Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 87–98

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Implementation of Blind Speech Separation for Intelligent Humanoid Robot using DUET Method

Alexander A S Gunawan[a]*, Albert Stevelino[b], Heri Ngarianto[b], Widodo Budiharto[b], Rini Wongso[b]

[a]*Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
[b] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*

Abstract

Nowadays, there are many efforts in building intelligent humanoid robot and adding advanced ability such as Blind Speech Separation (BSS). BSS is a problem of separation of several speech signals in a real world from mono or stereo audio record. In this research, we implement BSP system using DUET algorithm which allow to separate any number of sources by using only stereo (two) mixtures. The DUET (Degenerate Unmixing Estimation Technique) algorithm replaces our previous FastICA (Fast Independent Component Analysis) method only success in simulation but failed in the implementation. The main problem of FastICA is that it assumes instantaneous mixing without time delay in the recording process. To deals with audio record in the presence of inevitable time delays, it has to be replaced with DUET algorithm to separate well in real time. Finally, the DUET algorithm is implemented to humanoid robot which is developed using Raspberry Pi and equipped with RaspPi Cam to detect human face. Furthermore, the Cirrus Logic Audio Card is stacked to Raspberry Pi in order to record stereo audio. In our experiments, there are three controlled variables to evaluate algorithm performance, that is: distance, number of sources, and subject's name. Robot will record stereo audio for four seconds after face is detected by system. The recording is then separated by DUET algorithm and produce two source estimations with average computation time 1.8 seconds. With Google API, the recognition accuracy of separated speech is varying between 40%-70%.

Keywords: intelligent humanoid robot; DUET; blind speech separation; speech recognition

* Corresponding author. Tel.: +628175001010
E-mail address: aagung@binus.edu

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.014

1. Introduction

Service robots can be defined as robots that sense, think, and act to perform services to the well-being of humans [1]. The service robots have main advantage because they can work nonstop in well consistency. Currently, service robots are developed for various applications, including: edutainment, guidance and office works, inspection and surveillance etc. For a suitable service, human interaction capability is must, then face and voice detection and recognition are being the main features in service robots. Now, researches in detecting face and voice have been developed well over the world, but there is still problem for service robots to separate several voice inputs simultaneously. As result, robots are not capable to give feedback or corresponding output appropriately in natural environment. Actually, the voice sources are not coming just from one direction, but we know that human could separate and recognize the sources very well. Furthermore, the voice sources are also mixed with background noise in natural environment, thus we have to focus auditory attention on a particular source while filtering out the background noise. This phenomenon is called as the cocktail party effect [2]. The objective in here is to isolate the speech of one particular speaker from all the other sounds made by the rest of the party. The techniques for solving this problem, which need no prior knowledge about the mixing ratios of the various sources, is called as Blind Speech Separation (BSS) [3] and our service robots need to implement this technique.

In our initial attempt [4], we employ FastICA (Fast Independent Component Analysis) for solving BSS problem. FastICA is one of the decomposition methods which is capable of decomposing mixing signals into additive subcomponents in real time. In our previous experiment, FastICA method only success in simulation but failed in the implementation because of inevitable time delays. There is several improvements of FastICA, for example Takatani et al [5] proposed blind signal decomposition algorithm based on Single-Input Multiple-Output (SIMO) acoustic signal model using the extended ICA algorithm, called as SIMO-ICA. The separated signal of SIMO-ICA can maintain the spatial qualities of each speech source comparing to the conventional method. Nevertheless, all extensions of ICA algorithm suffer its fundamental assumption that is: the source signals must be independent of each other. The assumption causes the mixing signals must be instantaneous mixtures of sources without time delay. This requirement is impossible to be fulfilled in implementation stage because there is inevitable time delays during stereo input recording.

ICA will fails in case where any delay is present between the sources [6]. Therefore, to deal with the mixtures of sources containing delay, Degenerate Unmixing Estimation Technique (DUET) is developed [7] [8]. Rickard [8] states two or more source can be well separated from the mixing of sources contain small delay with DUET algorithm. The underlying framework of DUET is clustering method similar to van Hulle method [9]. The main different is DUET uses the time–frequency representations, beside van Hulle method employed the space representation and tried to estimate the mixing matrix like ICA approach. Furthermore, DUET can separate blindly an arbitrary number of sources given just two input mixtures [8].

In DUET experimental results [7], the algorithm works very well for synthetic mixtures similar to ICA method. Furthermore, DUET also works well in real mixtures of speech recorded in an anechoic room. Anechoic means it does not represent echoes, that is, reflection of sounds that arrive with a delay after the direct sounds. Because DUET is based on the assumptions of anechoic mixing, the quality of the demixing is reduced in an echoic room experiments. The experiment result also shows that DUET can estimate arbitrary sources using just two mixtures and work optimally at three source estimations. If more than three, the error will increase significantly.

Based on the experiments [7], we will be used DUET to estimate arbitrary sources in humanoid robot for being used as service robot. The robot is designed like a human based on Raspberry Pi platform and equipped with RaspPi Cam to detect human face. Furthermore, the Cirrus Logic Audio Card is stacked to Raspberry Pi in order to record stereo audio from two microphones. Our experiment design is similar to Baeck et al research [10] in order to evaluate the algorithm performance in real environment. The research was conducted to answer the following problem: how is the performance of humanoid robot to estimate the speech sources based on DUET algorithm and convert those sources into text? Hence, the purpose of this research is: (i) develop a humanoid robot that capable to estimate each sources and covert it to text, (ii) evaluate the text conversion accuracy.