人間型ロボットによる自律的な音韻獲得と物体動作の関連付け

金 国林[†] 鈴木 健嗣[†]

筑波大学大学院システム情報工学研究科[†]

1. INTRODUCTION

Most of the robot's word learning systems is paid attention to the association among visual images and the meaning of given words. However, these models are determined by the meanings of word in terms of other symbols. There also have a few models base on the learning of visually-grounded words and syntax without the aid of human annotations or transcriptions [1], but assumption inherent in the phonetic feature of words is that the speech is represented in terms of sequences of the given or pre-categorized phoneme.

We have been developing a robotic system with an autonomous phoneme acquisition [2]. Inspired by infants' early word learning, we consider that words are bind directly to non-symbolic perceptual physical features: such as visual features of the given frame and acoustic features of the given utterance, without any prior knowledge about the phonemes and images. Regarding the phoneme description, we proposed a method of categorical phonetic feature map (CPF map [2]) by employing Mel Frequency Cepstrum Coefficients obtained from spoken speeches. However, the system was able to deal with the static image, and image sequences of dynamic objects. In this paper, we report the current progress of the proposed learning system for word acquisition, in particular, an extended system which is capable of learning the object movement and its word representation.

2. SYSTEM OVERVIEW

The system aims at associating the object movement and its word representation. Human understands the object movement by observing it with its color and structure information, and typical movement is based on the translation and rotation. In this study, the system mainly consists of vision and listening modules, and those will be bind by given utterances according to the object movement. In order to understand the movement of the given object, we utilized the mean-shift algorithm. The central of the extracted area which include the target object is calculated in every frame, and examples of these detected sequences of the movement is illustrated in the left figures of Fig. 2.

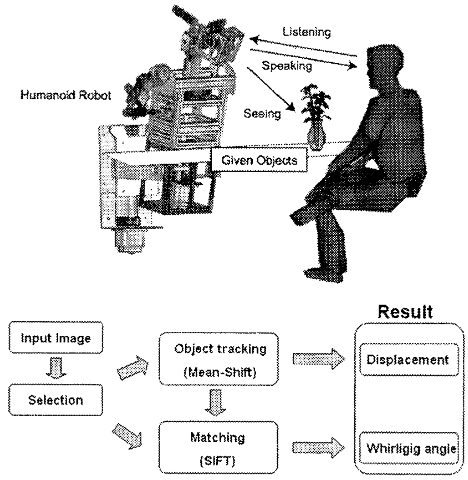


Fig 1. The processing of understanding object movement

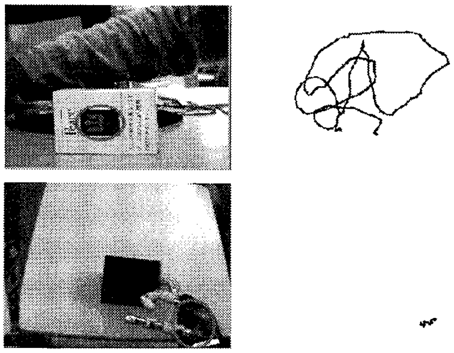


Fig. 2 Tracking object by Mean-Shift

In order to obtain the angular property of the movement, we use the first and last frame, and a matching between them with SIFT algorithm is introduced to describe the the angular property of the movement. By using SIFT, the system is capable to obtain the orientations $O_1(i)$and $O_2(i)$ that correspond the aspect of each feature point of each frame, respectively. The normalized rotational angle of the movement can be then calculated as follows.

$$Angle = \frac{1}{N}\sum_{i=1}^{N}(O_1(i) - O_2(i)) \qquad (1)$$

Here N is the number of the feature points used for the matching process. Hence the system obtains the visual features of the movement such as the pathway and orientations.

Autonomous phoneme acquisition and its linkage to object movement by humanoid robot

Gukleem Kim[†] Kenji Suzuki[†]
[†]Graduate School of Systems and Information Engineering, University of Tsukuba

Rotated Angle (counter-clockwise: 1.56) SIFT Feature Point : 15

Rotated Angle (counter-clockwise: 0.63) SIFT Feature Point : 6

Fig. 3 The result of SIFT matching

3. LISTENING AND SPEECH SYSTHESIS

In the listening module, the first step is to calculate MFCCs from the given speech in order to make the categorical phonetic feature map (CFP map)[2]. MFCCs are short-term spectral features, which are known as the dominant features for speech recognition. The mel-weighted cepstral coefficients are calculated after a nonlinear warping onto perceptual frequency scale. Thus, the speech signal $s(\eta)$ (a segmented word η) is represented as a set of vectors:

$$s(\eta) = \{c_i(\tau)\} \qquad (\tau = 1,2,...,L; i = 1,...,M) \tag{2}$$

$c_i(\tau)$ represents MFCCs. L and M denote the number of frames for each speech signal and the number of Mel cepstrum coefficients, respectively. The sampling rate is chosen at 11.050kHz. The frame length is 512(46ms). Every frame overlaps by 256 samples (23ms). Therefore, a M parameter representation of the Mel transformed spectrum is estimated from a 46 ms window of speech signal. At the sampling rate of 11kHz, the appropriate number of coefficient is suggested to around 15. The increase of number provides better characteristics without missing phonetic property.

In this experiment, we show 5 Japanese vowels of males and a female. Each subject is requested to make speech one-by-one like /a/, /i/, /u/, /e/, /o/, and totally 50seconds. We analyzed 100 seconds sampled sound in 13 dimensional data sequence. In order to visulaize the result, we use PCA to reduce from 13 dimensional data sequence to 2D data sequence. Fig. 4 shows the mapping result in 2 dimentional data sequence. The sympols which correspond to each vowel in Fig. 4 is shown in Table 1.

From Fig. 4, we can see that each vowel is categorized in different regions in 2 dimension coordinate. From the figure, we can also understand that the regions of same vowel but made by male and female positions more close to each other then different vowels. In addition, we inverse MFCCs sequence back to speech waveform by modulating a noise, which is illustrated in Fig. 5

Table 1: The Characters in Fig. 4

/a/	/i/	/u/	/e/	/o/
○	◊	□	△	⬡

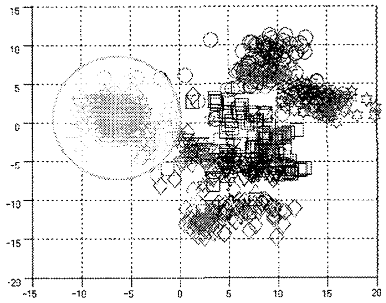
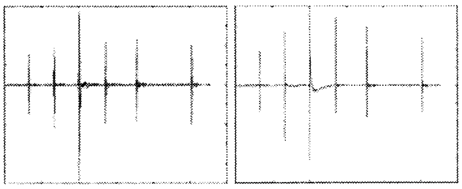


Fig. 4. The result of sound analysis, here red point is made from female subject and blue is made by male subject. The gray area are categorized as noise sound.



(a) (b)

Fig. 5. The result of speech sythesis by MFCCs, here the left fig. is original waveform and the right one is result that from MFCCs.

5. CONCLUTION AND FUTURE WORKS

We have been developing a robotic system which has an ability of autonomous phoneme acquitision. The developed system is able to calculate and make Categorical Phonetic Feature map from the utterance given by human. The robot can understand the word by seeing the movement of given object and listening to the utterance from human. In addition to that, the sysmte is able to utter the speech based on the CPF map based on MFCCs parameters.

From the result of the experiment, we consider that the system is able to understand the object movement solely by using the object's color information, and structural information. We also verified that it is possible to make CPF map and we also can make the speech waveform from the sequence of MFCCs. The future consideration is to associate among given speech and obtained the image sequence in order to improve the CPF map.

REFERENCES

[1] D. Roy, 2005, "Grounding Words in Perception and Action: Insights from Computational Models," *Trends in Cognitive Science, 9(8)*, pp.389-396

[2] K. Suzuki et al., "Learning from Long-term and Multimodal Interaction between Human and Humanoid Robot," IEEE IECON2008, USA, 2008.

[3] Lowe, D.G. (2004), "Distinctive Image Feature from Scale-Invariant Key points", *International Journal of Computer Vision.*