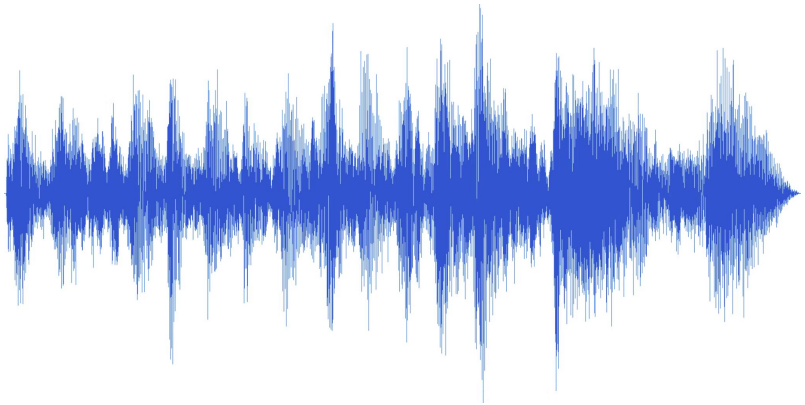
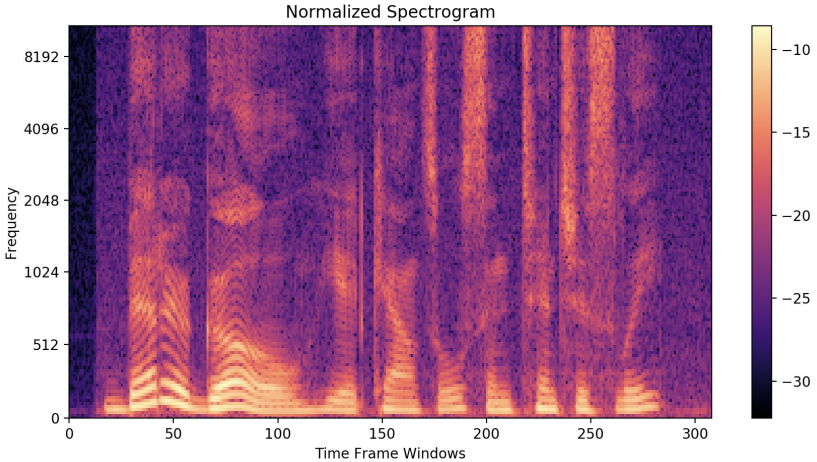
Audio Data Augmentation Techniques

Valerio Velardo

Raw audio vs spectrogram augmentation





Raw audio augmentation transformations

- Time shifting

Raw audio augmentation transformations

- Time shifting
- Time stretching

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition
- Impulse response addition

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition
- Impulse response addition
- Low/high/pass-band filters

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition
- Impulse response addition
- Low/high/pass-band filters
- Polarity inversion

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition
- Impulse response addition
- Low/high/pass-band filters
- Polarity inversion
- Random gain

Raw audio augmentation transformations

- Time shifting
- Time stretching
- Pitch scaling
- Noise addition
- Impulse response addition
- Low/high/pass-band filters
- Polarity inversion
- Random gain
- …

Spectrogram augmentation transformations

SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le*

Google Brain

{danielspark, williamchan, ngyuzh, chungchengc, barretzoph, cubuk, qvl}@google.com

Abstract

We present SpecAugment, a simple data augmentation method for speech recognition. SpecAugment is applied directly to the feature inputs of a neural network (i.e., filter bank coefficients). The augmentation policy consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps. We apply SpecAugment on Listen, Attend and Spell networks for end-to-end speech recognition tasks. We achieve state-of-the-art performance on the LibriSpeech 960h and Swichboard 300h tasks, outperforming all prior work. On LibriSpeech, we achieve 6.8% WER on test-other without the use of a language model, and 5.8% WER with shallow fusion with a language model. This compares to the previous state-of-the-art hybrid system of 7.5% WER. For Switchboard, we achieve 7.2%/14.6% on the Switchboard/CallHome portion of the Hub5'00 test set without the use of a language model, and 6.8%/14.1% with shallow fusion, which compares to the previous state-of-the-art hybrid system at 8.3%/17.3% WER.

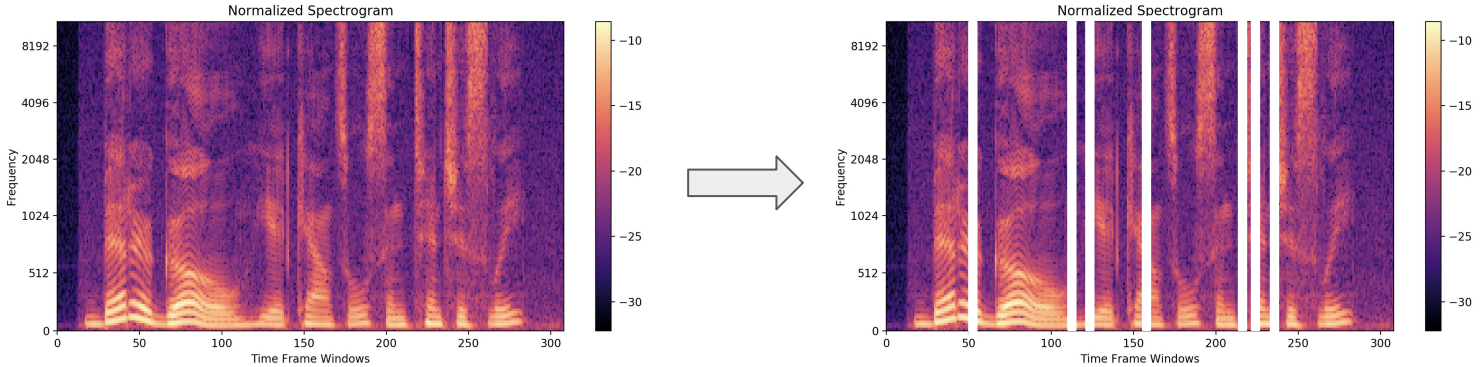
Index Terms: end-to-end speech recognition, data augmentation

require any additional data. We are thus able to apply SpecAugment online during training. SpecAugment consists of three kinds of deformations of the log mel spectrogram. The first is time warping, a deformation of the time-series in the time direction. The other two augmentations, inspired by "Cutout", proposed in computer vision [19], are time and frequency masking, where we mask a block of consecutive time steps or mel frequency channels.

This approach while rudimentary, is remarkably effective and allows us to train end-to-end ASR networks, called Listen Attend and Spell (LAS) [6], to surpass more complicated hybrid systems, and achieve state-of-the-art results even without the use of Language Models (LMs). On LibriSpeech [20], we achieve 2.8% Word Error Rate (WER) on the test-clean set and 6.8% WER on the test-other set, without the use of an LM. Upon shallow fusion [21] with an LM trained on the LibriSpeech LM corpus, we are able to better our performance (2.5% WER on test-clean and 5.8% WER on test-other), improving the current state of the art on test-other by 22% relatively. On Switchboard 300h (LDC97S62) [22], we

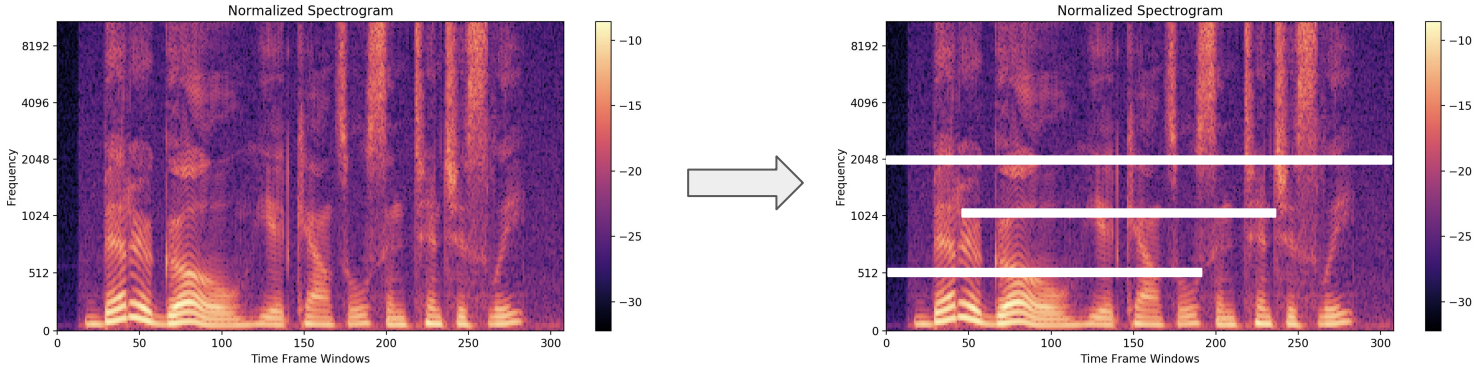
Spectrogram augmentation transformations

- Time masking



Spectrogram augmentation transformations

- Time masking
- Frequency masking



Spectrogram augmentation transformations

- Time masking
- Frequency masking
- Time stretching

Spectrogram augmentation transformations

- Time masking
- Frequency masking
- Time stretching
- Pitch scaling

Spectrogram augmentation transformations

- Time masking
- Frequency masking
- Time stretching
- Pitch scaling
- …

Join the community!

thesoundofai.slack.com

What’s next?

- Implement a few audio augmentation transforms in Python