Spectrogram Augmentation Using Image Processing Techniques

Hejung Yang
Yonsei University
hejung.yang@dsp.yonsei.ac.kr

Abstract

In speech processing, many data augmentation schemes exist regarding spectrogram as an image. These spectrogram augmentation plays an important role in improving performance, preventing overfitting and making up for sparsity of the existing clean dataset. In this paper, various image processing techniques are applied for several speech processing tasks. Specifically, both spatial-domain and frequency-domain filtering ideas from image processing domain are utilized. Additionally, commonly used image augmentation techniques are tested on the speech processing task. In order to see the flexibilities on importing image augmentation methodologies onto speech, both speaker identification and speaker verification tasks are considered as the target in the experiments. It could be seen that all of the techniques could introduce either performance gain or task-specific advantages, which increases anticipation on further applying augmentation ideas from image processing domain to the speech domain.

1. Introduction

In most supervised training of the deep neural network, the amount of training data heavily affects the model quality in real-world test sets. However, in many cases, there exists not enough amount of training data to get the satisfied results. Given sparse training set, the model tends to overfit with low generalization power, resulting low performance in evaluation. To mitigate the problem, various data augmentation schemes arose.

In speech processing related tasks, both time-domain, frequency-domain related augmentation methods are largely used to supplement the lack of dataset. Warping on either time or frequency axis is suggested from [3], [4]. [8] includes pitch shifting, dynamic range compression for further augmentation. When converting 1-dimensional signal into spectrogram by applying Short-Time Fourier Transorm and interpreting the signal as a mixture of temporal-frequency components, it is also possible to view the signal as an image. Some of the recent augmentation schemes perform augmentation base on the spectrogram, including SpecAugment [6] and SpecAugment++ [14].

This paper shows several spectrogram based augmentation schemes conducted with the help of image processing techniques, especially in spatial filter and notch filter. Furthermore, commonly used image augmentation methodologies are tested on speech processing domain.

2. Augmentation schemes

2.1. spatial filter on spectrogram

Joint bilateral filter ([2], [7]) is a variation of the bilateral filter first introduced to preserve precise edge components from the high-resolution depth image while smoothing out noisy artifacts.

Given input image I, reference image J with smoothing parameters σ_s, σ_r, joint bilateral smoothing works as follows:

$$JointBilateral(p, I, J, \sigma_s, \sigma_r) = \\ 1/K_p \sum_{q \in N(p)} exp(-\frac{||p-q||^2}{2\sigma_s^2} - \frac{||J(p)-J(q)||^2}{2\sigma_r^2})I(q) \tag{1}$$

$$K_p = \sum_{q \in N(p)} exp(-\frac{||p-q||^2}{2\sigma_s^2} - \frac{||J(p)-J(q)||^2}{2\sigma_r^2}) \tag{2}$$

where $N(p)$ stands for the set of neighboring pixels of p.

From the fact that joint bilateral filter utilizes reference image for better noise suppression, in case where both the original and noisy image are given, it is expected that the filter would generate the image which is almost identical to the original with appropriate adjustment of the parameters $sigma_s$ and $sigma_r$. However, as the nature of the bilateral filter is to "smooth out" the image, while it seems that the noisy components are completely removed, it is rather embedded all over the image. Thus it can be expected that the denoised image contains both the characteristics of the original image and the inherent noisy pattern. From this hypothesis, denoised spectrogram using the bilateral filter is used for training speech processing model.

Detailed process of augmenting the data, depicted in Fig. 1, is as follows: First, clean speech is added with noise

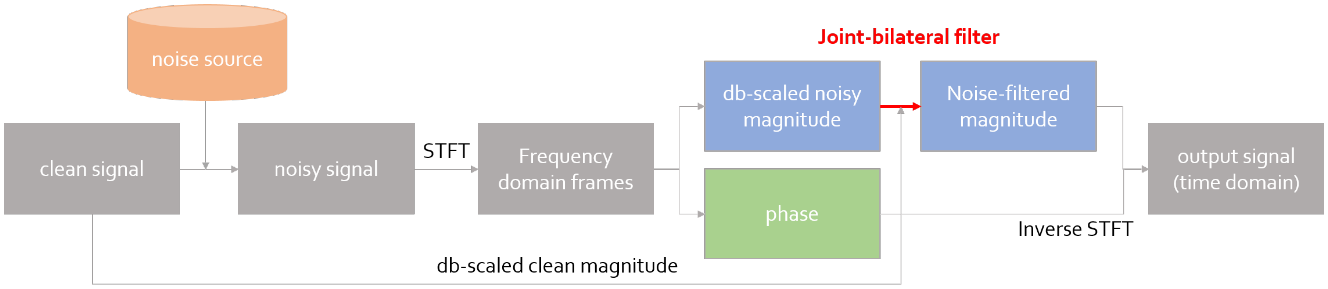

Figure 1. augmentation process using joint-bilateral filter. noised signal is filtered in magnitude-frequency domain for denoising.

sources given specific Signal-to-Noise ratio (SNR). Noise mixing is done in time domain of the signal. Next, the noisy signal is converted to complex-valued time-frequency components by Short-Time Fourier Transform. Joint-bilateral filter is applied on log-scaled magnitude of the frequency component after separating the complex value into magnitude and phase. Noise-filtered log-scale magnitude and the phase are then used to reconstruct time-domain signal.

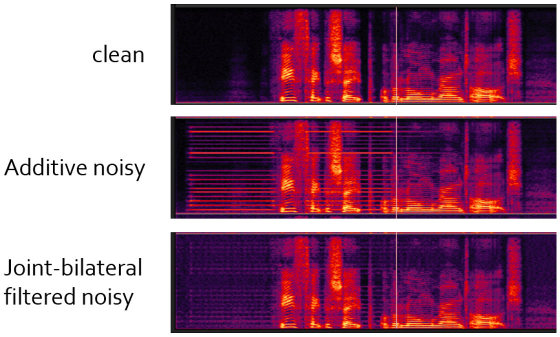

Figure 2. results of denoising noisy speech with joint-bilateral filter.

Fig. 2 shows spectral results of denoising noisy input with joint-bilateral filter on magnitude-frequency domain. Striped patterns shown along the whole frequency bins in noisy speech are largely suppressed after the denoising.

2.2. notch filter on spectro-temporal modulation

Modulation spectrum [5] is the evolution over time of the amplitude content of the various frequency bands ω of an STFT by a second Fourier Transform. From the definition of an STFT as $x(\omega, \tau)$, modulation spectrum $X(\omega, \Omega)$ is defined as follows:

$$x(\omega, \tau) = \frac{1}{\sqrt{2\pi}} \int_t x(t) h(\tau - t) e^{-j\omega t} dt \qquad (3)$$

$$X(\omega, \Omega) = \frac{1}{\sqrt{2\pi}} \int_\tau |x(\omega, \tau)| e^{-j\Omega\tau} d\tau \qquad (4)$$

where ω is the frequencies of the STFT, τ is the center-time of the STFT windows and Ω is the frequencies of the second Fourier Transform.

Spectro-temporal modulation ([1], [9]), also known as Modulation Power Spectrum (MPS), is the two-dimensional Fourier transform of the spectrogram which quantifies the power in temporal and spectral modulations. When regarding spectrogram as an image, MPS can be viewed as an output of Fourier transform of the image. From the idea of a notch filter from image domain, masking MPS on both temporal, spectral axis can help augmenting the signal.

Effect of masking speech in MPS domain can be found in Fig. 3. After applying 2-dimensional Fourier transform on the magnitude of the spectrogram, masking is done on either horizontal or vertical line with fixed thickness. Masked MPS is then inverse-transformed to attain new magnitude spectrogram. With the phase information on the clean spectrogram, final output of time-domain audio signal is generated. It can be seen in Fig. 3 that each of masking generates different periodic artifacts on the output spectrogram.

2.3. popular image augmentation techniques on spectrogram

Mixup [16] is a commonly used image augmentation scheme which adds several images pixel-wise and set its label as the smoothed version of the selected classes. It is proven in several image classification tasks that this simple idea of populating corpus improves performance with minimal implementational cost. One variation of [16] is CutMix [15] in which blockwise replacement of an image occurs, instead of pixel-wise mixing. It also outperformed the baseline model in both ImageNet and CIFAR tasks. In [16] simple speech augmentation result is also shown. In the experiment, the model is trained to classify a few of short voice commands including *yes, no, down* and mixing is done on padded spectrogram of the signals. In this paper, more sophisticated yet practical task is on target. Furthermore, several implementational variances in signal mixup schemes are compared, and analyzed.

One of the options in mixing speeches is to add two (or more) speeches in time domain, with each of the weights. Next option is to add magnitude-frequency of the speeches while preserving consistency on phase information; whilst

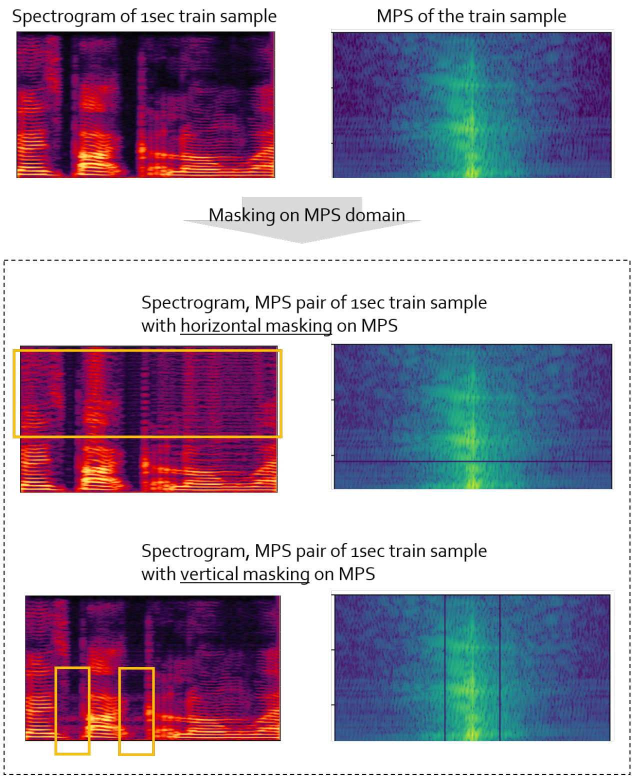

Figure 3. result of masking on MPS domain and convert into spectrogram. Noisy artifacts having periodicity on either time axis or frequency axis is added by the MPS masking.

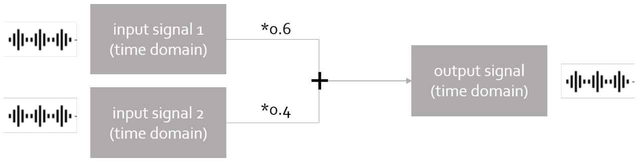

Figure 4. mixup of two speeches in time domain.

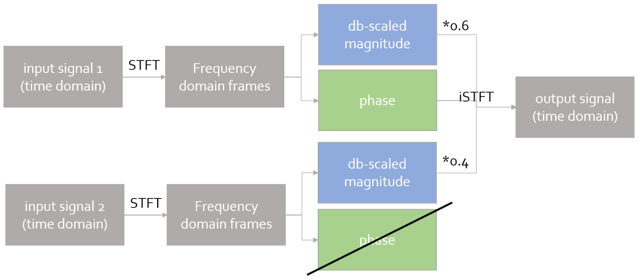

Figure 5. mixup of two speeches in magnitude-frequency domain.

applying weighted-sum on magnitude-frequency domain, phase of a speech having the largest weight is used for reconstruction of the output. It enables more conservative mixing than time domain sample-wise mixing. Fig. 4 and Fig. 5 shows the process of the two mixing schemes. It can be found from Fig. 6 that mixup in magnitude-frequency domain preserves most of the low-band component from Speech 1.

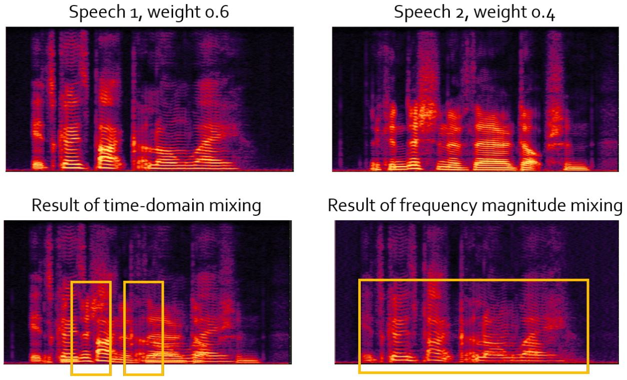

Figure 6. result after mixup in both time domain and magnitude-frequency domain.

3. Experiments

3.1. spatial filter on spectrogram

To see the effectiveness of spatially-denoised corpus, speaker identification task is used as the target. In speaker identification, the model is trained to classify the speaker id given the speech signal, assume that every speaker in test set is contained in training set.

Training and evaluation are done on 100 speakers of VCTK [13] corpus, so that the model classifies given signal to one of 100 speaker ids. For training, total 1.4 hours of the data is used and 30 minutes of the data is used for evaluation. Noise augmentation on clean testset is done to test noise robustness of the model. Noise sources for augmentation are from DEMAND [12] and MUSAN [10]. There exists several types of noises in the noise sources, including music, station, traffic, etc. As the baseline, simple Time-delayed neural network (TDNN) of size about 178KB is used.

Evaluation results can be found in Tab. 1. It can be seen that mixing clean corpus with additive noise augmented corpus makes the model robust to the noisy testset while the performance degrades on the clean testset. The tradeoff between the noise robustness and the performace on clean testset can be controlled by the SNRs on noise augmentation; as the SNR gets higher, the accuracy on clean testset increases while the noise robustness decreases. Using joint-bilateral noisy data does not harm the clean testset performance while handling the noisy testset.

3.2. notch filter on spectro-temporal modulation

Instead of speaker identification which is more suitable for checking noise-robustness of the model, speaker veri-

Training corpus	Accuracy on clean testset	Accuracy on noisy(10db) testset
Clean	89.0%	24.0%
Clean + additive(10db)	84.0%	61.6%
Clean + additive(20db)	85.2%	49.0%
Clean + additive(30db)	88.8%	38.0%
Clean + joint-bilateral	**89.4%**	54.4%
Clean + additive(10db) + joint-bilateral	85.6%	**61.8%**

Table 1. Results on several types of training corpus.

Training corpus	EER on testset	Threshold for the EER (range: 0-1)
Clean	**1.94719%**	0.784438
Clean + MPS0	2.44224%	0.789325
Clean + MPS1	2.11221%	0.788875
Clean + MPS0 + MPS1	2.0462%	**0.805046**

Table 2. Results on several types of MPS augmentations. Thresholds along with the EER scores are reported.

fication task is used to test the feasibility of notch filtering on MPS domain. Speaker verification model classifies whether an audio sample belongs to a predetermined person or not, thus evaluation can be done with speech of unseen speaker. For performance evaluation, Both Equal error rate(EER) and its threshold value are used. EER is the crosspoint where False acceptance rate equals False rejection rate, so lower EER stands for better overall verification performance.

For training and evaluation, 100 speakers on VCTK are splited by 80, 20 speakers respectively. Total 40 hours of speech data is used in training and 6 hours are used for evaluation. No external noise augmentation is used in the task. X-vector [11] of size 2.8MB is used as the base model.

Evaluation results can be found in Tab. 2, where MPS0 refers "masking on axis 0 of the MPS" and MPS1 refers "masking on axis 1 of the MPS". It is shown that adding MPS augmented speech degrades performance and applying MPS on axis 1 results better performance than on axis 0. However, it can be seen that the threshold on the EER point got increased, implying the fact that the model gets more robust to false-acceptance.

3.3. popular image augmentation techniques on spectrogram

As the above section, speaker verification task is used to prove the effectiveness of the augmentation schemes. Training, evaluation set are identical as before. X-vector of size around 178KB is first used to test various combinations of corpora, then combination recipes included in top-3 performances are further trained on bigger model of size 2.8MB.

Evaluation results can be found in Tab. 3. In the table, Mix2 means "mixing two different time-domain signals", while MixMag2 refers "mixing two different frequency magnitude signals" and MixCut2 refers "mixing two different time-domain signals with cutting scheme". Parentheses in a combination such as "Clean(2) + Mix2(1)" refers probabilities of sampling training batch from each corpus type.

It can be seen that EER on big model gets improved when adding augmented corpus. Model trained with clean corpus gets best EER on small model, but tends to overfit on bigger size. With additional augmentation, overfitting is alleviated on the big model.

4. Conclusions

In this paper, spatial filtering, frequency-domain augmentation(MPS) and mixup augmentation techniques are covered. It could be seen in spatial filtering experiments that noise suppressed corpus tends to maintain the accuracy on clean testset while helping model deal with noisy testset. Further different noise suppression techniques can be further checked besides of the joint-bilateral filter.

Results on MPS augmentation showed degradation of the performance, while the model becomes more robust to false-acceptance. Experiments on different masking scheme, with fine-tuned parameters on masking can be further investigated to improve the overall performance. ALso, false-acceptance resistance can be another main topic of the augmentation scheme depending on the application of the model. Finally, it is shown that common image augmentation techniques can improve the performance of speech processing applications. For future work additional image augmentation schemes proven to be useful in image domain can be applied to the speech domain.

References

[1] Taishih Chi, Yujie Gao, Matthew C Guyton, Powen Ru, and Shihab Shamma. Spectro-temporal modulation trans-

Training corpus	Model size 178KB	Model size 2.8MB
Clean	**2.34323%**	1.94719%
Clean(2) + Mix2(1)	2.64026%	-
Clean(2) + MixMag2(1)	2.50825%	1.91419%
Clean(2) + MixCut2(1)	3.49835%	-
Clean(4) + MixMag2(2) + MixMag4(1)	2.54125%	-
Clean(9) + MixMag2(3) + MixMag4(1)	2.47525%	**1.81518%**

Table 3. Results on several types of mixup, CutMix combinations.

fer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, 106(5):2719–2732, 1999. 2

[2] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM transactions on graphics (TOG)*, 23(3):673–678, 2004. 1

[3] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, page 21, 2013. 1

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*, 2015. 1

[5] Ugo Marchand and Geoffroy Peeters. The modulation scale spectrum and its application to rhythm-content analysis. In *DAFX (Digital audio effects)*, 2014. 2

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 1

[7] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG)*, 23(3):664–672, 2004. 1

[8] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283, 2017. 1

[9] Nandini C Singh and Frédéric E Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394–3411, 2003. 2

[10] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015. 3

[11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018. 4

[12] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035081. Acoustical Society of America, 2013. 3

[13] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017. 3

[14] Helin Wang, Yuexian Zou, and Wenwu Wang. Specaugment++: A hidden space data augmentation method for acoustic scene classification. *arXiv preprint arXiv:2103.16858*, 2021. 1

[15] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2