Bird sounds classification by combining PNCC and robust Mel-log filter bank features

PNCC와 robust Mel-log filter bank 특징을 결합한 조류 울음소리 분류

Alzahra Badi,[1] Kyungdeuk Ko,[1] and Hanseok Ko[1][†]

(알자흐라 바디,[1][†] 고경득,[1] 고한석[1][†])

[1]School of Electrical Engineering, Korea University
(Received November 7, 2018; accepted January 25, 2019)

ABSTRACT: In this paper, combining features is proposed as a way to enhance the classification accuracy of sounds under noisy environments using the CNN (Convolutional Neural Network) structure. A robust log Mel-filter bank using Wiener filter and PNCCs (Power Normalized Cepstral Coefficients) are extracted to form a 2-dimensional feature that is used as input to the CNN structure. An *ebird* database is used to classify 43 types of bird species in their natural environment. To evaluate the performance of the combined features under noisy environments, the database is augmented with 3 types of noise under 4 different SNRs (Signal to Noise Ratios) (20 dB, 10 dB, 5 dB, 0 dB). The combined feature is compared to the log Mel-filter bank with and without incorporating the Wiener filter and the PNCCs. The combined feature is shown to outperform the other mentioned features under clean environments with a 1.34 % increase in overall average accuracy. Additionally, the accuracy under noisy environments at the 4 SNR levels is increased by 1.06 % and 0.65 % for shop and schoolyard noise backgrounds, respectively.

Keywords: Acoustic event recognition, Environmental sound classification, CNN (Convolutional Neural Network), Weiner filter, PNCCs (Power Normalized Cepstral Coefficients)

PACS numbers: 43.60.Bf, 43.60.Uv

초 록: 본 논문에서는 합성곱 신경망(Convolutional Neural Network, CNN) 구조를 이용하여 잡음 환경에서 음향 신호를 분류할 때, 인식률을 높이는 결합 특징을 제안한다. 반면, Wiener filter를 이용한 강인한 log Mel-filter bank와 PNCCs(Power Normalized Cepstral Coefficients)는 CNN 구조의 입력으로 사용되는 2차원 특징을 형성하기 위해 추출됐다. 자연환경에서 43종의 조류 울음소리를 포함한 *ebird* 데이터베이스는 분류 실험을 위해 사용됐다. 잡음 환경 에서 결합 특징의 성능을 평가하기 위해 ebird 데이터베이스를 3종류의 잡음을 이용하여 4개의 다른 SNR (Signal to Noise Ratio)(20 dB, 10 dB, 5 dB, 0 dB)로 합성했다. 결합 특징은 Wiener filter를 적용한 log-Mel filter bank, 적용하 지 않은 log-Mel filter bank, 그리고 PNCC와 성능을 비교했다. 결합 특징은 잡음이 없는 환경에서 1.34 % 인식률 향상 으로 다른 특징에 비해 높은 성능을 보였다. 추가적으로, 4단계 SNR의 잡음 환경에서 인식률은 shop 잡음 환경과 schoolyard 잡음 환경에서 각각 1.06 %, 0.65 % 향상했다.

핵심용어: 음향 이벤트 인식, 환경음 인식, CNN(Convolutional Neural Network), Weiner filter, PNCCs(Power Normalized Cepstral Coefficients)

I. Introduction

Recently with the rapid development of AI (Artificial

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
Engineering Building Room 419, School of Electrical Engineering, Korea University Anam Campus 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

Intelligence), environmental sound classification has been a research focus of many applications from surveillance,[1] to environmental monitoring.[2-5] AI has been applied to the detection and classification of certain animal species through acoustic sounds to provide information used by

applications monitoring biodiversity and endangered species preservation. These environmental monitoring applications classify the sounds of animals ranging from marine life[3,4] to bats[5] or birds.[6,7]

To capture the information contained in acoustic event classification,[8,9] sound and automatic speech recognition[10–12] have used several features, such Mel-log filter bank and MFCCs (Mel Frequency Cepstral Coefficients), which are based on the human auditory system. Even though such features have dominated in acoustic applications, and continue to do so, their performance decreases with the amount of noise present in the signal. Therefore, several attempts to suppress noise without distorting the acoustic signal have been proposed using stationary noise suppression mechanisms to achieve system performance under environmental noisy conditions. These methods include the use of the Wiener filter, PNCCs (Power Normalized Cepstral Coefficients)[13] and RCGCCs (Robust Compressive Gamma-chirp filter bank Cepstral Coefficients).[14] While such features do improve performance, their effectiveness still depends on the types or characteristics of the noise present, such as whether it is non-stationary noise. Interestingly, recent research has shown that using combinations of features can boost overall system performance. For example, References [15], [16] show that combining MFCC features with PNCC features, which are both robust to noise, makes the overall system, which can then learn from and exploit both features, perform better under noisy environments.

In this work, we investigate combined robust feature performance under noisy environments using both PNCCs, and the log Mel-filter bank integrated with the Wiener filter, which both work with stationary noise but use different stationary noise suppression algorithms. Firstly, the Wiener filter is combined with the log Mel-filter bank to suppress stationary noise which is an optimal causal system where the power spectrum of the noise is estimated based on the present and previous signal frames to provide an estimate of the clean signal based on the mean square error. Integrating the Wiener filter will allow use of a log Mel-filter bank and also suppress noise using an optimal

estimation.[17] Secondly, the PNCCs are employed as they achieve good performance under noisy environments by applying the medium duration power bias subtraction algorithm, which is based on asymmetric filtering and temporal masking effects. Additionally, the PNCC uses the power law nonlinearity with gammatone filter instead of the log and the triangular filter used by the log Mel-filter bank.[13] Given the different characteristics of these features, we expect that combining them through a convolutional layer that enables the system to extract features from both will boost system performance for bird sound classification under both clean and noisy environments. The proposed extraction method is explained in more detail in section II and the experimental work and discussion in section III.

II. Proposed Method

This section describes the proposed method consisting of two principal stages, the feature extraction stage and the classification stage, which uses the AlexNet structure[18] to classify bird species. In the feature extraction stage, both the log Mel filter bank with Wiener filter and PNCC noise estimations and their characteristics will be elucidated in more detail as well as the combining procedure of both features that feeds into the AlexNet network. Fig. 1 illustrates the feature extraction steps of both features.
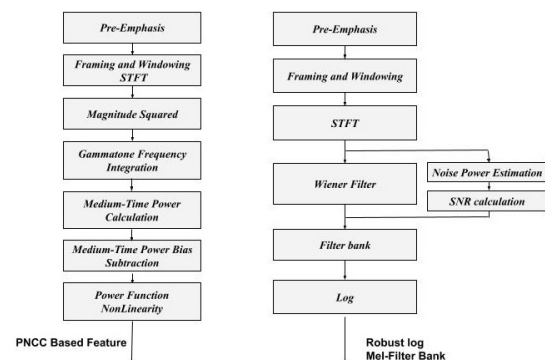

Fig. 1. Feature extraction for robust log Mel-filter bank and PNCC feature.

2.1 Feature Extraction

2.1.1 Log Mel-Filter Bank with Wiener filter

Mel-filter bank is been widely used and it is designed based on the human auditory system. In order to extract the enhanced or robust features, the optimal Wiener filter is used after obtaining the spectrum of the signal following the work and recommendation in Reference [19]. As we assuming an additive noise as in Reference [20], where the $Y(f)$, $S(f)$, $W(f)$ denotes the observed, desired, and noise signal in frequency domain. Also, the Wiener filter output or transfer function can be expressed in the frequency domain as in Eq. (2), clean or enhanced signal can be demonstrated using Eq. (3).

$$Y(f) = S(f) + N(f). \tag{1}$$

$$H_{wiener}(f) = \frac{P_{ss}(f)}{P_{ss}(f) + P_{nn}(f)}. \tag{2}$$

$$P_{ss}(f) = H_{wiener}(f) \ P_{YY}(s). \tag{3}$$

The P_{ss} and P_{nn} are the desired observed and noise power spectra, respectively. Therefore, the estimated signal can be obtained through Eq. (3) by first estimating the noise signal spectrum using the MMSE-SPP (Minimum Mean Square Eerror - soft Speech Presence Probability) algorithm proposed in Reference [21], which does not require bias correction or VAD (Voice Activity Detection) as the MMSE-based noise spectrum estimate approach would.

2.1.2 PNCCs (Power Normalized Cepstral Coefficients)

The PNCC features are designed for stationary noise suppression, a goal which can also be achieved using the log Mel-filter bank. However, the PNCC algorithm has three main differences. It uses the gammatone filter to replace the filter bank and applies power-law nonlinearity based on the hearing of Steven's law power[22] as this approach leads to close to zero output when the input is too small, in contrast to the log function that is used for the

Mel-log filter bank features. In addition, it performs a peak power normalization using the medium-time power bias subtraction method proposed in Reference [13]. This method does not estimate the noise power from non-speech frames, but instead removes the power bias that has information about the level of the background noise as assumed, and uses the ratio of the arithmetic mean to the geometric mean when determining the power bias. The final power can be obtained through Eqs. (4) - (6).

$$Q(i,j) = \frac{1}{2M+1} \sum_{j'=j-M}^{j+M} P(i,j'). \tag{4}$$

$$Q'(i,j|B(i)) = \max\big(Q(i,j) - B(i), d_0 Q(i,j)\big). \tag{5}$$

$$\tilde{P}(i,j) = \left(\frac{1}{2N+1} \sum_{i'=\max(1-N,1)}^{\min(i+N,C)} \frac{Q'(i,j|B(i))}{Q(i,j)}\right) P(i,j). \tag{6}$$

Hence, i is the channel index, j is the frame index, C is total number of channels, $P(i,j)$ is the power observed in a single analysis frame, $Q(i,j)$ is the average power of 7 frames ($M=3$), and the normalized power ($Q'(i,j|B(i))$) can be obtained by subtracting the level of background excitation ($B(i)$). In addition, the d_0 in (5) is constant to prevent the normalized power from becoming negative. Therefore, the final power ($\tilde{P}(i,j)$) can be obtained though Eq. (6). For more details refer to Reference [13].

2.1.3 Features Combination using Convlutional Network

This stage aims to find the best feature representation of data based the available features map using the convolutional neural network. Where, the combined features are obtained by concatenated both extracted features and create a 3-dimensional features (features, frames, 2) in order to be fed into a convolutional layer to get only one feature mapping representation ($k=1$) of both extracted feature type (features, frames, 1), whereas the mapping is done with a trainable filter size of (l, l, q) known as filter bank (W) that connect the feature map ($q=1$)
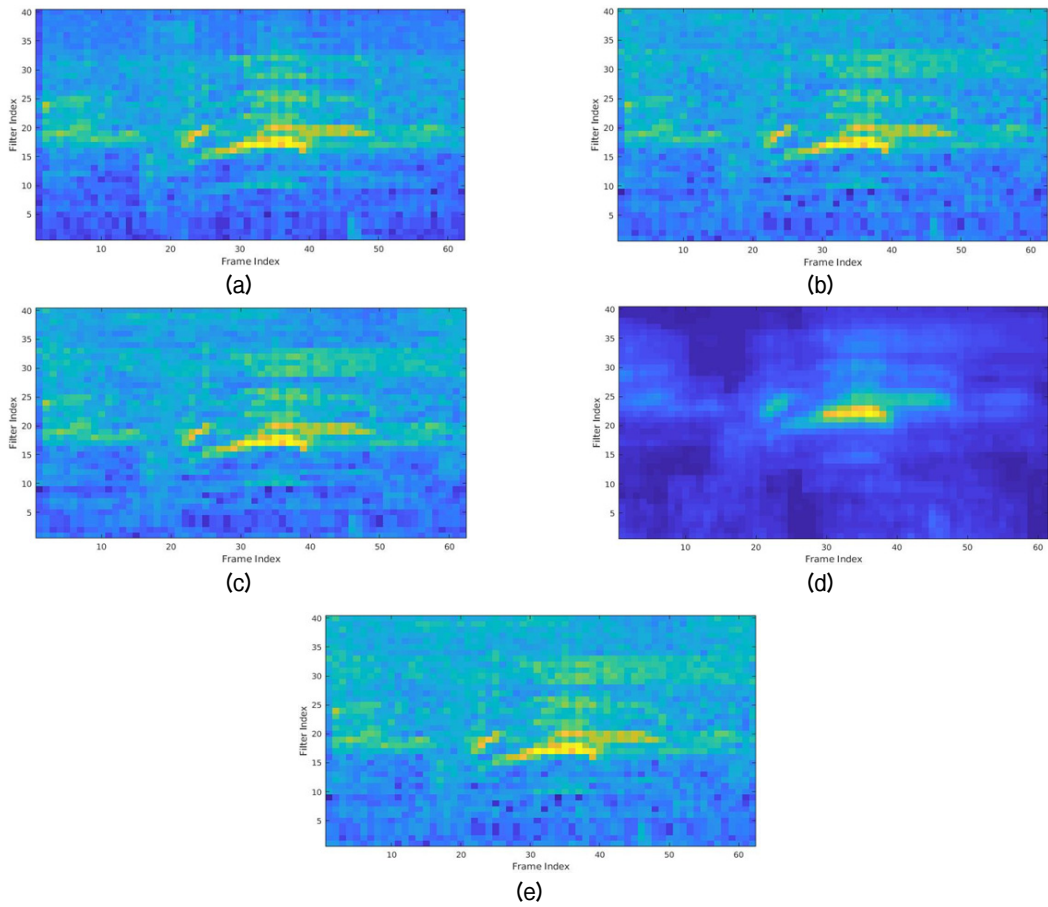
Fig 2. Power spectral density of the extracted features of (a) log mel-filter bank of clean signal, (b) log mel-filter bank under shop noise (10 dB), (c) robust log mel-filter bank under shop noise (10 dB), (d) PNCC under shop noise, and (e) combine features under noise (10 dB).
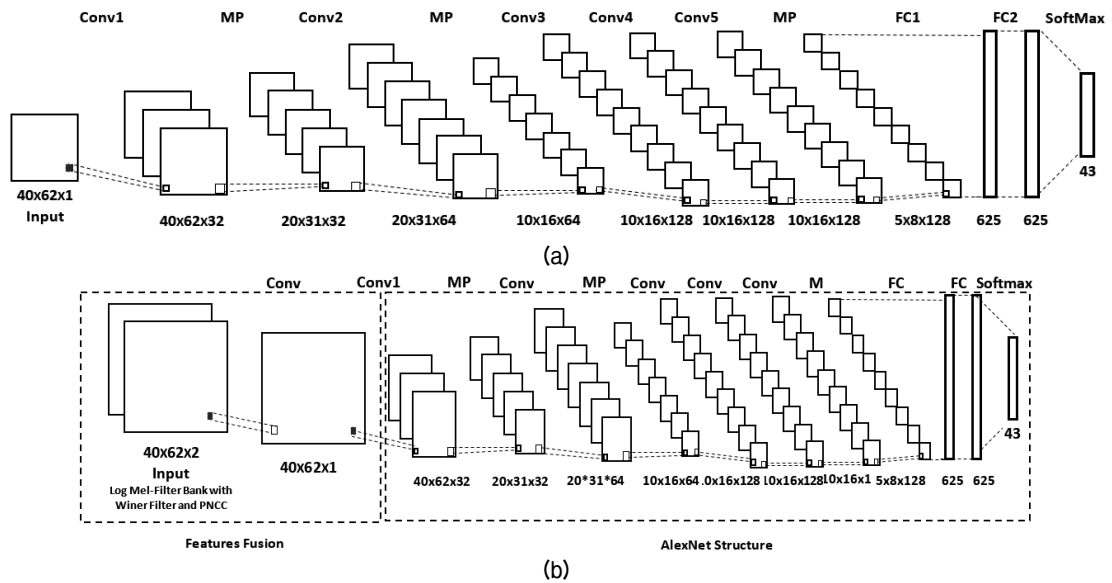

Fig. 3. Convolutional neural network architecture for classifying using (a) single feature (baseline) (b) combined features.

of the input to the layer into the output or desired map (k) which can be obtained using Eq. (7).

$$z^s = \sum_{t=1}^{q} W_t^{k*} x^{i_{s,t}} + b_s,\qquad(7)$$

where * is 2-dim convolution operator, b is bias and x^i denotes the features in each feature map.[23] This will allow the network to assign a certain weight or importance to each feature in the feature map for both the PNCCs and robust log Mel-filter bank. This leads to a better system performance by highlighting the most significant features that can post our system. Fig. 2 shows the power spectral density of the extracted features from the robust log Mel-filter bank, PNCCs and combined features under shop noise with an SNR (Signal to Noise Ratio) of 10 dB. Looking at the results we can observe that the combined extracted features seem to sum and exploit both features by keeping more information that are similar to the log Mel-filter bank and reducing a fraction of the added noise.

2.2 ALexNet

Based on References [18], [24], the CNN structure is composed of 5 convolutional layers with a 3×3 filter size and a stride of 1, and 3 max pooling layers. Max-pooling layers with a filter size of 2×2 and a stride of 2 were used after the first, second and the fifth convolutional layers. Three fully connected layers were used following the 5 convolutional layers separated by a dropout as used in References [24]. Fig. 3 illustrates the CNN structure for classification using the single feature and combined features.

III. Experimental Work

3.1 Dataset

The bird sounds were collected from *https://ebird.org* at a sample rate of 44.1 kHz, and down sampled to 16-bit resolution and segmented using the EPD (End-Point Detection) method[25] based on the procedure in Reference [7] for data processing and segmentation. This resulted in 0.719 s audio samples of 43 bird class species. The database was augmented with 3 types of background noise (café, shop and schoolyard) at 4 levels of SNR (20 dB, 10 dB, 5 dB and 0 dB) using ADDNOISE MATLAB.[26]

3.2 Experimental Setting

The baseline features are the log Mel-filter bank with and without the Wiener Filter, where the Wiener filter is applied after taking the spectrum of the signal, as well as the PNCCs. The features were extracted from 0.719 s audio sounds and resulted in 40 features with 62 frames per audio. These features were used as inputs to the AlexNet structure for classification as explained in Fig. 3(b). When using the proposed method, both the robust log Mel-filter bank (log Mel-filter bank with Wiener filter) and PNCCs were concatenated to get 3-D ($40 \times 62 \times 2$) arrays which were used as inputs to the network to extract one 3-D feature map, as explained in section 2 and illustrated in Fig. 3, using a filter size of (1, 1, 2). Moreover, the database was divided into training and test sets using 5-fold cross validation, resulting in 5 sets. Each feature vector was normalized by the mean and variance before being fed into the AlexNet

Table 1. Bird species classification in clean environment.

Folds	Log mel-filter bank (FB)	Log mel-filter bank with Wiener filter (FB&WF)	PNCCs	Combined feature (PNCCs, FB & WF)
Fold1	79.23 %	79.92 %	79.25 %	82.65 %
Fold2	81.24 %	80.41 %	79.78 %	82.00 %
Fold3	81.10 %	79.57 %	79.55 %	81.40 %
Fold4	79.57 %	80.27 %	78.47 %	81.68 %
Fold5	82.14 %	77.13 %	78.93 %	82.26 %
Avg. accuracy (%)	80.66 %	79.46 %	79.20 %	82.00 %

Table 2. Bird species classification accuracy in noisy environment.

Noise type	SNR (dB)	Log mel-filter bank (FB) avg. accuracy	Log mel-filter bank with Wiener filter (FB&WF) avg. accuracy	PNCC avg. accuracy	Combined feature (PNCC, FB&WF) avg. accuracy
Café	20	74.05 %	70.99 %	76.98 %	77.92 %
	10	61.96 %	57.32 %	68.99 %	69.02 %
	5	51.46 %	48.37 %	61.09 %	60.26 %
	0	38.03 %	37.85 %	48.85 %	47.46 %
Shop	20	77.38 %	74.62 %	76.97 %	78.31 %
	10	68.50 %	64.82 %	69.15 %	70.00 %
	5	60.10 %	57.17 %	61.66 %	62.75 %
	0	47.28 %	46.58 %	49.23 %	50.20 %
School yard	20	70.55 %	70.53 %	73.46 %	74.83 %
	10	50.60 %	53.33 %	56.77 %	57.49 %
	5	39.48 %	41.51 %	44.85 %	45.16 %
	0	29.03 %	29.42 %	31.67 %	31.87 %

for training. In the AlexNet structure, a dropout of 0.5 was used to reduce the effect of overfitting after the fully-connected layer, and ReLU (Rectified Linear Unit) activation function were used with batch size of 500.

3.3 Results and Discussion

Table 1 illustrates the overall accuracy for all 43 given classes for all 5 folds. The combined features gave the highest average accuracy of 82 % among all the features meaning an increase in average accuracy of 1.34 % over the Mel log-filter bank. This was followed by accuracies of 80.66 % and 79.46 % for the Mel log-filter bank and robust log Mel-filter bank, respectively. The PNCCs demonstrated the lowest average accuracy of 79.20 % with the clean data.

In addition, these features' performances were tested on the augmented data and the results are presented in Table 2. Notably, the combined feature almost always outperforms the others except the cases of the café augmented data with SNRs of 5 dB and 0 dB where the PNCCs give the highest accuracy rates of 61.09 % and 48.85 %, respectively. The overall percentage increases in accuracy with shop and schoolyard background noise, respectively, under all SNRs were 1.06 % and 0.65 %. However, there was a decrease in average overall accuracy over all SNR levels for the café background noise type of 0.31 %. Fig. 4 shows the confusion
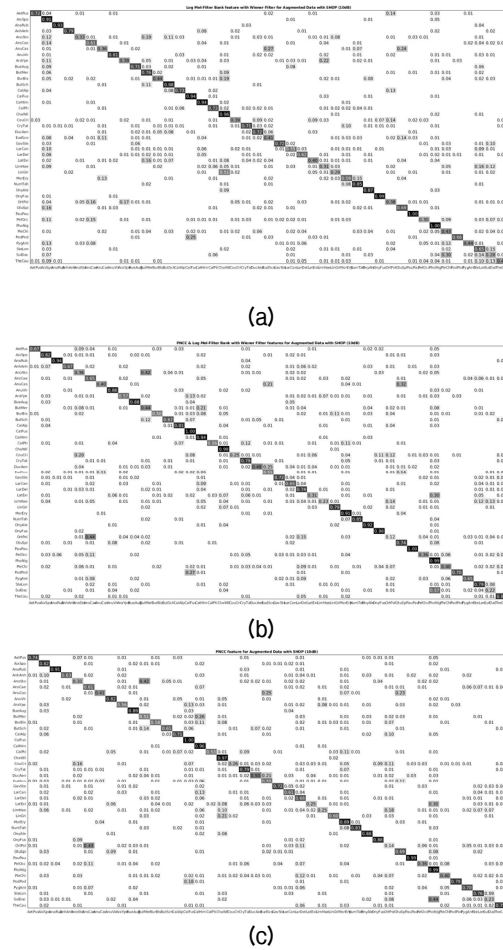

(a)

(b)

(c)

Fig. 4. The confusion matrix for noise with type shop under SNR of 10 dB for (a) robust log mel-filter bank (b) PNCC features and (c) combined features.

matrix of the combined features, PNCCs and robust log Mel-filter bank under shop background noise with an SNR of 10 dB.

However, the background noise is non-stationary noise and both features focus on a noise suppression feature that works with stationary noise. Therefore, the performance of the log Mel filter bank with the Wiener filter inevitably cannot give a better performance across all noise types and PNCCs enhanced the performance most effectively under the noisy environment. By combining these two features, the network seems to highlight and signify the most relevant features contained in both and therefore led to an increase in overall accuracy under both clean and noisy environments.

IV. Conclusions

In this work, we proposed 3-D combined robust features feeding to a convolutional layer followed by AlexNet for acoustic sound classification. The database of ebird.com was used to test the performance of the log Mel-filter bank, PNCC and combined features structure. The combined features structure outperformed the single features in most cases yielding an increase in accuracy by 1.34 % in a clean environment and 1.06 % and 0.65 % under shop and schoolyard background noise environments, respectively when averaged over 4 different SNR levels. These results illustrated that extracting these features from the combined ones using a convolutional neural network can exploit the complementarity of the combined features by making them accessible to the classification step, and thereby increase the recognition rate.

Acknowledgement

References

1. R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust., 158-161 (2005).

2. J. Salamon and J. P. Bello, "Deep convolutional neural networks and sata augmentation for environmental sound classification," IEEE Signal Process. Lett., **24**, 279-283 (2017).

3. F. R. González-Hernández, L. P. Sánchez-Fernández, S. Suárez-Guerra, and L. A. Sánchez-Pérez, "Marine mammal sound classification based on a parallel recognition model and octave analysis," Applied Acoustics, **119**, 17-28 (2017).

4. M. Malfante, J. Mars, M. D. Mura, C. Gervaise, J. I. Mars, and C. Gervaise, "Automatic fish sounds classification," J. Acoust. Soc. Am. **143**, 2834-2846 (2018).

5. O. M. Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, I. Pandourski, S. Parsons, J. Russ, A. Szodorary-Paradi, F. Szodoray-Paradi, E. Tilova, M. Girolami, G. Brostow, and K. E. Jones, "Bat detective-Deep learning tools for bat acoustic signal detection," PLoS Comput. Biol., **14**, e1005995 (2018).

6. F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach." J. Acoust. Soc. Am. **131**, 4640-4650 (2012).

7. K. Ko, S. Park, and H. Ko, "Convolutional feature vectors and support vector machine for animal sound classification," Proc. IEEE Eng. Med. Biol. Soc. 376-379 (2018).

8. R. Lu and Z. Duan, "Bidirectional Gru for sound event detection," Detection and Classification of Acoustic Scenes and Events (DCASE), (2017).

9. T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," Detection and Classification of Acoustic Scenes and Events (DCASE), (2016).

10. Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding," 2015 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2015, 167-174 (2016).

11. D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-End Attention-based large vocabulary speech recognition," Acoust. Speech Signal Process (ICASSP), 2016 IEEE Int. Conf., 4945-4949 (2016).

12. A. Ahmed, Y. Hifny, K. Shaalan, and S. Toral, "Lexicon

free Arabic speech recognition recipe," Advances in Intelligent Systems and Computing, **533**, 147-159 (2017).

13. C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 28-31 (2009).

14. M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," Digit. Signal Process., **29**, 147-157 (2014).

15. M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification," 4th Int. Work. Biometrics Forensics (IWBF), 1-6 (2016).

16. S. Park, S. Mun, Y. Lee, D. K. Han, and H. Ko, "Analysis acoustic features for acoustic scene classification and score fusion of multi-classification systems applied to DCASE 2016 challenge," arXiv Prepr. arXiv1807.04970 (2018).

17. N. Upadhyay and R. K. Jaiswal, "Single channel speech enhancement: using Wiener filtering with recursive noise estimation," Procedia Comput. Sci., **84**, 22-30 (2016).

18. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in neural information processing systems, 1097-1105 (2012).

19. P. M. Chauhan and N. P. Desai, "Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using Wiener filter," Green Computing Communication and Electrical Engineering (ICGCCEE), 1-5 (2014).

20. S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation theory* (PTR Prentice-Hall, Englewood Cliffs, 1993), pp. 400-409.

21. T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA), 145-148 (2011).

22. S. S. Stevens, "On the psychological law," Psychological Review, **64**, 153 (1957).

23. L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," IEEE Geosci. Remote Sens. Mag., **4**, 22-40 (2016).

24. K. Ko, S. Park, and H. Ko, "Convolutional neural network based amphibian sound classification using covariance and modulogram" (in Korean), J. Acoust. Soc. Kr. **37**, 60-65 (2018).

25. J. Park, W. Kim, D. K. Han, and H. Ko, "Voice activity detection in noisy environments based on double-combined fourier transform and line fitting," Sci. World J., **2014**, e146040 (2014).

26. ITU-T, ITU-T P.56, *Objective Measurement of Active Speech Level*, 2011.

Profile

▸ **Alzahra Badi (알자흐라 바디)**

She received the B.S. degree in Electrical and Electronics Engineering from Benghazi University in 2013. In 2013, she joined the Tatweer Research and Development Company, where she worked in ICT and Prosthetics and Orthotics (P&O) projects. In 2014, she studied an exclusive business program at Judge Business School, Cambridge University. Since 2017, she has been in the M.S. course in Department of Electrical and Computer Engineering from Korea University. Her interests include acoustic and image signal processing, geophysics signal processing, and artificial intelligence.

▸ **Kyungdeuk Ko (고경득)**

He received the B.S. degree in Biomedical Engineering from Yonsei University in 2017. Since 2017, he has been in the M.S. course in Department of Electrical and Computer Engineering from Korea University. His interests include acoustic/image signal processing, biomedical signal/image processing and artificial intelligence.

▸ **Hanseok Ko (고한석)**

He received B.S. degree from Carnegie Mellon University, in 1982, M.S. degree from the Johns Hopkins University, in 1988, and Ph.D. degree from the CUA, in 1992, all in electrical engineering. At the onset of his career, he was with the WOL, Maryland, where his work involved signal and image processing. In March of 1995, he joined the faculty of the Department of Electronics and Computer Engineering at Korea University, where he is currently Professor. His professional interests include speech/image signal processing for pattern recognition, multimodal analysis, and intelligent data fusion