Underwater Acoustic Target Classification with Joint Learning Framework and Data Augmentation

2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 978-1-6654-9913-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIBD55127.2022.9820117

Liang Chen
Institute of Ocean Engineering and Technology
Zhejiang University
Zhoushan, China
chl160@126.com

Feng Liu
National Innovation Institute of Defense Technology
Chinese Academy of Military Science
Beijing, China
liufeng_cv@126.com

Daihui Li
National Innovation Institute of Defense Technology
Chinese Academy of Military Science
Beijing, China
lidh_ai@sina.com

Tongsheng Shen
National Innovation Institute of Defense Technology
Chinese Academy of Military Science
Beijing, China
shents_bj@126.com

Dexin Zhao
National Innovation Institute of Defense Technology
Chinese Academy of Military Science
Beijing, China
zhaodx2008@163.com

***Abstract*—As speech recognition technologies have garnered increasing attention recently, the possibility of integrating related technologies into the field of underwater acoustic target classification (ATC) has piqued researchers' curiosity. However, the classification of underwater acoustic targets confronts several obstacles due to the time-varying and complicated nature of the underwater environment,, such as the low signal-to-noise ratio and background interference. To address these concerns, this paper proposes a combined feature extraction framework and a deep convolutional neural network learning framework. This combination features include MFCC, CQT, Gammatone, Log-mel which are extracted from the underwater acoustic signal and fed into the joint learning framework. The classifier has a deep learning architecture, which is a deep neural network system that can learn differentiating features from physical feature representations to abstract concepts in a hierarchical pattern. Multiple datasets under diverse settings have been used to verify the performance of our proposed architecture. Using the acoustic data gathered in several underwater experiments conducted in a shallow sea setting, the proposed framework achieves a recognition rate of 89.9% and a competitive classification accuracy in the underwater scenario.**

Keywords—underwater acoustic target classification, ensemble feature learning, convolutional neural network

I. INTRODUCTION

In recent years, ATC technology based on machine hearing has become a research hotspot. ATC technology is adapted to specific environmental conditions and outputs target discrimination services according to the context[1-8]. The classification of marine underwater targets has evolved into a critical technology for ocean exploration. Traditional underwater ATC technology relies heavily on time-domain and frequency-domain features . These feature extraction procedures rely heavily on human experience and subjective consciousness, making feature extraction technologies for underwater target classification difficult.

For this purpose, researchers extend the target feature extraction from time domain and frequency domain features to transform domain feature extraction. Generally, a feature vector that can reflect the attributes of the target category is extracted from the original signal that has undergone certain processing, or a transformation method is used to transform the original signal, and the feature vector that reflects the attribute of the target category is extracted from the transform domain. If the extracted feature vector contains enough difference information and enough interference information, the better the target classification performance will be.

Inspired by several representative feature extraction methods in speech recognition methods, such as CQT, MFCC, Log-mel, Gamma, some researchers have conducted research on related features in the classification of underwater targets. Han [9] and colleagues introduced the concept of differential MFCC coefficients and the corresponding feature extraction method and conducted simulation research and experimental analysis based on the MFCC feature extraction method for underwater targets. Thomas [10] et al. integrated the CQT feature extraction method with a parallel neural network, which can effectively obtain the time domain and frequency domain classification difference information in the speech signal. Hu [11] et al. employed Gammatone filter banks instead of the traditional triangular filter banks to simulate the human auditory model, and exponential compression instead of fixed logarithmic compression to simulate the non-linear characteristics of the human auditory model. In the research on the acoustic signal classification of ships and marine organisms, Li [12] et al. used Log-mel features with Gaussian mixture model to train and classify acoustic targets, and discussed different MFCC

dimension changes with different MFCC feature combinations to identify and classify the impact.

Erik [13] et al. combine low-order features like intensity and zero-crossing rate with MFCC power spectrum features, or transform MFCC into another form of features. Such as SPPCC (subspace projection cepstral coefficients) [14], PNCC power regularization to spectral coefficients or PLP coefficients. Lyon [15] et al. examined the history of research on the auditory filter model, as well as the benefits and drawbacks of each model. Li [16] et al. evaluated the research of the auditory peripheral computing model and its application in the field of speech recognition. Ma [17] et al. investigated the nonlinear compression of the auditory system and illustrated the limitations of logarithmic compression in the MFCC extraction procedure. The above several spectrum analysis methods are the process of changing from a linear spectrum to a nonlinear spectrum, and each has its own advantages and disadvantages. In this paper, the four features are merged, and the multi-feature fusion means the voting mechanism is adopted to classify underwater acoustic targets.

Deep learning [18-25] has the ability to learn features from a vast quantity of data automatically, but its application in the field of underwater acoustic signal classification is still in its infancy. Underwater acoustic signal recognition classification differs from speech recognition in that it has a limited sample size and data acquisition problem. In the training stage, the original signal should be extended to the transform domain for feature extraction, and different features have varying characterization capabilities. Judging from the current research situation, using the above-mentioned feature extraction methods to perform feature extraction, recognition, and classification of underwater acoustic target signals has proven to be a successful approach, but each has its scope of application. The four attributes are paired with deep learning approaches to allow it to mine deep-level data features.

In this work, we have three main contributions: (i) we propose a novel ensemble ATC framework; (ii) we investigate three alternative fully CNN models in the proposed ensemble system; (iii) we explore novel data augmentation strategies to reduce the device dependency of our models. Our experimental results are gathered on the underwater data set, and the proposed CNN models produce competitive results. Specifically, By combining the four separate features, we were able to get an overall ATC accuracy of 89.9% and a competitive classification accuracy in the underwater scenario.

II. The Ensemble Features in the Proposed Framework

Traditional underwater acoustic target detection and classification approaches require manual extraction of feature data with generalization and substantial generalization ability. The procedure is lengthy and difficult, as well as sophisticated and requiring human involvement. The recognition and classification process has strong human-computer interaction characteristics. For a long time, feature vector extraction methods have been the research focus of underwater acoustic target recognition and classification. According to the research basis of typical human auditory, analyzing auditory features for underwater acoustic targets and exploring automatic feature extraction methods that increase the capacity of underwater acoustic target recognition is a realistic research.. The underwater acoustic target feature extraction method based on auditory features is one of the research hotspots. The following are the typical auditory features discussed above.

A. MFCC Underwater Acoustic Target Feature Extraction

For a long time, MFCC feature extraction has been utilized in speech recognition, and it is a frequently used in automatic speech and speaker recognition. The Fig. 1 depicts the feature extraction procedure:

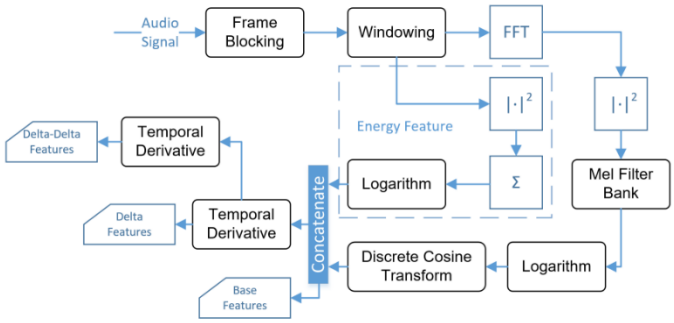

Fig. 1. The process of MFCC feature extraction.

The term "frame blocking" refers to the process of cutting variable-length sounds into fixed-length segments. The necessity for framing arises from the fact that the subsequent Fourier transform is best suited to analyze stationary signals, whereas audio signals vary frequently. There is overlap between frames to maintain the continuity between the signals of each frame, which is called "frame-shift". In this paper, the length of each frame is 46 milliseconds, and the frame-shift is 11 milliseconds.

Since the transformation of the signal in the time domain is usually difficult to see the characteristics of the signal, it is usually converted to the energy distribution in the frequency domain for observation. Different energy distributions can represent the characteristics of different target signals. The rapid Fourier transform is used to obtain the frame signal spectrum, and then the signal energy spectrum is obtained. The formula can be expressed as:

$$p(f) = |X(f)^2| = \left|FFT\left(x(n)\right)\right|^2 \tag{1}$$

In Eq. (1), x(n)is the input frame signal, and X(f) is the spectrum of input frame signal. Pass the obtained energy spectrum through the Mel filter bank, the formula can be expressed as

$$E(m) = \sum_{k=0}^{N-1} (p(f) \cdot H_m(f)) \tag{2}$$

In Eq. (2), N is the total number of points of each frame signal, and H_m (f) is the coefficient of the Mel filter bank. Then do logarithm operation on the filter outputs, and then do discrete cosine transform to get MFCC features.

$$s(m) = log \sum_{k=0}^{N-1} (p(f) \cdot H_m(f)) \tag{3}$$

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad n = 1,2,\dots,L \qquad (4)$$

In the above formula, n is the MFCC order, and M is the number of filters.

B. CQT Features

CQT stands for constant Q transform [26], which refers to a filter bank whose center frequency is distributed exponentially, with different filtering bandwidths, but the ratio of the center frequency to bandwidth is constant Q. It is commonly utilized in music signal processing, but due to the rise of deep learning, it's now being employed in other research fields like ATC. It differs from the Fourier transform in that its horizontal frequency spectrum is not linear, and the length of the filter window can be changed according to the frequency of the spectrum to obtain better performance. Since the distribution of CQT and scale frequencies is the same, by calculating the CQT spectrum of the music signal, the amplitude value of the music signal at each note frequency can be directly obtained, which is suitable for processing music signals.

The time domain speech signal is filtered using a set of constant Q filters. The constant Q here indicates that the center frequency and wideband frequency ratios are the same. The bandwidth is narrow at low frequencies and high at high frequencies since the ratio of each filter is the same, resulting in a non-linear frequency domain signal. It is a technique of converting a linear spectrum to a nonlinear spectrum, similar to MFCC.

C. GammaTone Features

When an external voice signal penetrates the cochlea's basement membrane, it will be decomposed according to its frequency and generate traveling wave vibration, which stimulates the auditory sensory cells [27-28]. GammaTone filter is a set of filter models used to simulate the frequency decomposition characteristics of the cochlea. It can be used for the decomposition of audio signals to facilitate subsequent feature extraction.

The frequency analysis approach of the peripheral auditory system is thought to be able to be mimicked to some extent using a set of band-pass filters. Various filter sets, such as roex filters, have been proposed for this purpose (Patterson and Moore 1986). In neuroscience, there is a calculation method called "reverse correlation" (de Boer and Kuyper 1968), which calculates the response of primary auditory nerve fibers to white noise stimulation and the degree of correlation, that is, before auditory neurons emit action potentials. The average superimposed signal, thereby directly estimating the shape of the auditory filter from the physiological state. This filter takes effect before the peripheral auditory nerve emits action potentials, so it is named "revcor function", which can be used as an estimate of the impulse response of the peripheral auditory filter to a certain limit, that is, the cochlea and other audio signals Band pass filtering.

Johannesma introduced the GammaTone filter (GTF) in 1972 as a mathematical analytic technique for approximating the recvor function. The mathematical expression for this filter bank is simple, and its many features may be simply examined. Since GTF is obtained from the measurement of the impulse response, it has complete amplitude and phase information. In contrast, Only single amplitude information, such as roex filter [29], can be measured in the psychoacoustic shielding experiment. Holdsworth [30] et al. further clarified the various characteristics of GTF and provided a digital IIR filter design scheme. This technology enables GTF to be implemented more easily and efficiently than FIR, paving the way for some important practical applications in the future.

$$h(t) = \begin{cases} ct^{n-1}e^{(-2\pi bt)}\cos(2\pi f_0 t + \phi) & , \ t > 0 \\ 0 & , \ t < 0 \end{cases} \qquad (5)$$

In the above formula, c is the proportional coefficient, n is the filter order, the larger the skewness, the thinner and taller the filter. b is the time attenuation coefficient, the larger the filter time is, the shorter the filter time. f_0 is the filter center frequency, and ϕ is the filter phase.

In this paper, we use the library librosa [29] to extract multiple features from the audio signal and propose the framework in the following way:

- Firstly, we believe that low-level features each contain valuable and complementary information, hence we develop a method to effectively combine different spectrograms input features, namely log-Mel, Gamma, MFCC and CQT spectrograms.
- To extract high-level features from a multi-spectrogram input, we propose a novel encoder-decoder architecture comprising an encoder front end of four parallel CNN-DNN paths (C-DNN). Each CNN block learns to map one spectrogram into high-level features, and we also combine these high-level features from the middle layers of the networks to form a combined feature.
- In terms of acoustic target classifier, we evaluated the combination of different features and models. We compared the performance of the combination features framework to the individual ones.

We evaluate our approach on hydrophone datasets. We will see that the performance of our proposed architecture is competitive with the individual feature model. The remainder of this paper is structured as follows. Section 3 explains why we chose a mixed multi-spectrogram architecture. Performance comparison and discussion are discussed in Section 4. We conclude in Section 5.

III. The Proposed Framework

The architecture of our framework is outlined in Fig. 2. The figure below is the overall framework of this article. In the beginning, spectrograms of the underwater acoustic signal from a certain channel in the hydrophone is represented. Then four different types of feature extractors are used to generate corresponding energy features. In order to suppress the problem of poor adaptability of a single feature in different environments and different target conditions, we combine the four features of log-Mel, Gammatone, CQT and MFCC spectrograms which are split into non-overlapping patches with the size of 128×128. The

configuration parameters we used in the process of these types of feature extraction are shown in Table I:

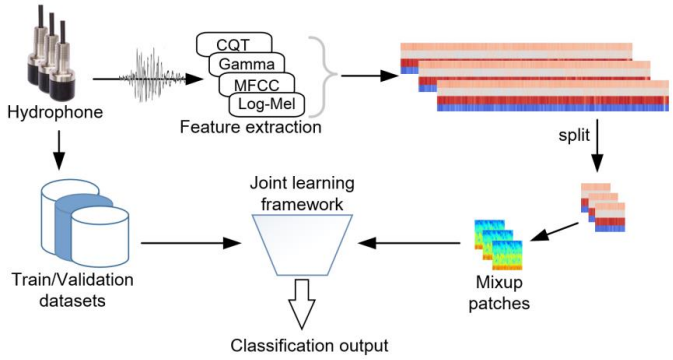

Fig. 2. The architecture of our framework.

TABLE I. THE FEATURE EXTRACTION SETTINGS IN THE FRAMEWORK

Parameters	*Settings*
Window size	2048
Hop size	512
Mel bins	40
Minimum frequency	50
Maximum frequency	14000
MFCC frames	431
Number of MFCC bins	40
MFCC hop size	512
Gammatone frames	499
Number of gammatone bins	64

A. Data Augmentation

Deep neural network training necessitates a significant amount of data, yet the cost of acquiring underwater sound data is substantial, and the amount of data available is restricted.. In order to improve the robustness of the overall framework, this paper uses the mixup data augmentation method [31-33]. Mixup is an algorithm that was originally used in computer vision to enhance images by mixing. It can mix images between different classes to expand the training data set. Suppose that b_{x1} is a batch sample, and b_{y1} is a label corresponding to the sample; meanwhile, b_{x2} is a batch sample, and b_{y2} is a label corresponding to the sample. λ is the mixing coefficient calculated from the beta distribution with the parameters α and β, so we can get the new sample and label from the following formula:

$$Mixup \begin{cases} \lambda = Beta(\alpha, \beta) \\ mixb_x = \lambda * b_{x1} + (1-\lambda) * b_{x2} \\ mixb_y = \lambda * b_{y1} + (1-\lambda) * b_{y2} \end{cases} \quad (6)$$

In the above formula, $Beta$ is the beta distribution, $mixb_x$ is the batch sample after mixing, and $mixb_y$ is the label corresponding to the mixed batch sample. In the framework, we let the value of α and β be 0.5.

B. Joint Learning Architecture of ATC

In order to evaluate individual and multiple spectrograms, we proposed a C-DNN network architecture as described in Table II and Fig. 3 and. CNN part is described by 4 Cv-Bn blocks, performed by Batchnorm (Bn), Convolutional (Cv[kernel size]), Rectified linear unit (ReLu), Dropout (Dr(Percentage dropped)), Average Pooling (Ap) layers as shown in the top of Table II. Global pooling is applied after the last convolutional layer to obtain fixed-length vectors, which is operated by global average pooling in the frequency axis and global max pooling in the temporal axis as Fig. 3 shown, C-DNN architecture comprises of CNN and DNN parts in order. Meanwhile, the DNN part in Fig. 3 is configured by three Fl blocks with Fully-connect (Fl), ReLu, Dropout (Dr(Percentage dropped)), and Softmax layers, as described at the bottom of Table II. It can be seen that CNN part helps to map input image patches to condensed and discriminative vectors, referred to as current individual features. Each feature vector presents 512 dimensions due to the number of kernels used in the final convolutional layer in Cv-Bn block 04. Next, the DNN part explores the high-level features, thus classified into 10 categories (the category number in the acoustics datasets) and reports the classification accuracy.

TABLE II. NETWORK STRUCTURE USED IN OUR FRAMEWORK

Parameters	*Settings*
Cv-Bn Block01	Cv [3×3]@64 - BN - ReLU
Pooling	Avg Pooling 4 × 2
Cv-Bn Block02	Cv [3×3]@128 - BN - ReLU
Pooling	Avg Pooling 4 × 2
Cv-Bn Block03	Cv [3×3]@256 - BN - ReLU
Pooling	Avg Pooling 2 × 2
Cv-Bn Block04	Cv [3×3]@512 - BN - ReLU
Pooling	Global Pooling
Fl Block01	Fl - ReLu - Dr
Fl Block02	Fl - ReLu - Dr
Fl Block03	Fl - Softmax

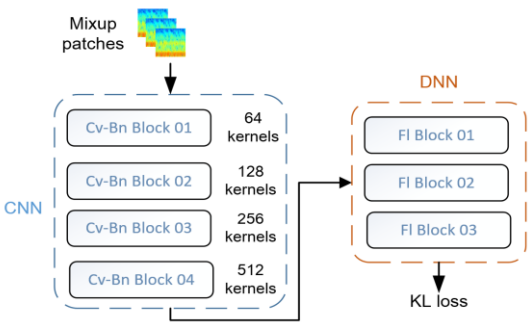

Fig. 3. Flowchart of blocks in the framework.

C. Hyper Parameters Setting

In this work, we adopt Pytorch [34] framework to build deep learning models with a learning rate of 0.001, a batch size of 64, epoch number of 15000, and the Adam method [35] for learning rate optimization. As using mixup data augmentation, the labels are not a one-hot format. Therefore, we use Kullback-Leibler (KL) divergence loss [36] instead of the standard cross-entropy loss as shown in Equation below:

$$Loss_{KL}(\theta) = \sum_{n=1}^{N} y_n \log(y_n/\hat{y}_n) + \frac{\lambda}{2}\|\theta\|_2^2 \qquad (7)$$

In Eq. (7), $Loss_{KL}(\theta)$ is KL-loss function, θ describes the trainable parameters of the network trained, λ denote the ℓ_2-norm regularization coefficient experimentally set to 0.001, N is the batch size, y_n and $\hat{y}_n$ are the ground-truth and the network recognized output, respectively.

IV. Performance Comparison and Discussion

Our datasets come from three different shallow sea environments. The marine environment at the time of data collection is quite variable due to the variations in location and time. So we label the data of three different sea areas as A, B, and C respectively. The total number of samples in the dataset is 3543, with 709 examples in the validation dataset accounting for 20% of the total. Each sample lasts around 10 seconds, and the sampling frequency is 44,100 Hz. As shown in Table III, the combination features method performs well in classifying practically all target types, particularly MotorboatA, which is just slightly behind in one category.

Initially, by putting all four individual spectrograms into the C-DNN network, we are able to compare category-wise performance on the ten target classifications. The validation set's results are represented in the form of a confusion matrix. In most categories, Gammatone and MFCC outperform Log-mel and CQT, as shown in Fig. 4. Gammatone comes out on top in terms of average accuracy, with a competitive score of 88.6 percent. Meanwhile, CQT has poor average scores when compared to other features, showing a 2.1 percent performance difference. The Gammatone spectrogram performed exceptionally well in ship categories such as FishingboatA and PassengerB.

TABLE III. The Accuracy of Individual Spectrograms and Their Combination on 3 Different Datasets

Type	*Log-mel*	*CQT*	*Gamma*	*MFCC*	*Combi-nation*
FishingboatA	0.933	0.844	0.956	0.933	0.956
MotorboatA	0.690	0.793	0.828	0.862	0.931
PassengersA	0.951	0.981	0.981	0.981	1.000
OceanlinerA	1.000	0.955	1.000	0.932	1.000
PassengerB	0.806	0.806	0.821	0.791	0.806
CargoB	0.741	0.724	0.690	0.741	0.707
YachtB	0.872	0.862	0.862	0.872	0.883
Othersboat	0.764	0.764	0.753	0.742	0.764
CargoshipC	0.960	0.920	0.980	0.980	0.940
MerchantC	0.992	1.000	0.992	1.000	1.000
Average accuracy	0.871	0.865	0.886	0.883	0.899

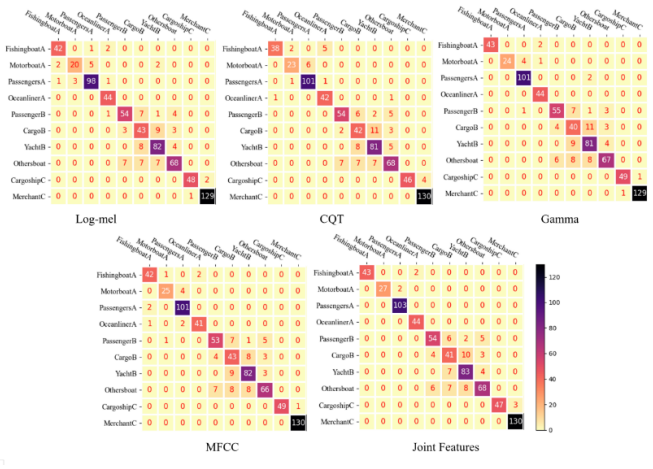

Fig. 4. Category-wise performance comparison among spectrograms and their combination.

The results on the four-spectrogram combination witness an increase of 1.3% to 3.4% on individual spectrogram. As regards the results in some items, such as PassengerB, CargoB, CargoshipC, the combination lags behind the Gammatone, by up to a maximum of 4 percentage in a single item, but staying ahead on the whole.

We summarize all types of spectrogram combinations, and highlight which achieves the best scores on all datasets. As results shown in Table III and Fig. 3, there is an increase in the accuracy rate when the combination of spectrograms are applied. In particular, Gammatone achieves the best performance among individual spectrograms, and its performance even surpasses combination in some cases. By combining all the features, it is possible to improve by 2.3% on average. Therefore, the results generally demonstrate the robustness of the combination.

V. Conclusion

This paper has presented a robust framework applying for ATC tasks. In front-end feature extraction, the idea of providing a comprehensive analysis of low-level spectrogram representation from drawing audio signals is able to figure out the effective types of individual spectrograms and their combination. As regards back-end classification, our innovative join learning network based on parallel convolutional recurrent architecture has facilitated learning both spatial and temporal structural features of spectrograms. In the underwater scenario, we attain competitive results leveraging multi-spectrogram input and the join learning network, compared to individual spectrograms.

References

[1] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, Wei Xiao, and Huy Phan, Continuous robust sound event classification using time-frequency features and deep learning, PloS one, vol. 12, no. 9, pp. e0182309, 2017.

[2] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim, Acoustic scene classification using parallel combination of LSTM and CNN, in Proc. DCASE, pp. 11-15, 2016.

[3] Ren Zhao, Kong Qiuqiang, Qian Kun, D.Plumbley Mark, and W.Schuller1 Bjorn, Attention-based convolutional neural networks for acoustic scene classification, in Proc. DCASE, pp. 39-43, 2018.

[4] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes, in Proc. ICASSP, pp. 56-60, 2019.

[5] Sai Phaye, Emmanouil Benetos, and Ye Wang, SubSpectralNet - Using sub-spectrogram based convolutional neural networks for acoustic scene classification, in Proc. ICASSP, pp. 825-829, 2019.

[6] Hongwei Song, Jiqing Han, Shiwen Deng, and Zhihao Du, Acoustic scene classification by implicitly identifying distinct sound events, in Proc. INTERSPEECH, pp. 3860–3864, 2019.

[7] Ian Vince McLoughlin, Speech and Audio Processing: a MATLAB-based approach, Cambridge University Press, 2016.

[8] Ian McLoughlin, Zhipeng Xie, Yan Song, Huy Phan, and Ramaswamy Palaniappan, Time-frequency feature fusion for noise robust audio event classification, J. Circuits Syst. Signal Proc., 2019.

[9] Han Xue. Feature Extraction of Underwater Target Based on Auditory Features. Harbin Engineering University, 2013.

[10] Thomas Lidy, CQT-Based Convolutional Neural Networks For Audio Scene Classification And Domestic Audio Tagging, Dcase2016, 3 Sept. 2016.

[11] Hu Fengsong, Cao Xiao yu, Auditory Feature Extraction Based on Gammatone Filter Bank, Computer Engineering, Vol.38, No.21, Nov. 2012.

[12] Li Xinxin. Feature extraction from underwater signals using wavelet packet transform. Harbin Engineering University, 2012.

[13] Erik Marchi, Dario Tonelli, Xinzhou Xu. Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification, in Proc. DCASE, pp. 65–69, 2016.

[14] Sangwook Park, Seongkyu Mun, Younglo Lee, and Hanseok Ko, Score fusion of classification systems for acoustic scene classification, Tech. Rep., DCASE2016 Challenge, 2016.

[15] Lyon, Richard F., Andreas G. Katsiamis, and Emmanuel M. Drakakis. History and future of auditory filter models. Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE, 2010.

[16] Li Zhaohui, Chi Huisheng. Progress in computational modeling of auditory periphery. ACTA Acustica, 31(5) : 449-465, 2006.

[17] Ma Yuanfeng, Chen Kean. A New Cepstrum Coefficents Applied to Acoustic Target Recognition. ACTA Armamentari, 30(11): 1477-1483, 2009.

[18] Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10): 2222-2232, 2017.

[19] Lam Pham, Ian McLoughlin, Huy Phan, Ramaswamy Palaniappan, and Yue Lang, Bag-of-features models based on C-DNN network for acoustic scene classification, in Proc. AES, 2019.

[20] Shengkui Zhao, Thi Ngoc Tho Nguyen, Woon-Seng Gan, and Jones Douglas L., ADSC submission for DCASE 2017: Acoustic scene classification using deep residual convolutional neural networks, Tech. Rep., DCASE2017 Challenge, September 2017.

[21] Wei Dai, Juncheng Li, Phuong Pham, Samarjit Das, and Shuhui Qu, Acoustic scene recognition with deep neural networks (DCASE challenge 2016), Tech. Rep., DCASE2016 Challenge, September, 2016.

[22] Ian McLoughlin, Yan Song, Lam Dam Pham, Huy Pham, Palaniappan Ramaswamy, and Lang Yue, Early detection of continuous and partial audio events using CNN, in Proc. INTERSPEECH, 2018.

[23] Huy Phan, Phillip Koch, Ian McLoughlin, and Alfred Mertins, Enabling early audio event detection with neural networks, in Proc. ICASSP, 2018.

[24] Huy Phan, Lars Hertel, Marco Maass, Philipp Koch, and Alfred Mertins, Label tree embeddings for acoustic scene classification, in Proc. ACM, pp. 486-490, 2016.

[25] Ekaterina Garmash and Christof Monz, Ensemble learning for multi-source neural machine translation, in The 26th International Conference on Computational Linguistics: Technical Papers, pp. 1409–1418, 2016.

[26] Thomas Lidy and Alexander Schindler, CQT-based convolutional neural networks for audio scene classification, in pro. DCASE, pp. 1032–1048, 2016.

[27] Ditter D, Gerkmann T. A multi-phase gammatone filterbank for speech separation via tasnet. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE: pp. 36-40, 2020.

[28] Ren J, Jiang X, Yuan J, et al. Sound-event classification using robust texture features for robot hearing. IEEE Transactions on Multimedia, 19(3): 447-458, 2016.

[29] McFee, Brian, Raffel Colin, Liang Dawen, Daniel. PW.Ellis, McVicar Matt, Battenberg Eric, and Nieto Oriol, librosa: Audio and music signal analysis in python, in Proceedings of The 14th Python in Science Conference, pp. 18–25, 2015.

[30] Al-Dayyeni W S, Sun P, Qin J. Investigations of auditory filters based excitation patterns for assessment of noise induced hearing loss. arXiv preprint arXiv:1705.10805, 2017.

[31] Holdsworth J, Nimmo-Smith I, Patterson R, et al. Implementing a gammatone filter bank[J]. Annex C of the SVOS Final Report: Part A: The Auditory Filterbank, 1: 1-5, 1988.

[32] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu, Mixup-based acoustic scene classification using multi-channel convolutional neural network, in Pacific Rim Conference on Multimedia, pp. 14–23, 2018.

[33] Thulasidasan S, Chennupati G, Bilmes J, et al. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. arXiv preprint arXiv:1905.11001, 2019.

[34] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703, 2019.

[35] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[36] Eguchi S, Copas J. Interpreting kullback-leibler divergence with the neyman-pearson lemma. Journal of Multivariate Analysis, 97(9): 2034-2040, 2006.