

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

Content

1. Introduction
2. Augmentation policy
3. Model
4. Experiments
5. Discussion

1. Introduction

Problem: Deep learning models employed in Automatic Speech Recognition (ASR) often tend to overfit and require substantial data for training.

Past Solutions: Data augmentation has been suggested as a method to generate additional training data for ASR systems.

- Artificial data augmentation for low-resource speech recognition tasks.
- Synthesis of noisy audio by superimposing clean audio with a noisy signal.
- Application of speed perturbation to raw audio for Large Vocabulary Continuous Speech Recognition (LVCSR) tasks.
- Implementation of feature drop-outs in training multi-stream ASR systems.
- ...

1. Introduction

This study introduces SpecAugment, a data augmentation technique that operates on the log mel spectrogram of the input audio, rather than the raw audio itself.

SpecAugment applies three types of deformations to the log mel spectrogram:

- Time Warping
- Time Masking
- Frequency Masking

1. Introduction

Advantages of SpecAugment:

1. Simple and computationally inexpensive to apply
 - a. Directly acts on the log mel spectrogram as if it were an image
 - b. Does not require additional data
2. Can be applied online during training.
3. Remarkable effective
 - a. Suprasses more complicated hybrid systems
 - b. Achieves state-of-the-art results even without the use of Language Models (LMs)

2. Augmentation Policy

- Time warping
 - Random point selected
 - Warped to left or right by a distance w
- Frequency masking
 - $[f_0, f_0+f)$ frequency channels are masked
- Time masking
 - $[t_0, t_0+t)$ time steps are masked

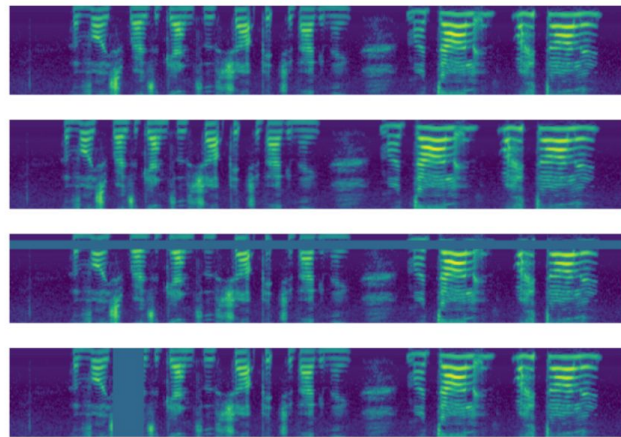


Figure 1: *Augmentations applied to the base input, given at the top. From top to bottom, the figures depict the log mel spectrogram of the base input with no augmentation, time warp, frequency masking and time masking applied.*

2. Augmentation Policy

Table 1: Augmentation parameters for policies. m_F and m_T denote the number of frequency and time masks applied.

Policy	W	F	m_F	T	p	m_T
None	0	0	-	0	-	-
LB	80	27	1	100	1.0	1
LD	80	27	2	100	1.0	2
SM	40	15	2	70	0.2	2
SS	40	27	2	70	0.2	2

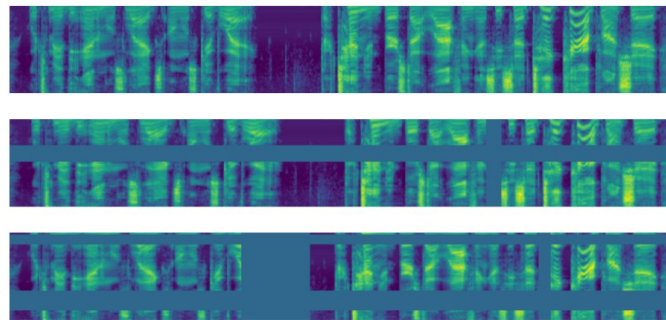


Figure 2: Augmentation policies applied to the base input. From top to bottom, the figures depict the log mel spectrogram of the base input with policies None, LB and LD applied.

3. Model

Listen, Attend, and Spell (LAS) Network

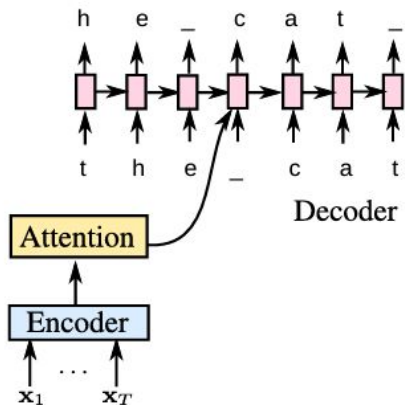


Figure 1: *LAS model*.

Figure 1 visualizes LAS with these two components. We provide more details of these components in the following sections.

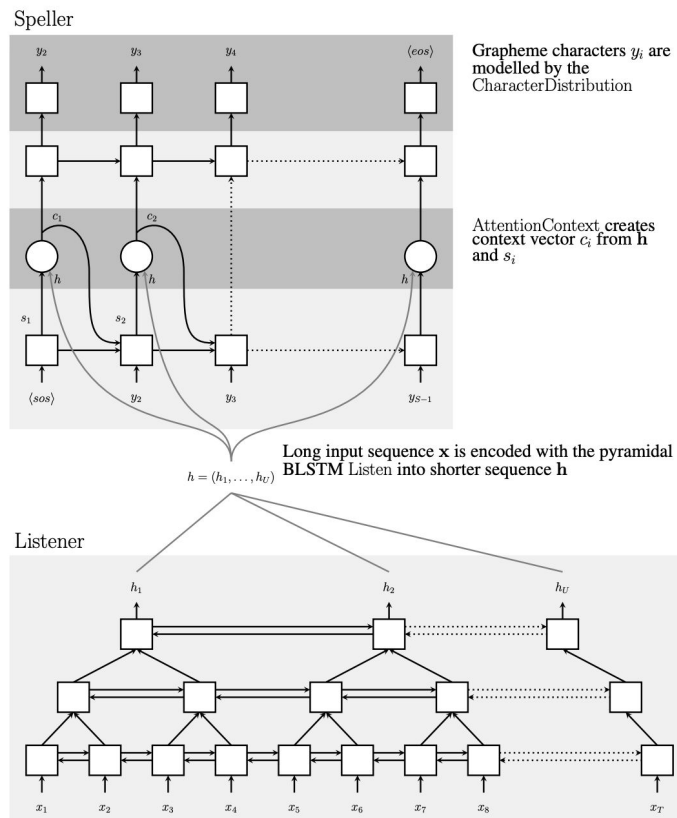


Figure 1: Listen, Attend and Spell (LAS) model: the listener is a pyramidal BLSTM encoding our input sequence x into high level features h , the speller is an attention-based decoder generating the y characters from h .

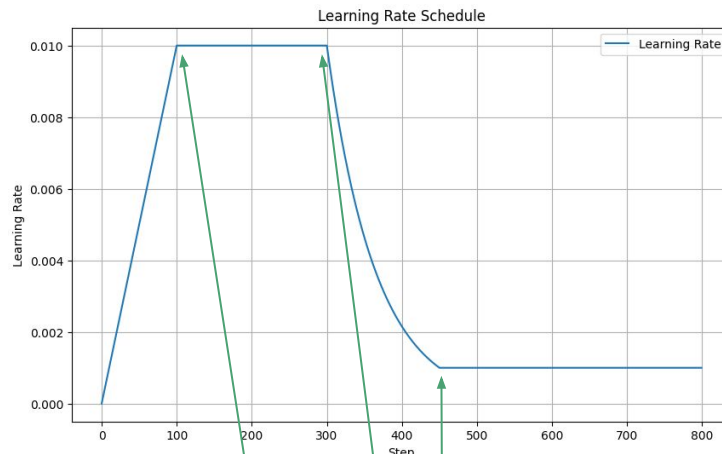
L(ong): $(s_r, s_{\text{noise}}, s_i, s_f) = (1\text{k}, 20\text{k}, 140\text{k}, 320\text{k})$

3. Model

Listen, Attend, and Spell (LAS) Network

Training:

- Learning rate scheduler
- Variational weight noise
- Uniform label smoothing (0.1 uncertainty)



Steps for variational weight noise

1. B(asic): $(s_r, s_{\text{noise}}, s_i, s_f) = (0.5\text{k}, 10\text{k}, 20\text{k}, 80\text{k})$
2. D(ouble): $(s_r, s_{\text{noise}}, s_i, s_f) = (1\text{k}, 20\text{k}, 40\text{k}, 160\text{k})$
3. L(ong): $(s_r, s_{\text{noise}}, s_i, s_f) = (1\text{k}, 20\text{k}, 140\text{k}, 320\text{k})$

3. Model

Listen, Attend, and Spell (LAS) Network

Training:

- Learning rate scheduler
- Variational weight noise
- Uniform label smoothing (0.1 uncertainty)

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}} (\log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_{LM}(\mathbf{y}))$$

Inference:

- Shallow fusion with LM(2 layer 1024-dim RNN) (**Pre-trained**)

4. Experiments

- Dataset
 - LibriSpeech
 - Switchboard
 - Switchboard includes 2,400 phone conversations across 543 U.S. speakers (302 male, 241 female), featuring 70 topics, with each speaker participating only once per topic and never repeating a conversational partner.
- Exploring:
 - **Performance compare to other system**
 - Different training schedules (strategies)
 - Different augmentation strategies
 - Performance of different model size
 - Effectiveness of Label smoothing

Performance Comparison

Table 3: LibriSpeech 960h WERs (%).

Method	No LM		With LM	
	clean	other	clean	other
HMM				
Panayotov et al., (2015) [20]			5.51	13.97
Povey et al., (2016) [30]			4.28	
Han et al., (2017) [31]			3.51	8.58
Yang et al. (2018) [32]			2.97	7.50
CTC/ASG				
Collobert et al., (2016) [33]	7.2			
Liptchinsky et al., (2017) [34]	6.7	20.8	4.8	14.5
Zhou et al., (2018) [35]			5.42	14.70
Zeghidour et al., (2018) [36]			3.44	11.24
Li et al., (2019) [37]	3.86	11.95	2.95	8.79
LAS				
Zeyer et al., (2018) [24]	4.87	15.39	3.82	12.76
Zeyer et al., (2018) [38]	4.70	15.20		
Irie et al., (2019) [25]	4.7	13.4	3.6	10.3
Sabour et al., (2019) [39]	4.5	13.3		
Our Work				
LAS	4.1	12.5	3.2	9.8
LAS + SpecAugment	2.8	6.8	2.5	5.8

Table 5: Switchboard 300h WERs (%).

Method	No LM		With LM	
	SWBD	CH	SWBD	CH
HMM				
Vesely et al., (2013) [41]			12.9	24.5
Povey et al., (2016) [30]			9.6	19.3
Hadian et al., (2018) [42]			9.3	18.9
Zeyer et al., (2018) [24]			8.3	17.3
CTC				
Zweig et al., (2017) [43]	24.7	37.1	14.0	25.3
Audhkhasi et al., (2018) [44]	20.8	30.4		
Audhkhasi et al., (2018) [45]	14.6	23.6		
LAS				
Lu et al., (2016) [46]	26.8	48.2	25.8	46.0
Toshniwal et al., (2017) [47]	23.1	40.8		
Zeyer et al., (2018) [24]	13.1	26.1	11.8	25.7
Weng et al., (2018) [48]	12.2	23.3		
Zeyer et al., (2018) [38]	11.9	23.7	11.0	23.1
Our Work				
LAS	11.2	21.6	10.9	19.4
LAS + SpecAugment (SM)	7.2	14.6	6.8	14.1
LAS + SpecAugment (SS)	7.3	14.4	7.1	14.0

4. Experiments

- Dataset
 - LibriSpeech
 - Switchboard
 - Switchboard includes 2,400 phone conversations across 543 U.S. speakers (302 male, 241 female), featuring 70 topics, with each speaker participating only once per topic and never repeating a conversational partner.
- Exploring:
 - Performance compare to other system
 - **Performance of different model size**
 - **Different training schedules (strategies)**
 - **Different augmentation strategies**
 - Effectiveness of Label smoothing

LibriSpeech

1. Augmentation always help
2. Larger model can be trained
3. Longer the Schedule, the better
4. The harsher augmentation, the better

Table 2: *LibriSpeech test WER (%) evaluated for varying networks, schedules and policies. First row from [25].*

Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5

LibriSpeech

1. **Augmentation always help**
2. Larger model can be trained
3. Longer the Schedule, the better
4. The harsher augmentation, the better

Table 2: *LibriSpeech test WER (%) evaluated for varying networks, schedules and policies. First row from [25].*

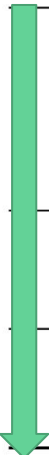
Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5

LibriSpeech

1. Augmentation always help
- 2. Larger model can be trained**
3. Longer the Schedule, the better
4. The harsher augmentation, the better

Table 2: *LibriSpeech test WER (%) evaluated for varying networks, schedules and policies. First row from [25].*

Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5



LibriSpeech

1. Augmentation always help
- 2. Larger model can be trained**
3. Longer the Schedule, the better
4. The harsher augmentation, the better

Table 2: *LibriSpeech test WER (%) evaluated for varying networks, schedules and policies. First row from [25].*

Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5

LibriSpeech

1. Augmentation always help
2. Larger model can be trained
- 3. Longer the Schedule, the better**
4. The harsher augmentation, the better

Table 2: *LibriSpeech test WER (%) evaluated for varying networks, schedules and policies. First row from [25].*

Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5

LibriSpeech

1. Augmentation always help
2. Larger model can be trained
3. Longer the Schedule, the better
- 4. The harsher augmentation, the better**

Table 2: *LibriSpeech* test WER (%) evaluated for varying networks, schedules and policies. First row from [25].

Network	Sch	Pol	No LM		With LM	
			clean	other	clean	other
LAS-4-1024 [25]	B	-	4.7	13.4	3.6	10.3
LAS-4-1024	B	LB	3.7	10.0	3.4	8.3
	B	LD	3.6	9.2	2.8	7.5
	D	-	4.4	13.3	3.5	10.4
	D	LB	3.4	9.2	2.7	7.3
	D	LD	3.4	8.3	2.8	6.8
LAS-6-1024	D	-	4.5	13.1	3.6	10.3
	D	LB	3.4	8.6	2.6	6.7
	D	LD	3.2	8.0	2.6	6.5
LAS-6-1280	D	-	4.3	12.9	3.5	10.5
	D	LB	3.4	8.7	2.8	7.1
	D	LD	3.2	7.7	2.7	6.5

4. Experiments

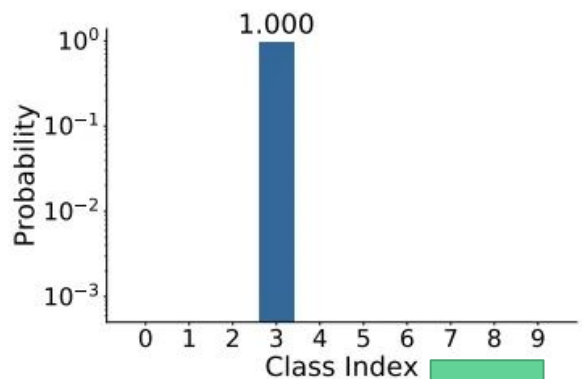
- Dataset
 - LibriSpeech
 - Switchboard
 - Switchboard includes 2,400 phone conversations across 543 U.S. speakers (302 male, 241 female), featuring 70 topics, with each speaker participating only once per topic and never repeating a conversational partner.
- Exploring:
 - Performance compare to other system
 - Different training schedules (strategies)
 - Different augmentation strategies
 - Performance of different model size
 - **Effectiveness of Label smoothing**

Switchboard

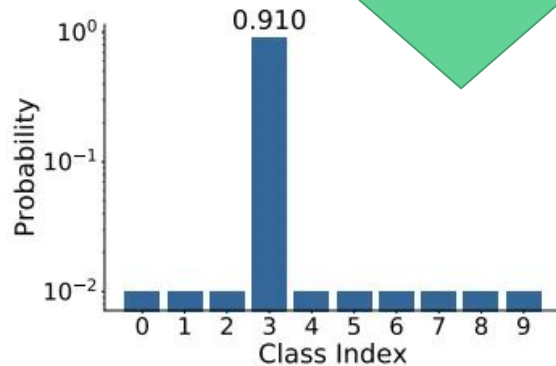
1. Label Smoothing helps

Table 4: *Switchboard 300h WER (%) evaluated for LAS-4-1024 trained with schedule B with varying augmentation and Label Smoothing (LS) policies. No LMs have been used.*

Policy	LS	SWBD	CH
-	×	12.1	22.6
	○	11.2	21.6
SM	×	9.5	18.8
	○	8.5	16.1
SS	×	9.7	18.2
	○	8.6	16.3



(a) Hard Label



(b) LS

Discussion

- Time warping contributes, but is not a major factor in improving performance
 - Most expensive with least influence

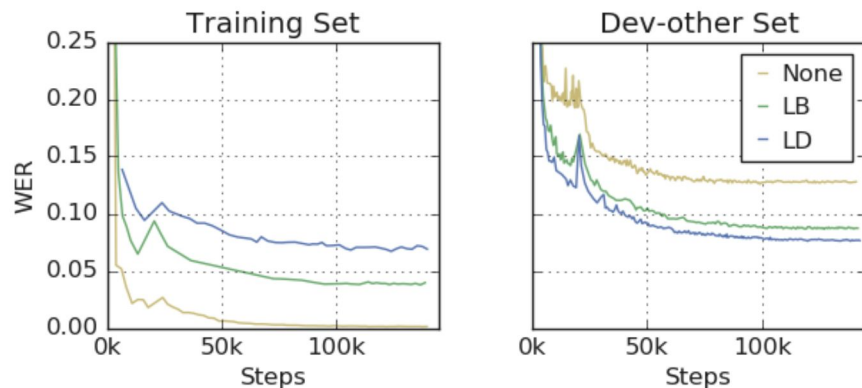
Table 6: *Test set WER (%) evaluated without LM for network LAS-4-1024 trained with schedule B.*

W	F	m_F	T	p	m_T	test-other	test
80	27	1	100	1.0	1	10.0	3.7
0	27	1	100	1.0	1	10.1	3.8
80	0	-	100	1.0	1	11.0	4.0
80	27	1	0	-	-	10.9	4.1

Time warping, time masking, frequency masking turned off

5. Discussion

- Label smoothing introduces instability to training
 - The proportion of unstable training runs increases for LibriSpeech when label smoothing is applied with augmentation
- Augmentation converts an overfitting problem into an underfitting problem
 - The networks during training not only under-fit the loss and WER on the augmented training set, but also on the training set itself when trained on augmented data



Deeper networks & training them with longer schedules

THE END :)