Separating Multi Speeches in Intelligent Humanoid Robot using FastICA

Heri Ngarianto[1], Alexander A S Gunawan[2], Widodo Budiharto[3]

***Abstract*—The main objective of our research is to develop an intelligent humanoid robot for teaching children by listening and answering the questions. In our previous research, we have designed a humanoid robot that can detect human face and receive commands by using speech recognition. Our robot is based on Bioloid GP robot and Raspberry Pi2 as control system. In this study, we would like to expand the capability of the robot system in order to isolate the speech of one speaker from all the other sounds. The problem for separating multi speeches from stereo audio record is called as Blind Speech Separation (BSS). We propose FastICA algorithm to solve the BSS problem. FastICA is an efficient algorithm to separate several signals based on Independent Component Analysis (ICA) algorithm. Some assumption must be met to use FastICA, that is the number of mixtures are equal to the number of sources and the sources are linearly independent from each other. To evaluate the algorithm, we use several simulations based on two speech sources and its mixing matrix. Our simulation shows FastICA algorithm can solve BSS problem by separating two sound signals, but its linearly independent assumption makes it difficult to implement in our humanoid robot.**

Keywords—humanoid robot, education, Bioloid GP, Raspberry Pi 2, FastICA, face detection, speech recognition, BSS.

I. Introduction

Recently, researcher in robot technologies focus to develop a robot which can emulate human capabilities. Our main research goal is to develop a humanoid robot for education purposes [1]. One of its ability is to listen and answer the questions, mainly in service robots [2]. In the education robot, the first task is to detect and recognize the student face, then the robot will recognize the student voice and respond to it. Nevertheless, there is still problem for an education robot to separate several voice inputs simultaneously. As result, the robot is not capable to give feedback or corresponding output correctly in natural environment.

This paper is our effort to isolate the speech of one speaker from all the other sounds. The problem for separating multi speeches from mixed audio record is called as Blind Speech Separation (BSS). In here, FastICA algorithm is proposed to solve the BSS problem. FastICA is efficient implementation of Independent Component Analysis (ICA) algorithm. ICA is the general technique of separating an original signal into its components by maximizing non-gaussianity. To use ICA there are some assumption to be fulfilled, that is: the number of mixtures is equal to the number of components and the components are statistically linear independent from each other [3]. While FastICA is a method of performing the ICA technique. FastICA employs Newton's method for approximating negentropy function [4]. Negentropy, in here, is the measurement of non-gaussianity. One promising application of ICA is blind source separation (BSS). Source, in here, means an original signal, and Blind means that we have very little knowledge on the mixing matrix, and make little assumptions on the source signals.

[1]Heri Ngarianto and Widodo Budiharto are with Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. E-mail: hngarianto@gmail.com

[2]Alexander A S Gunawan is with Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. E-mail: aagung@binus.edu.

In our case, the original signal is audio record signal and the components are the sources of the speech [5].

In this paper, we would like to duplicate the human capability in separating multi speeches in our robot by solving BSS problem through FastICA algorithm. The remainder of this paper is composed as follows: first we discuss robot architecture in section 2, and then is followed by FastICA algorithm, in section 3. In section 4, we report the experiment result in blind speech separation based on FastICA. Finally, we summarize the results and suggestion for our future research in section 5.

II. Method

A. Robot Architecture

The main components in our robot architecture are Bioloid GP robot and Raspberry Pi2 as controller. Bioloid GP is a programmable humanoid robot which have high quality motors, several sensors and aluminum structural frames [6]. Raspberry Pi2 is a small single-board computer based on Broadcom BCM2837 SoC with a 1.2 GHz 64-bit quad-core ARM Cortex-A53 processor [7]. In Figure 1, we show our robot architecture in three modules: (i) controller based on Raspberry Pi2, (ii) output and input devices, and (iii) humanoid robot Bioloid GP.

Raspberry Pi2, in our architecture, is used as controller to process data from audio microphone and camera. Furthermore, the captured video will be processed for face detection and the recorded audio will be processed speech recognition. Our new module for solving BSS problem should be placed before the speech recognition module. The audio results will be sent to speaker and the motion responds is sent to robot controller CM530. The CM530 will drive the actuators of the humanoid robot. Motion programming is done through the Bioloid software called as RoboPlus. It allows to design the robot movements and to automate actions in a simple manner. The program is then transferred on the CM-530 controller and the humanoid robot Bioloid GP can move autonomously.

The robot architecture has been designed in our previous research [1], which focuses on receiving command and doing simple speech recognition. The front view of our robot architecture, which consist of a humanoid robot, embedded controller, and web camera can be seen in figure 1. We develop the program module in Raspberry Pi2 based on Python programming language. In the figure 1, laptop is used to access Raspberry Pi2 controller remotely through Wi-Fi. Therefore, we can modify the code and monitor the results in humanoid robot through our laptop. Next subsection, we explain about speech recognition module, in which the new separating multi speeches should be placed. The main input, of course, are two microphones to get the stereo audio records. For this purpose, the Cirrus Logic Audio Card is stacked over Raspberry Pi2 because Raspberry Pi2, by default, can only receive mono audio input.

B. Speech Recognition Module

In speech recognition module, we process a sound signal input from the microphone. This module is developed based on python programming and use pyttsx library to access google translation API. The default spoken language is set to Indonesian language. The result of the audio processing is a wav file (namely output.wav). It will be presented by the humanoid robot through robot's speaker.

To activate the speech recognition module, our humanoid robot, firstly, have to detect the user's face. This activation is indicated by greeting the user and introduce itself (see figure 3). After that, the robot will ask about the user's intentions. The robot can tell a story or just sing a song (see figure 4). Furthermore, the robot also synchronizes its spoken response with its motions. All the conversations between robot and user will use Indonesian language.

III. SEPARATING MULTI SPEECHES

A. Blind Source Separation

In blind source separation (BSS), the observed values of x(t) correspond to a mixed signal. Then the components si(t) are called source signals, which are usually original signals or noise sources [8]. We assume the sources are statistically independent from each other, and thus the signals can be recovered from linear mixtures x(t) by finding a transformation in which the transformed signals linearly independent each other. In this paper, we propose FastICA to solve to BSS problem. Figure 5 illustrates the BSS problem. Next subsection describes how to solve the problem using ICA pproach step by step.

B. Independent Component Analysis (ICA)

Suppose that the mixed signal x(t) has elements $x_1, \ldots, x_n$ and the source signals si(t) where $i=1, \ldots, n$. Let denote matrix A with elements a_{ij}, the mixing model can be written as:

$$x(t) = \sum_{i=1}^{n} a_i s_i(t) = As(t) \quad (1)$$

The above mixing model is called as ICA model, which describes the observed data is generated by a process of mixing the source signals $s_i(t)$. The mixing matrix A is assumed to be unknown and the source signals is latent, that cannot be directly observed. After estimating the matrix A, it can be computed its inverse W and capture the independent components by:

$$s(t) = Wx(t) \quad (2)$$

The main idea to estimate the ICA model is non-gaussianity [9]. Without non-gaussianity, the estimation of ICA model is impossible. Thus, gaussian random variables, which is used in noise model, is excluded to ICA. Based on the Central Limit Theorem, that state a sum of independent random variables has distribution closer to gaussian comparing the original random variables, ICA algorithm will maximize the non-gaussianity to construct the original independent components. To find several independent components, we need to find all local optima.

C. FastICA

FastICA is an efficient implementation of ICA technique. It finds an orthogonal rotation of prewhitened data by maximizing a measure of non-gaussianity of the rotated components. To measure non-gaussianity is used negentropy function and to maximize this function is used Newton's method.

First, the input data matrix x(t) must be prewhitened, before applying the FastICA algorithm. It means the data will be centered and whitened. Centering the data means transforming each component of the input data by subtracting to its means. By using centering, the expected value of the input data will be 0. While whitening the data means transforming the centered data so that its components are uncorrelated and have variance one. It can be done by calculating eigenvalue decomposition on the covariance matrix of the centered data.

To find single component $s_i(t)$, it will be calculated the weight vector w_i (row vector of matrix W in equation 2) which maximizes a measure of non-Gaussianity of the projection w_i x(t). To measure non-Gaussianity, FastICA uses approximation of negentropy function [3], a nonquadratic nonlinear function f(u). In here, we choose f(u) as exponential function, which have simple derivation that is:

$$f(u) = -e^{-u^2/2} \quad (3)$$

and its first and second derivatives are:

$$f'(u) = ue^{-u^2/2} \quad (4)$$
$$f''(u) = (1-u^2)e^{-u^2/2}$$

We use Newton's method to extract the weight vector w_i iteratively. Next, the algorithm of FastICA can be written as following:

1. Choose initial random weight vector w_i
2. If E{} is expected value, then calculate:
   $w^+ = E\{x(t)\, f'(w_i\, x(t))\} - E\{f''(w_i\, x(t))\, w_i$
3. Let $w_i = w^+ / \|w^+\|$
4. If not converged, go back to 2

The above algorithm only extracts a single component. Thus, to estimate additional components requires repeating the algorithm to obtain n linearly independent components.

IV. Results and Discussion

In our simulation, we used just two source speech signals: one is man voice and the other is woman voice. There are several parameters in recording speeches, that is kind of speeches (a word or a query), distance (near or far) and angle to microphone (0^o or 45^o). Furthermore, we maintain the recording environment is free from noise interference. All recorded mixed speeches are in the form of wav file. Figure 6 is example of stereo recorded query: "nama kamu siapa" from man and woman voices.

The result of separating multi speeches using FastICA can be seen in figure 7. The red signals are the ground truth or original signals, and the green signals are the separating results from mixed signal. It can be noticed that the original signals can be recover well, although their amplitude in separating results are change.

Because of the changing amplitude in results, we also make experiments by manipulating the mixing matrix, so the mixed signals have two categories: balance (same loudness from each source signals) and unbalance (different loudness from each source signals).

The experiments are based on these categories together with three experiment parameters: kind of speeches, distance and angle, can be seen in below tables. To evaluate the separating results, we use qualitative assessment by using human auditory perception. If human cannot hear a separating result, it will be considered as bad or 0. On the other hand, if a separating result can be heard well, it will be considered as good or 1. Furthermore, we make ten experiments for each case and report in table as the average of successful results.

A. Experiment 1: Kind of Speeches

There are two kinds of speeches in our experiments, that is word and query. For word, we used simple Indonesian nouns words, such as "kopi", "teh" etc. While for query, we used simple Indonesian sentences, like "nama anda siapa", "saya ingin mendengar sebuah cerita" etc. For this experiment, we set the distance parameter as near and the angle as 0^o. Table 1 is the summary of experiment 1 about the effect of speeches type in the separating results.

From Table 1, we can see that the kind of speeches does not have influence on FastICA algorithm in separating multi-speeches. All the speech sources can be recovered well.

B. Experiment 2: Distance

The distance in here means distance between the man or woman who speak to the microphone. There are two setting in the experiment that is near (around 5 cm) and far (around 30 cm). For this experiment, we set the kind of speeches as a query and the angle as 0^o. Table 2 is the summary of experiment on distance effect in the separating results.

From Table 2, we can see that far distance really have influence on results of FastICA algorithm. The quality of recorded voice in far distance setting is not clear because some environment noise is included in the audio record. As stated in section 3, FastICA algorithm requires non-gaussianity assumption, that means it cannot handle such environment noise.

C. Experiment 3: Angle

The angle means angle between the man or woman who speak to the microphone. There are two setting in the experiment that is 0^o (in front microphone) and 45^o (alongside microphone). In here, we would like to evaluate the impact of microphone sensitivity in recording process. We used a low-cost microphone type that is electret condenser [10]. An electret microphone is a type of capacitor microphone which have good performance and ease of manufacture. Most microphones made nowadays are electret microphones. For this experiment, we set the kind of speeches as a query and the distance as near. Table 3 is the summary of experiment on angle effect in the separating results.

From Table 3, we can see that electret microphone is already suitable for our purpose. The quality of recorded voice is not too sensitive to the direction of voice sources. Thus, we can use the electret microphone in our future research without any worry.

V. Conclusion

In this paper, we would like to present the development of intelligent humanoid robot system by expanding the capability of the robot system to isolate the speech of one speaker from all the other sounds. The problem is called as Blind Speech Separation (BSS). We propose FastICA algorithm to solve the BSS problem. To evaluate FastICA algorithm, we use several simulations based on two speech sources and several research parameters Our simulation shows FastICA algorithm can solve BSS problem by separating two sound signals, but its linearly independent assumption makes it difficult to implement in our humanoid robot. Furthermore, FastICA cannot compensate the environment noise. For future work, it is planned to find other algorithm to solve the BSS problem without linearly independent assumption.

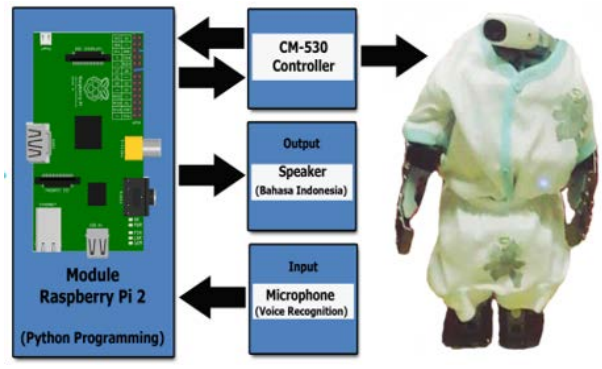

Figure 1. Architecture of our humanoid robot.

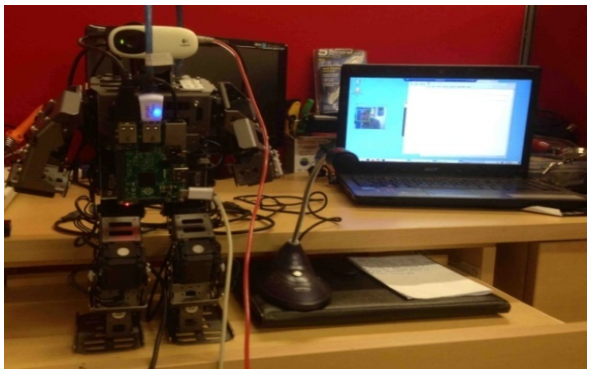

Figure 2. Front view of our humanoid robot.

Figure 3. Face detection for module activation.

Figure 4. User ask to the humanoid robot.

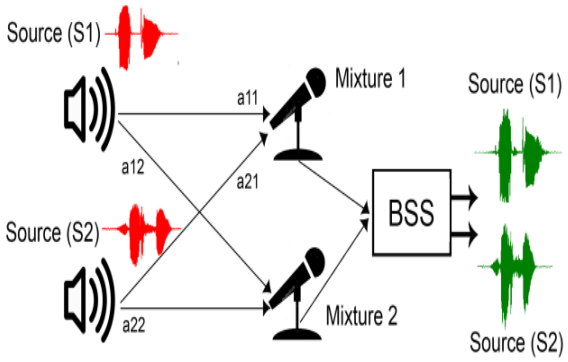

Figure 5. Blind Source Separation (BSS) problem.

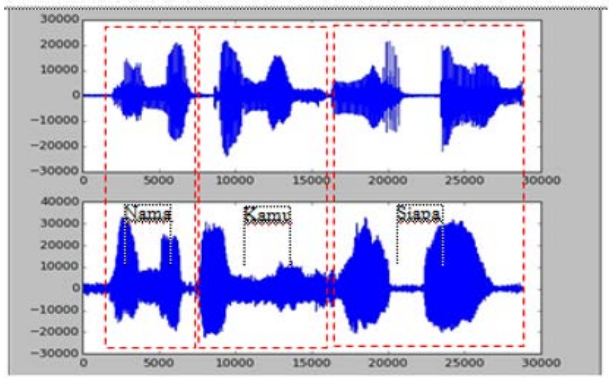

Figure 6. Example of mixed signals in stereo record.

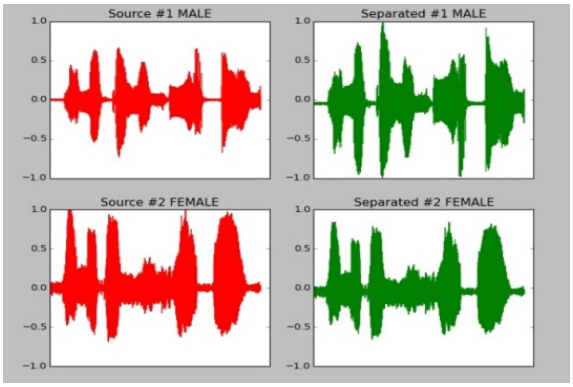

Figure 7. FastICA result

TABLE 1.
KIND OF SPEECHES EXPERIMENT

	Kind of Speeches	
	a word	**a query**
Balance	100%	100%
Unbalance	100%	100%

TABLE 2.
DISTANCE EXPERIMENT

	Distance	
	near	**far**
Balance	100%	60%
Unbalance	100%	40%

TABLE 3.
ANGLE EXPERIMENT

	Angle	
	0º	**45º**
Balance	100%	100%
Unbalance	100%	100%

REFERENCES

[1] W. Budiharto, A. Agung, A. C. Sari, and H. Ngarianto, "Designing of Humanoid Robot with Voice Recognition Capability," in *Proceedings of the 3rd IIAE International Conference onIntelligent Systems and Image Processing*, 2015, pp. 202–205.

[2] Z. Teresa, "History of Service Robots," in *Service Robots and Robotics: Design and Application*, Pennsylvania, USA: IGI Global, 2012, pp. 1–14.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications.," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–30.

[4] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[5] X. Yu, D. Hu, and J. Xu, *Blind source separation : theory and applications*. Singapore: John Wiley, 2014.

[6] ROBOTIS, "ROBOTIS GP." [Online]. Available: http://en.robotis.com/model/board.php?bo_table=print_en&wr_id=31. [Accessed: 24-Jan-2017].

[7] The Raspberry Pi Foundation, "Raspberry Pi 2 Model B -." [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-2-model-b/. [Accessed: 24-Jan-2017].

[8] G. R. Naik and W. Wang, *Blind source separation : advances in theory, algorithms and applications*. Berlin Heidelberg: Springer, 2014.

[9] P. Comon and C. Jutten, *Handbook of blind source separation : independent component analysis and applications*. Oxford, UK: Academic Press, 2010.

[10] G. M. Sessler and J. E. West, "Self-Biased Condenser Microphone with High Capacitance," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1787–1788, Nov. 1962.