Journal of
Marine Science and Engineering

MDPI

Article

Joint Representation and Recognition for Ship-Radiated Noise Based on Multimodal Deep Learning

Fei Yuan, Xiaoquan Ke and En Cheng *

Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, Xiamen University, Xiamen 361005, China; yuanfei@xmu.edu.cn (F.Y.); 23320171153169@stu.xmu.edu.cn (X.K.)
* Correspondence: chengen@xmu.edu.cn

Received: 23 September 2019; Accepted: 22 October 2019; Published: 27 October 2019

Abstract: Ship recognition based on ship-radiated noise is one of the most important and challenging subjects in underwater acoustic signal processing. The recognition methods for ship-radiated noise recognition include traditional methods and deep learning (DL) methods. Developing from the DL methods and inspired by audio–video speech recognition (AVSR), the paper further introduces multimodal deep learning (multimodal-DL) methods for the recognition of ship-radiated noise. In this paper, ship-radiated noise (acoustics modality) and visual observation of the ships (visual modality) are two different modalities that the multimodal-DL methods model on. The paper specially designs a multimodal-DL framework, the multimodal convolutional neural networks (multimodal-CNNs) for the recognition of ship-radiated noise. Then the paper proposes a strategy based on canonical correlation analysis (CCA-based strategy) to build a joint representation and recognition on the two different single-modality (acoustics modality and visual modality). The multimodal-CNNs and the CCA-based strategy are tested on real ship-radiated noise data recorded. Experimental results show that, using the CCA-based strategy, strong-discriminative information can be built from weak-discriminative information provided from a single-modality. Experimental results also show that as long as any one of the single-modalities can provide information for the recognition, the multimodal-DL methods can have a much better multiclass recognition performance than the DL methods. The paper also discusses the advantages and superiorities of the multimodal-Dl methods over the traditional methods for ship-radiated noise recognition.

Keywords: ship-radiated noise recognition; pattern recognition; multimodal deep learning; canonical correlation analysis

1. Introduction

Generally, when a ship moves on the water, it produces noise, called ship-radiated noise. The ship-radiated noise, along with marine mammals' voice, and natural ambient noise, constitutes most of the acoustic sound in oceans, which makes the underwater acoustic sound informative and also very complicated. Generally speaking, ship-radiated noise of different types of ships or different ships contains different acoustic characteristics, so it is possible to recognize different types of ships by analyzing the ship-radiated noise. Ship-radiated noise recognition is aimed at detecting and recognizing the marine vessels with the radiated acoustic signals recorded by passive sonars. It has many important applications in ocean engineering, such as automatic target recognition (ATR) and marine monitoring. Ship recognition based on ship-radiated noise is one of the most important and challenging subjects in underwater acoustic signal processing.

Information in the real world comes through multiple modalities. Images are associated with captions and tags, videos contain visual and audio signals, and sensory perception includes simultaneous inputs from visual, auditory, motor, and haptic pathways. Each modality is characterized by very distinct statistical properties, which makes it difficult to disregard the fact that they come from different input channels. Useful representations can potentially be learned for such data by combining the modalities into a joint representation [1]. In [2], Jiquan et al. have represented a series of audio–video speech recognition (AVSR) tasks for *multimodal deep learning* (multimodal-DL) methods by jointly modeling on different modalities. In [2], the audio inputs and video inputs are the two different modalities that the multimodal-DL methods model on. A moving ship can also be jointly learned from the ship-radiated noise (acoustics observation) as well as visual observation of the ship, because both of these two observations are simultaneously generated from the same moving ship. The acoustics observation and visual observation can also be treated as two different modalities from the ships. Inspired by AVSR [2] based on multimodal-DL methods, the paper also introduces multimodal-DL methods to jointly model on different modalities from the ships for ship-radiated noise recognition. In this paper, the ship-radiated noise (*acoustics modality*) and visual observation of the ships (*visual modality*) are the two different modalities that the multimodal-DL methods model on.

In recent years, convolutional neural networks (CNN) architectures have been successfully applied to many pattern recognition tasks with local connectivity and weight sharing [3]. Using CNN architectures, approaches developed for image recognition can be extended to ship-radiated noise recognition by regarding the time–frequency representations (T–F representations) of the signals, as contributed in [4]. Inspired by [4], the paper specially designs a CNN-based multimodal-DL framework (multimodal-CNNs) to jointly model on the two different modalities (acoustics modality and visual modality) from the ships.

The most import issue of multimodal-DL methods for ship-radiated noise recognition is to figure out the *joint representation* problem, that is to build a more discriminative joint representation over the two different modalities. The more discriminative joint representation over the two modalities can not be built by simply combining these two modalities. Inspired by feature fusion using canonical correlation analysis (CCA) [5], the paper proposes a *CCA-based strategy* to build more discriminative representations. The CCA method can be used in pattern recognition applications for fusing the features extracted from multiple modalities or combining different feature vectors extracted from a single-modality [5]. CCA is to establish the correlation criterion function between the two groups of feature vectors, to extract their canonical correlation features according to this criterion, and to form effective discriminative vectors for recognition [6]. This method uses correlation features between two groups of feature vectors as effective discriminative information, so it not only is suitable for information fusion, but also eliminates the redundant information within the features [6]. All in all, the goal of the CCA-based strategy is to combine relevant information from two modalities into a single-modality with more discriminative power than any of the input modalities.

The main contributions of the paper are summarized as follows:

1. Based on DL methods, the paper further introduces multimodal-DL methods for ship-radiated noise recognition, and advantages and superiorities of the multimodal-DL methods over the DL methods and traditional methods are demonstrated.
2. The paper proposes the multimodal-DL framework, the multimodal-CNNs to simultaneously model on the two modalities.
3. The paper proposes the CCA-based strategy to build a more discriminative joint representation and recognition on the two single-modality.

The paper contains six sections. Section 2 briefly introduces other related work of ship-radiated noise recognition, including traditional methods and DL methods. Section 3 explains technical details of the multimodal-DL methods for ship-radiated noise recognition. Section 4 conducts experiments

and discusses experimental results. Section 5 contains a brief summary of our work and the last section is the future perspective of our work.

2. Related Work

Recognition of ship-radiated noise also belongs to pattern recognition problems. Though the recognition methods for ship-radiated noise have been developed for decades, performance of these methods still cannot satisfy practical demands. Typically, traditional methods of recognition of ship-radiated noise is extracting hand-crafted features then followed by classification. Numerous techniques have been proposed for the feature extraction of ship-radiated noise. For example, Jian et al. in [7] extracted line spectrum and line spectrum density features from ship-radiated noise and fed them into support vector machine (SVM) classifiers. In [8], Wei et al. introduced an approach for extracting ship-radiated noise based on $1\frac{1}{2}D$ spectrum features. In [9], Mel-frequency cepstral coefficients (MFCCs) features were extracted and statistical classification of these features were based on Gaussian mixture models (GMMs). MFCCs, along with first-order differential MFCCs, and second-order differential MFCCs features were also used in [10] to recognize underwater target. Meng and Yang in [11] designed a combined feature vector of zero-crossing wavelength, peek-to-peek amplitude, and zero-crossing-wavelength difference for the recognition of ship-radiated noise. In [12], linear predictive coding (LPC) features were extracted, in [13], energy distribution features in the blocks of wavelet packet (WP) coefficients were extracted, and in [14], Hilbert spectral features were extracted for the recognition of ships.

Compared with traditional ship-radiated noise recognition methods based on a priori knowledge, DL methods are able to hierarchically learn high-level features from a large number of ship-radiated noise data, and the extracted deep features are more robust to invariants. DL methods require less engineering skill, domain expertise and prior knowledge, but utilizing DL methods can also achieve competitive performance even outstanding performance compared with the traditional methods. For instance, in [15], Cao et al. utilized Stacked Autoencoder (SAE) to extract high-level features for ship recognition with short time frequency transform, which provided competitive performance. Ke et al. in [16] utilized deep Autoencoder and SVM to classify two classes of underwater acoustic targets. In [17,18], deep belief networks (DBNs) were utilized to the recognition of ship-radiated noise. The deep convolutional neural network (CNN) has been applied to ship-radiated noise recognition in [4]. Note that one of the advantages of the DL methods is that some of these methods can additionally make use of unlabeled ship-radiated noise data. For example, in [18], unsupervised pre-training models such as DBNs can make use of unlabeled data, and only a small number of labeled ship-radiated noise data are required.

3. Multimodal Deep Learning Methods for Ship-Radiated Nose Recognition

This section fully explains technical details of the multimodal-DL methods for ship-radiated noise recognition. Application scenario of the multimodal-DL methods is introduced before explanation of the multimodal-CNNs framework. Then training method for the multimodal-CNNs framework will be explained. Finally, the CCA-based strategy used for building a joint representation and recognition will be explained in detail.

3.1. Application Scenario

The application scenario that the multimodal-DL methods based on is depicted in Figure 1.

As shown in Figure 1, visual observation and acoustics observation of the ships are collected by high-performance camera and hydrophone, respectively. Then both of these two observations are sent

to monitor station for further analysis. In this paper, the purpose of the analysis is to recognize the class of ship[1]. Harbor monitoring and securing can also apply to this scenario.

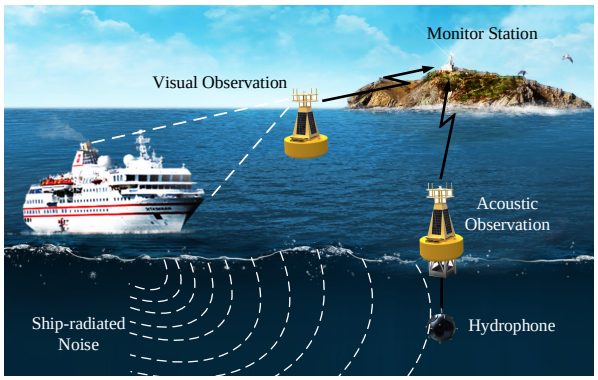


Figure 1. Application scenario of the multimodal-deep learning (DL) methods for ship-radiated noise recognition.

Different from other DL methods that only make use of one of the two modalities from the ships, the multimodal-DL methods simultaneously make use of the both two modalities. The core idea of the multimodal-DL methods is that each of the modalities is assumed to reflect different characteristics of the ships, and via jointly modeling on the two modalities, we can achieve a better joint representation and recognition of the ships compared with other DL methods that only make use of a single-modality.

3.2. The Multimodal-CNNs Framework

The multimodal-CNNs framework is fully explained in this section. A similar multimodal-DL framework for AVSR has been depicted in [19].

The number of layers, kernel size of each layer, kernel number, kernel stride, and so on are hyper-parameters of the multimodal-CNNs framework that need to be manually adjusted according to previous work, experiences, or intuitions. We adjust these hyper-parameters according to previous work and experiences and we conduct corresponding experiments to verify the performance of the framework of these adjustments. In other words, we explore the detailed structures of the framework and we conduct corresponding experiments to figure out whether these structures can improve the performance (mainly refers to recognition accuracy) of the framework. Through many experiments, the hyper-parameters that lead to the best performance are chosen. The process of adjusting the hyper-parameters is similar to parameter optimization in [20] where conducts many experiments to choose the optimal parameters that lead to the best performance of the network.

To our best knowledge, the multimodal-CNNs has never been introduced to any other ship-radiated noise recognition. The multimodal-CNNs framework is depicted in Figure 2.

---

[1] We recognize 10 classes of ships in this paper, and the recognition of a specific ship is similar.

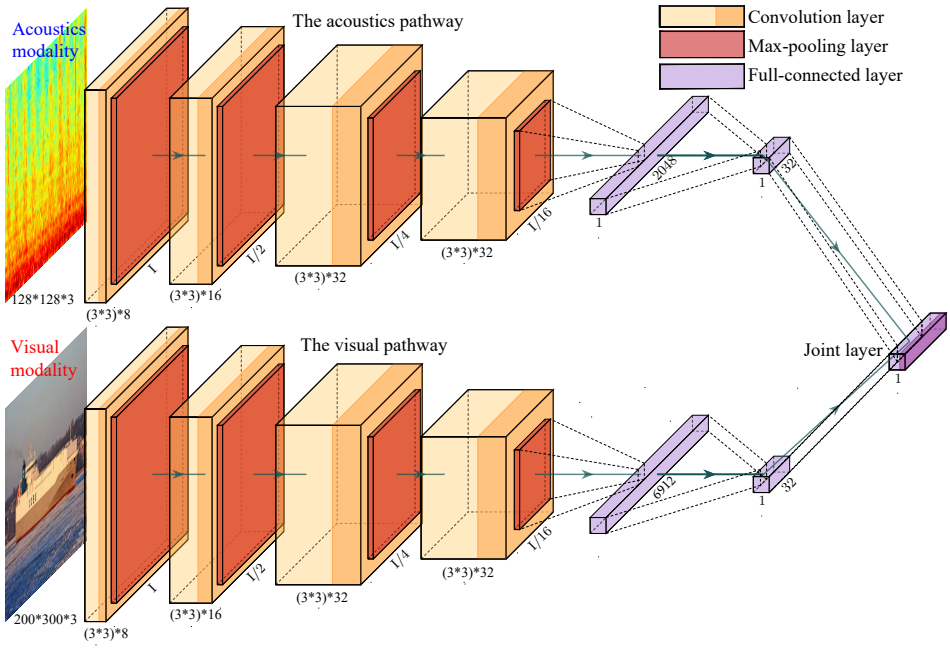


Figure 2. The multimodal convolutional neural networks (multimodal-CNNs) framework for ship-radiated noise recognition.

3.3. Training Method for Multimodal-CNNs Framework

Assume that the top-level feature vectors of each pathway of the multimodal-CNNs passing through a softmax activation layer output:

$$softmax\left(\alpha\right)_i = \frac{e^{\alpha_i}}{\sum_{i=1}^{10} e^{\alpha_i}}, i = 1, 2, \cdots, 10. \tag{1}$$

Equation (1) means that the softmax activation layer will turn the feature vectors into a probability distribution in which each element corresponds to a specific class (e.g., a class of ships). Objective of the training is to minimize cross-entropy between the probability distribution and ship class label distribution:

$$H\left(y_i, \frac{e^{\alpha_i}}{\sum_{i=1}^{10} e^{\alpha_i}}\right) = -\sum_{i=1}^{10} y_i \log\left(\frac{e^{\alpha_i}}{\sum_{i=1}^{10} e^{\alpha_i}}\right), \tag{2}$$

where $y = (y_1, y_2, \cdots y_i, \cdots, y_{10})$ is the ship class label distribution. If we minimize Equation (2), the probability distribution will become much closer to the ship class label distribution. The training method for multimodal-CNNs framework can be summarized as follows:

$$\begin{aligned} &\arg\min_{\alpha_i} H\left(y_i, \frac{e^{\alpha_i}}{\sum_{i=1}^{10} e^{\alpha_i}}\right) \\ =&\arg\min_{\alpha_i} \left[-\sum_{i=1}^{10} y_i \log\left(\frac{e^{\alpha_i}}{\sum_{i=1}^{10} e^{\alpha_i}}\right)\right]. \end{aligned} \tag{3}$$

3.4. CCA-Based Strategy

In multivariate statistical analysis, correlation problem of two random vectors often needs to be studied, that is to convert the correlation research of two random vectors into that of a few pairs of variables, which are uncorrelated [6]. CCA is one of the valuable multi-data processing methods, which has been widely used to analyze the mutual relationships between two sets of variables [5]. Assume that $X \in \mathbb{R}^{p \times n}$ and $Y \in \mathbb{R}^{q \times n}$ denote two different modalities from the same patterns, where

p is feature dimension of the first modality, q is feature dimension of the second modality, and n is sample number of each modality.

Suppose that $S_{XX} \in \mathbb{R}^{p \times p}$ and $S_{YY} \in \mathbb{R}^{q \times q}$ denote the within-sets covariance matrices of X and Y, and S_{XY} denotes the between-sets covariance matrix. The overall $(p+q) \times (p+q)$ covariance matrix S can be combined as follows:

$$\begin{aligned} S &= \begin{pmatrix} cov(X) & cov(X,Y) \\ cov(Y,X) & cov(Y) \end{pmatrix} \\ &= \begin{pmatrix} D(X) & E\left(XY^T\right) \\ E\left(YX^T\right) & D(Y) \end{pmatrix} \\ &= \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix}. \end{aligned} \tag{4}$$

CCA aims to find linear combinations, $X^* = W_x^T X$ and $Y^* = W_y^T Y$ that maximize pair-wise correlations across the two modalities:

$$\arg\max \frac{cov(X^*, Y^*)}{\sqrt{D(X^*)}\sqrt{D(Y^*)}}, \tag{5}$$

where:

$$\begin{aligned} cov(X^*, Y^*) &= cov\left(W_x^T X, W_y^T Y\right) \\ &= E\left(\left(W_x^T X\right)\left(W_y^T Y\right)\right) \\ &= W_x^T E\left(XY^T\right) W_y \\ &= W_x^T S_{XY} W_y, \end{aligned} \tag{6}$$

$$D(X^*) = W_x^T E\left(XX^T\right) W_x = W_x^T S_{XX} W_x, \tag{7}$$

$$D(Y^*) = W_y^T E\left(YY^T\right) W_y = W_y^T S_{YY} W_y. \tag{8}$$

Equation (5) can be transformed to:

$$\begin{aligned} &\arg\max \frac{cov(X^*, Y^*)}{\sqrt{D(X^*)}\sqrt{D(Y^*)}} \\ =&\arg\max_{W_x, W_y} \frac{W_x^T S_{XY} W_y}{\sqrt{W_x^T S_{XX} W_x}\sqrt{W_y^T S_{YY} W_y}} \end{aligned}. \tag{9}$$

Conclusively, projective matrices W_x and W_y can be obtained by solving the following optimization problems:

$$\begin{cases} \arg\max\limits_{W_x, W_y} W_x^T S_{XY} W_y \\ W_x^T S_{XX} W_x = 1 \\ W_y^T S_{YY} W_y = 1 \end{cases}. \tag{10}$$

Optimization of Equation (10) is usually performed using Lagrange multipliers [5]. Note that the linear combinations X^* and Y^* obtained by solving Equation (10) are also known as *canonical variates* [6]. Following the canonical variates:

$$Z_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix}, \tag{11}$$

$$Z_2 = X^* + Y^* = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix}, \tag{12}$$

as the combinatorial feature projected respectively [6], used for pattern recognition, while the transformation matrix are $W_1 = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}$ and $W_2 = \begin{pmatrix} W_x \\ W_y \end{pmatrix}$.

We also call W_1 and W_2 as the canonical projective matrix (CPM), Z_1 and Z_2 as the canonical correlation discriminant features (CCDFs) [6], and transformation Equations (11) and (12) as the CCA feature fusion strategy I (CCA-FFS I) and CCA feature fusion strategy II (CCA-FFS II), respectively.

4. Experiments and Discussion

4.1. Experiment Setting

All data of ship-radiated noise come from a database called ShipsEar [9]. During 2012 and 2013 the sounds of many different classes of ships were recorded on the Spanish Atlantic coast and were included in the ShipsEar database (available at http://atlanttic.uvigo.es/underwaternoise/). The recordings were made with autonomous acoustic digitalHyd SR-1 recorders, manufactured by MarSensing Lda (Faro, Portugal). Each originally recorded signal in the dataset is framed using a window of length 52,734. With sampling frequency $fs = 52{,}734$ Hz, each sample lasts 1 s. Table 1 shows the classes of ship-radiated noise and their number of samples in training set and testing set. The ID in Table 1 is the index of the ship-radiated noise in the ShipsEar database.

Table 1. Training set and testing set of classes of ship-radiated noise.

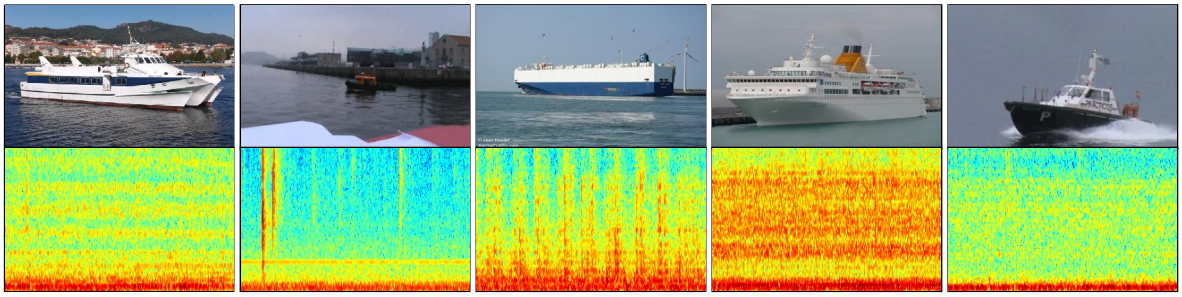
		Training Set	Testing Set
	ID	**Number of Samples**	**Number of Samples**
Class 1: Passenger ferries	60, 61, 62	554	252
Class 2: Tugboats	15, 31	143	63
Class 3: RO-RO vessels	18, 19, 20	789	297
Class 4: Ocean liners	22, 24, 25	159	76
Class 5: Pilot boats	29, 30	105	33
Class 6: Motorboats	50, 51, 52, 70, 72, 77, 79	487	229
Class 7: Mussel boats	46, 47, 48, 49, 66	497	233
Class 8: Sailboats	37, 56, 57, 68	282	126
Class 9: Fishing boats	73, 74, 75, 76	366	148
Class 10: Dredgers	80, 93, 94, 95, 96	188	74
Total		3570	1531

Note that the training set and testing set are not totally intersected, and training samples of each class are unbalanced. We randomly pick up some spectrogram and pictures in each class of ships and show them in Figure 3.

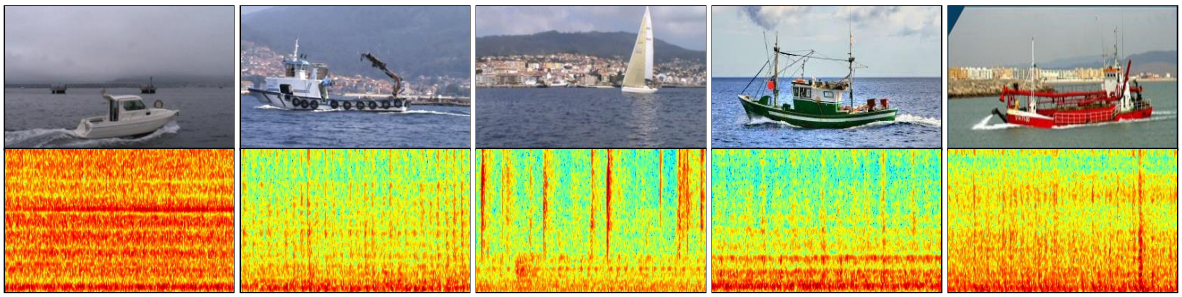
The sampling frequency is 52,734 Hz, thus the analyzed bandwidth of spectrogram is 0–26,367 Hz. However, according to [9] that only the first 8 kHz turned out to be useful for the best classification rate, we only use 0–8000 Hz of spectrogram for analysis.

Assume that X represents the output of the acoustics pathway, and Y represents the output of the visual pathway, then we can define:

Definition 1. *Acoustics (modality):* X
Visual (modality): Y
FFS I-Acoustics-Visual: $X + Y$
FFS II-Acoustics-Visual: $\begin{pmatrix} X \\ Y \end{pmatrix}$
CCA-FFS-Acoustics: $W_x^T X$
CCA-FFS-Visual: $W_y^T Y$
CCA-FFS I-Acoustics-Visual: $X^* + Y^* = W_x^T X + W_y^T Y$
CCA-FFS II-Acoustics-Visual: $\begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix}$



(**a**) Class 1. (**b**) Class 2. (**c**) Class 3. (**d**) Class 4. (**e**) Class 5.



(**f**) Class 6. (**g**) Class 7. (**h**) Class 8. (**i**) Class 9. (**j**) Class 10.

Figure 3. Random samples of spectrogram and pictures in each class of ships.

4.2. Single-Modality Consideration

This section considers the acoustics modality and the visual modality, respectively. The acoustics modality contains Acoustics and CCA-FFS-Acoustics, while the visual modality contains Visual and CCA-FFS-Visual. In the recognition experiments, the training set was used to train SVM classifiers and the testing set was used to obtain the multi-class recognition accuracies.

4.2.1. The Acoustics Modality

In this subsection, we consider multi-class recognition accuracies of Acoustics and CCA-FFS-Acoustics, respectively. Note that in CCA-FFS-Acoustics, correlations between the acoustics modality and the visual modality have been considered, as can be seen in Equation (6), "marginal correlations" of the visual modality on the acoustics modality are represented by W_x^T, while Acoustics is totally independent from the visual modality. The multi-class recognition accuracies of the acoustics

modality under different signal-to-noise ratio (SNR[2]) are depicted in Figure 4. The comparison between Acoustics and CCA-FFS-Acoustics is also depicted in Figure 4. As can be seen from Figure 4, the multi-class recognition accuracies of Acoustics and CCA-FFS-Acoustics improve with the increasing of SNR. However, no essential difference can be seen between the recognition performance of Acoustics and CCA-FFS-Acoustics, as shown in Figure 4. This is partly due to the fact that the acoustics modality contains much more discriminative information than the visual modality. Therefore, it is relatively difficult for the visual modality to provide discriminative enough "marginal correlations" on the acoustics modality. The used ship-radiated noise database provides much more various acoustics modality than the visual modality, as the acoustics modality is recorded using a high-quality hydrophone for a long time while the visual modality is only represented by ship pictures with less variation.

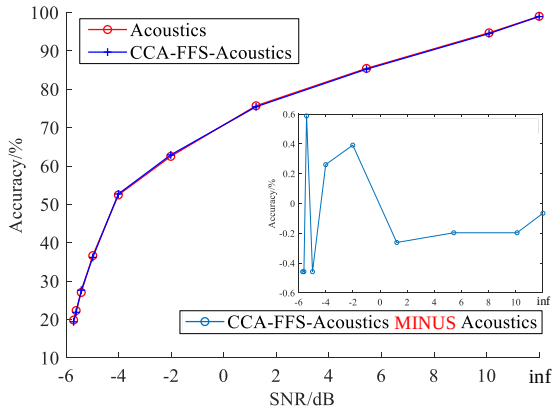


Figure 4. Recognition accuracies of the acoustics modality.

4.2.2. The Visual Modality

We only consider the visual modality in this subsection. The visual modality includes Visual and CCA-FFS-Visual. Just like the acoustics modality, multi-class recognition accuracies of Visual and CCA-FFS-Visual under different peak signal-to-noise ratio (PSNR[3]) are depicted in Figure 5. Their comparison is also depicted in Figure 5. Similarly, "marginal correlations" of the acoustics modality on the visual modality are represented by W_y^T in Equation (6). With the increasing of PSNR, the recognition accuracies of Visual and CCA-FFS-Visual both improve, just like the acoustics modality. However, as depicted in Figure 5, under the situation of PSNR from 10 to 15 dB, the recognition accuracies of CCA-FFS-Visual are much higher than that of Visual, especially under the situation of when PSNR is 10.9821 dB (up to 10%), which means that there is an essential difference between the recognition performance of Visual and CCA-FFS-Visual. One of the explanations of the essential difference is that as in Equation (6), when the correlations between the acoustics modality and the visual modality are maximized, the "marginal correlations" provided from the acoustics modality can make CCA-FFS-Visual become more discriminative. In this case, we can draw an important conclusion that by using CCA-based strategy to maximize the correlations between the acoustics modality and the visual modality, even though only one modality is used for recognition, CCA-FFS-Visual is still able to use the additional "marginal correlations" provided from the acoustics modality to improve the recognition performance.

---

[2] To obtain the samples of different SNR, different amplitudes of Gaussian white noise are added to the original samples. The "inf" in Figure 4 means that no noise is added to the original samples.

[3] We also add different Gaussian white noise to the original pictures to obtain the pictures of different PSNR, the "inf" also means that no noise is added to the original pictures.

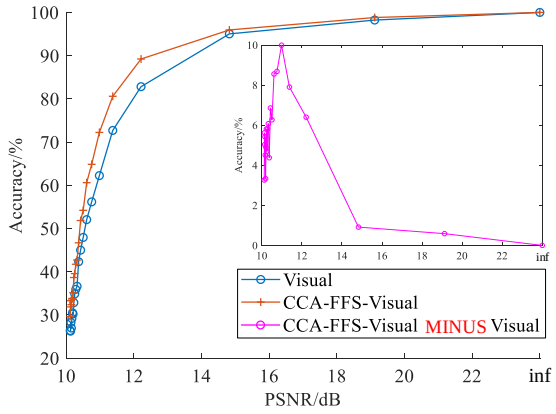
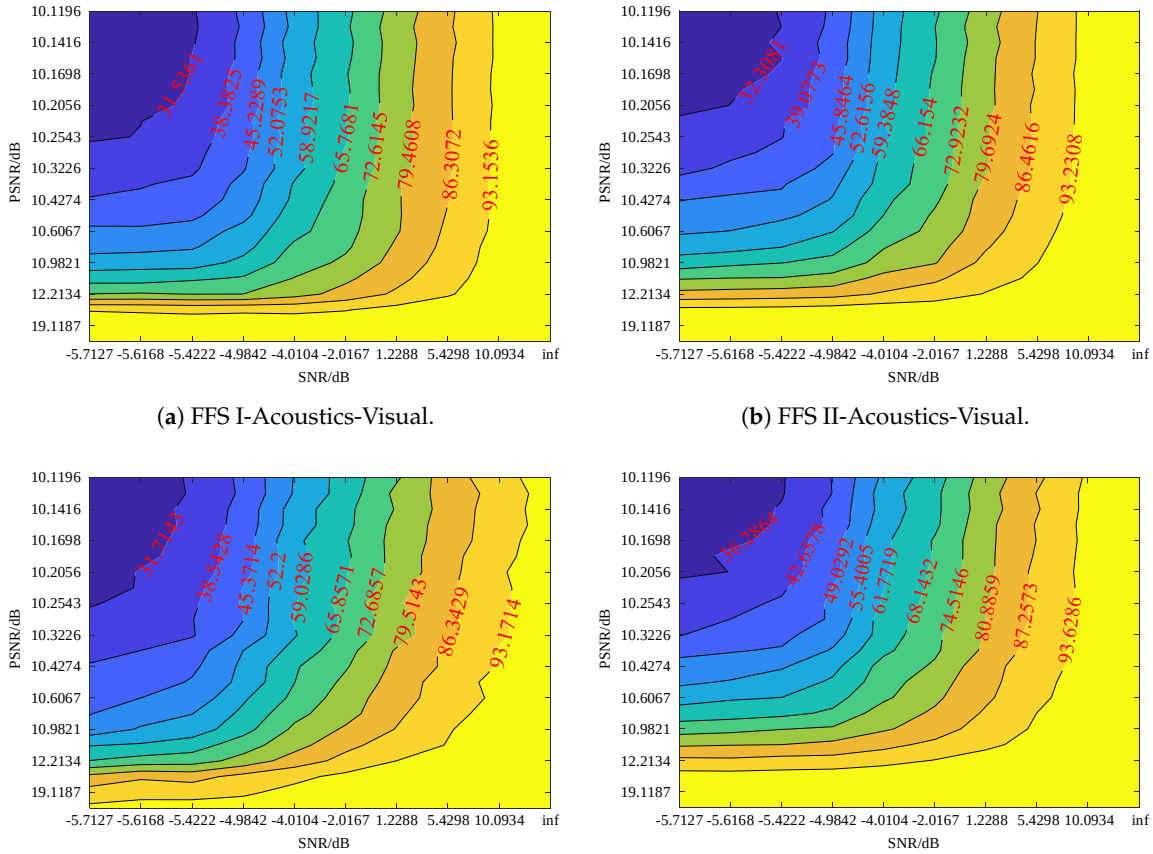


Figure 5. Recognition accuracies of the visual modality.

4.3. Multi-Modalities Consideration

In this case, we simultaneously consider the acoustics modality and the visual modality: FFS I-Acoustics-Visual, FFS II-Acoustics-Visual, CCA-FFS I-Acoustics-Visual, and CCA-FFS II-Acoustics-Visual. Two-dimensional multi-class recognition accuracies under different SNR and PSNR are depicted in Figure 6.



(**a**) FFS I-Acoustics-Visual.

(**b**) FFS II-Acoustics-Visual.

(**c**) CCA-FFS I-Acoustics-Visual.

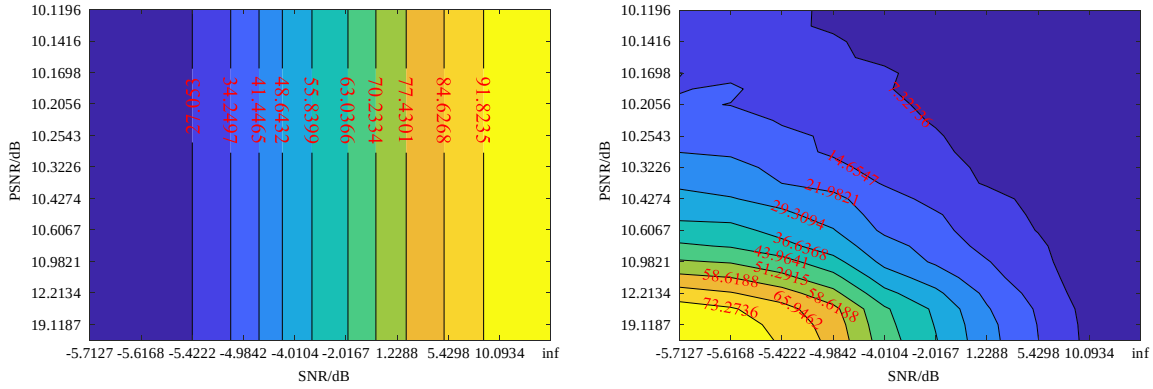
(**d**) CCA-FFS II-Acoustics-Visual.

Figure 6. Two-dimensional recognition accuracies (%) of multi-modalities.

As shown in Figure 6, no matter which multi-modalities, the recognition accuracies improve with the increasing of SNR or PSNR. From Figure 6, we can also know that CCA-FFS II-Acoustics-Visual achieves the best two-dimensional multi-class recognition accuracies. In CCA-FFS I-Acoustics-Visual

and CCA-FFS II-Acoustics-Visual, the CCA not only allows one modality to use the additional "marginal correlations" provided from the other modality, but also allows information to flow more readily between the acoustics modality and the visual modality, and by combining these two modalities, the recognition accuracies are improved. The dimension of combined feature space of CCA-FFS II-Acoustics-Visual is 64, while that of CCA-FFS I-Acoustics-Visual is 32, thus more information are provided and recognition accuracies of CCA-FFS II-Acoustics-Visual are higher than that of CCA-FFS I-Acoustics-Visual. In the following experiments, we focus on CCA-FFS II-Acoustics-Visual.

We further illustrate the advantages of multi-modalities over single-modality. The comparison between Acoustics and FFS II-Acoustics-Visual is depicted in Figure 7.



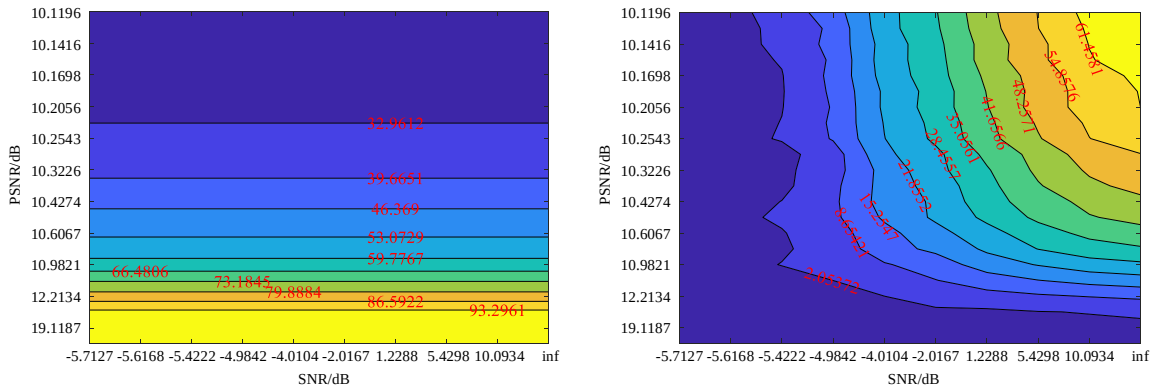
(**a**) Two-dimensional recognition accuracies of Acoustics.

(**b**) FFS II-Acoustics-Visual *MINUS* Acoustics.

Figure 7. The comparison between Acoustics and FFS II-Acoustics-Visual.

As shown in Figure 7, the recognition accuracies of FFS II-Acoustics-Visual are significantly higher than that of Acoustics under whatever SNR or PSNR. From Figure 7b we can know that especially in low SNR, when the acoustics modality can not provide any information for recognition, the visual modality of FFS II-Acoustics-Visual still can independently support the recognition.

Similar to Figure 7, the comparison between Visual and FFS II-Acoustics-Visual is depicted in Figure 8.



(**a**) Two-dimensioanl recognition accuracies of Visual.

(**b**) FFS II-Acoustics-Visual *MINUS* Visual.

Figure 8. The comparison between Visual and FFS II-Acoustics-Visual.

As shown in Figure 8, similarly, the recognition accuracies of FFS II-Acoustics-Visual are significantly higher than that of the Visual under whatever SNR or PSNR, especially in low PSNR,

when the visual modality can not provide any information for recognition. In this case, we can draw a conclusion that by using the CCA-based strategy to combine the two different modalities, when one modality is missing or can not provide any information for the recognition, the other modality is still able to independently support the recognition. This conclusion confirms our expectation, because the joining of the two modalities can provide much more information than any single-modality, and by using the CCA-based strategy to maximize the correlations between the two single-modalities, each single-modality is able to use the "marginal correlations" provided from the other modality to simultaneously make the information become far more discriminative.

4.4. Joint Representation

In this subsection, we consider representations of single-modality and joint representations of multi-modalities under different SNR and PSNR. T-distribution stochastic neighbor embedding (t-SNE) [21] is used to visualize the representations. The visualization technique t-SNE can map high-dimensional representations into 2 dimensions and show the distributions of the high-dimensional representations. The representations of two single-modality and joint representations of multi-modalities under several SNR and PSNR are depicted in Figure 9.

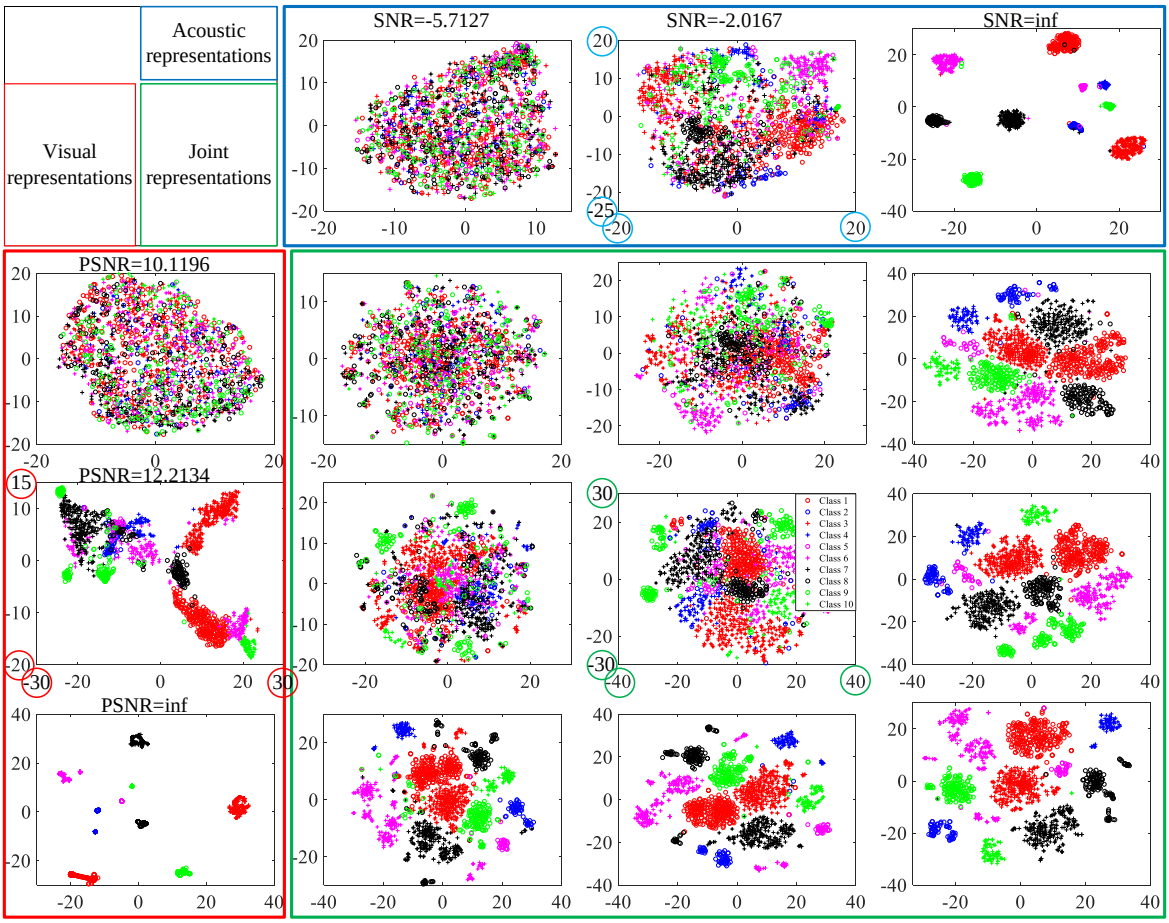


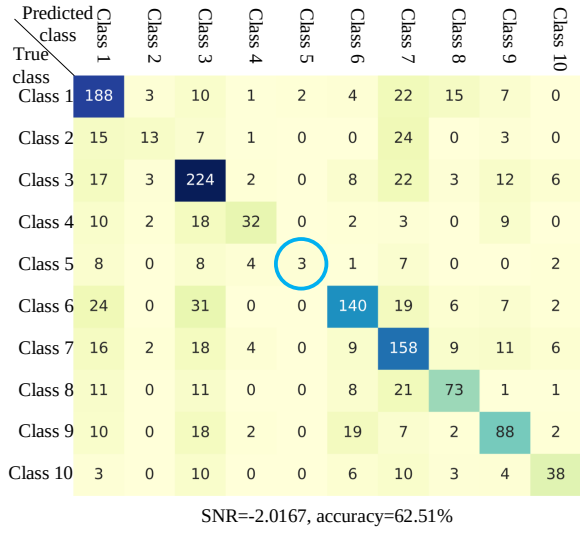
Figure 9. Representations of single-modality and joint representations of multi-modalities.

As shown in Figure 9, the acoustics representations inside each class become more concentrated and between each class become more separated with the increasing of SNR, and a similar trend can be seen from the visual representations but with the increasing of the PSNR. Considering the joint representations, the distance between each class tends to increase more than that of any of the single-modalities, which can improve multi-class recognition accuracies, especially under the situation of when SNR is -2.0167 dB and PSNR is 12.2134 dB.

4.5. Joint Recognition

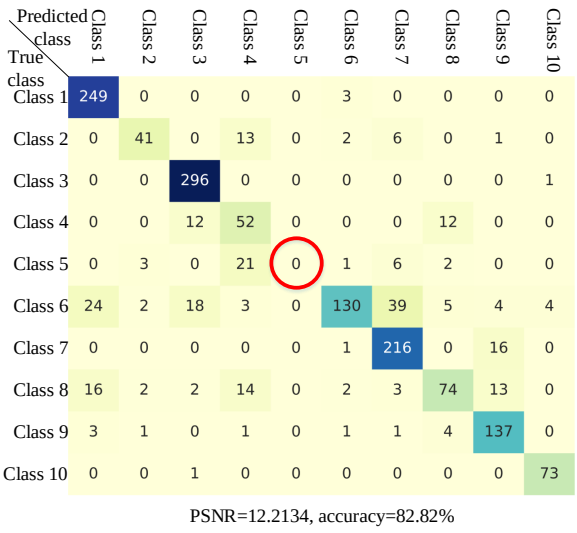
In this subsection, the joint recognition of multi-modalities is considered. Confusion matrices of single-modality and multi-modalities under a certain value of SNR and PSNR is depicted in Figure 10.

It can be clearly seen from Figure 10 that the recognition accuracy of FFS II-Acoustics-Visual is higher than any single-modality, and what is the most important is that Class 5, which can not be truly predicted by any of the single-modality, is truly predicted by the joint recognition of FFS II-Acoustics-Visual, which indicates that multi-modalities can simultaneously utilize the information provided from both single-modalities, in order to improve recognition accuracy. This is not simply using the information of one modality to compensate the other, but instead *generating* more discriminative information simultaneously from two single-modalities, in other words, strong-discriminative information can be generated from weak-discriminative information provided from two single-modalities.



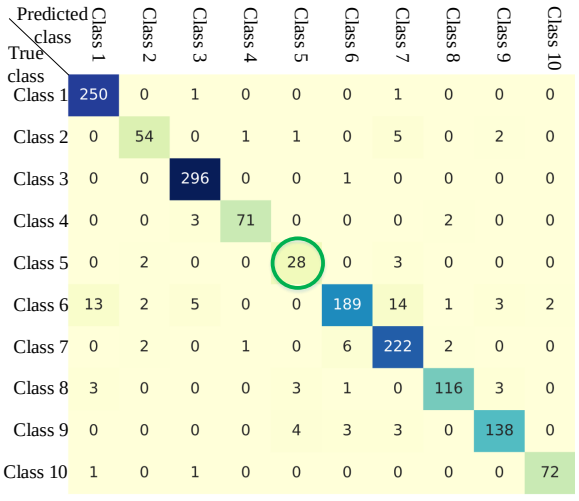
SNR=-2.0167, accuracy=62.51%

(**a**) Confusion matrix of Acoustics.



PSNR=12.2134, accuracy=82.82%

(**b**) Confusion matrix of Visual.



SNR=-2.0167, PSNR=12.2134, accuracy=89.75%

(**c**) Confusion matrix of FFS II-Acoustics-Visual.

Figure 10. Confusion matrices of single-modality and multi-modalities.

4.6. Overall Comparison

This section compares recognition accuracies of single-modality and multi-modalities in all, as well as a traditional method for ship-radiated noise recognition.

4.6.1. The Acoustics Modality

In Figure 11, we depict multi-class recognition accuracies of Acoustics and FFS II-Acoustics-Visual of different PSNR.

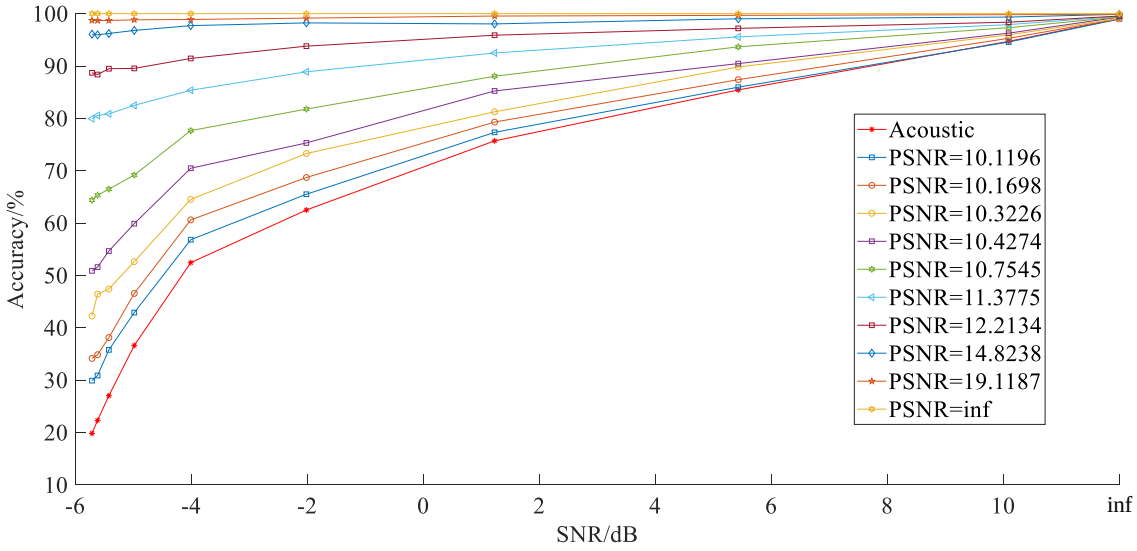


Figure 11. Recognition accuracies of Acoustics and FFS II-Acoustics-Visual of different peak signal-to-noise ratio (PSNR).

Considering the multi-modalities, with aid of the visual modality, FFS II-Acoustics-Visual have better recognition accuracies than Acoustics under any SNR. As long as the visual modality can provide enough information for the recognition, FFS II-Acoustics-Visual can have a good enough recognition performance even when the acoustics modality is under a very low SNR. This is due to the fact that in very low SNR, when the acoustic modality can not provide information for ship recognition, the recognition of the ship will depend on the visual modality. To ensure a good enough recognition of ship, at least one modality should be provided.

4.6.2. The Visual Modality

Similar to last subsection, we depict multi-class recognition accuracies of Visual and FFS II-Acoustics-Visual of different SNR in Figure 12.

Figure 12 shows that FFS II-Acoustics-Visual have much better multi-class recognition accuracies than Visual under whatever PSNR. Considering the multi-modalities, the result is very similar to that in Figure 11: Once the acoustics modality can support the recognition, FFS II-Acoustics-Visual is able to have a good enough recognition performance even under very low PSNR of when the visual modality no longer provides any information for the recognition.

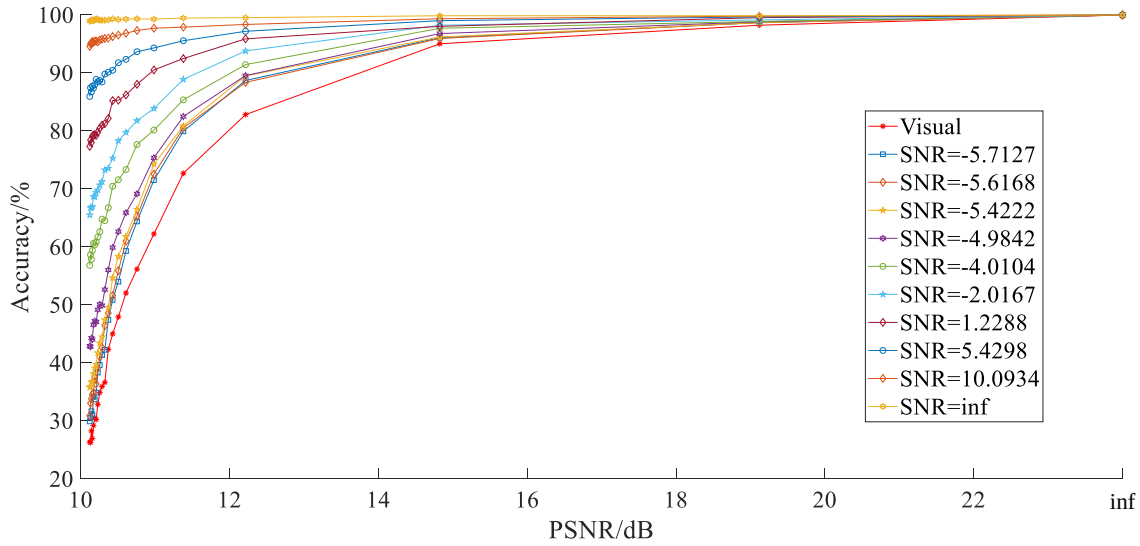


Figure 12. Recognition accuracies of Visual and FFS II-Acoustics-Visual of different SNR.

4.6.3. The Baseline

We also compare FFS II-Acoustics-Visual with a baseline (MFCCs + GMMs) provided in [9] which adopts the same ship-radiated noise database. We draw receiver operating characteristics (ROC) curves with area under curve (AUC) values to further evaluate the recognition performance, which are depicted in Figure 13. The one that has the larger AUC value has the better recognition performance.

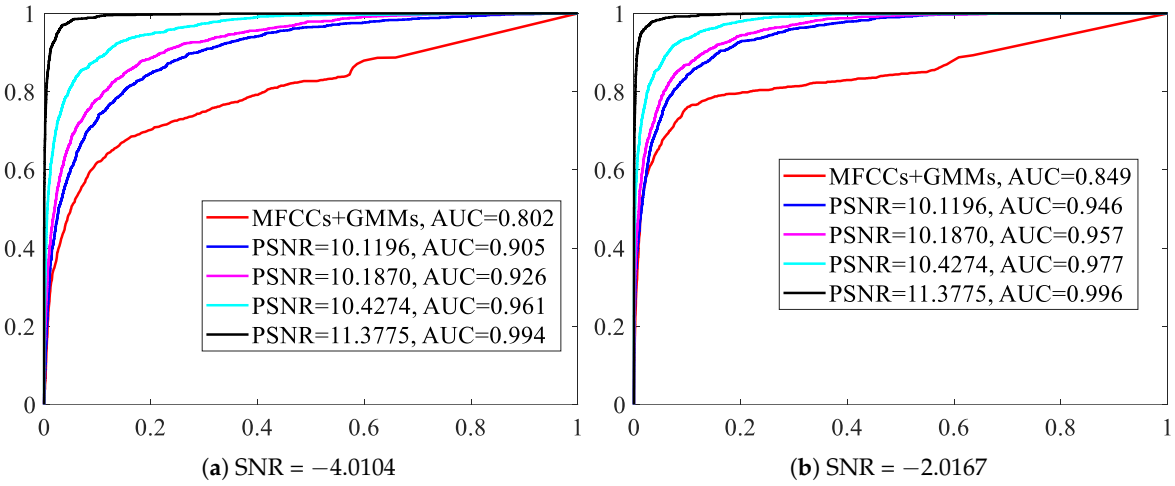


Figure 13. Receiver operating characteristics (ROC) curves of FFS II-Acoustics-Visual and Mel-frequency cepstral coefficients (MFCCs) + Gaussian mixture models (GMMs).

As shown in Figure 13, with the increasing PSNR, the recognition performance of FFS II-Acoustics-Visual improves. For the visual modality, the higher PSNR will provide more discriminative information for ship recognition, thus a better recognition performance is achieved. Figure 13 also shows that with the aid of the visual modality, the recognition performance of FFS II-Acoustics-Visual is better than that of the MFCCs + GMMs under no matter which PSNR. To some extent, an additional modality can indeed improve the recognition performance.

5. Conclusions

Inspired by AVSR based on multimodal-DL methods, we also introduce multimodal-DL methods for joint representation and recognition of the ship-radiated noise. We propose the multimodal-CNNs

framework specially designed for the recognition of ship-radiated noise. We propose the CCA-based strategy to build the joint representation and recognition simultaneously on the acoustics modality and the visual modality. Using the CCA-based strategy, strong-discriminative information can be generated from weak-discriminative information provided from any single-modality. Experimental results show that the multimodal-DL methods for joint recognition of ship-radiated noise have much more advantages over the DL methods, and much more superiorities than the traditional methods. In multimodal-DL methods, as long as one of the single-modalities still provides information for the recognition, joint recognition will have a good enough recognition performance when compared to the recognition of any single-modality.

6. Future Work

Preliminarily, a CCA-based strategy is utilized to build a discriminative joint representation and recognition of the acoustics modality and the visual modality. Actually, a more robust and promising strategy can be explored to build the joint representation and recognition, such as the powerful multimodal-DBMs on which our future study will focus. In addition, more powerful and state-of-art deep neural networks can be explored to represent the multiple modalities.

Author Contributions: F.Y. and X.K. contributed to the idea of the incentive mechanism and designed the algorithms; F.Y. and E.C. were responsible for some parts of the theoretical analysis. X.K. designed and performed the experiments. F.Y., E.C. and X.K. contributed with the structure, content and the paper check.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61571377, 61771412 and 61871336) and the Fundamental Research Funds for the Central Universities (Grant No. 20720180068).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Srivastava, N.; Salakhutdinov, R. Multimodal Learning with Deep Boltzmann Machines. *J. Mach. Learn. Res.* **2014**, *15*, 2949–2980.
2. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, Washington, DC, USA, 28 June–2 July 2011; pp. 689–696.
3. Swietojanski, P.; Ghoshal, A.; Renals, S. Convolutional neural networks for distant speech recognition. *IEEE Signal Process. Lett.* **2014**, *21*, 1120–1124.
4. Cao, X.; Togneri, R.; Zhang, X.; Yu, Y. Convolutional Neural Network With Second-Order Pooling for Underwater Target Classification. *IEEE Sens. J.* **2018**, *19*, 3058–3066. [CrossRef]
5. Haghighat, M.; Abdel-Mottaleb, M.; Alhalabi, W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1984–1996. [CrossRef]
6. Sun, Q.S.; Zeng, S.G.; Liu, Y.; Heng, P.A.; Xia, D.S. A new method of feature fusion and its application in image recognition. *Pattern Recognit.* **2005**, *38*, 2437–2448. [CrossRef]
7. Liu, J.; He, Y.; Liu, Z.; Xiong, Y. Underwater Target Recognition Based on Line Spectrum and Support Vector Machine. In *2014 International Conference on Mechatronics, Control and Electronic Engineering (MCE-14)*; Atlantis Press: Paris, France, 2014. [CrossRef]
8. Wei, X. On feature extraction of ship radiated noise using 11/2 d spectrum and principal components analysis. In Proceedings of the 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Hong Kong, China, 5–8 August 2016; pp. 1–4.
9. Santos-Domínguez, D.; Torres-Guijarro, S.; Cardenal-López, A.; Pena-Gimenez, A. ShipsEar: An underwater vessel noise database. *Appl. Acoust.* **2016**, *113*, 64–69. [CrossRef]
10. Zhang, L.; Wu, D.; Han, X.; Zhu, Z. Feature extraction of underwater target signal using Mel frequency cepstrum coefficients based on acoustic vector sensor. *J. Sens.* **2016**, *2016*. [CrossRef]
11. Meng, Q.; Yang, S. A wave structure based method for recognition of marine acoustic target signals. *J. Acoust. Soc. Am.* **2015**, *137*, 2242. [CrossRef]

12. Azimi-Sadjadi, M.R.; Yao, D.; Huang, Q.; Dobeck, G.J. Underwater target classification using wavelet packets and neural networks. *IEEE Trans. Neural Netw.* **2000**, *11*, 784–794. [CrossRef] [PubMed]
13. Averbuch, A.; Zheludev, V.; Neittaanmäki, P.; Wartiainen, P.; Huoman, K.; Janson, K. Acoustic detection and classification of river boats. *Appl. Acoust.* **2011**, *72*, 22–34. [CrossRef]
14. Wang, S.; Zeng, X. Robust underwater noise targets classification using auditory inspired time–frequency analysis. *Appl. Acoust.* **2014**, *78*, 68–76. [CrossRef]
15. Cao, X.; Zhang, X.; Yu, Y.; Niu, L. Deep learning-based recognition of underwater target. In Proceedings of the 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, China, 16–18 October 2016; pp. 89–93.
16. Ke, X.; Yuan, F.; Cheng, E. Underwater Acoustic Target Recognition Based on Supervised Feature-Separation Algorithm. *Sensors* **2018**, *18*, 4318. [CrossRef] [PubMed]
17. Yang, H.; Shen, S.; Yao, X.; Sheng, M.; Wang, C. Competitive deep-belief networks for underwater acoustic target recognition. *Sensors* **2018**, *18*, 952. [CrossRef] [PubMed]
18. Kamal, S.; Mohammed, S.K.; Pillai, P.S.; Supriya, M. Deep learning architectures for underwater target recognition. In Proceedings of the 2013 Ocean Electronics (SYMPOL), Kochi, India, 23–25 October 2013; pp. 48–54.
19. Mroueh, Y.; Marcheret, E.; Goel, V. Deep multimodal learning for audio-visual speech recognition. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 2130–2134.
20. Tuma, M.; Rørbech, V.; Prior, M.K.; Igel, C. Integrated optimization of long-range underwater signal detection, feature extraction, and classification for nuclear treaty monitoring. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3649–3659. [CrossRef]
21. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).