International Journal of Internet, Broadcasting and Communication Vol.11 No.3 106-110 (2019)
http://dx.doi.org/10.7236/IJIBC.2019.11.3.106

IJIBC 19-3-14

Deep Learning-Based Sound Localization Using Stereo Signals Based on Synchronized ILD

Hyeon Tae Hwang*, Deokgyu Yun**, Seung Ho Choi***

*, ***Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea
*gusxo3975@naver.com, ***shchoi@snut.ac.kr
**Dept. of Electronic Engineering, Seoul National University of Science and Technology, Seoul, Korea
**deokkyuyun@gmail.com

Abstract

The interaural level difference (ILD) used for the sound localization using stereo signals is to find the difference in energy that the sound source reaches both ears. The conventional ILD does not consider the time difference of the stereo signals, which is a factor of lowering the accuracy. In this paper, we propose a synchronized ILD that obtains the ILD after synchronizing these time differences. This method uses the cross-correlation function (CCF) to calculate the time difference to reach both ears and use it to obtain synchronized ILD. In order to prove the performance of the proposed method, we conducted two sound localization experiments. In each experiment, the synchronized ILD and CCF or only the synchronized ILD were given as inputs of the deep neural networks (DNN) [1-2], respectively. In this paper, we evaluate the performance of sound localization with mean error and accuracy of sound localization. Experimental results show that the proposed method has better performance than the conventional methods.

Keywords: *Synchronized ILD, Sound localization, Stereo signals, Deep neural networks*

1. Introduction

Sound localization is a technique for estimating the position of a sound source and is used for the hearing function of a robot, a virtual reality, and an artificial intelligence CCTV. Conventionally, the interaural time difference (ITD) and the interaural level difference (ILD) are used to estimate sound localization [3-5]. Since the method of obtaining the conventional ILD simply calculates the level difference with respect to the signal reaching both ears, an error due to the inaccurate level difference caused by the time difference may occurs as shown in Figure 1.

Manuscript Received: July. 24, 2019 / Revised: July. 29, 2019 / Accepted: August. 5, 2019
Corresponding Author: shchoi@snut.ac.kr
Tel: +82-02-970-6461
Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea

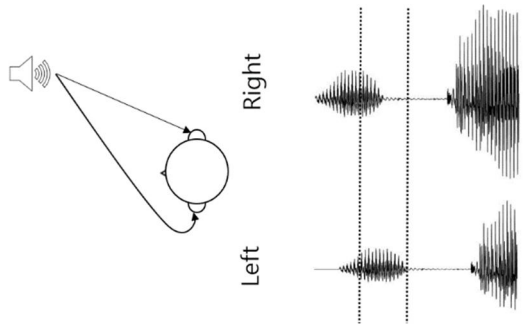

Figure 1. An example of error in ILD due to time difference.

If the ILD is obtained in the situation where the sound source is located on the right side of the person as shown in Figure 1, the energy of the signal reaching the left ear is larger even though the actual sound source is located on the right side. Therefore, the position of the sound source is estimated to the left. To resolve this error, we propose a sound localization technique based on synchronized ILD.

2. Synchronized ILD

In order to obtain the synchronized ILD, the time difference can be obtained through the cross-correlation function (CCF). Let $s_l(n)$ and $s_r(n)$ be the signals arriving both ears, and let *N* be the length of the signal. Also, let time difference τ_d is obtained from Equation (1).

$$\tau_d = \underset{\tau}{argmax} \sum_{\tau=-(N-1)}^{N-1} s_l(n) s_r(n+\tau) \tag{1}$$

If the stereo signals are synchronized as the estimated time difference τ_d, it becomes as shown in Figure 2. Looking at the displayed frame shown in Figure 2, the energy of the signal reaching the right ear is greater than that of the left ear. Since the actual position is on the right side, it is necessary to obtain the ILD after synchronization processing in order to correctly estimate the position of the sound source. After synchronizing the signals in this way, the ILD is obtained through Equation (3) by expressing the difference in energy of both ears shown in Equation (2) on a dB scale.

$$E_{s_l} = \sum_{n=0}^{N} (s_l(n))^2 , E_{s_r} = \sum_{n=0}^{N} (s_r(n))^2 \tag{2}$$

$$ILD = 10log \frac{E_{s_l}}{E_{s_r}} \tag{3}$$

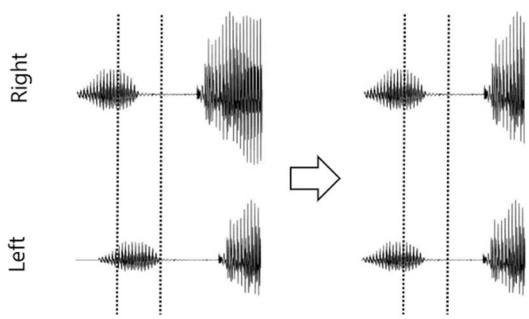

Figure 2. An example of stereo signals after synchronization.

3. Experiment and result

In this research work, the sound localization experiments are based on a deep neural networks (DNN) in order to evaluate the performance of synchronized ILD as shown in Figure 3. The database used in the experiment is the VCTK speech database [6], and the experiment was performed by dividing the sound source for training and test. We also used the PKU-IOA-HRTF database as a head-related impulse function [7]. Speech signals are sampled at 16 kHz, and the feature vector consists of 33 CCFs and one synchronized ILD [8-10]. Two feature vectors corresponding to stereo signals are used as DNN inputs and the output includes 72 nodes divided the angle of 360 degrees by 5 degrees on horizontal azimuth.

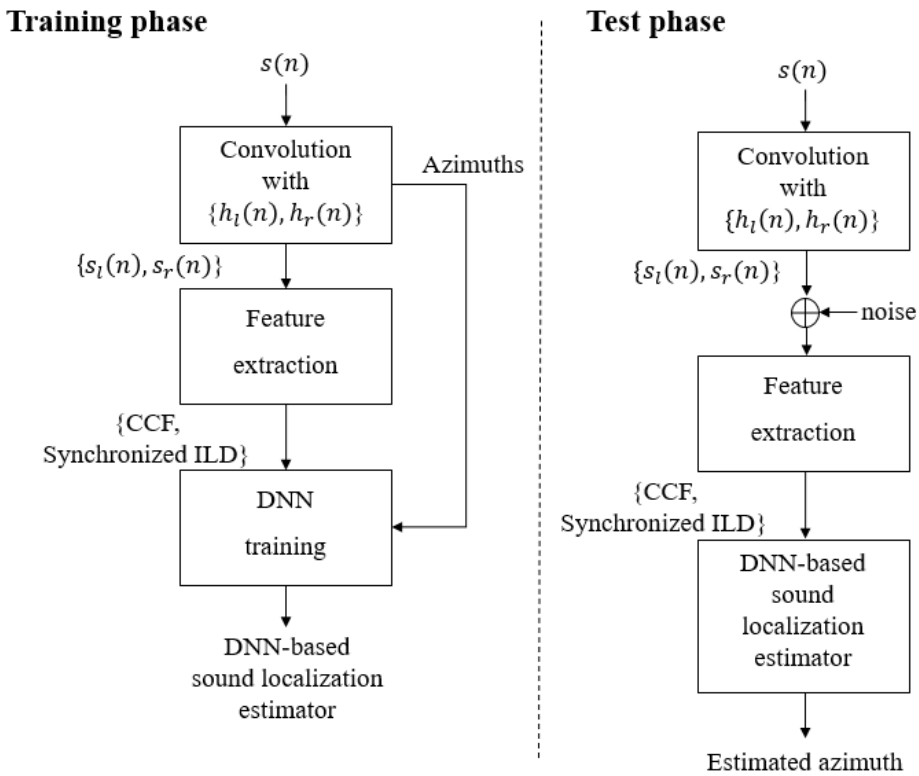

Figure 3. The block diagram of the proposed method.

The performance of the proposed synchronized ILD is compared with the conventional ILD as in Table 1. Table 2 shows the result using one ILD and the CCFs as the input of DNN. As in Tables, the proposed method is superior to the conventional method.

Table 1. Result of sound localization experiment using ILD.

Method	Accuracy [%]		Mean error [degree]
	Frame	Sentence	
Conventional ILD	20.27	63.05	11.08
Synchronized ILD	20.43	64.17	8.35

Table 2. Result of sound localization experiment using CCFs and ILD.

Method	Accuracy [%]		Mean error [degree]
	Frame	Sentence	
Conventional ILD	56.35	97.50	1.29
Synchronized ILD	62.61	98.61	0.35

4. Conclusion

In this paper, we proposed the synchronized ILD based sound localization method, which removes the time difference that reaches both ears in order to mitigate inaccurate localization of the conventional ILD that simply calculates the level difference with respect to the signal reaching both ears. In order to prove that the proposed method is superior to the existing method, we experimented by using synchronized ILD and CCF or only synchronized ILD as input for DNN-based sound localization. Through these experiments, we confirmed that the proposed method outperforms the conventional method. In the future, we will confirm whether the proposed method produces the results we expect even in the noisy environments. Therefore, we will experiment with DNN based sound localization using speech signals with several kinds of noise. Furthermore, we will find other feature vectors that can replace CCF and confirm if the performance improves when used as input to the DNN with synchronized ILD.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2016-0-00144, Moving Free-viewpoint 360VR Immersive Media System Design and Component Technologies).

References

[1] Geoffrey Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine* 29, Nov. 2012.
DOI: https://doi.org/10.1109/MSP.2012.2205597

[2] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
DOI: https://arxiv.org/abs/1303.5778

[3] Amir Avni and Boaz Rafaely, "Sound localization in a sound field represented by spherical harmonics," *International Symposium on Ambisonics and Spherical Acoustics*, May 2010.

[4] William M. Hartmann and Zachary A. Constan, "Interaural level differences and the level-meter model," *The Journal of the Acoustical Society of America* 112.3: 1037-1045, 2002.
DOI: https://doi.org/10.1121/1.1500759.

[5] S. T. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," *International Conference on Acoustics, Speech, and Signal Processing,* 2005.
DOI: https://doi.org/10.1109/ICASSP.2005.1416207.

[6] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.
DOI: https://doi.org/10.7488/ds/1994.

[7] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Trans. on Audio, Speech and Language Processin*g, vol. 17, no. 6, pp. 1124-1132, 2009.
DOI: https://doi.org/ 10.1109/TASL.2009.2020532.

[8] Ning Ma, Tobias May, and Guy J. Brown, "Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 25.12: 2444-2453, 2017.
DOI: https://doi.org/ 10.1109/TASLP.2017.2750760.

[9] Frederic L. Wightman and Doris J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*105.5: 2841-2853, 1999.
DOI: https://doi.org/10.1121/1.426899.

[10] Tobias May, Steven van de Par, and Armin Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 19, no. 1, pp. 1–13, Jan. 2011.
DOI: https://doi.org/10.1109/TASL.2010.2042128.