DOI 10.31891/2307-5732-2024-341-5-54
УДК 004.8

ПІРКО АНДРІЙ
Національний лісотехнічний університет України
https://orcid.org/0009-0007-9056-0413
e-mail: pirko.andrii@nltu.lviv.ua
БОРЕЦЬКА ІРИНА
Національний лісотехнічний університет України
https://orcid.org/0000-0002-6767-104X
e-mail: boretska@nltu.edu.ua

# МЕТОДИ АУДІО-АУГМЕНТАЦІЇ У МОДЕЛЯХ МАШИННОГО НАВЧАННЯ

*У цій статті розглядається вплив технік аудіо-аугментації на класифікацію гітарних акордів. Аудіо-аугментація, як метод розширення навчальних датасетів шляхом модифікації аудіосигналів, є важливим інструментом для покращення стійкості моделей до різних варіацій сигналів. Після застосування методів аугментації, таких як додавання шуму, зміна швидкості, реверберація та часовий зсув, було проведено навчання згорткової нейронної мережі (CNN) на розширеному датасеті гітарних акордів. Результати експерименту продемонстрували значне підвищення точності класифікації в порівнянні з моделями, навченими на неаугментованих даних. Отримані дані свідчать про те,, що вибір конкретних технік аугментації залежить від типу завдання, а їх впровадження в моделі машинного навчання відкриває нові можливості для підвищення ефективності аудіоаналізу.*

*Ключові слова: аудіо-аугментація, машинне навчання, нейронні мережі, обробка аудіосигналів.*

PIRKO ANDRII, BORETSKA IRYNA
Ukrainian National Forestry University

## AUDIO AUGMENTATION METHODS IN MACHINE LEARNING MODELS

*This paper explores the role of audio augmentation techniques in enhancing the classification of guitar chords using machine learning. In the field of audio analysis, especially for tasks like chord classification, obtaining a sufficiently large and diverse dataset can often be challenging. Audio augmentation addresses this issue by synthetically increasing the size and diversity of the training dataset, thereby allowing models to generalize better to unseen data. By modifying audio signals in specific ways, such as adding noise, altering speed, applying reverb, and shifting the timing of signals, augmentation enables the creation of varied versions of the original audio. This helps in simulating real-world scenarios, where audio inputs can be distorted due to various factors such as environmental noise, recording equipment limitations, or differences in instrument performance.*

*The study employs a convolutional neural network (CNN) architecture for the classification task, a choice motivated by CNNs' effectiveness in learning spatial hierarchies and patterns, which are crucial for recognizing features in audio spectrograms. The dataset of guitar chords, initially limited in scope, was augmented with various techniques, each chosen to mimic different types of distortions or variations that a chord signal might encounter in practice. For instance, noise addition simulates interference or background sound, speed modification accounts for variations in tempo, reverb mimics the effects of different acoustic environments, and time shifting introduces subtle timing variations often seen in live recordings.*

*These transformations expand the dataset, ensuring the model is exposed to a broad spectrum of variations, which enhances its ability to generalize to new, unseen audio samples. The CNN trained on this augmented dataset exhibited significantly higher classification accuracy compared to models trained on the original, non-augmented dataset. This finding underscores the importance of data diversity in training machine learning models, particularly for audio classification tasks where real-world data often contains unpredictable variations.*

*Keywords: audio augmentation, machine learning, neural networks, audio signal processing.*

## Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

В сучасних умовах розвиток систем штучного інтелекту та машинного навчання тісно пов'язаний з обробленням великих обсягів даних, зокрема аудіосигналів. Однак аудіодані можуть бути вразливими до різних видів спотворень та шумів, що знижує ефективність навчання моделей та їхню здатність узагальнювати знання на нові, незнайомі зразки. Проблема полягає в тому, що реальні аудіодані часто є недостатньо репрезентативними для всіх можливих сценаріїв, а це, в свою чергу, може призводити до перенавчання моделей або їх поганої продуктивності при роботі з новими даними.

Техніки аудіо-аугментації можуть суттєво покращити якість навчання моделей машинного навчання, дозволяючи створювати синтетичні варіації даних і таким чином збільшувати обсяг і різноманітність навчальних наборів. Це особливо важливо у випадках, коли отримання реальних аудіоданих є складним, дорогим або обмеженим. Таким чином, застосування аудіо-аугментації сприяє підвищенню стійкості моделей до шуму та непередбачуваних спотворень, що є ключовим науковим завданням у сфері розпізнавання мовлення, обробки музики, класифікації звуків та інших завдань.

З практичної точки зору, аудіо-аугментація забезпечує поліпшення роботи таких систем, як голосові асистенти, системи безпеки на основі звуку, автоматизоване розпізнавання мовлення, музичні рекомендаційні системи, а також системи моніторингу та аналізу звукових подій. Використання цих технік є важливим кроком у напрямку до створення більш адаптивних, точних і надійних систем штучного інтелекту для роботи з аудіоданими в реальних умовах.

Упродовж останніх років дослідження у сфері аудіо-аугментації активно розвиваються, оскільки це ключовий інструмент для поліпшення ефективності моделей машинного навчання, зокрема в задачах розпізнавання мовлення та класифікації звуків.

**Аналіз досліджень та публікацій**

Одним із перших досліджень, яке привернуло увагу до аудіо-аугментації, є робота автора Yu Zhang [1], який запропонував інноваційну техніку SpecAugment, яка полягає у модифікації спектрограм шляхом часового та частотного маскування. Цей підхід дозволив значно поліпшити результати розпізнавання мовлення на таких наборах даних, як LibriSpeech, і став основою для подальших досліджень в аудіо-аугментації.

Також слід відзначити роботу [2], яка надає огляд різних технік аугментації аудіоданих, таких як зміна швидкості, додавання шуму та перекривання сигналів. Автори розглянули ефективність цих методів у задачах класифікації звукових подій і показали, що аугментація значно підвищує продуктивність нейронних мереж у реальних умовах.

У роботі [3] досліджується використання глибоких нейронних мереж для оброблення аудіосигналів, включаючи аугментацію даних. Книга містить практичні рекомендації щодо інтеграції аугментації в процес навчання моделей, особливо у таких завданнях, як розпізнавання мовлення та музичний аналіз.

Дослідження [4] вивчає вплив аугментації на стійкість моделей до різних видів шуму. Автори пропонують використовувати кілька технік, таких як додавання білого шуму та зміну гучності для покращення якості розпізнавання мовлення в реальних умовах, де сигнал часто спотворений зовнішніми факторами.

Ці дослідження заклали фундамент для подальших експериментів і впровадження нових методик аудіо-аугментації. Вони демонструють, що такий підхід є критично важливим для підвищення точності та стійкості моделей глибокого навчання, особливо в контексті завдань, де реальні аудіодані часто мають шум або непередбачувані спотворення.

**Виклад основного матеріалу**

Техніки аудіо-аугментації набули великого значення в машинному навчанні, особливо при вирішенні задач, пов'язаних із звуковими даними. Використовуючи ці техніки можна значно розширити навчальні вибірки та покращити узагальнювальну здатність моделей, адаптуючи їх до різних акустичних сценаріїв [5]. Основною метою цієї роботи є огляд та аналіз існуючих методів аудіо-аугментації, оцінка їхнього впливу на моделі нейронних мереж, інтеграція з сучасними архітектурами, а також рекомендації щодо застосування цих методів у реальних проєктах.

Аудіо-аугментація передбачає модифікацію оригінальних аудіосигналів за допомогою різних технік для створення нових, змінених версій цих сигналів. Ключовими техніками аугментації є:

1. **Додавання шуму**. Один із найбільш поширених методів, де до оригінального сигналу додається білий, кольоровий або будь-який інший тип шуму. Ця техніка допомагає підвищити стійкість моделі до реальних умов, де аудіосигнали часто спотворені зовнішніми шумами.
2. **Зміна гучності**. Маніпуляція рівнем гучності аудіосигналу, яка дозволяє моделі працювати із сигналами різної інтенсивності.
3. **Зміна швидкості відтворення**. Ця техніка дозволяє змінювати тривалість аудіо без зміни його висоти, що допомагає моделі бути стійкою до сигналів з різною тривалістю.
4. **Зсув у часі**. Зміщення аудіосигналу вперед або назад у часовій шкалі, що імітує природні варіації запису.
5. **Частотне та часове маскування**. Метод, запропонований у SpecAugment, який полягає в маскуванні частотних (часових) діапазонів спектрограми, що підвищує стійкість моделей до пропусків або втрат інформації.
6. **Спотворення та реверберація**. Ці техніки додають ефекти, які змінюють акустичні властивості сигналу, симулюючи реальні умови запису.

Класифікація методів аугментації може бути розділена за типом маніпуляцій: маніпуляції у часовій, частотній або інтенсивній областях. Для кожної техніки можна використовувати різні набори параметрів, що дозволяє створити велику кількість варіацій одного аудіосигналу.

Аугментація відіграє вирішальну роль у задачах розпізнавання мовлення та класифікації звуків. Вона дозволяє моделі краще адаптуватися до різних реальних сценаріїв, де якість аудіосигналів може значно відрізнятися. Дослідження [1-4] показали, що використання технік, таких як додавання шуму або частотне маскування, покращує здатність моделей справлятися з шумами та спотвореннями.

Класифікація звуків, таких як розпізнавання музичних інструментів або звукових подій, є важливою задачею в області аудіоаналітики, яка може бути суттєво покращена за допомогою технік аудіо-аугментації. Аугментація аудіосигналів, яка включає в себе зміни швидкості, часовий зсув, додавання шуму та інші методи обробки, дозволяє моделі стати більш стійкою до варіацій у даних [6]. Це особливо важливо у реальних записах, де аудіосигнали можуть мати різні характеристики через вплив навколишнього середовища, апаратних засобів або навіть способу виконання звукових подій.

В контексті класифікації гітарних акордів, які можуть значно відрізнятися залежно від типу інструменту, манери гри, акустичних умов та інших факторів, застосування аудіо-аугментації є критично важливим для підвищення надійності та точності класифікаційної моделі [7]. Використання методів аугментації забезпечує кращу генералізацію моделі, дозволяючи їй успішно працювати зі складними і нестандартними аудіоданими, що виникають у реальних ситуаціях [8].

Це дослідження має на меті вивчення ефективності різних методів аудіо-аугментації у задачі класифікації гітарних акордів. Зокрема, було створено спеціальний датасет, що включає всі діатонічні акорди, записані за допомогою акустичної гітари (таблиця 1).

Таблиця 1

**Початковий датасет**

| Назва акорду | Кількість записів | Назва акорду | Кількість записів |
|---|---|---|---|
| C | 28 | Cm | 27 |
| C# | 26 | C#m | 26 |
| D | 26 | Dm | 28 |
| D# | 27 | D#m | 26 |
| E | 28 | Em | 28 |
| F | 28 | Fm | 28 |
| F# | 26 | F#m | 26 |
| G | 27 | Gm | 27 |
| G# | 27 | G#m | 27 |
| A | 26 | Am | 28 |
| A# | 28 | A#m | 26 |
| B | 26 | Bm | 26 |

Для навчання нейронної мережі була розроблена модель з архітектурою згорткової нейронної мережі (CNN), реалізована на основі бібліотеки PyTorch (лістинг 1).

Лістинг 1

```
import torch.nn as nn
import torch.nn.functional as F

class ChordClassifierCNN(nn.Module):
    def __init__(self):
        super(ChordClassifierCNN, self).__init__()
        self.conv1 = nn.Conv2d(1, 32, kernel_size=3, stride=1, padding=1)
        self.conv2 = nn.Conv2d(32, 64, kernel_size=3, stride=1, padding=1)
        self.conv3 = nn.Conv2d(64, 128, kernel_size=3, stride=1, padding=1)
        self.pool = nn.MaxPool2d(kernel_size=2, stride=2, padding=0)
        self.fc1 = nn.Linear(128 * 16 * 16, 256)
        self.fc2 = nn.Linear(256, 64)
        self.fc3 = nn.Linear(64, 24)

    def forward(self, x):
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        x = self.pool(F.relu(self.conv3(x)))
        x = x.view(-1, 128 * 16 * 16)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

Для тестування було створено окремий датасет, що містить записи електрогітари, де кожен діатонічний акорд представлений десятьма екземплярами. Результати навчання, а також перевірки моделі на тестовому датасеті, наведені у таблиці 2, демонструють, що збільшення кількості циклів навчання призводить до підвищення точності на навчальних даних, але не сприяє покращенню точності на тестовому датасеті.

Таблиця 2

**Результати моделі на початковому датасеті**

| Кількість циклів навчання | Точність моделі на даних для навчання | Точність моделі на тестових даних |
|---|---|---|
| 10 | 60% | 20% |
| 100 | 87% | 32% |
| 1000 | 92% | 27% |

Це явище вказує на можливе перенавчання моделі, що робить її менш ефективною при роботі з новими, невідомими даними. Для вирішення цієї проблеми передбачається застосування методів аудіо-аугментації:

- додавання білого шуму;
- зміна швидкості;
- часове маскування;
- застосування спотворення та реверберації

Приклади спектрограм та осцилограм аудіосигналу акорду до мажор після застосування методів аудіо-аугментації наведено на рис. 1 – 5. Ці приклади ілюструють, як різні методи аугментації змінюють вихідний сигнал, і як такі зміни можуть вплинути на процес класифікації.

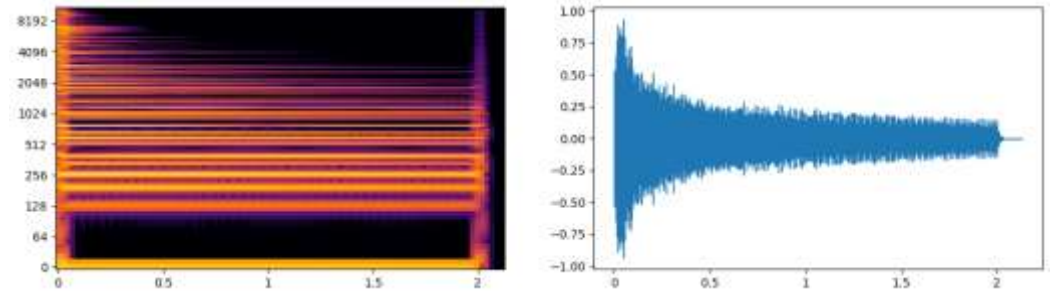

**Рис. 1. Спектограма та осцилограма акорду до мажор без застосування аудіо-аугметації**

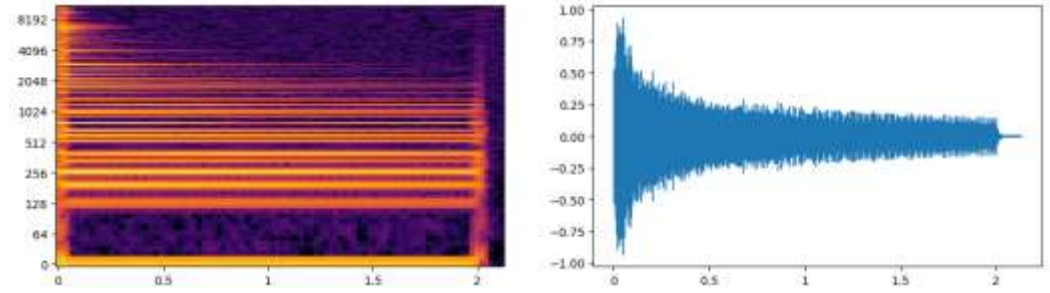

**Рис. 2. Спектограма та осцилограма акорду до мажор із доданим білим шумом**

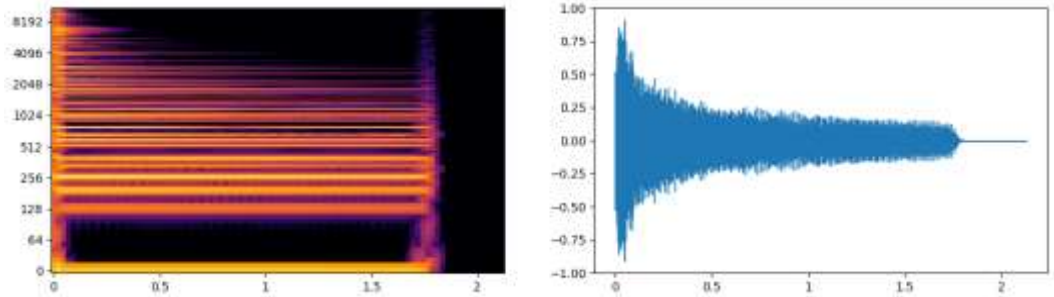

**Рис. 3. Спектограма та осцилограма акорду до мажор із застосування зміни швидкості**

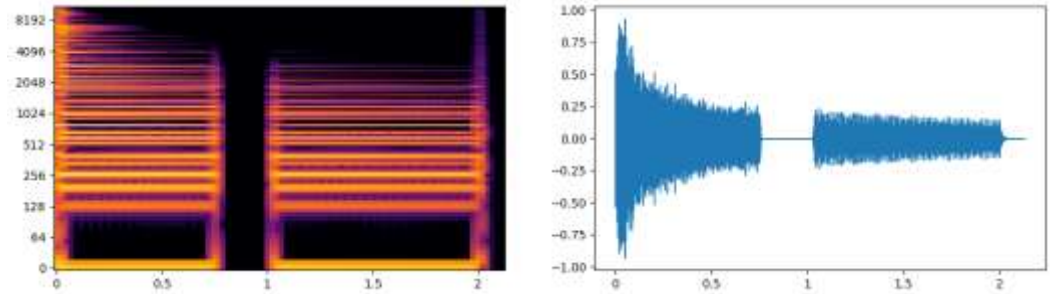

**Рис. 4. Спектограма та осцилограма акорду до мажор із застосування часового маскування**

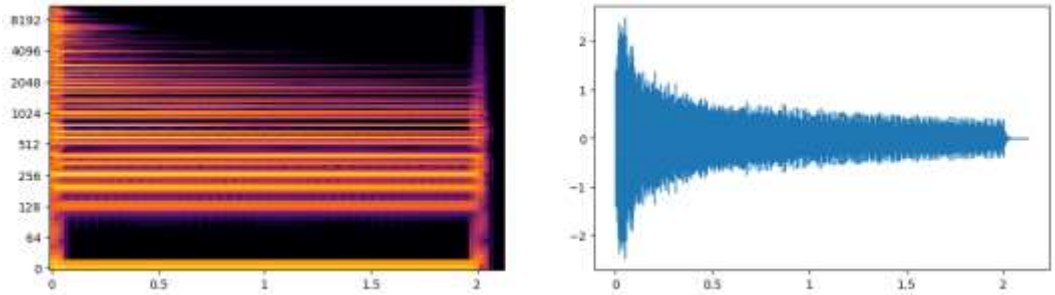

**Рис. 5. Спектограма та осцилограма акорду до мажор із застосуванням спотворення та реверберації аудіосигналу**

Після проведення перетворення аудіосигналів з використанням технік аудіо-аугментації було значно розширено тренувальний датасет, що дозволило підвищити різноманітність даних та зробити модель більш стійкою до різних варіацій аудіосигналів. Наступний етап дослідження включав повторне навчання моделі на розширеному датасеті та її перевірку на тестових даних, результати наведено у таблиці 3.

Таблиця 3

**Результати моделі на розширеному датасеті**

| Кількість циклів навчання | Точність моделі на даних для навчання | Точність моделі на тестових даних |
|---|---|---|
| 10 | 84% | 72% |
| 100 | 96% | 81% |
| 1000 | 97% | 88% |

На основі проведеного експерименту, можна зробити висновок, що перетворення аудіосигналів за допомогою технік аудіо-аугментації значно покращило точність моделі з архітектурою CNN для розпізнавання гітарних акордів.

**Висновки**

Аудіо-аугментація є незамінним інструментом у сучасних системах машинного навчання, особливо коли йдеться про оброблення аудіоданих. Вона не лише дозволяє створювати більш різноманітні й узагальнені набори даних, але й допомагає нейронним мережам краще адаптуватися до змінних реальних умов. Завдяки таким технікам, як додавання шуму, зміна тону, часове та частотне маскування, а також синтетичне збільшення наборів даних, моделі стають стійкішими до різних видів спотворень і шумів.

На основі проведеного експерименту можна зробити кілька ключових висновків. По-перше, перетворення аудіосигналів з використанням технік аудіо-аугментації значно покращило точність моделі з архітектурою CNN для розпізнавання гітарних акордів. Це підтверджує гіпотезу про те, що аугментація здатна підвищити ефективність моделей машинного навчання, особливо у випадках, коли доступ до великих обсягів оригінальних даних обмежений.

По-друге, вибір конкретних технік аугментації залежить від характеру завдання та типу аудіоданих. Наприклад, для задач розпізнавання мовлення найбільш ефективними є техніки додавання шуму та частотного маскування. Ці методи підвищують стійкість моделі до зовнішніх спотворень, таких як фоновий шум або незначні зміни тембру голосу. Для музичних застосунків або класифікації звукових подій корисними є зміна швидкості та тривалості аудіосигналу, оскільки ці техніки імітують природні варіації звуків у реальних умовах. Наприклад, у випадках роботи з музичними записами ці варіації можуть відображати різні темпи виконання, зміни гучності або резонансу.

По-третє, для мовних моделей, які використовуються в задачах синтезу мовлення або аналізу розмовних даних, техніки реверберації та зсуву в часі дозволяють моделі адаптуватися до різних акустичних середовищ. Це робить моделі більш гнучкими та здатними до роботи в умовах, що можуть значно відрізнятися від тих, в яких проводилося навчання.

Ще однією важливою областю застосування аудіо-аугментації є обробка послідовних аудіоданих. Тут ефективними є техніки, що змінюють часові характеристики сигналів, такі як зміщення або розтягування у часі. Ці методи особливо корисні для задач синтезу мовлення або генерації звуків, де послідовність і тривалість сигналів має критичне значення.

Перспективи розвитку аудіо-аугментації тісно пов'язані з подальшим вдосконаленням архітектур нейронних мереж та підвищенням обчислювальних потужностей. Зокрема, у майбутньому можна очікувати вдосконалення технік аугментації за рахунок використання генеративних моделей, таких як Generative Adversarial Networks (GANs), для створення синтетичних аудіоданих. Це може значно розширити можливості аугментації та дозволить створювати високоякісні синтетичні дані, що точно відображають властивості оригінальних сигналів.

Крім того, інтеграція аудіо-аугментації в самонавчальні системи, де моделі самостійно генерують варіації даних під час тренування, може відкрити нові горизонти у створенні моделей, які здатні до адаптації в реальних умовах. Розроблення нових методів аугментації, орієнтованих на індивідуальні особливості аудіосигналів, дозволять створювати спеціалізовані рішення для роботи з конкретними типами аудіо, такими як музика, мовлення або навколишні звуки.

Загалом, аудіо-аугментація вже зараз є важливим інструментом у розробленні штучного інтелекту і її подальший розвиток сприятиме підвищенню точності та надійності аудіомоделей, розширюючи їх можливості для застосування у різноманітних сферах.

**Література**

1. Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le (2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." *Proceedings of the Annual Conference of the International Speech Communication Association*.

2. Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5901–5905. IEEE, 2019.

3. Tom Ko, Vijayaditya Peddinti, Daniel Povey, S. Khudanpur (2015). "Audio augmentation for speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

4. Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pages 309–314. IEEE, 2013.

5. Justin Salamon, Juan Pablo Bello (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24, 279–283.

6. Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling". Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 100–104.

7. Rafael L. Aguiar., Yandre M.G. Costa, Carlos N. Silla (2018). Exploring data augmentation to improve music genre classification with convnets. Proceedings of the International Joint Conference on Neural Networks, 1–8.

8. Navdeep Jaitly, Geoffrey E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition". International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing, 2013.

**References**

1. Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le (2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." *Proceedings of the Annual Conference of the International Speech Communication Association*.

2. Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5901–5905. IEEE, 2019.

3. Tom Ko, Vijayaditya Peddinti, Daniel Povey, S. Khudanpur (2015). "Audio augmentation for speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

4. Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pages 309–314. IEEE, 2013.

5. Justin Salamon, Juan Pablo Bello (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24, 279–283.

6. Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling". Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 100–104.

7. Rafael L. Aguiar., Yandre M.G. Costa, Carlos N. Silla (2018). Exploring data augmentation to improve music genre classification with convnets. Proceedings of the International Joint Conference on Neural Networks, 1–8.

8. Navdeep Jaitly, Geoffrey E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition". International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing, 2013.