RELATED WORK

Введение

Исследование методов аугментации актуально в настоящее время. Понятию аугментации сложно дать точное определение, в данной работе под ней будем понимать создание новых данных на основе уже имеющихся. С помощью аугментации можно существенно расширить объем обучающей выборки, что особенно хорошо в тех случаях, когда исходных данных не очень много. Применение аугментации способно повысить обобщающую способность модели. Аугментация данных используется при решении многих задач глубинного обучения, связанных с обработкой изображений, звуковых данных, текстов.

В данной работе исследуются методы аугментации применительно к звуковым данным, а именно к мел-спектрограммам. Мел-спектрограммы подаются на вход нейронной сети как изображения, поэтому многие подходы к аугментации изображений применимы и к аудиоданным. Например, метод Random Erasing [1], сводящийся к вырезанию случайных прямоугольников из изображения, может быть использован в задаче аудиоклассификации [2]. Также в задаче классификации звуковых данных применяются такие методы аугментации, как Shift Augmentation [3] - сдвиг мел-спектрограммы влево или вправо, Noise Augmentation [3] - добавление Гауссовского шума, Loudness Augmentation [3] - регулирование громкости, Speed augmentation [3] - ускорение или замедление аудиозаписи.

SpecAugment [4] - один из наиболее известных методов аугментации аудиоданных, который показал свою эффективность в задаче автоматического распознавания речи. Политика аугментации SpecAugment определяется 3 возможными преобразованиями: Time warping, Frequency masking, Time masking. В настоящее время известны некоторые модификации SpecAugment: SpliceOut [5], SpecAugment++ [6], FilterAugment [10].

Методы аугментации, основанные на mixup [7], также нашли применение в задачах, связанных с аудиоданными: MIXSPEECH [8], SPATIAL MIXUP [9]. Также

возможны комбинации mixup [7] с другими методами аугментациями:
SpecMix [11], Cutmix [12].

Существуют подходы [13], [14] к аугментации звуковых данных, основанные на GAN [15]. Однако в данной работе подобные методы рассматриваться не будут.

Методы аугментации мел-спектрограмм не ограничиваются лишь преобразованиями соответствующих изображений. Существуют алгоритмы применения этих преобразованиях. В [16] предложена адаптивная весовая схема для аугментации временных рядов. Подобные алгоритмы применимы и для аугментации мел-спектрограмм.

Существующие методы аугментации

В этом разделе представлены известные методы аугментации аудиоданных, которые будем исследовать и на основе которых будут предложены некоторые из новых подходов.

Здесь и далее считаем, что FreqSize – размерность мел-спектрограммы по частотной оси, TimeSize – размерность мел-спектрограммы по временной оси. S – матрица значений мел-спектрограммы.

Также введем матрицу $M(I, J)$, где I, J - множества индексов:

$$M(I,J) = \{M(i,j)\} = \begin{cases} 0, & (i,j) \in I \times J, \\ 1, & \text{иначе.} \end{cases}$$

Стоит рассматривать только случаи, когда в представленных ниже аугментациях значения t, f, shift ненулевые. В противном случае ($t = 0$, или $f = 0$, или shift $= 0$) мел-спектрограмма никак не изменяется.

1. TimeMasking[1] [4]

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.

   В результате применения аугментации:

   $S \to S \cdot M(\{0, \ldots, \text{FreqSize} - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

2. FreqMasking[2] [4]

   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, F - параметр аугментации.

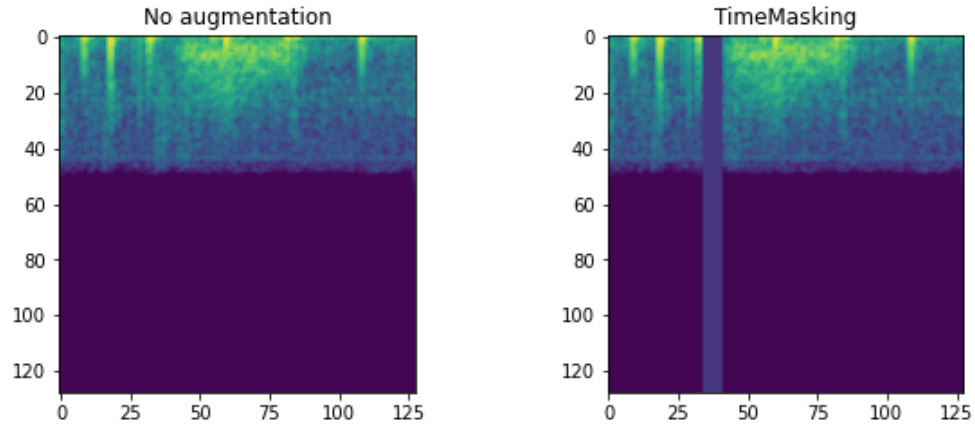


Рис. 1: TimeMasking

В результате применения аугментации:

$S \to S \cdot M(\{f_0, \ldots, f_0 + f - 1\}, \{0, \ldots, \text{TimeSize} - 1\})$

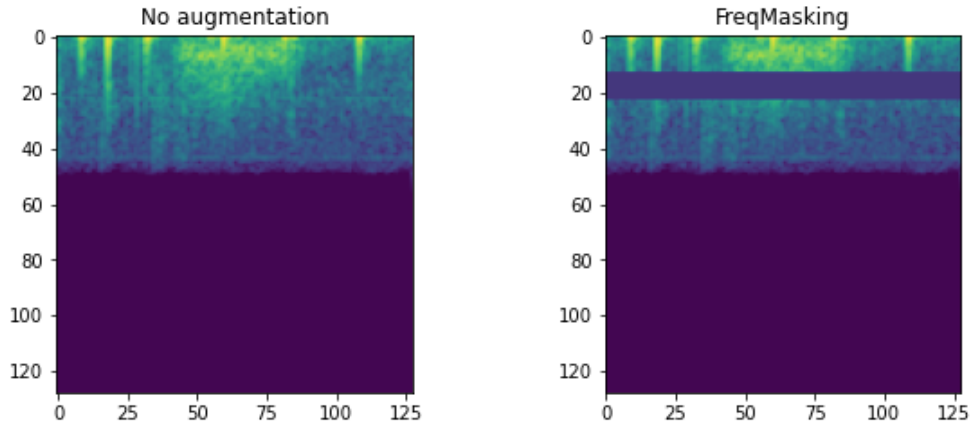


Рис. 2: FreqMasking

3. Noise[3] [3]

   К каждому значению в мел-спектрограмме добавляется $g \sim N(0, \sigma)$ (для каждого значения мел-спектрограммы генерируется свое g), где σ - параметр аугментации (в данной работе $\sigma = 0.01$).

4. TimeShift[4] [3]

   Сдвигаем все значения мел-спектрограммы относительно временной оси влево или вправо на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига определяется знаком shift: если

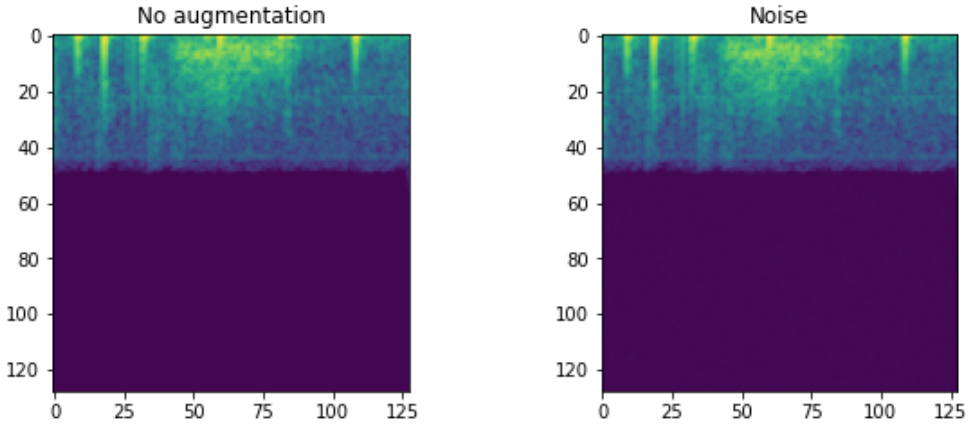


Рис. 3: Noise

shift > 0, происходит сдвиг вправо, если shift < 0 - влево. Пустая область, образующаяся в результате сдвига, заполняется нулями.

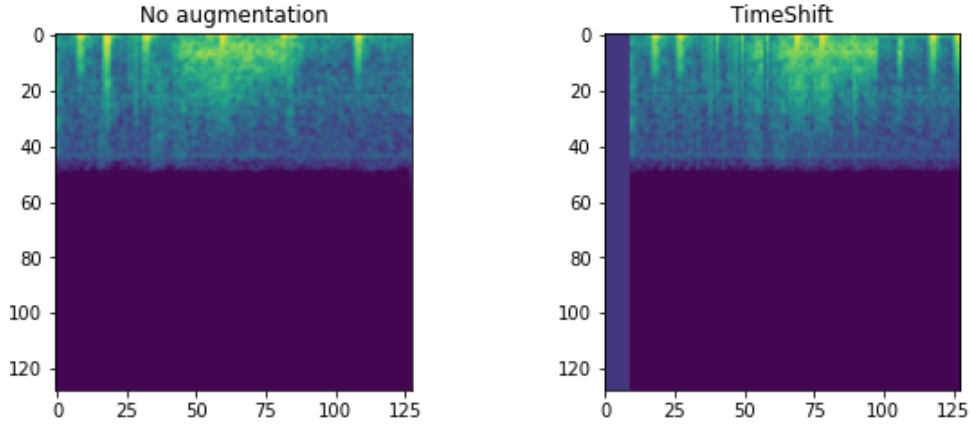


Рис. 4: TimeShift

5. RandomErasing[5] [1]

   Случайное вырезание прямоугольника в мел-спектрограмме.

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - t - 1\},$

   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - f - 1\}$, T, F - параметры аугментации.

   В результате применения аугментации:

   $S \to S \cdot M(\{f_0, \ldots, f_0 + f - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

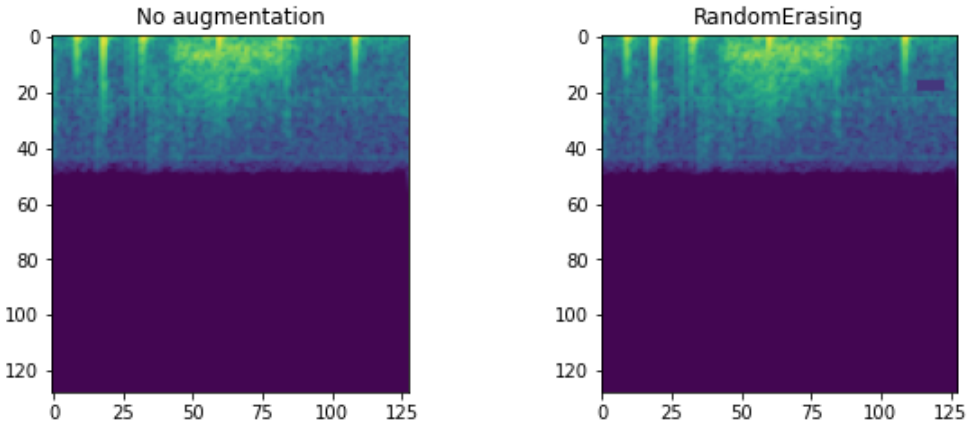


Рис. 5: RandomErasing

6. Mini-batch based mixture masking[6] (MM) [6]

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.

   $M_1 = M(\{0, \ldots, \text{FreqSize} - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, F - параметр аугментации.

   $M_2 = M(\{f_0, \ldots, f_0 + f - 1\}, \{0, \ldots, \text{TimeSize} - 1\})$

   $M = M_1 \cdot M_2$

   Y - матрица значений мел-спектрограммы другого объекта из того же мини-батча

   В результате применения аугментации:

   $S \to S \cdot M + (1 - M) \cdot \frac{S+Y}{2}$

7. Mini-batch based cutting masking[7] (CM) [6]

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.

   $M_1 = M(\{0, \ldots, \text{FreqSize} - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, F - параметр аугментации.

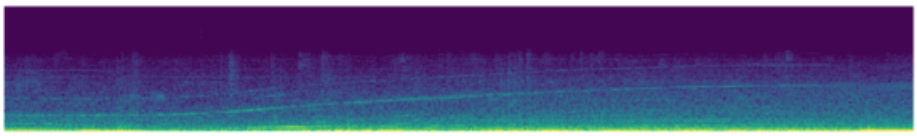
   $M_2 = M(\{f_0, \ldots, f_0 + f - 1\}, \{0, \ldots, \text{TimeSize} - 1\})$

   $M = M_1 \cdot M_2$

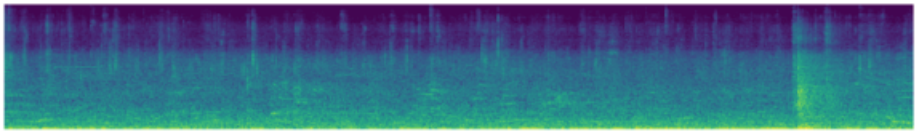
   Y - матрица значений мел-спектрограммы другого объекта из того же мини-батча

   В результате применения аугментации:

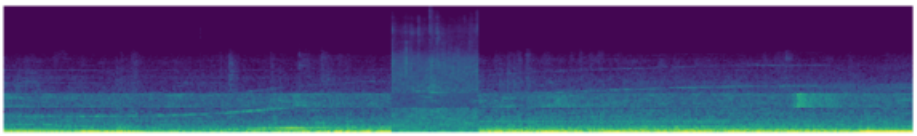
   $S \to S \cdot M + (1 - M) \cdot Y$



The target sample.

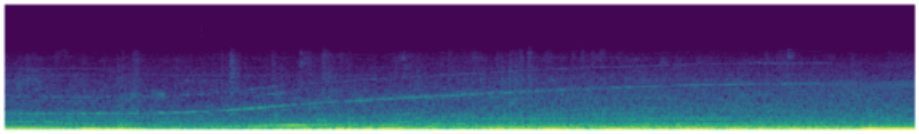


Another sample within the same mini-batch.

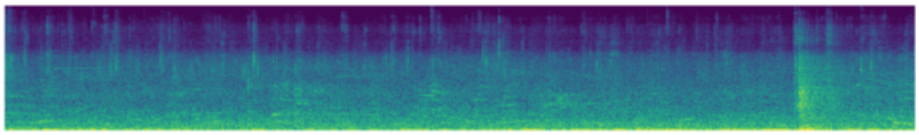


The augmented sample by MM.

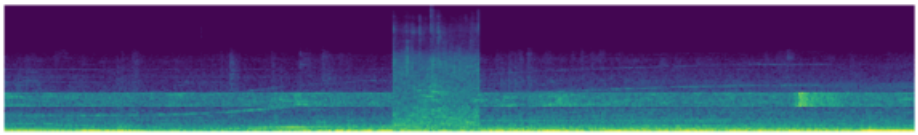
Рис. 6: Mini-batch based mixture masking



The target sample.



Another sample within the same mini-batch.



The augmented sample by CM.

Рис. 7: Mini-batch based cutting masking

8. Mixup[8] [7]

   Есть 2 объекта обучающей выборки (x_i, y_i), (x_j, y_j)

   Тогда новый сэмпл (x_{new}, y_{new}) получается следующим образом:

$x_{new} = \lambda x_i + (1 - \lambda)x_j$

$y_{new} = \lambda y_i + (1 - \lambda)y_j$, где $\lambda \sim Beta(\alpha, \alpha)$,

α - фиксированный параметр

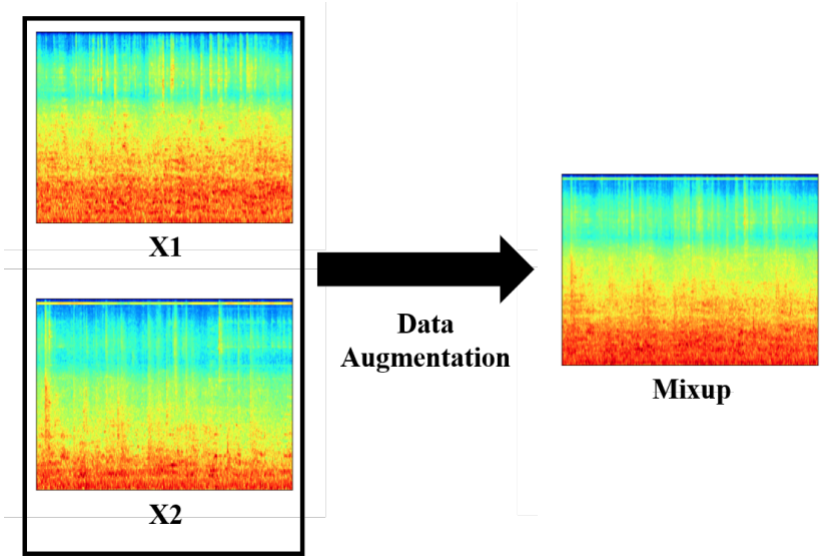


Рис. 8: Mixup

9. Cutmix[9] [12]

   Есть 2 объекта обучающей выборки (x_i, y_i), (x_j, y_j)

   Тогда новый сэмпл (x_{new}, y_{new}) получается следующим образом:

   $M_0 = M(\{r_x, \ldots, r_x + r_w - 1\}, (\{r_y, \ldots, r_y + r_h - 1\})$

   $x_{new} = M_0 \cdot x_i + (1 - M_0) \cdot x_j$

   $y_{new} = \lambda y_i + (1 - \lambda)y_j$, где $\lambda \sim Beta(\alpha, \alpha)$,

   α - фиксированный параметр,

   $r_x \sim Unif(0, W)$, $r_w = [W\sqrt{1 - \lambda}]$,

   $r_y \sim Unif(0, H)$, $r_h = [H\sqrt{1 - \lambda}]$

   Данный метод представляет собой комбинацию Mixup [7] и RandomErasing [1].

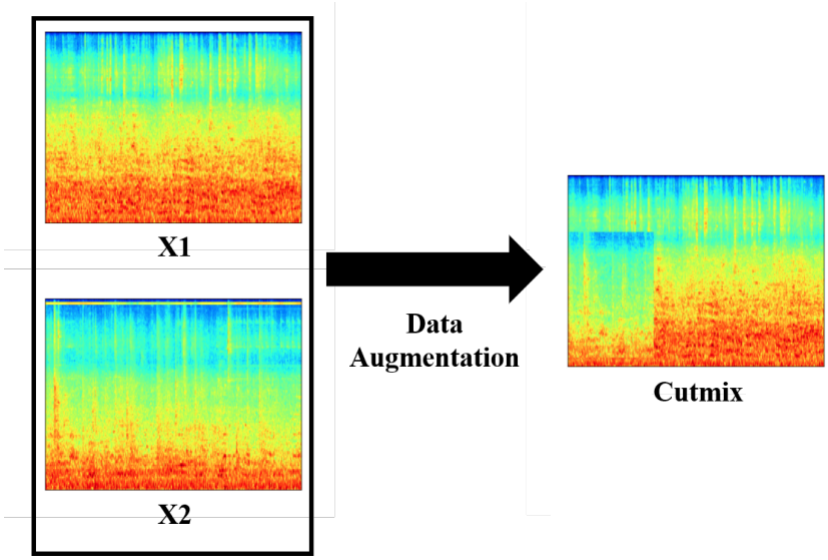


Рис. 9: Cutmix

10. SpecMix[10] [11]

    Есть 2 объекта обучающей выборки (x_i, y_i), (x_j, y_j)

    Тогда новый сэмпл (x_{new}, y_{new}) получается следующим образом:

    $x_{new} = M \cdot x_i + (1 - M) \cdot x_j$

    $y_{new} = \lambda y_i + (1 - \lambda) y_j$, где

    матрица M определяется 2 стратегиями маскирования:

    - Time masking [4]
    - Frequency masking [4]

    λ - доля единиц среди всех значений матрицы M

    Данный метод представляет собой комбинацию Mixup [7] и SpecAugment [4].

Стоит отметить, что в данной работе мел-спектрограммы нормализуются следующим образом:

$value = \frac{value - mean}{std}$, где mean – математическое ожидание значений мел-спектрограммы, std – стандартное отклонение.

Поэтому замена некоторых значений мел-спектрограммы на 0 в результате применения аугментации – это замена на математическое ожидание.

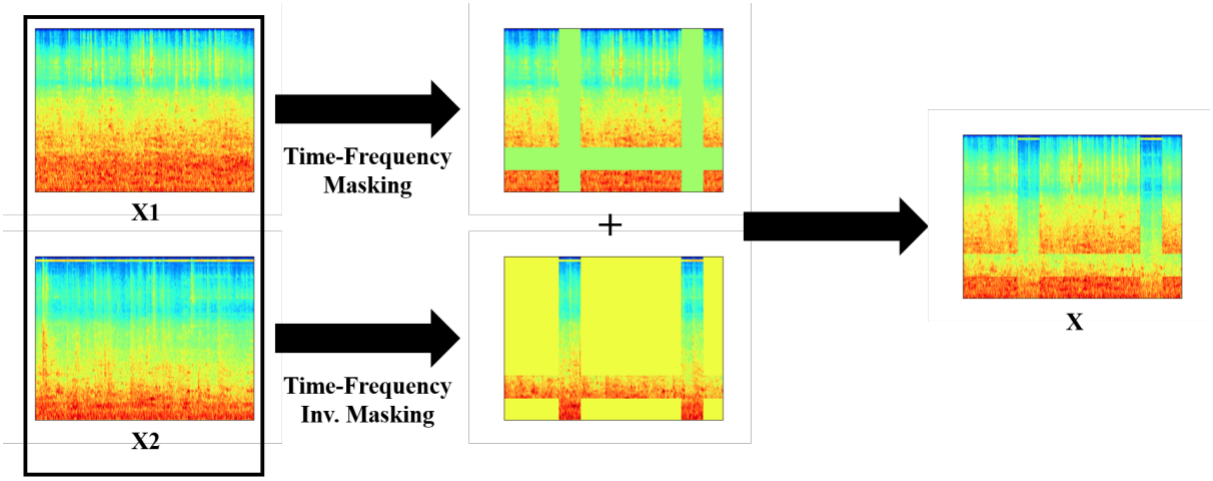


Рис. 10: SpecMix

Список литературы

[1] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, Yi Yang, "Random Erasing Data Augmentation". 2017.
https://arxiv.org/pdf/1708.04896.pdf

[2] Haiwei Wu, Lin Zhang, Lin Yang, Xuyang Wang, Junjie Wang, Dong Zhang, Ming Li, "Mask Detection and Breath Monitoring from Speech: on Data Augmentation, Feature Representation and Modeling". 2020.
https://arxiv.org/pdf/2008.05175.pdf

[3] Steffen Illium, Robert Muller, Andreas Sedlmeier and Claudia Linnhoff-Popien, "Surgical Mask Detection with Convolutional Neural Networks and Data Augmentations on Spectrograms". 2020.
https://arxiv.org/pdf/2008.04590.pdf

[4] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". 2019.
https://arxiv.org/pdf/1904.08779.pdf

[5] Arjit Jain, Pranay Reddy Samala, Deepak Mittal, Preethi Jyoti, Maneesh Singh, "SpliceOut: A Simple and Efficient Audio Augmentation Method". 2021.
https://arxiv.org/pdf/2110.00046.pdf

[6] Helin Wang, Yuexian Zou, Wenwu Wang, "SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification". 2021. https://arxiv.org/pdf/2103.16858.pdf

[7] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz, "mixup: Beyond Empirical Risk Minimization". 2017. https://arxiv.org/pdf/1710.09412.pdf

[8] Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, Bo Xu, "MIXSPEECH: DATA AUGMENTATION FOR LOW-RESOURCE AUTOMATIC SPEECH RECOGNITION". 2021. https://arxiv.org/pdf/2102.12664.pdf

[9] Ricardo Falcon-Perez, Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Yuki Mitsufuji, "SPATIAL MIXUP: DIRECTIONAL LOUDNESS MODIFICATION AS DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION". 2021. https://arxiv.org/pdf/2110.06126.pdf

[10] Hyeonuk Nam, Seong-Hu Kim, Yong-Hwa Park, "FILTERAUGMENT: AN ACOUSTIC ENVIRONMENTAL DATA AUGMENTATION METHOD". 2021. https://arxiv.org/pdf/2110.03282.pdf

[11] Gwantae Kim, David K. Han, Hanseok Ko, "SpecMix : A Mixed Sample Data Augmentation method for Training withTime-Frequency Domain Features". 2021. https://arxiv.org/pdf/2108.03020.pdf

[12] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features". 2019. https://arxiv.org/pdf/1905.04899.pdf

[13] Nhat Truong Pham, Duc Ngoc Minh Dang, and Sy Dzung Nguyen, "Hybrid Data Augmentation and Deep Attention-based Dilated Convolutional-Recurrent Neural Networks for Speech Emotion Recognition". 2021. https://arxiv.org/pdf/2109.09026.pdf

[14] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, Helen Meng, "Adversarial Data Augmentation for Disordered Speech Recognition". 2021.
https://arxiv.org/pdf/2108.00899.pdf

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Nets". 2014.
https://arxiv.org/pdf/1406.2661.pdf

[16] Elizabeth Fons, Paula Dawson, Xiao-jun Zeng, John Keane, Alexandros Iosifidis, "Adaptive Weighting Scheme for Automatic Time-Series Data Augmentation". 2021.
https://arxiv.org/pdf/2102.08310.pdf