DOI: 10.61782/fa.2023.0588

forumacusticum 2023

TARGETED BEAMFORMING ACTIVE NOISE CONTROL BASED ON DISTURBANCE METRICS

Panagiotis Zachos[1*] **George Moiragias**[1] **John Mourjopoulos**[1]
[1] Audio and Acoustic Technology Group, Wire Communications Laboratory,
Department of Electrical Computer Engineering, University of Patras, Greece

ABSTRACT

This work proposes a novel headphone Active Noise Control scheme based on the targeted attenuation of sources that are deemed disturbing for listeners. Initially, a novel listening test determines the disturbance of distinct noises that coexist within background noise based on their class. The listening test is based on such predetermined spatial scenes, binaurally auralized and presented via generic headphones. The disturbance metric derived from such test, guides the operation of the proposed headphone ANC: any complex auditory scene is subsequently analyzed and via a Sound Event and Localization model, a beamformer is steered to the source deemed to be the most disturbing. A Time-Domain Beamformer, driven by a phased array is formed by the two already existing reference microphones commonly found in the outer shell of ANC-Enabled headphones and guides the multi-reference ANC controller in order to provide an improved attenuation of the primary disturbing source, while also significantly attenuating the background noise field to acceptable levels.

Keywords: *Active Noise Control, Headphones, Sound Event Localization and Detection, Listening Test, Beamforming*

1. INTRODUCTION

Even though ANC technology has been around for many decades, it has only recently been applied to headphones, which are the most popular personal listening devices, with such market expected to further increase in the coming years. The most popular ANC applications utilize the Filtered-x Least Mean Squares (FxLMS) algorithm and its extensions [1], boasting high performance, a simple structure, a robust operation and low computational complexity [2–4]. A common approach between such algorithms is that they largely ignore the spatial sensitivity of such systems [5] by attempting to minimize the total external noise.

To more significantly and robustly attenuate the most disturbing source a novel Targeted through Beamforming ANC (TBANC) approach has been proposed in [6], utilizing a Time Domain Beamformer (TDBF). While also significantly attenuating the background noise field to acceptable levels, this method is limited in single source scenarios.

An extension to the TBANC approach is proposed in this work, hereby denoted by TBANC-D, where multi-source scenarios are considered in the presence of diffuse noise. The proposed TBANC-D is able to operate in scenes where none, one or multiple sources exist. In the case of multiple sources a single source that is deemed exceptionally disturbing, guided by a novel Disturbance metric that determines the severity of distinct noises in a complex sound field, is more strongly attenuated, while the remaining sources are attenuated as part of the remaining background noise.

[*] ***Corresponding author****: p_zachos@upnet.gr.*

Funding: This research has been co-financed by the ERDF of EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the special actions “Aquaculture”, “Industrial Materials”, and “Open Innovation in Culture” (project code: T6YP-00490, MIS 5052072).

2. METHODS

The ANC proposed in this work is based upon the assumption that when a listener resides in a complex sound field, a specific source can exist that is dominant and disruptive thus being associated with a higher Disturbance metric, hereby denoted as the primary noise source.

In practice the presence of such a source can be detected by a simple differential energy detector, i.e. when the level of the noise out of one microphone exceeds a predetermined threshold as described in [6].

In the proposed approach, the ANC controller can be mainly focused on the attenuation of the primary noise source without completely disregarding the remaining sound field. To achieve the attenuation of the complete sound field with special focus on the primary noise source, the Multi-Reference ANC (MRFANC) [7] is chosen to be extended with the use of a beamformer, in this work realized in the form of a straightforward TDBF.

Additionally, the multi-reference strategy practically allows the summation of two anti-noise signals. One predominantly based on the primary noise with the goal of further attenuating such an undesired stimulus and another effectively based on the diffuse sound-field.

The proposed method consists of a Sound Event Localization and Detection (SELD) machine learning model, TDBF and a MRFANC controller [7]. TBANC-D is driven by a Disturbance metric, which is derived from a subjective listening test, and is used to guide the operation of the proposed ANC scheme. An overview of the proposed system is shown in Fig. 1, with the signals captured by the left and right microphones being used to extract the features which are then fed into the SELD model. The SELD model consequently outputs the Sound Event Detection (SED) estimates, showing which source classes are active in the current scene, as well as the Direction of Arrival Estimates (DOAE), showing the direction of arrival of each source. Subsequently, the results of the listening test drive the disturbance system given the SED and DOA estimates, which finally outputs the DOA of the primary disturbing source $\hat{p}$.

2.1 Problem formulation

This work, assumes that the headphone user exists within a complex sound-field consisting of diffuse, spatially white noise, along with a primary noise source that emits a particularly disturbing to the listener signal $p(n)$ and a less disturbing secondary noise source $s(n)$, as shown in Fig. 2. In such a scenario, the right ear which is closest to the primary noise source is assumed to be the *Good-Ear* in the sense that the energy of $p(n)$ and $s(n)$ is much more prevalent in the signal recorded by the respective reference microphone.

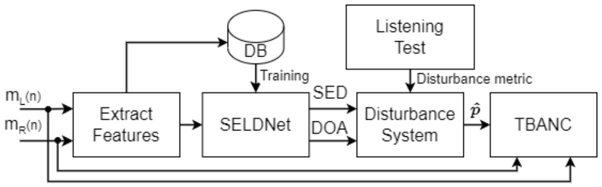

Figure 1. Block diagram of the proposed TBANC-D system. $\hat{p}$ denotes the DOA of the primary disturbing source.

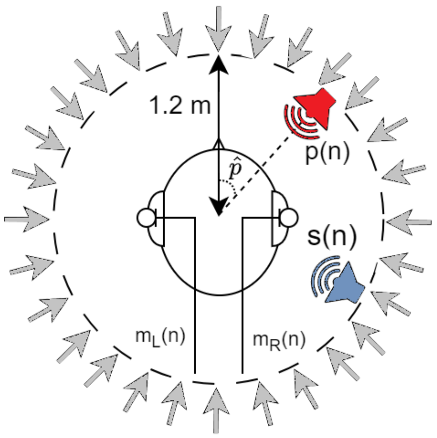

Figure 2. Schematic of the ANC setup formulation. The primary noise source $p(n)$ is denoted by a red speaker with its respective DOA given by $\hat{p}$, while the secondary source $s(n)$ is denoted by a blue speaker. Broadband white noise sources, denoted by the gray arrows, are placed with a 5^o spacing in order to simulate a diffuse noise field.

The signals captured by the two reference microphones $m_L(n)$ and $m_R(n)$, especially the signal that corresponds to the Bad-Ear, are heavily affected by the head shadowing effect. Since the acoustic shadow is dependent upon the angle of incidence of the source [8] it can be calculated as the Relative Transfer Function between the 2

Head Related Transfer Function (HRTF) channels as follows

$$HS(\omega,\theta) = \frac{\text{HRTF}_{BE}(\omega,\theta)}{\text{HRTF}_{GE}(\omega,\theta)} \tag{1}$$

where θ indicates the sound angle of incidence, HRTF_{GE} and HRTF_{BE} indicate the transfer functions of the Good-Ear and Bad-Ear respectively.

The noise captured by the reference microphone at the side of the Good-Ear due to a noise source emitting a signal $\xi(n)$ (primary or diffuse) is assumed to arrive as-is with a simple delay, while the noise captured by the microphone at the side of the Bad-Ear is affected by the respective Head-Shadow filter calculated by Eq. (1). The resulting signals are then given by

$$\begin{aligned} m_{GE}(n) &= \xi(n-N) \\ m_{BE}(n) &= m_{GE}(n) * HS(\omega,\theta) \end{aligned} \tag{2}$$

where m_{GE} and m_{BE} denote the signals captured by the Good-Ear and Bad-Ear respectively, $\xi(n-N)$ denotes a noise source placed at an angle θ in the horizontal plane arriving with a delay N due to its distance from the Good-Ear. Depending on the angle θ of the noise source, the role of the Good-Ear is assumed by the right-ear for $-180^o \leq \theta \leq 0^o$ or by the left-ear for $180^o < \theta < 0^o$ with 0^o being directly in front of the listener and θ increasing in a counter-clockwise fashion.

2.2 Disturbance System

The aim of the Disturbance System is to determine which of the active sound events corresponds to the primary noise source, based on the respective Disturbance metrics evaluated by human assessors through a listening test, and then provide TBANC with the angle of the primary noise source $\hat{p}$. The listening test procedure consisted of 24 assessors, all self-reported as normal hearing listeners. Three noise excerpts from each monophonic recording described in Sec. 2.4 with a duration of 10 seconds were selected for the experiment, that were loudness normalized at -22 LUFS according to [9]. Each of the 9 noise samples was reproduced via headphones and the assessors were asked to register the level of disturbance, which corresponds to the Disturbance metric of each noise sample via a multi-stimulus procedure, by answering the question "How disturbing is each noise sample?". For the evaluation of disturbance, a 9-point differential scale was used [10], with the upper limit representing high disturbance, while the lower limit representing low disturbance. Therefore, pairwise comparisons of the disturbance ratings of the sound events will reveal the sound event with the higher Disturbance metric and hence, define the primary noise source in different scenarios.

2.3 Neural network features

The Spatial Cue-Augmented Log-Spectrogram for Polyphonic SELD (SALSA) feature [11] has been employed in this work. This feature was chosen due to the high performance achieved by SELD models in the DCASE 2022 challenge utilizing SALSA for microphone recording inputs, as is also the case for this work, where sources have to be detected and localized through the use of signals captured by the left and right reference microphones m_{GE} and m_{BE}. Specifically, a more computationally efficient version of the SALSA feature, called SALSA-Lite [12] was used.

The $2M-1$ channel SALSA-Lite is comprised by the M channel log-spectrogram features from the M reference microphones and the $M-1$ channel frequency-normalized inter-channel phase differences (NIPD) [13], where M is the number of reference microphones. The NIPD is given by

$$\mathbf{\Lambda}(t,f) = -c(2*\pi*f)^{-1}\text{arg}[X_1^*(t,f)\boldsymbol{X}_{2:M}(t,f)] \tag{3}$$

where t and f are the time and frequency indices; $c \approx 343\text{ms}^{-1}$ is the speed of sound and $X_i(t,f)$ is the short-time Fourier transform (STFT) of the i^{th} microphone signal.

SALSA-Lite has been shown to be excellent for resolving overlapping sound events [12], as is the case in this work, due to it having an exact TF alignment between the log-spectrograms and the NIPD channels.

2.4 Dataset

The dataset used to train this work consisted of monophonic recordings from the DEMAND dataset [14] that were spatialized using the method described by Eq. (2). Three different monophonic recordings were selected i.e. 3 sound classes, corresponding to a busy café that consisted of babble and cutlery noise along with strong gusts of wind, a heavy street traffic scene and a relatively quiet subway station environment. These recordings were selected because they were representative of the environments that the proposed system would be used in.

The spatialized recordings consisted of two simultaneous active sources, each placed in a different azimuth angle on the right hemiplane, with a step of 5^o and a minimum spacing of 30^o. The first 30 seconds of each recording were used to generate the spatialized recordings for the training stage, while the following 10 seconds were used for the inference stage, resulting in 36 hours of data for training and 12 hours for inference. Such recordings consist of the signals captured by the left and right reference microphones m_{GE} and m_{BE} and are sampled at 24kHz for the purposes of training the SELD network. Only cases for the right ear were considered, since resolving the hemiplane where the sources reside is straightforward, in using a simple energy based approach as described in [6]. The case where the two sources reside on opposite hemiplanes is a point of future work.

To further enhance the available dataset 2 data augmentation techniques were applied to all the features during training: Random Cutout (RC) [15,16] and Frequency Shifting (FS) [11]. In RC either random cutout or TF masking via SpecAugment [15] was applied on all channels of the input features, by producing a rectangular or a cross-shaped mask on the spectrograms respectively. For FS the input features were randomly shifted up or down by up to 10 bands.

2.5 Network architecture

The CRNN network employed in this work is based on the SELD network employed in [11], modified to suit the requirements of this work is shown in Fig. 3. The SELD network consists of an encoder block [17], the SED branch is formulated as a multiclass multilabel classification, while the DOAE branch is formulated as a one dimensional regression problem.

2.6 Hyperparameters

The STFT window length was set to 512 samples, with a hop size of 300 samples, a Hann window and 512 FFT points. A cutoff frequency of 9kHz was selected to compute the features resulting in 192 bins for the SALSA-Lite features. 8-second audio chunks were used during training, with an overlap of 0.5 seconds while the full audio clip was used during the inference stage. The Adam optimizer [18] was used with a learning rate of 0.0003 and the network was trained for 2 epochs with a batch size of 16.

2.7 Targeted Beamforming ANC

The block diagram for the right-ear controller of TBANC incorporating the proposed beamforming scheme is

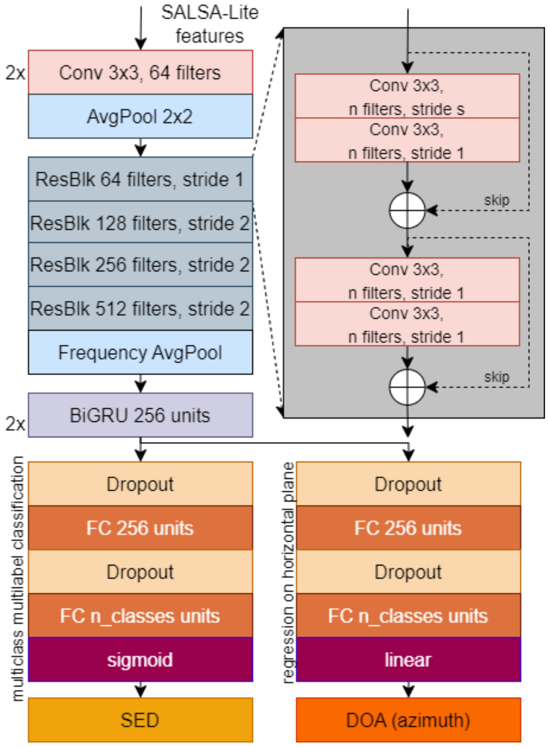
SALSA-Lite features
2x
Conv 3x3, 64 filters
AvgPool 2x2
ResBlk 64 filters, stride 1
ResBlk 128 filters, stride 2
ResBlk 256 filters, stride 2
ResBlk 512 filters, stride 2
Frequency AvgPool
2x
BiGRU 256 units
Conv 3x3, n filters, stride s
Conv 3x3, n filters, stride 1
skip
Conv 3x3, n filters, stride 1
Conv 3x3, n filters, stride 1
skip
multiclass multilabel classification
Dropout
FC 256 units
Dropout
FC n_classes units
sigmoid
SED
regression on horizontal plane
Dropout
FC 256 units
Dropout
FC n_classes units
linear
DOA (azimuth)

Figure 3. Block diagram of the SELD Network. Image adapted from [11].

shown in Fig. 4 . The signal fed to the control loudspeaker can be expressed as the summation of the control signals generated by filtering both the left and right reference signals m_R and m_L respectively. Without loss of generality, the anti-noise signal driving the right loudspeaker is given by

$$\begin{aligned} y'_R(n) =& y_{RR}(n) + y_{LR}(n) \\ =& w_{RR}(n) * BF_o(n) + w_{RL}(n) * m_L(n) \end{aligned} \tag{4}$$

where * denotes the convolution operation, $BF_o(n)$ is the beamformer output, that drives the w_{RR} adaptive filter, w_{RR} and w_{RL} are weights of the adaptive control filters, with w_{RR} corresponding to the filter whose input comes from the right reference microphone and w_{RL} corresponding to the filter whose input comes from the left reference

Demos of the proposed ANC system can be found at http://audiogroup.ece.upatras.gr/tools/TBANCD.php

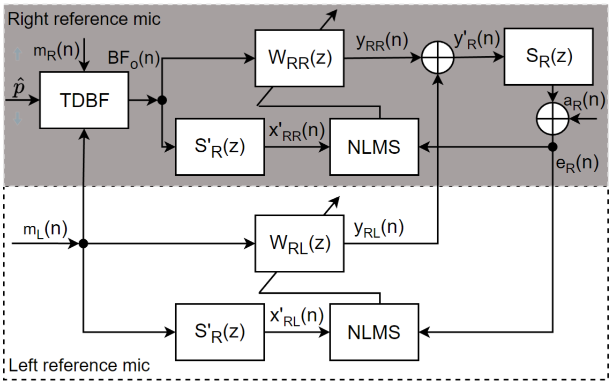

Figure 4. Right ear controller overview of the proposed targeted ANC with beamforming. $\hat{p}$ denotes the DOA of the primary disturbing source. An equivalent control algorithm operates independently for the left ear.

microphone and is used to drive the right ear error sensor. The weight update procedure of the Normalized LMS controller can be expressed by:

$$\begin{aligned} w_{RR}(n+1) &= w_{RR}(n) + \mu\frac{e_R(n)x'_{RR}(n)}{r+||x'_{RR}(n)||^2} \\ w_{RL}(n+1) &= w_{RL}(n) + \mu\frac{e_R(n)x'_{RL}(n)}{r+||x'_{RL}(n)||^2} \end{aligned} \tag{5}$$

where μ is the step size, r is the regularization factor, $e_R(n)$ is the error captured by the error microphone inside the right ear cup, $x'_{RR}(n)$ and $x'_{RL}(n)$ are the noise captured by the right reference microphone, filtered by the estimate of the secondary path, $S'(z)$ and the noise captured by the left reference microphone, filtered by the estimate of the secondary path, $S'(z)$ respectively. The error signal is given by

$$e_R(n) = a_R(n) + y'_R(n) * S_R(n) \tag{6}$$

where $a_R(n)$ is the total ambient noise captured at the left error microphone attenuated by the headphone shell, $y'_R(n)$ is the control signal driving the right headphone driver and $S_R(n)$ is the impulse response corresponding to the right headphone driver.

For the subsequent analysis, a typical example case is examined, and it is assumed that both the primary and secondary noise sources reside to the right of the listener in the horizontal plane, so the right-ear is chosen as the *Good-Ear*. This signifies that due to the head shadowing effect, the energy of the right reference microphone signal $m_R(n)$ contains significantly more energy than the respective left reference microphone signal $m_L(n)$. In order to further emphasize this, the output of the TDBF BF_o, is used to drive the adaptive filter w_{RR} and its respective controller.

2.8 Evaluation metrics

The performance of the SELD model was evaluated using the metrics of Precision, defined as the ratio between the number of correctly classified sound events to the total number of sound events classified as active, Recall, defined as the ratio between the number of correctly classified sound events to the total number of sound events present in the audio clip, F1-score, which is the harmonic mean of Precision and Recall for the SED task, and the Mean Absolute Error for the DOAE task.

High *Precision*, *Recall* and *F1-score* values indicate that the model is able to correctly classify the events, while low *MAE* values indicate that the model is able to accurately predict the DOA of the events.

The performance of the ANC is evaluated using the I_e metric, which is given by

$$I_e = 10log_{10}(\frac{\sum_{n=0}^{N} e^2_{\text{MRFANC}}(n)}{\sum_{n=0}^{N} e^2_{\text{PROP}}(n)}) \tag{7}$$

where I_e denotes the improvement in dB calculated as the ratio between the steady-state errors of the two methods, namely the proposed TBANC and the traditional MRFANC approach i.e. when no beamformer or source targeting system is active. To guarantee that both approaches have already converged, the last 20s of the simulations are used in the above calculations.

3. RESULTS

3.1 Disturbance metric

The disturbance ratings obtained from the listening test, described in Sec. 2.2, were statistically analyzed to determine whether the type of noise (café, traffic and subway station) had an effect on the evoked disturbance of the listeners. Due to the ordinal nature of the disturbance ratings and the fact that the distributions of ratings for each type of noise were not normal, the non-parametric Friedman test was conducted along with post hoc analysis [19], to examine the statistically significant differences between the disturbance ratings of each type of noise. It was found that there were statistically significant differences in the distributions of disturbance ratings of the three difference types of noise ($\chi^2(2, n = 72) = 22.525, p < .001$). Dunn's pairwise tests were carried out using a Bonferroni correction, and statistically significant differences were found between traffic and café noise ($p < .001$) as well as traffic and subway station noise ($p < .001$). No significant differences were found between café and subway station noise ($p = .145$). Therefore, there is strong evidence that the traffic noise has higher Disturbance metric compared to subway station and café noise, while no statement can be made for the evaluated Disturbance metric of café noise compared to the subway station noise.

3.2 Sound Event Detection Results

The results for the SED task are shown in Tab. 1 for each of the three classes, along with the average precision, recall, and F1-score. The results show that the proposed method achieves a high precision and recall for all classes, with an average precision of 82% and recall of 89%. The lower precision in the café class is attributed to the fact that strong gusts of wind are also included in the respective recordings, thus making the café class more difficult to detect.

Table 1. Classification results for the SED task. The precision, recall, and F1-score are reported for each class along with their respective averages.

	Precision	Recall	F1-score
café	0.73	0.93	0.82
traffic	0.91	0.78	0.84
subway station	0.82	0.96	0.89
Avg.	0.82	0.89	0.85

3.3 Direction of Arrival Estimation Results

In this section, the results for the DOAE task are presented, given a correct SED classification. The results regarding the achieved MAE is shown across different SNRs in Fig. 5 for the 3 different SNR cases evaluated in this work.

The proposed method achieves a low MAE for all classes, with a total average MAE of 0.34°. The average MAE achieved for the café, traffic and subway station classes was $0.72°$, $0.20°$, $0.08°$ respectively.

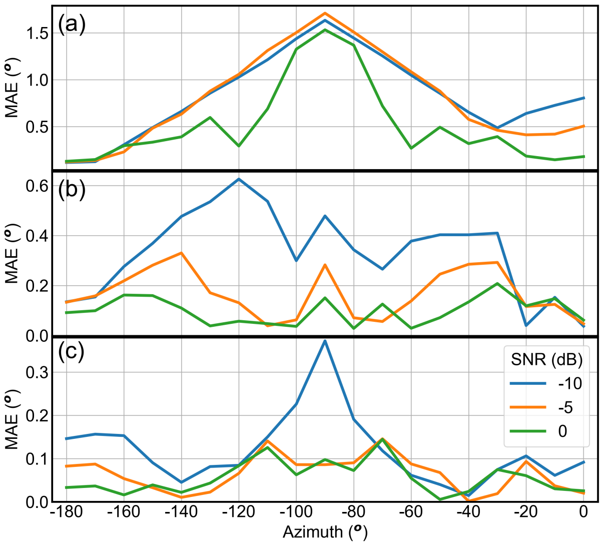

Figure 5. MAE for the DOA estimation task for the different classes averaged across SNRs: (a) café; (b) traffic; (c) subway station.

3.4 Convergence speed

Due to the difference between the two signals $m'_R(n)$ and $m_L(n)$ used to drive the respective adaptive filters the convergence speed of the MRFANC algorithm is negatively affected as can be seen in Fig. 6, with the proposed TBANC scheme converging to the steady-state error after $2.5s$ compared to the original MRFANC algorithm that converges after $1.3s$. The primary-to-diffuse SNR had no effect on the convergence speed of the algorithms.

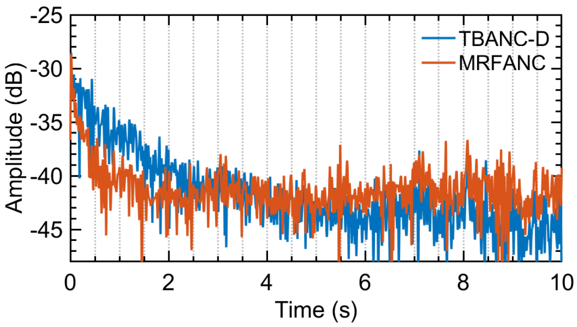

Figure 6. Error measured inside the right earcup for the proposed TBANC-D scheme (blue), compared to the original MRFANC approach (red).

3.5 Steady state performance

The steady state frequency domain results of the proposed TBANC-D approach can be observed in Fig. 7. The performance of MRFANC and the proposed approach are similar for frequencies $\leq$ 200Hz, but in the frequency range between 200Hz and 10kHz a significant improvement is observed by the TBANC-D reaching up to $20dB$ improvement in the $3-5$kHz region.

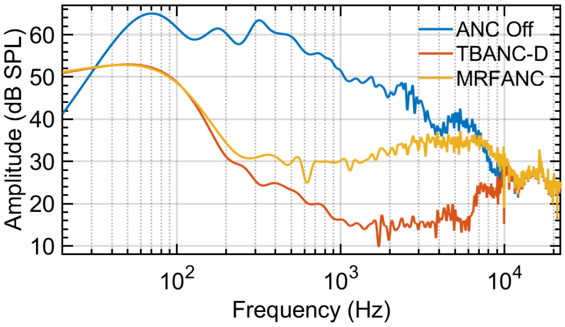

Figure 7. Spectra of the passive noise attenuation performance of the headphone shell (blue); TBANC-D approach (red); and the original MRFANC (yellow). The spectra are from the last 20s period.

In Fig. 8 the steady-state error improvement I_e (in dB) achieved by TBANC-D compared to MRFANC for different mixing scenarios is shown. The results show that the proposed approach achieves a significant improvement in the steady-state error for all cases, except for when the traffic noise plays the role of the secondary source i.e. the cafe-traffic scenario where performance deteriorates and the subway-traffic scenario where the performance improvement is negligible. This result however is compensated by the fact that according to the results presented in Sec. 2.2, this will never be the case, since the traffic noise consistently plays the role of the primary source due to the related disturbance metric.

4. SUMMARY & DISCUSSION

In this work TBANC-D, a novel Targeted through Beamforming ANC approach is proposed, which utilizes a novel Disturbance metric in order to steer a TDBF to the primary disturbing source using a SELD neural network, to more significantly attenuate the most disturbing source in a complex auditory scene, while also significantly attenuating the background noise field to acceptable levels.

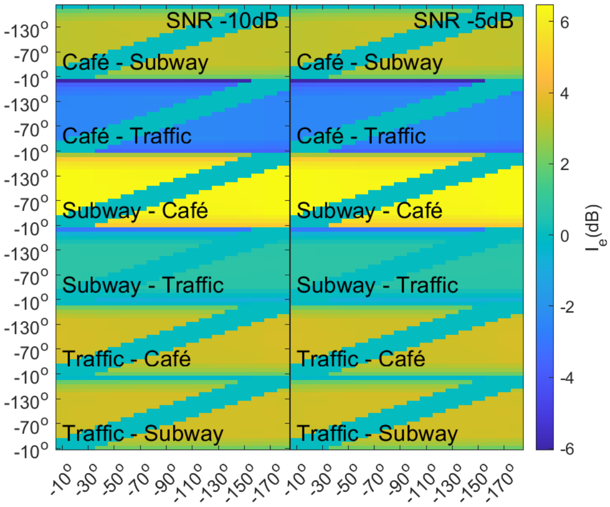

Figure 8. Steady-state error improvement I_e (in dB) achieved by TBANC-D compared to MRFANC for different *Primary* - *Secondary* source mixing scenarios.

The performance of the system was evaluated through the simulation of a diffuse sound field in the presence of up to 2 noise sources. Specifically the Precision, Recall and F1-Score were used to gauge the performance of the SED system, and the MAE of the DOAE system, with both achieving high performance in their respective tasks with the average F1-Score being 0.85 and the average MAE being 0.33^o. The TBANC component was evaluated with respect to the steady-state noise attenuation, as well as in terms of convergence speed compared to the established MRFANC approach, with the proposed TBANC approach achieving an improvement of up to 20dB in the $3-5kHz$ region, being especially important since human listeners have a significantly increased sensitivity in such region.

The results presented in this work, concern up to 2 active noise sources whose location remained the same throughout the simulation. It is important to note, that a scenario where the recordings utilized in this work would coincide to form an acoustic scene would be highly unlikely, however such noises are excellent representations of the types of noises that are encountered in real life scenarios, such as babble noise, wind, cutlery etc. Future research includes extending the proposed method in order to accommodate any number of sources with varying location, through the use of the employed SELD network, with similar works achieving exceptional results [11] in such scenarios. However, in such cases the available microphones would have to increase in order to achieve more directive beamforming. Furthermore, a personalized Disturbance metric would ideally be developed, in order to further improve the experience of the headphone listener.

5. REFERENCES

[1] L. Lu, K.-L. Yin, R. C. de Lamare, Z. Zheng, Y. Yu, X. Yang, and B. Chen, "A survey on active noise control in the past decade—part i: Linear systems," *Signal Processing*, vol. 183, p. 108039, 2021.

[2] J. Lorente, M. Ferrer, M. de Diego, and A. Gonzalez, "The frequency partitioned block modified filtered-x nlms with orthogonal correction factors for multichannel active noise control," *Digital Signal Processing*, vol. 43, pp. 47–58, 2015.

[3] M. T. Akhtar and W. Mitsuhashi, "Improving performance of fxlms algorithm for active noise control of impulsive noise," *Journal of Sound and Vibration*, vol. 327, no. 3-5, pp. 647–656, 2009.

[4] I. T. Ardekani and W. H. Abdulla, "Theoretical convergence analysis of fxlms algorithm," *Signal Processing*, vol. 90, no. 12, pp. 3046–3055, 2010.

[5] S. Liebich, J.-G. Richter, J. Fabry, C. Durand, J. Fels, and P. Jax, "Direction-of-arrival dependency of active noise cancellation headphones," *ASME 2018 Noise Control and Acoustics Division Session presented at INTERNOISE 2018*, 2018.

[6] P. Zachos and J. Mourjopoulos, "Beamforming headphone ANC for targeted noise attenuation," *Audio Engineering Society Convention 154*, May 2023.

[7] J. Cheer, V. Patel, and S. Fontana, "The application of a multi-reference control strategy to noise cancelling headphones," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3095–3103, 2019.

[8] C. Oberzut and L. Olson, "Directionality and the head-shadow effect," *The Hearing Journal*, vol. 56, no. 4, pp. 56–58, 2003.

[9] R. EBU-Recommendation, "Loudness normalisation and permitted maximum level of audio signals," *European Broadcasting Union*, 2011.

[10] N. Zacharov, *Sensory evaluation of sound*. CRC Press, 2018.

[11] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[12] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "Salsa-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720, IEEE, 2022.

[13] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[14] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," June 2013. Supported by Inria under the Associate Team Program VERSAMUS.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13001–13008, 2020.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.