



Deep-sound field analysis for upscaling ambisonic signals

Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M. Hegde

► To cite this version:

Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M. Hegde. Deep-sound field analysis for upscaling ambisonic signals. EAA Spatial Audio Signal Processing Symposium, Sep 2019, Paris, France. pp.1-6, 10.25836/sasp.2019.14 . hal-02275176

HAL Id: hal-02275176

<https://hal.archives-ouvertes.fr/hal-02275176>

Submitted on 30 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DEEP-SOUND FIELD ANALYSIS FOR UPSCALING AMBISONICS SIGNALS

Gyanajyoti Routray, Sourya Basu, Pranay Baldev, and Rajesh M Hegde

Department of Electrical Engineering

Indian Institute of Technology, Kanpur, India

{groutray, souryab, bpranay, rhegde}@iitk.ac.in

ABSTRACT

Higher Order Ambisonics (HOA) is a popular method for rendering spatial audio. However, the desired sound field can be reproduced over a small reproduction area at lower ambisonic orders. This problem can be handled by upscaling B-format signals using several methods both in the time and frequency domain. In this paper, a novel Sequential Multi Stage DNN (SMS-DNN) is developed for upscaling Ambisonic signals. The SMS-DNN allows for training of a very large number of layers since training is performed in blocks consisting of a fixed number of layers. Additionally, the vanishing gradient problem in DNN with a large number of layers is also effectively handled by the proposed SMS-DNN due to its sequential nature. This method does not require prior estimation of the source locations and works in multiple source scenarios. Reconstructed sound field analysis, subjective and objective evaluations conducted on the upscaled Ambisonic sound scenes indicate reasonable improvements when compared to the benchmark HOA reproduction.

1. INTRODUCTION

Spatial sound reproduction using Higher Order Ambisonics (HOA) is one of the most promising techniques for spatial audio reproduction [1, 2]. The knowledge of spherical harmonic decomposition is used to render spatial sound herein. But such a rendering has the limitation of low spatial resolution. Spatial resolution can be improved by increasing the number of loudspeakers during reproduction [3–5]. The preferred number of loudspeakers (L) is computed using the inequality $L \geq (N + 1)^2$, N being the order of HOA [5]. Additionally, a large number of loudspeakers results in an under-determined system of equations. Consequently increase in number of loudspeakers is not a good choice [6]. However, it can be improved by the combined effort of upscaling the Ambisonic order and increasing the number of loudspeakers during sound reproduction. The upscaling can be done using compressed

sensing technique [7] in time domain. Alternatively, in the frequency domain it can be performed as in [8]. These are the sparsity based methods where the source and its direction are computed using an overcomplete spherical harmonics dictionary. The accuracy of this method depends on selection of the dictionary and estimation accuracy of the source location. These techniques have a limitation on the number of sources that can be rendered accurately. Real time upscaling can also be performed from multi channel recordings using spherical microphone array (SMA) such as Eigenmike[®] [9] and Visisonics[®] [10]. The number of microphones available and the design of spherical microphone array limits the order of HOA to 4 and 7 respectively.

In this work, a Sequential Multi Stage Deep Neural Network (SMS-DNN) for upscaling the order of Ambisonic signals is proposed and developed. The source sound is recorded using a tetrahedron microphone which gives a B-format (order-1 ambisonics) encoded signal [11]. The recordings are represented as four channels: W (omnidirectional) and (X, Y, Z) channels (bidirectional sounds in the direction of x , y , and z axes) respectively. These signals are upscaled into order- N HOA encoded plane wave sounds using the SMS-DNN in this work. Subsequently a complete framework is developed for upscaling Ambisonic signals using the SMS-DNN.

The rest of the paper is organized as follows. In Section 2 the problem for upscaling the lower order Ambisonics is described. The proposed SMS-DNN for upscaling Ambisonic signals is described in Section 3. The performance of the proposed method is evaluated in Section 4. Section 5 concludes the paper.

2. PROBLEM FORMULATION


A source vector consisting of P plane waves is given by

$$\mathbf{s} = [s_1, s_2, \dots, s_P]$$

Where $(\theta_{s_1}, \phi_{s_1}), (\theta_{s_2}, \phi_{s_2}), \dots, (\theta_{s_P}, \phi_{s_P})$, represent the location of the plane wave sources. Individually, θ represents the elevation, and ϕ the azimuth measured anticlockwise from x -axis. In general the HOA encoded signal is computed as [3]

$$\mathbf{B} \triangleq \mathbf{Y}\mathbf{s} \quad (1)$$

where $\mathbf{Y} = [Y_{nm}(\theta_{s_1}, \phi_{s_1}), \dots, Y_{nm}(\theta_{s_P}, \phi_{s_P})]$ defines the spherical harmonics matrix, $n = 0, \dots, N$ and $m =$

 © Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M Hegde. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Gyanajyoti Routray, Sourya Basu, Pranay Baldev, Rajesh M Hegde. "Deep-Sound Field Analysis for Upscaling Ambisonics Signals", 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019.

$0, \dots, n$ being the order and degree of spherical harmonic coefficients respectively. The spherical harmonic coefficients are defined as

$$Y_{nm}(\theta, \phi) = \frac{1}{2} \sqrt{\frac{(2n+1)(n-|m|)!}{\pi(n+|m|)!}} P_n^{|m|}(\cos \theta) e^{im\phi} \quad (2)$$

$P_n^{|m|}(\cdot)$ is defined as the normalized associated Legendre function of degree n and order m . From (2), it can be observed that every increase in order adds a pair of lobes. A simple way of decoding the encoded HOA is to place speakers in alignment with the direction of spherical harmonic functions and assign gains proportional to the directivity pattern of the source [6]. The sound field produced in such a decoding method is influenced by the interference width of the directivity pattern. For higher order Ambisonics the directive patterns are narrower, which results in improved spatial resolution. In the case of loudspeakers which are arranged uniformly in an icosahedron pattern on a sphere, the simplest way of obtaining the loudspeaker feeds (decoding) using HOA method is given by

$$\mathbf{g} = \mathbf{D}_{nm} \mathbf{B} \quad (3)$$

where $\mathbf{D}_{nm} = [Y_{nm}(\theta_{l_1}, \phi_{l_1}), \dots, Y_{nm}(\theta_{l_L}, \phi_{l_L})]^\dagger$, L being the number of loud speakers satisfying $L \geq (N+1)^2$, and \dagger representing the pseudo inverse of a matrix. Increasing the number of loudspeakers results in an undetermined system of equations. Hence for any improvement in the area of reproduction it is required to increase the ambisonic order. Due to the limitation on the number of microphones in spherical array processing, the order cannot be arbitrarily increased. In case of lower order ambisonic (B-format) audio recordings, the order can be up-scaled only if the source direction is known. Additionally this method is limited to single source scenarios only. In this context the problem of upscaling can be modeled as a transfer $\mathfrak{F}(\cdot)$, that transfers the lower order spherical harmonics $Y_{nm}(\theta, \phi)|_{N=1}$ to the higher order spherical harmonics $Y_{nm}(\theta, \phi)|_{N>1}$. Due to the linear dependency between the HOA coefficients and the source signal the up-scaling process can be defined as

$$Y_{nm}(\theta, \phi)|_{N=1} \mathbf{s} \xrightarrow{\mathfrak{F}} Y_{nm}(\theta, \phi)|_{N>1} \mathbf{s} \quad (4)$$

For multiple sources this can be formulated as

$$\mathbf{Y}_{nm} \mathbf{s} \xrightarrow{\mathfrak{F}} \mathbf{Y}_{nm}^{\text{upscaled}} \mathbf{s} \quad (5)$$

Where \mathbf{Y}_{nm} is the spherical harmonic matrix corresponding to the location of sources. In-order to develop a flexible, scalable, and high resolution method for upscaling Ambisonic signals, a Sequential Multi Stage DNN (SMS-DNN) is proposed and developed in this work. The additional novelty of this method also lies in the fact that no prior estimation of source locations is required even in multiple source scenarios.

3. SMS-DNN FOR HOA ENCODING

The SMS-DNN consists of sequentially stacked DNNs, where each of the stacked DNNs upscales the order of the signal by 1. One of the most important properties of spherical harmonics is that the components of the signals are independent of each other. Additionally for a particular (θ, ϕ) , increase in the order of signal only adds coefficients to the higher order Ambisonics, keeping the lower order Ambisonic coefficients unchanged. This property motivated the training of the DNNs independently for each order upscaling. Fig.1 shows the structure of the DNN for upscaling a signal from order 1 to order N .

3.1 Training the DNN

In this section, development of the dataset for training the DNN and subsequent upscaling is discussed. An algorithm is also detailed for the proposed method of upscaling Ambisonic signals.

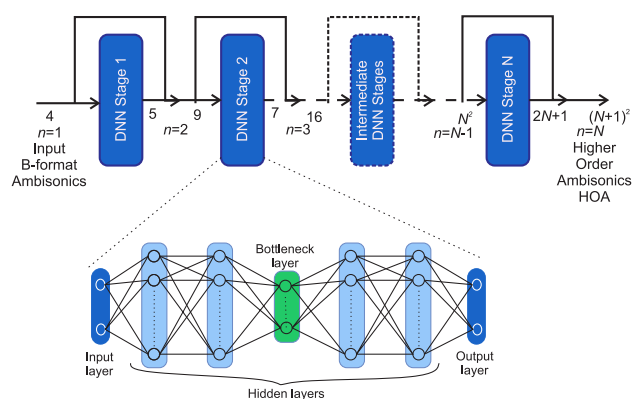


Figure 1. Sequential Multi Stage Deep Neural Network structure for upscaling Ambisonic signals.

3.1.1 Dataset for Training

The input training data of the deep neural network consists of an order N_l encoding of a mixture of sound signals located at K random locations (with random θ and ϕ). The output training data consists of a higher order encoding of the same mixture of sound sources with similar locations as the input data. For upscaling a signal from order N_l to N_u , we generate $N_u - N_l$ sets of training data, where the j^{th} dataset corresponds to the upscaling of the signal from order $N_l + j - 1$ to $N_l + j$. Additional details on the sequential training of the DNN is as follows.

3.1.2 Algorithm development for upscaling Ambisonic signals using SMS-DNN

Training a deep neural network for upscaling is a non-trivial problem, since the input to the DNN is a mixture of signals generated at different locations, hence there can exist multiple solutions for θ_s, ϕ_s and the amplitude of the sound sources given only the mixture. However, such solutions can be complex for a simple feed-forward network to compute. Also, we note that, for $N_u > N_l$, the order N_u

signal of length $(N_u + 1)^2$ has the initial $(N_l + 1)^2$ values exactly equal to an order N_l signal. Further, note that the essential information that a DNN should extract from the input is a very small number of unknowns, such as location and amplitude parameters of the sound source for reproducing the entire higher order signal. The proposed SMS-DNN model for Ambisonic upscaling is capable of using a large number of layers without facing the typical problems such as vanishing gradient faced while increasing the number of layers in a single DNN [12]. The proposed method also uses a sequential approach to train the neural network. If the required upscaling task is from order N_l to N_u , then $N_u - N_l$ separate networks are trained independently. The j^{th} DNN is used to upscale from order $N_l + j - 1$ to $N_l + j$. Further, while upscaling a signal from order $N_l + j$ to $N_l + j + 1$, only the last $2(N_l + j) + 3$ entries of the upscaled values are required to be trained. Thus only $2(N_l + j) + 3$ output nodes are required in the j^{th} DNN, which makes the DNN fast and efficient even for large values of N_l and N_u . Finally, since the relevant information that needs to be extracted from the input layers is very small, a bottleneck hidden layer is introduced at the middle of the neural network with a smaller number of nodes [13, 14]. It was observed that using the bottleneck layer helps in faster convergence of the DNN compared to standard feed-forward DNN.

Algorithm 1: SMS-DNN for upscaling ambisonics signal to order N

1 Training:

- 2 Generate B-format dataset for K randomly located sound sources.
- 3 **for** $j = 1 : N - 1$ **do**
- 4 Train the j^{th} DNN with order j sound signals as input and last $2j + 3$ elements of order $j + 1$ signal as desired output.
- 5 Concatenate the $2j + 3$ output to the order j input signal to form the order $j + 1$ signal.
- 6 The order $j + 1$ signal obtained from the concatenation is the input for the next DNN.

7 **Return:** Trained SMS-DNN.

8 Upscaling:

- 9 **Input:** Order 1 encoded signal of K sound sources.
- 10 **for** $j = 1 : N - 1$ **do**
- 11 Feed the j^{th} DNN with the order j input and get $2j + 3$ output elements.
- 12 Concatenate the $2j + 3$ output to the order j input signal to form the order $j + 1$ signal.
- 13 The order $j + 1$ signal obtained from the concatenation is the input for the next DNN.

14 **Return:** Order N signal.

4. PERFORMANCE EVALUATION

Performance of the proposed method is evaluated using sound field reconstruction analysis, subjective and objective evaluations.

4.1 Experimental Conditions

The proposed method produces high resolution Ambisonics encoded signals from B-format encoded signals. Hence three sound scenes are created having five sounds each to evaluate the method. The scenes are created such that at any interval all the five sounds are overlapping. Each of the sound scenes are of length 10 to 15 seconds. B-format ambisonics signal of all the three sound scenes are upscaled to order 2 -order 7 using the SMS-DNN¹.

At the same time upscaled signal is also obtained using (1), which is referred to as benchmark encoded HOA signal. The mean square error between these two encoded signals is obtained and plotted in Figure 2. From the Figure 2, it is clear that the error propagates with the upscaling of ambisonic signals but it is bounded to $-5dB$.

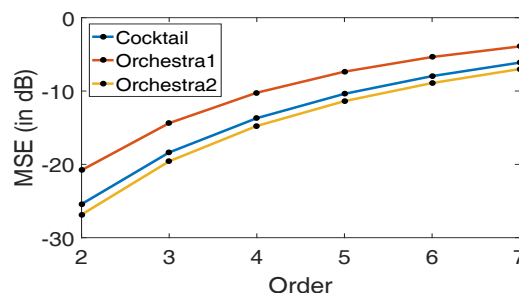


Figure 2. MSE between the reference and upscaled ambisonic encoded signals of various sound scenes for order 2 to 7.

4.1.1 Architecture of SMS-DNN

The proposed DNN for upscaling Ambisonic signals consists of a sequence of $N-1$ fully connected feed-forward neural networks, where each network is trained separately [15]. Each of the DNNs have 5 hidden layers, which consists of 300 nodes except the 3rd layer where only 20 nodes used to introduce bottleneck DNN structure.

4.1.2 Training Dataset

SMS-DNN was trained using 4×10^5 number of sample data points, where each of the training data is represented as a mixture of five randomly located sound sources. The elevation and azimuth were chosen randomly in the interval $\theta \in (0, \pi/2)$ and $\phi \in (0, \pi)$ respectively.

4.2 Analysis of Reconstructed Sound Field

A monochromatic source of frequency 2kHz, with a reproduction area of $0.64m^2$ centered around the receiver is considered. Spatially decoded signals are obtained using the conventional HOA decoder for the arrangement of loudspeaker as described in section 4.3. The sound density plots are shown in Figure 3 which illustrates improvement in spatial resolution as the ambisonics order increases. Average Error Distribution (AED) were calculated for the reproduced sound field using the proposed SMS-DNN by

¹ http://home.iitk.ac.in/~groutray/upscale_demo.html

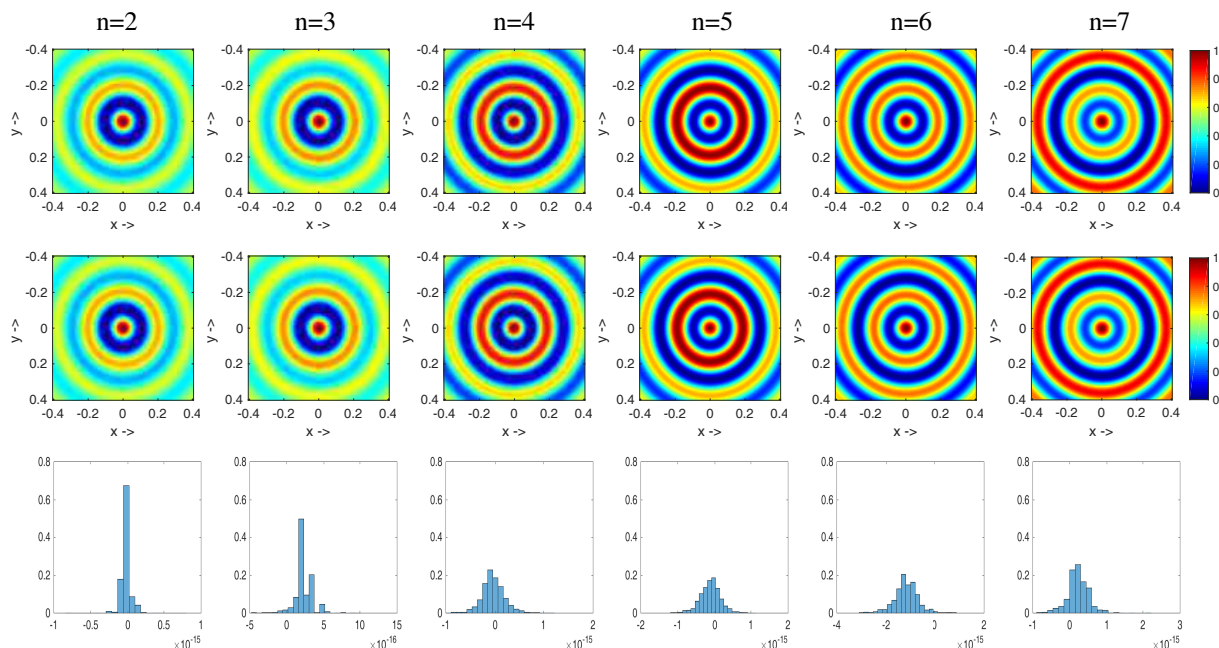


Figure 3. Sound pressure plots for order $n = 2$ to 7 (left - right) for a frequency 2kHz. Row-1 and row-2 represent the reference and SMS-DNN upscaled sound fields respectively. Row-3 illustrates Average Error Distribution for the reconstructed sound fields for order $n = 2$ to 7 .

varying the order of n , from 2 to 7, as shown in 3rd row of Figure 3. From the Figure 3, it can be observed that the error of reconstruction reduces as the order of ambisonic is increased. Figure 3 shows significant improvement in the area of error free reproduction as the order increases.

4.3 Subjective and Objective Evaluations

For the analysis of spatial sound reproduction quality, a conventional HOA decoding procedure is adopted. For this 12, 24, 32, 42, 52, and 64 number of loudspeakers are used for orders from 2 to 7 respectively. The loudspeaker positions are found using the spherical t-design structure [16], such that the spherical harmonic matrix is well conditioned. Both subjective and objective evaluations were conducted for all the three sound scenes created earlier.

4.3.1 Subjective Evaluation

MULTI Stimulus test with Hidden Reference and Anchor (MUSHRA) [17] test is conducted for perceptual evaluation. First order Ambisonics is used as the anchor. The reproduced signal using the VisiSonics [18] spherical microphone array is used as the reference signal for the test. Fifteen participants were asked to rate the three scenes in a progressive scale of 0 for bad to 5 for excellent. The results of the MUSHRA test are shown in Figure 4. From Figure 4 it is observed that as the order of ambisonics increases the mean opinion scores for these three sound scenes also increases. The scores for the order 5 to 7 are clearly indicate an improvement in perceptual quality in comparison to the lower orders. The improvement is due to the fact that the area of perceptual reception of spatial audio increases as order increases. Hence it can be anonymously conveyed

that as order increases, the quality perception of the spatial audio also increases.

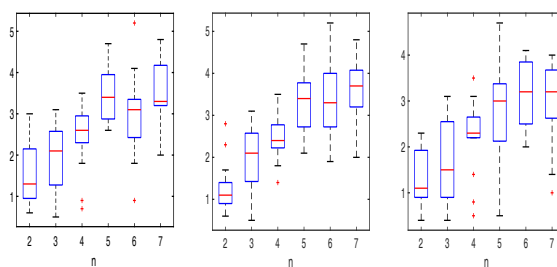


Figure 4. Mean perception score for various scenes (a) Cocktail (b) Orchestra1, and (c) Orchestra2.

4.3.2 Objective Evaluation

For objective evaluation, the methods as in [19, 20] are adopted. The perceptual quality of the sound was measured in terms of Perceptual Evaluation of Audio Quality (PEAQ) and Perceptual Similarity measure (PSM). The results for these tests are listed in Table 1. PEAQ is an objective measure in a scale 0 (for excellent) to -5 (for annoying). The measure Distortion index (DI) that compares the observed audio with the reference audio to find the distortion between the two audios. The absolute value closer to unity represents reduced distortion in the reproduced audio. PSM measures the similarity between the observed and reference audio. PSMt represents the fifth percentile of the sequence of instantaneous audio quality. In both the situation unity represents the perfect matching between the observed audio and reference audio. From the Table-1, it is observed that at higher orders, especially $n=7$

for PEAQ and PSMt show performance score decreasing towards annoying, while subjective results show increasing performance as order is increasing. The objective performance tends to annoying as order increases due to the propagation of error. But the area of error free reproduction increases as the order increases. Hence the perceptual quality improves as the spatial resolution is improved due to upscaling.

Table 1. Objective evaluation scores for various sound scenes

N	Sound Scene	PEAQ	DI	PSM	PSMt
3	Cocktail	-1.8840	-	0.9630	0.8428
	Orchestra1	-2.5177	-0.1146	0.9577	0.7566
	Orchestra2	-2.1330	-	0.9781	0.8145
4	Cocktail	-2.5968	-1.0520	0.9312	0.7316
	Orchestra1	-3.0070	-1.3530	0.9232	0.6274
	Orchestra2	-2.7488	-1.0696	0.9559	0.6937
5	Cocktail	-2.9960	-1.9506	0.9002	0.6390
	Orchestra1	-3.2704	-2.0800	0.8923	0.5209
	Orchestra2	-3.0970	-1.9139	0.9317	0.5913
6	Cocktail	-3.2382	-2.5109	0.8674	0.5674
	Orchestra1	-3.4343	-2.5387	0.8643	0.4448
	Orchestra2	-3.3090	-2.4445	0.0982	0.5087
7	Cocktail	-3.3956	-2.8780	0.8428	0.5103
	Orchestra1	-3.5407	-2.8367	0.8440	0.3939
	Orchestra2	-3.4480	-2.8030	0.8889	0.4458

5. CONCLUSION

In this work a sequential multi stage DNN is proposed and developed for upscaling ambisonics signals. HOA encoded signal obtained from the trained SMS-DNN is compared with the reference HOA signal for evaluation. Analysis of sound field shows that the proposed upscaling technique improves the spatial resolution and reduces the error variance as observed from AED at varying higher ambisonic orders. This work can be extended to provide various perceptual quality improvements by infusing novelty into the training methodology followed. The extension of this approach to model based and parametric methods of spatial audio reproduction is currently been investigated.

6. ACKNOWLEDGEMENT

This work was funded by the SERB-DST under project no. SERB/EE/2017242.

7. REFERENCES

- [1] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI, France, 2000.
- [2] J. Daniel and S. Moreau, "Further study of sound field coding with higher order ambisonics," in *Audio Engineering Society Convention 116*, Audio Engineering Society, 2004.
- [3] A. Wabnitz, N. Epain, A. van Schaik, and C. Jin, "Time domain reconstruction of spatial sound fields using compressed sensing," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 465–468, IEEE, 2011.
- [4] D. Excell, "Reproduction of a 3d sound field using an array of loudspeakers," Master's thesis, 2003.
- [5] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.
- [6] S. Moreau, J. Daniel, and S. Bertet, "3d sound field recording with higher order ambisonics—objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, pp. 20–23, 2006.
- [7] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 1–4, IEEE, 2011.
- [8] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 385–388, March 2012.
- [9] "The Eigenmike Microphone Array [online]." Available: <http://www.mhacoustics.com>.
- [10] "The Visisonics Microphone Array [online]." Available: <https://visisonics.com>.
- [11] P. G. Craven and M. A. Gerzon, "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," Aug. 16 1977. US Patent 4,042,779.
- [12] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
- [13] B. Zhang, L. Xie, Y. Yuan, H. Ming, D. Huang, and M. Song, "Deep neural network derived bottleneck features for accurate audio classification," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*, pp. 1–6, IEEE, 2016.
- [14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *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.

- [15] Y. Bengio *et al.*, “Learning deep architectures for ai,” *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [16] N. Sloane, R. Hardin, and P. Cara, “Spherical designs in four dimensions,” in *Information Theory Workshop, 2003. Proceedings. 2003 IEEE*, pp. 253–258, IEEE, 2003.
- [17] I. Recommendation, “1534-1: Method for the subjective assessment of intermediate quality level of coding systems,” *International Telecommunication Union*, 2003.
- [18] C. B. Barley and J. B. Roach, “Surveillance camera with rapid shutter activation,” Dec. 27 2012. US Patent App. 13/169,818.
- [19] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, “Peaq-the itu standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.
- [20] R. Huber and B. Kollmeier, “Pemo-qa new method for objective audio quality assessment using a model of auditory perception,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.