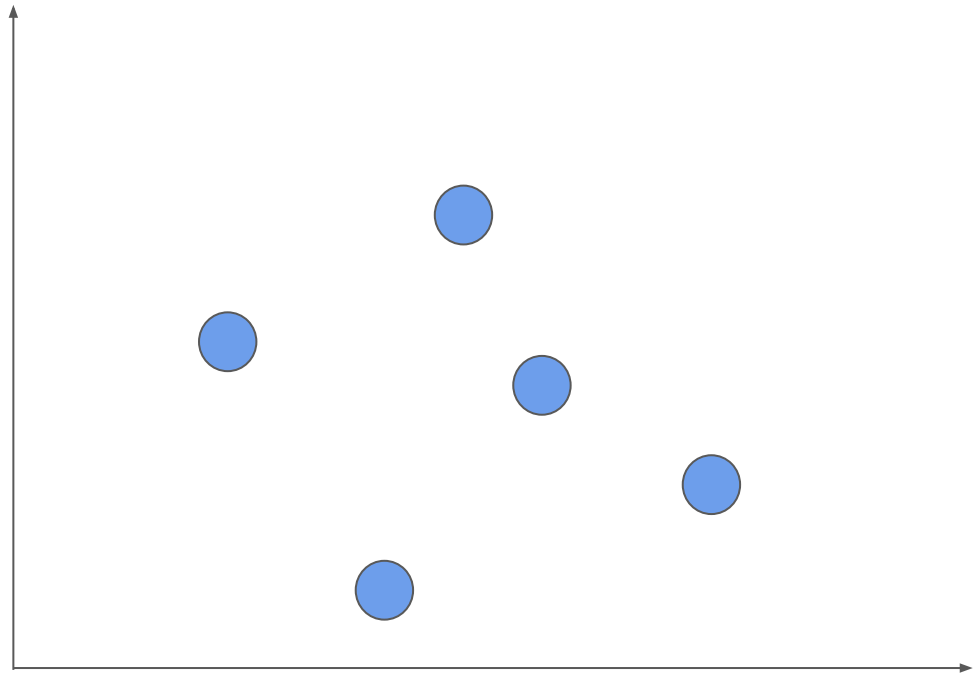
Audio Data Augmentation Is All You Need

Valerio Velardo

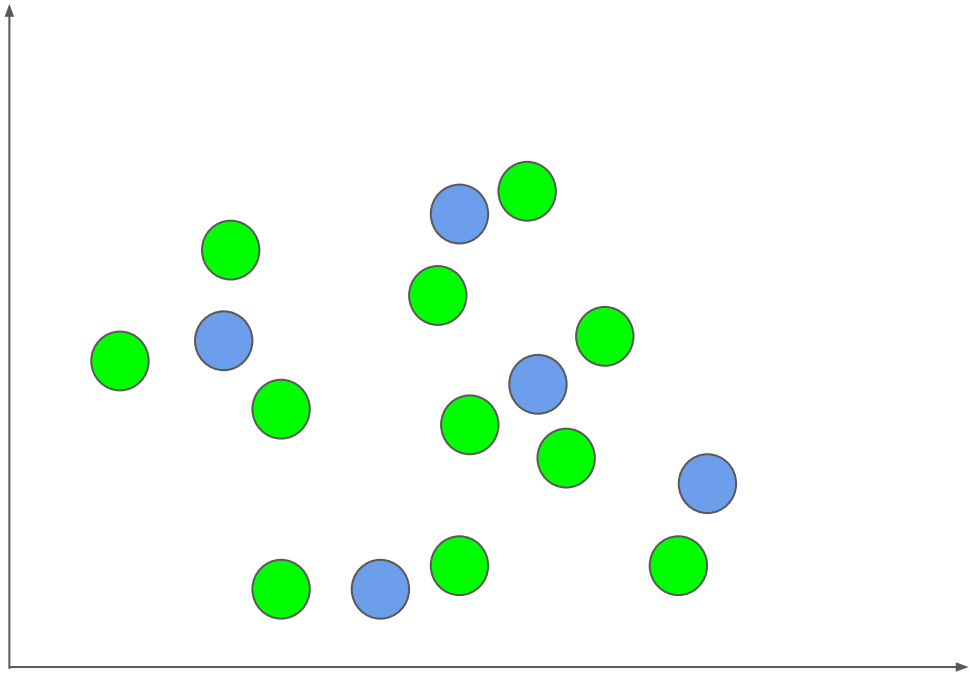
What’s data augmentation?

Technique used to increase the number of samples an ML model sees during training

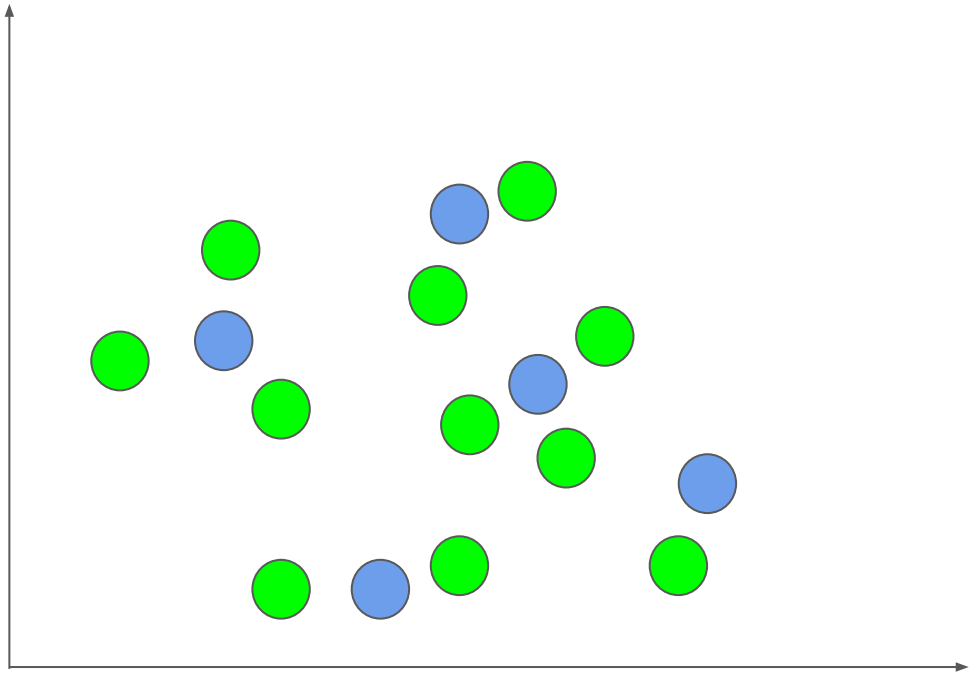
What’s the goal of data augmentation?



What’s the goal of data augmentation?



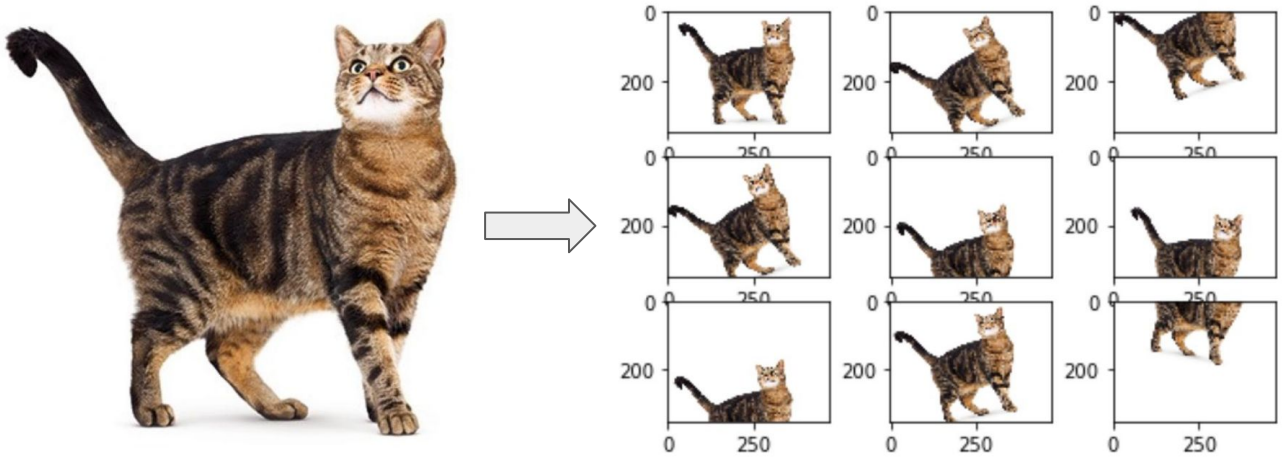
What’s the goal of data augmentation?



- Cover the problem space as much as possible

How does data augmentation work?

Image data augmentation



Data augmentation

Computer vision

Data augmentation

Audio

Why should I care about audio data augmentation?

- Address data scarcity
- Increase models’ robustness
- Improve models’ accuracy
- Reduce overfitting
- Save resources to collect and label data

Augmented data

is NOT as good as

Additional data

Use cases for audio data augmentation

Use cases for audio data augmentation

EXPLORING DATA AUGMENTATION FOR IMPROVED SINGING VOICE DETECTION WITH NEURAL NETWORKS

Jan Schlüter and Thomas Grill
Austrian Research Institute for Artificial Intelligence, Vienna
jan.schlueter@ofai.at thomas.grill@ofai.at

ABSTRACT

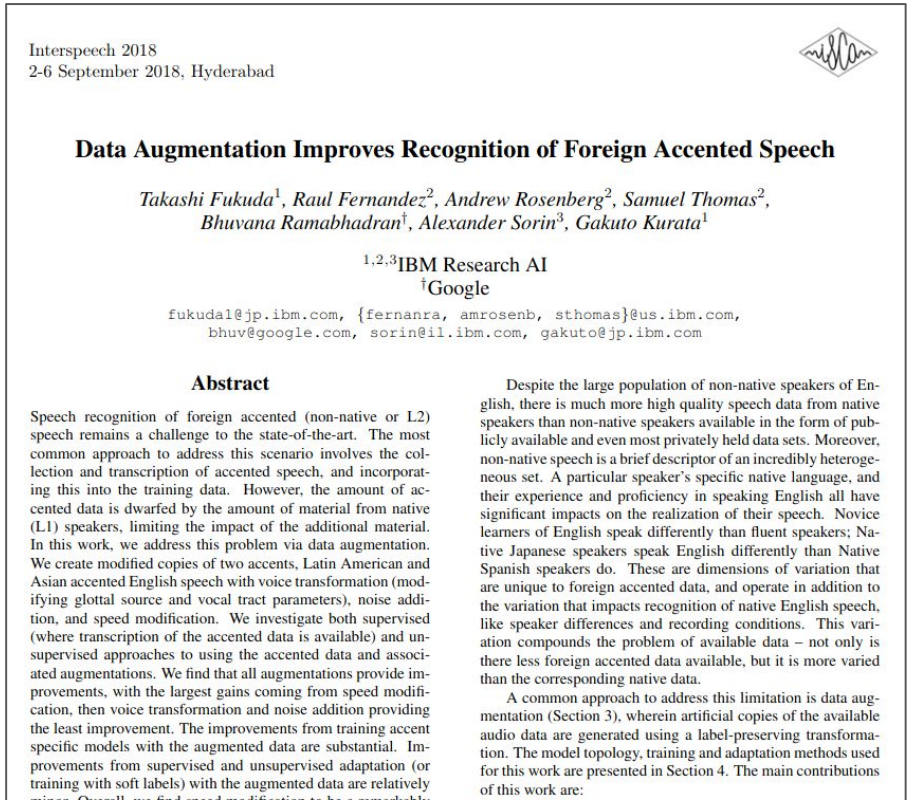
In computer vision, state-of-the-art object recognition systems rely on label-preserving image transformations such as scaling and rotation to augment the training datasets. The additional training examples help the system to learn invariances that are difficult to build into the model, and improve generalization to unseen data. To the best of our knowledge, this approach has not been systematically explored for music signals. Using the problem of singing voice detection with neural networks as an example, we apply a range of label-preserving audio transformations to assess their utility for music data augmentation. In line with recent research in speech recognition, we find pitch shifting to be the most helpful augmentation method. Combined with time stretching and random frequency filtering, we achieve a reduction in classification error between 10 and 30%, reaching the state of the art on two public data-

music information retrieval (MIR) – has picked it up as well [9], we could only find anecdotal references to it in the MIR literature [8, 18], but no systematic treatment.

In this work, we devise a range of label-preserving audio transformations and compare their utility for music signals on a benchmark problem. Specifically, we chose the sequence labelling task of singing voice detection: It is well-covered, but best reported accuracies on public datasets are around 90%, suggesting some leeway. Furthermore, it does not require profound musical knowledge to solve, making it an ideal candidate for training a classifier on low-level inputs. This allows observing the effect of data augmentation unaffected by engineered features, and unhindered by doubtable ground truth. For the classifier, we chose CNNs, proven powerful enough to pick up invariances taught by data augmentation in other fields.

The following section will review related work on data

Use cases for audio data augmentation



Interspeech 2018
2-6 September 2018, Hyderabad

Data Augmentation Improves Recognition of Foreign Accented Speech

Takashi Fukuda[1], *Raul Fernandez*[2], *Andrew Rosenberg*[2], *Samuel Thomas*[2],
Bhuvana Ramabhadran[†], *Alexander Sorin*[3], *Gakuto Kurata*[1]

[1,2,3]IBM Research AI
[†]Google

fukudal@jp.ibm.com, {fernanra, amrosenb, sthomas}@us.ibm.com,
bhuv@google.com, sorin@il.ibm.com, gakuto@jp.ibm.com

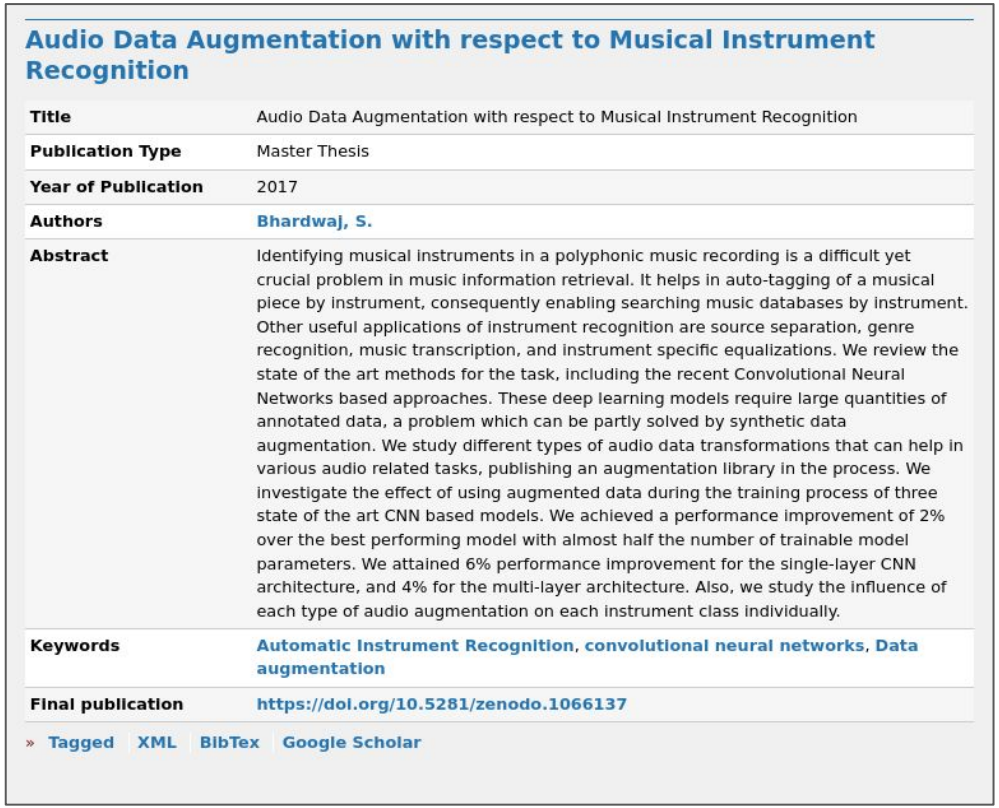
Abstract

Speech recognition of foreign accented (non-native or L2) speech remains a challenge to the state-of-the-art. The most common approach to address this scenario involves the collection and transcription of accented speech, and incorporating this into the training data. However, the amount of accented data is dwarfed by the amount of material from native (L1) speakers, limiting the impact of the additional material. In this work, we address this problem via data augmentation. We create modified copies of two accents, Latin American and Asian accented English speech with voice transformation (modifying glottal source and vocal tract parameters), noise addition, and speed modification. We investigate both supervised (where transcription of the accented data is available) and unsupervised approaches to using the accented data and associated augmentations. We find that all augmentations provide improvements, with the largest gains coming from speed modification, then voice transformation and noise addition providing the least improvement. The improvements from training accent specific models with the augmented data are substantial. Improvements from supervised and unsupervised adaptation (or training with soft labels) with the augmented data are relatively

Despite the large population of non-native speakers of English, there is much more high quality speech data from native speakers than non-native speakers available in the form of publicly available and even most privately held data sets. Moreover, non-native speech is a brief descriptor of an incredibly heterogeneous set. A particular speaker's specific native language, and their experience and proficiency in speaking English all have significant impacts on the realization of their speech. Novice learners of English speak differently than fluent speakers; Native Japanese speakers speak English differently than Native Spanish speakers do. These are dimensions of variation that are unique to foreign accented data, and operate in addition to the variation that impacts recognition of native English speech, like speaker differences and recording conditions. This variation compounds the problem of available data – not only is there less foreign accented data available, but it is more varied than the corresponding native data.

A common approach to address this limitation is data augmentation (Section 3), wherein artificial copies of the available audio data are generated using a label-preserving transformation. The model topology, training and adaptation methods used for this work are presented in Section 4. The main contributions of this work are:

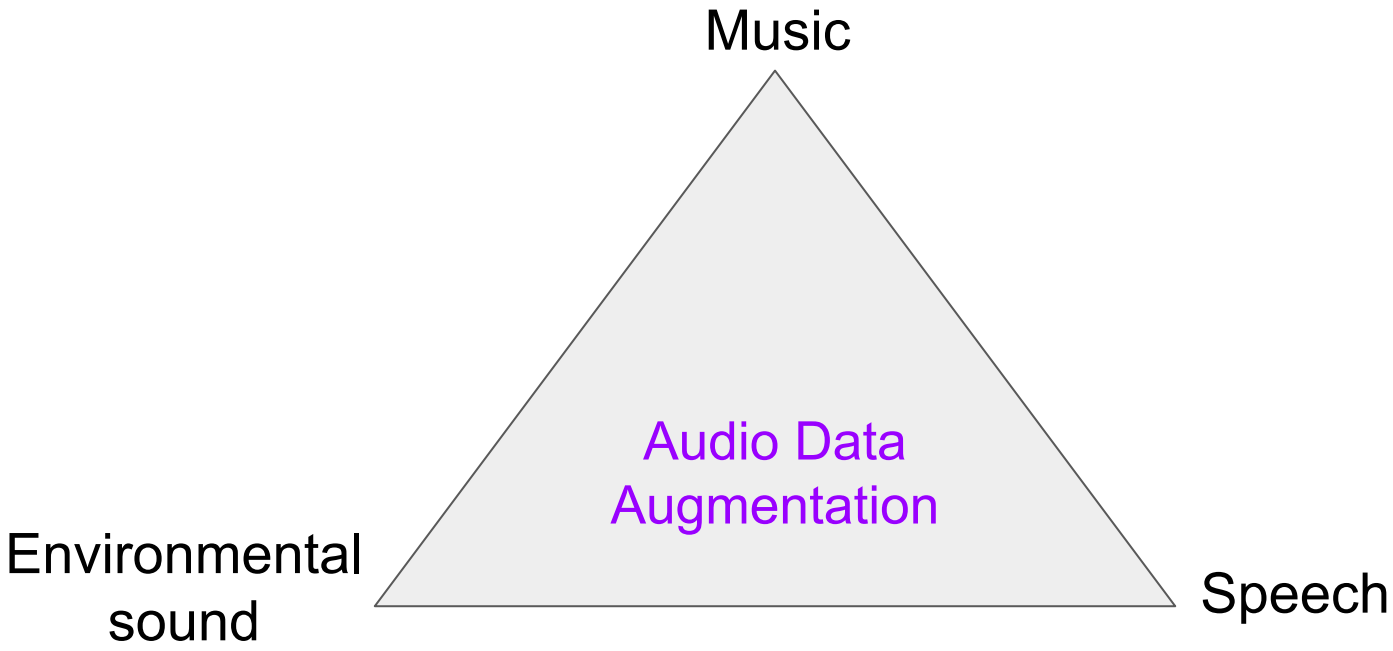
Use cases for audio data augmentation



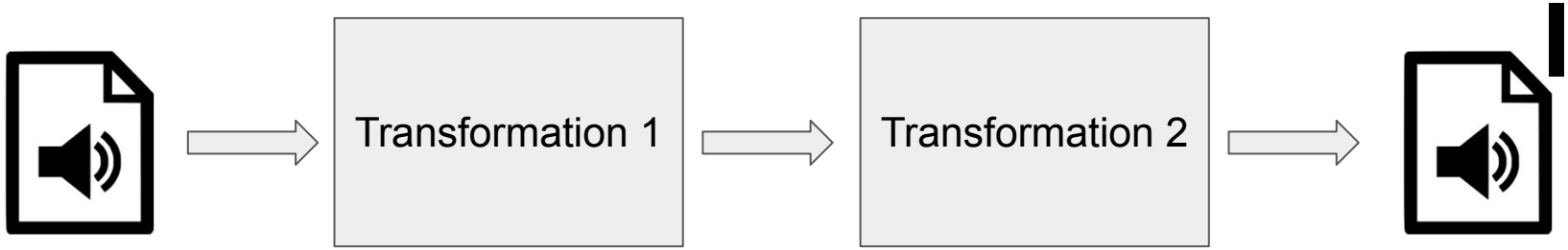
Audio Data Augmentation with respect to Musical Instrument Recognition

Title	Audio Data Augmentation with respect to Musical Instrument Recognition
Publication Type	Master Thesis
Year of Publication	2017
Authors	**Bhardwaj, S.**
Abstract	Identifying musical instruments in a polyphonic music recording is a difficult yet crucial problem in music information retrieval. It helps in auto-tagging of a musical piece by instrument, consequently enabling searching music databases by instrument. Other useful applications of instrument recognition are source separation, genre recognition, music transcription, and instrument specific equalizations. We review the state of the art methods for the task, including the recent Convolutional Neural Networks based approaches. These deep learning models require large quantities of annotated data, a problem which can be partly solved by synthetic data augmentation. We study different types of audio data transformations that can help in various audio related tasks, publishing an augmentation library in the process. We investigate the effect of using augmented data during the training process of three state of the art CNN based models. We achieved a performance improvement of 2% over the best performing model with almost half the number of trainable model parameters. We attained 6% performance improvement for the single-layer CNN architecture, and 4% for the multi-layer architecture. Also, we study the influence of each type of audio augmentation on each instrument class individually.
Keywords	**Automatic Instrument Recognition**, **convolutional neural networks**, **Data augmentation**
Final publication	**https://doi.org/10.5281/zenodo.1066137**

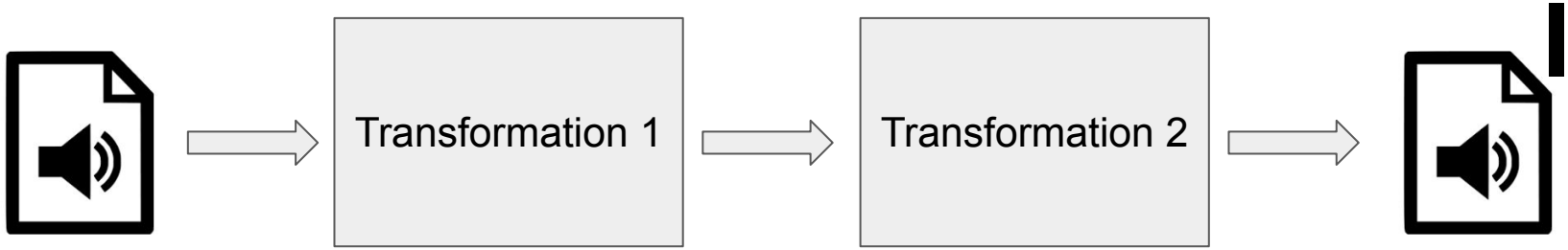
» **Tagged** **XML** **BibTex** **Google Scholar**


Music
Audio Data
Augmentation
Environmental
sound
Speech

Augmentation chains



Augmentation chains



- Randomly select transformations to apply
- Randomly select transformation parameters


WHEN YOU AUGMENT
THE VALIDATION SET
imgflip.com

What data should I augment?

- Augment the train set only
- Augmenting validation / test set -> data leakage

What data should I augment?

- Augment the train set only
- Augmenting validation / test set -> data leakage

1. Split your dataset in train / validation / test sets first
2. Apply augmentation to train set only

When should I augment data?

Offline augmentation

- Precompute transformations before training
- Save computation in the long run
- Augmentation code separate from model code
- Done on CPU -> slower
- More storage required

When should I augment data?

Offline augmentation

- Precompute transformations before training
- Save computation in the long run
- Augmentation code separate from model code
- Done on CPU -> slower
- More storage required

Online augmentation

- Apply transformations at training time
- Carried out with DL libraries (e.g., *torchaudio.transforms*)
- Done on GPU -> faster
- Model deployment is easier
- Augmentation module coupled with model code
- Computationally expensive in the long run

The golden rule of augmentation

The augemented audio samples MUST be “credible”

Join the community!

thesoundofai.slack.com

What’s next?

- Waveform vs spectrogram augmentation
- Audio transformations [theory]