PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Neural Network-based Broadband Beamformer with Less Distortion

Mitsunori MIZUMACHI[1]

[1] Kyushu Institute of Technology, Japan

ABSTRACT

Beamforming has been one of the important issues in the field of multi-channel signal processing including acoustic signal processing. A wide variety of beamformers have been proposed for each application. In general, acoustic beamforming deals with broadband signals such as speech signals compared to narrowband beamforming for antenna array and radar applications. Recently, neural network-based non-linear beamformers become popular but have a problem that causes an annoying non-linear distortion on the output signal. In the case of speech enhancement, it is a serious problem because our auditory system is highly sensitive to artificial non-linear distortion on speech signals. This paper proposes to solve the problem with the relaxed dual cost functions in the neural network-based beamformer for speech enhancement. The primary cost function aims at sharpening the beam-pattern, and the second cost function is introduced to achieve decreasing speech distortion. Those cost functions are alternatively used for optimizing the beam-pattern in the frequency range of speech signals. The feasibility of the proposed method is confirmed by carrying out a listening test.

Keywords: Beamforming, neural network, non-linear distortion, cost functions, spectral distortion

1. INTRODUCTION

Beamforming is a representative means for noise reduction and signal enhancement [1]. It enables to detect and enhance target signals in adverse conditions. A wide variety of beamformers have proposed for several decades. The traditional beamformers have been designed analytically and adaptively [1]. A neural network can be an alternative approach to optimizing the beamformers. Kobatake et al. proposed a pioneering super-directive beamformer with the three-layered neural network (NN) structure [2]. The NN-based beamformers became popular for narrow-band antenna applications [3-5]. It is, however, difficult for those beamformers to deal with wide-band acoustical signals, although various non-linear beamformers with learning schemes based on NNs have been investigated for acoustical applications [6-9].

Non-linear beamformers could achieve sharp beampatterns using the proposer training data, but cause annoying distortion on the output target signals. It is difficult to decrease the distortion on the target signal due to non-linear activation functions in each layer of the NN. The author proposed a deep neural network(DNN)-based beamformer with relaxed cost functions, which aim at decreasing the non-linear distortion [10]. The sub-band neural networks are individually trained using band-limited training data when the microphone arrangement is optimized for the frequency range of speech [11]. The discontinuity in the frequency domain causes another type of non-linear distortion. Bonafonte *et al.* proposed an end-to-end speech enhancement scheme with a generative adversarial network [12]. It succeeds in neural network-based speech enhancement with less distortion. However, it requires a huge amount of training data due to its higher degree of freedom. Koizumi *et al.* proposed a DNN-based source enhancement scheme, where the DNN is trained to increase objective sound quality assessment scores such as the perceptual evaluation of speech quality (PESQ).

In this paper, a NN/DNN-based beamformer with the relaxed dual cost functions is proposed for speech enhancement. The proposed beamformer employs the different cost functions based on the directivity and spectral distortion in the spatial and spectral domains, respectively. The primary cost

[1] mizumach@ecs.kyutech.ac.jp

function based on the spatial features used in end-to-end NN-based beamformer achieves sharpening the beam-pattern, and the second cost function in the spectral domain is introduced to achieve decreasing speech distortion. Those cost functions are alternatively used for optimizing the beam-pattern in the frequency range of speech signals. The feasibility of the proposed method is confirmed by computer simulation with a small amount of training data, which include sinusoidal signals, random noises, and speech signals. The feasibility of the proposed beamformer is evaluated by carrying out a listening test.

2. NON-LINEAR NEURAL NETWORK-BASED BEAMFORMER

Beamforming can be achieved by linear signal processing, where multiple observed signals are phase-adjusted and summed up such as the delay and sum beamformer [1]. The target signal coming from the desired direction is not distorted by the linear beamforming. On the other hands, interference signals coming from the undesired directions are weakened by phase interference. However, the delay-and-sum beamformer is not superior in controlling the directivity pattern compared with the state-of-the-art beamformers. The delay-and-sum beamformer needs a number of microphones to form the sharp main lobe, especially in the low-frequency range and does not turn attention to the directions except the look direction.

Kobatake *et al*. proposed a novel framework of non-linear beamforming, where a three-layered NN was employed to achieve superdirectivity [2]. The NN is trained as an end-to-end autoencoder. The NN is allowed to output the input signal as it is, only when the signal comes from the target direction. When signals come from non-target directions, the NN is trained not to output any signal in the training phase. It can achieve superdirectivity for the narrowband signal such as sinusoidal signals. However, the non-linear activation functions used in the NN-based beamformers cause the non-linear distortion on the target signal. Non-linear distortion should be reduced for wide-band acoustic applications.

3. NEURAL NETWORK-BASED BEAMFORMER WITH LESS DISTORTION

The conventional NN-based beamformers merely aimed at forming the sharp main lobes using the directivity-based cost functions. NN is a flexible framework for system optimization. Then, another cost function is additionally introduced based on spectral distortion, which can optimize the beam-pattern in minimizing the spectral distortion of the target signal. Those cost functions based on directivity and spectral distortion enable to reduce interfering signals and decrease the spectral distortion of the target signal, respectively. The dual cost functions are alternatively used in the training process.

The proposed beamformer is trained using a sinusoidal signal of 1.7 kHz and speech database, which consists of 27 English speakers [14], for the cost functions based on directivity and spectral distortion, respectively. In this paper, the target signals are speech signals so that the frequency range is limited from 300 Hz to 3.4 kHz.

For speech data set uttered by each speaker, the proposed beamformer is trained with the cost function based on spectral distortion after the pre-training is completed with the directivity-based cost function. In the pre-training phase, the sinusoidal signal is given to the neural network-based beamformer sample by sample with the sampling frequency of 44.1 kHz with 16 bits in the accuracy. In the later training phase, the amplitude spectra obtained by the short-term Fourier transform with 256 samples with the sift of 20 samples are prepared as the training data. The resultant parameter sets are used as the initial values for the training with another speech dataset by the different speaker.

4. PERFORMANCE EVALUATION

4.1 Experimental Conditions

The experimental conditions are summarized in Table 1. The target male speech signal came from the front (0 degrees), and either pink noise or urban noise, which was recorded in a pinball parlor [15], come from 45 degrees. Both the target and interference signals were band-limited in 300 Hz to 8 kHz. Noisy observation data were prepared by the phase adjustment in a computer at -10 dB in the target-to-interference ratio. Noisy signals were obtained using an 8-ch linear microphone array, of which neighboring microphone spacing was 10 cm, in the far field condition.

Table 1 – Experimental conditions

Target signal	Male utterance (300 Hz to 8 kHz)
Interference signal	Pink noise and urban noise (300 Hz to 8 kHz)
Arrival direction of target signal	0 degrees
Arrival direction of interference signal	45 degrees
Target-to-interference ratio	-10 dB
Signal accuracy	44100 Hz / 16 bits
Number of microphone	8
Microphone spacing	10 cm

In this paper, three kinds of the proposed methods were prepared as the three-layered NN-based beamformer with spatial and spectral cost functions in order, the same beamformer with the dual cost functions in the inversed order, and the five-layered DNN-based beamformer with dual cost function in the normal order. In addition, the conventional delay-and-sum beamformer, five-layered DNN-based end-to-end beamformer [9], and the distortion-less DNN-based beamformer [10] were prepared as references.

4.2 Procedure

In the listening test, Thurstone's paired comparison was carried out to subjectively confirm the feasibility of the proposed method. 20 undergraduates and graduate students with normal hearing participated in the listening test. They were required to choose the better speech sample in between two noise-reduced speech samples in terms of the noisiness of the interference noise, clearness of the target speech, and easiness in speech perception.

The noise-reduced samples were presented through the headphone amplifier with USB-DAC (Marantz HD-DAC1) and headphone (Sennheiser HD650). Each speech sample was played at 81 dB on an average in A-weighted sound pressure level. The sound pressure level was measured using the ear simulator (B&K Type 4153) and the handheld analyzer (Aco Type 6240).

4.3 Results

The results of the listening test are given in Figs. 1 and 2 for the pink noise and urban noise, respectively. In each panel, six speech samples are plotted including three proposed methods: Proposed-BF (3-layered), Proposed-BF (5-layered), and Proposed-BF (inversed CF).

The statistical difference between the two speech samples was confirmed by the binomial test at a significant level of 5 %. The pair of speech samples with the significant differences are overplotted in each panel.

4.4 Discussion

In Fig. 1 in the pink noise condition, the proposed methods are evaluated with higher scores compared with the five-layered DNN-based beamformer. However, the distortion-less DNN-based beamformer [10] obtains the best score in terms of the clearness of the target speech. The delay-and-sum beamformer is also evaluated better than the proposed methods. The proposed methods need to be improved to decrease the distortion of the target signal under wide-band noise conditions.

In Fig. 2 in the urban noise condition, the three proposed methods are prior to the other method. Among the proposed methods, the three-layered NN beamformer is better than the five-layered DNN beamformer There are significant differences against five-layered DNN-based end-to-end beamformer [9] or the distortion-less DNN-based beamformer [10]. The urban noise has the dominant power in the lower frequency range below 1 kHz. The resultant noise component of the NN/DNN beamformer output might be distorted, but does not affect the speech perception.

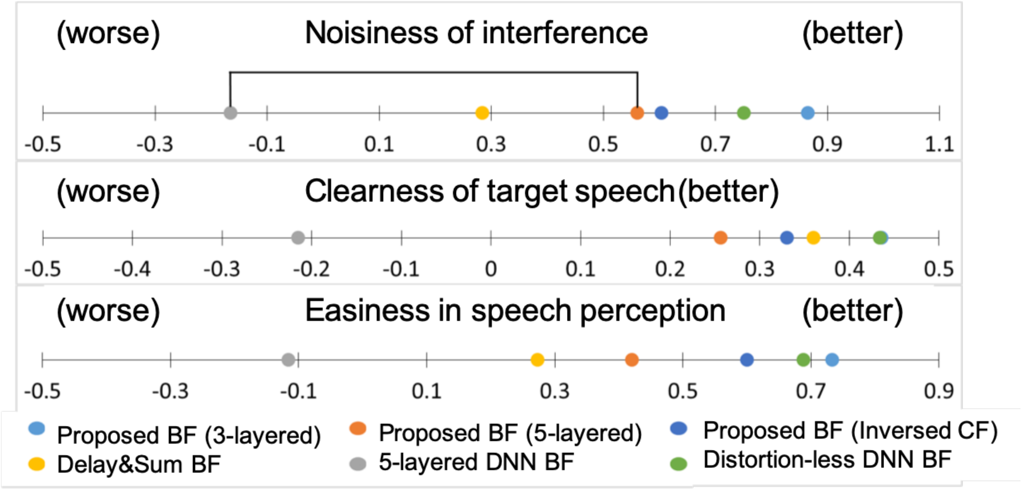

Figure 1 – Results of listening test in pink noise condition.

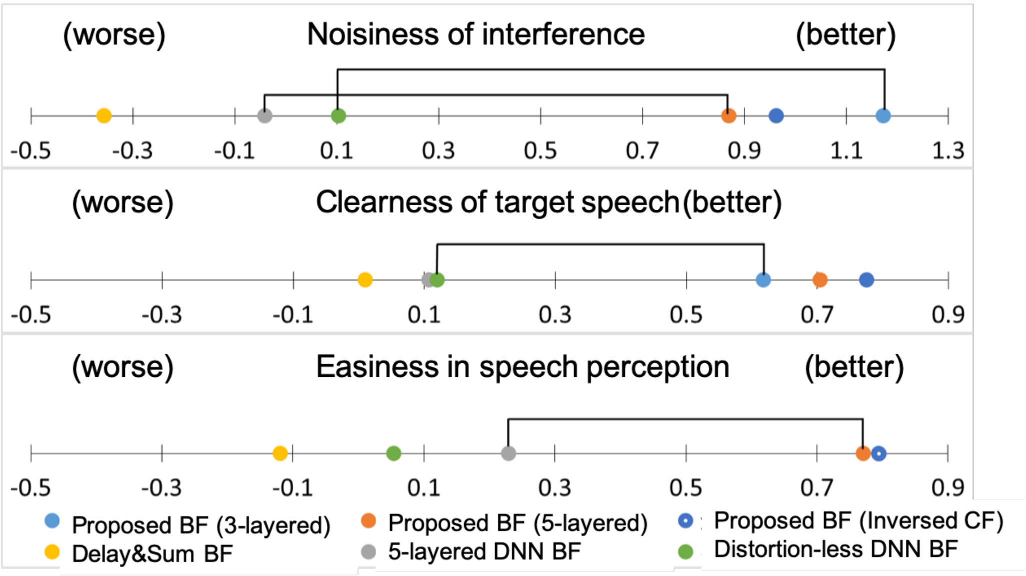

Figure 2 – Results of listening test in urban noise condition.

5. CONCLUSIONS

In this paper, a neural network-based beamformer with dual cost functions is proposed aiming at improving the nonlinear distortion of the beamformer output. The cost functions consist of the spatial and spectral cost functions for sharpening the main lobe and decreasing the spectral distortion within the speech band. The cost functions are alternatively used in the training process with the network structure of three and five layers. The feasibilities of the proposed methods are confirmed by carrying out a listening test. The proposed three-layered neural network-based beamformer is superior to conventional beamformers. Future works include the further investigation of decreasing the distortion on the target signal under wide-band noise conditions.

ACKNOWLEDGEMENTS

This work was partially supported by KAKENHI 16H01734.

REFERENCES

1. Brandstein M, W. Darren W. (eds.). *Microphone arrays: signal processing techniques and application.* Springer; 2013.
2. Kobatake H, Morita W, Yano Y. Super directive sensor array with neural network structure. Proc. ICASSP; 1992; vol. 2, pp. 321-324.
3. Chang P-R, Yang W-H, Chan K-K. A neural network approach to MVDR beamforming problem. IEEE Trans. Ant. and Propag. 1992; 40(3): 313-322.
4. Southall H.L, Simmers J.A, O'Donnell T. H. Direction finding in phased arrays with a neural network beamformer. IEEE Trans. Ant. and Propag. 1995; 43(12): 1369-1374.
5. Zooghby A.H.E, Christodoulou C.G, Georgiopoulos M. Neural network-based adaptive beamforming for one- and two-dimensional antenna arrays. IEEE Trans. Ant. and Propag. 1998; 46(12): 1891-1893.
6. Dahl M, Claesson I, A neural network trained microphone array system for noise reduction. Proc. IEEE Signal Processing Society Workshop; 1996; pp. 311-319.
7. Song X, Wang J, Han Y, Tian D, Neural Network-Based Robust Adaptive Beamforming. Proc. International Joint Conference on Neural Network Proceedings; 2006; pp. 1758-1763.
8. Iseki A, Ozawa K, Kinoshita Y, Neural network-based microphone array learning of temporal-spatial patterns of input signals. Proc. IEEE Global Conference on Consumer Electronics; 2014; pp. 88-89.
9. Mizumachi M, Origuchi M, Advanced delay-and-sum beamformer with deep neural network. Proc. ICA2016; 2016; Paper ID: 696.
10. Nishijima Y, Origuchi M, Mizumachi M, Sub-band optimization of neural network-based broadband beamformer. Proc. Youngnam-Kyushu Joint Conference on Acoustics 2017 (YKJCA2017); 2017: Paper ID: J09.
11. Hayashi H, Mizumachi, M, Speech enhancement by non-linear beamforming tolerant to misalignment of target source direction. Journal of the Institute of Industrial Applications Engineers 2013; 1(2), 97-104.
12. Bonafonte S, A. Bonafonte, Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. Proc. Interspeech 2017; 2017; pp. 3642-3646.
13. Koizumi Y, Niwa K, Hioka Y, Kobayashi K, Haneda Y. DNN-based Source Enhancement to Increase Objective Sound Quality Assessment. IEEE/ACM Trans. ASLP. 2018; 26(10), 1780–1792.
14. Valentini-Botinhao C, Wang X, Takaki S, Yamagishi J. nvestigating rnn-based speech enhancement methods for noise robust text-to-speech. Proc. ISCA Speech Synthesis Workshop; 2016; pp. 146–152.
15. Kawai K, Fujimoto K, Iwase T, Yasuoka H, Sakuma T, Hidaka Y, Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004 (Sound Material in Living Environment 2004). Proc. ICA, 2004; 1561-1564.