BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Jesper Haahr Christensen
Technical University of Denmark
jehchr@elekro.dtu.dk

Sascha Hornauer
UC Berkeley / ICSI
saschaho@icsi.berkeley.edu

Stella Yu
UC Berkeley / ICSI
stellayu@berkeley.edu

1. Introduction

We present a method for learning associations between binaural sound signals and visual scenes. Our task is to train a machine learning system that can turn binaural sound signals to 1) 3D depth maps and 2) grayscale images of plausible layout of the scene ahead.

Solving this task can benefit robot navigation and machine vision with complementary information or enable a new sensor modality in no-light conditions.

Our inspiration for this work comes from nature, where bats, dolphins and whales utilize acoustic information heavily. They adapted to environments where light is sparse. Bats have evolved advanced ears (pinnae) that provides vision in the dark known as *echolocation*: They sense the world by continuously emitting ultrasonic pulses and process echos returned from the environment and prey. Likewise, humans suffering from vision loss have shown to develop capabilities of echolocation using palatal clicks similar to dolphins, learning to sense obstacles in the 3D space by listening to the returning echoes [6, 10].

Trying to harness sound for artificial systems, previous work also mimics parts of biological systems. By using an artificial pinnae pair of bats, highly reflecting ultrasonic tar-

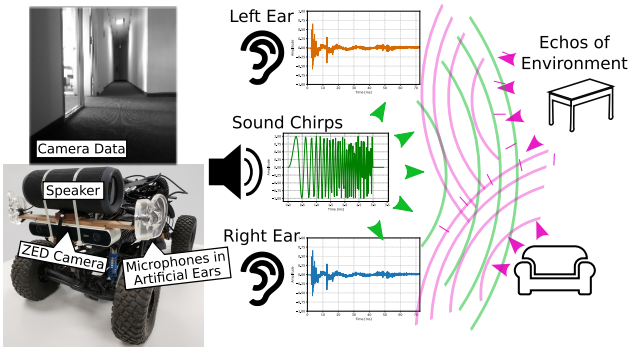

Figure 1. The *BatVision* [1] system learns to generate visual scenes by just listening to echos with two ears. Mounted on a model car, the system has two microphones embedded into artificial human ears, a speaker, and a stereo camera which is *only used during training* for providing visual image ground-truth. The speaker emits sound chirps in an office space and the microphones receive echos returned from the environment. The camera captures stereo image pairs, based on which depth maps can be calculated.

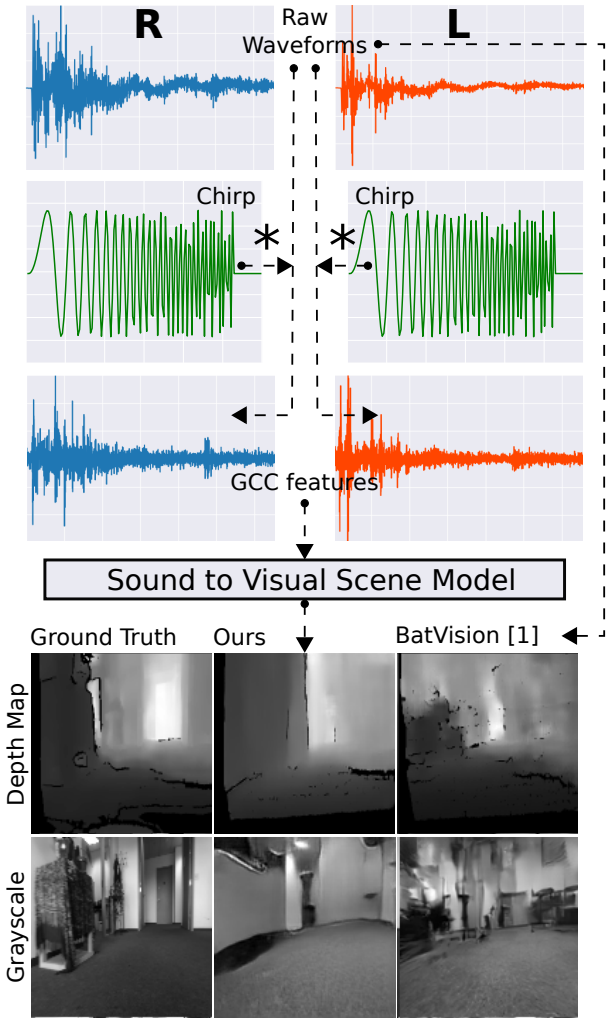

Figure 2. Generalized Cross-Correlation Features (GCC) contribute largely to improved reconstruction performance. Each left and right waveform is independently correlated ($*$) with the sent chirp signal (see equation 1). Using only these features as input in our network shows clearer depth and grayscale images with less artifacts and a more plausible room layout.

Figure 3. Overview of our dataset by [1]. Training and validation data is collected in separate areas of the same floor, whereas the test data comes from another floor and has different obstacles and decorations.

gets in the 3D space were located. The ears act as complex direction-dependent spectral filters and head-related transfer functions have been modelled to better mimic how a particular ear (left or right) receives sound from a point in space [9, 7].

We investigate how to visualize the full 3D layout ahead only from binaural echos, recorded from microphones in artificial ears. Sound chirps are played from a speaker into the environment which we also record with a stereo camera. With the time-paired data of generated depth-images and echos, we train a network to predict the former from the latter. As a proof of concept we also predict monocular grayscale images with the objective of generating plausible layout of free space and obstacles. We show an overview of our proposed system in Fig. 1.

Our contribution is an enhanced sound-to-visual system using generalized cross-correlation (GCC) features which we compare to raw waveforms and spectrograms as input encoding (cmp. Fig. 2). We further show the advantage of Residual-in-Residual Dense Blocks [11] for the generator in our architecture. We also introduce spectral normalization [8] to the PatchGAN [4] discriminator to replace batch normalization and empirically observe a more stabilized training process.

2. Audio-Visual Dataset

We use the same dataset as in [1], containing time-synchronized binaural audio, RGB images and depth maps for learning associations between sound and vision. The data has been collected using off-the-shelf, low-cost hardware fitted to a small model car, as shown in Fig. 1. Training, Validation and Test data was collected at different locations of an indoor office with hallways, conference rooms, offices and open areas. The data and collection locations are shown in Fig. 3. We refer to [1] for more details on signal generation, data collection, hardware and preparation.

GCC Features. We calculate generalized cross-correlation features for pairs of one input channel (left or right ear) respectively and our chirp source waveform:

$$G_l(f) = \frac{X_l(f)S(f)}{|X_l(f)S(f)|}; \qquad G_r(f) = \frac{X_r(f)S(f)}{|X_r(f)S(f)|}; \tag{1}$$

where $X_l(f)$ and $X_r(f)$ are our left and right waveform represented in the frequency domain, $S(f)$ is our chirp source (described in [1]) in the frequency domain padded to the same length as $X_l(f)$ (* denotes the complex conjugate). Transformations between the original time-domain and the frequency-domain are obtained by applying the Fourier transformation. The time-domain generalized cross-correlation values are then obtained by applying the inverse Fourier transformation on $G_l(f)$ and $G_r(f)$. This is currently a pre-processing step carried out using the gccphat tool in MATLAB. In Fig. 2 we show a paired raw waveform sample and its corresponding GCC feature values. The time-series cross-correlation values are then fed to the network, concatenated along the channel dimension.

3. Proposed Method

Network Architecture. As shown in our network architecture overview in Fig. 4, we keep the high-level design of BatVision. We suggest the following modifications for improving the model and obtaining a more stable training process. First, we modify the input to the audio encoder to generalized cross-correlation features rather than raw waveforms or spectrograms of binaural audio signals. Second, we re-model the generator and base it on residual learning using Residual-in-Residual Dense Blocks [11]. Third, we replace batch normalization in the discriminator with spectral normalization [8] and propose a suitable weight factor for the adversarial loss. With these modifications, we observe improved reconstruction results with less artifacts and a more stable training process than in the original model. Please see [1] for details on the original architecture.

The full learning objective of our model is:

$$\min_G \max_D \; L_{GAN}(D) + L_{GAN}(G) + L_{L_1}(G), \tag{2}$$

L_{GAN} is a least-squares adversarial loss, L_1 a L_1 regression loss and λ a weight factor.

Evaluation Metrics. For evaluating our predicted depth maps we use a common evaluation method for depth measurements as proposed in [2]. It consist of five evaluation indicators:

$$\text{Abs Rel} = \frac{1}{|N|}\sum_{i \in N} \frac{|d_i - d_i^*|}{d_i^*},$$

$$\text{Sq Rel} = \frac{1}{|N|}\sum_{i \in N} \frac{\|d_i - d_i^*\|^2}{d_i^*},$$

Table 1. Depth results on our test dataset.

	Lower is better				Accuracy: higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE Log	$< 1.25^1$	$< 1.25^2$	$< 1.25^3$
BatVision [1] + Waveforms	1.670	0.488	0.259	3.118	0.249	0.359	0.484
BatVision [1] + Spectrograms	1.544	0.398	0.241	3.177	0.256	0.369	0.521
BatVision [1] + GCC	1.782	0.464	0.252	3.231	0.236	0.330	0.454
Ours + Waveforms	1.839	0.472	0.245	3.253	0.252	0.357	0.471
Ours + GCC	1.542	0.454	0.235	3.168	0.290	0.424	0.556

Figure 4. Our sound to vision network architecture. The temporal convolutional audio encoder A turns the binaural input into a latent audio feature vector, based on which the visual generator G predicts the scene depth map. The discriminator D compares the prediction with the ground-truth and enforces high-frequency structure reconstruction at the patch level.

$$\text{RMSE} = \sqrt{\frac{1}{|N|}\sum_{i \in N} \|d_i - d_i^*\|^2},$$

$$\text{RMSE Log} = \sqrt{\frac{1}{|N|}\sum_{i \in N} \|\log(d_i) - \log(d_i^*)\|^2},$$

$$\text{Accuracies: \% of } d_i \text{ s.t. } \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = < thr,$$

where $|N|$ is the total number of pixel with real-depth values, d_i is the predicted depth value of pixel i and d_i^* is the ground truth depth value. Finally, thr denotes a threshold.

During training we check generated samples constantly in addition to the L_1 regression loss of our generator because better perceptual quality of the predictions does not always correspond to lower L_1. This is especially true when training in the GAN framework.

4. Experiments

We perform all experiments using the same model and hyper-parameters for depth map and grayscale prediction. Also, while training with waveforms or GCC-features we apply the same two input augmentations. 1) We select a window of constant size from the input which start position is randomized in time by 30%. The center start position is chosen so that always one complete chirp and its echos is captured by the window. We follow the design choices on the length of the window as in [1] 2) We add Gaussian noise $X \sim N(\mu = 0, \sigma^2 = [0, 0.1])$ to the signals. The ground truth and predicted output of the model have a spatial size of 128×128. For our generator, we use a total of 8 Residual-in-Residual Blocks in the low-resolution domain. We follow with a set of up-sampling and convolutional layers until the output resolution is reached. Up-sampling of feature maps is by nearest-neighbor interpolation. We have chosen a batch-size of 16 per GPU, a weight factor of $1 \cdot 10^{-1}$ and a learning rate of $1 \cdot 10^{-4}$ for both the generator and discriminator. For optimization, we use Adam [5] with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We alternately update the discriminator and generator until the model generates accurate and visually pleasing results (approx. 100 iterations). We implement our model in the PyTorch framework and train using NVIDIA RTX 2080 TI GPUs.

For a complete comparison we also evaluate the original BatVision network architecture with our proposed GCC-features and our proposed architecture with raw waveforms as input. Using the depth evaluation metrics, we compare depth prediction on the test set of all GAN model combinations and show the results in Table 1. A comparison of all trained models and input types is given in Table 2.

We observe that our model improves the original work of BatVision in nearly all metrics and generates qualitatively more accurate and less noisy predictions. In Fig. 5, we present examples generated by our best model as indicated by Table 2 and 1. Note that it is not possible to predict the exact grayscale image of a scene because not all information about appearance can be transported by sound. Rather the goal is to reconstruct an image which shows plausible layout in terms of free space and obstacles. Finally, when training both approaches we empirically find that our proposed method is more stable during training and less affected by small changes in hyper-parameters compared to BatVision.

5. Conclusions

We evaluated generalized cross-correlations features over raw waveforms as input modality and novel model con-

Table 2. L_1 loss on the test set for depth map and grayscale generation for different network configurations.

Arch. + Input	L_1 Loss	
Depth Map	*Gen. Only*	*GAN*
BatVision [1] + Waveforms	0.0880	0.0930
BatVision [1] + Spectrograms	0.0742	0.0878
BatVision [1] + GCC	0.0678	0.0758
Ours + Waveforms	0.0698	0.0773
Ours + GCC	**0.0645**	**0.0732**
Grayscale	*GAN*	
BatVision [1] + Waveforms	0.2018	
BatVision [1] + Spectrograms	0.1841	
Ours + GCC	**0.1770**	

Figure 5. *Test sample reconstructions.* Columns 1 and 4 show the ground truth depth map and grayscale scene image. The remaining columns show predictions from result from our method and BatVision [1]. Overall, our generations show correct mapping of close and distant areas with less noise and more smooth reconstruction than the BatVision method.

figurations for BatVision [1]. With *Residual-in-Residual* Dense Blocks in the generator and spectral normalization in the discriminator we achieve major quantitative and qualitative improvements. Apart from better scores on the evaluation metric, reconstructed depth and grayscale images show significantly better perceptual quality. The results in this work show as proof-of-concept the potential information, contained in sound. Complementary to vision we argue it can be useful in many tasks, either as exclusive or additional sensor input or to guide machine learning, as recently well presented in concurrent work [3].

References

[1] Jesper Haahr Christensen, Sascha Hornauer, and Stella Yu. Batvision: Learning to see 3d spatial layout with two ears, 2019.

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2366–2374. Curran Associates, Inc., 2014.

[3] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616*, 2020.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. pages 5967–5976, 07 2017.

[5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[6] R Kuc and Victor Kuc. Modeling human echolocation of near-range targets with an audible sonar. *The Journal of the Acoustical Society of America*, 139:581–587, 02 2016.

[7] Ikuo Matsuo, Junji Tani, and Masafumi Yano. A model of echolocation of multiple targets in 3d space from a single emission. *The Journal of the Acoustical Society of America*, 110(1):607–624, 2001.

[8] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018.

[9] Filips Schillebeeckx, Fons De Mey, Dieter Vanderelst, and Herbert Peremans. Biomimetic sonar: Binaural 3d localization using artificial bat pinnae. *I. J. Robotic Res.*, 30:975–987, 07 2011.

[10] Jascha Sohl-Dickstein, Santani Teng, Benjamin Gaub, Chris C. Rodgers, Crystal Li, Michael R. DeWeese, and Nicol S. Harper. A device for human ultrasonic echolocation. *IEEE transactions on bio-medical engineering*, 62, 01 2015.

[11] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: enhanced super-resolution generative adversarial networks. *CoRR*, abs/1809.00219, 2018.