INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

SpecMix : A Mixed Sample Data Augmentation method for Training with Time-Frequency Domain Features

Gwantae Kim[1], *David K. Han*[2] *and Hanseok Ko*[1]

[1]Korea University, South Korea
[2]Drexel University, USA

kgt1103211@korea.ac.kr, dkh42@drexel.edu, hsko@korea.ac.kr

Abstract

A mixed sample data augmentation strategy is proposed to enhance the performance of models on audio scene classification, sound event classification, and speech enhancement tasks. While there have been several augmentation methods shown to be effective in improving image classification performance, their efficacy toward time-frequency domain features of audio is not assured. We propose a novel audio data augmentation approach named "Specmix" specifically designed for dealing with time-frequency domain features. The augmentation method consists of mixing two different data samples by applying time-frequency masks effective in preserving the spectral correlation of each audio sample. Our experiments on acoustic scene classification, sound event classification, and speech enhancement tasks show that the proposed Specmix improves the performance of various neural network architectures by a maximum of 2.7%.

Index Terms: Acoustic scene classification, Data augmentation, Deep neural networks, Sound event classification, Speech enhancement

1. Introduction

Deep learning has shown remarkable successes on various audio processing tasks, such as speech enhancement [1, 2], Automatic Speech Recognition(ASR) [3, 4], sound classification [5–8], and speech synthesis [9, 10]. To further enhance their performance, many research efforts have focused on designing better network architectures for specific tasks. While improving the architectures may deliver better performances, these methods tend to overfit easily and require large amounts of training data [11, 12]. To avoid this problem, there have been some efforts in exploring data augmentation and regularization strategies.

For augmenting the audio dataset, there are two main approaches: time-domain waveforms and time-frequency domain features, such as spectrogram, mel-spectrogram, and mel-frequency cepstral coefficient. For the waveform data, the data augmentation strategies may include noise injection, changing pitch, changing speed, shifting time, and speed perturbation [13] of the waveform, to expand the data set without disturbing salient information therein. For the time-frequency domain features, Specaugment [11] proposed time warping, frequency-masking, and time-masking data augmentation strategies. Although Specaugment is successfully applied to ASR, its application to other tasks has been limited [14]. For example, in the speech enhancement task, zero-masking on the time and frequency axis tends to degrade the performance.

Corresponding Author:Hanseok Ko.

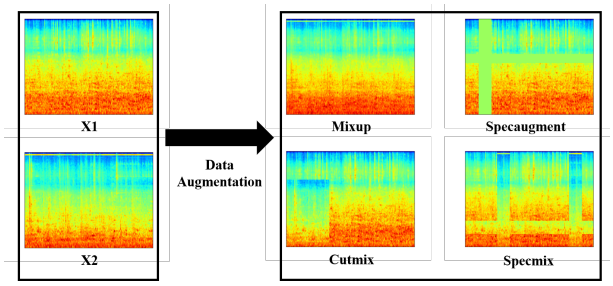

Figure 1: *Overview of the data augmentation methods:Mixup, Cutmix, Specaugment and our Specmix.*

Since the time-frequency domain features are two dimensional and can be projected as a 2D image, data augmentation strategies, particularly of Mixed Sample Data Augmentation (MSDA) type in the computer vision domain, have been applied to the time-frequency domain features as shown in Fig. 1. Mixup [15, 16] blends two images of the audio features and labels by varying a random parameter γ. Its performance has been shown to be effective in the image classification tasks, however, due to the way it mixes magnitudes of spectrograms from different source components together it is difficult to disentangle them in the audio domain. Thus, the performance from the Mixup approach has been limited. Cutout [17] and Specaugment [11] employ zero-masking to the image and spectrogram, respectively. Although these methods can be applied to images and spectrogram successfully, salient audio information can be lost due to zero-masking. Cutmix [18] randomly attaches a part of an image to another image. It applies a randomly generated mask for cutting a spectrogram region and pasting it randomly to another spectrogram region. While Cutmix can preserve magnitude information of X_1 and X_2, the time-frequency information taken from one image is randomly shifted to another resulting in a frequency shift.

As these methods have been adopted from image domain applications, their approach of mixing two different spectrograms has to be tailored for audio signals to preserve data distribution within a spectrogram structure for maintaining salient time-frequency correlations. Therefore, a different masking policy that preserves salient frequency features for MSDA is needed. Inspired by the previous data augmentation strategies in the speech and vision domains, we propose a novel audio data augmentation strategy, named SpecMix, for training with time-frequency domain features. The proposed method expands the idea of Cutmix, which attempted to cut-and-mix two data samples. However, their masking policy is based on image data, therefore it is not necessarily suitable for time-frequency domain features. To address this problem, we modified the mask-

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-103

ing policy tailored to time-frequency domain features. The proposed method can be integrated to ResNet, U-Net, and other state-of-the-art architectures for acoustic scene classification, sound event classification, or speech enhancement tasks.

2. Specmix Policy

In this section, we describe the Specmix algorithm in detail. We aim to construct an MSDA policy that directly acts on the time-frequency domain features, which improves the generalization of the model using time-frequency domain features.

2.1. Algorithm on classification tasks

Let $x \in \mathbb{R}^{F \times T \times C}$ and y denote a time-frequency domain features and its label, respectively. F denotes the number of the frequency bin, T denotes the number of the time bin and C denotes the number of time-frequency domain features. The goal of Specmix is to generate a new training sample $(\tilde{x}, \tilde{y})$ by combining two training samples (x_A, y_A) and (x_B, y_B). We define the combining operation as

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \quad (1)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (2)$$

where $\mathbf{M} \in \{0, 1\}^{F \times T}$ denotes a binary mask indicating where to drop out and fill in from two images, $\mathbf{1}$ is a binary mask filled with ones, and $\odot$ is element-wise multiplication. The combination ratio λ between two data points is the number of pixels of x_A in $\tilde{x}$.

In each training iteration, a mixed sample $(\tilde{x}, \tilde{y})$ is generated by combining two training samples selected from two mini-batches according to Equation (1) and (2).

2.2. Algorithm on speech enhancement tasks

Let $x \in \mathbb{R}^{F \times T \times C}$ and $z \in \mathbb{R}^{F \times T \times C}$ denote a time-frequency domain features of the noisy signal and related time-frequency domain features of the clean signal, respectively. The new training sample pair, $(\tilde{x}, \tilde{z})$, can be constructed by combining two training samples (x_A, z_A) and (x_B, z_B). We define the combining operation as

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B \quad (3)$$

$$\tilde{z} = \mathbf{M} \odot z_A + (\mathbf{1} - \mathbf{M}) \odot z_B \quad (4)$$

where $\mathbf{M} \in \{0, 1\}^{F \times T}$ denotes a binary mask indicating where to drop out and fill in from two images, $\mathbf{1}$ is a binary mask filled with ones, and $\odot$ is element-wise multiplication. In each training iteration, a mixed sample pair, $(\tilde{x}, \tilde{z})$, is generated by combining two training samples selected from two mini-batches according to Equation (1) and (2).

2.3. Masking

Figure 2 outlines the masking process of mixing audio features from two different samples. From frequency masking, we generate frequency bands up to three different segments while the time masking does the same in generating up to three different temporal bands. The starting frequencies and time are chosen from a random process of a uniform distribution from 0 to 1 while their widths are defined by the user. These bands form the binary mask enabling the mixing of two audio spectrograms.

1. Frequency masking: Up to three frequency bands can be masked depending on a random number f_{times} sampled from a uniform distribution from 0 to 3 in an integer value. In each frequency band selection, the starting band frequency f_{start} is selected by a random number from a uniform distribution from 0 to F. Next, the ending band frequency is calculated by the equation $f_{end} = f_{start} + \gamma F$ with γ chosen by a user-defined parameter between 0 to 1. Repeat this process f_{times} times.
2. Time masking: Again, up to three time bands can be masked in the identical process utilized for the Frequency masking described above. Again, γ, chosen by the user controls the width of the time band, and the random processes control the starting time and the number of temporal bands.

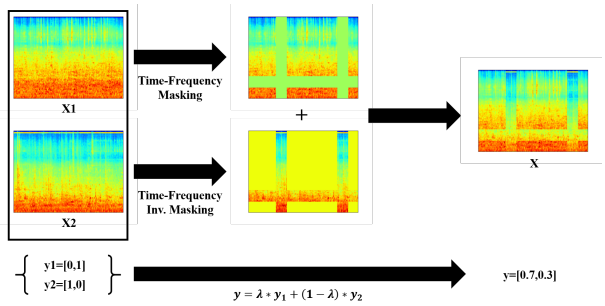

Figure 2: *Augmentation policies applied to the base input, given X_1 and X_2. To generate new sample X, applying a time-frequency mask to X_1, inverted time-frequency mask to X_2, and summation of them.*

3. Models

In this section, we introduce the preprocessing methods and models that were used to evaluate the proposed data augmentation strategy.

3.1. Acoustic scene classification & Sound event classification

The input time-frequency domain features are mel-spectrogram, its delta and delta-delta features. The sampling rate is 44.1kHz, nfft is 2048 and hop length is 1024, and the number of mel filters is 128. The shape of processed input is $[B, F, T, C] = [B, 128, T, 3]$, where B denotes the batch size, F denotes the number of frequency bins, T denotes the total time bins and C denotes the total number of channels. T depends on the audio length.

We used Resnet-101 [19] on acoustic scene classification and sound event classification tasks. Adam [20] optimizer and cross-entropy loss function were used to train our model with a batch size of 32. We employed learning rate decay from 1e-3 to 1e-7. We also employed the model and training procedure proposed in [5], considered state-of-the-art acoustic scene classification model, to evaluate the generalization performance of the proposed method.

3.2. Speech enhancement

During training, the input waveform was cut or padded to make the waveform length of 32768 samples. The input time-frequency domain feature is a spectrogram. The sampling rate

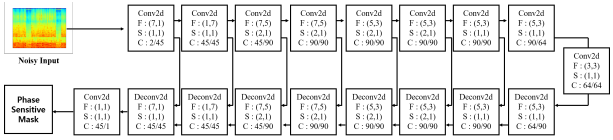

Figure 3: *A U-Net model for solving speech enhancement task.*

Table 1: *Comparison of state-of-the-art data augmentation methods for time-frequency domain features on TAU Urban Acoustic Scenes 2020 Mobile benchmark.(Model : ResNet-101)*

Model : ResNet-101	Accuracy(%)
No augmentation	59.60
Mixup [15]	58.15
Cutmix [18]	59.54
Specaugment [11]	57.68
Specmix $\gamma = 0.3$	**62.13**

is 16kHz, nfft is 512, and hop length is 256. The shape of the processed input is $[B, F, T, C] = [B, 256, T, 2]$, where B denotes the batch size, F denotes frequency bins, T denotes time bins and C denotes channels. T depends on the audio length. The first channel is the real part of the spectrogram and the second channel is the imaginary part of the spectrogram.

We build a. U-Net [21] style speech enhancement model illustrated in Fig. 3. The model consists of 8 encoder layers, 1 mid-level layer, 8 decoder layers, and 1 last convolution layer. To deal with arbitrary length inputs, the strides of all the layers on the time axis are not squeezed. The model predicts phase sensitive mask [22] and reconstructs a clean spectrogram by applying the mask to a noisy spectrogram. During training, Adam [20] optimizer and a mean squared error loss function is used to train our model with a batch size of 6. We employ learning rate decay from 1e-2 to 1e-5.

4. Experiments

In this section, we evaluate Specmix on acoustic scene classification, sound event classification, and speech enhancement tasks. The effects of Specmix on acoustic scene classification and sound event classification are examined first. We then compare the result with changing γ and masking policies. We also show that Specmix can improve noise reduction performance.

4.1. Acoustic scene classification

We evaluate Specmix on TAU Urban Acoustic Scenes 2020 Mobile benchmark [23], the dataset containing recordings from 12 European cities in 10 different acoustic scenes using 4 different recording devices. The dataset consists of 10 classes of urban acoustic scene recordings with 13965 labeled clips for training and 2970 clips for the test. The evaluation metric is accuracy.

Results with ResNet-101 model are given in Table 1. We observe that Specmix achieves the best result (62.13% accuracy) among the considered augmentation strategies. Specmix outperforms Mixup, Cutmix and Specaugment by +3.98%, +2.59% and +4.45%, respectively. Interestingly, Mixup, Cutmix, and Specaugment performed worse than with no augmentation. We believe that these augmentation strategies resulted in some information loss, as mentioned in Section 1, leading to poor performances.

We evaluate Specmix with $\gamma \in \{0.1, 0.3, 0.5, 0.7\}$ and $\mathcal{U}[0, 1]$ with ResNet-101 model. The results are given in Fig. 4. For all γ values considered, Specmix improves the performance over the case with no augmentation. The best performance is achieved when $\gamma = 0.3$.

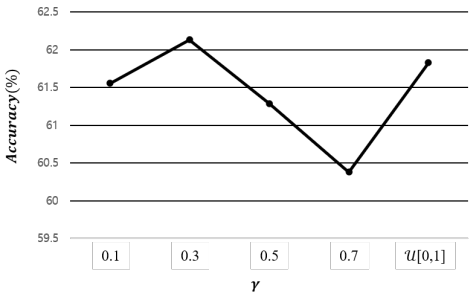

Figure 4: *Impact of γ on TAU Urban Acoustic Scenes 2020 Mobile benchmark.*

Table 2: *Comparison of state-of-the-art data augmentation methods for time-frequency domain features on TAU Urban Acoustic Scenes 2020 Mobile benchmark.(Model : [5])*

Model : [5]	Accuracy(%)
No augmentation	68.90
Mixup [15]	71.29
Cutmix [18]	70.92
Specaugment [11]	69.07
Specmix $\gamma = \mathcal{U}[0, 1]$	**71.60**

Results with Specmix and the other augmentation methods applied to [5]'s model are given in Table 2. We observe that Specmix improves the performance by +0.31%, +0.68%, +2.53% compared to Mixup, Cutmix, and Specaugment, respectively. We also performed an ablation study to observe the impacts of the masking strategies. The results are given in Table 3. Random masking is a masking strategy that randomly selects N pixels filled by X1 while the remaining pixels are filled by X2. Specmix(time only) only applies time masking and Specmix(freq. only) only applies frequency masking. The results show that both time masking and frequency masking improve performance.

4.2. Sound event classification

We evaluate on SECL_UMONS benchmark [24]. SECL_UMONS is a real sound recording dataset for simultaneous classification and localization of sound events. The data samples in the SECL_UMONS are recording in two indoor room environments. The room condition of the first one is RT60 of 0.7s with the room dimensions of $7.8 \times 3.6 \times 2.45$[m] and the second one is RT60 of 0.9s with room dimensions of $9.4 \times 7.5 \times 4.85$[m]. The dataset consists of 11 indoor event classes, such as *chair_movement*, *cup_drop_off*, *furniture_drawer*, *hand_clap*, *keyboard*, *knock*, *phone_ring*, *radio*, *speaker*, *step*, and *whistle*. The dataset contains 2178 sequences for the training set and 484 sequences for the validation set. We used single-channel waveform for our comparison. Evaluate metric is accuracy. Results with the ResNet-101 model are given in Table 4. Specmix achieves 97.107% accuracy on SECL_UMONS, +1.04% higher than the no augmentation case.

Table 3: *Impact of masking strategies on TAU Urban Acoustic Scenes 2020 Mobile benchmark.(Model : [5])*

Model : [5]	Accuracy(%)
Random masking	70.05
Specmix(time only)	70.42
Specmix(freq. only)	70.52
Specmix	**70.79**

Table 4: *Comparison of state-of-the-art data augmentation methods for time-frequency domain features on SECL_UMONS benchmark.(Model : ResNet-101)*

Model : ResNet-101	Accuracy(%)
No augmentation	96.07
Mixup [15]	95.87
Cutmix [18]	95.66
Specaugment [11]	96.90
Specmix $\gamma = \mathcal{U}[0,1]$	**97.11**

4.3. Speech enhancement

We evaluated Specmix on Voicebank + Diverse Environments Multichannel Acoustic Noise Database(DEMAND) benchmark, which is proposed by [25]. Noisy and clean speech recordings were provided from the DEMAND [26] and the Voice Bank corpus [27], respectively with each recorded with the sampling rate of 48kHz. A total of 40 different noise conditions are considered in the training set and 20 different conditions are considered in the test set. Finally, the training and test set contained 11572 and 824 noisy-clean speech pairs, respectively. Note that the speaker and noise classes were uniquely selected for the training and test sets. Evaluation metrics are Perceptual Evaluation of Speech Quality(PESQ), mean opinion score predictor of signal distortion(CSIG), background noise intrusiveness(CBAK), overall signal quality(COVL), and Segmental Signal to Noise Ratio(SSNR).

Results with the U-Net model are given in Table 5, where NA denotes No Augmentation, MU denotes Mixup, CM denotes Cutmix, and SA denotes SpecAugment. Mixup and Specaugment have failed to improve the speech enhancement performance over the no augmentation. However, Cutmix improves the performance of speech enhancement, and Specmix achieves the best results compared to the other data augmentation strategies.

We evaluate Specmix with $\gamma \in \{0.1, 0.3, 0.5, \mathcal{U}[0,1]\}$ with U-Net model. The results are given in Fig. 5. Although it is not shown, all the cases with γ values utilized in Specmix improved performance over the case with no augmentation. The best performance is achieved when $\gamma = 0.3$.

5. Conclusion

In this paper, we proposed a data augmentation strategy, named Specmix, for training with time-frequency domain features. While there are augmentation strategies for mixing two different audio sources, our proposed method was shown to preserve spectral information throughout the augmentation process. The proposed Specmix was shown easy to be incorporated into existing training pipelines and it has a low computational cost. From several evaluations of the proposed augmentation, the method improved the performance over various models on audio scene classification, sound event classification, and speech enhancement tasks. We expect that Specmix can be applied to various machine learning tasks using time-frequency domain features.

Table 5: *Comparison of state-of-the-art data augmentation methods for time-frequency domain features on Voice bank + DEMAND benchmark.(Model : U-Net).*

	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.67
NA	2.50	3.44	3.18	2.95	9.26
MU	2.44	3.39	3.16	2.89	9.43
CM	2.52	3.50	3.22	2.99	9.48
SA	2.41	3.48	3.16	2.93	9.35
Specmix	**2.54**	**3.60**	**3.24**	**3.05**	**9.57**

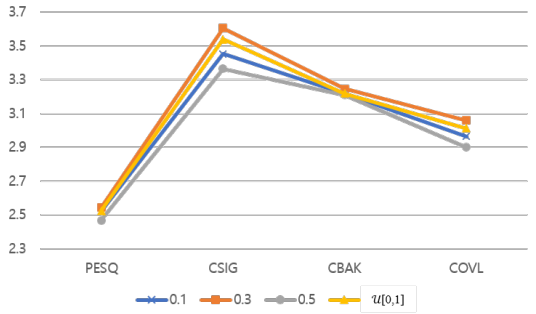

Figure 5: *Impact of γ on Voice Bank + DEMAND benchmark.*

6. Acknowledge

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-19-1-4001.

7. References

[1] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," *arXiv preprint arXiv:1911.04697*, 2019.

[2] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] M. D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 141–145.

[6] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[7] D. Kim, J. Park, D. K. Han, and H. Ko, "Dual stage learning based dynamic time-frequency mask generation for audio event classification," *Proc. Interspeech 2020*, pp. 836–840, 2020.

[8] Y. Lee, S. Park, and H. Ko, "A time delay convolutional neural network for acoustic scene classification," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2018, pp. 1–3.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[12] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6879–6883.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[16] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with cnn variants," DCASE2020 Challenge, Tech. Rep, Tech. Rep., 2020.

[17] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[18] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[24] M. Brousmiche, J. Rouat, and S. Dupont, "Secl-umons database for sound event classification and localization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 756–760.

[25] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[26] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[27] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.