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

The Signal Processing in a Brain and a Programmable Voice Recognition System

Shinji Karasawa[†] Masatoshi Iwamoto[‡]

† Sendai-shi Aoba Invention Club, 1-16-18 Kokubuntyou, Aoba-ku, Sendai-shi, Miyagi, Japan, 980-0803

‡Tohoku Gakuin University, 13-1, Tyuou-1-tyoume, Tagajou-shi, Miyagi, Japan, 985-8537

E-mail: † shinji-karasawa@cup.ocn.ne.jp, ‡ masa@tjcc.tohoku-gakuin.ac.jp

Abstract This paper describes brain mechanism for speech recognition from viewpoint of digital signal processing and a programmable voice recognition system. Here, these processors deal with plural activated reactions. Timing of an affair in a real world and timing of the recognition is different. Many candidates for recognition will be activated to understand an affair in the real world. Interaction among activated junctions in a nerve network contributes to construct the next activities. The unconsciously activated plural activities are unified through those impulsive reactions and the decision is intermittently renewed. This mechanism was adopted at a speech recognition system.

Keyword: Organization of agent, Brain, Production of information, Haar wavelet transform, Speech recognition.

1. Introduction

The intelligent mechanisms are available for the engineering of a robot. Investigations about the origin of intelligence reveal the universal characteristics of recognition. Although the knowledge is necessary for the science, the intelligence is necessary for the engineering. The engineering of intelligence is the action of “supply to the demand”.

If we can present the way to manage activated plural demands, we can carry out such intelligent operation by using a digital computer. “The concept of activity [1]” is powerful tool to describe the intelligence that responds to the changes of outer world at real time [2]. The primitive recognition is a part of behavior control. The intermittent interactions are carried out at the activated junctions in a nerve network [3].

The essence of first life is unification of plural activities. There are many kinds of intermediates in the reaction in organic compounds. The activated plural activities are unified through another reaction and different reactions take place concurrently.

A compact speech recognition system is designed from view point of engineering i.e. supply according to demand. The segmentation of analyses has to adjust to the utterance. There are reports on wavelet based feature extraction for speech processing [4],[5],[6]. Although a signal is represented as a sum of sinusoids in Fourier transform, it is represented as a sum of rectangles in Haar Wavelet Transform (H-WT). Hamming distance on the sum of absolute value of H-WT coefficient (SWC) is used for an evaluation of pattern matching [7],[8]. In this paper, results of a command detector by a program are reported.

2. Origin of intelligence

2.1. The first life is organized by organic materials

Hydrocarbons and carbohydrates were produced from carbon dioxide in the early earth. The carbon dioxide is soluble in the sea water. The sea water circulates to deep sea. There is the possibility that the oxygen atom forms oxygen molecule of O_2 owing to the strong water pressure, because the size of oxygen is very large compared with that of hydrogen. We tried the experiment under the high pressure with low temperature on the ice in which CO_2 and NaCl are solved. Many bubbles appeared in the ice.

The quantity of dissolved oxygen is increased in the sea under deeper than 1,000m. The molecule of O_2 discharges to the outside from the sea, and methane-hydrates are produced. Every organic material in the earth was produced from carbon dioxide in the early earth.

There are many organic materials in the sea water. The organic molecules form many kinds of large molecules by the polymerization. The speed of polymerization is slow. Interactions among intermediates of polymers make many kinds of huge organic molecules. The polymer possesses a function of copy production. It is an organic catalyst. Those are materials to produce a first life.

2.2. To organize reactions is the first intelligence

A gathering of many kinds of organic molecules makes possible to form a system of creature. A life is the system that is able to continue activities as shown in Fig.1.

Each reaction is activated by preconditions of the reaction. If circulating of a series of reactions forms a continuous chain of activities, this system is able to be active continuously.

Many kinds of reactions for a system of continuous activities

Surroundings ⇒ Reaction(1) ⇒ Renewal of surroundings ↓
↑ Renewal of surroundings ⇒ Reaction(2) ⇒ Surroundings

Ecology; adapts behavior of each creature to the environment

Fig.1. A system of life is supported by many kinds of polymers and catalysts in order to continue the reaction. Here, the function of a catalyst corresponds to a subroutine in a program.

The reaction accompanies some transportation of materials, and it changes the situation. So, a creature lives together with other lives. The creatures form an ecosystem. The process of self-organization of those reactions is shown in Fig. 2.

Rules for self-organization = Principle of intelligence

Demands ⇒ Supply ↓ (Agent 1)
Demands⇒ Supply ↓ (Agent 2)
Demands ⇒ Supply ↓ (Agent 3)

Communication originates in organization

Fig.2. Self-organization is carried out by a rule of “Supply to Demand”.

2.3. Evolution; that is caused by activities

If a creature makes action, the situation of it will change. The individual biological environment is not fixed. Extinct creatures were not adaptive to the new environment. Since individual biological environment changes, the creature possesses some attributes of development. As the result, the creature that is survived possesses a mechanism of subjective reaction.

Every animal makes action in order to continue activities of life. Every creature works for the next activities. The gene is a tool box for the activities. The programs in the gene are tools for activities. This concept of activity explains the evolution of creature as follows.

New function is acquired by trial and error

↳ Attributes of mutation in evolution

Results of evolution possesses the rule of development

1. **Earthworm ; Photoreceptors distributed around body**
2. **Planaria ; Eye as an arrangement of photoreceptors**
3. **Nautilus ; Eye as a pin hole camera**
4. **Squid ; Eye as a camera with lens**

Fig.3. Mechanism of evolution (There are steps of development in the evolution. If a different animal that lives with similar life style, its structure will resemble.)

New parts are prepared before the new system. The new development is achieved by the method of try and error. So, the evolution indicates an attribute of mutation. The intelligence of a creature inevitably becomes adaptable. The innovation of creatures diffuses through the mechanism of copy.

3. Signal processing in a brain

3.1. Communication without linguistic activities

The activated plural activities in a creature are unified through the biochemical reactions and each decision is intermittently renewed. The nerve circuit is formed in a multicellular animal in order to improve performance of behavior control.

The organized system of activities is localized and it forms a colony. When there are many candidates in a colony, a junction for the next reaction is decided by a mechanism of winner-take-all. The algorism of winner-take-all is achieved by inhibition of each other at the colony.

The reaction of neuron is impulsive, and the nerve system coordinates plural actions at each reaction as shown in Fig.4.

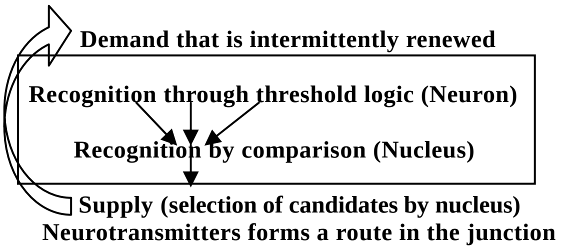
Demand that is intermittently renewed
Recognition through threshold logic (Neuron)
Recognition by comparison (Nucleus)
Supply (selection of candidates by nucleus)
Neurotransmitters forms a route in the junction

Fig.4. Coordination of concurrent activities due to a time sharing operation of the network

Each reaction in a nerve network plays a primitive function of communication. Each impulsive reaction represents an activity and a neuron corresponds to an agent. The pattern of discharged neurotransmitters forms the junction of nerve network concurrently. The pattern stored in a nerve network can be used as pins in a music box.

3.2. Use of memory; that is additional activities

The world of information does not include what will happen in the future. The knowledge is a result of experience. The world of information does not change, if it does not implement. The usage of information in memory is an additional operation for a life.

The additional organ of neocortex is the organ for the memory. As tools for ad hoc reactions, many modules of reactions are prepared in the neocortex. The information is implemented at the occasion of activity according to progress of the real time.

The mammal that possesses the neocortex appeared on earth about two hundred million years ago in the Mesozoic era. The fur to keep warm and the memory to get food in the night are necessary for the life style of nocturnal.

3.3. Filter carries out function of focused attention

There are many areas on visual recognition in a sight. The filter that eliminates candidates not concerned is necessary in order to recognize an image in the vision.

The information from the real world at each moment is limited, but stored information will be accumulated. The increase of memory needs a mechanism of focused attention. The mechanism of focused attention has been developed to economize the process of decision making.

3.4. Control of activities according to demand

The priority of reaction depends on the demand at the situation. The mechanism that assigns priority area among activated agents is available to recognize speech voice. Fig.5 shows the block diagram of the control in a brain.

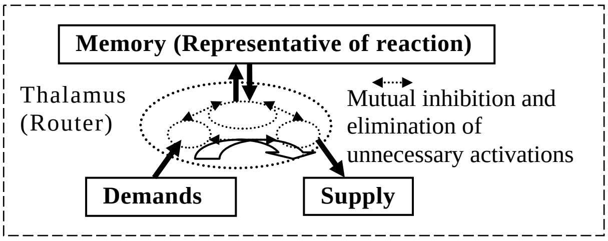

Fig.5. Control of activated areas in a brain (This control depends on demand. It is subjective.)

The transmission of activated portions is able to add structures to an already existing working system. These functions of brain can be materialized by means of digital technologies.

3.5. Organization by intermittent operations

Although the information obtained by sampling does not change, the cognitive function in a brain is carried out through a time-shared operation of impulsive activities.

The intermittent reaction is able to change the behavior and it makes possible to adapt the behavior to the changing environment. The plural events are checked by means of time sharing scanning that makes possible to recognize unpredictable events. The behavior control is able to repeat and a part of it is always rewritten.

3.6. Nucleus as a shift register for pattern of data

Each impulse could not stay at the same state in a nerve network. A sheet of neurons is able to operate as a shift register for a pattern of impulses, because the pattern of impulses is always shifted in the sheet of delay elements. Each sheet of outputs in a nucleus are referred to parallel candidates concurrently.

3.7. Why recognition depends on segmentation

The meaning of reaction depends on the segmentation, because each neuron makes action of recognition through a process of data-matching. Here, the system of sharing elements economizes the elements. So, the neuron that operates similar function is linked together. The linkage forms a module and a layered structure.

The possible states of junctions those link to the candidates are eliminated by the distance between demands. To select the candidate with shortest distance is to carry out “Winner-take-all”.

3.8. Language as a tool of memory

There is an innate structure in order to continue the activities in a life. The origin of information is a tool to satisfy the demand. Human expresses the information by language. The linguistic expression is a representative.

The surface structure of language is implemented at the time of experience i.e. the acquisition of native language is heuristic. It can be considered from view point of engineering that the faculty of language is an artificial tool to supply for the demand.

4. A compact speech recognition system

4.1. Segmentation; that is adjusted to utterance

By comparison of results we confirmed that the processing of speech recognition must be adjusted to utterance. We decided the data for processing as follows.

The length of a frame was 51.2msec that was adjusted to an action. The time shift of frame is 25.6msec. The whole length of time for a processing on one utterance was 252msec of 10 steps.

4.2. Haar wavelet transform (H-WT)

Traditional speech recognition is based on Fourier transform in which signals are represented as a sum of sinusoids. But Haar wavelet transform (H-WT) is computationally simple compared with Fourier transform.

We use H-WT in which signals are represented as a sum of rectangles as shown in Fig.6.

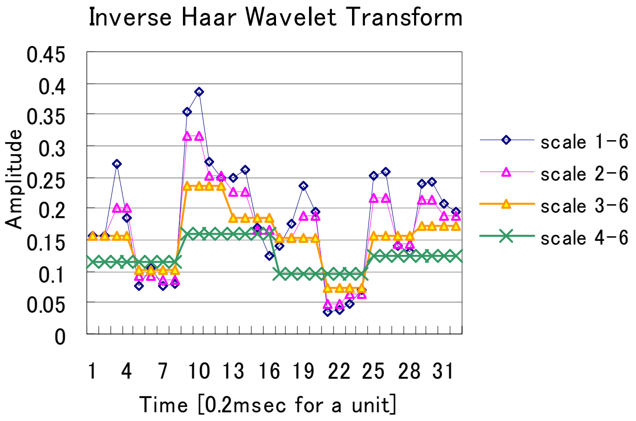

Fig.6. Haar wavelet transform (H-WT) in which a wave is represented as a sum of rectangles

Haar wavelet is one cycle of square wave. The frequency characteristic is reflected in the sum of absolute value of Haar wavelet transform coefficient.

We use 256 pieces of data those are picked up during 51.2msec. Here, the sampling frequency is 5 kHz. If we use SWC as a representative of frequency spectrum, 256 pieces of data are compressed to 8 pieces of SWC.

4.3. Preprocessing for template-matching (TM)

The technology of template matching (TM) depends on the pre-processing.

The speed of utterance does not change during one utterance. The time scale contraction in which each time span on one utterance is normalized linearly is carried out in order to adjust the speaking speed fluctuation. This Dynamic Time Warping (DTW) was carried out through changing address of original data.

4.4. Waveforms of speech voice

The calculation in this report were carried out by means of Visual Basic for Application (VBA) as a macro of Excel. The processing was written in a series of program.

The waveforms to be decoded are shown in Fig.7 and Fig.8. There are 3 times repeated of [Ma-e], and 3 times repeated of [A-to]. Here, [Ma-e] and [A-to] are Japanese, and [Ma-e(前)] means front. [A-to(後)] means back.

The waveform of Fig.7 is uttered by T (a boy). The sampling frequency is 5 kHz. The number of data for an utterance is 20000. That corresponds to the time span of 4 sec. The waveform of Fig.3 is uttered by S (another boy). The condition of data gathering is the same in case of T.

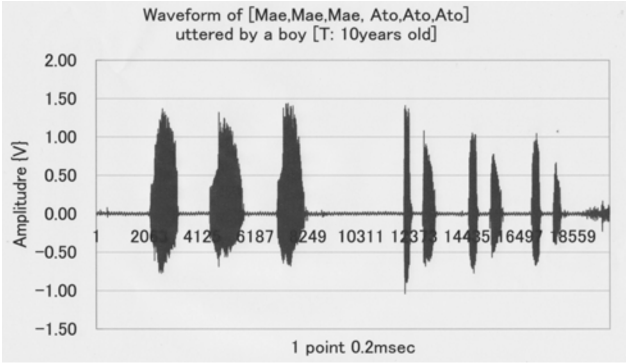

Fig.7. Waveform of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by T

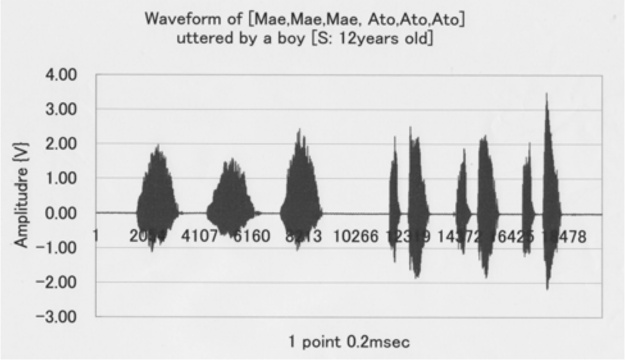

Fig.8. Waveform of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by S

4.5. Segmentation of voice for template-matching

The segmentation was decided automatically by using average values of amplitude over time span of 10msec shown in Fig.9 and Fig.10.

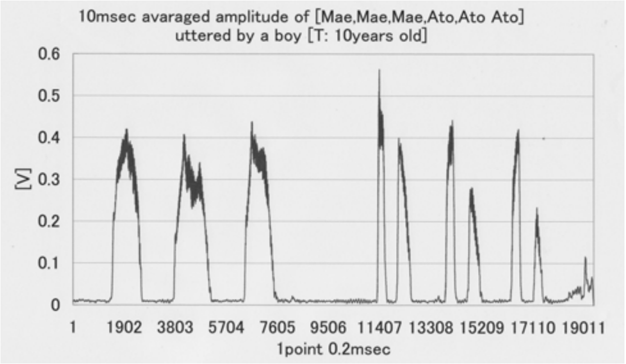

Fig.9. Averaged amplitude of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by T

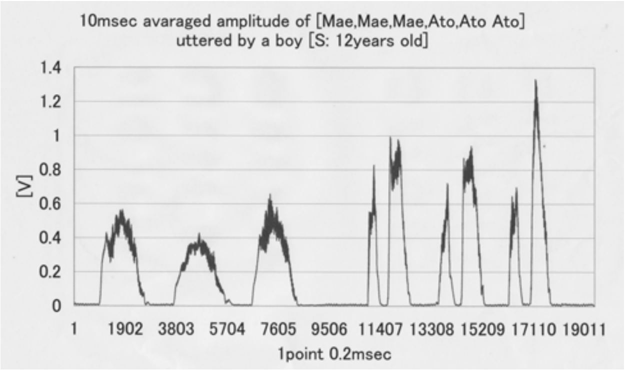

Fig.10. Averaged amplitude of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by S

The values shown in Fig.9 and Fig.10 are used to decide the areas to be analyzed as shown in Fig.11 and Fig.12.

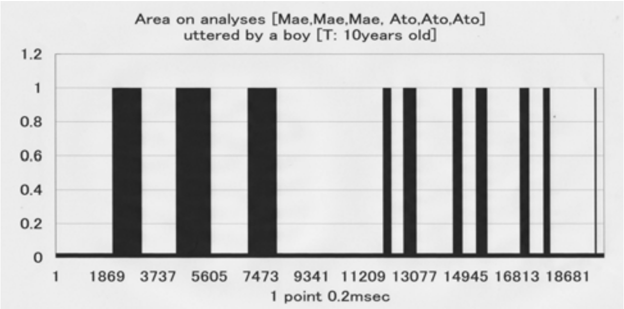

Fig.10. Area of analyses for voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by T

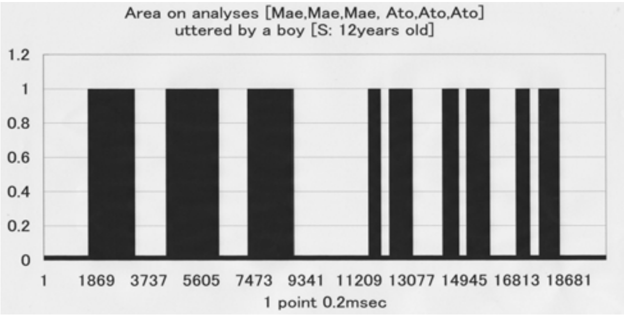

Fig.11. Area of analyses for voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by S

4.6. SWC as a representative of frequency characteristics

The sum of absolute value of H-WT coefficient (SWC) is influenced by frequency characteristics [5], [6].

Principal constituents of speech voices are 0.8msec (1.25kHz),0.4msec (2.5kHz), 1.6msec (625Hz), 3.2msec (312.5Hz), 6.4msec (156Hz). Fig.12 and Fig.13 show SWC those were obtained from the wave shown in Fig.7 and Fig.8.

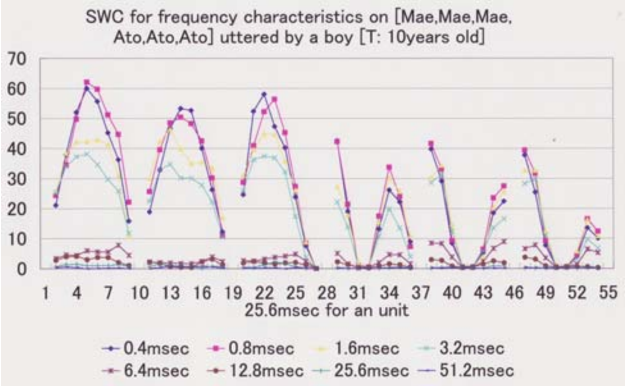

Fig.12. SWC of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by T

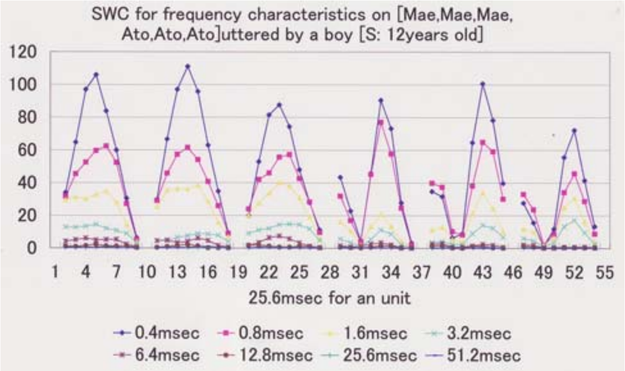

Fig.13. SWC of voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] uttered by S

4.7. Speaker dependent speech recognition

The template matching is evaluated by means of the value of distances between inputs (Demand) and references (Supply).

Fig.14 and Fig.15 shows the Hamming distances are calculated by using the data shown in Fig.7. In Fig.14, the first line (value＝0) is the reference. In Fig.15, the last line (value＝0] is the reference.

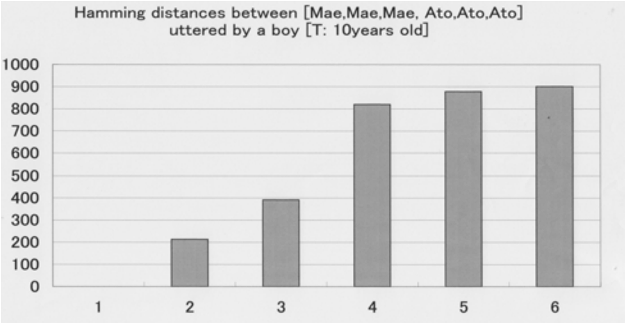

Fig.14. Hamming distances between voice on [Ma-e] and [Ma-e, Ma-e, A-to, A-to, A-to] uttered by T

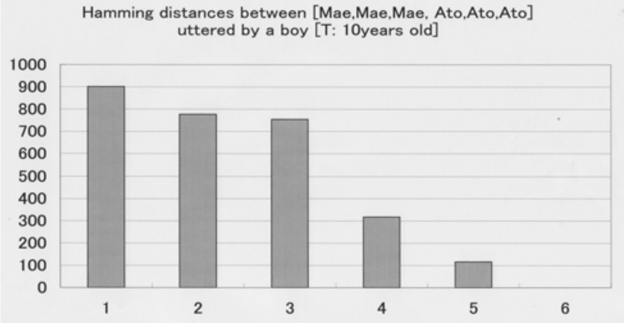

Fig.15. Hamming distances between voice on [Ma-e, Ma-e, Ma-e, A-to, A-to] and [A-to] uttered by T

4.8. Speaker independent speech recognition

A performance of speaker independent voice decoder is evaluated by using the data those are uttered by the other person as references

Fig.18 and Fig.15 show the hamming distance on TM by using SWC data uttered by different person for the reference.

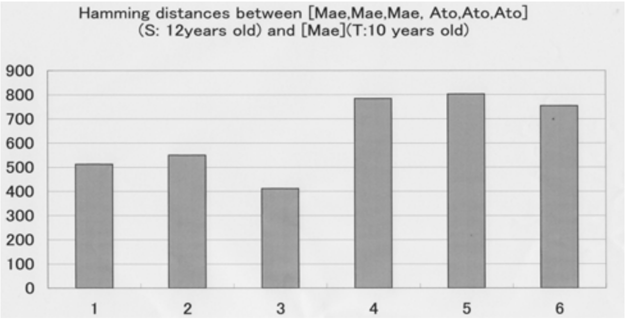

Fig.18. Hamming distances between voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] by S and [Ma-e] by T

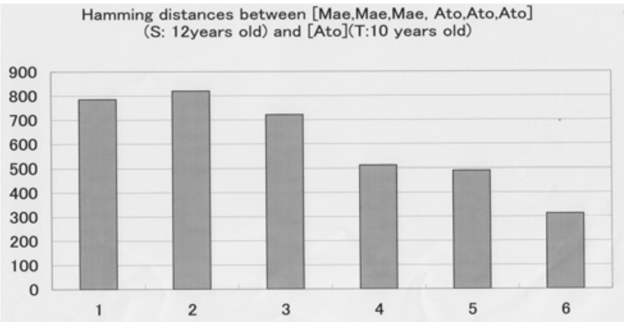

Fig17. Hamming distances between voice on [Ma-e, Ma-e, Ma-e, A-to, A-to, A-to] by S and [A-to] by T

5. Conclusions

The brain mechanism has been investigated from a viewpoint of activity. Although every kind of our behavior possesses some meaning, our reasoning depends on the occasion. Every reaction will be changed by results of the action. The satisfaction depends on the state of demands. An immutable purpose of engineering is "to supply for necessary economically".

There are reports on uses of wavelet transforms for speaker independent speech recognition. But in this paper, the use of H-WT for task oriented treatment of automatic voiced command decoder is reported.

The simple calculation of H-WT for characteristic extraction of voice is attractive. It economizes the calculations of speech recognition. We hope that this research will bring good harvests for the task oriented approach of speech recognition.

References

[1] S. Karasawa, "Brain mechanism on understanding of information explained by concept of activity", IEICE Technical Report TL2007-2, pp.5-10, ISSSN 0913-5685, 2007.

[2] S. Karasawa, "Activity transfer models for associative activities in a brain", Proceedings of Language Sense on Computer, part of PRICAI-04, ISBN: 1-877314-32-3, pp.18-25, Auckland, New Zealand, 2004.

[3] S. Karasawa, J. Oomori, Impulse circuits for a distributed control inspired by the neuro-anatomical structure of a cerebellum, pp.185-190, Intelligent Engineering System through Artificial Neural Networks, Vol.10, ASME Press series, 2000.

[4] B.T.Tan, M. Fu, A. Spray, F. Dermody, "The Use of Wavelet Transforms in Phoneme Recognition", Inter. Conference on Spoken Language Processing, 1996.

[5] C.J.Long, S.Datta, "Wavelet Based Feature Extraction for Phoneme Recognition", Inter. Conference on Spoken Language Processing, 1996.

[6] Bing-Fei Wu, Kum-Ching Wang, Voice activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator, Computational Linguistics and Chinese Language Processing, Vol.11, No.1, pp.87-100, March 2006.

[7] S. Karasawa, H. Sakuraba, "Use of Haar Wavelet Transform Based Multiple Template Matching for Analyses of Speech Voice", Euro American Conference on Telematics and Information Systems, No.78. Faro, Portugal, 2007.

[8] S. Karasawa, H. Sakuraba, "Making Use of Wavelet Transform in Template Matching for Phoneme Analyses of Japanese Voice", IEICE Technical Report, ISSN 0913-5685, SP2006-98, pp.77-82, 2007.

The software on Haar wavelet transform was referred to Y.Okawa "Mathematical processing for abstraction of characteristics on wave from, ISBN4-7898-3715-7, pp.66". The software on A/D conversion was referred to Y.Okawa "Excel for data processing, ISBN4-7775-1010-7, pp.182".