INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Towards Speech Robustness for Acoustic Scene Classification

Shuo Liu[1], *Andreas Triantafyllopoulos*[1,2], *Zhao Ren*[1], *Björn W. Schuller*[1,2,3]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]audEERING GmbH, Gilching, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

shuo.liu@informatik.uni-augsburg.de

Abstract

This work discusses the impact of human voice on acoustic scene classification (ASC) systems. Typically, such systems are trained and evaluated on data sets lacking human speech. We show experimentally that the addition of speech can be detrimental to system performance. Furthermore, we propose two alternative solutions to mitigate that effect in the context of deep neural networks (DNNs). We first utilise data augmentation to make the algorithm robust against the presence of human speech in the data. We also introduce a voice-suppression algorithm that removes human speech from audio recordings, and test the DNN classifier on those denoised samples. Experimental results show that both approaches reduce the negative effects of human voice in ASC systems. Compared to using data augmentation, applying voice suppression achieved better classification accuracy and managed to perform more stably for different speech intensity.

Index Terms: Acoustic scene classification, speech robustness, voice suppression, computational auditory scene analysis

1. Introduction

Acoustic scene classification (ASC) is the problem of classifying an audio sample to the type of environment in which it has been produced [1, 2, 3, 4]. It has been applied in the context of smartphones [5, 6, 7], robots [8, 9, 10], and wearable devices including medical equipment like hearing aids [11, 12]. Oftentimes, the goal of ASC is to enable adaptation with respect to the recognised environments.

In latest years, one of the most prominent scientific challenges for this topic is the Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). It contains tasks for scene classification, sound event detection and localisation, and sound source tagging [13, 14, 15]. Typically, the first task is concerned with the basic problem of ASC, with the challenge participants competing for the highest classification accuracy in assigning audio snippets to the environment they were recorded in.

Sakashita and Aono [16] used an ensemble of neural networks, each of them taking as input a log-Mel spectrum. This approach achieved 81 %, which was the best result of the 2018 edition of the challenge. Chen and Zhang [17] utilised a data augmentation scheme based on generative adversarial networks (GANs), to further improve the classification performance, and achieved an accuracy of 85 %, which was the top performance on the 2019 edition of the challenge. Both approaches are based on convolutional neural networks (CNNs), which have been shown to outperform other approaches for the ASC task [18, 19, 20].

All the ASC tasks of past DCASE challenges measure system performance only in clean conditions without the presence of human voice. That also allows the data set to conform to data privacy requirements. However, in real-life applications, ASC systems will most likely have to work in the presence of human speech, either because of the nature of the application itself, e. g., real-time denoising for human-to-human communication or human-computer interaction. In addition, contemporary and future privacy requirements might not allow constant monitoring of the surrounding environment, in which case the system will only have access to audio after a human has explicitly requested it to do so [21, 22]. Hence, in real-life situations the presence of speech will be the norm, and ASC systems will have to perform well under these more challenging conditions, too.

In this work, we investigate the effects of human voice, which in the context of ASC can be considered as noise, on two deep neural network (DNN) based classification algorithms. The first is the 2019 DCASE baseline model, and the second is an attentive atrous CNN used in our previous work [23]. Both algorithms are described in detail in Section 3. We used the official 2019 DCASE data set, and artificially added human speech from the Edinburgh noisy speech database [24]. Results show that both algorithms suffer from the presence of human speech.

We propose two alternative methodologies for overcoming this limitation. First, we introduce data augmentation as a means to expose the models to the conditions encountered at test-time. Our second approach is based on a voice-suppression frontend. We formulate voice-suppression as the exact opposite of speech enhancement, which aims to reduce the presence of environment sounds and enhance the human speech components in an audio recording [25, 26]. In contrast, we seek to reduce the presence of human speech components while preserving the signal characteristics necessary to perform the ASC task. The voice-suppression architecture is described in detail in Section 4.2.

The rest of the paper is organised as follows. In Section 2, we first introduce the data sets we use in this work. In Section 3, we present the architectures used for ASC and discuss their robustness to human voice. In Section 4, we introduce the voice-suppression architecture.

2. Data sets

As our ASC data, we use the data set provided for the SubTask-1A of the 2019 DCASE challenge. The data set consists of 40 hours of clean stereo recordings, each recorded with the same device for 10 known acoustic scene classes in 10 different cities. The data is split into 9 185 segments in the training set and 4 185 in the test set, with each segment having a duration of 10 seconds.

To simulate the presence of human voice in scene recordings, we artificially mixed the clean scene recordings of the

http://dx.doi.org/10.21437/Interspeech.2020-2365

2019 DCASE test set with clean speech utterances from the Edinburgh speech database [24]. The Edinburgh database was recorded from 56 speakers (28 female and 28 male) of different accent regions in Scotland and the United States, and for each speaker about 400 utterances are available. The sampling rate is 48 kHz which is identical to that of 2019 DCASE recordings.

3. Acoustic scene classification architectures

In this work, we use two different ASC architectures. The first is the official 2019 DCASE baseline architecture that consists of two convolutional layers, followed by batch normalisation, rectified linear unit (ReLU) activation, dropout and max pooling, and two fully-connected layers. The architecture is depicted in Figure 1(a). The baseline model achieves a classification accuracy of 62.20 % on the official test set of the 2019 DCASE challenge.

In addition, as outlined we employ an CNN architecture combining atrous convolution with spatial attention mechanism that we used in our previous work [23]. The difference of an atrous CNN compared to the standard architecture is that instead of using pooling layers, it extends the receptive field for each convolutional layer by exploiting dilation settings which controls the spacing between the kernel points [27].

As shown in Figure 1(b), our attentive atrous CNN is comprised of 4 atrous convolutional layers, on top of which 2D attention values are learnt and allocated to each of the feature maps. Again, batch normalisation is applied for each convolutional layer. The attentive feature maps are then averaged across all locations and projected to the scene labels via a dense layer. The architecture has been reported to be effective for acoustic scene classification and achieved 69.0 % on the evaluation set of the official 2018 DCASE challenge. In this work, we develop and test the attentive atrous CNN on the 2019 DCASE data, and obtain a classification accuracy of 77.51 % on the evaluation set.

For both architectures, we extract log-Mel spectra using a frame size of 40 ms and hop size of 20 ms, leading to 40×500 Mel-band spectrum as the input for the networks.

3.1. Robustness under speech noise

We investigate the robustness of both ASC architectures in the presence of human speech by training them on the original training data, and testing them on test data corrupted with speech noise. For each clean scene recording in the 2019 DCASE test set, a speech utterance randomly sampled from the Edinburgh database is first truncated or extended to 10 seconds, and then mixed into the clean scene recording with different signal-to-noise ratios (SNRs).

The results shown in the "Noisy" column of Table 2 for each scene classifier demonstrate the influence of speech on the classification performance under different SNR conditions. As expected, the performance of both algorithms deteriorates as the signal energy of the human voice relative to that of the acoustic scene increases.

We observe that both architectures do not seem to suffer from noise in high SNR conditions. However, as the SNR drops, classification performance does fall down to almost chance performance as for the case of -10 dB. Even for the relatively high SNR of 10 dB, performance drops by 15 % and 23 %, respectively.

In real-life conditions, an SNR of -10 dB, indicating the speech is 10 dB stronger than environment in energy, is very common, especially for audio recordings made on mobile and wearable devices. These conditions would render both algorithms completely ineffective for the ASC task.

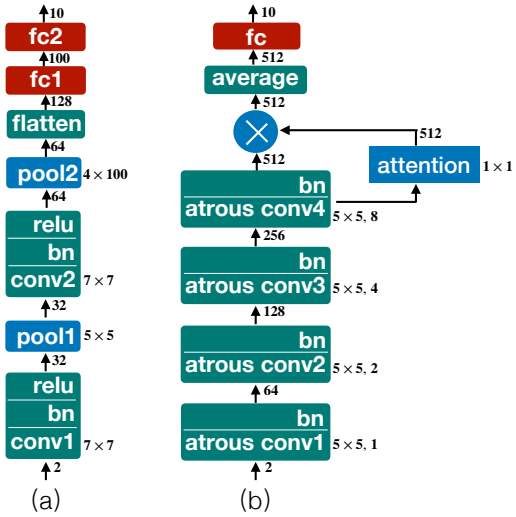

Figure 1: *Acoustic scene classifiers used in this work: (a) Baseline model of DCASE2019. Kernel size for each convolutional layer is labelled besides the block, and the number of channels is denoted between each block transition. (b) Attentive atrous CNN. Dialation size for each atrous convolutional layer is given next to the kernel size.*

4. Improving speech robustness of ASC classifiers

4.1. Approach 1: Data Augmentation

A straightforward way to mitigate the effect of human voice on ASC classifiers is to improve their ability to generalise in the presence of speech using data augmentation. To this end, both the baseline and the attentive atrous CNNs are trained on contaminated scene recordings generated by mixing clean training speech utterances from the Edinburgh database into the clean scene recordings from the training set of the 2019 DCASE challenge.

We first test this approach under *matched SNR* conditions, in which the training set is augmented with data mixed in the exact same SNR as the test set. We also test the more realistic *multi-SNR* case, in which instances in the training data are mixed with an SNR randomly selected to be one of $-10, -5, 0, 5, 10, 20$ or 30 dB.

As can be seen in Table 2, using data augmentation substantially improves the performance of both models, and they are able to recover most of the lost accuracy due to the mismatch between training and testing conditions. The baseline model performs better for the matched SNR conditions in all cases. For the atrous CNN model, we observe that the model performs better in the multi-SNR case for low SNR conditions. This indicates the generalisation benefit arising from training with multiple different SNRs.

4.2. Approach 2: Voice Suppression

The experiments in Section 3.1 show that both models are able to deal with small perturbations in the input signals, as performance does not degrade substantially for SNRs up to 20 dB.

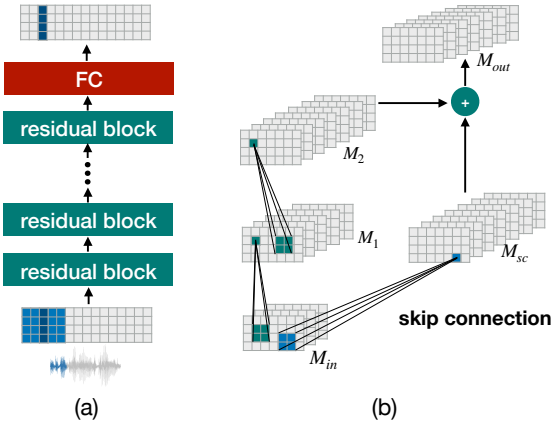

Figure 2: *(a) Voice suppression model based on a sequence of 8 residual blocks. (b) Residual block architecture.*

In addition, previous work in other audio tasks has shown that it can be beneficial to use a denoising architecture as a preprocessing frontend instead of augmenting the training data with noise [28, 29, 30]. Therefore, we investigate the potential of a denoising architecture. In this context, the architecture would need to remove all speech components, rather than all environment components as is traditionally the case in the speech enhancement field. Thus, we use the term *voice suppression* system to refer to architectures of this kind. A similar definition has been found in the music source separation task as introduced by zapata2013. However, in this work, source separation approaches were exploited to concurrently estimate the singing voice and its accompaniments in a music segment, which can fall short of clean removal of voice components in the estimated accompaniments. Hence, we propose a voice suppression system, using a spectral-mapping scheme that can directly estimate the surrounding environments in an audio, and discard the voice to the greatest possible extent.

Our proposed voice suppression system processes a spectrum segment of the contaminated scene recording, and recovers the enhanced scene spectrum segment. The system architecture consists of a sequence of 8 residual blocks as displayed in Figure 2. Each basic residual block, shown in Figure 2(b), consists of two convolutional layers in the main path to process the input feature maps. Using a 1×1 convolution in the skip-connection path, the input feature maps are converted to the same size as the second convolutional layer's output, and then added to the main path to produce the output feature maps. Batch normalisation is applied for each convolutional layer, activated by the ReLU function. The skip-connections enable a substantially deeper CNN architecture, which has been proven very successful in both the computer vision and audio domains [31, 32, 33]. Each residual block of the voice suppression system has its own kernel size, stride, and number of channels as appear in Table 1.

The input to the network is a neighbourhood of 35 log short-time Fourier transform (STFT) frames in the frequency domain, from which the network learns to directly predict the clean central frame. To compute the STFT frames, we used a 25 ms Hanning window shifted by 10 ms. The model is trained by minimising mean square error (MSE) between the estimated scene frame and the target frame, which is the central frame of the clean scene spectrum segment. We optimise the network parameters using stochastic gradient descent (SGD) with a learning rate of 0.01.

Table 1: *The speech suppression model specifications.*

Block	Kernel	Stride	#Channels
1	$(4, 4)$	$(1, 1)$	64
2	$(4, 4)$	$(1, 1)$	64
3	$(4, 4)$	$(2, 2)$	128
4	$(4, 4)$	$(1, 1)$	128
5	$(3, 3)$	$(2, 2)$	256
6	$(3, 3)$	$(1, 1)$	256
7	$(3, 3)$	$(2, 2)$	512
8	$(3, 3)$	$(1, 1)$	512

The voice suppression model is trained with artificially mixed data using the training set of the 2019 DCASE data set, and clean speech recordings from the training set of the Edinburgh data set. The data were mixed with randomly selected SNRs in the range [-20, -10 -5 ,0, 5, 10, 15] dB.

Figure 3 shows the effect of the trained voice suppression model on a test instance corrupted with speech. The spectrogram of the clean scene recording and that of the speech utterance are depicted on the first two panels, their mixture leads the noisy scene, whose spectrogram is shown on the third panel. Our proposed voice suppression model processes the noisy scene and produces the denoised scene, which is expected to be as close as the clean scene. It is observed that the voice in the noisy scene is explicitly suppressed, while the scene components are successfully preserved.

Applying voice suppression to all noisy scene recordings of the test set, we see that the log Mel-band energies extracted from the denoised scene audios have less deviation from that extracted from the clean scene audios, leading to reliable input for ASC classifiers. The deviation can be measured by computing the average MSE between the two feature sets, as depicted in Figure 4. The blue curve represents the deviation between the noisy scene signals and the clean scene signals in terms of feature sets, while the red curve stands for the deviation between the enhanced scene signals and the clean scene signals. Comparing to the noisy scene recordings, the feature sets from the enhanced scene recordings are less affected by speech, especially for the higher SNR cases.

5. Discussion

For each test scene recording contaminated by speech, the scene spectrum is processed by the voice suppression system to produce each frame of the enhanced scene spectrum. The log mel-band energies are then extracted from the enhanced scene spectrum and fed into an acoustic scene classifier to predict the associated environment labels. Experimental results, under the column "Denoised" of Table 2, imply that voice suppression assists with acoustic scene classifiers in achieving considerable improvements in terms of classification accuracy for the large SNR cases. For the higher SNR values, e. g., above 20 dB, where, due to the minimal presence of human voice, the original classifiers perform similarly for the noisy scene recordings as for the clean recordings, we observe only a slight improvement. With the assistance of voice suppression, the baseline model works stably across all SNR cases and achieves around 60 % classification accuracy. The attentive atrous CNN perfor-

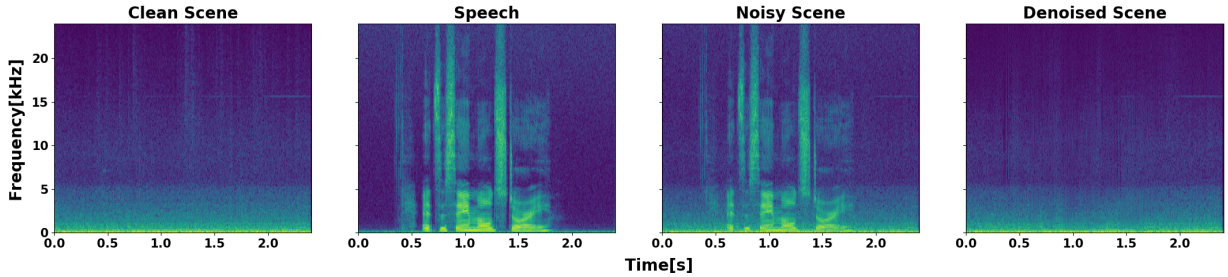

Figure 3: *Voice suppression processing on an example of noisy scene recording. The clean scene recording is deteriorated with the speech utterance, leading to the noisy scene recording. The system processes the noisy scene recording and produces the denoised scene recording.*

Table 2: *ASC accuracy[%] results*

	2019 DCASE Baseline				Attentive Atrous Model			
SNR	**Noisy**	**Multi-SNR**	**Matched SNR**	**Denoised**	**Noisy**	**Multi-SNR**	**Matched SNR**	**Denoised**
Clean	62.20	–	–	–	77.51	–	–	–
30 dB	61.60	41.46	59.89	61.67	76.58	58.78	65.93	77.16
20 dB	59.40	42.15	59.90	60.81	71.95	59.93	62.22	67.00
10 dB	47.38	44.87	59.42	57.35	54.84	60.86	61.51	62.08
5 dB	36.27	46.81	60.22	56.42	41.17	61.15	63.23	60.45
0 dB	25.19	49.68	58.97	57.08	28.29	61.74	58.87	61.91
-5 dB	17.11	52.21	56.27	57.71	21.48	61.31	59.73	63.25
-10 dB	14.27	53.41	57.37	57.71	17.99	60.05	58.23	61.95

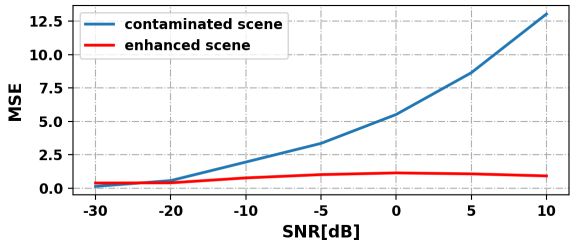

Figure 4: *Deviation between the extracted log mel-band energies from denoised scene audio and clean scene over different SNRs.*

mance is less stable across different SNRs. As voice intensity increases, the classification performance gets worse, but voice suppression nevertheless assists in reducing the effect of human voice, and enhances the acoustic scene classification for contaminated scene recordings by a great margin.

The scene classifiers with voice suppression as its front-end processor to improve audio scene quality have achieved reliable classification accuracy for contaminated scene recordings, but the models still do not achieve the same performance as for clean scene recordings. This can be attributed to two reasons:

- first, although we attempt to suppress the human voice in noisy scene recordings to the greatest extent, there still exists some residual voice components in the processed scene audio, and the speech leftover occasionally turns to very low fizzer sounds, which, though not always audible, may still lead to classification performance degradation;
- second, the voice suppression model attempts to eliminate all speech contents in scene recordings, and therefore may be aggressive towards the scene context itself, especially when the scene recordings originally contain some human voice, leading to a loss of environmental information.

6. Conclusion

In this paper, we investigated the effects that human voice can have on neural network based acoustic scene classification, and concluded that the presence of speech in real-world recordings can be detrimental to system performance.

To tackle the issue, we investigated two alternative approaches:

1. augmenting the training set with speech data, and
2. training a voice suppression model to be used as a pre-processing frontend.

In almost all cases, the voice suppression architecture achieved superior performance in making the classifiers robust to speech noise compared to simply using data augmentation. Nevertheless, there still remains a substantial gap compared to system performance under clean conditions. This illustrates that the problem of acoustic scene classification remains very challenging in real-world conditions.

In addition, more work should be done to extract features from contaminated scene recordings that are more robust to speech. Future work should be directed in collecting more realistic ASC data sets "in-the-wild", so that they cover the case of humans interacting in the "foreground" of the scene. Furthermore, since the use of pre-processing architectures seems promising compared to vanilla data augmentation, it would be beneficial to further investigate such architectures, drawing inspiration from related speech enhancement literature, but extending them for the task of removing, rather than preserving, human speech.

7. References

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] S. Gharib, H. Derrar, D. Niizumi, T. Senttula, J. Tommola, T. Heittola, T. Virtanen, and H. Huttunen, "Acoustic scene classification: A competition review," in *Proc. MLSP*, Aalborg, Denmark, 2018, pp. 1–6.

[3] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[5] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proc. UbiComp*, Osaka, Japan, 2015, pp. 283–294.

[6] M. Green and D. Murphy, "Environmental sound monitoring using machine learning on mobile devices," *Applied Acoustics*, vol. 159, 2019, 8 pages.

[7] M. Won, H. Alsaadan, and Y. Eun, "Adaptive audio classification for smartphone in noisy car environment," in *Proc. ACM Multimedia*, Mountain View, CA, 2017, pp. 1672–1679.

[8] B. Kuehn, A. Belkin, A. Swerdlow, T. Machmer, J. Beyerer, J. Beyerer, and K. Kroschel, "Knowledge-driven opto-acoustic scene analysis based on an object- oriented world modeling approach for humanoid robots," in *Proc. ISR/ROBOTIK*, Munich, Germany, 2010, pp. 1–8.

[9] S. Aziz, M. Awais, T. Akram, U. Khan, M. Alhussein, and K. Aurangzeb, "Automatic scene recognition through acoustic classification for behavioral robotics," *Electronics*, vol. 8, no. 5, pp. 483–500, 2019.

[10] S. Chu, S. Narayanan, and C. J. Kuo, "Content analysis for acoustic environment classification in mobile robots," in *Proc. AAAI Fall Symp: Aurally Informed Performance*, Arlington, VA, 2006, pp. 16–21.

[11] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2991–3002, 2005.

[12] D. Fabry and J. Tchorz, "Results from a new hearing aid using acoustic scene analysis," *The Hearing Journal*, vol. 4, no. 58, pp. 30–36, 2005.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups," in *Proc. DCASE2019*, New York, NY, 2019, pp. 164–168.

[14] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[15] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[16] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," in *Proc. DCASE2018*, Woking, Surrey, UK, 2018, 5 pages.

[17] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," in *Proc. DCASE2019*, New York, NY, 2019, 5 pages.

[18] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Acoustic scene classification: from a hybrid classifier to deep learning," in *Proc. DCASE2017*, Munich, Germany, 2017, 5 pages.

[19] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *Proc. IJCNN*, Anchorage, Alaska, 2017, pp. 1547–1554.

[20] T. Zhang, J. Liang, and B. Ding, "Acoustic scene classification using deep cnn with fine-resolution feature," *Expert Systems with Applications*, vol. 143, no. 2020, 2020.

[21] B. Gulmezoglu, A. Zankl, C. Tol, S. Islam, T. Eisenbarth, and B. Sunar, "Undermining user privacy on mobile devices using AI," in *Proc. Asia CCS*, Auckland, New Zealand, 2019, pp. 214–227.

[22] I. H. Hann, K. L. Hui, S. Y. T. Lee, and I. P. Png, "Overcoming online information privacy concerns: An information-processing theory approach," *Journal of Management Information Systems*, vol. 24, no. 2, pp. 13–42, 2007.

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

[24] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," University of Edinburgh. School of Informatics. Centre for SpeechTechnology Research (CSTR), 2017.

[25] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2005.

[26] N. Saleem and M. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *International Journal of Speech Technology*, vol. 22, no. 2019, pp. 1051–1075, 2019.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, San Juan, PR, 2016, pp. 1–13.

[28] A. Moore, P. Peso Parada, and P. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Computer Speech Language*, vol. 46, pp. 574–584, 2016.

[29] F. Zhao, H. Li, and X. Zhang, "A robust text-independent speaker verification method based on speech separation and deep speaker," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 6101–6105.

[30] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1691–1695.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, 2016, pp. 770–778.

[32] H. K. Vydana and A. K. Vuppala, "Residual neural networks for speech recognition," in *Proc. EUSIPCO*, Kos island, Greece, 2017, pp. 543–547.

[33] H. Jung, M. K. Choi, J. Jung, J. H. Lee, S. Kwon, and W. Young Jung, "Resnet-based vehicle classification and localization in traffic surveillance systems," in *Proc. CVPR*, Honolulu, HI, 2017, pp. 61–67.