

FIA 2020/22

XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA

XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC

Florianópolis, SC, Brasil

Analysis of a neural network approach to linear beamforming

Kuno, Y.M.¹; Masiero, B.S.²; Madhu, N.³

^{1,2} University of Campinas, Brazil, yugo4k@decom.fee.unicamp.br,
masiero@unicamp.br

² Ghent University – imec, Belgium, nilesh.madhu@ugent.be

Abstract

Beamforming algorithms are applied to signals captured by sensor arrays to enhance signals arriving from desired directions while reducing background noise and other interfering signals. Beamforming coefficients can be optimally derived, usually satisfying some second-order statistical criterion such as the mean noise variance or mean output power, if the direction of arrival of the desired and interfering sources, as well as a statistical description of the background noise, is available. However, this is usually not the case. In an attempt to blindly estimate optimal beamforming coefficients directly from the array signals, we proposed the use of a neural network architecture. The network is composed of several blocks of neural network layers, each block responsible for processing one discrete frequency bin in the short-time discrete Fourier representation of the signals. The input to each block is the microphone signal vector at the corresponding discrete frequency, and for a single time frame. The output is the set of beamforming coefficients for that frequency and time frame. By analyzing the array's beampattern when the position of both the target source and the interferer varies we observed that the proposed neural network can track both sources, keeping the array pointed to the target source while nulling out the direction of the interferer. Thus it effectively serves as an adaptive beamformer. This paper evaluates this architecture rigorously based on several different performance metrics for the case of a single target plus localised or non-localised interference sources.

Keywords: beamforming, neural network, blind source separation, speech enhancement.

PACS: 43.60.Fg, 43.60.Np, 43.60.Mn, 43.60.Hj.



1. INTRODUCTION

Beamforming is a collection of techniques that use the spatial information extracted from an array of microphones to filter the signal originating from specific positions or directions. The most simple of these techniques is known as time domain delay and sum beamforming [1]; it simply uses the different propagation delays between the desired source position and the microphone positions to shift the signal in each of the microphone channels in time, so that the signal from the desired source is aligned. Because the signal is usually digital, and therefore discrete, it becomes impossible to create a perfect alignment between the channels, which introduces quantization errors. Reducing these errors is possible with signal interpolation, but this is computationally intensive and introduces a new source of errors; the simpler solution is to align the microphone channels in the frequency domain instead.

The frequency domain delay and sum beamformer (DSB) [2] solves the quantization problem since shifting the signal in time is no longer restricted to discrete steps. Nevertheless, this beamformer still suffers from spatial aliasing: some frequency components of signals originating from different directions have the same phase differences between microphone channels, making it impossible to align the signals from the target direction without also aligning the frequency components from interfering signals from different directions. Signals from directions that are close to the target direction are also difficult to filter out: at lower frequencies the DSB has lower angle discrimination, and it becomes narrower and sharper as the frequency increases. As such, the DSB's beampattern exhibits a significant low pass behavior for interfering signals that is dependent on their proximity to the target direction.

The minimum variance distortionless response (MVDR) beamformer [2], instead of maximizing the signal from the target direction, uses the information from the geometric model and from the background noise to minimize the energy from directions outside of the target direction (it's necessary to estimate or model the background noise information). This makes the high

energy interfering bands in the beampattern become generally lower than the ones present in the DSB's beampattern, with a narrower acceptance range in the target direction for lower frequencies, but it still displays mainlobe narrowing as the frequency grows higher. Also, since the MVDR beamformer is a special case of the speech distortion weighted multichannel Wiener filter (SDW-MWF) [3] weighted so that the response is strictly distortionless, it produces a significant residual noise level. This beamformer can introduce cancellations in the direction of the interfering sources, but doing so requires precise knowledge of those directions along with the diffuse noise model.

The relative transfer function based minimum variance distortionless response (RTF-MVDR) beamformer follows the same design as the MVDR beamformer, but instead of using a geometric model of interference and noise sources, it uses the relative transfer function (RTF) generated from the separated interference and noise components of the mixture signal to minimize their energy [4]. To actually implement this beamformer in a practical scenario, a very robust activity detection system is generally required, since its performance will depend entirely on the premise that the signal used on the RTF is sufficiently representative of the undesired sources and that it is not contaminated by the desired signal. On the other hand, this beamformer can be useful as a benchmark oracle knowledge beamformer, since the components of the mixture signal can be taken as part of the desired oracle knowledge.

The relative transfer function based minimum power distortionless response (RTF-MPDR) beamformer is very similar to the RTF-MVDR beamformer, the difference being that the whole mixture signal is used to generate the RTF to minimize the energy from the undesired sources [4]. Because the minimum power distortionless response (MPDR) design is based on minimizing the power of the whole mixture (instead of only that of the undesired sources) while still providing a distortionless response that depends only on the correct steering of the beamformer, no activity detection system is required, making it simpler to apply in practice.

While the MVDR, RTF-MVDR and RTF-MPDR beamformers require and use *a priori* information about the precise direction of desired and interference sources, as well as the background noise, the machine learning based nMLP beamformer proposed by the author in [5] goes one step further: it uses the complete information available from the microphone array captured signal as input to generate a beamforming weight vector. This vector changes according the mixture signal characteristics for each time frame. Because the beamforming vector may vary from frame to frame, the nMLP beamformer is able to use information about the target and interfering directions (which is encoded in the recorded data) to adapt to varying source directions and statistical characteristics. The datacentric model of machine learning (ML) also makes it simple to train this neural network (NN) to filter signals that come from a range of directions instead of a single target, resulting in a beampattern that has an acceptance direction range that is nearly constant for all frequencies (other than the very lowest ones), with a sharper and stronger attenuation for the directions outside of that range.

The details of the proposed network architecture and the training method used are discussed in the coming section. Following that, in Section 3, the results obtained with the proposed method are presented and compared with classical beamformers using different evaluation metrics.

2. MACHINE LEARNING BASED BEAMFORMING

The ML based beamforming approach used in this neural network architecture follows the same model of other frequency-domain linear beamforming techniques: the input data is analyzed and a set of beamforming coefficients is generated for each frequency. The following subsections go into the details of the proposed architecture and the training method used.

2.1 Neural network architecture

The nMLP starts by normalizing the energy of the complex valued audio data $\mathbf{x}$ for each time frame t_b and each frequency f_k is used as the input of a multilayer perceptron (MLP) block.

This block is composed of two perceptron layers (with weight matrices $\mathbf{P}_1$, $\mathbf{P}_2$ and bias vectors $\mathbf{p}_1$, $\mathbf{p}_2$), using a non-linear activation function g with a trained slope parameter λ . The output of the MLP block is a set of beamforming coefficients $\mathbf{w}_{NN}$ for the same time frame and frequency of the inputs. These coefficients are then conjugate transposed and multiplied to the original audio input array as usual in linear-beamforming, generating the beamformed signal $\hat{\mathbf{y}}$. The breakdown of these operations, for a single time frame and frequency, is shown in the Equation set (1):

$$\begin{aligned}\bar{\mathbf{x}} &= \frac{\mathbf{x}}{\|\mathbf{x}\|} \\ \mathbf{q} &= \mathbf{P}_1 \bar{\mathbf{x}} + \mathbf{p}_1 \\ \mathbf{r} &= g(\mathbf{q}, \lambda) \\ \mathbf{w}_{NN} &= \mathbf{P}_2 \mathbf{r} + \mathbf{p}_2 \\ \hat{\mathbf{y}} &= \mathbf{w}_{NN}^H \mathbf{x}\end{aligned}\quad (1)$$

The non-linear activation function g is a sigmoid function that acts only on the amplitude of the outputs of the first MLP layer, so that the phase doesn't change. If the activation function is applied to the real and imaginary components separately, the resulting phase doesn't follow a straightforward pattern, and the phase information is lost. The MLP block for each frequency also has its own sigmoid slope parameter λ , which is fine tuned during training, and the function is of the form shown in Equation (2). Figure 1 shows how the shape of that sigmoidal function changes according to the variation of the λ parameter — the third row shows that there's no difference between the phase distribution of the input and of the output.

$$g(\mathbf{q}, \lambda) = \frac{\mathbf{q}}{\sqrt{\lambda^2 + \|\mathbf{q}\|_2^2}} \quad (2)$$

Each frequency f_k has a specific MLP block to analyze the data and generate its beamforming coefficients, such that this neural network beamforming architecture is composed of as many MLP blocks as the resolution of the short time Fourier transform (STFT) applied to the signal. Appropriately, the proposed neural network

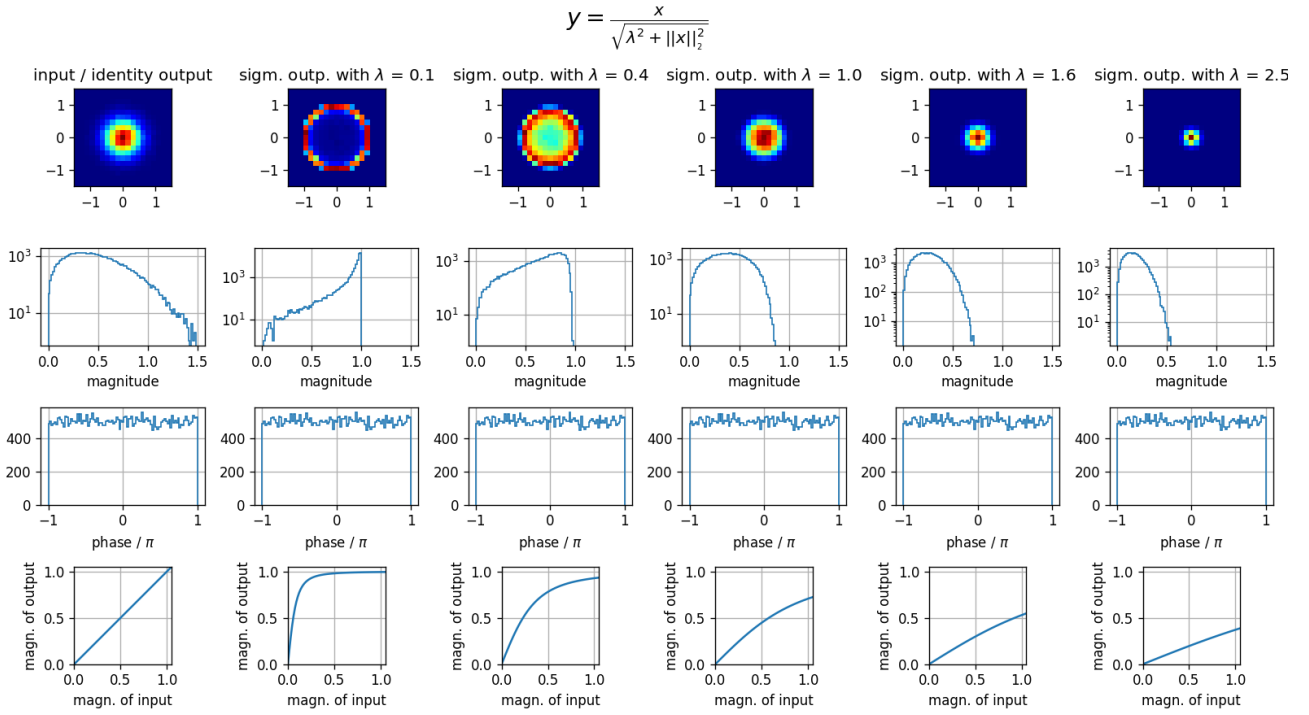


Figure 1: Examples of how the magnitude sigmoidal activation function works on complex values. From left to right, the first column shows the characteristics of the values used as input to the activation function, while the remaining columns show the outputs of this sigmoidal activation function for values of λ as follows: 0.1, 0.4, 1.0, 1.6 and 2.5. From top to bottom, the first row shows the complex value histograms (real values on the x axis and imaginary values on the y axis), the second row shows the magnitude distribution, the third row shows the phase distribution (they're all identical by design), and the fourth row shows the magnitude transformation provided by the sigmoid function (input on the x axis and output on the y axis).

beamformer was named nMLP beamformer, and a diagram of its architecture and the generation of the output signal is shown in the Figure 2.

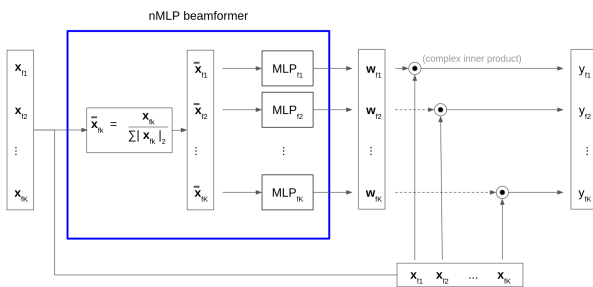


Figure 2: Architecture of the neural network beamformer, composed of multiple MLP blocks, each one generating linear beamforming coefficients for one frequency component of the STFT of the input signal.

2.2 Training data

An ideal training dataset for a ML based beamformer would be composed of recordings of multiple speakers, recorded in all possible acoustic conditions, and using all conceivable microphone arrays. This is clearly an impossible task,

and as such the compromise made to generate the training dataset was to synthesize speech signals that weren't tied to any single room acoustic condition and that were made specifically to one microphone array per training.

Each NN instance is trained for one microphone array geometric configuration; this configuration is used to generate near-field impulse responses for free-field conditions. The choice of free-field propagation was made so that the training result didn't overfit to any particular choice of specific acoustic conditions – at this stage of development the ML beamformer isn't trained with signals with any sort of reverberation.

For the results presented here, the chosen array was a hexagonal grid planar array of seven microphones of the model UMA-8 USB mic array - V2.0, from miniDSP, which has its microphones are spaced 4.3 cm apart from each other [6]. The microphone array is show in Figure 3.

These impulse responses were then convolved

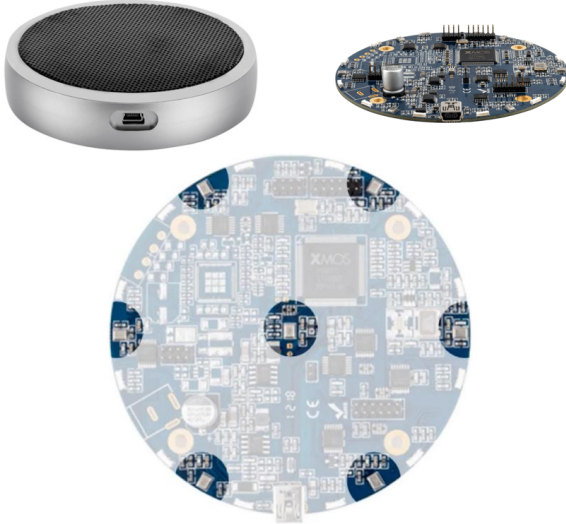


Figure 3: Microphone array used in the results presented. Top left shows the array with its casing, top right shows the array itself, and bottom part shows a highlight of the seven microphone positions.

with semi-anechoic monaural speech signals obtained from the CSTR VCTK English Multi-speaker Corpus dataset [7]. These convolved signals have their sources set to positions inside the desired direction regions or outside of them, which are then used to train the neural network to preserve desired signals (the ones inside) and filter out interference signals (the ones outside).

Moreover, spatial uncorrelated noise is added to the input signal so that the nMLP beamformer learns to not only filter out electronic noise, but also limit the directional gain from directions where there is neither desired signal nor interference signal.

3. RESULTS

To understand how the NN architecture of the nMLP beamformer works, beampatterns generated directly from the weight matrices and bias vectors are presented and discussed. Beampatterns generated from the output beamforming coefficients are then presented to demonstrate the tracking capabilities of the proposed beamformer, displaying not only its target tracking but also its interference tracking. Finally, the nMLP beamformer performance is compared to that of classical beamforming algorithms by a set of evaluation metrics.

3.1 Interpreting how the nMLP beamformer NN works

The first layer of the architecture is applied directly to the normalized complex valued audio data and, as such, its main function is to analyze this input data. Figure 4 shows some example beampatterns generated from the first layer weight matrices for each of the frequencies: because the signal may come from various directions, the resulting beampatterns have a number of lobes with varying positions to sense the directions of arrival in the input mixture. Since the architecture does not take into consideration any correlation between the the MLPs for each frequency component, there is a large amount of variation between the lobes for each frequency in the beampatterns shown.

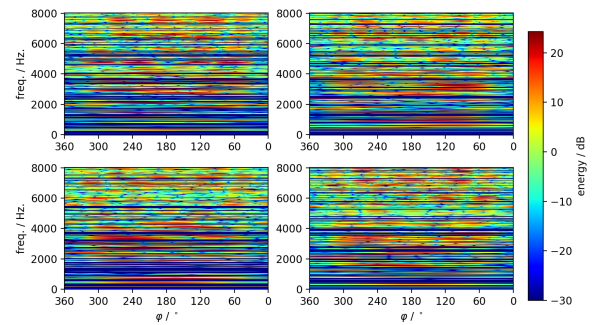


Figure 4: Beampatterns generated from 4 samples of multiplicative weights from the first layer of parameters of the nMLP.

The second layer bias vectors provide a static additive component to the beamforming coefficients output of the nMLP beamformer. It follows that the beampattern generated from these parameters will display the basis of the beamforming behavior of the proposed architecture, a starting point on top of which the input data will adjust the beampattern to the specific source positions. Figure 5 displays exactly that: a wide mainlobe that covers the target direction range of 60° with a mostly uniform gain, showing little narrowing for higher frequencies, and limited broadening for lower frequencies. This is possible because the NN training allows the beamformer to vary the gain intensity across microphones, allowing control of the mainlobe width over different frequencies.

The second layer weight matrices are responsible

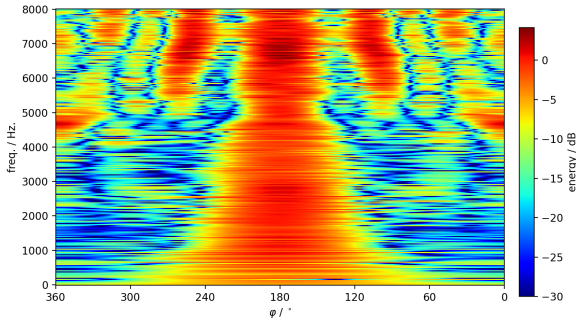


Figure 5: Beampattern generated from the bias weights from the second layer of parameters of the nMLP.

for those adjustments to the final beampattern of the output coefficients, since they're multiplied by the result of the first layer operation over the input signal. Considering that these adjustments are made over the second layer bias basis, it makes sense that Figure 6 shows a number of lobes with varying directions, but all of them display low gain in the target direction range — because the basis provided by the bias of the second layer already generates a part of the mainlobe over the target direction range, there's no need for this component to reproduce a behavior that will be mostly stable for the audio input data. Instead, these parameters are better used making fine adjustments to tweak the mainlobe or cancel the directions of interference, as will be seen in the next subsection.

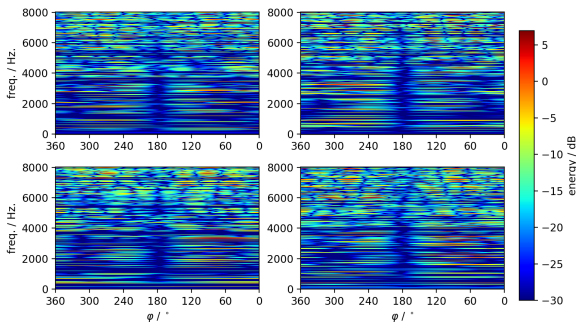


Figure 6: Beampatterns generated from 4 samples of multiplicative weights from the second layer of parameters of the nMLP.

3.2 Tracking

As just explained in the previous section, Figure 7 shows how the nMLP beamformer is able to create an attenuation region in the direction from the active interference source (left average beampattern, where there is a column of low gain

at the interference direction of -30°). Not only that, but the mainlobe is tweaked to keep a sharp gain border at the left side, while sacrificing the right side with more gain, since there is no active interference in that direction range.

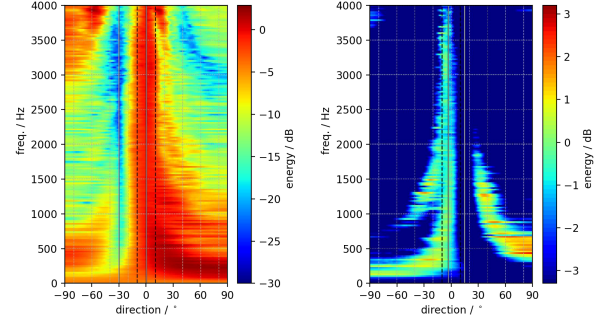


Figure 7: Average beampatterns obtained for a uniform linear microphone array, for scenarios with one desired and one interference source positions, which are displayed with white lines in both plots. The target direction range, between -10° and $+10^\circ$, is delimited by black dashed lines. In the left plot, the desired signal is at 0° and the interference is at -30° . In the right plot, the desired signal is at -3° and the interference is at $+15^\circ$. Please note that the plots have different color scales.

The colormap of the subplot on the right side of Figure 7 was zoomed in to a range between -3 dB and +3 dB to highlight the direction of highest gain and display how the actual target direction (in this case -3°) is prioritized over the whole target direction range (which goes from -10° to $+10^\circ$). This allows a lower gain even inside the target region so that the interference source at $+15^\circ$ is better attenuated, as long as that doesn't compromise the specific direction of the current target source.

3.3 Speech metrics

To evaluate the nMLP beamformer's performance, a set of metrics were chosen to contrast results against a group of oracle knowledge beamformers, and these tests were conducted using real-world acoustic conditions. The chosen metrics are the delta segmental signal to interference ratio (Δ sSNIR), the noise attenuation ratio, the perceptual evaluation of speech quality (PESQ) and the short-time objective intelligibility (STOI) measure.

The Δ sSNIR evaluates the average signal to noise plus interference ratio (SNIR) improvement pro-

vided by the beamformer compared to the original mixture signal, and it is measured as shown in Equation (3)

$$\Delta \text{sSNIR} = \frac{1}{B} \sum_b (\text{sSNIR}_{\text{output}}(b) - \text{sSNIR}_{\text{input}}(b))$$

$$\text{sSNIR}(b) = \left[10 \log_{10} \left(\frac{\sum_{k,m} (S_d(b,k,m))}{\sum_{k,m} (S_{n+i}(b,k,m))} \right) \right]_{-20\text{dB}}^{+35\text{dB}} \quad (3)$$

where b is the time index of the signal frame and B is the total number of frames, S_d is the desired signal component and S_{n+i} is the interference and noise component of the mixture, k is the frequency component index and m is the microphone channel index. In the case of the $\text{sSNIR}_{\text{input}}$, this m index goes through all microphone channels, while in the case of the $\text{sSNIR}_{\text{output}}$ there is only one output channel for the index to go through. The $\text{sSNIR}(b)$ values are limited between -20 dB and +35 dB so that outlier values don't have too great an effect over the average beamforming results.

The noise attenuation metric simply compares the average channel energy of the input noise and interference signals over those of the output signal, as shown in Equation (4)

$$\text{NA} = \frac{\frac{1}{M} \sum_{b,k,m} (|S_{n+i}(b,k,m)|^2)}{\sum_{b,k} (|\hat{S}_{n+i}(b,k)|^2)} \quad (4)$$

where M stands for the number of microphone channels in the input signal, S_{n+i} stands for the input noise and interference signal component of the mixture, and $\hat{S}_{n+i}$ stands for the output signal generated from beamforming only those same noise and signal components.

The PESQ and STOI measures are broadly used in acoustics and are respectively described in [8] and [9].

The beamformers chosen for benchmarking were the MVDR, the RTF-MVDR, and the RTF-MPDR. These beamformers are considered to

have oracle knowledge because

- the MVDR has perfect knowledge of the position of the desired and interference sources: the steering vector for the desired source used was calculated considering the near-field propagation from the source position, and the covariance matrix was obtained considering the near-field propagation vectors from the interference sources and an ideal diffuse noise distribution.
- the RTF-MVDR has perfect knowledge of the position of the desired source and perfect knowledge of the statistics of interference and noise: the steering vector is calculated as the one for the MVDR beamformer, while the covariance matrix is generated from the RTF obtained from interference and diffuse noise components of the mixture input signal.
- and the RTF-MPDR has perfect knowledge of the position of the desired source and perfect knowledge of the statistics of mixture input signal: it is generated in the same manner as the RTF-MVDR except for the covariance matrix, which uses the RTF obtained from the complete mixture input signal.

The real-world conditions referred to were produced by using RIRs recorded in one of the classrooms located in Ghent University. For each source, semi-anechoic speech signals were convolved with one of these RIRs, generating the directional signals arriving at the microphone array. These signals are considered part of the desired signal if they come from the target direction range, and interference signals if outside of that same range. Additionally, diffuse white noise and spatially uncorrelated noise are added to generate the mixture signal that is the input passed to the beamformers.

3.3.1 Test configuration 1: one desired source and one interference source

The first test uses only one active source in the target direction range and only one active interference source outside of it, while the diffuse and spatially uncorrelated noises are set to an SNR



of 10 dB. Using a reference frame centered at the central microphone of the array, the desired source is positioned one at 170° , the interference source is positioned at 250° , and both are at a 1 meter distance. This configuration is shown in Figure 8.

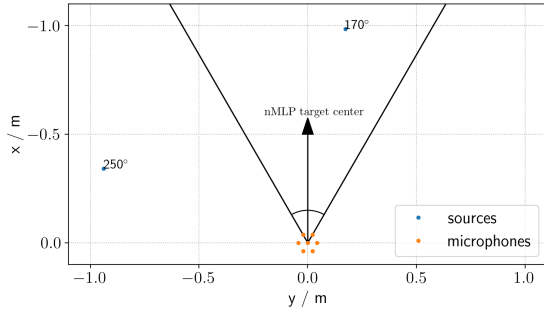


Figure 8: Map for the test configuration 1, with one target source and one interference source as shown. The target position is 10° off-center in regards to the beamformer direction range, which goes from 150° to 210° .

All benchmark beamformers are steered to the precise location of the desired source, while the nMLP beamformer direction range goes from 150° to 210° (so that the desired source is 10° off-center).

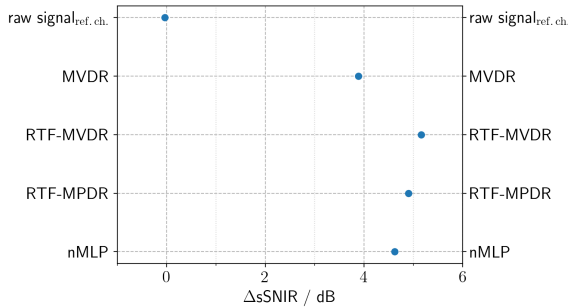


Figure 9: $\Delta s\text{SNIR}$ results for test configuration 1.

Figure 9 shows that all beamformers provide improvements above 3.5 dB of $\Delta s\text{SNIR}$ in regards to the raw input signal. While the RTF-MVDR and RTF-MPDR results surpass the proposed nMLP beamformer as expected (since they are oracle knowledge beamformers) the nMLP beamformer results are superior to those of the MVDR beamformer, even though the latter is directly pointed to the desired source.

Figure 10 shows that while the benchmark beamformers provide 5 to 6 dB of noise attenuation,

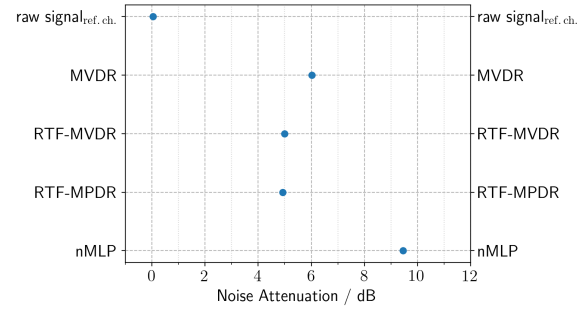


Figure 10: Noise attenuation results for test configuration 1.

the same result for the nMLP beamformer is above 9 dB. This result might seem surprising at first, but it is explained by the fact that the nMLP neural network wasn't trained to generate a distortionless response as the benchmark beamformers, but to minimize the error between the output signal and the desired signal. As such, it does attenuate the output of the desired signal where that compromise produces an output that is closer to the one desired, while the distortionless response beamformers will output interference and noise when it is physically impossible to separate them from the desired signal.

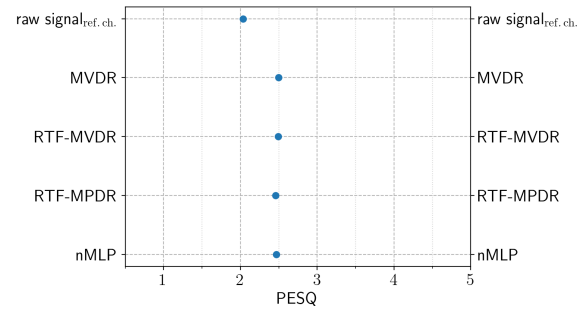


Figure 11: PESQ results for test configuration 1.

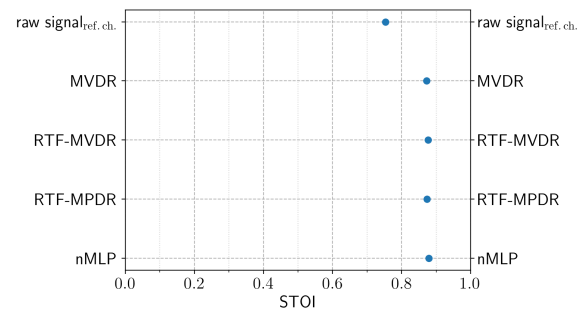


Figure 12: STOI results for test configuration 1.

Figures 11 and 12 show that all beamformers tested provide improvements to speech quality

and speech intelligibility that have no significant difference, raising the PESQ score from ≈ 2.0 to ≈ 2.5 , and the STOI score from ≈ 0.75 to ≈ 0.88 .

3.3.2 Test configuration 2: one desired source and two interference sources

This second test still uses only one active source, but it has two active interference sources independent from each other. The diffuse and spatially uncorrelated noises are also set to an SNR of 10 dB. With the same reference frame used for the previous configuration, Figure 13 shows that the desired source is still positioned at 170° , while the two interference sources are positioned at 110° and 250° , all of them at a 1 meter distance from the center of the array.

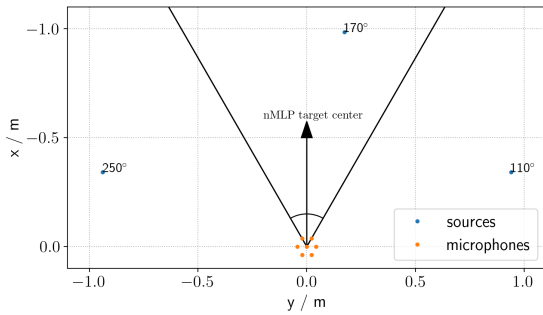


Figure 13: Map for the test configuration 2, with one target source and two interference sources as shown. The target position is 10° off-center in regards to the beamformer direction range, which goes from 150° to 210° .

As before, the benchmark beamformers are steered to the precise location of the desired source, while the nMLP beamformer direction range goes from 150° to 210° (so that the desired source is 10° off-center).

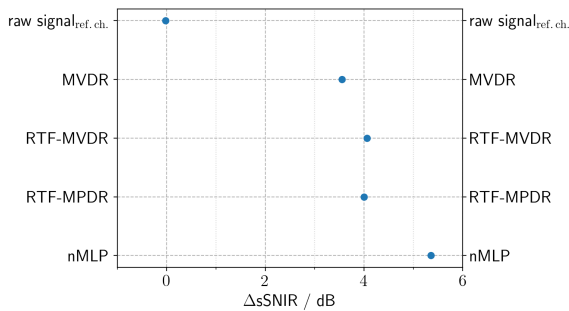


Figure 14: $\Delta s\text{SNIR}$ results for test configuration 2.

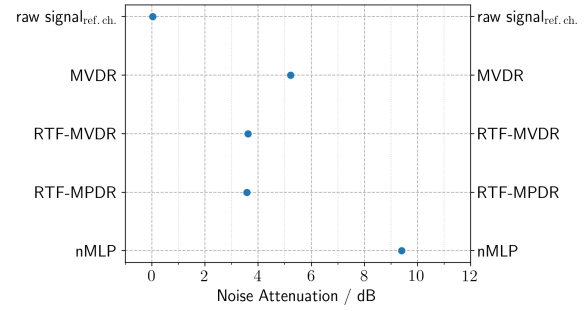


Figure 15: Noise attenuation results for test configuration 2.

The improvement of the $\Delta s\text{SNIR}$ and noise attenuation by the nMLP beamformer is even more marked against these two interference sources. The benchmark beamformers improve only around 3.5 to 4.1 dB to the $\Delta s\text{SNIR}$, while the nMLP beamformer enhance it above 5 dB. The noise attenuation is limited between 3.5 and 5.2 dB for the benchmark beamformers, and above 9 dB for the nMLP beamformer. This is also a reflection of how the beamformer proposed is able to achieve a better compromise between preserving the desired signal and attenuating interference and noise signals, which is more important for intelligibility than providing a perfectly distortionless response.

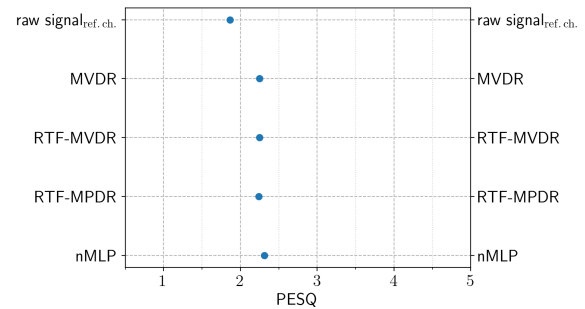


Figure 16: PESQ results for test configuration 2.

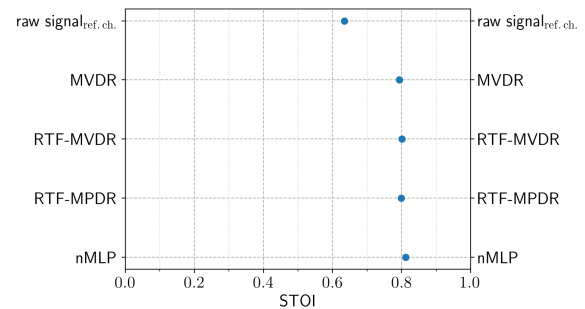


Figure 17: STOI results for test configuration 2.



As before, the improvements on the PESQ and STOI scores are nearly the same for all beamformers, as it can be seen in Figures 16 and 17.

3.3.3 Test configuration 3: two desired sources and two interference sources

The third and final test is only supposed to compare the performance of the proposed beamformer against the benchmark beamformers in scenarios in which there are multiple independent active sources, so that the desired signal may come from two different sources simultaneously. Since the benchmark beamformers used are designed to point only to a single direction, the compromise chosen for this test was to point the steering vector to a direction in the middle of the two desired source directions. This test also uses two independently active interference sources, and the diffuse and spatially uncorrelated noises are also set to an SNR of 10 dB. With the same reference frame used for the previous configurations, Figure 18 shows that the desired sources are positioned at 170° and 190° , while the two interference sources are positioned at 110° and 250° as before, all of them at a 1 meter distance from the center of the array.

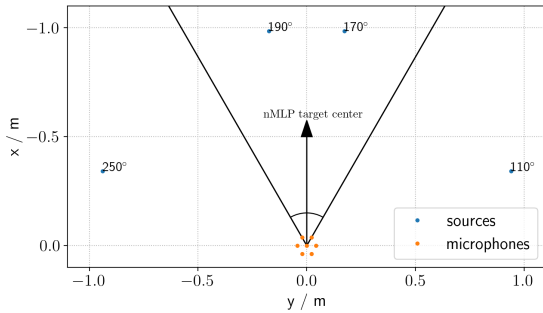


Figure 18: Map for the test configuration 3, with two target source and two interference sources as shown. The target positions are 10° off-center in regards to the beamformer direction range, which goes from 150° to 210° .

While the benchmark beamformers are now steered to the compromise direction of 180° , the nMLP beamformer direction range still goes from 150° to 210° (so that both desired source are 10° off-center).

Figures 19 and 20 show that the steering compromise is very detrimental to the performances of both the RTF-MVDR and the RTF-MPDR

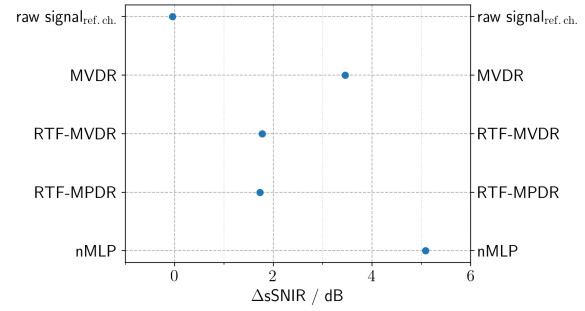


Figure 19: $\Delta s\text{SNIR}$ results for test configuration 3.

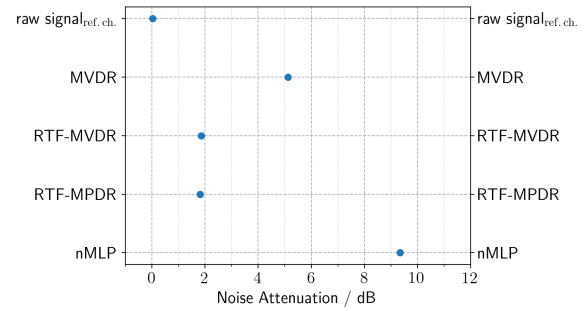


Figure 20: Noise attenuation results for test configuration 3.

when compared to the MVDR results, while the nMLP beamformer still shows an excellent performance. Both RTF beamformers provide an improvement of only ≈ 1.8 dB to the $\Delta s\text{SNIR}$, while the MVDR result is ≈ 3.4 dB and the nMLP result is above 5 dB. The noise attenuation is a little below 2 dB for the RTF beamformers, above 5 dB for the MVDR and above 9 dB for the nMLP beamformer.

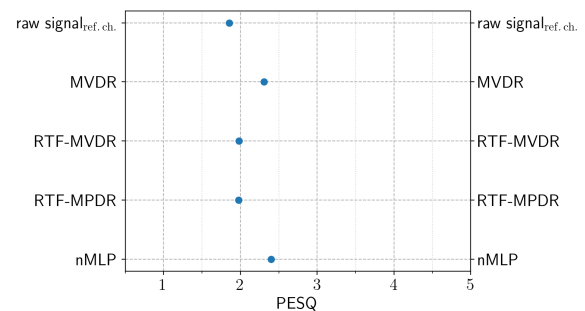


Figure 21: PESQ results for test configuration 3.

The double source configuration is also detrimental for the PESQ and STOI results of the RTF beamformers, though both the MVDR and nMLP beamformers show similar results.

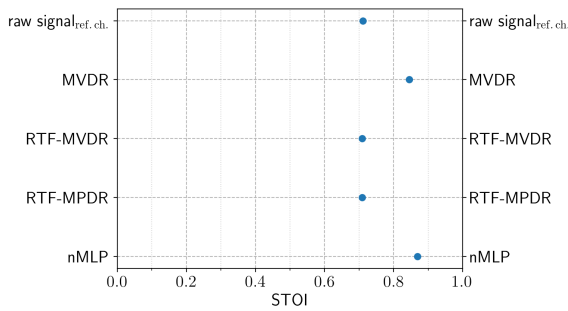


Figure 22: STOI results for test configuration 3.

4. CONCLUSION

In this contribution we showed that it is possible to construct a linear beamformers using neural networks trained for a specific microphone array, and these ML based beamformers are able to track target and interference sources based on a target direction range set during training. The results provided by the nMLP neural network beamformer show that the architecture is able to adapt to a varying the number and position of sources, and that the resulting beampattern is optimized by its innate source tracking. Furthermore, it displays a performance that is comparable to that of oracle beamformers without the necessity of additional methods to detect the direction of arrival (DOA), detect active speech from the desired sources, do preliminary source separation or do signal segmentation.

Even though the proposed beamformer exhibited results comparable to oracle-based beamformers, the architecture is still extremely simple, based on very basic two-layer MLP blocks that do not correlate information across frequencies or time frames. This shortcoming limits the beamformer performance, since the beampattern for each frequency component and time frame can vary wildly, which gives rise to audible artifacts.

These drawbacks will be addressed in a future NN beamformer design based on convolutional neural networks. These networks architectures are able to analyze correlations between neighboring elements, such as different frequency components and time frames from the STFT. Because the DOA may be better estimated by taking into account information extracted from multiple frequencies, and adjacent time frames

can help confirm and stabilize the beampattern of the output coefficients, it is expected that such an architecture will provide even more significant results.

5. ACKNOWLEDGMENTS

The contributions received from Alexander Bohlender, Anthony McAtear, Yanjue Song and Sushmita Thakallapalli are greatly appreciated by the authors.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and in part by the São Paulo Research Foundation (FAPESP), grant #2017/08120-6.

REFERENCES

- [1] Michael Brandstein, Darren Ward. *Microphone Arrays*. Springer, 2001. ISBN 978-3-662-04619-7.
- [2] Krim, H. e Viberg, M. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996. doi:[10.1109/79.526899](https://doi.org/10.1109/79.526899).
- [3] Gannot, Sharon; Vincent, Emmanuel; Markovich-Golan, Shmulik e Ozerov, Alexey. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017. doi:[10.1109/TASLP.2016.2647702](https://doi.org/10.1109/TASLP.2016.2647702).
- [4] H. Moore, Alastair; Hafezi, Sina; R. Vos, Rebecca; A. Naylor, Patrick e Brookes, Mike. A compact noise covariance matrix model for mvdr beamforming. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, págs. 1–14, 2022. doi:[10.1109/TASLP.2022.3180671](https://doi.org/10.1109/TASLP.2022.3180671).
- [5] Kuno, Yugo M.; Masiero, Bruno S. e Madhu, Nilesh. A neural network approach to broadband beamforming. Em *Proc. 23rd International Congress on Acoustics*, Aachen, Germany, 2019.
- [6] *UMA-8 v2 Product datasheet*. miniDSP, 5 2018. v2.
- [7] Yamagishi, Junichi; Veaux, Christophe e MacDonald, Kirsten. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019.
- [8] Rix, A.W.; Beerends, J.G.; Hollier, M.P. e Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Em *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, págs. 749–752 vol.2, 2001. doi:[10.1109/ICASSP.2001.941023](https://doi.org/10.1109/ICASSP.2001.941023).



- [9] Taal, Cees H.; Hendriks, Richard C.; Heusdens, Richard e Jensen, Jesper. A short-time objective intelligibility measure for time-frequency weighted noisy speech. Em *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, págs. 4214–4217, 2010. doi:[10.1109/ICASSP.2010.5495701](https://doi.org/10.1109/ICASSP.2010.5495701).