tactile sense of his 6 fingers by a help of interpreters as shown in figure 6(lower left). This tactile communication method is called "finger Braille" that he first investigated in the world.

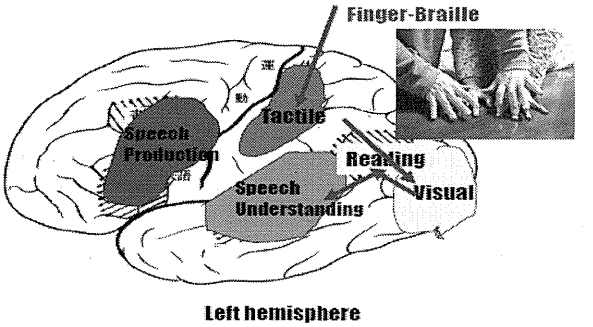

Figure 6: *Finger-Brail information routs acquired by "plasticity" of the human brain.*

From recent neuroscience researches, it is anticipated that the finger Braille patterns would be reached to the visual cortex through the tactile cortex and then the patterns would be associated with language understanding cortex as shown in figure 6. By the "plasticity" of neural network in the human brain, the lost functions might be compensated by the other sensory cortex. Actually, in 1998, by using a functional MRI, Swedish brain researchers showed that tactile stimulation activates the auditory cortex of the acquired deaf [4]. Furthermore, it has also been found that the visual cortex is activated by Braille stimulation on a fingertip of acquired blind subjects [5].

Encouraged by a professor Fukushima and the neuro-science results, we re-designed a new model of the tactile vocoder in 2006.

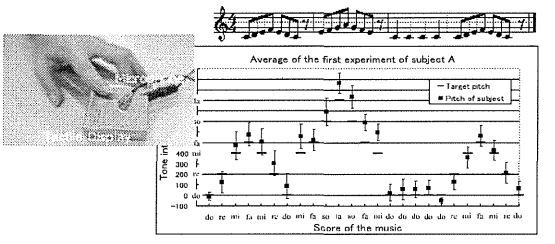

Figure 7: *Left: New tactile vocoder made of a microprocessor and a vibrator array. Right: Pitch pattern of a deaf-blind' produced by a "Frog's song"*

The new device consisted of a DSP and a piezo-electric vibrator array so that the signal processing can be changed only by software. For the deaf-blind use, the new device was set to transform a voice pitch frequency into musical scale that corresponds to vibrating point of a fingertip.
A lady, who has been deaf-blind since she was 40 years old and is now 67 years old, attended at our experiment. As she was a teacher of Japanese musical instrument and folk songs until she lost her visual and auditory senses, we expected that she can easily handle the tactile pitch display and can sing songs by using a feedback of the songs through her fingertip. Actually, after she leant musical scale using the tactile device for short time, she could sing some Japanese songs as shown figure 6. This example shows that the tactile information might help to get a feedback of melodies of the songs [6]. However, congenital deaf-blind people were not able to sing any song.
These findings give us many suggestions to analyze whether or not the tactile information could be transmitted to auditory cortex and speech understanding area in the brain. It is ascertained how brain and cognitive researches are important to design assistive tools.

2.2 Captioning system for the hearing impaired

The second example is a voice recognition system to assist listening of the acquired deaf who have lost their hearing when they became adult. We were asked to design the voice typewriter by the acquired deaf group in 1975.
Fortunately, all Japanese voices are represented as a series of 68 monosyllables (/a/, /ka/, /sa/, /ta/, /na/.....) each of which consists of 5 vowels and 14 consonants and represented by one of Japanese Kana. This simplicity makes easy to design a Japanese voice typewriter that converts each monosyllable into a corresponding Kana.

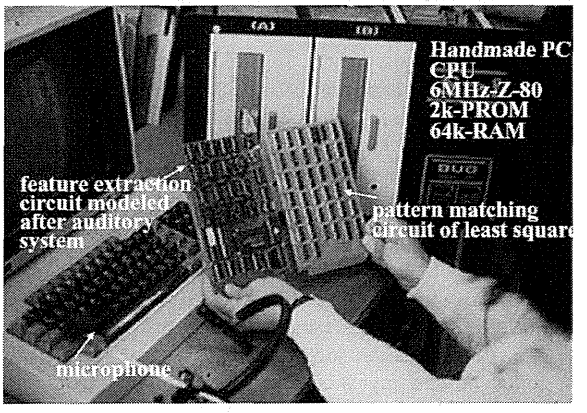

Figure 8: *Monosyllabic voice typewriter using a Z80 microprocessor and 32kbyte memories.*

However, Japanese sentences are represented by combining Japanese Kana and Chinese characters that have different pronunciations. This complexity makes difficult to design a Japanese voice typewriter.
In 1976, we designed a monosyllabic voice typewriter system using a Z80 microprocessor and 32kbytes IC memories, as shown in figure 8, which was used together with a Japanese word processor so that Japanese monosyllabic voices are automatically converted into both Japanese and Chinese letters.
Our voice typewriter could recognize separately pronounced 68 Japanese monosyllabic voices into Kana at a correct recognition rate of 96% in the case of well trained speakers. However, as the voice typewriter system was too expensive for deaf users to buy, it was only used for an input device of printing machines in 1977 **[7]**, not for an assistive tool for the deaf.
In 2001, 25 years after of the voice typewriter research, we were asked to design a captioning system by DPI (Disabled Peoples' International) conference held in 2002. As about 3000 disabled people from more than 100 countries attend at the conference, we had to design the captioning system that can automatically convert various languages into both Japanese and English captions. As the speech recognition technology was not perfect, we expected an ability that the hearing impaired can often guess the correct meaning of spoken sentences by observing a speaker's face and lip-movement.
With this in mind, we designed the captioning system in such a way that both the series of letters and the speaker's face simultaneously appear on a large screen. Where, we used a commercially available speech recognition software "Via Voice". Furthermore, we adopted "re-speak" method that the speaker's speeches are sent to a well trained re-speaker who repeats the speeches. In our system, the recognized outputs were checked by humans and then the corrected letters were displayed on a screen. The system was first used at a pre-conference in 2001 and then at the DPI in 2002. Both Japanese and International sign languages, both Japanese and English captions as well as speaker's face were displayed on a same screen as shown in figure 9.

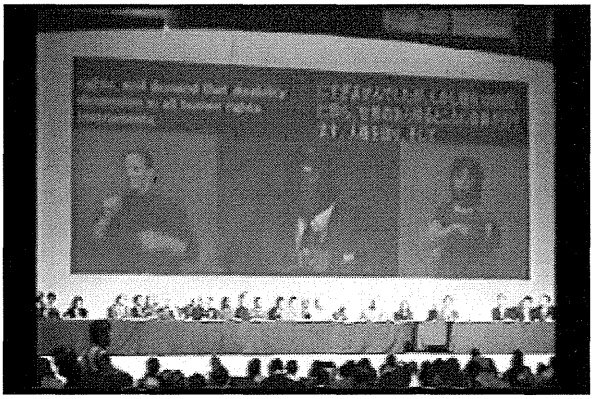

Figure 9: *Captions (English and Japanese) presented on a screen together with sign languages and a speaker's face at DPI conference in 2002.*

From the analysis of captioning results, the accuracy of the caption was about 98% and captioning speeds were 4 sec from English to English, 11 sec from Japanese to Japanese, 12 sec from Japanese to English and 17 sec from English to Japanese.
After the use at DPI, we investigated whether or not such non-verbal information as a speaker's lip-movements and facial expressions as shown in figure 10 could improve the comprehension of spoken sentences that contain incorrect words.
From the results of the speech comprehension, the improvement was observed only when the facial expression was displayed roughly one second after the incomplete presentation of a speaker's words. This fact was indirectly

proven by an experimental result that eye balls of the deaf subject mainly stayed on the captions, not on the speaker's face [8].

Figure 10: *left: Caption only, center: Caption + face, right: Caption + lip-movement.*

No improvement was obtained for people with normal hearing. It is ascertained that combining both incomplete verbal with non-verbal information is indeed significant in facilitating comprehension for the hearing impaired .
After the DPI, the caption system was used in the various conferences in Japan through internet (ISDN). For example, at International Conference of Universal Design held at Yokohama in 2002, both English or Japanese speeches picked up at Yokohama were simultaneously interpreted, and then the interpreted speeches were sent to Sapporo where the speeches were converted into both Japanese and English text information which were displayed on screens at Yokohama as shown in figure 11.

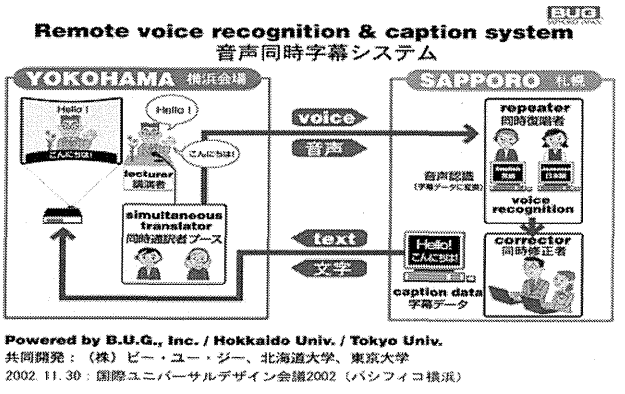

Figure 11: *Captioning system and mobile phone (lower right) using internet*

Furthermore, we tried to apply the captioning system for communication aid for mobile phone in 2004 in order to be used by general people. These experiences suggest that the assistive tools should be widely used as possible as we can not only for deaf users but also for general users in order to make the system cheap and better model.

3. Assistive tools for speech disorders

Speech disorders are roughly divided into three causes shown in figure 12. The first is a caused mainly by laryngeal cancer. The laryngeal cancer patients have sometime lost their larynx by surgical operation, losing their vocal folds. Therefore, they become "laryngectomee" who lost a sound source for speech. The second is articulation disorders who are difficult to control their speech organ such as tongue, jaw and lips. The third is an aphasia that is caused by a disorder of brain nervous system in Broca area in the cortex.

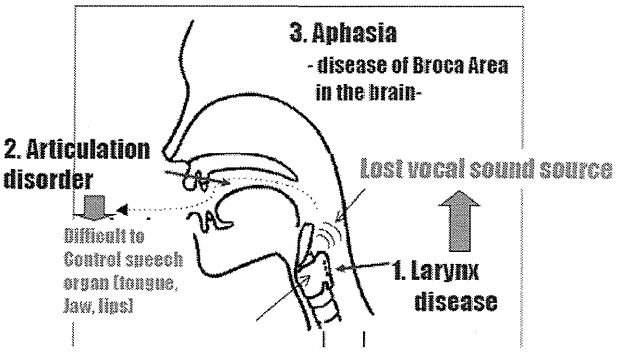

Figure 12: *Three causes of speech disorders*

3.1. Artificial electro-larynx for laryngectomee

An electronic artificial larynx is one of substitute vocalization methods. This method applies a vibrator to the lower jaw, which sends a vibrating sound into a voice tract through soft subcutaneous tissue. We focused on the electro-larynx because this method can be easily learned, although the voice quality is very poor like a buzzer sound.

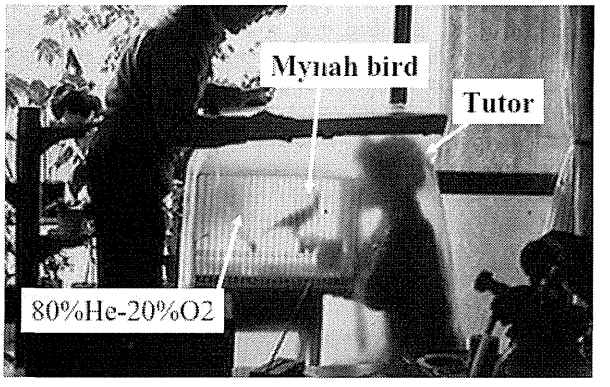

Figure 13: *Mynah bird and her tutor in He-O2 gas*

The ability to generate voice sounds is not limited to human beings. Among the animals which can generate voice sounds are certain birds, including mynahs, parrots, and parakeets. For researchers of speech assistive technology, it is a mystery as to why these birds, whose mouths and ears are shaped in a totally different way from humans, are capable of distinguishing and vocalizing certain human words. Especially the imitations of a mynah's voice sounds quite smooth and natural. Therefore, a research of the vocalization mechanism of mynahs might lead to new ideas for improvements of the electro-larynx.
From an analytical study of the mynah bird's voice imitations using He-O2 gas, about 30 years ago as shown in figure 13, we found that the mynah's imitations in the He-O2 gas are heard as the same as in normal air. Since it was reported that mynah bird has two sound sources (syrinx) corresponds to a human vocal fold, it was ascertained that he produces speech signals by combining the two sounds sources. Therefore, a formant pattern of the mynah was quite different from the human's formant pattern. We also found the reason why the mimicking of the mynah can be heard as natural voices is because she is able to faithfully imitate the fluctuation and intonation of the human voice [9].

After completing this basic research regarding similarity between the human voice and the mynah imitation, we designed a new electro-larynx composed of an air flow sensor, a small computer and a vibrator. An important role in the creation of voice sounds is played by the flow of air from the lungs to the larynx, determining a metrics such as a sound level, an accent and an intonation. In our electro-larynx, the intonation can be controlled by the exhalation flow detected by the air flow sensor as shown in figure 14 [10].

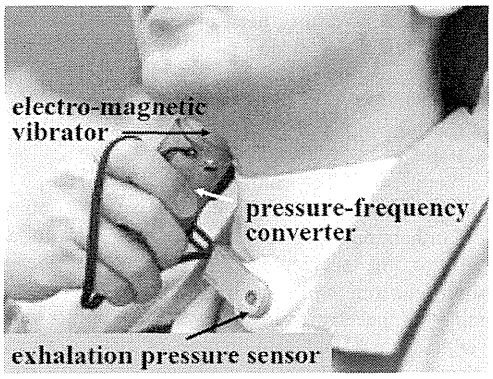

Figure.14: *Pitch controlled electro-larynx*

The electro-larynx voices can be perceived like the human voices with the intonation although the formant structure obtained by the voices is very different from the human's voices. The electro-larynx with pitch-controlled function has been manufactured and more than 4,000 patients have been using the device since 1998 in Japan.
However, since this method requires the use of a hand, it is still somewhat restricting. Many patients have strongly requested us to design a hands-free electro-larynx so that they can use it in daily life and in their office without their hands bound by the control of the electro-larynx while using it.

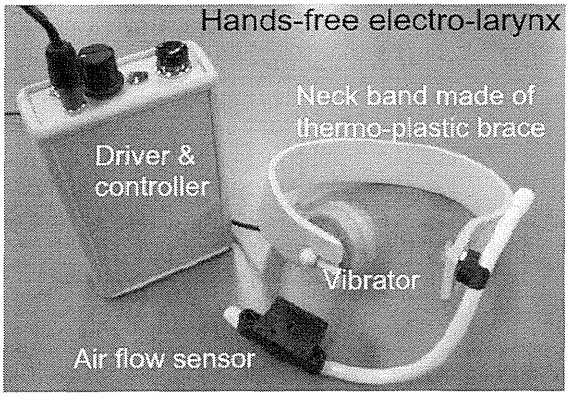

Figure 15: *Hands-free pitch controlled electro-larynx*

Furthermore, the electro-larynx sound is still not human-sounding voice. It was supposed that fluctuations of speech waves are needed to add to the electro-larynx voice.
In order to answer their request, we have designed a hands-free electro-larynx that can be attached to patient's neck. By comparing the various materials for the neck band, a thermo-plastic brace was selected. It becomes soft about at a temperature of 70 degrees centigrade while being hard at a body temperature so that it is suitable for adjusting individual differences of a shape and hardness of the neck tissue.
We made a smaller vibrator (32mm diameter × 19mm thick and 32g weight) and a much more sensible flow sensor than the conventional one. The vibrator and the sensor can be mounted to the thermo-plastic brace. Figure 15 shows the prototype of the hands-free electro-larynx worn by a patient.
From usability tests, it was confirmed that the electro-larynx voices are heard more naturally than the conventional one and that the activity of the patients significantly improves in daily life. Furthermore, by adding a 1/f fluctuation of the pitch frequency [11] to the electro-larynx, the voice quality apparently exceeded the mynah bird's imitation [12].

3.2. Voice synthesizer for the aid of articulation disorders

In order to assist articulation disorders, we designed a speech synthesizer that can be handled by a finger as if a user is playing a musical instrument. One of students, who is difficult to control his speech organ because he has been suffering from muscle dystrophy, designed the synthesizer for himself. He has been disappointed in ordinary speech aids that do not work in real time or can not produce non-verbal information such as intonation and emotional expressions.
Our method was modeled after a speech production mechanism of a parakeet and a ventriloquist that can produce normal speech without moving their mouth. By analyzing their speech production, it was found that they produce speech sounds only by rapidly moving tongue inside the mouth.
This fact means that they can produce bilabial consonants such as /pa/, /ba/, and /ma/ by rapidly moving the tongue without closing their lips. As an analytical result of the ventriloquism vocalization mechanism, it was hypothesized that the ventriloquist quickly moves his/her tongue so that /a/ of /pa/ can be produced after putting the tongue at a front tooth; an articulation point of a consonant /t/, in the case of ventriloquist's /pa/ as shown in figure 16. Human can hear only /pa/ by this tongue movement even if the sound of the consonant /t/ exists. Almost the same mechanism as /pa/ was ascertained in the case of /ba/ and /ma/. This result means that /pa/, /ba/ and /ma/ can be heard only by the tongue movement inside a mouth.

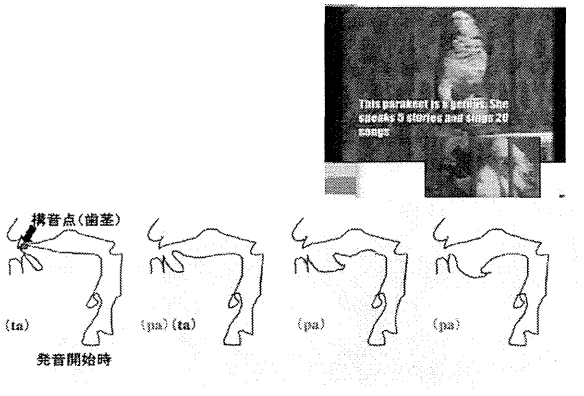

Figure 16: *Parakeet has a flexible and thick tongue (upper), so that she can produce various vowel formants by changing resonant frequencies. Ventriloquist quickly moves a tongue so that /a/ of /pa/ can be produced after putting the tongue at an articulation point of /t/ (lower).*

An idea of our speech synthesizer for the disorders is mainly based on a fact that human recognizes the bilabial consonants produced by the ventriloquist. Parameters of first formant frequency (F1) and the second formant frequency (F2) of a formant synthesis software were controlled by a position and a motion of a user's index finger without any key input.
A touch pad was adopted for the input device that detects the fingertip's position and movement. The F1 and F2 that correspond to the position of the tongue were two dimensionally assigned to the plane of the touch pad as shown in figure 17.

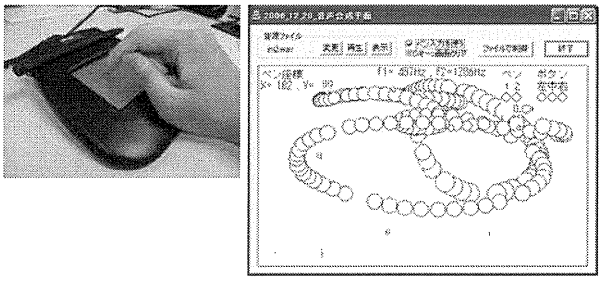

Figure: 17 *A touch pad (left) where F1 and F2 frequencies were two dimensionally assigned to the plane (right).*

A prototype model was designed on a personal computer using formant synthesis software to get the optimal speech synthesis method for the articulation disorders [13]. In our method, in order for the users to easily find the formant transitions to produce the desired consonants only by changing the formant transitions, "Guiding Lines" were superimposed on the pen-tablet display. Direction and length of each guiding line was assigned at a starting point and a locus of the formant transition on the F1-F2 plane. The starting point and the formant transition locus were decided according to an "Expanded Locus Theory" that we modified based on "Locus Theory" [14].
In the evaluation of sentence recognition, continuous speech sounds were synthsized by tracing the touch pad using the index finger as shown in figures 17 (left). After a few hours training to produce some Japanese sentences such as /ohayo-gozaimasu/ (good morning), /kon'nichiwa/ (good afternoon), /kombanwa/ (good evening) and /arigato-gozaimasu/ (thank you), the consonants /ha/, /go/, /za/, /su/, /ko/, /ni/, /chi/, /wa/, and etc. were apparently heard in the continuous sentences although the synthesized voices had no consonant sounds.

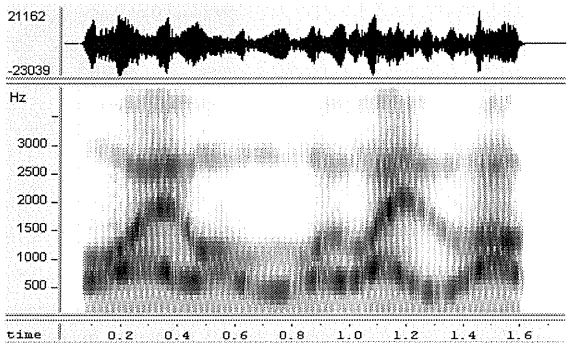

Figure 18: *A sound spectrogram of /ohayo-gozaimasu/ obtained by our synthesizer for the speech disorders.*

Figure 18 shows a sound spectrogram of /ohayo-gozaimasu/. From the evaluation results, it was ascertained that the users may more easily produce the arbitral sentences by tracing the guide lines than conventional one [15].
Next we carried out evaluation test of randomly synthesized 100 words composed of 4 monosyllables. Although the identification of vowels and voiced consonants such as semi-vowels, nasals, showed high score around 70 %, the average identification rate showed only 33 %. It was still difficult to perceive some consonants that need random noise component such as /s/, /h/ and /z/. We have been investigating how the noise component should be added and controlled by user's finger.
By using a touch pad with a pressure sensor and by assigning the detected touching pressure to the pitch frequency, the subject could produce some emotional voices such as "laughing", "surprising" and "disappointment" by controlling the pressure of the touch and the speed of finger movement. Furthermore, by connecting this tool with a musical key-board, the subject could sing some Japanese songs after short time training.
We are planning the speech production method to apply a tool for people with aphasia and also to one of new musical instruments for general users.

4. Assistive tools supporting "Seeing"

There are three approaches for visual substitutes as shown in figure 19. One is to convert visual information into tactile stimulations, the second is to electrically stimulate the surviving visual nerves, and the third is to convert text to speech. In this session, two examples to assist the visually disabled are mentioned.

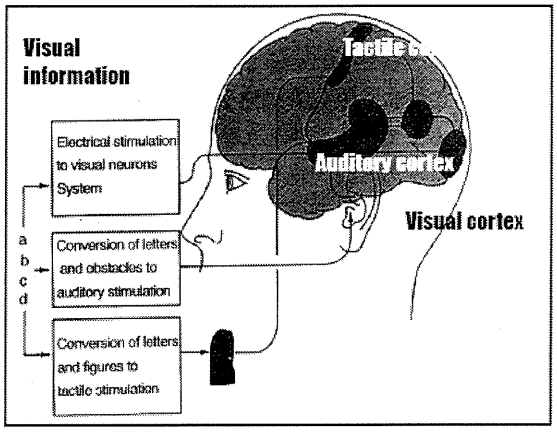

Figure 19: *Three approaches supporting "seeing".*

4.1. Screen reader for the blind

In 1996, we developed a screen reader, called '95-reader', which can convert text information into synthesized voices for Japanese visually impaired people as shown in figure 20 [16]. However, operating systems based on graphical user interfaces (GUIs) have spread rapidly in personal computer environments. Consequently, it is inherently difficult for visually impaired people to use the screen reader, since they are often unable to detect the necessary visual information.
With this problem in mind, we have proposed a tactile jog-dial (TAJODA) which can control the speech rates of the screen reader while displaying rich text features, such as various

fonts, letter sizes, and bold script onto the tactile sense of a fingertip.

For the design of the TAJODA, first, maximum listening tests were performed on blind users in order to investigate how many morae per minute the blind can correctly recall the presented sentences. From the experimental results obtained by about 20 subjects, it was found that the maximum recall rate for experienced users was about 1,400 morae per minute, which is 2.6 times faster than the average listening rate (550 morae per minute) of the sighted, as shown in figure 21 [17].

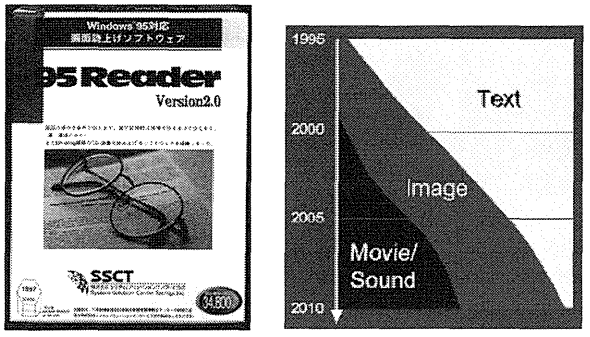

Figure 20: *Screen reader software (left) and increase of graphical user interface (right).*

ext, based on the psychophysical experiments, we determined what kind of rich texts should be presented as tactile patterns. As a result, we have determined some parameters of the prototype model that can control speech rates in real time by use of a finger while touching the rich texts with another finger, as shown in figure 22 [18].

The use of the TAJODA was well accepted by all of the blind subjects, since it allowed them to actively and flexibly navigate through the speech information by dynamically controlling the speech rate. The tactile cues also helped them to look for particular information within the overall speech information. It has been proven that document recognition rate is 2-3 faster than the conventional screen reader.

In a recent model that has been commercially available in Japan, the speech rate can be changed according to monosyllabic voice units by clicking upper button for "more quickly" or the lower button for "more slowly" using a thumb finger of the right hand. An index finger can touch the tactile.

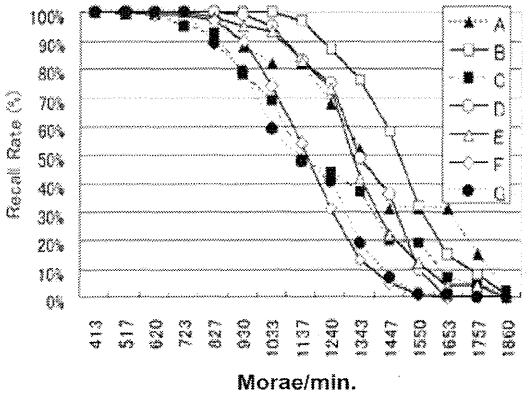

Figure: 21 *Recall rate in percent as a function of speech rate in morae/min.*

display which is arranged in two columns of 8 pins each, using 32 piezo-vibrator pins [19].

One of blind subjects reported "The tactile cues act like guideposts on a road. The road is like a string of speech information and the guideposts are the tactile cues". The subject also reported that the one dimensional speech information felt as if two dimensional documents were presented. This fact anticipates that the visual cortex is activated while using the TAJODA.

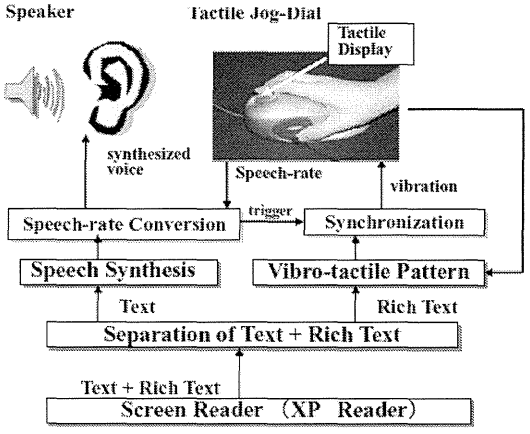

Figure 22: *Block-diagram of the TAJODA interface. Speech rate can be controlled by the thumb. Rich text is displayed as vibro-tactile patterns on an indicating fingertip.*

Furthermore, we have been designing a Tactile Mobile Phone for assisting "seeing", "hearing" and "speaking for the deaf-blind" by combining the tactile aid with the TAJODA as shown in figure 23.

In the mobile phone, image and direction information are sensed by a camera and an electric compass inside the mobile phone, which signals are transmitted to a tactile display. Environment sounds including speech can also be detected by a microphone attached to the mobile phone. Furthermore, since the tactile display is utilized as a tactile matrix sensor, tactile information touched and traced by a finger can be displayed onto a receiver's fingertip through the mobile phone. This function will be useful for the deaf-blind as well as normal users, especially in the condition when the users' eyes and ears are bounded.

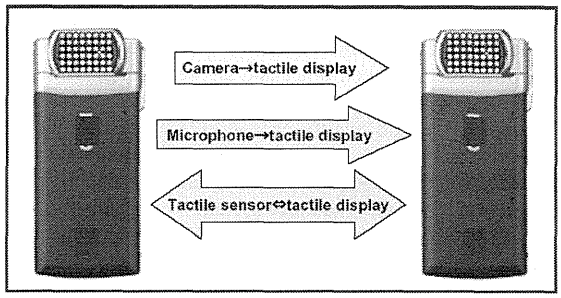

Figure 23: *Tactile Mobile Phone for assisting "seeing", "hearing" and "speaking"*

4.2. Tow-dimensional image display method using

Most visually impaired people can detect obstacles by using an ability known as "obstacle sense". We have been investigating the mechanism of the obstacle sense based on psychophysical experiments using visually impaired people .
By analyzing and modeling the mechanism of the obstacle sense, the resultant model could be utilized in new concepts for image display aids as well as mobility aids for the blind.
From our past research [20], we have ascertained that the reason why they can detect the obstacle and guess the distance is due to their ability to discriminate tiny changes of sound spectral patterns produced by the existing of obstacle. The spectral changes are caused by phase interference between directly reached sound to ears and reflected sounds from the obstacle as shown in figure 24. This spectral change is called "coloration". If environmental noise-like sounds do not exist, the ability of obstacle sense decrease or does not work because of a luck of the phase interference.

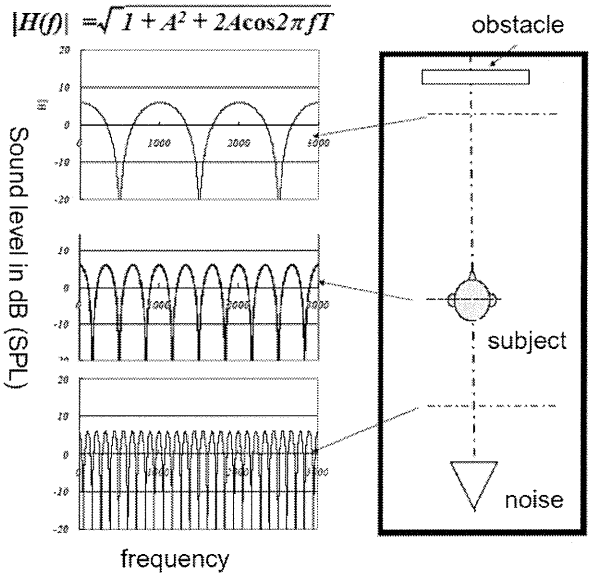

Figure 24: *Spectral changes caused by phase interference between direct sounds and reflected sounds from the obstacle.*

From basic experiments related to obstacle sense by the blind, we found that the discrimination ability of coloration changes is no difference between blind and sighted groups. However, the blind may detect the obstacle by perceiving the coloration difference as "sound quality change" [21]. It means that the obstacle sense would be acquired by their experience. In another word, they acquired the obstacle sense by "brain plasticity" in the central auditory system.
It is supposed that two dimensional sound images displayed using a loud speaker matrix would be more accurately perceived by the blind than by the sighted for a short training time. With this supposition in mind, we tried to convert the images into two dimensional sound images by using a loud-speaker matrix that can produce temporally and spatially controlled sound sources more than 30 years ago as shown [22].
We have investigated again how the moving sound images as shown in figure 25 are perceived by hearing in order to determine the optimal arrangement of the loudspeakers. Although this study is very preliminary, we may obtain some suggestions how image information should be displayed in addition to speech information for the blind [23].

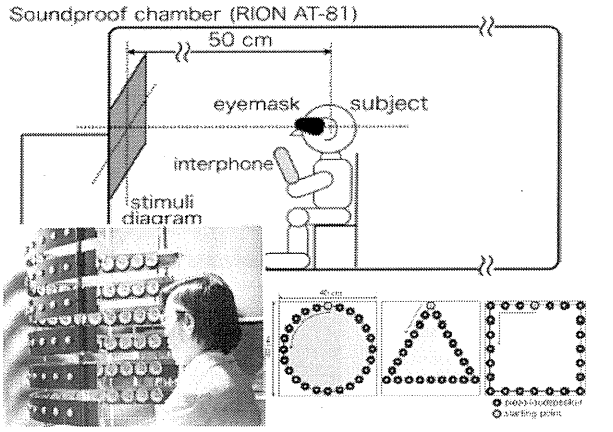

Figure 25: *Speaker array for displaying two dimensional images in addition to speech information*

5. Conclusion

By applying the sound-based assistive tools to the hearing impaired, the visual impaired and the speech disorders, certain basic hypotheses concerning the brain function have been lead and also the tools have lead to the design of human interfaces for general users. Furthermore, the assistive tools designed for the young disabled will be also useful for supporting the elderly with disabilities by modifying them. The research approach mentioned in this report may contribute to open new markets and to benefit a greater number of people.

6. Acknowledgements

The author would like to thank Dr. Tatsuya Hirahara, Dr. Norihiro Uemi, Dr. Chikamune Wada, Dr. Shuichi Ino, Dr. Mitsuo Hshiba, Dr. Yosikazu Seki, Professor Tadayuki Sasaki, Dr. Hayato Kuroki, Dr. Ken'ichi Yabu, Dr. Tetsuya Watanabe, Dr. Chieko Asakawa, Mr. Msanobu Sakajiri, Mr. Takahiro Miura, Dr. Takahiro Ueda and Dr. Satoshi Fukushima who directly involved in this study as well as many engineers helped us to design assistive tools. Their help in running the experiments and designing the devices has been very invaluable. Most fundings were supported by Grant-in-Aid from the Ministry of Education and Science.

7. References

[1] Ifukube, T., "Sound-based Assistive Technology for the Disabled", Corona Publishing Co., (250 pages.in Japanese) (1997)

[2] Ifukube. T., "A Neuroscience-based Design of Intelligent Tools for the Elderly eeeand Disabled," ACM press, New York, 31-36 (2002).

[3] Ifukube, T., Yoshimoto, C., "A sono-tactile deaf-aid made of piezoelectric vibrator array," J. Acoust. Soc. Jpn., 30 (8) 461-462 (1974).

[4] Levanen, S., Jousmaki, V., Hari, R., "Vibration-induced auditory-cortex activation a congenitally deaf adult," Current Biology, 8(15), 869-872 (1998).

[5] Sadato, N., and et al., "Activation of the primary visual cortex by Braille reading in blind subjects," Nature, 380, 526-528 (1996).

[6] Sakajiri, M., Miyoshi, S., Nakamura, K., Fukushima, S and Ifukube, T.," Voice Pitch Control by using Tactile Feedback for the Deafblind or the Hearing Impaired Persons to Assist their Singing", 2010 IEEE Internatuional Conference on Systems, Man, and Cybernetics (in press).

[7] Ifukube, T., "Design of a Voice Typewriter", CQ Publishing Co., (195 pages, in Japanese), (1984)

[8] Nakano, S., Makihara, T., Kanazawa, T., Kuroki, H., Ueda, K., Ino, S. and Ifukube, T., "A Study on the Ease of Real-Time Speech-to-Caption System for the Hearing Impaired(1): The Effect of Line Break," The Transactions of Human Interface Society, Vol.10, No.4, pp.51-60 (2008).(in Japanese)

[9] Hirahara, T., Ifukube, T., and Yoshimoto, C., "An articulation model of a talking bird "Mynah"-An analysis of music voices pronounced in He-O_2 atmosphere-," J.Acoust.Soc.Jpn, 38(6):321–329 (1982) (in Japanese)

[10] Uemi, N., Ifukube, T., Design of a new electro larynx having a pitch control function, in: IEEE Inter. Workshop on Robot and Human Communication, 198-203 (1994) (in Japanese)

[11] Aoki, N. and Ifukube, T., "Analysis and perception of spectral 1/f characteristics of amplitude and period fluctuations in normal sustained vowels," J.Acoust.Soc.Am, 106(1) (1999).

[12] Hashiba, M., Sugai, Y., Izumi, T., Ino, S. and Ifukube, T., "Development of a hands-free electro larynx for persons whose larynx have .been removed", J.Acoust. Soc. Amer. 120(5), p.3351 (2006)

[13] Yabu, k., Aonuma, S., Ifukube, T., "A speech synthesis device for voice disorders.- Its research approach and design concept – " IEICE Technical Report, 106(613), 25-30 (2007)

[14] Pierre Delattre, Franklin S. Cooper, Alvin M.Liberman, and Louis Gerstman:Acoustic Loci and Transitional Cues for Consonants; The Journal of the AcousticalSociety of America 26(1), p137 (1954)

[15] Yabu, k., Aonuma, S., Ifukube, T., "Proposal of a speech–synthesis interface for speech disorders by using a pointing device, The Trans. Of Human Interface Sciety, 11(4), 135-146 (2009) (in Japanese)

[16] Watanabe, T., Okada, S. And Ifukube, T., "Development of a GUI screen reader for blind persons, Systems and Computers in Japan. 29, 13, 18-27 (1998) (in Japanese).

[17] Asakawa, C., Takagi, H., Ino, S. and Ifukube. T., "Maximum listening speeds for the blind," Proc. ICAD, 2003.

[18] Asakawa, C., Takagi, H., Ino, S. and Ifukube. T., "TAJODA: Proposed Tactile and Jog-Dial Interface for the Blind," IEICE Trans. on Information and Systems, Vol.E87-D, No.6, pp.1045-1014 (2004).

[19] Ifukube,T. and Kimura, S., "Screen Reader Interface Dynamically Hastens Speech while Giving Emphasized Information to the Tactile Sense,"4th Joint Meeting of The Acoust. Soc. Am. and the Acoust. Soc. Jpn (2006).

[20] Seki, Y. amd Ifukube, T., "Relation between the reflected sound localization and the obstacle sense of the blind," J.Acoust.Soc.Jpn, 50(4):289–295, (1994) (in Japanese)

[21] T. Miura, T. Muraoka, and T. Ifukube. "Comparison ofobstacle sense ability between the blind and the sighted:A basic psychophysical study for designs of acoustic assistivedevices," Acoust. Sci. Tech., 31(2):137–147, 2010.

[22] Ifukube, T., Sasaki, T. and Peng, C., A blind mobility aid modeled after echolocation of bats, in: IEEE Trans., BME38(5) 461-465 (1991)

[23] Suzuki, ., Miura, ., Tsuchiya, ., Ueda, K., and Ifukube, T., "Design of a loudspeaker-matrix which presents two-dimensional patterns: Perception of sound-image trajectory", T. SICE Vol.43 No.3 in press (2010) (in Japanese)

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2010-96,SP2010-52,WIT2010-40(2010-10)

発話障害者支援のための連続タッチ平面で操作する音声生成器
― 子音改善のための基礎的検討 ―

藪 謙一郎† 伊福部 達†

†東京大学先端科学技術研究センター 〒153-8904 東京都目黒区駒場 4-6-1
E-mail: †{yabu,ifukube}@human.rcast.u-tokyo.ac.jp

あらまし 我々は、発話障害者支援のために、連続なタッチ平面上をペンや指でなぞることで楽器のように音声をリアルタイムに生成する音声生成器を提案し、試作器の開発をしている。これにより、自由な間(ま)やリズムなどの非言語的な情報を自由に表現可能となることが期待できる。先の研究で、試作器による実験から、2 つのホルマント周波数をペンや指の位置や動きの入力によって制御するだけで、不明瞭ながらも簡単な単語や子音に近い音声を出せることを示してきた。本稿ではさらにその明瞭度改善のために、摩擦音などに必要な雑音成分を、同様のペン操作で付加できるようにして、聴取実験を行いその効果を確認したので報告する。また、了解度が低かった単語について原因を考察し述べる。

キーワード ホルマント, 子音, 音声合成, 発話障害者

A speech synthesis device for speech disorders controlled by continuous plane touching.
—Basis examination for an improvement of intelligibilities of consonants.—

Ken-ichiro Yabu† Tohru Ifukube†

†Research Center for Advanced Science and Technology, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8904 Japan
E-mail: †{yabu,ifukube}@human.rcast.u-tokyo.ac.jp

Abstract We have proposed a speech production method in which the voice synthesis was controlled by a position and a motion of a pen or a finger put on a pointing device in real-time for speech disorders, and have developed it. Nonverbal information like rhythms and intervals of voices can be freely controlled by using this method. Our previous study with a prototyping model which we developed showed that voices that sound like consonants were produced by controlling two formant frequencies by a movement of a pen or a finger. In this paper, we described about an experiment with a improved production method for consonants, and discussed about problems.

Keyword Formant, Consonant, Voice Synthesis, Articulation Disorder.

1. はじめに

発話音声はヒトが日常生活を送る上で、最も基本的で簡便なコミュニケーション手段である。その内容は、言語情報の伝達以外にも、抑揚や間(ま)、話すテンポやリズムなどの表現によって、様々な情報を伝える役割を持っており、文字や単語の記述だけでは表現できない非言語的な情報が多く含まれている。また、このような非言語的な情報は、日常の中で他者との関係を円滑に送るために大変重要な役割を持っている。

既存の、文字や単語の入力から音声を出力する発話障害者を支援する機器[1], [2]は、発話障害者が自身の意思を他者に伝える上で非常に強力な手段である。しかし、文字入力や特定のボタンからの入力から話す内容を入力するという制約から、数秒以下の単位での即時の発話や、あいまいな発話や、自由なリズムの表出に困難さが残る。

そこで我々は、文字や単語に頼らずに、指やペンで面をなぞる動きで、楽器のように使用者自身がリアルタイムに音声をコントロールしながら音声を生成できる音声生成器を提案している[3]。

具体的には、ヒトの母音知覚に必要な第 1 ホルマントと第 2 ホルマントの周波数(以下、F1,F2)[4]を、音声生成器の操作面の縦軸と横軸とに割り当て、利用者が指やペンを操作面に置いた瞬間に、そのホルマント周波数を持った音を発するというものである。この方法では、操作面に指やペンを置いた瞬間に声が出ることから、入力に伴う時間遅れが殆ど無く、自由なリズム

をつけることが可能である。また、母音と母音の間の中間的な音を連続的に変化させて出すことも可能である。

我々は先の研究で、汎用パソコン上で動作する音声生成器の試作器を用いて、「おはよう」「こんにちは」などの簡単な言葉であれば母音部分の F1、F2 を指やペンの操作で再現するだけで表現可能であることを示した[5]。さらに、ヒトの発話音声中の子音開始部分に存在するホルマントの急激な変化(ホルマント遷移)に着目して、これを再現するための「導線」を提案した[6]。

しかし、文脈のない 100 語の 4 モーラ日本語音声の生成実験と聴取実験では、「導線」によって改善が見られたものの、子音の了解度の点では課題が残った。

そこで本稿では、とくに摩擦音や破擦音などの子音の了解度改善を目的として、子音部の雑音成分を自動付加するように試作器を改良したので報告する。その際の操作方法は、新しいボタンスイッチ等を加えずに、これまでの「導線」を用いた操作方法とほぼ同じになるようにした。

本稿では、まず我々が開発している音声生成器の原理と概要を述べた後、新しく加えた擬似子音の付加の原理と、その操作方法を述べる。そして、擬似子音の影響を見るための聴取実験について報告し、考察と展望を述べる。

2. 音声生成器の概要

2.1. 音声生成処理部

図 1 に、本研究の音声生成処理のブロック図を示す。音声の生成処理は大別すると、声帯振動による音を再現する母音処理部と、摩擦音・破裂音・破擦音などに含まれる雑音成分(以下、乱流音と呼ぶ)を処理する乱流音処理部にわかれる。図 1 の上部が母音処理部、下部が乱流音処理部で、ディジタル共振器と反共振器が四角で示されている。

母音生成部では、原音となる音声波形の信号が、各ホルマントに対応する Fv1 から Fv4 のディジタル共振フィルタを通って特徴付けられた後、出力される。フィルタのうち、Fv1 と Fv2 の特性が、使用者の操作面からの入力によってリアルタイムに制御される。その他のフィルタの特性は固定とした。また高い周波数成分を強調するために、緩やかなハイパスフィルタを設けてあり、図中の HPF と示している。

母音生成部の原音波形には、声の振幅やピッチのゆらぎ**エラー! 参照元が見つかりません。**を再現するため、筆者の「あー」という約 3 秒間の声をサンプリング周波数 16kHz で録音し、30 次の LPC 逆フィルタ処理で補正した音声を用いていている。

また、後述の擬似子音の付加では、乱流音処理部を用いた。乱流音処理部は、音源として C 言語の乱数発

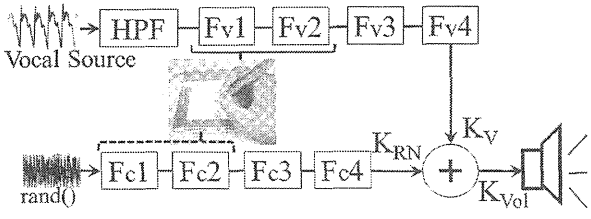

図 1 音声生成器のブロック図

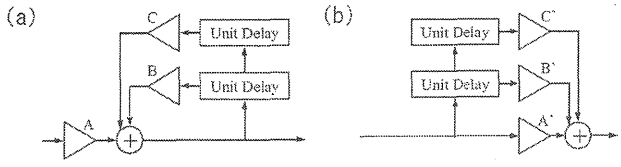

図 2 使用したディジタル共振器(a)と反共振器(b)

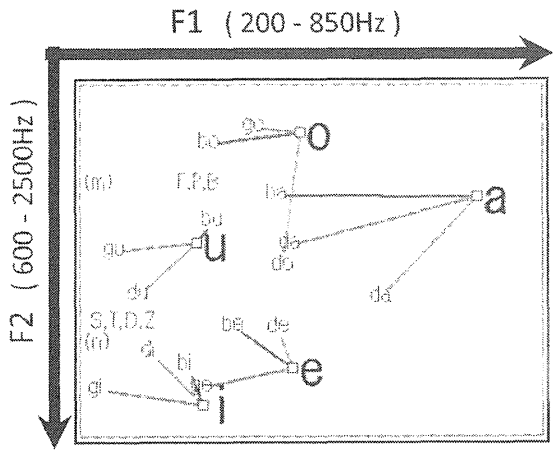

図 3 子音生成のための導線

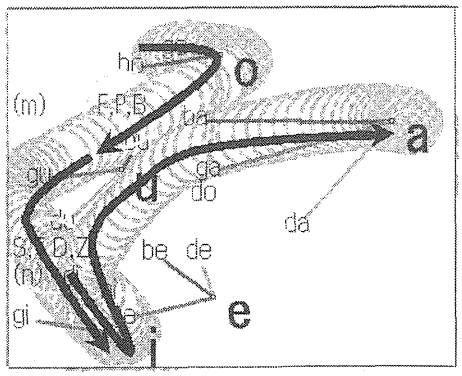

図 4「こんにちは」生成時のペンの軌跡

生関数によって生成された雑音を使用した以外は、母音処理部とほぼ同じである。

共振・反共振器フィルタについては、Klatt らと同様のフィルタ[8]を使用した(図 2)。HPF には、反共振フィルタの反共振周波数をゼロとしたものを用いた。

サンプリング周期を T の共振フィルタの出力 $y(nT)$ は、入力 $x(nT)$に対して式(1)で表される。A,B,C は、共振周波数 F とバンド幅 Bw から式(2)で求められる係数である。

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T) \quad (1)$$

$$\begin{cases} C = -\exp(-2\pi B_w T), \\ B = 2\exp(-\pi B_w T)\cos(2\pi F T), \\ A = 1 - B - C \end{cases} \quad \cdots\cdots\cdots\cdots \quad (2)$$

反共振器は式(3)で示され、係数は式(4)によって求められる。

$$y(nT) = A'x(nT) + B'x(nT - T) + C'x(nT - 2T) \quad (3)$$

$$\begin{cases} C' = -C/A \\ B' = -B/A \\ A' = 1/A \end{cases} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4)$$

2.2. 音声生成器の入力部

2.2.1. 母音操作面

我々が提案している音声生成器の操作面を図 3 に示す。ペンタブレットの縦軸と横軸とを XY 座標とし、X 座標(横軸) に F1 の 200～850Hz を、Y 座標(縦軸) に F2 の 600～2500Hz を対応させ、任意の位置にペンを置くとその座標に対応する母音が生成され、ペンを離すと母音が停止する。操作面上には、“ a ”,“ i ”,“ u ”,“ e ”,“ o ”という文字を各母音のホルマント周波数に対応した位置へ表示させた。図 3 は、ペンタブレットの操作面に示すと同時に、コンピュータの画面上にも、同様のものを表示させた。また、ペンの筆圧を音声の強度に対応させ、筆圧が強ければ音が大きく、低ければ音が小さくなるようにした。ペンの位置は、画面上の対応する位置に小円形でリアルタイムに表示させ、筆圧を円の大きさで表示させた。

通常、ヒトがペンを扱う場合、机などの面に手を置いた状態で作業することが多く、その状態が最も楽にかつ正確に操作できると考えられる。そこで、手を面に固定した状態で操作面全体をタッチできるような、小型のペンタブレット(図 4、表 1)を採用した。

また、ペンの位置が母音の範囲から大きく外れると異音が生じるため、図 4 に示すように、所定の範囲を切り取った厚さ約 0.4mm の PET 樹脂を、貼りつけ、ペンが範囲外にでないようにするガードとした。

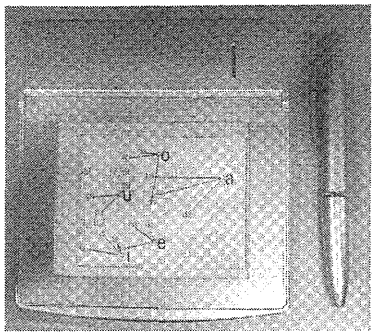

図 5 導線とガード取り付け後のペンタブレット
(P-Active XP-3300A)

表 1 使用したペンタブレットの仕様[9]

型名	XP-3300A
読取方式	電磁誘導
読取範囲	82.5×63.5mm
読取分解能	0.025mm
読取高さ	10mm
読取速度	100 ポイント/秒以上
筆圧レベル	1024 レベル
外形寸法	143(D)×155(W)×5(H)

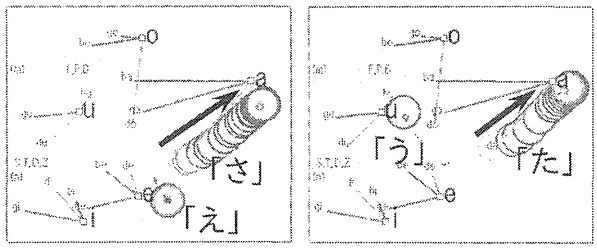

図 6「えさ」生成時(左)と「うた」
生成時(右)のペンの軌跡

2.2.2. 子音のための導線

先の研究から、「おはよう」「こんにちは」などの簡単な単語であれば、図 3 で示した操作面中の、母音だけをペンでたどるだけで、十分に認識可能な音声を生成することができることが確かめられた。しかし、動かし方を工夫すると、さらに子音に近づいた音声を生成することができる。そのような子音に近い音声を生成できる理由として、ヒトの発話音声の子音開始部分に存在しているホルマント周波数の急激な変化(ホルマント遷移)が再現されていることが考えられる。

そこで我々は、ヒトの発話音声を分析し、個々のホルマント周波数を再現できるような軌跡を簡易的に操作面に表示した「導線」を提案している。図 3 の直線がそれである。

子音は音声学的に、子音生成時の舌の位置(構音位置)と子音の発音の方法(構音方式)とによって分類できる。ホルマント遷移は主に舌の動きによって生じるため、同じ構音位置の子音は似たホルマント遷移を持っている。

そこで、同じ構音位置の子音では同じ「導線」を共通に使用することとし、数を減らして表示させた。すなわち、/k/と/g/、/s/と/z/と/t/と/d/、/p/と/b/では、同じ導線をたどる。また、/h/は構音位置が声門であるため、ホルマント遷移がほとんどないため、表示していない。さらに、語頭でない無声子音の場合には、ペンを一瞬だけ操作面から離し、約 100～200ms の無音区間を子音の前に付けると、子音に近い音声に聞こえる。

たとえば、「こんにちは」と出力したい場合には、図4のような軌跡をたどる。「こ」は"g"と"o"を結ぶ直線をなぞり、「ち」は"d"と"i"を結ぶ直線をなぞる。また、「ち」の子音は無声子音であるため、子音の直前で一瞬ペンを浮かせる。

3. 擬似子音の付加

先の研究で行われた音声生成実験と聴取実験では、導線を頼りにしたペンの操作による母音部の制御だけでも、文脈無く提示された 100 語のうち 1～2 割程度の単語了解度を得ることができた。しかし、特に無声子音を多く含んだ単語は了解度が低く、課題となった。

そこで本報告では、試作器に新しく乱流音生成部を加え、これまでとほぼ同じ操作方法で、擬似子音である乱流音を生成できるようにした。すなわち、ペンを操作面から離してから 300ms の間を乱流音モードに自動的に切り替わるように設定し、その間にペンが置かれた場合に、乱流音が生成されるように改良した。

操作方法は、語頭でない無声子音の場合には、2.2 節で述べた音声操作方法とほとんど同じとなり、一瞬ペンを離しすぐに置くと乱流音が生成され、擬似子音が付加される。語頭で乱流音を生成させたい場合には、一瞬だけ弱くペンを置いて話したあとに、ペンを置くと、乱流音を生成することができる。

また、ペンを置くタイミングを調整することにより、乱流音部分の長さを調整することができ、破擦音や破裂音に近づける[3]ことが可能である。また、乱流音生成時も乱流音がない場合と全く同じにように導線をたどる。

この方式を用いて、「えさ」と「うた」を生成させた時のペンの軌跡と、筆圧の変化を図6と図7に示す。筆圧が 0 に鳴っている箇所が、ペンが操作面から離れている部分である。また、「さ」と「た」だけを生成した時の筆圧の変化を図 8 に示し、生成された「さ」の音声波形とスペクトログラムを図 9 に示す。

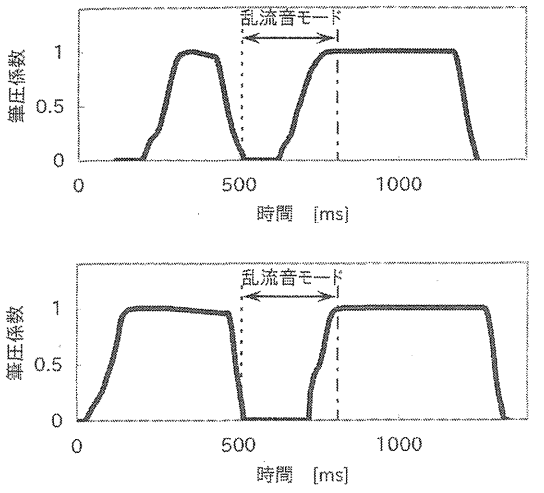

図 7「えさ」生成時(上)と「うた」生成時(下)の筆圧変化

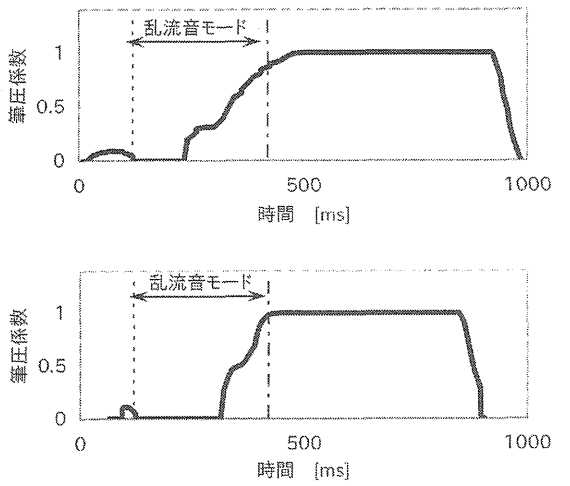

図 8「さ」生成時(上)と「た」生成時(下)の筆圧変化

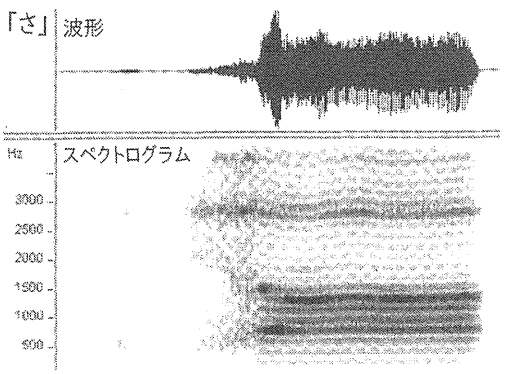

図 9 生成された「さ」の波形とスペクトログラム

4. 聴取実験

4.1. 音声の生成

先に述べた音声生成器の乱流音付加の効果を調べるため、音声の生成実験および聴取実験を行った。

音声の生成は、本インタフェースの開発に関わり使用方法を熟知している 1 名(以下、生成者と呼ぶ)が行った。生成者は、左耳 50dBHL、右耳 30dBHL 程度の聴力レベルであり、器質性および運動性の構音障害を持つ。手や指の運動機能は正常である。

対象とする単語として、坂本らによる“単語了解度試験用単語リスト[10],[11]”中の、親密度 7.0～5.5 の単語リスト中から、無作為に選んだ 2 枚のリスト(1 単語 4 音節からなる計 100 単語) を採用した。日常生活においては文脈や前後の単語の並びから不明瞭な音声であってもかなり推測が可能である一方で、生成作業では文脈が全く無いということを考慮し、評価対象としては親密度の高い単語を採用した。

生成者は、ランダムに並びかえられた対象の単語を見ながら、3 節で述べた方法で乱流音を含んだ音声を順に生成した。それぞれの単語の生成回数は自由とし、生成者自身が納得できる音声を生成できた時に、次の単語へ移るようにした。さらに、音声生成中のペンの軌跡と全く同じ軌跡を再現できるように、ペンの筆圧と座標をミリ秒単位の時間変数とともに記録した。

生成作業中の音声は、イヤホンで生成者へ提示すると同時に、PC 上で動作する音声波形エディタ Wavesurfer によって録音された。以下、この音声を「乱流音あり生成音」と呼ぶ。

そして、生成者が 100 語の音声を生成し終えた後に、記録されたペンの筆圧と座標を元に、乱流音生成機能の無い状態で、全く同じ入力を再現し、生成された音声を同様に録音した。以下、この音声「乱流音なし生成音」と呼ぶ。

4.2. 生成音の聴取

上記の、生成された乱流音あり音声と乱流音なし音声を対象音声として、聴取実験を行った。

聴取者は全員、聴力に関する大きな病気をしたことがなく正常な聴力を持つ 20 代から 30 代の男女 6 名(男性 5 名、女性 1 名) とした。

聴取実験は、聴取者へ、録音された生成音 1 語ずつをランダムに提示し、聴取者自身が PC のキー入力によって、聞こえた単語が何であるかを記入する方法で行った。音声の提示は、汎用 PC に接続されたアームレスヘッドホン Victor HP-AL201 を通して行われた。

聴取者に対しては、実験の前に提示音声が日本語の 4 音節の単語であることを教示し、単語が推測可能な場合には、聞こえたままの文字を記入するのではなく、その単語を答えるように指示した。

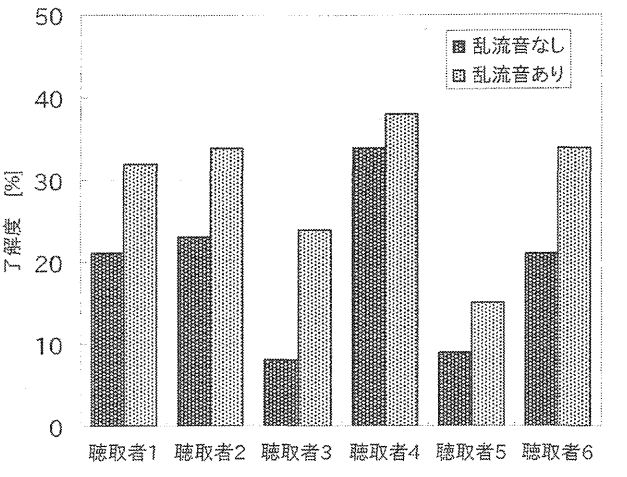

図 10 聴取実験結果

また、聴取実験の順序効果の影響を相殺するため、生成された 100 語の単語を、無作為に A 群と B 群とに分け、A 群の乱流音あり生成音と B 群の乱流音なし生成音をペアとして音声群 I とし、B 群の乱流音あり生成音と A 群の乱流音なし生成音をペアとして音声群 II の音声とした。すなわち、音声群 I と II は、それぞれ全く同じ単語が含まれ、乱流音の有無が入れ替わった群となる。聴取実験では、3 名の聴取者には音声群 I の実験の後に音声群 II の実験を行い、残り 3 名聴取者には、音声群 II の実験のあとに音声群 I の実験を行った。

4.3. 結果

図 10 に聴取実験の結果を示す。6 人の単語了解度の平均は、乱流音あり生成音で 29.5%、乱流音なし生成音で 19.3%であった。また、いずれの聴取者においても乱流音あり生成音のほうが乱流音なし生成よりも了解度が高かった。

単語了解度が高かったものとして、「おやゆび」があげられ、乱流音の有無にかかわらず全員が正答した。一方で、「すうはい」という単語については、乱流音なし生成音では誰も正しく解答できなかったのに対して、乱流音あり生成音では全員が正答した。

乱流音なしに対して乱流音ありで正答者数が増えた単語は 38 単語存在し、逆に乱流音ありで正答者数が減った単語は 18 単語存在した。ただし、乱流音ありで正答者数が、逆に 3 人以上増えた単語は 10 単語存在し、逆に 3 人以上減った単語は無かった。

単語了解度が低かったものに着目すると、正答者が一人もいなかった単語は、乱流音なしで 40 単語あったのに対し、乱流音ありでは 28 単語であった。

4.4. 考察

実験では、乱流音ありで約 3 割、乱流音なしで約 2 割の了解度であった。いずれの聴取者でも乱流音ありのほうが了解度が高かったことから、新しく加えた乱流音付加による効果が、ある程度合ったといえる。

ただし、2〜3 割という了解度は一般には高いとは言えず、更なる改良が求められる。了解度が低かった単語を見ると、「ま」や「な」等の鼻音を含む単語や、有声子音を多く含んだ単語が多く見られた。鼻音については、鼻腔の影響があるためホルマント遷移だけでは十分に再現できなかったと考えられる。また、有声子音については、乱流音部分と母音部分の切り替わりのタイミングや振幅係数の制御や、またホルマント遷移速度が不十分であることが、原因として考えられる。また、操作のしやすさの観点から、乱流音長さを比較的長くとった点も、ひとつの要因であろう。

今後は、これらの処理を利用者が操作できるように、改良を加えていく必要があると考えている。

5. まとめ

我々は発話障害者支援のための音声生成器として、文字や単語に頼らずに、指やペンで面をなぞる動きで、楽器のようにリアルタイムに音声をコントロールしながら音声を生成できる音声生成方式を提案している。具体的には、母音生成に必要な第 1 ホルマントと第 2 ホルマントの周波数を、操作面の横軸と縦軸とに取り、その操作面を指やペンでなぞって、音声を制御する。

汎用パソコン上で動作する試作器による過去の研究では、母音部分をたどるだけでも、簡単な単語を生成できることを示した。また、「導線」をなぞることにより、子音の直前のホルマント遷移を再現しやすくして、子音に近い音声を生成する方法を提案してきた。しかし、その生成音の明瞭さは高くはなく、特に無声子音が不明瞭であるという課題があった。

本稿では、この音声生成器の子音部分の明瞭さを上げるため、ペンタブレットからの入力のみで、摩擦音・破擦音・破裂音の雑音部分を簡易的に付加できるように改良し、その音声生成と聴取実験について報告した。

文脈の無い 100 語の生成音声の聴取実験から、乱流音の付加の効果がある程度見られた。しかし、単語了解度は全体として約 3 割にとどまったことから、更なる改良が必要であることが示唆される。

本研究は、音声をできるだけ記号化せずにペンかの入力で、音声生成パラメータを連続的に制御して、楽器のように声を生成することに主眼を置いている。一方で、語了解度を上げるためには、入力されたペンの動きに対して、ある程度の自動処理を付加して、音声パラメータを制御することが必要と思われる。

今後も、使用者ができるかぎり多様な思い通りの音声を自由に生成できるように、最低限の入力手段と処理を加える方法を探りながら、改良を加えていきたい

文　献

[1] ㈱ナムコ トーキングエイド
http://www.p-supply.co.jp/comaid/voca/talkingaid/

[2] VOCA アルカディアのボイスエイド
http://www.arcadia.co.jp/VOCA/

[3] 藪,謙一郎 伊福部,達 青村,茂, "発話障害者支援のための音声合成器の基礎的設計(聴覚・音声・言語とその障害,一般)", 電子情報通信学会技術研究報告. SP, 音声, 105, pp:59-64, 2006.

[4] ジャック・ライアルズ著, 今富摂子, 荒井隆行, 青原勉監訳/新谷敬人, 北川裕子, 石原健訳: 音声知覚の基礎.海文堂出版, 東京, 2003

[5] 藪 謙一郎, 伊福部 達, 青村 茂, ポインティングデバイスを利用した音声生成方式-発話障害者のための支援機器として-, 日本保健科学学会誌 12(1), pp.49-57, 2009

[6] 藪 謙一郎, 青村 茂, 伊福部 達, ポインティング・デバイスで操作する発話支援インタフェース ヒューマンインタフェース学会誌, Vol.11 No.4 (135-146) 2009

[7] 伊福部, 達, 橋場参生, 松島純一. 母音の自然性における「波形ゆらぎ」の役割. 日本音響学会誌 47, no. 12: 903-910., 1991

[8] Klatt,D.H.: Software for a cascade/parallel formant synthesizer: Journal of the Acoustical Society of America,pages 971-995, volume 67, number 3 March 1980.

[9] ピー・アクティブ P-Active : XP-3300A ポケットサイズペンタブレット
http://www.p-active.com/product/pt/3300a.htm

[10] 難聴者のための単語了解度試験用単語リスト
http://www.ais.riec.tohoku.ac.jp/lab/wordlist/indexj.html

[11] 坂本修一, 鈴木陽一, 天野成昭, 近藤公久: 親密度と単語の音韻バランスを統制した単語了解度試験用リストの構築: 東北大学電通談話会記録第, 69(2), pp.21-34, 2000

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2010-126 (2011-3)

発話障害者支援のためのペン入力座標による リアルタイム音声生成方式
－鼻子音出力の操作方法と音声生成方法の検討－

藪 謙一郎† 伊福部 達†

†東京大学先端科学技術研究センター 〒153-8904 東京都目黒区駒場 4-6-1
E-mail: †{yabu,ifukube}@human.rcast.u-tokyo.ac.jp

あらまし 我々は、発話障害者の日常会話を支援する音声生成支援機器として、ペンタブレットやタッチパッドで入力された座標値を音声生成パラメータへ直接対応させた、リアルタイムの音声生成器を提案・開発している。この方式では、使用者が音声を楽器のようにコントロールするため、自由な間(ま)やリズムなどの非言語的な情報を表現可能となる。試作器による実験から、簡単な単語であれば2つのホルマントをペンで操作するだけでも生成できることが分かっている。しかし、それだけでは明瞭度が不十分であるため先の研究では、明瞭度改善のために、ペンの軌跡と筆圧による操作だけで、摩擦子音などの阻害音も簡易的に生成できる方法を提案してきた。本稿ではさらに、鼻子音を生成する方法を提案し、その実験について報告する。

キーワード ホルマント, 子音, 音声合成, 発話障害者、鼻音

A real-time speech synthesis method to support speech disabilities by pen-operated coordinate representation.
－An approach of manipulation and generation method for nasal consonants.－

Ken-ichiro Yabu† Tohru Ifukube†

† Research Center for Advanced Science and Technology, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8904 Japan
E-mail: †{yabu,ifukube}@human.rcast.u-tokyo.ac.jp

Abstract We have proposed and developed a real-time speech synthesis device in which pen input coordinate values are directly corresponding the speech synthesis parameters, to assist daily communication for speech disability persons. The method of this device enables expressions of nonverbal information such as time intervals or rhythms, because the user can control sound like playing a musical instrument. It have been shown that voices that sound like consonants were produced by controlling only two formant frequencies, but it was unclear sounds. In our previous experiment, we have proposed a method to generate simplified obstruent consonant only by a pen input. In this paper, we propose a method to generate nasal consonants, and describe about an experiment of that improved device.

Keyword Formant, Consonant, Voice Synthesis, Articulation Disorder., nasal consonant.

1. はじめに

発話障害者向けの音声による市販の支援機器として現在、TTSを基本とした機器が最も一般的となっている。これらの機器は、文字や単語から音声を生成するため、使用者が自分の要求や意思を確実に伝える上で、有力な手段となっている。

一方で、健常者を含めて一般に発話音声には、言語情報の以外にも、間(ま)や話すテンポや抑揚、リズムなどの表現が含まれている。それによって、文字や単語の記述だけでは表現できない非言語的な情報の伝達が可能となる。このような非言語的な情報は、日常生活の中で他者との関係を円滑に送るために大変重要な役割を持っている。発話障害者にとっても、その重要性は同様である。

我々は、個性を持った多彩な表現を可能とする機器には、ヒトが構音器官を巧みに操って自由に音声を出すのと同じように、身体の一部の動きと出力音声とが連動するような方式が必要であると考えている。

そこで、文字や単語に頼らずに、指やペンで面をなぞる動きで、楽器のように使用者自身がリアルタイムに音声をコントロールしながら音声を生成できる音声生成器を提案している[1]。

具体的には、ヒトの母音知覚に必要な第1ホルマントと第2ホルマントの周波数(以下、F1,F2)[2]を、音声生成器の操作面の縦軸と横軸とに割り当て、使用者が指やペンを操作面に置いた瞬間に、そのホルマント周波数を持った音を発するというものである。この方法では、操作面に指やペンを置いた瞬間に声が出ることから、入力に伴う時間遅れが殆ど無く、自由なリズムをつけることが可能となる。また、母音と母音の間の中間的な音を連続的に変化させて出すことも可能であり、多様な音韻が表現できる。

先の研究では、「おはよう」「こんにちは」などの簡単な言葉であれば、我々が試作した汎用パソコン上で動作する音声生成器で、母音部分の F1、F2 を指やペンの操作で再現するだけで表現可能であることを示した[3]。また、ヒトの発話音声中の子音開始部分に存在するホルマントの急激な変化(ホルマント遷移)に着目して、これを再現するための「導線」を提案した[4]。そして、摩擦音や破擦音などの子音の明瞭度の改善を目的として、子音部の雑音成分を自動付加するように改良してきた[5]。子音部の自動付加は、新しいボタンスイッチ等を加えずに、「導線」を目印としてペンを操作するだけで可能となるように、極力簡易的なものにしている。

この方式を用いた実験では、「単語了解度試験用単語リスト」[6]の親密度 7.0-5.5 の単語リストに含まれる 4 音節からなる 100 単語において、約 3 割程度の了解度となった。特に、/m/や/n/の鼻子音を含む単語の了解度が低い傾向にあった。

そこで本稿では、鼻子音の明瞭化のために、操作面上に鼻子音領域を加える改良を行い、音声の生成、聴取実験を行ったので、報告する。

2. 試作器の概要

2.1. 操作面と入力装置

音声生成器の操作面を図1に示す。ペンタブレットの縦軸と横軸とを XY 座標とし、X 座標(横軸) に F1 の 200～850Hz を、Y 座標(縦軸) に F2 の 600～2500Hz を対応させている。使用者が任意の位置にペンを置くとその座標に対応する母音が即時に生成され、ペンを離すと母音が停止する。操作面上には、“ a ”,“ i ”,“ u ”,“ e ”,“ o ”という文字を各母音のホルマント周波数に対応した位置へ表示させた。この文字は、ペンタブレットの操作面に示すと同時に、コンピュータの画面上にも、同様に表示させた。また、ペンの筆圧を音声の強度に対応させ、筆圧が強ければ音が大きく、低ければ音が小さくなるようにした。ペンの位置は、画面上の対応する位置に小円形でリアルタイムに表示させ、筆圧を円の大きさで表示させた。

通常、ヒトがペンを扱う場合、机などの面に手を置

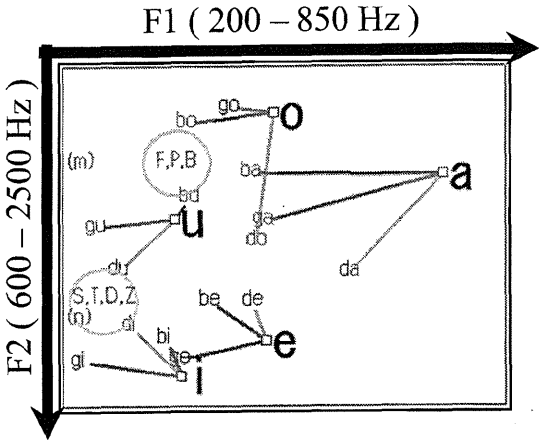

図 1 入力操作面

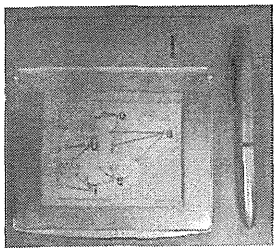

図 2 導線とガード取り付け後のペンタブレット
(P-Active XP-3300A)

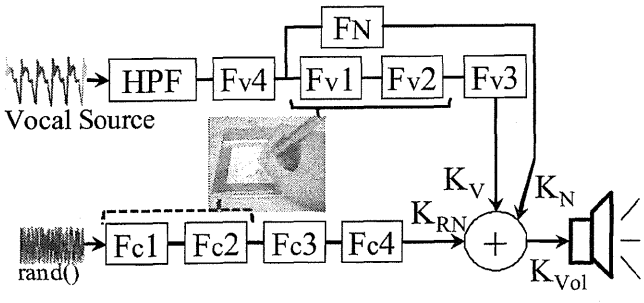

図 3 音声生成器のブロック図

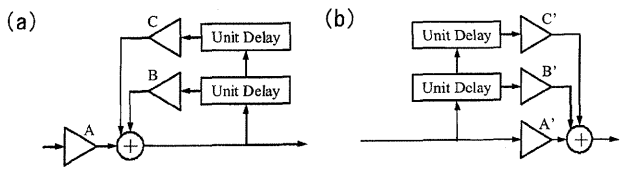

図 4 使用したディジタル共振器(a)と反共振器(b)

いた状態で作業することが多く、その状態が最も楽にかつ正確に操作できると考えられる。そこで、手を面に固定した状態で操作面全体をタッチできるような、小型のペンタブレット(図 2、表 1)を採用した。

ペンの位置が母音の範囲から大きく外れると異音が生じるため、図 3 に示すように、所定の範囲を切り抜いた厚さ約 0.4mm の PET 樹脂を貼りつけ、ペンが範囲外に出ないようにするガードとした。

2.2. 生成処理部

音声生成処理のブロック図を図 3 に示す。音声の生成処理は大別すると、声帯振動による音を再現する母音処理部と、摩擦音・破裂音・破擦音などに含まれる雑音成分(以下、乱流音と呼ぶ)を処理する乱流音処理部で構成されている。母音処理部と乱流音処理部は、各々、ディジタル共振器と反共振器で構成されている。

母音生成部では、原音となる音声信号が、各ホルマントに対応する Fv1 から Fv4 のディジタル共振フィルタを通って特徴付けられた後、出力される。フィルタのうち、Fv1 と Fv2 の特性が、使用者の操作面からの入力によってリアルタイムに制御される。その他のフィルタの特性は固定とした。また高い周波数成分を強調するために、緩やかなハイパスフィルタを設けてあり、図中の HPF と記している。

FN は、後述の鼻音生成のためのフィルタで、係数 K_N で振幅調整されてから合成される。

母音生成部の原音波形には、声の振幅やピッチのゆらぎを再現するため、筆者の「あー」という約 3 秒間の声をサンプリング周波数 16kHz で録音し、30 次の LPC 逆フィルタ処理で補正した擬似声帯音声を用いていている。

乱流音処理部は、次に述べる擬似子音の付加に用いている。乱流音処理部は、音源が乱数発生関数によって生成された雑音である以外は、母音処理部とほぼ同じ構成である。

共振・反共振器フィルタについては、Klatt らと同様のフィルタ[9]を使用した(図 4)。HPF には、反共振フィルタの反共振周波数をゼロとしたものを用いた。図 4 の *A,B,C* は、共振周波数とバンド幅から求められる係数である。

2.3. 子音のための導線と擬似子音

先に述べたように、簡単な単語であれば、図 3 で示した操作面中の、母音だけをペンでたどるだけでも、十分に認識可能な音声を生成することができることができる。また、動かし方を工夫すると、さらに子音に近づいた音声を生成することができる。これは、ヒトの発話音声の子音開始部分に存在しているホルマント周波数の急激な変化(ホルマント遷移)が再現されるためであると考えられる。

表 1 使用したペンタブレットの仕様[7]

型名	XP-3300A
読取方式	電磁誘導
読取範囲	82.5×63.5mm
読取分解能	0.025mm
読取高さ	10mm
読取速度	100 ポイント/秒以上
筆圧レベル	1024 レベル
外形寸法	143(D)×155(W)×5(H)

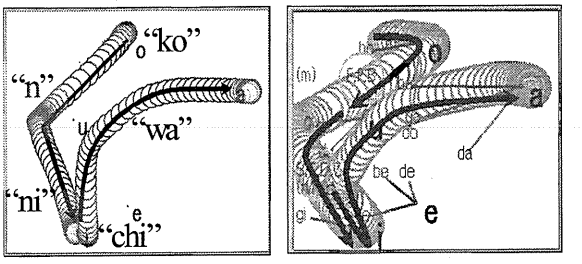

図 5「こんにちは」生成時のペンの軌跡

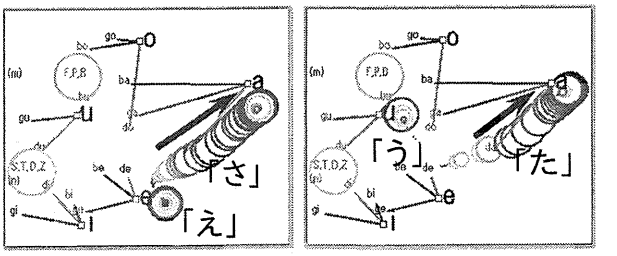

図 6「えさ」生成時(左)と「うた」生成時(右)のペンの軌跡

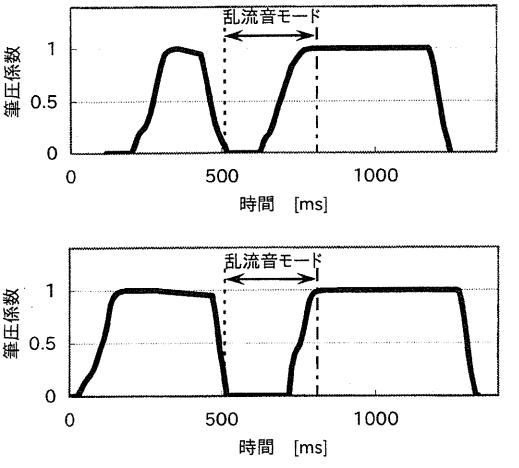

図 7「えさ」生成時(上)と「うた」生成時(下)の筆圧変化

そこで、子音に近い音を出すことが可能なペンの軌跡を使用者に提示するため、「導線」を提案した。図 3 の直線が「導線」である。

各線は、ヒトの発話音声の分析から、求めたものであるが、全ての子音について別々の線を提示するとは事用上不可能である。そこで、子音生成時の舌の位置(構音位置)と子音の発音の方法(構音方式)による、古典的な音声学的分類に基づき、同じ構音位置の子音では同じ「導線」を共通に使用することとし、数を減らして表示させた。すなわち、/k/と/g/、/s/と/z/と/t/と/d/、/p/と/b/では、同じ導線をたどる。これは、同じ構音位置の子音はホルマント遷移が似ていることによる。ただし、/h/は構音位置が声門であるため、ホルマント遷移がほとんどないため、表示していない。導線を用いて「こんにちは」と出力したときのペンの軌跡を図 5 右図に示す。

さらに、音声出力中にペンが一瞬だけ操作面から離されたとき、ペンが操作面から離されてから 300ms の間を乱流音モードに自動的に切り替わり、その間にペンが置かれた場合に、乱流音が生成されるようにした。語頭以外で乱流音による子音を出す際には、図 6 に示すように導線に沿ってペンを動かすと同時に、図 7 に示すようにペンを一瞬浮かせると乱流音が付加される。語頭の場合には、図 8 のようにペンを弱くタップしてからペンで導線をなぞる。ペンを置くタイミングを調整することにより、乱流音部分の長さを調整することができ、破擦音や破裂音に近づけられる。

3. 鼻音領域の付加

3.1. 鼻音生成の原理と数値制御による実験

本稿の実験では、新しく図 3 に示される、鼻子音の処理を加えた。それに先立ち、鼻子音/m/, /n/を生成可能な簡略化された制御パラメータを得るため、まず数値制御による実験を行った。

まず、成人男性 1 名の音声の周波数を分析し、必要な周波数特徴量を求めた。それを図 9 に示すような、線形的に遷移するようなデータとして、音声生成器へ数値列によって入力した。実験者による音の生成と聴取の繰り返しにより各係数の値とタイミングを定め、共振器は共振周波数 250Hz の共振器とし、図 9 のようなデータをモーラごとに得た。その際、Fv 1 ～Fv4 については、各鼻子音と同じ構音位置の「導線」に使用したものと同じ周波数変化となるようにした。

得られた制御パラメータによって単モーラ音声を生成し、聴取実験によりその明瞭さを確認した。

鼻音以外のモーラとの混合による、聴取者 8 名での実験で、実験の結果、/ma/, /mi/, /mu/, /me/, /mo/については平均で 63 %、/na/, /ni/, /nu/, /ne/, /no/については、平均で 53 %の正答率を得られた。このことから、/m/

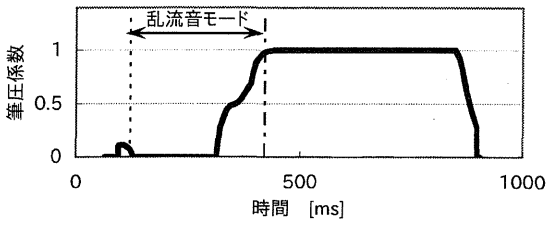

図 8「た」生成時の筆圧変化

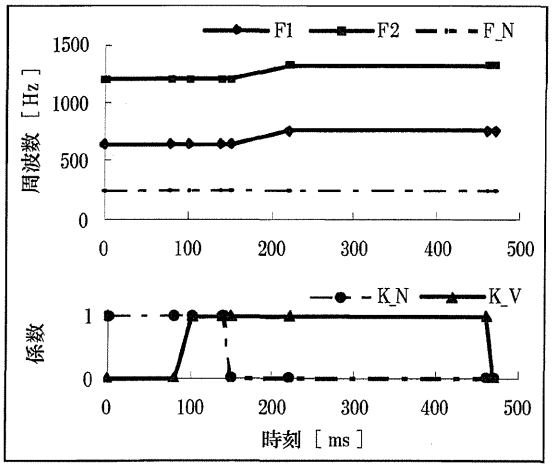

図 9 数値制御による鼻音の生成の概略

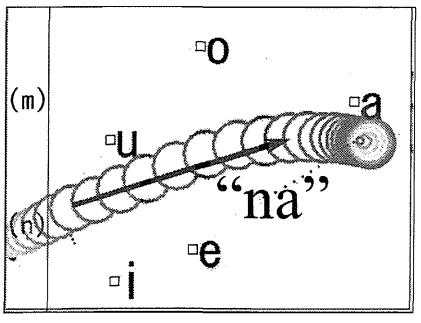

図 10「な」生成時のペンの軌跡

についてはその他の両唇構音の子音（/p/及び./b/)、/n/についてはその他の歯茎構音の子音（/s/及び/z/, 他）と同じ F1, F2 の遷移で、鼻子音用の共振器による音を加えるだけで、鼻子音を簡易的に生成できることが示唆された。この知見に基づいて、鼻子音を制御する操作面の実装を行った。

3.2. 操作面への実装

一般的に、第 1 歩ルマント周波数について着目すると、狭母音は F1 が低く、広母音は F1 が高い。一方で、鼻子音/m/と/n/は、両唇や舌で閉鎖を行って発音する。そこで、/m/, /n/について狭母音の延長上にあるものと捕らえ、図 10 に示すように、操作面上の左端（F1 が極端に低い部分）に鼻音領域を定め、この領域へペンの座標がくると、係数 K_N によって鼻音用共振器が機能するように設定した。/m/と/n/の違いは F2 に対応する上下の位置となる。

4. 聴取実験

音声生成器の鼻子音領域の付加の効果を調べるため、音声の生成実験と聴取実験を行った。

4.1. 音声の生成

生成は、本インタフェースの開発に関わり使用方法を熟知している 1 名(以下、生成者と呼ぶ)が行った。生成者は、左耳 50dBHL、右耳 30dBHL 程度の聴力レベルであり、器質性および運動性の構音障害を持つ。手や指の運動機能は正常である。

対象とする単語は、先の研究[5]で用いた 100 語の単語のうち、特に了解度が低かった鼻子音を含んだ単語 36 語を使用した。このうち 28 単語は、了解者が 6 人中 1 人以下であった単語である。いずれも「単語了解度試験用単語リスト」[6][10]中の、親密度 7.0-5.5 のリストに含まれる 1 単語 4 音節からなる単語である。日常生活においては文脈や前後の単語の並びから不明瞭な音声であってもかなり推測が可能である一方で、生成作業では文脈が全く無いということを考慮し、評価対象としては親密度が高めの単語を採用している。

生成者は、対象の単語を見ながら、鼻子音領域を含んだ操作面上でペンを使って音声を生成した。それぞれの単語の生成回数は自由とし、生成者自身が納得できる音声を生成できた時に、次の単語へ移るようにした。さらに、音声生成中のペンの軌跡と全く同じ軌跡を再現できるように、ペンの筆圧と座標をミリ秒単位の時間変数とともに記録した。

生成作業中の音声は、イヤホンで生成者へ提示すると同時に、PC 上で動作する音声波形エディタ Wavesurfer によって録音された。以下、この音声を「鼻音あり生成音」と呼ぶ。

そして、生成者が 100 語の音声を生成し終えた後に、記録されたペンの筆圧と座標を元に、乱流音生成機能の無い状態で、全く同じ入力を再現し、生成された音声を同様に録音した。以下、この音声「鼻音なし生成音」と呼ぶ。

4.2. 生成音の聴取実験

上記の、生成された乱流音あり音声と乱流音なし音声を対象音声として、聴取実験を行った。

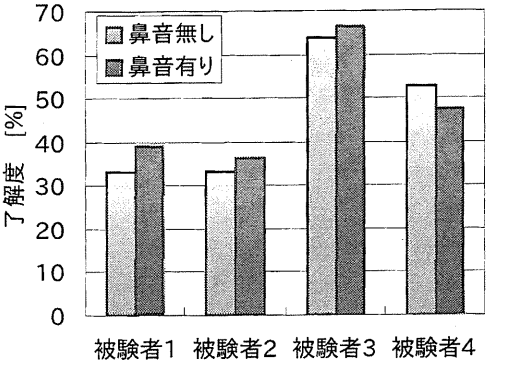

図 10 聴取実験結果

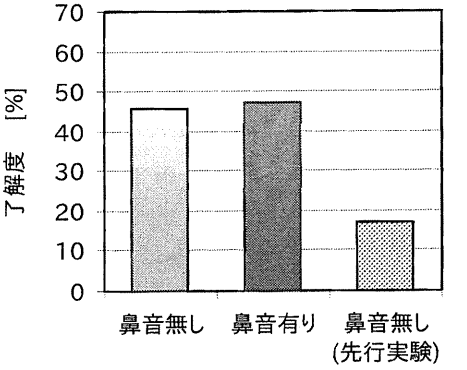

図 11 鼻音の有無による了解度の比較

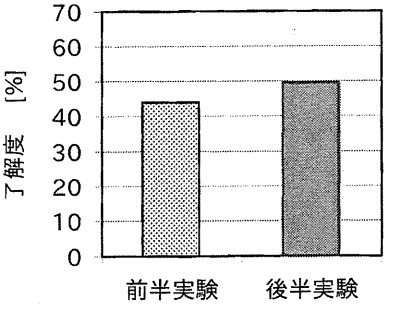

図 12 前半実験と後半実験との了解度の比較

聴取者は全員、聴力に関する大きな病気をしたことがなく正常な聴力を持つ20代から30代の男女4 名(男性 3 名、女性 1 名) とした。

録音された生成音 1 語ずつを聴取者へランダムに提示し、聴取者自身が PC のキー入力によって、聞こえ