SOUND SOURCE LOCALIZATION IN A MULTIPATH ENVIRONMENT USING CONVOLUTIONAL NEURAL NETWORKS

Eric L. Ferguson[*], *Stefan B. Williams*

Australian Centre for Field Robotics
The University of Sydney, Australia

Craig T. Jin

Computing and Audio Research Laboratory
The University of Sydney, Australia

ABSTRACT

The propagation of sound in a shallow water environment is characterized by boundary reflections from the sea surface and sea floor. These reflections result in multiple (indirect) sound propagation paths, which can degrade the performance of passive sound source localization methods. This paper proposes the use of convolutional neural networks (CNNs) for the localization of sources of broadband acoustic radiated noise (such as motor vessels) in shallow water multipath environments. It is shown that CNNs operating on cepstrogram and generalized cross-correlogram inputs are able to estimate more reliably the instantaneous range and bearing of transiting motor vessels when the source localization performance of conventional passive ranging methods is degraded. The ensuing improvement in source localization performance is demonstrated using real data collected during an at-sea experiment.

***Index Terms*—** source localization, DOA estimation, convolutional neural networks, passive sonar, reverberation

1. INTRODUCTION

Sound source localization plays an important role in array signal processing with wide applications in communication, sonar and robotics systems [1]. It is a focal topic in the scientific literature on acoustic array signal processing with a continuing challenge being acoustic source localization in the presence of interfering multipath arrivals [2, 3, 4]. In practice, conventional passive narrowband sonar array methods involve frequency-domain beamforming of the outputs of hydrophone elements in a receiving array to detect weak signals, resolve closely-spaced sources, and estimate the direction of a sound source. Typically, 10-100 sensors form a linear array with a uniform interelement spacing of half a wavelength at the array's design frequency. However, this narrowband approach has application over a limited band of frequencies. The upper limit is set by the design frequency, above which grating lobes form due to spatial aliasing, leading to ambiguous source directions. The lower limit is set one octave below the design frequency because at lower frequencies the directivity of the array is much reduced as the beamwidths broaden.

An alternative approach to sound source localization is to measure the time difference of arrival (TDOA) of the signal at an array of spatially distributed receivers [5, 6, 7, 8], allowing the instantaneous position of the source to be estimated. The accuracy of the source position estimates is found to be sensitive to any uncertainty in the sensor positions [9]. Furthermore, reverberation has an adverse effect on time delay estimation, which negatively impacts sound source localization [10]. In a model-based approach to broadband source localization in reverberant environments, a model of the so-called early reflections (multipaths) is used to subtract the reverberation component from the signals. This decreases the bias in the source localization estimates [11].

Using a single sensor, the instantaneous range of a broadband signal source is estimated using the cepstrum method [12]. This method exploits the interaction of the direct path and multipath arrivals, which is observed in the spectrogram of the sensor output as a Lloyds mirror interference pattern [12]. Generalized cross-correlation (GCC) is used to measure the TDOA of a broadband signal at a pair of sensors which enables estimations of the source bearing. Furthermore, adding another sensor so that all three sensor positions are collinear enables the source range to be estimated using the two TDOA measurements from the two adjacent sensor pairs. The range estimate corresponds to the radius of curvature of the spherical wavefront as it traverses the receiver array. This latter method is commonly referred to as passive ranging by wavefront curvature [13]. However, its source localization performance can become problematic in multipath environments when there is a large number of extraneous peaks in the GCC function attributed to the presence of multipaths, and when the direct path and multipath arrivals are unresolvable (resulting in TDOA estimation bias). Also, its performance degrades as the signal source direction moves away from the array's broadside direction and completely fails at endfire. Note that this is not the case with the cepstrum method with its omnidirectional ranging performance being independent of source direction.

Recently, Deep Neural Networks (DNN) based on supervised learning methods have been applied to acoustic tasks such as speech recognition [14, 15], terrain classification [16], and source localization tasks [17]. A challenge for supervised learning methods for source localization is their ability to adapt to acoustic conditions that are different from the training conditions. The acoustic characteristics of a shallow water environment are non-stationary with high levels of clutter, background noise, and multiple propagation paths making it a difficult environment for DNN methods.

A CNN is proposed that uses generalized cross-correlation (GCC) and cepstral feature maps as inputs to estimate both the range and bearing of an acoustic source passively in a shallow water environment. The CNN method has an inherent advantage since it considers all GCC and cepstral values that are physically significant when estimating the source position. Other approaches involving time delay estimation typically consider only a single value (a peak) in the GCC or cepstogram. The approach adopted in this paper uses a minimum number of sensors (no more than three) to localize the source. The CNNs are trained using real, multi-channel acoustic recordings of a surface vessel underway in a shallow water environ-

[*]Work supported by Defence Science and Technology Group Australia.

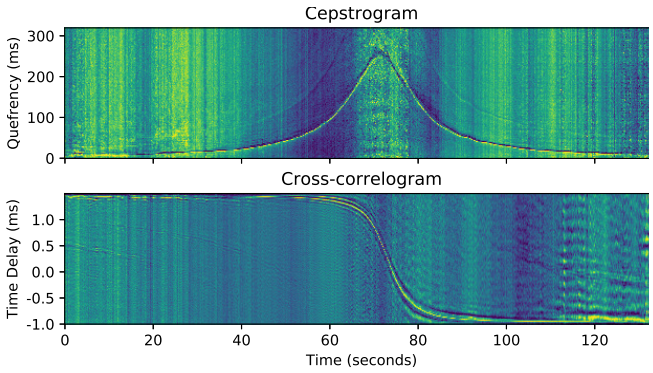

Fig. 1. **a)** Cepstrogram for a surface vessel as it transits over a single recording hydrophone located 1 m above the sea floor, and **b)** the corresponding cross-correlogram for a pair of hydrophones.

ment. CNNs operating on cepstrum or GCC feature map inputs only are also considered and their performances compared. The proposed model is shown to localize sources more reliably than conventional passive sonar localization methods that use TDOA measurements. Generalization performance of the networks is tested by ranging another vessel with different radiated noise characteristics.

The original contributions of this work are:

- Development of a multi-task CNN for the passive localization of acoustic broadband noise sources in a shallow water environment where the range and bearing of the source are estimated jointly with improved performance over traditional methods;
- Range and bearing estimates are continuous, allowing for improved resolution in position estimates when compared to other passive localization networks which use a discretized classification approach [17, 18]; and
- A novel loss function based on localization performance in physical space, where bearing estimates are constrained for additional network regularization when training;

2. ACOUSTIC LOCALIZATION CNN

A neural network is a machine learning technique that maps the input data to a label or continuous value through a multi-layer non-linear architecture. Neural networks have been successfully applied to image and object classification [19, 20], hyperspectral pixel-wise classification [21] and terrain classification using acoustic sensors [16]. CNNs learn and apply sets of filters that span small regions of the input data, enabling them to learn local correlations.

2.1. Architecture

Since the presence of a broadband acoustic source is readily observed in a cross-correlogram and cepstrogram, Fig. 1, it is possible to create a unified network for estimating the position of a vessel relative to a receiving hydrophone array. The network is divided into sections, Fig 2. The GCC CNN and cepstral CNN operate in parallel and serve as feature extraction networks for the GCC and cepstral feature map inputs respectively. Next, the outputs of the GCC CNN and cepstral CNN are concatenated and used as inputs for the dense layers, which outputs a range and bearing estimate.

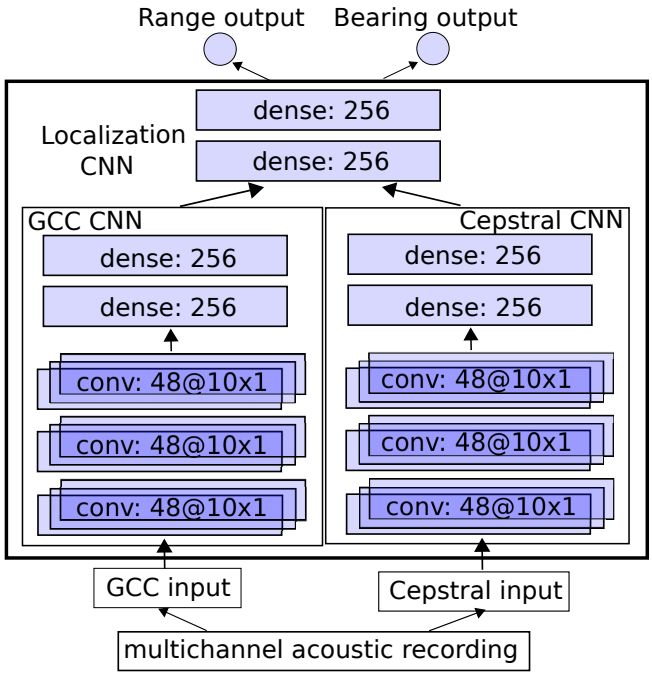

Fig. 2. Network architecture for the acoustic localization CNN

For both the GCC CNN and cepstral CNN, the first convolutional layer filters the input feature maps with 48 kernels of size $10 \times 1 \times 1$. The second convolutional layer takes the output of the first convolutional layer as input and filters it with 48 kernels of size $10 \times 1 \times 48$. The third layer also uses 48 kernels of size $10 \times 1 \times 48$, and is followed by two fully-connected layers. The combined CNN further contains two fully-connected layers that take the concatenated output vectors from both of the GCC and cepstral CNNs as input. All the fully-connected layers have 256 neurons each. A single neuron is used for regression output for the range and bearing outputs respectively. All layers use rectified linear units as activation functions. Since resolution is important for the accurate ranging of an acoustic source, max pooling (or any other down-sampling between network layers) is not used in the network's architecture.

2.1.1. Input

To localize a source using a hydrophone array, information about the time delay between signal propagation paths is required. Although such information is contained in the raw signals, it is beneficial to represent the time delay information more directly so that it can be readily learned by the network. Here, the cepstral feature map contains the mulitpath time delay information between the direct path arrival and a multipath arrival at a single sensor. The GCC feature map contains the time delay information of the direct path arrivals (together with the multipath arrival) at a pair of sensors.

A cepstrum can be derived from various spectra such as the complex or differential spectrum. For the current approach, the power cepstrum is used and is derived from the power spectrum of a recorded signal. It is closely related to the Mel-frequency cepstrum used in automatic speech recognition tasks [14, 15], but has linearly spaced frequency bands rather than bands approximating the human auditory system's response. The cepstral representation of the signal is neither in the time nor frequency domain, but rather, it is in the

quefrency domain [22]. Cepstral analysis is based on the principle that the logarithm of the power spectrum for a signal containing echoes has an additive periodic component due to the echoes from multi-path reflections [23]. Where the original time waveform contains an echo the cepstrum will contain a peak and thus the TDOA between propagation paths of an acoustic signal can be measured by examining peaks in the cepstrum [24]. The cepstrum has application when strong multipath reflections are present, which can degrade the performance of GCC methods [25]. The cepstrum $\hat{x}(n)$ is given by the inverse Fourier transform of the logarithm of the power spectrum:

$$\hat{x}(n) = F^{-1}\big(log|S(f)|^2\big), \tag{1}$$

where $S(f)$ is the Fourier transform of a discrete time signal $x(n)$.

For a given source-sensor geometry, there is a bounded range of quefrencies useful in source localization. As the source-sensor separation distance decreases to a minimum, the multipath time delay values (position of peaks in the cepstrum) will tend to a maximum value, which occurs when the source is at the closest point of approach to the sensor. TDOA values greater than this maximum are not physically realizable and are excluded. Cepstral values near zero are dominated by extraneous quefrencies and are also excluded.

GCC is used to measure the TDOA of a signal at a pair of hydrophones and is useful in situations of spatially uncorrelated noise [26]. For a given array geometry, there is a bounded range on useful GCC information. For a pair of recording sensors, a zero relative time delay corresponds to a broadside source, whilst a maximum relative time delay corresponds to an endfire source. TDOA values greater than the maximum bound have no physical significance and are excluded [27, 12]. The windowing of CNN inputs has the added benefit of reducing the number of parameters in the network. A cepstrogram and cross-correlogram for a surface vessel transit is shown in Fig. 1.

2.1.2. Output

For each example, the network predicts the range and bearing of the acoustic source as a continuous value (each with a single neuron regression output). This differs from other recent passive localization networks which use a classification-based approach where range and bearing predictions are discretized, putting a hard limit on the resolution of estimations that the networks are able to provide [17, 18].

2.2. Multi-task Joint Training

The objective of the network is to predict the range and bearing of an acoustic source relative to a receiving array from reverberant and noisy multi-channel input signals. Since the localization of an acoustic source involves both a range and bearing estimate, both the range and bearing output loss components are jointly minimized using a loss function based on localization performance in physical space. This additional regularization is expected to improve localization performance when compared with minimizing range loss and bearing loss separately.

The total objective function E minimized during network training is given by the weighted sum of the polar-distance loss E_p and the bearing loss E_b, such that:

$$E = \alpha E_p + (1-\alpha)E_b, \tag{2}$$

where E_p is the L_2 norm of the polar distance given by:

$$E_p = y^2 + t^2 - 2yt\cos(\theta - \phi) \tag{3}$$

and E_b is the L_2 norm of the bearing loss only, given by:

$$E_b = (\theta - \phi)^2 \tag{4}$$

with the predicted range and bearing output denoted as t and ϕ respectively, and the true range and bearing denoted as y and θ respectively. The inclusion of the E_b term encourages bearing predictions to be constrained to the first turn, providing additional regularization and reducing parameter weight magnitudes. The two terms are weighted by hyper-parameter α so each loss term has roughly equal weight. Training uses batch normalization [28] and is stopped when the validation error does not decrease appreciably per epoch. In order to further prevent over-fitting, regularization through a dropout rate of 50% is used in all fully connected layers when training [29].

3. EXPERIMENTAL RESULTS

Passive localization on a transiting vessel was conducted using a multi-sensor algorithmic method described in [30], and CNNs with cepstral and/or GCC inputs. Their performances were then compared. The generalization ability of the networks to other broadband sources is also demonstrated by localizing an additional vessel with a different radiated noise spectrum and source level.

3.1. Dataset

Acoustic data of a motor boat transiting in a shallow water environment over a hydrophone array were recorded at a sampling rate of 250 kHz. The uniform linear array (ULA) consists of three recording hydrophones with an interelement spacing of 14 m. Recording commenced when the vessel was inbound 500 m from the sensor array. The vessel then transited over the array and recording was terminated when the vessel was 500 m outbound. The boat was equipped with a DGPS tracker, which logged its position relative to the receiving hydrophone array at 0.1 s intervals. Bearing labels were wrapped between 0 and π radians, consistent with bearing estimates available from ULAs which suffer from left-right bearing ambiguity. Twenty-three transits were recorded over a two day period. One hundred thousand training examples were randomly chosen each with a range and bearing label, with examples uniformly distributed in range only. A further 5000 labeled examples were reserved for CNN training validation. The recordings were preprocessed as outlined in Section 2.1.1, using 0.1 s of recorded multi-channel data per example. The networks were implemented in TensorFlow and were trained with a Momentum Optimizer using a NVIDIA GeForce GTX 770 GPU. The gradient descent was calculated for batches of 32 training examples. The networks were trained with a learning rate of 3×10^{-9}, exponential learning rate decay of 0.96, network parameter weight decay of 1×10^{-5} and momentum of 0.9. Additional recordings of the vessel were used to measure the performance of the methods. These recordings are referred to as the test dataset and contain 9980 labeled examples. Additional acoustic data were recorded on a different day using a different boat with different radiated noise characteristics. Acoustic recordings for each transit started when the inbound vessel was 300 m from the array, continued during its transit over the array, and ended when the outbound vessel was 300 m away. This dataset is referred to as the generalization set and contains 11714 labeled examples.

3.2. Input of Network

Cepstral and GCC feature maps were used as inputs to the CNN and they were computed as follows. For any input example, only a select

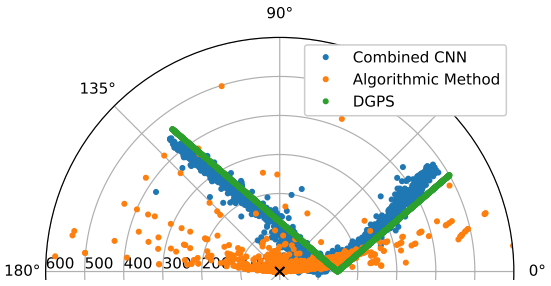

Fig. 3. Estimates of the range and bearing of a transiting vessel. The true position of the vessel is shown relative to the recording array, measured by the DGPS.

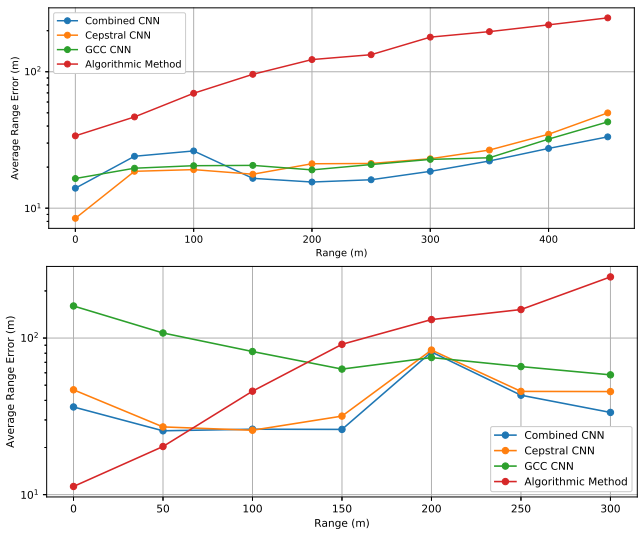

Fig. 4. Comparison of range estimation performance as a function of the vessels true range for the **a)** test dataset and **b)** generalization dataset.

range of cepstral and GCC values contain relevant TDOA information and are retained - see Section 2.1.1. Cepstral values more than 1.4 ms are discarded because they represent the maximum multipath delay and occur when the source is directly over a sensor. Cepstral values less than 84 μs are discarded since they are extraneous. Thus, each cepstrogram input is liftered and samples 31 through 351 are used as input to the network only. A cepstral feature vector is calculated using 0.1 s of audio for each recording channel, resulting in a 320 x 3 cepstal feature map. Due to array geometry, the maximum time delay between pairs of sensors is $\pm$9.2 ms. A GCC feature vector is calculated using 0.1 s of observations for two pairs of sensors, resulting in a 4800 x 2 GCC feature map. The GCC map is further sub-sampled to size 480 x 2, which reduces the number of network parameters.

3.3. Comparison of Localization Methods

Algorithmic passive localization was conducted using the methods outlined in [30]. The TDOA values required for algorithmic localization were taken from the largest peaks in the GCC. Nonphysical results at ranges greater than 1000 m are discarded. Other CNN architectures are also compared. The GCC CNN uses the GCC CNN

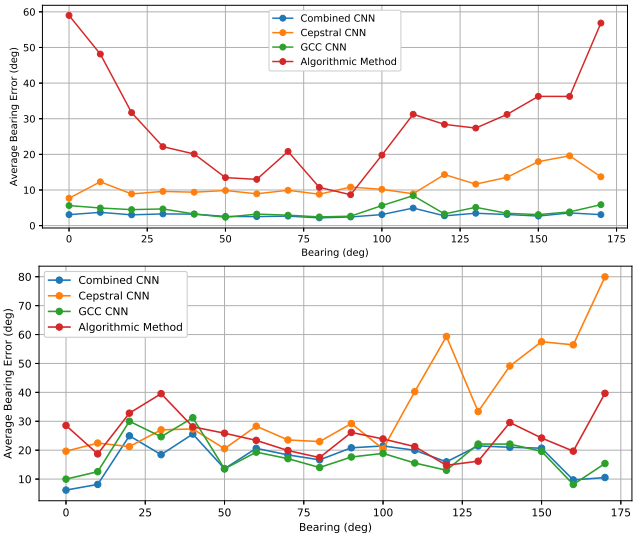

Fig. 5. Comparison of bearing estimation performance as a function of the vessels true bearing for the **a)** test dataset and **b)** generalization dataset.

section of the combined CNN only, and the Cepstral CNN uses the Cepstral CNN section of the combined CNN only, both with similar range and bearing outputs, Fig 2. Fig. 3 shows localization results for a vessel during one complete transit. Fig. 4 and Fig. 5 show the performance of localization methods as a function of the true range and bearing of the vessel for the test dataset, and the generalization set respectively. The CNNs are able to localize a different vessel in the generalization set with some impact to performance. The performance of the algorithmic method is degraded in the shallow water environment since there are a large number of extraneous peaks in the GCC attributed to the presence of multipaths, and when the direct path and multipath arrivals become unresolvable (resulting in TDOA estimation bias). Bearing estimation performance is improved in networks using GCC features, showing that time delay information between pairs of spatially distributed sensors is beneficial. The networks show improved robustness to interfering multipaths. Range estimation performance is improved in networks using cepstral features, showing that multipath information can be useful in determining the sources range. The combined CNN is shown to provide superior performance for range and bearing estimation.

4. CONCLUSIONS

In this paper we introduce the use of a CNN for the localization of surface vessels in a shallow water environment. We show that the CNN is able to jointly estimate the range and bearing of an acoustic broadband source in the presence of multipath interference. Several CNN architectures are compared and evaluated. The networks are trained and tested using cepstral and GCC feature maps as input formed from real acoustic recordings. Networks are trained using a novel loss function based on physical localization performance with additional constraining of bearing estimates. The inclusion of both cepstral and GCC inputs facilitates robust passive acoustic source localization in reverberant environments, where conventional algorithmic methods are less reliable.

5. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[2] M. Viberg, B. Ottersten, and T. Kailath, "Detection and estimation in sensor arrays using weighted subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2436–2449, 1991.

[3] X. Zeng, M. Yang, B. Chen, and Y. Jin, "Low angle direction of arrival estimation by time reversal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 3161–3165.

[4] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[5] G.C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, pp. 463–470, 1981.

[6] G.C. Carter, Ed., *Coherence and time delay estimation*, IEEE Press, New York, 1993.

[7] Y.T. Chan and K.C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. on Signal Process.*, vol. 42, pp. 1905–1915, 1994.

[8] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech and Audio Process.*, vol. 12, no. 5, pp. 509–519, 2004.

[9] E.L. Ferguson, "Application of passive ranging by wavefront curvature methods to the localization of biosonar click signals emitted by dolphins," in *Proc. of International Conf. on Underwater Acoust. Measurements*, 2011.

[10] J. Chen, J. Benesty, and Y.A. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP J. on Adv. in Signal Process.*, vol. 2005, no. 1, pp. 498964, 2005.

[11] J.R. Jensen, J.K. Nielsen, R. Heusdens, and M.G. Christensen, "DOA estimation of audio sources in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 176–180.

[12] E.L. Ferguson, R. Ramakrishnan, S.B. Williams, and C.T. Jin, "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 2657–2661.

[13] E.L. Ferguson, "A modified wavefront curvature method for the passive ranging of echolocating dolphins in the wild," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 3972–3972, 2013.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M.L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 5745–5749.

[15] J. Heymann, L. Drude, Christoph Boeddeker, Patrick Hanebrink, and R. Haeb-Umbach, "Beamnet: end-to-end training of a beamformer-supported multi-channel asr system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 5325–5329.

[16] A. Valada, L. Spinello, and W. Burgard, "Deep feature learning for acoustics-based terrain classification," in *Robotics Research*, pp. 21–37. Springer, 2018.

[17] S. Chakrabarty and E.A.P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," *arXiv preprint arXiv:1705.00919*, 2017.

[18] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 2217–2221.

[19] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. in neural information process. systems*, 2012, pp. 1097–1105.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recog.*, 2014, pp. 580–587.

[21] L. Windrim, R. Ramakrishnan, A. Melkumyan, and R. Murphy, "Hyperspectral CNN classification with limited training samples," in *British Machine Vision Conf.*, 2017.

[22] B.P. Bogert, "The quefrency analysis of time series for echoes: Cepstrum pseudo-autocovariance, cross-cepstrum, and saphe cracking," *Time Series Analysis*, pp. 209–243, 1963.

[23] K.W. Lo, B.G. Ferguson, Y. Gao, and A. Maguer, "Aircraft flight parameter estimation using acoustic multipath delays," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, no. 1, pp. 259–268, 2003.

[24] A.V. Oppenheim and R.W. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Process. Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[25] Y. Gao, M. Clark, and P. Cooper, "Time delay estimate using cepstrum analysis in a shallow littoral environment," *Conf. Undersea Defence Technology*, vol. 7, pp. 8, 2008.

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

[27] E.L. Ferguson, R. Ramakrishnan, S.B. Williams, and C.T. Jin, "Deep learning approach to passive monitoring of the underwater acoustic environment," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 3351, 2016.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conf. on Machine Learning*, 2015, pp. 448–456.

[29] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *J. Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] H.C. Schau and A.Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, 1987.