A Functional Model for Acquisition of Vowel-like Phonemes and Spoken Words Based on Clustering Method

Tomio Takara, Eiji Yoshinaga, Chiaki Takushi, and Toru Hirata*
* University of the Ryukyus, Okinawa, Japan
E-mail: takara@ie.u-ryukyu.ac.jp Tel: +81-98-895-8718

***Abstract*— A new born baby can gradually acquire spoken words in the condition where he/she is merely exposed to many linguistic sounds. In this paper, we propose a functional model of such acquisition of spoken words, in which first, vowel-like phonemes are automatically acquired and then the words are acquired using the words represented by these quasi-vowels. This model was applied to command words used for a robot. We implemented this model into a new clustering algorithm for word HMMs. Using this model, the acquisition of spoken words was performed with reasonably high recognition score even though a few phonemes were used. Then the proposed model was shown to represent the early stage of human process of spoken words' acquisition.**

I. INTRODUCTION

Human infants become to be able to discriminate basic phonemes such as vowels without instruction. They are merely exposed to speech sound of their mother language [1]. This is thought to be done by a self- learning effectively using statistical feature of the speech sound.

We model this infants' acquisition process into an engineering algorithm, in which an infant acquires phonemes using only the statistical feature of the speech, then acquires words expressed with these phonemes. The self-leaning without teaching can be modeled into the clustering that is also called an unsupervised leaning. Using the model, we test whether words can be acquired by only using the statistical feature even though the distribution of speech parameters is very complicated.

In ACORNS research project, they model the acquisition process of spoken language from a view point of emphasizing infant's skill of word detection from continuous speech [2].

In the above research, however, the acquisition process of phonemes is not adopted explicitly in the model. We think that the acquisitions of phonemes and words are different processes because the phonemes are acquired also by creatures other than human [1], and the word has meaning, then which is only for human.

Therefore, in this research, we construct and study a model in which, first some phonemes are acquired in the unsupervised learning and then words, which are expressed with these phonemes, are acquired in the supervised leaning. We expect the first acquired phonemes will be vowel-like one. We adopt Hidden Markov model (HMM) for a data structure of words and for a fundamental recognition algorithm. We evaluated the model in robot's acquisition of instructional words using digit words.

We showed experimentally that the quasi-phonemes can be acquired automatically only using statistical feature of speech sound, and the spoken words represented by these quasi-phonemes can be artificially acquired only assuming the pointing skill. .

II. ACQUISITION OF PHONEMES

Human infants become to be able to discriminate vowels without teaching where they are merely exposed to speech sound of their mother language [1].

This process of vowel acquisition is explained that prototypes are detected from statistical distribution in the feature space of speech parameters, and categories are constructed with the magnet from the prototypes. Not only human but also the other creatures have the skill of this categorization [1].

Automatic categorization can be modeled into the clustering [3] in engineering, in which correct discrimination is done by itself without supervision.

We model the infant's acquisition of vowel into the clustering which uses a statistical distribution of speech spectra. In other words, we hypothesize that only infants' skill of categorization is needed for the acquisition of phonemes. The model is that, after an infant clusters the well listened sound, as a result, vowels are first acquired from his/her language. The well listened sound is considered here lauder, continuous and higher pitch voices which are characteristics of voice that mothers speak to their babies [4]. In this study we adopted such a lauder voice that exceeds a threshold of speech power (C0) and such a continuous speech whose Euclidian distance between neighbor frames falls below a threshold.

A. Acquisition of vowel like-phonemes

Speech parameters used in this study are MFCC and FMS [5], which is Fourier transform of a spectrum expressed by Mel-scale frequency and Sone-scale amplitude. The clustering algorithms are K-means clustering and hierarchical clustering. Speech database is "Tohoku University and Panasonic

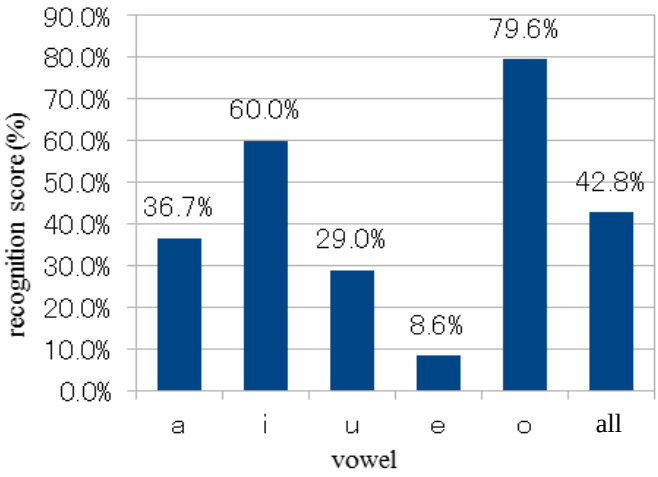

Fig. 1: Results of the hierarchical clustering (recognition score [%])

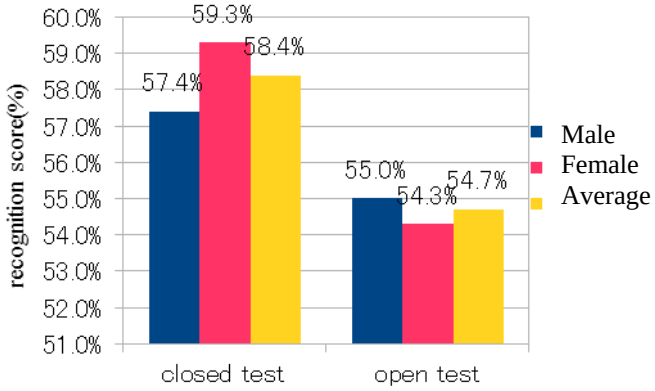

Fig. 2: Results of the nearest neighbor recognition method (recognition score).

isolated spoken word database", which has phoneme balanced 212 words whose frames are labeled with phonemes [6]. We used 10% frames from these data.

The sampling frequency is 11.025Hz and the quantization is 16 bit. The FMS analysis is done with a frame length 25.6ms and a frame shift is 10ms. The frame length of MFCC analysis is 16ms and a frame shift is 10ms.

B. Clustering algorithm

In the *K*-means clustering, first we set the number of clusters *K*. Starting from arbitrary cluster centers, each pattern is attached to a cluster whose cluster center is the nearest to this pattern. A cluster center is calculated to be an averaged vector at each resultant new cluster. Next, each pattern is attached to the new cluster centers. These procedures are repeated until the cluster centers do not change.

In hierarchical clustering, first all patterns are regarded as clusters with one member. Euclidean distances are calculated among all patterns, and then a new cluster is made by combining the nearest pair clusters. This procedure is repeated until the number of clusters becomes the value already set. We set this value so that the largest five clusters include 75 % of all training patterns.

C. Experimental result

The well listened speech was detected automatically using the MFCC analysis and the threshold of speech power C0 and the above mentioned parameter of the continuity. The detected frames were analyzed to be FMSs and used for the hierarchical clustering. The result is shown in Figure 1, where the correct rate represents percentage of the indicated vowel in the five largest clusters. The average correct rate was 42.8 % whereas it was 44.0 % using the MFCC parameter. Some vowels have very low scores in Figure 1 because better prototypes may be in the other clusters than the largest five. We use the word quasi-vowel after this because they are not correct vowels but vowel-liken one in a precision or correctness of 42.8%.

The cluster centers were calculated for the clusters made by the hierarchical clustering using newly selected speech data frames labeled as vowels. These cluster centers will be used after this for the

prototype vector of the clusters in the acquisition model of words. We evaluated these prototype vectors whether they are reasonable feature parameter of vowels in the nearest neighbor recognition method. The speech data were uttered by three males and three females. The result is shown in Figure 2, where the closed test means the experiment in which the tested data are uttered by the same speaker as training and the open test uses that of different speakers.

III. ACQUISITION OF WORDS

A. Process of the model of word's acquisition

Human's process of acquisition of language in early stage is divided as follows.

Pre-linguistic period: after the birth until 12 months old
One word uttering period: 12 months to 18 months old
Two words sentence period: after 18 months old

In this study, we model the acquisition process of words in the pre-linguistic period into the unsupervised leaning and the supervised leaning, and the process in the one word uttering period into the active learning.

In pre-linguistic period, infants are exposed to speech sound uttered by mothers and the others. They gradually discriminate some speech sounds. And then they understand meaning of some words.

We model this phenomenon into the unsupervised learning. The unsupervised leaning can be implemented by using the clustering algorithm. We adopt, in this study, Hidden Markov Model (HMM) for a data structure of words and for a fundamental recognition algorithm. And we propose the declining threshold method that is a new clustering algorithm for the unsupervised learning of HMM.

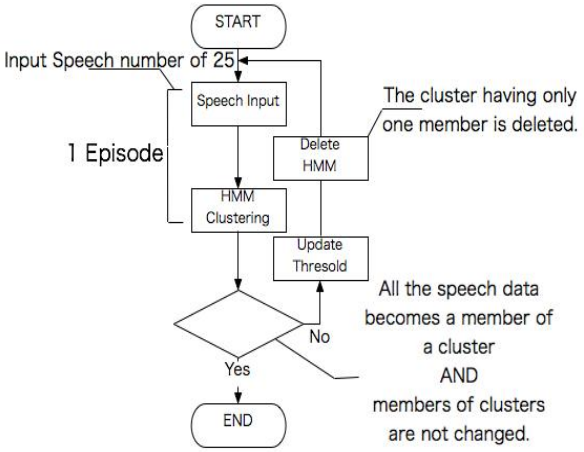

Fig. 3: HMM constructing method using declining threshold.

The supervised learning links a spoken word (HMM) to a meaning (an action in a robot's case). A supervised learning needs a pointing method for dialoging people to concentrate their conscious to one object in a same time. We adopt a word YOSHI (OK) for such a pointing in this study. We hypothesize this word to be recognized inherently.

In our model of acquisition of words, we only hypothesize the pointing skill of human infants, which the other animals also have.

B. Spoken word represented by vowel sequence

As mentioned above, in the pre-linguistic period, phonemes are acquired in the condition where an infant is merely exposed to speech. Vowels are thought to be acquired in an early stage. We model these processes into expressing words in the prototype vectors (the averaged vector at a cluster) of vowel like-phonemes acquired by the method mentioned in the previous chapter. In other words, we adopt the above mentioned prototype vectors in place of usual code vectors of the vector quantization.

C. Unsupervised learning in the pre-linguistic period

The unsupervised leaning can be implemented by using the clustering algorithm. We adopt HMM for a data structure of words and for a fundamental recognition algorithm. We propose the declining threshold method that is a new clustering algorithm for the unsupervised learning of HMM.

First we explain about the HMM clustering method with a static threshold. The threshold is fixed and speech data are inputted at random. In the beginning, a HMM for the inputted speech data is created and used as a representative of the cluster. From the second input, its likelihood is calculated using HMM of each cluster. When the likelihood of inputted data exceeds the threshold, the cluster with the highest likelihood is selected. An inputted data is added as a new member of the cluster, and the HMM is updated. If there is no cluster with an enough likelihood that exceeds the threshold, a new HMM is created and its new cluster is formed. This flow is repeated for all data, renewing the HMM of the cluster.

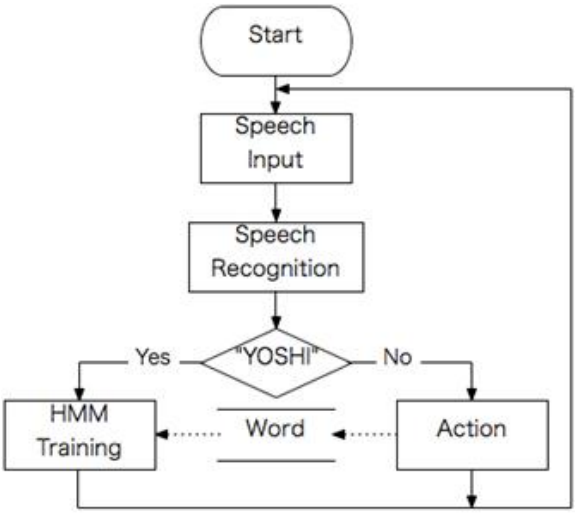

Fig. 4: Supervised learning in the pre-linguistic period

We tested this algorithm using some deferent threshold. As results, the number of clusters with only one member is decreasing when the threshold is lowered. However, when we checked the member of the clusters, the same words have gathered but another words also included. As a result, we found that we cannot get correct clusters using the static threshold method.

Therefore, we propose a new HMM constructing method using declining threshold shown in Figure 3. Speech data are inputted and the clustering is performed using the representative HMM. We define this process as one episode. The threshold is updated whenever one episode is finished, and a cluster with only one member is deleted. The episode is repeated until all the speech data become members of clusters and members of clusters are not changed.

We tested this algorithm using five words' vocabulary. The cluster was formed with the same word while the episode was repeated. Moreover, five clusters were formed by using all the inputted words.

This clustering method does not need to specify a final number of clusters because the number of clusters is decided automatically. In this study, we adopt this clustering method as the model of unsupervised learning for spoken words

D. Unsupervised learning of meaning (action)

In our model, unsupervised leaning is performed also in meaning's space. Our model is that the supervised learning is done fast because the number of categories is decreased by the unsupervised learning in the meaning's space.

The "meanings" of this study are actions made by a robot which are represented by vectors consisting of angles of robot's stepping motors. The clustering is performed using these vectors. After this in the supervised learning, words' labels are attached to these clusters.

The clustering using the vectors was performed with the simple clustering and *K*-means algorithm. Then 40 clusters were constructed.

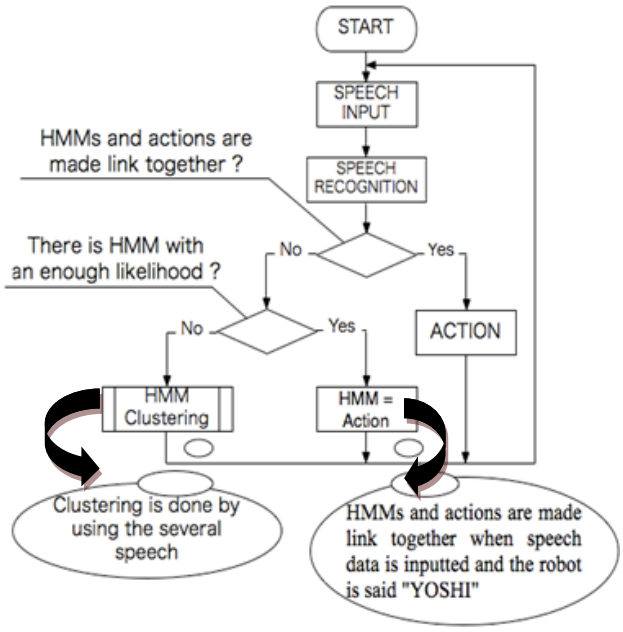

Figure 5: Unsupervised and the supervised learning

It may be no simple relation between the result of this clustering and the meanings of words thought by a human. So we performed classification of meanings (actions) using human's observation. As a result, the cluster of the vectors corresponded 67 % with the human's classification.

E. Supervised learning in the pre-linguistic period

Spoken words are acquired by the flow shown in Figure 4. First, speech is inputted to a robot. The robot can recognize the word. Then the robot selects an action. The selection is done randomly because the robot doesn't know the correct answer. After the selection is done, the robot acts. Additionally, the spoken word is temporarily saved. If the robot's action is incorrect for the inputted word, a user may input the same word again. If the robot performed the correct action, the user says "YOSHI" (OK). The robot recognizes "YOSHI" and trains the HMM using the word temporarily saved. Thereafter, when the robot listens to this word, the robot acts suitably for this word.

Figure 5 shows the flow of acquisition of spoken words using the unsupervised and the supervised learning. Speech data are inputted at random. The HMM clustering method using declining threshold described in the previous section is executed. The created HMM will be used for speech recognition, but these HMMs don't correspond to action at this unsupervised stage.

HMMs and actions are linked together when speech data is inputted and the robot is said "YOSHI" (OK). This means that the robot can label an action as an inputted word. The robot can act correctly if the inputted speech corresponds to the action. If the inputted speech doesn't correspond to an action yet, the robot keeps the speech until the robot is said "YOSHI". Moreover, because HMM is trained enough at HMM clustering stage or the unsupervised learning stage, a spoken word has been acquired at the supervised learning stage. This method needs less inputted data to attain correct recognition score than the method without unsupervised learning.

There are useless speech inputs because the action of the robot is selected at random. Therefore, we propose the action selecting algorithm that uses Yes-No-List. The Yes-No-List consists of two lists. The one list memorizes incorrect action from past inputted words. The other list memorizes actions that have been already linked to the words. Because the meaningless action can be omitted by using these lists, the acquiring time can be reduced very much compared to the method which selects an action randomly.

F. Active learning

A human infant increases explosively his/her vocabulary at about 18 month age. It is because he/she can ask a name of something him/herself. We think that the learning can be done at once and fast because he/she prepares the meaning and gets word which is its label by asking him/herself.

We define the active training as a robot acts itself and learns after a user utters a correct word for the action. In other words, it is the model that a robot links correctly the meaning (action) to a spoken word (HMM) by asking "What is this?"

G. Self-training for speaker independent task

Many HMMs for each speaker are constructed by the declining threshold method when the task is for the speaker independent.

We propose a clustering algorithm in which all non-meaning HMMs are attached to the nearest HMMs with the meaning in the supervised learning. In recognition, the likelihoods of an input are calculated for all HMM, then the recognition result is decided to be the meaning of a cluster that includes the HMM giving the largest likelihood. This is the multiple standard patterns method in pattern recognition theory. Then, sometimes the open tests are better the closed test.

IV. RECOGNITION EXPERIMENT

In order to confirm the effectiveness of this model, we performed recognition experiments. Speech data were Japanese 10 digit words /itʃi/, /ni/, /san/, /jon/, /go/, /roku/, /nana/, /hatʃi/, /kju/, /rei/ uttered 4 times by 6 male speakers, totally 240 tokens. We performed the procedure that, in the pre-linguistic period learning, the HMM clustering was done using the 10 digits words, and 5 digits word /itʃi/ to /go/ were acquired first. The other 5 digits words were acquired in the active learning. Because the result of HMM clustering depends on the order of input, we prepared 10 random patterns of input order and averaged the results.

In the closed test, 30 tokens uttered by one speaker were used for training, and the other 10 tokens uttered by the same speaker were used for test. The tests were repeated 24 times by changing the tested tokens and speakers. The tested tokens were 240. In the open test, 200 tokens uttered by five speakers were used for training, and 40 tokens uttered by another

speaker were used for test. The tests were repeated 6 times by changing the tested speaker. The tested tokens were 240.

For the comparison to the proposed method with prototype vectors, we prepared a code book with code book size 5 whose code vectors were constructed by the hierarchical clustering method or were randomly selected form the original code book with size 64.

The experimental results are shown in Table 1. The recognition score of the usual ASR may be over 97% [7] with monophone and a discrete HMM. From this table, we can see that the proposed method with the prototype vectors acquired by the *K*-means clustering attains a recognition score better than that of the method with code book size 5 and comparable to that of the traditional code size 64. The method using the hierarchical clustering attains the recognition score better than that of the method with the random code book size 5. The prototype of the hierarchical clustering could not attain better score than the method with the hierarchically constructed 5 codes at this stage. We think this method needs one more clustering stage. After the improvement, the method will attain the score comparable to that of the *K*-means method.

Table1. Experimental results (recognition score [%]).

	closed test	Open test
Prototype: *K*-means method	83.2	83.6
Prototype: hierarchical method	65.2	67.0
Code book size 64	83.9	82.6
Code book size 5 (hierarchically constructed)	72.5	72.1
Code book size 5 (randomly selected)	50.9	44.9

V. CONCLUSIONS

We proposed and studied a model in which vowel-like phonemes are acquired first in the unsupervised learning and then words expressed with these quasi-vowels are acquired in the supervised leaning. We adopted HMM structure for word's data structure and fundamental recognition algorithm. We evaluated the model in robot's acquisition of command words using the spoken digit words' recognition.

First we found that vowel-like phonemes can be acquired automatically with a recognition accuracy of 42.8% as a result of the model of phoneme acquisition process in a clustering of spectra. Next, we expressed spoken words with these only five quasi-vowels and applied the words to spoken words recognition. As a result, a high recognition score 83.6% was obtained in a speaker open test.

We showed experimentally that quasi-phonemes can be acquired automatically only using statistical feature of speech sound, and spoken words represented by these quasi-phonemes can be artificially acquired only assuming the pointing skill. The proposed model was shown to represent early stage of human process of spoken words' acquisition.

REFERENCES

[1] Kuhl, P.K. et al,"Phonetic Learning as a Pathway to Language: New Data and Native Language Magnet Theory Expanded (NLM-e)", Phil. Trans. R. Soc. B, 363, 979–1000, 2008.

[2] ACORNS, "An overview; results of the first two years", http://www.acornsproject.org/documents/index.html#Top

[3] Bow, S-T., "Clustering Analysis and Nonsupervised Learning", Pattern Recognition Application to Large Data-Set Problem, Marcel Dekker, Inc., 98-156, 1984.

[4] Kuhl, P. K.., "Early Language Acquisition: Cracking the Speech Code", Nature Reviews, Neuroscience, Volume 5, 831-843, 2004.

[5] Takara, T, Higa, K, Nagayama, I, "Isolated Word Recognition Using the HMM Structure Selected by the Genetic Algorithm", IEEE ICASSP, 967-970, 1997.

[6] Makino, S., Niyata, K., Mafune, M., Kido, K., "Tohoku University and Panasonic isolated spoken word database", Acoustical society of Japan, 42, 12, 899-905, 1992.

[7] Takara, T., Matayoshi, N., Higa, K.: "Connected Spoken Word Recognition Using a Many-State Markov Model", International Conference on Spoken Language Processing, 235-238, 1994.