Regression versus classification for neural network based audio source localization

Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain Serizel, Alexandre Guérin

REGRESSION VERSUS CLASSIFICATION FOR NEURAL NETWORK BASED AUDIO SOURCE LOCALIZATION

Lauréline Perotin,[1,2] *Alexandre Défossez,*[3,4] *Emmanuel Vincent,*[2] *Romain Serizel,*[2] *Alexandre Guérin*[1]

[1] Orange Labs, Cesson-Sévigné, France
[2] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[3] Faceboook AI Research, Paris, France
[4] INRIA/ENS PSL Research University, Paris, France

ABSTRACT

We compare the performance of regression and classification neural networks for single-source direction-of-arrival estimation. Since the output space is continuous and structured, regression seems more appropriate. However, classification on a discrete spherical grid is widely believed to perform better and is predominantly used in the literature. For regression, we propose two ways to account for the spherical geometry of the output space based either on the angular distance between spherical coordinates or on the mean squared error between Cartesian coordinates. For classification, we propose two alternatives to the classical one-hot encoding framework: we derive a Gibbs distribution from the squared angular distance between grid points and use the corresponding probabilities either as soft targets or as cross-entropy weights that retain a clear probabilistic interpretation. We show that regression on Cartesian coordinates is generally more accurate, except when localized interference is present, in which case classification appears to be more robust.

***Index Terms*—** Direction-of-arrival, training criterion, cost-sensitive classification, soft target, angular loss

1. INTRODUCTION

Direction-of-arrival (DOA) estimation is of major importance for various applications such as speech enhancement [1, 2] and automatic speech recognition (ASR) [3, 4]. Classical methods rely on physical modeling of the acoustic scene. Many are based on the time difference of arrival between microphones, for instance via the generalized cross-correlation with phase transform (GCC-PHAT) [5]. Steered response power (SRP) algorithms build acoustic maps by scanning the space with a beamformer [6]. Subspace algorithms such as MUSIC [7] and ESPRIT [8] use the eigenvalue decomposition of the covariance matrix of the signal to separate the contributions of point sources and diffuse noise. Other methods rely on clustering of the acoustic intensity vector [9]. These methods are not able to model the sound scene in real-world situations with reverberation, ambient noise, and where sources are not perfectly punctual, resulting in degraded localization performance [10, 11].

To overcome the limitations of physical modeling, data-driven approaches propose to use supervised learning in order to grasp the complexity of acoustic phenomena. Pioneer works made use of kernel estimators [12], ridge regression [13], support vector machines [14] or Gaussian mixture models [15].

The breakthrough of neural networks brought drastic performance improvements in ASR [16], speech enhancement [17], and more lately DOA estimation [18]. In a supervised learning framework, the problem can be formulated either as a regression or as a classification problem. When the output space is not structured and is discrete, for example in the case of image recognition, classification is an obvious choice. On the contrary, for problems with a structured and possibly continuous output space, both formulations have assets and drawbacks. This has been discussed for audio source counting [19], where classification slightly outperformed regression; for object localization in images, where classification coupled with Bayesian probabilistic modeling proves to be more accurate than regression [20]; for audio generation, where the SING algorithm [21] uses regression to achieve a significant gain in complexity and better rendering than Wavenet [22], which uses classification.

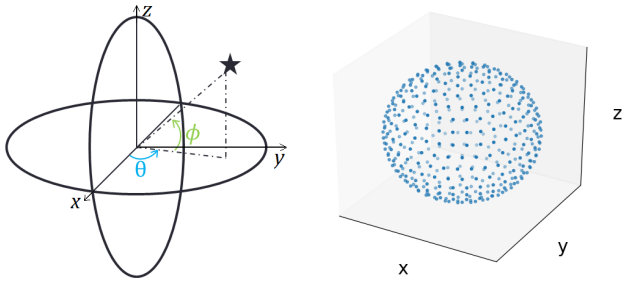
In the context of DOA estimation of audio sources, although the output space is highly structured, most neural network based systems rely on multi-label binary classification on the discretized unit sphere [18, 23–25], as this was reported to perform better than regression in a footnote in Xiao et al. [18]. Two notable exceptions can be found: He et al. [26] reintroduced a structure in the output DOA space by likelihood-based encoding of the output of the network, and Adavanne et al. [27] proposed a regression based formulation where the output of the network is formed with the Cartesian coordinates on the unit sphere corresponding to the target DOA. However, none of these works compare their performance with the more common multiclass formulation. To our knowledge, no such comparative study has been led for DOA estimation so far.

In this article, we investigate the impact of the framework (classification versus regression) on DOA estimation with neural networks. We build on the convolutional and recurrent neural network (CRNN) architecture designed in our previous work on single-source DOA estimation for Ambisonics recordings [25]. We compare several classification and regression approaches by using targets and loss functions that are adapted to the geometry of the problem. We notably propose a classification network that accounts for the spherical structure of the output space while enabling a clear probabilistic interpretation.

In Section 2, we clarify the formulations of DOA estimation as regression and classification problems. Section 3 presents different systems for each formulation. We describe our experiments in Section 4 and their results in Section 5. We conclude in Section 6.

2. FRAMEWORKS FOR THE LOCALIZATION PROBLEM

The goal of this article is to compare various formulations of the DOA estimation problem as regression or classification. We restrict



(a) (b)

Figure 1: (a) Target for regression: azimuth θ and elevation ϕ of the speaker seen from the microphone array. (b) Target for classification: discretization of the unit sphere corresponding to (2).

our study to the situation where a single static speaker needs to be localized in a reverberant environment with ambient noise. We seek to estimate the DOA, that is to say the azimuth and elevation of the speaker with respect to the center of the microphone array. Estimating the distance between the source and the microphone is beyond the scope of this article.

2.1. Regression

In the regression formulation, the goal is to directly recover an estimate of the azimuth θ and elevation ϕ corresponding to the DOA of the source in the spherical coordinate system centered on the microphone array (see Fig. 1a):

$$\begin{cases} \theta \in (-180^\circ, 180^\circ] \\ \phi \in [-90^\circ, 90^\circ] \end{cases}. \tag{1}$$

2.2. Classification

In the classification formulation, the neural network outputs a score for each class on the discretized unit sphere. The class that is the closest to the actual DOA (θ, ϕ) should get the highest score.

In this work, we use the grid of points ψ_{ij} (see Fig. 1b):

$$\begin{cases} \phi_i &= -90 + \frac{i}{I} \times 180 \qquad \text{with } i \in \{0, \dots, I\} \\ \theta^i_j &= -180 + \frac{j}{J^i+1} \times 360 \text{ with } j \in \{0, \dots, J^i\}, \end{cases} \tag{2}$$

where $I = \lfloor \frac{180}{\alpha} \rfloor$ and $J^i = \lfloor \frac{360}{\alpha} \cos \phi_i \rfloor$ with α the desired grid resolution in degrees. The resulting grid contains $n_{\text{DOA}} = \sum_{i=0}^{I} (J^i + 1)$ points. Any other quasi-uniform grid could be used.

3. PROPOSED SOLUTIONS

3.1. A shared neuronal basis

We use the input features and the neural network architecture described in our previous work [25]. We consider that the signals are available in the first-order Ambisonics format [28], which consists of 4 channels called W, X, Y, and Z, and we use the 6-channel normalized intensity vector as input. This was proved to be suitable for robust localization, including in real-life scenarios [11]. Only the last layer and the cost function vary between the baseline [25] and the systems proposed in the following.

The shared architecture of the networks is depicted in Fig. 2. It takes several frames of input at a time in the short-time Fourier transform (STFT) domain. The first part is made of three convolutional blocks. Each of them consists of a convolutional layer across time and frequency with a rectified linear unit (ReLU) activation function, followed by a batch normalization layer and max-pooling across frequencys. A second part is made of two bidirectional long short-term memory (BiLSTM) layers with hard-sigmoid recurrent activation and hyperbolic tangent kernel activation. It is followed by two fully-connected feed-forward (FF) layers. The first FF layer has a linear activation. The activation of the last layer depends on the system.

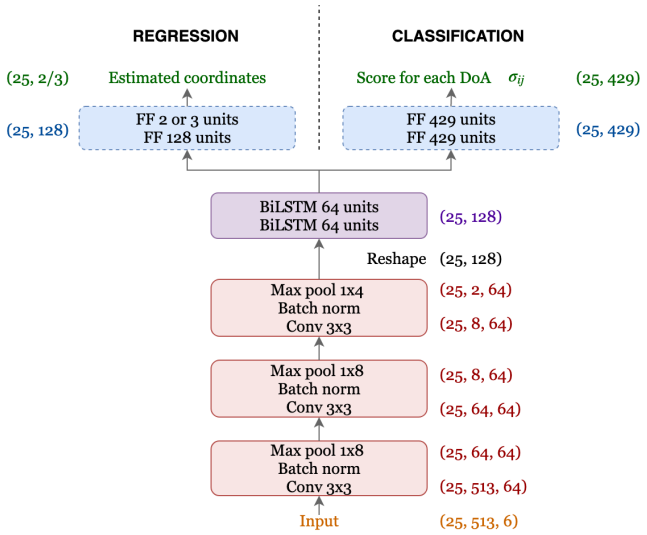


Figure 2: Architecture of the neural networks for localization. Right: with the output for regression. Left: with the output for classification.

3.2. Configurations for regression

We designed three neural networks to estimate the DOA directly by regression.

- **Regression with spherical target**: this network has two output units with θ and ϕ as targets. Mean-square error (MSE) is used for the loss. This loss does not account for the spherical geometry of the outputs.
- **Regression with spherical target and angular loss**: this network is identical to the previous one, except that the function used for training is the angular distance δ between the prediction $\hat{\psi} = (\hat{\theta}, \hat{\phi})$ and the actual DOA $\psi = (\theta, \phi)$:

$$\begin{aligned} \delta(\hat{\psi}, \psi) = \arccos\{ &\sin(\hat{\phi}) \sin(\phi) \\ &+ \cos(\hat{\phi}) \cos(\phi) \cos(\hat{\theta} - \theta)\}. \end{aligned} \tag{3}$$

  This loss accounts for the spherical geometry of the output space. However, mapping from directions on the unit sphere to azimuth and elevation coordinates is unstable near the poles.
- **Regression with Cartesian target**: in order to solve the latter issue, we design a network that targets the three Cartesian coordinates of the unit vector pointing towards the DOA, as in the work of Adavanne et al. [27]. MSE is used for the loss. In this case, it actually represents a geometrical distance between the prediction and the true DOA.

For all these networks, the labels are scaled between 0 and 1 and a sigmoid activation function is applied on the last layer. We also tried linear activation, with or without clipping, with worse results. As a post-processing in the prediction step, we average the outputs on all frames of the sequence, in order to return one prediction per sequence.

3.3. Configurations for classification

We designed three more neural networks to predict the DOA by classification using the architecture defined in Fig. 2. They are all based on the previously defined grid (2) and hence have n_{DOA} output units.

- **Cross-entropy (CE) with one-hot encoded target**: we use a softmax activation function at the output layer of the network and optimize a CE loss. The target distribution is taken as the one-hot in the grid point which is closest to the actual DOA.
- **MSE with soft Gibbs target**: we can induce some structure on the output space with a softer target: a Gibbs distribution [29] with energy taken as the angular distance between grid points ψ_{ij} and the true DOA ψ, similarly to He et al. [26]:

$$\mathcal{G}(\psi_{ij}) = e^{-\delta[\psi_{ij},\psi]^2/\beta^2} \tag{4}$$

  where δ is the angular distance (3) and β defines an angular neighborhood. For this network, we use sigmoid activations in the last layer and the MSE loss. The interpretability of the output as a probability distribution is thus lost.
- **Gibbs-weighted loss**: to keep the structure of the output space while allowing a clear probabilistic interpretation, we integrate the Gibbs distribution in the cross-entropy loss using cost-sensitive weights:

$$\text{loss} = -\log(\sigma_{ij}) - \sum_{\substack{(i',j') \\ \neq (i,j)}} (1 - \mathcal{G}(\psi_{i'j'}))\log(1 - \sigma_{i'j'}) \tag{5}$$

  where ψ_{ij} is the grid point that is the closest to the actual DOA and σ_{ij} is the output of the network for the class ψ_{ij} after a sigmoid activation. For $\beta = 0$, if the true DOA is on the grid, we recover the cross-entropy loss with one-hot encoded target.

In the prediction step, as for regression, the outputs σ_{ij} of the network are averaged over all frames of a sequence for all grid points. The estimated DOA then corresponds to the class with the highest global score.

4. EXPERIMENTS

4.1. Data

Training samples are generated from a set of spatial room impulse responses (SRIRs) simulated with the image method, thanks to an adaptation of Habets' generator [30] for Ambisonics. 42,900 rooms were generated with random dimensions in $[2.5, 10] \times [2.5, 10] \times [2, 3]$ m and a reverberation time RT60 uniformly drawn within $[0.2, 0.8]$ s. In each room, a microphone array is placed randomly at a minimum distance of 50 cm from any wall. The distance between the microphones and the sources is uniformly drawn between 1 and 3 m. Three SRIRs are generated in each room, leading to a total of 128,700 SRIRs. The first SRIRs in each room are enforced to be uniformly spread on the sphere. This ensures that all DOAs are significantly represented in the dataset. To synthesize the audio signal for learning, each SRIR is then convolved with a 1 s speech signal extracted from a subset of the French corpus Bref [31] composed of 44 speakers. Ambient noise is generated by convolving babble noise from Freesound[1] with a diffuse SRIR, made by averaging the diffuse parts of two real SRIRs randomly picked among 42 SRIRs recorded from the center of a real reverberant room. The noise is added with a signal-to-noise ratio (SNR) comprised between 0 and 20 dB. The validation set is constructed in the same manner with 1,287 unseen simulated SRIRs, unseen speakers from Bref and unseen babble noise.

The three datasets used for testing the performance of the systems are described in our previous work [11]. The simulated SRIRs dataset is made similarly to the training and validation sets, with 1,287 new SRIRs, English speakers from the SiSEC challenge [32] and unseen noise. The real SRIRs dataset is constructed the same way, except that the SRIRs come from a dataset of 576 SRIRs measured in a real room with 16 loudspeaker and 36 microphone array positions. Finally, the real recordings dataset is made of 3 speakers reading texts from 14 different positions in a living room. A TV was also recorded independently. It is used to create another version of this dataset with additional TV noise at 10 dB SNR.

4.2. Algorithm parameters

All signals are sampled at 16 kHz. The STFT is performed with a 1024-point sine window and 50% overlap for both analysis and synthesis. Each 1 s utterance is split into two sequences of 25 frames with 12 overlapping frames between sequences. The input dimensions for the networks are hence (25, 513, 6).

The convolutional layers use 64 filters of size 3×3. Max-pooling is performed along 8 frequency bands for the first two layers and 4 bands for the third one. BiLSTM layers include 64 hidden units. For classification, the grid is built with a step $\alpha = 10^\circ$, resulting in $n_{\text{DOA}} = 429$ DOAs. In that case, FF layers include 429 hidden units. For regression, the first FF layer includes 128 units while the second layer has as many units as there are values to estimate (2 or 3). In all cases, the Nadam optimizer is used for learning with an initial learning rate of 10^{-3}. 50% dropout is applied after each convolutional block, each FF layer and on the recurrent weights of the BiLSTM layers. Training is stopped after the performance on the validation set has stopped increasing for 20 epochs.

For the classification networks with a soft Gibbs distribution target or with a Gibbs-weighted loss, we set the neighborhood parameter $\beta = 2\alpha = 20^\circ$.

4.3. Performance measurement

The performance is measured in terms of angular accuracy with respect to given thresholds, that is to say the percentage of sequences where the angular distance (3) between the prediction and the actual DOA is below 5°, 10° or 15°. For classification, the grid is such that some DOAs are 7° apart from the closest point on the grid. In those cases, 5° accuracy is impossible to achieve. However, the resolution of the grid is one of the limitations of the classification approach and should not be disregarded. We additionally report the mean and median angular errors.

[1] `https://freesound.org`

(a) Real SRIRs

	Accuracy (%)			Ang. err. (°)	
Algorithm	<5°	<10°	<15°	mean	med.
Reg. sph. MSE	28.5	60.2	81.2	11.4	8.1
Reg. sph. ang. loss	29.7	64.3	80.0	11.4	7.9
Reg. Cartesian	**37.1**	**72.9**	**88.3**	**8.5**	**6.4**
Class. one-hot	26.6	66.8	85.3	**9.7**	8.0
Class. Gibbs target	26.0	67.7	**86.4**	**9.3**	**7.5**
Class. Gibbs loss	27.6	67.7	**84.5**	**9.7**	**7.2**

(b) Recordings

Reg. sph. MSE	25.4	71.7	87.6	11.4	7.1
Reg. sph. ang. loss	**49.8**	**90.8**	95.9	**7.1**	**5.0**
Reg. Cartesian	46.8	87.7	93.3	9.1	**5.3**
Class. one-hot	22.6	85.5	**98.0**	**7.5**	5.9
Class. Gibbs target	25.0	83.0	**98.3**	**7.3**	5.9
Class. Gibbs loss	27.3	88.8	**98.0**	**6.9**	5.7

Table 1: Performance (a) on the test sets built with the real SRIRs, (b) on real recordings without TV. The best results are in bold. 95% confidence intervals vary between $\pm 0.3\%$ and $\pm 2.8\%$ for accuracy, and $\pm 0.6°$ and $\pm 1.6°$ for angular error.

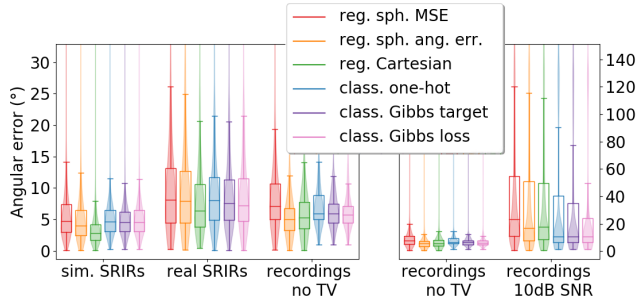


Figure 3: Violin plots of the angular errors for all algorithms and each test set. The y-axis scale changes between left and right plots. The boxes show the first and third quartiles as well as the median. The lower (resp. higher) ends of the whiskers correspond to the lowest (resp. highest) values within 1.5 interquartile range (IQR) of the lower (resp. upper) quartile.

5. RESULTS

The violin plots (Fig. 3) show that on the simulated SRIRs test set, which is similar to the training set, all three classification networks perform equally well and better than the regression networks targeting the spherical coordinates. Regression with Cartesian target performs best. On real SRIRs (Table 1a), the overall results worsen, but the ranking of the networks remains unchanged. The difference of performance between the three classification networks and the Cartesian regression network is hardly significant. On real recordings, the regression network with the angular loss slightly outperforms Cartesian regression, with 49.8% of the sources localized with less than 5° error against 46.8%. For a tolerance of 15°, classification networks perform better, with accuracies between 98.0 and 98.3% against 95.9% for the regression network with angular loss.

The right plot of Fig. 3 shows the results on the real recordings with an interfering TV at 10 dB SNR. Although it is not a point source, it is more localized than the diffuse babble noise used in training. The number of outliers increases for all systems, but classification appears to be more robust than regression in this situation. This is confirmed when observing the outputs of the networks (Fig. 4). Regression results are biased in the direction of the TV, while for classification, the scores of the classes in the directions of the TV increase, but in many cases they do not overtake the score of the class corresponding to the DOA of the speaker.

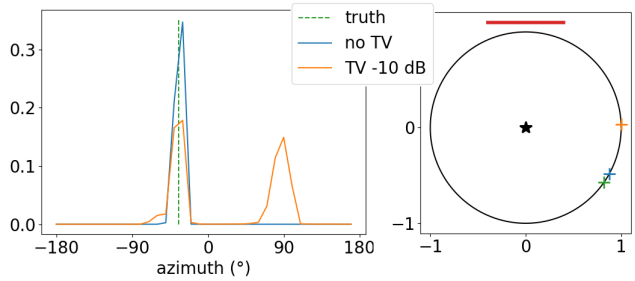


Figure 4: Outputs of the one-hot classification (left) and Cartesian regression (right) networks for a specific situation, with (orange) and without (blue) TV. The true DOA is shown in green and the approximate position of the TV in red. We restrain the plots to the median plan, in which all positions are roughly included.

For all test sets, classification with a Gibbs distribution as the target is slightly better than with a one-hot target. Using Gibbs-weighted loss further improves the performance, especially for real recordings where the number of outliers significantly decreases. It additionally speeds up the training, with only 60 epochs needed instead of 100 with a comparable computation time per epoch. This could be related to the vanishing gradient issue observed in [33].

6. CONCLUSION

We have compared three regression networks trained to recover the spherical coordinates (with MSE or angular loss) or the Cartesian coordinates of the source and three classification networks. We have introduced a simple framework for cost-sensitive classification using Gibbs weights, enabling efficient training and more accurate results than its one-hot target counterpart. This approach can easily be extended to any objective with a distance between labels.

Before drawing any general conclusion, we would like to highlight the fact that no single system stands out. The results are tightly linked to the evaluation scenario and metric. However, we can state that regression (preferably on Cartesian coordinates) is as legitimate as classification, although it was disregarded in the majority of previous works. It even appears to be more accurate in scenarios with diffuse interference. On the contrary, classification seems more robust to localized interference.

In this work, we have studied the ability of a neural network to deal with a regression or a classification formulation for the localization of a single source. For multiple sources, an implementation issue arises for regression: the number of outputs would depend on the number of sources (or at least the maximum number). In addition, label ambiguity should be handled during training. Classification might then be a preferable solution, if future works show that the performance is equally good or better than with a regression formulation.

7. REFERENCES

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[4] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[6] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of ICASSP*, vol. 1, 1997, pp. 375–378.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[8] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoustics, Speech, Sig. Proc.*, vol. 37, no. 7, pp. 984–995, 1989.

[9] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *JAES*, vol. 53, no. 12, pp. 1115–1127, 2005.

[10] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer Science & Business Media, 2008.

[11] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE JSTSP*, pp. 22–33, 2019.

[12] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail-party processor," in *Proc. of NIPS*, 2004, pp. 1425–1432.

[13] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Lang.*, vol. 14, no. 6, pp. 2156–2164, 2006.

[14] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. of IWAENC*, 2014, pp. 99–103.

[15] T. May, S. Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang.*, vol. 19, no. 1, pp. 1–13, 2011.

[16] G. Hinton, L. Deng, D. Yu, G. E. Dahl, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[17] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[18] X. Xiao *et al.*, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. of ICASSP*, 2015, pp. 2814–2818.

[19] F. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *Proc. of ICASSP*, 2018, pp. 436–440.

[20] M. Everingham and A. Zisserman, "Regression and classification approaches to eye localization in face images," in *Proc. of FGR*, 2006, pp. 441–446.

[21] A. Défossez, N. Zeghidour, N. Usunier, L. Bottou, and F. Bach, "SING: Symbol-to-instrument neural generator," in *Proc. of NIPS*, 2018, pp. 9041–9051.

[22] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," *arXiv:1609.03499 [cs]*, 2016, arXiv: 1609.03499.

[23] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. of WASPAA*, 2017, pp. 136–140.

[24] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 12, pp. 2444–2453, 2017.

[25] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *Proc. of IWAENC*, 2018, pp. 241–245.

[26] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. of ICRA*, 2018, pp. 74–79.

[27] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *J-STSP*, 2018, arXiv: 1807.00129.

[28] M. A. Gerzon, "Periphony: with-height sound reproduction," *JAES*, vol. 21, no. 1, pp. 2–10, 1973.

[29] J. W. Gibbs, *Elementary principles in statistical mechanics*. Charles Scribner's Sons, 1902.

[30] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., 2006.

[31] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proc. of Eurospeech*, 1991, pp. 505–508.

[32] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation," in *Proc. of ICA*, 2009, pp. 734–741.

[33] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs squared error training: a theoretical and experimental comparison," in *Proc. Interspeech*, vol. 13, 2013, pp. 1756–1760.