ロボット聴覚を目的とした深層学習による音源定位と音源分離の精度向上

Enhancing Sound Source Localization and Separation Using Deep Learning Models for Robot Audition

5115EE02-1 YALTA SOPLIN Nelson Enrique Prof. OGATA Tetsuya

ABSTRACT: This study proposes the use of deep learning (DL) models into two stages of the robot audition, sound source localization (SSL) and sound source separation (SSS), to enhance conventional models used for a robot audition system in each stage, or replace them. Robots have become part of everyday life, and by using their audition system they should interact with humans in reverberant and noisy environments. Recently, DL approaches have performed better at different signal processing tasks, showing that they could perform better than humans. In this study, DL models are employed to replace a conventional method in a SSL task and to enhance common method employed in SSS tasks. DL models surpass the results of conventional methods, and because of their shorter processing time they can be implemented at real-time tasks.

Keywords: Enhancement, Robot Audition, Sound Source Localization, Sound Source Separation, Deep Learning.

1. INTRODUCTION

For naturals interactions with humans, robots should have auditory functions [1]. In a speech interaction, humans look in the direction of the source(s) and filter the desired sound information to continue interacting. In audio processing, the search for the location of the source (i.e., sound source localization (SSL)) and the filter of a desired sound from mixed sounds (i.e., sound source separation (SSS)) are generic problems formulated as parts of the cocktail party effect, which humans can naturally solve. However, Environments add noise to sound signals of a conversation, and its reverberation affects the sound directional characteristics. These environmental characteristics complicate SSL and SSS tasks.

To function effectively, robots must be able to locate, and separate sound source information and reduce the noise from the sound source. Conventional methods proposed to solve these problems are based on:

a) Multiple signal classification (MUSIC), which is used to perform SSL, employs the representations of the energy of signals as well as the time difference of the arrival of those signals to locate sound source(s),

b) Geometrically constrained High-order Decorrelation based Source Separation with Adaptive Step-size control (GHDSS-AS) is a SSS technique proposed in [2]. GHDSS-AS can be applied to track the dynamic changes for robot audition system.

The use of both methods allows to robots *hear with their own ears* [1]. However, the performance of MUSIC methods is related to the number of microphones used for the task and for some implementations of MUSIC [2] is required the pre-calculation of the correlation matrix to perform SSL in noisy environments. In addition, SSS and noise reduction (NR) conventional methods attempt to process voice signals which have been affected by noise, as linear signals. However, these SSS methods have some constraints because of the nonlinearity of a voice signal.

Recently, deep learning (DL) approaches have led to major breakthroughs in different signal processing fields and they showed that can surpass the human performance. DL methods have performed a robust classification in image tasks and it performance is not affected by the noise. Furthermore, speech tasks using DL methods can be implemented because of DL's nonlinear characteristics.

This study proposes the use of DL:

a) To replace a MUSIC method in SSL tasks and perform a robust SSL and,

b) To enhance conventional methods used in SSS tasks by reducing the noise and enhancing a previous separated sound.

2. PROPOSED MODEL

The observation model of a captured audio stream using distant microphones is expressed as:

$$x(\omega) = \mathbf{D}(\omega).\mathrm{s}(\omega) + \mathrm{n}(\omega). \quad \textbf{...}\ (1)$$

Here, $x(\omega)$ is an observed microphone signal vector with M observations at frequency ω, $\mathbf{D}(\omega)$ is a transfer function matrix between the array of microphones and a sound source, $s(\omega)$ is a clean speech signal vector with N sources, and $n(\omega)$ is a noise vector with diffuse and dynamically changing colored noise. The observation model (1) is employed as input of a DL method which performs a SSL, SSS and NR tasks.

2.1. SSL using Deep Residual Networks

This study employs a Deep Residual Networks (DRN) to replace a MUSIC method in SSL tasks. First, (1) is extended fit the observed microphone signal respect with its location:

$$x(\omega) = \mathbf{D}(\omega, \phi, \theta_1).\ s(\omega, \theta_1) + \mathbf{n}(\omega, \phi) \text{.... } (2)$$

Here, the transfer function $\mathbf{D}$ is denoted in term of the direction of sound arrival (DOA) θ and the referential orientation of the robot's head ϕ.

The calculation of each DOA (θ_1) (i.e. SSL) for sound sources using DL models can be formulated as:

$$DOA(\theta_l) = argmax_{0degree}^{359degree}\ p(y = \theta_l | x(\omega)) \text{.... } (3)$$

DL models, which attempt to model high-level abstractions, have perform well in several signal processing tasks, such as computing image; and in image classification tasks, studies claimed that deep convolutional neural networks (DCNN), which is a DL model, can surpass human performance. To perform challenging tasks with DCNN, recently research has revealed that the network depth is critical. However, the training accuracy of deeper network saturates depending on the depth, and then rapidly degrades. This degradation problem is addressed by using deep residual networks (DRN) [3]. DRN have shown the best have shown the best performance on the ImageNet Large Scale Visual Recognition Challenge 2015. The residual learning of DRN is denoted as:

$$F(x) := H(x) - x. \qquad \dots (4)$$

Where, $H(x)$ is a desired underlying map of an input x. Residual learning (RL) allows the layers to fit a residual mapping $F(x)$ instead of hoping that each few stacked layers directly fit the desired underlying map. The original mapping is then reformulated as:

$$H(x) = F(x) + x. \qquad \dots (5)$$

This formulation can be implemented by a feedforward neural network using shortcut connections (Fig. 1) and is called *residual block*.

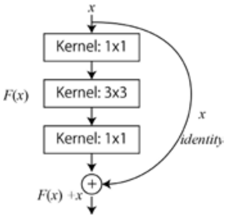

Fig. 1. Residual block - SSL proposed model

2.2. SSS and NR using Denoising Autoencoders

For SSS and NR tasks, this study propose the use of deep denoising autoencoder (DDA) to enhance the performance of a separated sound obtained by conventional methods. Deep neural networks (DNN) is an artificial neural network with multiple hidden layers of units between the input and the output layers that can model complex nonlinear representations. A principal characteristic of this DL model is the ability to self-organize sensory features from a large amount of training data, in which the recognition performance exceeds the performance of conventional models for a speech recognition. Moreover, DNNs demonstrate high performance as a model for NR tasks.

A DDA (Fig. 2) is an implementation of a DNN that attempts to learn the representation or principal features of data. Typically, this representation has a minor dimension to the input. In real-world applications, data can be corrupted or partially destroyed, which makes representation difficult. A DDA was proposed as a solution [4] to overcome the problems of partial destruction and noisy information.

3. IMPLEMENTATION

In SSL, a DCNN with residual blocks (ResNet1 – ResNet2) performs the tasks. This model is trained with supervised learning, using as input the power features from the Fourier transform of a multiple channel audio stream with noise, and the angle location of the sound source as the target. The training process is performed using ADAM optimization and the a SoftMax cost (i.e. loss) function is employed to reduce the error of the model. A plain network with the same number of layers as ResNet1 and trained at a same number of iterations is employed to compare and evaluate the residual learning.

In SSS and NR, a DDA with supervised learning is used perform the tasks. The mel-filter bank (MFB) features is used for modeling the separation filter. Depending on the task, a multiple channel noisy audio stream is used as input of the DDA to perform SSS and NR, or an audio stream obtained from a previous SSS and NR implementation is employed for a NR task.

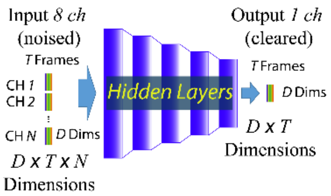

Fig. 2. DDA – SSS & NR proposed model

For training and evaluation of the tasks, the Acoustic Society of Japan-Japanese Newspaper Article Sentence (ASJ-JNAS) Corpora is employed to prepare the audio stream inputs. The impulse response of a room obtained from a HEARBO robot is used to simulate the reverberation and to synthesize the input.

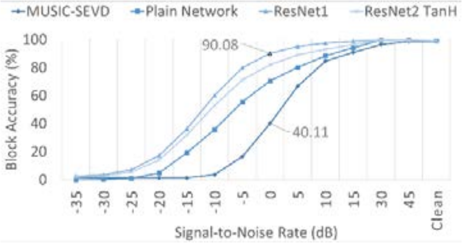

Fig. 3. SSL Results

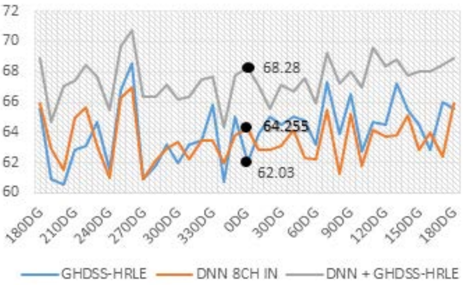

Fig. 4. SSS Results

4. RESULTS

Fig 3. shows that a DL model trained with RL (ResNet1 and ResNet2) can replace a MUSIC method (SEVD-MUSIC) and perform SSL with a robust accuracy, double of the accuracy at 0 dB SNR. Furthermore, the accuracy is better compared with a plain network, showing the effectiveness of RL.

Fig. 4. shows that DDA perform a SSS and SSL task with a similar accuracy compared to conventional method (GHDSS-HRLE). However, it also shows that the combination of a DDA with conventional methods perform better than used each one.

5. CONCLUSION

This study used DL models to improve the accuracy in two stages and adapt robot audition to dynamic environments, SSL and SSS. SSL implemented with DL model and using only the power performed a robust accuracy. DL model can perform well SSS and NR, but using it with a conventional method perform better.

References:

[1] K. Nakadai, T. Lourens, G. O. Hiroshi, and H. Kitano, “Active audition for humanoid,” *Proc. Natl. Conf. Artif. Intell.*, pp. 832–839, 2000.

[2] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, “Robot audition for dynamic environments,” *2012 IEEE Int. Conf. Signal Process. Commun. Comput. ICSPCC 2012*, pp. 125–130, 2012.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015.

[4] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” *Proc. 25th Int. Conf. Mach. Learn.*, pp. 1096–1103, 2008.