Blind Speech Separation System for Humanoid Robot with FastICA for Audio Filtering and Separation

Widodo Budiharto*, Alexander Agung Santoso Gunawan**
School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
* wbudiharto@binus.edu; ** aagung@binus.edu

ABSTRACT

Nowadays, there are many developments in building intelligent humanoid robot, mainly in order to handle voice and image. In this research, we propose blind speech separation system using FastICA for audio filtering and separation that can be used in education or entertainment. Our main problem is to separate the multi speech sources and also to filter irrelevant noises. After speech separation step, the results will be integrated with our previous speech and face recognition system which is based on Bioloid GP robot and Raspberry Pi 2 as controller. The experimental results show the accuracy of our blind speech separation system is about 88% in command and query recognition cases..

Keywords: humanoid robot, blind speech separation, speech recognition, face recognition, fastICA

1. INTRODUCTION

As the time goes by, robot technologies are already familiar to find in many fields. Recently, there are many researches focusing in how to build a robot that can emulate the human abilities. The main features of humanoid robot in enganging with users; has proven to be a challenging research area, for example the service robot which can ask and deliver an order to its customers. In the service robot, the first task is to detect and recognize its customers. For object detection, robot is expected to detect the customer face and then track its position. While in object recognition, robot is designed to recognize the face based on the customer database. Generally, face recognition is done in the controlled variables, such as light intensity, poses, and expression. Uncontrolled variables would decrease the performance of face recognition. The uncontrolled variables could be expression, wearable accessories (glasses, etc), lightness and object motion. The similar steps are applicable when the service robot deals with sound to identify the customer order. The main problems related to sound is to separate the multi speech sources and also to filter unrrelevant noises, both in moving and stationary sound sources. Real-world applications of service robot should consider these problems. However, most studies assume only dealing to one stationary sound source [1].

Independent Component Analysis (ICA) is one of the decomposition methods which is capable of decomposing multivariate signals into additive subcomponents, based on the assumption that the source components are statistically independent from each other, and follow a non-gaussian distribution. The challenge in how to extract one specific source signal from a mixture of many signal sources is known as the Cocktail Party Effect [2]. On a hypothetic cocktail party, there are various speakers, who speak more or less randomly. The objective is to isolate the speech of one specific speaker from all the other sounds made by the rest of the party. A group of techniques for solving this problem, which need no prior knowledge about the mixing ratios of the various sources, are called Blind Source Separation or Blind Signal Separation (BSS) [3]. In Takatani et al research [4], they addressed a blind decomposition problem of binaural mixed signals observed at the ears of humanoid robot, and introduced blind signal decomposition algorithm using single-input multiple-output-model-based ICA (SIMO-ICA). The SIMO-ICA consists of multiple ICAs and reliability controller. Each ICA will run in parallel under the reliability control of the entire separation system.

In this paper, the research is based on a low cost humanoid robot Bioloid GP robot and Raspberry Pi 2 model B as controller. The speech and face recognition system is built the Linux-based Raspian OS, OpenCV 2.4.10 as vision library and Python 2.7 as default programming language. In Fig. 1(a), it is shown our robot architecture called as BeeHumanoid Ver 2.0, which divided into three modules that are: controller using Rasberry Pi 2 [5], audio and visual sensors, and humanoid robot using Bioloid GP. While the sides view appearance of our robot architecture including humanoid robot, embedded controller and visual sensor can be seen in Fig 1(b). The robot architecture has been designed in our previous research [6], which focuses on face and speech recognition.

First International Workshop on Pattern Recognition, edited by Xudong Jiang, Guojian Chen, Genci Capi, Chiharu Ishii, Proc. of SPIE Vol. 10011, 1001113

(a) (b)

Fig.1. Our humanoid robot called as BeeHumanoid (a) robot architecture (b) side view appearance

2. AUDIO FILTERING AND SEPARATION

2.1 Independent Component Analysis (ICA)

Independent component analysis (ICA) is applied to separate a multivariate signal into additive subcomponents. FastICA [7] is an efficient and popular algorithm for ICA, by exploiting a fixed-point iteration scheme. The fastICA algorithm have a number of desirable properties when compared with standard ICA, mainly due to its convergence is cubic (or at least quadratic). For our research, it is assumed two persons are speaking simultaneously and then the mixed speech is recorded into two signals. These signals are denoted as $x_1(t)$ and $x_2(t)$, with x_1 and x_2 the amplitudes, and t the time index. If the speech signals from the two speakers, denoted by $s_1(t)$ and $s_2(t)$, then the mixed signals could be expressed as a linear equation:

$$x_1(t) = a_{11}\, s_1(t) + a_{12}\, s_2(t) \quad (1)$$

$$x_2(t) = a_{21}\, s_1(t) + a_{22}\, s_2(t) \quad (2)$$

Where a_{11}, a_{12}, a_{21}, and a_{22} are some parameters which depends on the distances between the microphones and the speakers. The main problem in here is to estimate the two original speech signals $s_1(t)$ and $s_2(t)$, based on the recorded signals $x_1(t)$ and $x_2(t)$. In our research, it is used fastICA [6] which is based on a fixed-point iteration for finding weight vector $\mathbf{w}$ in order to get maximum of the nongaussianity of $\mathbf{w}^T\mathbf{x}$. Finally, we propose an algorithm for blind speech separation which is incorporated to speech and face recognition system in algorithm 1:

Algorithm 1: Speech and Face recognition System for Humanoid Robot

Get input image from the camera
Detect and recognize face using PCA
If face detected then
  Do
    Get audio from user
    Separating speech process using fastICA
    Recognize each speech using Google Speech Recognition API
    Response with action
  Loop
Else
  Robot standby
Endif

3. EXPERIMENTAL RESULTS

The approach proposed in this paper was implemented and tested on a humanoid Robot named BeeHumanoid Ver 2.0 which can been in Fig 1. First experiment is to test the system ability for filtering non-gaussian noise. The result is

shown in Fig 2, where the mixed audio (Source#1) consist of speech signal and non-gaussian noise. The graphic shows that the fastICA algorithm able to separate speech signal from the non-gaussian noise, which in the form of sinus signal.

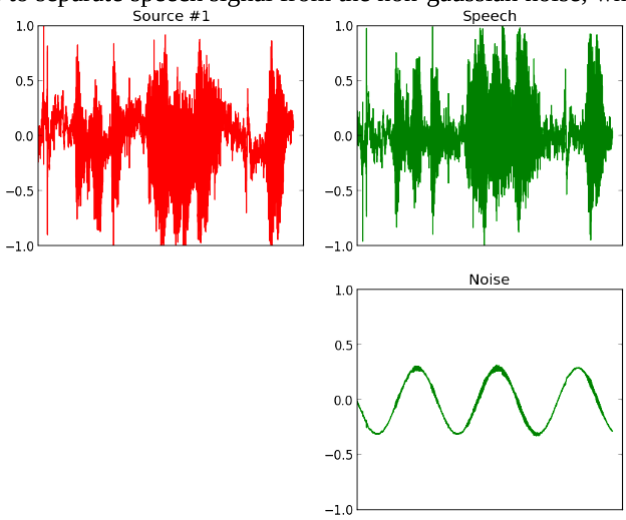

Fig 2. Filtering the non-gaussian noise

Second experiment is to test the ability of blind speech separation from 2 peoples (Source#1 and Source#2) that speak together as shown in Fig 3:

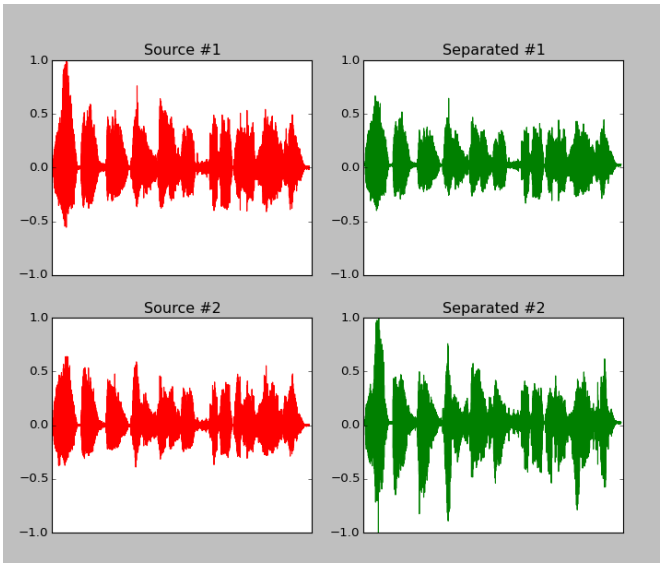

Fig 3. The blind speech separation using fastICA

The design of our controlled experiments is robot initially greets the user and asks the user's query, and user will give a short question in the middle of noisy chattered environment. In the end, robot recognize this query in two conditions: without filtering and with filtering data using fastICA. In the second experiment, robot is given one word command by the user which have to be responded, such as move, walk, stop etc. And finally robot try to recognize this command in two above conditions. The result of this experiment (Table 1) is counted from the number of true positive result from 10 trial of five persons. Of course, the '*command with filtering*' case tends to be the best case in our experiments. First, because of the nature of command is short and clear, then it is easier to be recognized. Second, the extracted speech recognition of filtering data from noisy chit-chat is more consistent to be true.

Table 1. Result of experiment with speech recognition

Subject	Actions			
	Query with filtering	Query without filtering	Commands with filtering	Command without filtering
1	9	7	9	7
2	8	7	9	7
3	9	8	10	8
4	8	6	9	7
5	8	7	9	7
Average	84%	70%	92%	72%

In addition, the ability of face recognition system which using Principal Component Analysis (PCA) is also tested and its accuracy result is about 93%. PCA is a standard technique that is used in statistical pattern recognition and signal processing for data reduction and extraction features [8]. Vision system which is one of the main sources for environment information, is very challenging since there are many uncontrolled visual variables in environment [9]. The illustration of our face recognition system is shown in Fig 4.

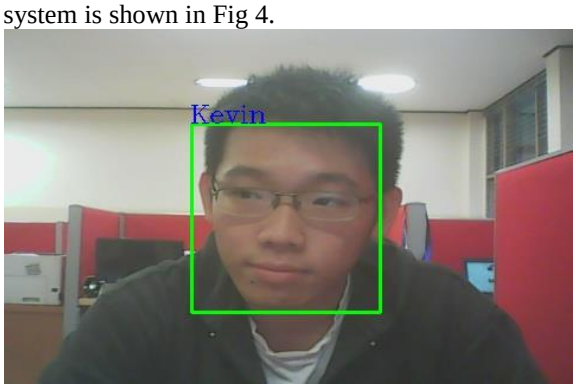

Fig.4. Recognized Face

4. CONCLUSION

In this paper, we introduced the architecture on our humanoid robot based on Bioloid GP and Raspberry Pi 2 that has reliable ability for speech and image processing. Based on this architecture, our experiments by integrating fastICA in command and query recognition will give the accuracy about 88% . For the future work, our humanoid robot will be developed to able to detect and recognize multiple faces and multiple commands.

REFERENCES

[1] Nakadai, K., Nakajima, H., Hasegawa, Y., Tsujino, H., “Sound source separation of moving speakers for robot audition”, IEEE International Conference on Acoustics, Speech and Signal Processing, 3685 - 3688. (2009). DOI: 10.1109/ICASSP.2009.4960426.

[2] Barry Arons, “A Review of The Cocktail Party Effect”, Journal of the American Voice I/O Society, vol. 12, (1992).

[3] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park and Soo-Young Lee, “Blind source separation and independent component analysis: A review”, Neural Information Processing-Letters and Review, v6. 1-57, (2005).

[4] Takatani, T., Ukai, S., "Blind sound scene decomposition for robot audition using SIMO-model-based ICA ", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), 2247 - 2252, (2005)

[5] Raspberry Pi Foundation , "Raspberry Pi 2 Model B", 2015, https://www.raspberrypi.org/products/raspberry-pi-2-model-b/. (January 2016).

[6] Widodo Budiharto, Alexander A S Gunawan, Azani C Sari, Heri Ngarianto, “Designing of Humanoid Robot with Voice Recognition Capability”, International Conference on Intelligent Systems and Image Processing (ICISIP), Fukuoka – Japan, (2015).

[7] Hyvarinen, A. "Fast and robust fixed-point algorithms for independent component analysis". IEEE Transactions on Neural Networks, 10 (3): 626–634, (1999)

[8] Turk, M., and Pentland, “Eigenfaces for recognition”, Journal of Cognitive and Neuroscience, vol 3(1), pp. 72-86, (1991).

[9] Szeliski, R., “Computer vision: algorithms and applications”, Springer-Verlag New York Inc, (2010).