ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747680

FILTERAUGMENT: AN ACOUSTIC ENVIRONMENTAL DATA AUGMENTATION METHOD

Hyeonuk Nam, Seong-Hu Kim, Yong-Hwa Park

Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Korea

ABSTRACT

Acoustic environments affect acoustic characteristics of sound to be recognized by physically interacting with sound wave propagation. Thus, training acoustic models for audio and speech tasks requires regularization on various acoustic environments in order to achieve robust performance in real life applications. We propose *FilterAugment*, a data augmentation method for regularization of acoustic models on various acoustic environments. FilterAugment mimics acoustic filters by applying different weights on frequency bands, therefore enables model to extract relevant information from wider frequency region. It is an improved version of frequency masking which masks information on random frequency bands. FilterAugment improved sound event detection (SED) model performance by 6.50% while frequency masking only improved 2.13% in terms of polyphonic sound detection score (PSDS). It achieved equal error rate (EER) of 1.22% when applied to a text-independent speaker verification model, outperforming model used frequency masking with EER of 1.26%. Prototype of FilterAugment was applied in our participation in DCASE 2021 challenge task 4, and played a major role in achieving the 3rd rank.

***Index Terms*—** data augmentation, acoustic environment, acoustic model, sound event detection, text-independent speaker verification

1. INTRODUCTION

Training deep neural networks (DNNs) requires enormous high-quality dataset in order to regularize models and achieve robust performances. However, collecting such dataset requires tremendous time and cost. Many data augmentation methods have been proposed in order to effectively utilize limited size of dataset. Data augmentation methods increase dataset size by providing various "views" of the same data [1, 2, 3]. By training neural networks with data from different views every epoch, they can be regularized to learn information shared across different views.

Application of deep learning (DL) methods in audio and speech tasks such as sound event detection (SED), speaker recognition and automatic speech recognition (ASR) have been adopting techniques from DL methods developed for computer vision domain [2, 3, 4, 5, 6]. Applying short-time Fourier transform (STFT) [7] on audio data can change data dimension from 1D waveform (time) to 2D spectrogram (time and frequency), which can be treated like image data [5, 6, 8]. Some of data augmentation methods proposed for computer vision tasks such as mixup [9] are actively adopted in audio and speech domain. However, most of the image data augmentation methods including rotation, flip, shear and crop [1] result in irrelevant transform of audio data when applied on spectrograms. Therefore, data augmentation methods consistent with acoustics and signal processing domain knowledge are required to effectively train *acoustic models* in audio and speech domain. The term acoustic model usually refers to the encoder structures in ASR models for extracting acoustic information from speech audio signals. In this paper, we call encoder structures in neural networks for audio and speech tasks as acoustic models, as their common purpose is to effectively extract useful acoustic information from audio signals.

Early works introducing DL methods on audio and speech tasks used conventional audio signal processing methods [2, 3] for data augmentation. Although conventional audio signal processing methods do help increasing dataset size, they are not easy to utilize without sufficient understanding in acoustics and signal processing domain. This problem was resolved by SpecAugment [10], involving time masking and frequency masking those simply mask a small time and frequency range of mel spectrograms. Time and frequency masking can be easily applied to train acoustic models as their algorithms are simple and straightforward, but they are brutal in the sense that they completely remove certain information from the data.

In this work, we propose *FilterAugment*, an improved version of frequency masking from SpecAugment [10]. FilterAugment is proposed to regularize acoustic models over various acoustic environments by mimicking acoustic filters. Sound could be heard differently in various acoustic environments such as conference room, shower room, performance hall, cave, etc. Although such acoustic characteristics derived from different physical contexts vary a lot, human can recognize sound events, speakers, or spoken words regardless of acoustic characteristics. These highly variable acoustic characteristics can be modeled using acoustic filters [11], which FilterAugment aims to mimic in a simplified way so that acoustic models could be trained to recognize the sound contents in various acoustic environments. FilterAugment approximates acoustic filters by applying random weights on randomly determined frequency bands. Although applying FilterAugment does not make resulting to sound as natural as results of applying acoustic filters, it effectively regularizes acoustic models by extracting sound information from wider range of frequency. Prototype of FilterAugment was applied on our participation in Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 challenge task4, resulting in the 3rd rank [12]. Considering that most of other teams above 5th rank added major modifications on model architecture [13, 14, 15, 16, 17] while we did not, FilterAugment is proven to be a powerful data augmentation method. The official implementation code for FilterAugment applied on SED is shared on GitHub[1].

2. AUDIO DATA AUGMENTATIONS

Data augmentation methods in audio and speech domain includes conventional audio signal processing methods such as time stretching, pitch shift, clipping, suppressing, adding noise, adding reverber-

This work was supported by "Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20204030200050)

[1] https://github.com/frednam93/FilterAugSED/

ation, etc. [2, 3]. These methods reflect domain knowledge in acoustics and signal processing, thus they have been frequently adopted for data augmentation purpose. However, data augmentation using conventional audio signal processing methods could introduce some inefficiencies when training acoustic models. Applying conventional audio signal processing methods requires prior knowledge to appropriately handle audio data. In addition, these methods may involve more computations in expense for more natural sound, which does not even guarantee to train acoustic models better. Such inefficiencies hinder optimal training of acoustic models. Therefore, we need data augmentation methods that are simple, intuitive, yet effective for training acoustic models to learn to extract information from audio data.

SpecAugment [10] is one of the most powerful and widely used data augmentation methods in audio and speech domain. Instead of applying data augmentation on waveform, it proposed time warping, time masking, and frequency masking those could be directly applied on log mel spectrogram. As it is applied directly on the input feature space, it is easy to comprehend and use. Intuitively, applying time warping on audio would sound like the audio played faster in some points and slower in some other points. Time masking would sound that some parts are not played for short duration. Frequency masking would sound like some part of frequency range is missing. As long as these distortions are not too severe, human can recognize the content of audio data after these processing, and trained acoustic models should do as well. Although these methods do not sound as natural as conventional audio processing methods when transformed back in waveform, it helps training acoustic models more effectively with extreme cases.

3. PROPOSED METHOD

3.1. Motivation

FilterAugment can be explained in two different but related points of view. In the viewpoint of acoustics and signal processing, FilterAugment regularizes acoustic models to various acoustic environments by mimicking acoustic filters. From the viewpoint of acoustic model training, FilterAugment learns to effectively extract acoustic information from wide frequency ranges while training. We will first explain motivation of FilterAugment in terms of acoustics and signal processing viewpoint, then discuss its significance in terms of acoustic model training.

When we hear sound events or speeches, we can recognize their contents regardless of acoustic environments unless the environments are too noisy or too echoic. It is because our auditory system is trained to understand the sound contents regardless of the acoustic environments. Acoustic environment refers to the physical objects surrounding the sound source, receiver (ear or microphone) and the air surrounding them (medium of the sound wave propagation). These interact with sound wave and change the acoustic characteristics of sound perceived by receiver with absorption, reflection, scattering, etc. [11]. Such change in acoustic characteristics appears as relative change in energy on different frequency range. For example, when the sound source is far away from the receiver, high-frequency energy reduces as it dampens more than low-frequency energy does while propagating in the air. Similarly, when there is a wall or any object blocking between the receiver and the sound source, high-frequency energy reduces as it does not diffract easily thus does not propagate to the receiver much. In addition, room's walls and furnishings cause reverberation, and early reverberation causes coloration which alters acoustic characteristics of the sound perceived. Such change in energy on different frequency ranges can be simulated by designing appropriate types of filter: high pass filter, low pass filter, band pass filter, notch filter, etc. [7, 11]. However, designing and applying such filters for data augmentation purpose requires understanding in acoustics and signal processing. In addition, applying filters to training audio data takes time to compute filters' impulse responses and convolute them with audio data. Although training time might not increase that much, it will complicate training and optimization process. Therefore, we propose FilterAugment, a simpler alternative data augmentation method to mimic filter effect. FilterAugment randomly increases or decreases energy of random frequency ranges of log mel spectrograms. Such increase or decrease of energy in random frequency range is equivalent to application of random filters. Although it might sound unnatural compared to acoustic filters as it induces discrete filter design, FilterAugment is much easier to comprehend and use.

From the viewpoint of acoustic model training, randomly weighting on random frequency bands of log mel spectrogram enables training of acoustics models to extract sound information from wider frequency regions. Without FilterAugment, acoustic model is likely to learn to recognize frequency ranges that exhibit dominant and distinctive feature of desired labels. However, we can recognize the sound content regardless of the acoustic environment that might even drastically reduce the frequency region with the most distinctive feature. It means that we still can recognize sound content from the other less distinctive frequency ranges. This would be the reason why applying frequency masking [10] improves training acoustic models as well. As frequency masking removes information from certain random frequency range, it helps to train acoustic model to infer the sound information from less distinctive frequency regions too. However, frequency masking completely removes certain part of energy that might help inferring the sound information. Such brutal damage on spectrum not only rarely happens in real situations but also causes the model to be trained to forcibly extract information from indistinct frequency ranges. Therefore, FilterAugment weakens some parts of frequency range while strengthening other parts instead. Lowering energy instead of removing it would at least let acoustic models to infer the information from that frequency region. In addition, increasing other frequency range energy would train acoustic models to recognize sound information from various frequency region as they will be trained with the same data highlighted on different frequency region every epoch. Therefore, FilterAugment helps training acoustic models to extract information from the wider range of frequency regardless of each frequency's relative significance composing the sound information.

3.2. Algorithm

We propose three types of FilterAugment: step, linear and mixed type. Detailed algorithm for step type FilterAugment is as follows.

1. Randomly choose number of frequency bands n within hyperparameter *band number range*.
2. Randomly choose $n-1$ mel frequency bins between 0 and F (number of mel frequency bins in mel spectrogram), and include 0 and F to form $n+1$ frequency boundaries. These frequency boundaries are separated from each other at least by hyperparameter *minimum bandwidth.*
3. Randomly choose n different weights within hyperparameter *dB range*.
4. Add chosen n weights on n frequency bands of log mel spectrogram defined by each set of subsequent frequency boundaries respectively.

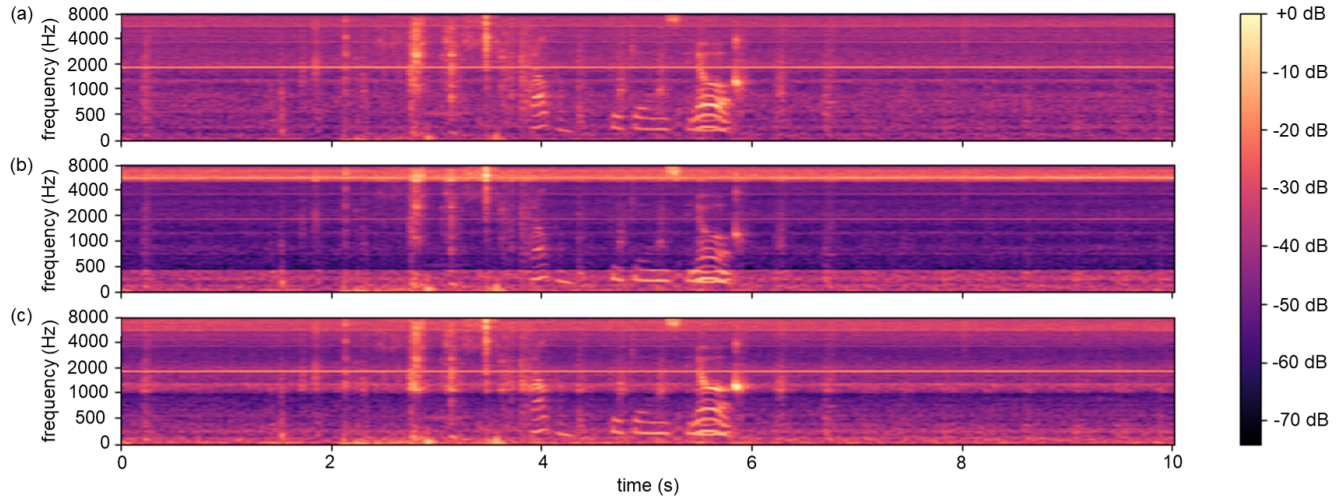

Fig. 1. Illustrations on application of FilterAugment on an example audio clip. (a) is log mel spectrogram of the original audio clip, (b) is log mel spectrogram applied with step type FilterAugment, (c) is log mel spectrogram applied with linear type FilterAugment.

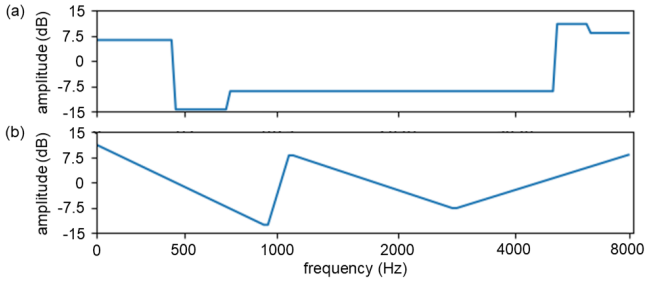

Fig. 2. Filters applied on the example audio clip. (a) is step type filter that resulted on Fig. 1. (b), and (b) is linear type filter that resulted on Fig. 1. (c).

As a result, mel spectrogram's energy is amplified in some frequency bands while reduced in other bands. Note that *Minimum bandwidth* is adopted to prevent applying weights too locally. Amplifying or reducing energy of a frequency band with too narrow bandwidth would cause negligible change in how audio clips sound, so we set minimum bandwidth to make sure each weight cause significant change in sound. Step type FilterAugment is the simplest type of FilterAugment composed of series of step functions. An example of step type FilterAugment is shown in Fig. 1. (b). The filter applied on the original audio clip's log mel spectrogram Fig. 1. (a) to produce it is shown in Fig. 2. (a). It can be observed that when compared to the original log mel spectrogram, the augmented result has higher energy below 400 Hz and above 5 kHz while the rest of frequency regions have lower energy. Also, the abrupt change in energy across frequency range can be seen as clear horizontal lines on frequency boundaries at 400 Hz and 5 kHz on Fig. 1. (b).

Abrupt energy changes at boundary frequencies on step type FilterAugment cause unnatural sound. To make more naturally augmented audio data, linear type FilterAugment is proposed. Algorithm for linear type FilterAugment shares 1st and 2nd steps with step type FilterAugment. Rest of the algorithm is as follows.

3. Randomly choose $n + 1$ different weights corresponding to $n + 1$ frequency boundaries within hyperparameter *dB range*.
4. Linearly interpolate weights within frequency boundaries.
5. Add resultant interpolated weights on log mel spectrogram.

While step type FilterAugment applies discontinuous filter that is composed of series of step functions, linear type applies continuous (although not differentiable) filters that are composed of series of linear functions. An example of linear type FilterAugment is shown in Fig 1. (c). It is produced by applying filter Fig. 2. (b) to original log mel spectrogram Fig. 1. (a). Compared to the original log mel spectrogram, peaks around 0 Hz, 1.2 kHz and 8 kHz and dips around 900 Hz and 3 kHz can be observed and the gradual change between these peaks and dips can be observed as well. Linear type FilterAugment shows smoother change in energy across frequency axis in Fig. 1. (c) when compared step type FilterAugment in Fig. 1. (b) that shows abrupt change in energy across frequency axis.

As step and linear type FilterAugment are expected to have different effect on training acoustic models, mixed type FilterAugment is proposed to train acoustic models to be regularized on both step and linear type of FilterAugment. Hyperparameter *mix ratio* determines the probability of using step type FilterAugment. For example, if mix ratio is 0.7, then there is 70% chance that step type FilterAugment is applied to the batch and 30% chance that linear type FilterAugment is applied to the batch.

4. EXPERIMENTS

4.1. Implementation Details

FilterAugment algorithm was tested on SED and text-independent speaker verification: one task each from audio and speech domains. We first optimized hyperparameters of FilterAugment and frequency masking on SED, then applied them with the optimized hyperparameters on speaker verification. Frequency masking involves a hyperparameter *maximum masking ratio* which determines the maximum ratio of mel frequency bins to be randomly masked during the training to F. The performances of the baseline models with and without FilterAugment and frequency masking are compared.

SED baseline model in this work is an upgraded version of baseline model for DCASE 2021 challenge task 4 [18, 19], which is the same with optimized model in [12] without the prototype FilterAugment. From DCASE baseline [19, 20], dimensions of convolutional recurrent neural network (CRNN) are doubled. Activation functions in convolutional neural network (CNN) structure are replaced by context gating [21]. Waveforms are normalized so that their absolute maximum equals to one. Time masking [10] is added with optimized masking range within 7 – 30 frames (0.11 – 0.48 seconds). Weak prediction masking is applied to test predictions [12].

Table 1. Performance of SED models trained with and without frequency masking and FilterAugment.

Methods	$PSDS_1$ ↑	$PSDS_2$ ↑	CB-F1 ↑	IB-F1 ↑
baseline	0.387	0.598	0.477	0.708
freq masking	0.396	0.610	0.470	0.710
step FiltAug	0.412	0.634	0.474	0.712
linear FiltAug	**0.413**	**0.636**	**0.490**	**0.735**

Table 2. Performance of SED models trained using mixed type FilterAugment. None of these surpasses neither step type nor linear type FilterAugment.

mix ratio	$PSDS_1$ ↑	$PSDS_2$ ↑	CB-F1 ↑	IB-F1 ↑
0.9	0.407	0.628	0.473	0.719
0.7	0.401	0.606	0.469	0.709
0.5	0.395	0.602	0.476	0.713
0.3	0.401	0.610	0.472	0.710
0.1	0.408	0.622	0.470	0.711

We compared the performance of baseline model with and without frequency masking and step/linear/mixed type FilterAugment. Evaluation metrics measured include polyphonic sound detection score (PSDS) criteria on DCASE 2021 challenge task 4 [18, 19, 22] for two situations ($PSDS_1$ and $PSDS_2$), macro collar-based F1 score [23] and macro intersection-based F1 score [24]. $PSDS_1$ penalizes more on inaccurate time localization while $PSDS_2$ penalizes more on confusions between classes. These four metrics are higher with better SED performance. Hyperparameters for frequency masking and FilterAugment are optimized for the highest $PSDS_1$ + $PSDS_2$ which is official evaluation score on DCASE 2021 challenge task 4. F1 scores are listed for reference.

We applied the optimized settings of frequency masking and FilterAugment to text-independent speaker verification baseline model, which is the model without data augmentation from [25]. Then, only dB range was re-optimized because the datasets [26, 27] are composed of interviews from YouTube videos recorded in controlled acoustic environments. Although they might have some noises, speaker-microphone distances are usually close and the microphones' recording qualities are good enough to keep speeches' acoustic characteristics almost constant. Therefore, dB range of FilterAugment is narrowed to match the variance of acoustic characteristics of speaker verification task. The baseline model is ResNet-34 with SE module and attentive statics pooling (ASP) [28]. It is trained using 5994 speakers of Voxceleb2 dataset [27] with combined loss function composed of Angular Prototypical (AP) loss [29] and vanilla softmax loss. Speaker embeddings are extracted from each utterance of Voxceleb1 dataset [26] and compared using cosine similarities for validation. For evaluation metrics, we used equal error rate (EER) and minimum detection cost function (MinDCF) with C_{miss} = 1, C_{fa} = 1 and P_{target} = 0.05 [26, 30]. Lower EER and MinDCF values imply better speaker verification performance.

4.2. Results and Analysis

Optimized hyperparameter for frequency masking is *maximum masking ratio* = 1/16. Optimized hyperparameters for step type FilterAugment (listed as step FiltAug in Table 1) are *dB range* = (-6, 6), *band number range* = (2, 5), and *minimum bandwidth* = 4. Optimized hyperparameters for linear type FilterAugment (listed as linear FiltAug in Table 1) are *dB range* = (-6, 6), *band number range* = (3, 6), and *minimum bandwidth* = 6. Mixed type FilterAugment uses hyperparameters optimized for step and linear type FilterAugment above. Macro collar-based F1 score and macro intersection-based F1 score are listed in Table 1 as CB-F1 and IB-F1 respectively. The optimized models' metric values are listed in Table 1, and these values chosen to be displayed are the maximum values of each metric upon three independently trained SED models. Since each training results in separate student model and teacher model by mean teacher method [31], these results are the maximum values among the results of 6 models. The results show that FilterAugment improved SED model performance and significantly outperformed model trained using frequency masking. Linear type FilterAugment shows slightly better performance than step type does. The difference is not significant, but more realistic simulation of acoustic filter might helped training acoustic models better. Note that optimized linear FilterAugment requires more frequency bands and wider minimum bandwidth. This should be required to give more distortion on the spectrogram, as linear type FilterAugment tend to cause spectrogram less distortion due to the linear interpolation between two weights on frequency boundaries. Mixed type FilterAugment performs worse than both step and linear type FilterAugment, as shown in Table 2. It can be observed that as mixing ratio is closer to 0.5 meaning uniform mixing, the performance worsens. It can be concluded that using different data augmentation method on different batchs during training could result in inconsistent training thus degrade performance. In the end, the best score is achieved by linear FilterAugment. Frequency masking surpassed baseline model performance by 2.13%, while linear type FilterAugment surpassed baseline model performance by 6.50%.

Table 3. Text-independent speaker verification performances with and without frequency masking and FilterAugment.

Methods	EER (%) ↓	MinDCF ↓
baseline (no data aug) [25]	1.29	0.091
frequency masking	1.26	0.092
FilterAugment	**1.22**	**0.088**

We compared the text-independent speaker verification performance with and without data augmentation methods, and the results are shown in Table 3. FilterAugment used for speaker verification follows the same setting of linear type FilterAugment for SED, with re-optimized *dB range* = (-1.5, 1.5). It is shown that FilterAugment shows the better performance than the models without augmentation and with frequency masking. Although Voxceleb1 and 2 have constrained acoustic environments, FilterAugment still outperformed frequency masking.

5. CONCLUSION

FilterAugment is an audio data augmentation method that enables effective training of acoustic models in audio and speech domain tasks. It regularizes acoustic models over various acoustic environments by learning to extract sound information from wider frequency range. On both SED and text-independent speaker verification, we showed that FilterAugment outperforms not only models without data augmentation, but also models with frequency masking which uses similar approach with FilterAugment. In conclusion, FilterAugment is simple yet one of the most powerful audio data augmentation methods, having contributed largely to winning 3rd rank in DCASE 2021 task 4.

6. ACKNOWLEDGEMENTS

We would like to thank Junhyeok Lee from MINDs Lab Inc. and Won-Ho Jung from KAIST for valuable discussions.

7. REFERENCES

[1] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition," in *Proc. Interspeech*, 2016, pp. 2982–2986.

[3] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[4] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, 2019.

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, Ar. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[6] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing: International Edition*, pp. 522–525, 850–851, Pearson, 3rd edition, 2010.

[8] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, "Deep learning based cough detection camera using enhanced features," *arXiv preprint arXiv:2107.13260*, 2021.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[10] D. S. Park, W., Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[11] L.E. Kinsler, A.R. Frey, A.B. Coppens, and J.V. Sanders, *Fundamentals of Acoustics*, pp. 149, 210, 224, 291–296, 333–334, Wiley, 4th edition, 2000.

[12] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," Tech. Rep., DCASE Challenge, 2021.

[13] X. Zheng, H. Chen, and Y. Song, "Zheng ustc team's submission for dcase2021 task4 – semi-supervised sound event detection," Tech. Rep., DCASE Challenge, 2021.

[14] N. K. Kim and H. K. Kim, "Self-training with noisy student model and semi-supervised loss function for dcase 2021 challenge task 4," Tech. Rep., DCASE Challenge, 2021.

[15] R. Lu, W. Hu, D. Zhiyao, and J. Liu, "Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios," Tech. Rep., DCASE Challenge, 2021.

[16] J. Ebbers and R. Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," Tech. Rep., DCASE Challenge, 2021.

[17] G. Tian, Y. Huang, Z. Ye, S. Ma, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, K. Ouchi, J. Ebbers, and R. Haeb-Umbach, "Sound event detection using metric learning and focal loss for dcase 2021 task 4," Tech. Rep., DCASE Challenge, 2021.

[18] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[19] DCASE, "Dcase 2021 challenge task4: Sound event detection and separation in domestic environmentse," Accessed on: 2021, Oct 2. [Online]. Available: http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments.

[20] N. Turpault, "Dcase2021 task4 baseline," GitHub. Available: https://github.com/DCASE-REPO/DESED_task.

[21] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," 2018.

[22] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 61–65.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[24] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen, and S. Krstulović, "Improving sound event detection metrics: Insights from dcase 2020," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 631–635.

[25] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from voxsrc 2020," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5809–5813.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[27] J. S. Chung., A. Nagrani., and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[28] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[29] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S.n Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[30] M. McLaren, L. Ferrer, D. Castán, and A. D. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Proc. Interspeech*, 2016, pp. 818–822.

[31] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.