ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096422

CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR THE CLASSIFICATION OF CETACEAN BIOACOUSTIC PATTERNS

Dimitris N. Makropoulos[1,2] Antigoni Tsiami[1] Aristides Prospathopoulos[2] Dimitris Kassis[2]
Alexandros Frantzis[4] Emmanuel Skarsoulis[3] George Piperakis[3] Petros Maragos[1]

[1] National Technical University of Athens, Greece
[2] Hellenic Centre for Marine Research (HCMR), Institute of Oceanography, Anavyssos, Greece
[3] Foundation of Research and Technology – Hellas, IACM, Heraklion-Crete, Greece
[4] Pelagos Cetacean Research Institute, Vouliagmeni 16671, Greece

{d.makropoulos, aprosp, dkassis}@hcmr.gr, {antsiami, maragos}@cs.ntua.gr, {eskars,piperak}@iacm.forth.gr, afrantzis@otenet.gr

ABSTRACT

In this paper we focus on the development of a convolutional recurrent neural network (CRNN) to categorize biosignals collected in the Hellenic Trench, generated by two cetacean species, sperm whales (*Physeter macrocephalus*) and striped dolphins (*Stenella coeruleoalba*). We convert audio signals into mel-spectrograms and forward the input into a deep residual network (ResNet), designed to capture spectral patterns. Next, ResNet's output is reshaped into a time-distributed layer and fed into recurrent network variants, Long Short-Term Memory (LSTMs) or Gated Recurrent Units (GRUs), able to recognize long-term time dependencies on extracted features. The hybrid network perfectly classifies audio signals into three categories (dolphins, sperm whales, ambient noise) while it also exhibits high learning ability on recognising intraclass representations of overlapping acoustic patterns (clicks vs whistles and clicks, both emitted by dolphins). The proposed scheme outperforms traditional Machine Learning (ML) techniques, baseline ResNet and LSTM architectures or their deep parallel combinations.

Index Terms— Machine learning, residual networks, pattern recognition, bioacoustic patterns, cetacean vocalization.

1. INTRODUCTION

Classification of cetacean acoustic calls is a fundamental challenge in the study of marine mammals' bioacoustics, motivated by the need to build a reliable tool for protection of endangered species. Moreover, cetaceans' identification has been a growing area of research especially since Machine Learning (ML) algorithms -supported by growing storage capacity and computational power- emerged, accessing large training databases. Most recently, hybrid networks comprising Convolutional Neural Networks (CNNs) connected in series or in parallel with recurrent layers are designed to solve problems on computational bioacoustics [1], acoustic scene classification [2] and polyphonic sound event detection [3].

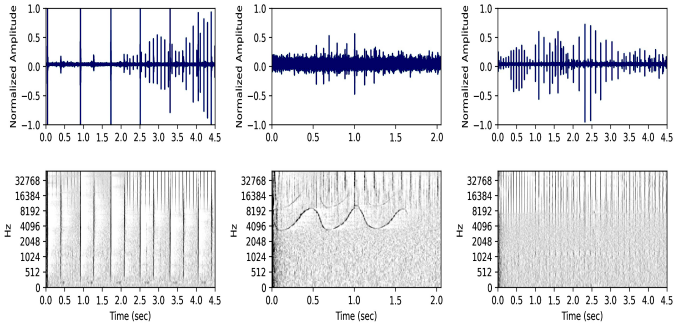


Fig. 1. Raw waveforms (up) and spectrograms (down) of sperm whale clicks (left) vs striped dolphin clicks and whistles (center) and striped dolphin clicks (right) recorded at Pylos station, Ionian Sea; sea Section 3.

CNNs architectures are inspired by studies in modern biology about how stages are developed in human vision processing. In fact, low-level vision representations such as lines or edges detection are considered to be an early stage in visual processing [4],[5] followed by a 'mid-level' grouping mechanism where virtual cortex recognizes intermediate forms such as motifs of plaid patterns and curved contours [6]. Finally, on a 'high level' representation, image components are bounded to form a semantic and coherent perception of objects and scenes in their environment [7] integrating attention mechanisms too. Hence, a typical CNN, mimicking biological systems, mainly comprises convolutional layers (CLs) functioning as feature extractors of progressively increasing abstraction level layer after layer, gradually recognizing from data,

The first author was supported by the projects DRESSAGE (MIS:5045792) and "Monitoring and recording the situation of the marine sub-regions of Greece /Upgrading and functional updating of the MSFD monitoring network" (MIS:5010880), under NSRF 2014-2020.

depth-dependent patterns critical for discrimination. Following pooling layers (PLs) perform sub-sampling, reducing output's sensitivity to distortions or shifts, providing thus invariance to translations and enhancing selectivity [8].

On the other hand Recurrent Neural Networks (RNNs) can learn temporal dynamics due to recurrent connections between layers allowing previous or future inputs to affect present outputs. Due to their ability to process sequential information of variable length they have been widely used for modeling speech recognition, time series prediction, music improvisation, image captioning, etc. Nevertheless, vanishing and exploding gradients prevent RNN from learning long-term dependencies, that is, correlations between temporally distant inputs. To overcome this issue, LSTM variant was proposed, allowing information persistence through time, utilizing as fundamental units recurrently connected blocks comprising memory cells acting as accumulators, consisting of forget, input and output gates the role of which is to decide information components that will be removed, updated and forwarded. Finally, GRU networks were developed mainly differing from LSTMs on simultaneously controlling through a single gate, forget and input gate mechanisms when updating the state unit [9].

In this paper we attempt to develop a combined Convolutional and Recurrent Neural Network (CRNN) algorithm i.e. a consecutive structure of a deep Residual Network, followed by special types of recurrent networks, such as LSTMs and GRUs to recognize cetacean bioacoustic calls collected in the Hellenic Trench across a diverse range of environmental and noise conditions. Constructing a hybrid network able to extract spectro-temporal features is justified by the fact that vocalisation calls -echolocation clicks or/and continuous whistles [Fig.1]- form specific patterns characteristic of species both on frequency (bandwidth, predominate frequencies) and time domain (call duration, inter-click intervals). Our study focuses on the development of deep learning (DL) techniques, attempting to extract appropriate feature vectors optimally capturing information from cetacean biosignals. In this framework, we have tested several candidate input spaces (spectrograms and scalograms) to evaluate methodological aspects on NN architectural design, visualizing subsequently extracted features on low dimensional space using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) techniques. Finally, we compare alternative ML and DL models in terms of performance, in contrast to most research works in cetacean bioacoustics which are specialized either on the development of sophisticated DL models or the utilization of traditional ML Networks.

2. RELATED WORK

Various approaches have been followed in the design of algorithms aiming to classify cetaceans' biosignals including both traditional ML techniques trained on extracted features and modern deep learning methods. Thus, Gaussian Mixture Models (GMMs) and Support Vector Machines (SVM) algorithms attempting to model boundaries between species' cepstral feature distribution, have been used to construct a classifier in [10] from the following species: Blainville's beaked whales *(Mesoplodon densirostris)*, short-finned pilot whales *(Globicephala macrorhynchus)*, and Risso's dolphins *(Grampus griseus)*. As demonstrated in [11], Hidden Markov Models (HMMs) treating data as sequence of separate GMM states -taking thus into account both spectral and temporal structure- can automatically identify killer whales *(Orcinus orca)* calls from a sample of known individuals. In recent studies it has also been demonstrated that deep Convolutional Neural Networks (CNN) trained on spectrograms generated from cetacean calls are capable to detect and classify sperm whale clicks [12] or humpback whale songs [13]. Moreover, Siamese Neural Networks (SNN) used to measure feature vectors' similarities in the input space were found to outperform CNNs when utilized to categorize different types of blue whale songs [14]. In addition, self-supervised learning approaches implemented by LSTM and GRU networks have been also successfully utilized to classify sperm whale coda types, categorize different vocal dialects, and identify individual whales [12].

3. MATERIALS AND METHODS

Origin and characteristics of cetacean data - Experimental Design - Data preprocessing

The data examined were provided from three different sources: i) a Passive Aquatic Listener (PAL) with a sampling frequency of 100 kHz was deployed at Pylos station of the POSEIDON buoy network in the Ionian Sea at 500 m depth from November 2008 to March 2009, approximately 10 km off the West Peloponnese coast [15]; ii) acoustic data were collected through a towed array during cetacean surveys of the Pelagos Cetacean Research Institute along the Hellenic Trench during the period 2001-2020 with a sampling frequency 48 kHz [16]; iii) acoustic recordings were carried out at frequencies up to 100 kHz in summer 2020 and 2021 from the 'SAvE Whales observatory' (System for the Avoidance of Collisions with Endangered Whales), consisting of three acoustics stations with one hydrophone each suspended at a depth of 100 m, and deployed in the Bay of Sougia, Southern Crete, about 2 km offshore and 1-2 km apart in an area of 500 m depth [17].

Sperm whale calls are mainly made up of broadband and highly directional impulsive signals named clicks, involved in echolocation and characterized by a centroid frequency of approximately 15 kHz [18], in contrast to also emitted stereotyped calls termed 'codas' [19], consisting of repetitive series of clicks of lower centroid frequency ranging of approx-

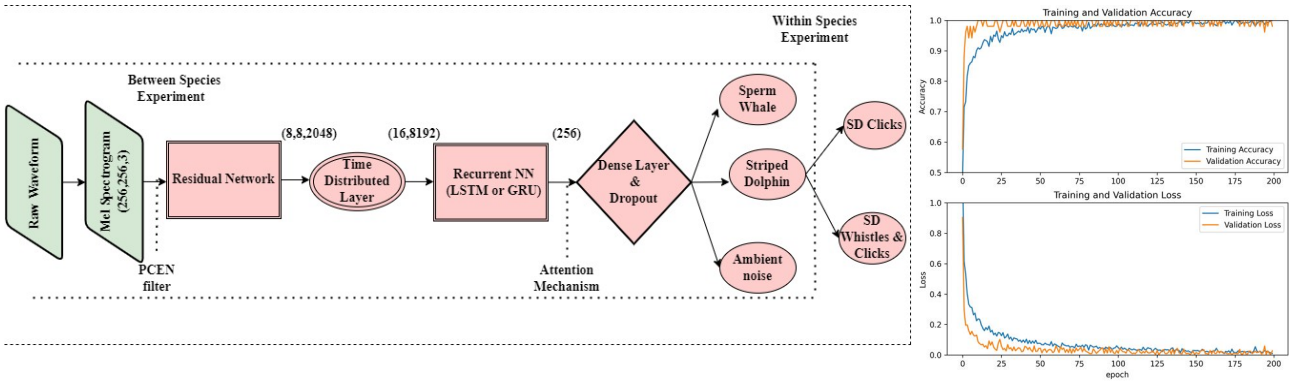


Fig. 2. General architecture of the proposed hybrid network

Fig. 3. Accuracy and Loss of a Residual-BiLSTM model on a between species -three classes- categorization problem

imately 5 kHz [20]. On the other hand, sounds generated by delphinids have in generally been classified into three distinct categories comprising echolocation clicks, burst pulse clicks or whistles. Dolphin's clicks are broadband signals of varying frequency from few tens of kHz to well over 100 kHz [21], while burst pulse clicks are closely spaced broadband click trains. Finally, whistles are continuous narrowband signals of fundamental frequency ranging from 2 to 30 kHz [21] of duration's range between 100ms and just over 4 s [22].

The dataset in this study is composed of audio signals belonging into three different groups: 291 recordings of sperm whale calls, 90 ambient-noise audio signals characterized by absence of clicks or whistles and a class consisting of 284 striped dolphin calls for the first between species experiment. Latterly a second experiment was designed dividing dolphin's class into two discrete -partially overlapping- subclasses (135 files representing clicks and 149 recordings composed of whistles and clicks). We have proceeded with this more detailed categorization increasing complexity of the input space, in order to better evaluate efficiency among various architectures, given that on -between species- experiments consisting of classifying dolphins' or sperm whales' vocalizations against ambient noise, accuracies exceeding 99% have been achieved for the majority of models.

Part of the challenge on solving bioacoustic classification problems comes from the wide range of anthropogenic and environmental noise incorporated into the audio signals. We have applied on audio files a Butterworth filter with cutoff at 1 kHz to remove low-frequency noise. Every audio clip has been then transformed into its mel-spectrogram representation applying the windowed Fourier transform, using a Hanning window with a size of 512 (corresponding to a window of 0.5ms duration) and an overlap of 50%. In contrast to normal spectrograms, mel-spectrograms have unequal spacing in the frequency bands, utilizing a logarithmic spacing above 1 kHz and linear frequency spacing below 1 kHz [23]. Next, Per Channel Energy Normalization (PCEN) has been applied to suppress stationary, narrowband electronic noise caused by the equipment itself leading to horizontal lines in mel-spectrograms and has successfully managed to enhance contrast between background and foreground transient events as in [24] such as cetacean clicks.

4. EXPERIMENTS AND DISCUSSION

Convolutional Recurrent Neural Networks - Parameter Optimization - Traditional ML techniques - Results

In this paper the proposed architecture consists of a consecutive structure of a Residual Network followed by an LSTM or GRU variant based network. In particular, we use a residual network (ResNet-101) as a baseline model, pretrained on ImageNet dataset, consisting of multiple blocks that are connected to each other in series while the added layers perform identity mapping, allowing signal's info to flow without losses through layers. We employ on different experiments several variants of RNNs such as LSTMs, GRUs or their bidirectional versions (BiLSTMs, BiGRUs) able to capture temporal dependencies in both the previous as well the next time steps. Figure 2 illustrates the proposed hybrid classification network. The first block of the network utilizes a ResNet101 architecture processing mel-spectrograms as inputs of shape 256x256x3 and produces spectral feature maps the output of which are of shape 8x8x2048. Feature maps are reshaped on a time-distributed layer in order to obtain appropriate dimensions 16x8192, and are subsequently fed into a second block comprising a recurrent network based variant with 256 units extracting temporal features which can be finally transferred to an attention mechanism layer (weighting important parts of the sequence) followed by a fully connected and a soft-max layer on the top for classification purposes.

In our experimental setup, the dataset was divided into two subsets of training and validation data with the proportions of 80%, 20% respectively. We have used K-Fold cross validation methods consisting of training on K-1 folds of data and using rest for validation purposes. The final result is cal-

Table 1. Performance of different NN architectures

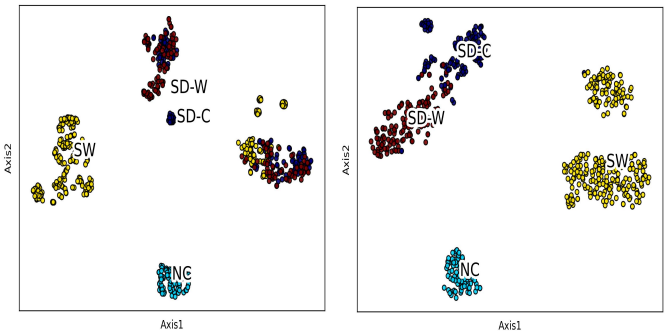
Models	Results on a test set (Mean values)		
	Parameters	***Accuracy***	***Precision***
MFCC-SVM (RBF)	-	83.0%	73.4%
MFCC-kNN	-	75.45%	73.4%
ResNet	1.0M	87.0%	84.7%
ResNet-LSTM	9.77M	91.3%	89.9%
ResNet-BiLSTM	18.5M	90.1%	89.1%
ResNet-GRU	7.6M	90.9%	89.8%
ResNet-BiGRU	14.2M	88.7%	88.0%
ResNet-LSTM-Attention	9.8M	90.4%	89.9%
Parallel ResNet-LSTM	8.2M	89.2%	88.7%

culated as the average accuracy and precision of the K experiments. For this study, we set K = 5 and the model is trained for 150 epochs with a batch size of 64. An Adam optimizer was used with a learning rate of 10^{-3} while categorical cross entropy for loss function was preferred during the optimization process. A dropout regularization of 0.4 was applied to the neural classifier to prevent from potential overfitting.

For comparison purposes we also developed a traditional ML network based on Mel Frequency Cepstral Coefficients (MFCCs), a feature vector widely used on traditional ML methods followed by an SVM or KNN classifier. MFCCs' extraction has been realized, applying on a frame-by-frame basis the Discrete Cosine Transform (DCT) on the log-Mel spectrogram of every audio clip. We have selected to retain the first 13 cepstral coefficients per frame incorporating most of the signal information, then excluded the zeroth coefficient representing the average log-energy of the signal, augmenting the feature vector by calculating their first derivatives. Next, we have calculated the average of MFCC features across all frames to build a feature vector composed of 24x1 (MFCC and ΔMFCC) coefficients for every vocalization.

Figure 3 illustrates performance evaluation of a ResNet-BiLSTM model during training and validation phase for a three groups classification problem. Table 1 shows accuracy scores for different utilized architectures on a 4-classes experiment with respect to the number of the training parameters of each model, corresponding to different computational complexity. We have repeated the analysis replacing LSTM with GRU layers comparing performances. We conclude that: (a) baseline DL models outperform traditional ML methods; (b) hybrid networks can achieve higher accuracies than baseline ResNets (91.3% vs 87.0% respectively); (c) bidirectional networks do not increase performance; (d) all architectures have succeeded to solve a -between species- classification problem while hybrid architectures have demonstrated a comparative advantage on differentiating intraclass overlapping patterns.

In terms of architectural design, the multiplicity of hyperparameters or layers to use, makes often hard to argue about the appropriate selected model in terms other than accuracy



(a) Cepstral features (b) CRNN features

Fig. 4. Implementation of PCA and t-SNE visualization of an MFCC feature vector vs an extracted deep feature vector with the four class representing: SW: Sperm whale clicks, NC: No Clicks, SD-C: Striped dolphins clicks, SD-W: Striped dolphins whistles.

metrics or computational complexity. This is why we attempt through this study to emphasize on the visualization of extracted features on low dimensional spaces as relative intuitive measure of different architectures efficiency. In the frame of the proposed CRNN architecture we remove the classification head and extract deep features, then employ PCA to reduce feature's space dimensionality from 256 to 10, followed by t-SNE to further decrease dimensionality from 10 to 2 as in [12] and plot extracted features on the Euclidean plane. Above presented Fig.4 reveals DLs approach improved capacity to shatter feature space relatively to MLs method respective results. In the frame of realized ablation studies, we have extensively utilized t-SNE visualization on features extracted from various architectures as a qualitative mechanism enhancing model selectivity before validating the architecture here mentioned.

5. CONCLUSION

In this paper, a hybrid network consisting of consecutive ResNet and RNN variants has been proposed for classification of cetacean bioacoustic calls collected in the Hellenic Trench. Each of the architectures we experiment with, solves quite successfully a -between species- categorization task confirming findings from recent studies on supervised learning ability to recognize patterns of bioacoustic events in time-frequency representations. Additionally, our study shows that networks' generalization ability is more accurately evaluated on spaces of higher complexity, where performance of different architectures deviates significantly. In fact, when dividing collection of dolphins' biosignals into partially overlapping subclasses, advantages in utilizing structures of hybrid networks versus both traditional ML techniques or baseline ResNet and LSTM networks have been observed.

6. REFERENCES

[1] Dan Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PEERJ*, 2022.

[2] Tao Zhang, Jinhua Liang, and Biyun Ding, "Acoustic scene classification using deep CNN with fine-resolution feature," *Expert Syst. Appl.*, vol. 143, 2020.

[3] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, February 2017.

[4] William McIlhagga, "Estimates of edge detection filters in human vision," *Vision Research*, vol. 153, pp. 30–36, 2018.

[5] Sarah M. Szymkowicz, Nicole R. Nissim, and Adam J. Woods, *Edge Detection*, pp. 1268–1269, Springer International Publishing, Cham, 2018.

[6] Jonathan W. Peirce, "Understanding mid-level representations in visual processing," *Journal of Vision*, vol. 15, no. 7, pp. 5, June 2015.

[7] Kiper D. and Carandini M., "Neural basis of pattern vision," *Encyclopedia of. Cognitive Science*, 2002.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[9] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[10] M. Roch, M. Soldevilla, R. Hoenigman, S. Wiggins, and J. A. Hildebrand, " Comparison of Machine Learning Techniques for the Classification of Echolocation Clicks from Three Species of Odontocetes," *Canadian Acoustics*, vol. 36(1), pp. 41–47, 2008.

[11] J.C. Brown, P. Smaragdis, and A. Nousek-McGregor, " Automatic identification of individual killer whales," *The Journal of the Acoustical Society of America*, vol. 128(3), 2010.

[12] P. Bermant, M. Bronstein, R. Wood, S. Gero, and D. Gruber, "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," *Scientific Reports*, vol. 9, pp. 1–10, Aug. 2019.

[13] Ann N. Allen, Matt Harvey, Lauren Harrell, Aren Jansen, Karlina P. Merkens, Carrie C. Wall, Julie Cattiau, and Erin M. Oleson, "A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset," *Frontiers in Marine Science*, vol. 8, 2021.

[14] M. Zhong, M. Torterotot, T.A. Branch, K.M. Stafford, J-Y. Royer, R. Dodhia, and J. Lavista Ferres, "Detecting, classifying, and counting blue whale calls with siamese neural networks," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3086–3094, 2021.

[15] J. Nystuen, M. Anagnostou, E. Anagnostou, and A. Papadopoulos, "Monitoring greek seas using passive underwater acoustics," *Journal of Atmospheric and Oceanic Technology*, vol. 32, pp. 334–349, Feb. 2015.

[16] A. Frantzis, P. Alexiadou, and K.C. Gkikopoulou, "Sperm whale occurrence, site fidelity and population structure along the hellenic trench (Greece, Mediterranean sea).," *Aquatic Conserv: Mar. Freshw. Ecosyst.,*, vol. 24, pp. 83–102, 2014.

[17] E. Skarsoulis, G. Piperakis, E. Orfanakis, P. Papadakis, D. Pavlidi, M. Kalogerakis, P. Alexiadou, and A. Frantzis, "A Real-Time Acoustic Observatory for Sperm-Whale Localization in the Eastern Mediterranean Sea," *Frontiers in Marine Science*, vol. 9, pp. 873888, May 2022.

[18] B. Møhl, M. Wahlberg, PT. Madsen, A Heerfordt, and A. Lund, "The monopulsed nature of sperm whale clicks.," *The Journal of the Acoustical Society of America*, vol. 114(2), pp. 1143–54, 2003.

[19] P. Madsen, R. Payne, N Kristiansen, M. Wahlberg, I. Kerr, and B. Møhl, "Sperm whale sound production studied with ultrasound time/depth-recording tags," *The Journal of experimental biology*, vol. 205, pp. 1899–906, Aug. 2002.

[20] Stefan Huggenberger, Michel André, and Helmut H. A. Oelschlager, "The nose of the sperm whale: overviews of functional design, structural homologies and evolution," *Journal of the Marine Biological Association of the United Kingdom*, vol. 96, pp. 783 – 806, 2014.

[21] J. Oswald, S. Rankin, J. Barlow, and M. Lammers, "A tool for real-time acoustic species identification of delphinid whistles," *The Journal of the Acoustical Society of America*, vol. 122, pp. 587–95, Aug. 2007.

[22] E. Papale, M. Azzolin, I. Cascão, A. Gannier, and Lammers et al., "Geographic variability in the acoustic parameters of striped dolphin's (*Stenella coeruleoalba)* whistles," *The Journal of the Acoustical Society of America*, vol. 133, pp. 1126–1134, Feb. 2013.

[23] K.S. Rao and A.K. Vuppala, *Speech Processing in Mobile Environments*, Springer Cham, 2014.

[24] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and JP Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.