KULLBACK–LEIBLER DIVERGENCE FREQUENCY WARPING SCALE FOR ACOUSTIC SCENE CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

Yuhong Yang[1,3(✉)], *Huiyu Zhang*[1], *Weiping Tu*[1], *Haojun Ai*[2,3(✉)], *Linjun Cai*[1], *Ruimin Hu*[1], and *Fei Xiang*[4]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Hubei, China
[2] School of Cyber Science and Engineering, Wuhan University, Hubei, China
[3] Collaborative Innovation Center of Geospatial Technology, China
[4] Xiaomi, Beijing, China

ABSTRACT

Most of current best performing Acoustic Scene Classification (ASC) systems utilize Mel scale spectrograms with Convolutional Neural Networks (CNNs). Mel scale is a common way to suit frequency warping of human ears, with strict decreasing frequency resolution on low to high frequency range. However, we find that significant frequency bins are located at mid to high frequency range for some acoustic scenes, such as travelling by bus, tram or train. In this paper, we show that a better frequency warping scale for ASC can be automatically learned from raw spectrograms, using Kullback-Leibler (KL) divergence scale. Our KL scale spectrograms with CNN method is evaluated on two public ASC datasets. The results show that we outperform the Mel scale method on both datasets. In addition, we also employ a Conditional Generative Adversarial Nets (Conditional-GAN) model for data augmentation, to prevent overfitting problem and allow further improvements on ASC.

Index Terms— KL divergence, Frequency Warping, Acoustic Scene Classification, CNN, Conditional-GAN

1. INTRODUCTION AND MOTIVATION

Acoustic Scene Classification (ASC) allow devices to sense and understand the surrounding environment, using audio signals. ASC has been applied to mobile terminals and wearable devices for customized services. For example, the wheelchair will automatically switch between two service modes according to whether the environment is indoor or outdoor [1]. The mobile phone will perceive the surroundings, and adjust settings to provide a better user experience.

The growing interest in ASC has motived the IEEE AASP Detection and Classification of Acoustic Scenes and Events (DCASE) challenge in 2013, 2016, 2017, and 2018 [2]-[5]. In the latest DCASE challenges, most of the best performing ASC systems utilize spectrogram with CNNs, or combined CNNs with other models. Six among them input Mel scale spectrogram into CNNs [6]-[11], one input Constant-Q-Transform (CQT) spectrogram [12].

Both Mel and CQT scales are static for all systems. CQT uses a series of constant logarithmically spaced filters, which is well suited for music data with harmonic structures[13]. However, since acoustic environments recordings are mostly non-music signals, CQT features are uncommon in CNN for ASC. Mel scale uses the perceptual based filter bank judged by listeners, to suit frequency warping of human ears[14], which is signal independent and perhaps the most widely used scale for spectrogram down sampling. Both scales are associated with strict decreasing frequency resolution from low to high frequency range.

However, we find that for some acoustic scenes, such as travelling by bus, tram or train, the most discriminative frequency bins are located at mid to high frequency range. When above mentioned scales are employed for spectrogram down sampling, they tend to focus on the low frequency range, which emphasize the less significant signal content, while may loss the significant details. This may introduce other factor of variation into training and inference. We hope to find a better frequency warping scale, to help more accurately discriminate the acoustic scenes.

Solomon Kullback and Richard Leibler introduced the Kullback–Leibler (KL) divergence, to measure how one probability distribution diverges from a second probability distribution [15]. And Bisot uses a KL divergence scale for NMF decomposition in ASC[16], to find the bases of discrimination for the input of Deep Neural Network (DNN), without intention to preserve the time frequency pattern of the raw spectrogram.

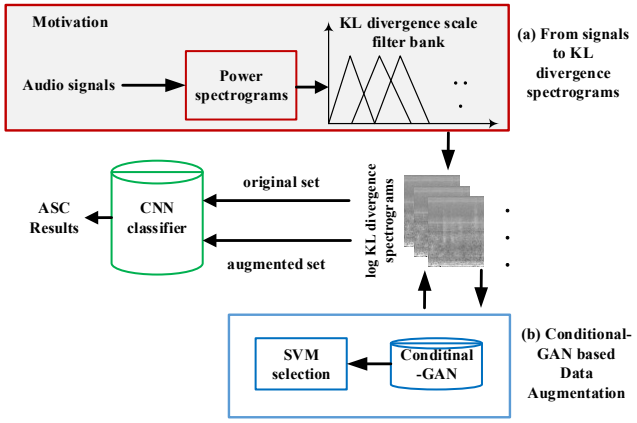

Fig. 1. Proposed KL divergence based CNN framework for ASC with data augmentation

In this paper, we proposed a KL divergence based frequency warping scale for ASC, which can be automatically learned from raw spectrograms and preserve the original time-frequency pattern. Then we use KL divergence scale filter bank to obtain down sampled spectrogram. Our KL scale spectrogram with CNN method is evaluated on DCASE 2016 and 2017 datasets available (DCASE 2018 is not included, since the ground truth labels for evaluation dataset are not released yet). The results show that we outperform the Mel and CQT scale method on both datasets. The

proposed KL system is also compared to the recent published work on ASC, including CNN based [9][11] and others [16]-[19].

In addition, we employed a Conditional Generative Adversarial Nets (Conditional-GAN) [20] model for data augmentation, to prevent overfitting problem and allow further improvements on KL based ASC. Finally, we summarized our work.

2. FRAMEWORK DESCRIPTION

The proposed KL divergence scale approach with data augmentation is depicted in Fig. 1. Details of the KL divergence filter bank and Conditional-GAN based data augmentation are illustrated as follows.

2.1. From Signals to KL Divergence Spectrograms

The audio signals are transformed into raw power spectrograms. Then we learn a KL divergence scale from the spectrograms in the training set, based on the one-vs.-rest KL divergences for each class. The learning diagram is illustrated in Fig.2.

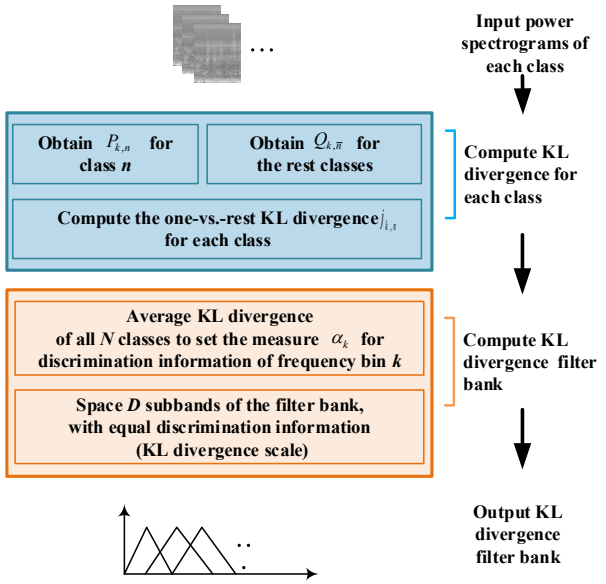

Fig. 2. Learning KL divergence scale from raw spectrograms

2.1.1. Compute KL divergence for each class

We take the normalized logarithm power spectrograms as inputs to compute the KL divergence. The one-vs.-rest Jensen–Shannon divergence [21] $j_{k,n}$ based on KL divergence for each class n are calculated at frequency bin k as:

$$j_{k,n} = \frac{1}{2}[D_{KL}(P_{k,n} \| \frac{P_{k,n}+Q_{k,\bar{n}}}{2}) + D_{KL}(Q_{k,\bar{n}} \| \frac{P_{k,n}+Q_{k,\bar{n}}}{2})] \quad (1)$$
$$k = 0...L-1, \quad n = 1...N$$

Where $P_{k,n}$ and $Q_{k,\bar{n}}$ refer to the numerical probabilistic distribution of the normalized logarithmic power for class n and the rest classes at frequency bin k, respectively. For discrete probability distributions, P and Q, the KL divergence from Q to P is defined as:

$$D_{KL}(P \| Q) = \sum_x P(x) \cdot (\log P(x) - \log Q(x)) \quad (2)$$

According to $j_{k,n}$, we find that for some acoustic scenes, such as travelling by bus, tram or train, the most discriminative frequency bins are located at mid to high frequency range.

2.1.2. Compute KL-divergence filter bank

For N classes, we average the N divergence values $j_{k,n}$ to obtain KL divergence measure J_k at frequency bin k:

$$J_k = \frac{1}{N}\sum_{n=1}^{N} j_{k,n}, \quad k = 0...L-1, \quad n = 1...N \quad (3)$$

To get the KL divergence filter bank with M overlapped subbands, we first have to choose the bandwidth interval $[f(m-1), f(m+1)]$ $(m = 1...M)$. Since $f(0) = 0$, and $f(M+1) = L-1$, M additional frequency bins $f(m)$ are calculated as:

$$\sum_{k=f(m-1)}^{f(m)} J_k = \frac{1}{M+1}\sum_{k=0}^{L-1} J_k, \quad m = 1...M \quad (4)$$

Now we can create KL divergence filter bank $H_m(k)$ as Mel scale filter bank:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m-1) \le k \le f(m) \\ 0, & k > f(m+1) \end{cases} \quad (5)$$
$$m = 1...M, \quad k = 0...L-1$$

2.2. Conditional-GAN Based Data Augmentation

To improve ASC performance, we use Conditional-GAN[20] to generated KL spectrogram extracted from the development set. Since it is not clear whether every sample generated by Conditional-GAN would have equal impact in classification performance, we select hard to classified samples by the SVM hyper plane for each class as in [8]. Finally, the augmented datasets are about two times of the original datasets, which contain both real samples from the development set and fake samples from generators selected by SVM. They are input into the CNN network for training and evaluation.

3. EXPERIMENTAL SETUP

3.1. Dataset and Evaluation Metrics

Our experiments were evaluated on the DCASE2016 and 2017 dataset, containing 15 different acoustic scenes (included in Table 2). The total amount of recordings in DCASE challenge were partitioned into development and evaluation subsets. The development set was further partitioned into four folds of training and testing sets, provided by the DCASE challenge organizer. Each audio scene has the same number of audio segments.

The development set in dcase2017 has 4680 segments of audio files, evaluation set has 1620 segments, and each segment of audio duration is 10 seconds. The development set in dcase2016 has 1170 segments, evaluation set has 390 segments, and each segment of audio duration is 30 seconds. Audio signals are all sampled at 44.1kHz and down mixed into mono for further tests.

We train our model using a cross validation on the 4-fold development set. The final results were obtained by averaging over all 4 folds, and evaluate on the evaluation set.

3.2. Spectrograms Down Sampling

All signals are pre-emphasized with a factor of 0.95. Then, the pre-emphasis signal is framed, windowed and Fourier transformed to obtain the power spectrogram, wherein the frame length is 40ms with 50% hop size, the window function is the Hamming window, and the number of Fourier transform bins is 2048.

The raw power spectrograms are down sampled by KL filter banks with 128 subbands, and the output KL spectrogram are logarithmically normalized. CQT and Mel spectrograms are obtained as KL, with filter banks of 128 subbands, using the cqt and mel function provided by the librosa library[22].

Finally, the normalized logarithm spectrograms (KL, Mel or CQT) are split into short sequences of 100 frames each. Therefore, the inputs of the CNN or Conditional-GAN's discriminator are gray images of size 128×100.

3.3. CNN model architecture

Since we mainly concerned the influence of CNN inputs, the employed deep CNN architecture simply followed [10] with some modifications. The specifications are shown in Fig. 3.

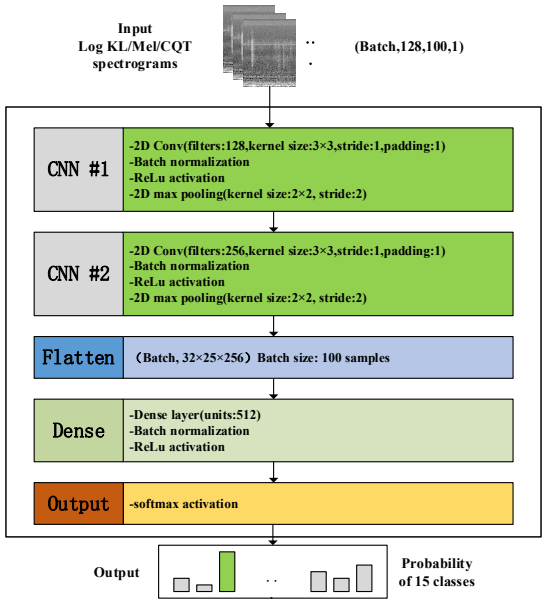

Fig. 3. The CNN model specifications (Weight and bias initial method: Xavier initializer; Optimizer: Adam)

3.4. Conditional-GAN model architecture

We followed [23] with some modifications to build our deep Conditional-GAN architecture, as shown in Fig.4.

After samples generation, we use a soft margin SVM to select the discriminative samples, the regularization factor in SVM's loss function is 0.5.

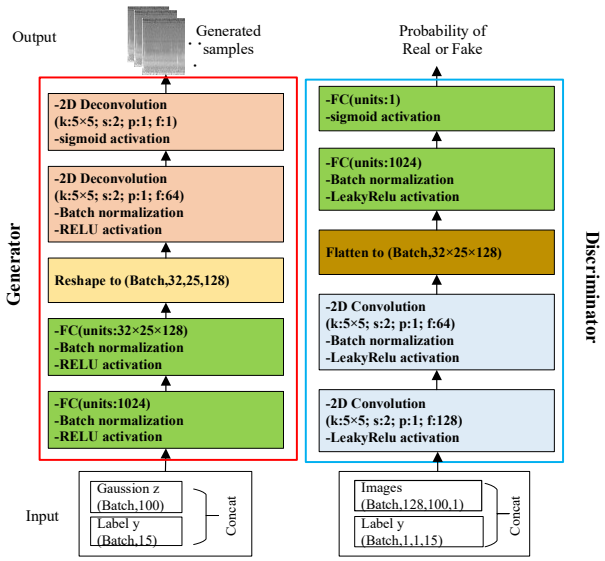

Fig. 4. The CGAN model specifications. (we use the audio label y as the conditional information, with batch size:100 samples, where k, f, s, and p refer to kernel size, filters, stride, and padding respectively; Weight and bias initial method: Xavier initializer; Optimizer: Adam.)

4. RESULTS

4.1. Influence of Spectrogram Down Sampling

In this section we propose to study the effects of input KL, Mel or CQT spectrogram into the CNN based ASC. The results on DCASE 2016 and DCASE 2017 are shown in Table 1, with accuracy on each four-fold cross-validation and evaluation datasets. Average accuracy of four-fold cross-validation is also presented for fair comparison.

Table 1. Comparison of experimental results with KL divergence Mel scale, and CQT with 4-fold cross-validation on the development dataset, and accuracy on the evaluation dataset.

Dateset	Feat.	Cross-Validation Acc(%)				Ave Acc (%)	Eva Acc (%)
		1	2	3	4		
DCASE 2017	CQT	80.7	75.8	78.9	75.1	77.6	64.9
	Mel	82.3	79.6	80.0	79.3	80.3	67.5
	KL	**84.4**	**81.0**	**83.8**	**80.8**	**82.5**	**69.3**
DCASE 2016	CQT	86.6	85.9	86.4	86.2	86.3	82.3
	Mel	88.1	86.3	87.6	86.8	87.2	84.9
	KL	**88.6**	**88.3**	**87.6**	**88.9**	**88.4**	**86.7**

The results show that KL spectrogram inputs outperform Mel and CQT on both data sets. Auto learning from the raw spectrogram of ASC datasets helps to improve the accuracy both in the cross-validation and evaluation. And Mel spectrogram outperform CQT on both datasets. This can attribute to the Mel scale associated with

the frequency warping of human ears, which is more general when compared with CQT for music signals in ASC systems.

In Table 2, performance degradations are found for some scenes such as Beach, Café, and Home, but we do find that scenes such as travelling by Bus, Car, Train, and Tram are improved as expected. And average accuracy of KL outperforms Mel approach.

Table 2. Class-wise accuracy comparison between Mel and KL on the evaluation set (highlight numbers indicate higher accuracy for the scene class).

Acc.	Dcase2017		Dcase2016	
(%)	Mel	KL	Mel	KL
Beach	**25.9**	25.0	**89.3**	88.5
Bus	43.5	**54.6**	82.3	**92.3**
Café	**74.1**	61.1	**71.5**	61.5
Car	81.5	**84.3**	100.0	100.0
City	93.5	93.5	**92.3**	88.5
Forest	96.3	**97.2**	100.0	100.0
Groce.	**76.9**	75.9	88.5	88.5
Home	**93.5**	79.6	**88.5**	84.6
Lib.	43.5	43.5	49.2	**65.4**
Metro.	74.1	**89.8**	84.6	**88.5**
Office	77.8	**86.1**	100.0	100.0
Park	36.1	**38.0**	**96.2**	92.3
Resid.	66.7	**76.9**	73.1	**84.6**
Train	78.7	78.7	57.7	**65.4**
Tram	50.0	**55.6**	100.0	100.0
Avg.	67.5	**69.3**	84.9	**86.7**

4.2. Influence of data augmentation

In order to further improve the ASC performance, we use the Conditional-GAN introduced in section 2.2 to generated, and use SVM to selected the KL spectrograms respectively.

The augmented datasets are about twice the size of the original ones. We compare the average ASC accuracies over 4-folds and the evaluation accuracy for the CNN classifier trained with or without data augmentation. The results are shown in the upper two rows of Table 3.

After data augmentation, the accuracies of KL approach are improved on the evaluation set for both datasets, with 2.8% on DCASE 2016, and with 4.3% on DCASE 2017. Conditional-GAN data augmentation can help to alleviate the overfitting to a certain extent and improve the performance of the model.

4.3. Comparison with Other ASC Systems

Table 3 compares the proposed KL approach to the latest published papers on the DCASE 2016 and DCASE 2017 datasets with mono inputs. The results include average cross validation accuracy (short for Dev in Table 3) over all folds, being the measure for most previous works. Evaluation accuracy (short for Eva in Table 3) is also included.

For DCASE 2016 dataset, we include two top CNN system in DCASE 2016 challenge with mono Mel spectrogram inputs[10][11]. Other published ASC results using DNN[16][19] and random forest[17] are also included for comparison. It can be seen from the table that the proposed KL approaches with or without data augmentation outperform other systems both in cross validation and evaluation.

Table 3. Comparison with other systems (DA: Data Augmentation)

System	D2016 Acc(%) Dev/Eva	D2017 Acc(%) Dev/Eva	Features	Classifier
Proposed KL	88.4/86.7	82.5/ 69.3	KL	CNN
Proposed KL with DA	**94.3**/ **89.5**	**89.8**/ **73.6**	KL	CNN
DCASE16 baseline[3]	72.5/ 77.2	-	MFCC	GMM
Valenti[10]	79.0/ 86.2	-	Mel	CNN
Lee[11]	83.1/ 84.6	-	Mel	CNN
Mun[19]	86.3/ -	-	MFCC	DNN
Abidin[17]	85.0/ -	-	LBP/HOG	Random Forest
Bisot[16]	82.5/ -	-	KL-NMF	DNN
DCASE17 baseline[4]	-	74.8/ 61.0	Mel	MLP
Piczak[24]		82.4/ 70.6	Spectrogram	CNN
Jimenez[18]	-	78.6/ -	emoco-nf	SVM

For DCASE 2017 dataset, we include one top CNN system in DCASE 2017 challenge with mono inputs and data augmentation[24], and one published paper with ASC results [18] on DCASE 2017. It can be seen from Table 3 that the proposed KL approach with data augmentation is also the best performing system for both measures.

5. CONCLUSIONS

This paper proposed a novel approach to obtain KL divergence scale spectrograms for CNN based acoustic scene classification. By learning KL divergence scale from raw spectrograms in ASC training datasets, we can get the specified spectrogram with better discrimination. Our algorithm can be readily extended to other audio and speech processing systems for spectrogram down sampling, simply by modifying the KL divergence scale on the associated training datasets. And we verified that Conditional-GAN based data augmentation can help to prevent overfitting problem and allow further improvements on ASC. Our results clearly show the proposed approach achieved state of the art performance on two public datasets with mono inputs. However, recent work can achieve further improvements on combining the CNN with other models[8], applying CNN ensembles[9], or using binaural inputs[8][12]. These approaches will be considered for our future work.

6. ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China (2016YFB0502204), National Natural Science Foundation of China (61702472, 61671335).

7. REFERENCES

[1] S. Chu, S. Narayanan, C. C. Kuo, and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," in *2006 IEEE International Conference on Multimedia and Expo*, IEEE, 2006: 885-888.

[2] DCASE2013, http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge, last accessed 2018/07/11.

[3] DCASE2016,http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification, last accessed 2018/07/11.

[4] DCASE2017, http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification, last accessed 2018/07/11.

[5] DCASE2018, http://dcase.community/challenge2018/index, last accessed 2018/10/25.

[6] D. Matthias, "Acoustic Scene Classification with Fully Convolutional Neural Networks and I-Vectors," in *2018 IEEE AASP Detection and Classification of Acoustic Scenes and Events,* 2018

[7] S. Yuma, "Acoustic Scene Classification by Ensemble of Spectrograms Based on Adaptive Temporal Divisions," in 2018 *IEEE AASP Detection and Classification of Acoustic Scenes and Events*, 2018.

[8] S. Mun, S. Park, D. Han, K. Hanseok, "Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane," in *DCASE2017 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2017.

[9] Y. Han, P. Jeongsoo, and L. Kyogu. "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *DCASE2017 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2017.

[10] M. Valenti, A. Diment, G. Parascandolo, et al., "DCASE 2016 acoustic scene classification using convolutional neural networks," in *DCASE2016 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2016.

[11] Y. Han, K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," in *DCASE2016 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2016.

[12] Z. Weiping, Y. Jiantao, X. Xiaotao, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *DCASE2017 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2017.

[13] J C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, 1998, 89(1):425--434.

[14] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in 1983 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 1983, 8: 93-96.

[15] S. Kullback, R.A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22(1), 79-86(1951).

[16] V. Bisot, R. Serizel, S. Essid, et al., "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, 25(6): 1216-1229.

[17] S. Abidin, X. Xia, R. Togneri et al., "Local Binary Pattern with Random Forest for Acoustic Scene Classification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2018: 1-6.

[18] A. Jiménez, B. Elizalde, and B. Raj, "Acoustic Scene Classification Using Discrete Random Hashing for Laplacian Kernel Machines," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018: 146-150.

[19] S. Mun, S. Shon, W. Kim, et al., "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 2017: 796-800.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv*:1411.1784, 2014.

[21] M. Naghshvar, T. Javidi, and M. Wigger, Extrinsic Jensen–Shannon Divergence: Applications to Variable-Length Coding[J]. *IEEE Transactions on Information Theory*, 2013, 61(4):2148-2164.

[22] LibROSA, https://librosa.github.io/librosa/index.html, last accessed 2018/10/29.

[23] C. Xi, D. Yuan, H. Rein, et al., "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets," *arXiv preprint arXiv*: 1606.03657, 2016.

[24] K. J. Piczak, "The details that matter: Frequency resolution of spectrograms in acoustic scene classification," in *DCASE2017 Workshop on Detection and Classification of Acoustic Scenes and Events,* 2017.