SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition

Pavel Fakanov

Motivation

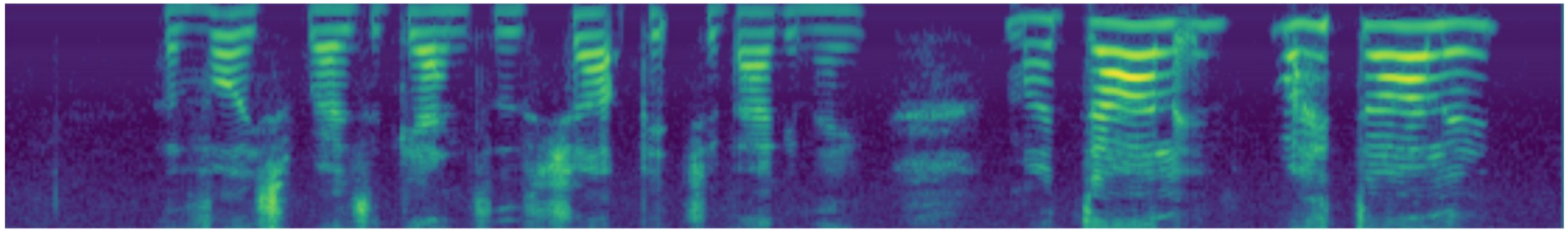
- ASR models overfit easily
- ASR models require large amounts of training data

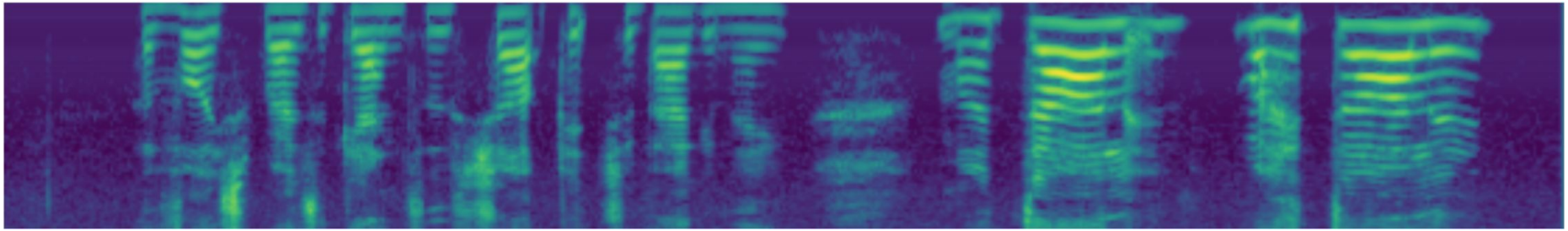
1 Augmentations

Augmentations

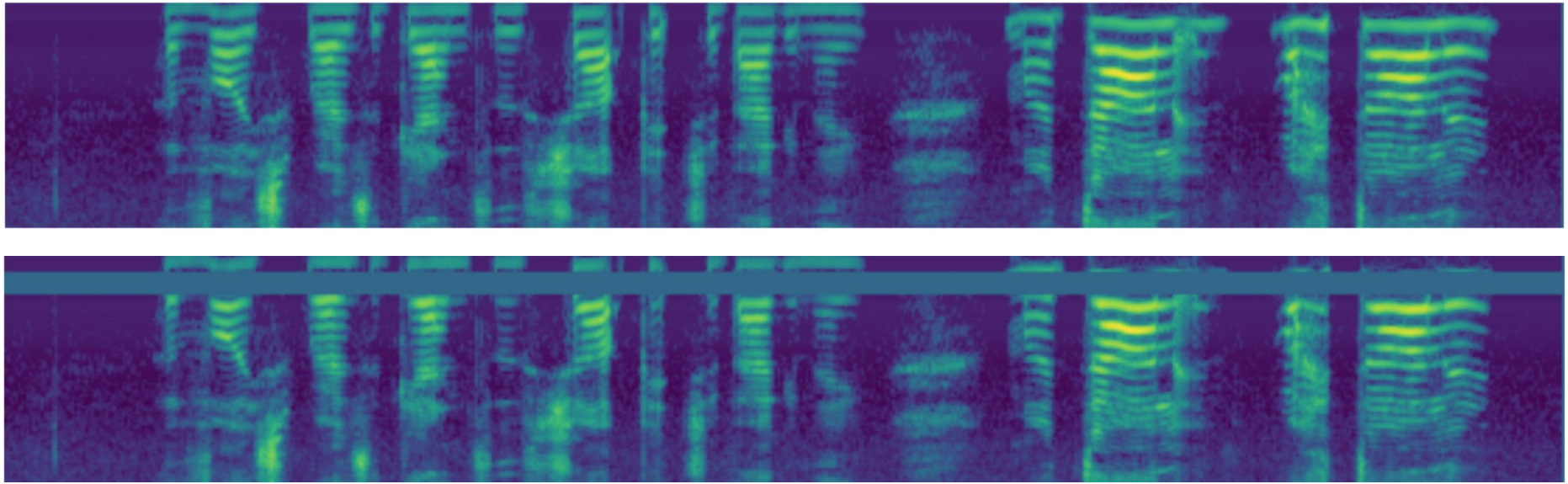
- Time wrapping
- Time masking
- Frequency masking

Time wrapping

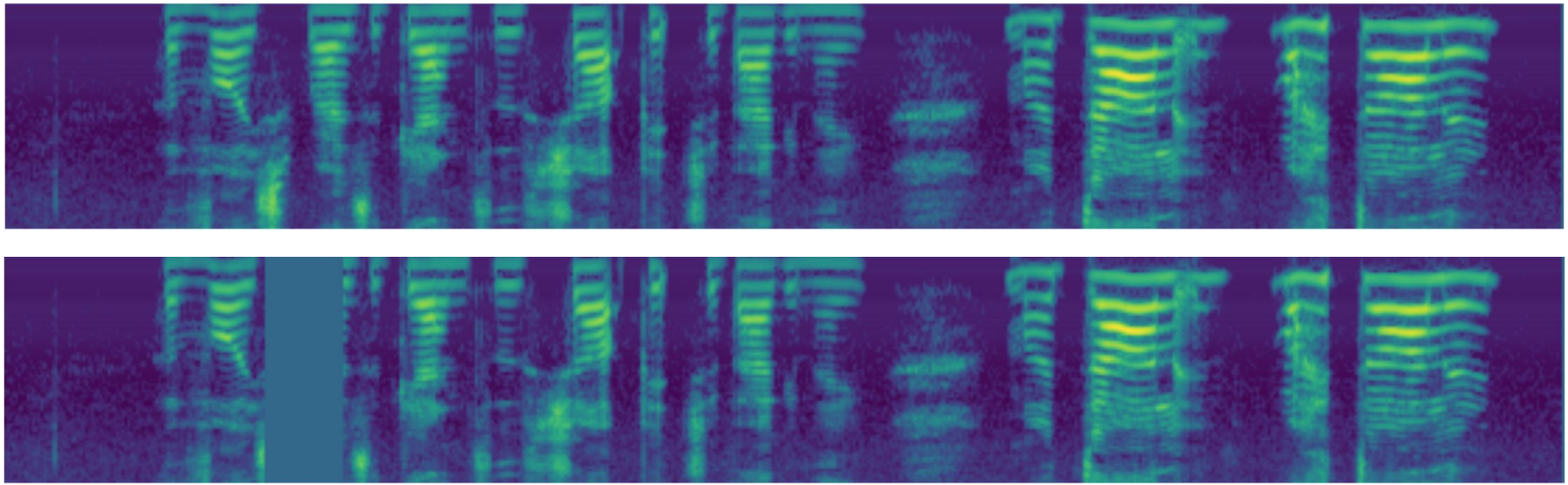




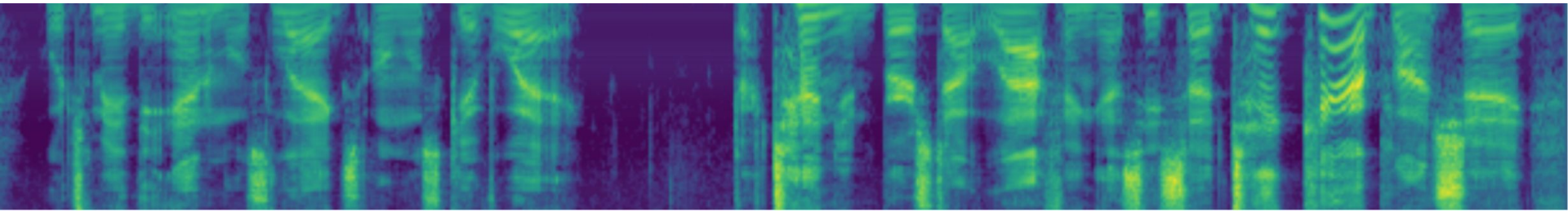
Time masking

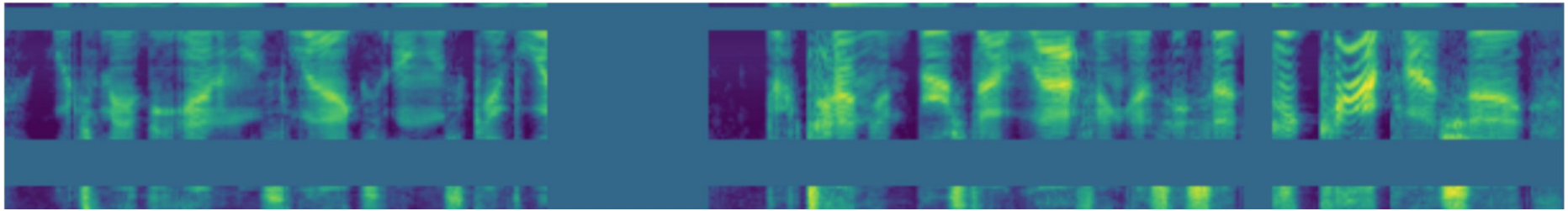


Frequency masking



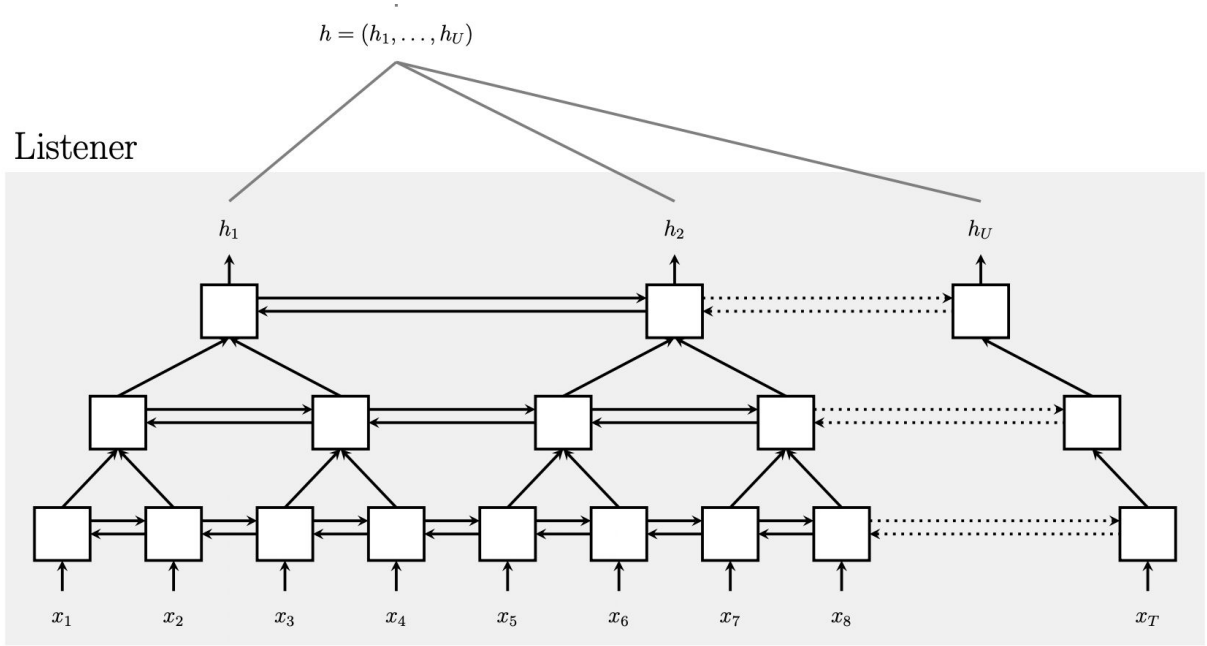
Combination of augmentations



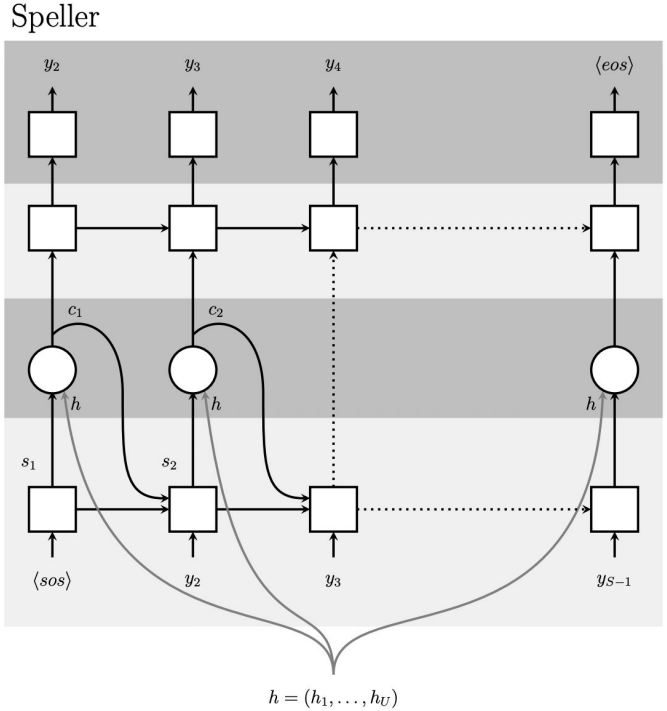


2 Model

LAS Network. Listener



LAS Network. Speller



Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from $\mathbf{h}$ and s_i

Learning Rate Schedules

- S_r – ramp-up is complete
- S_{noise} – turn on variational weight noise
- S_i – exponential decay start
- S_f – exponential decay stops

Shallow Fusion with Language Models

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

2 Experiments

LibriSpeech

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**

Switchboard 300h

Method	No LM		With LM	
	SWBD	CH	SWBD	CH
HMM				
Veselý 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**

Discussion

- Time warping, being the most expensive as well as the least influential
- Label smoothing introduces instability to training
- Augmentation converts an over-fitting problem into an under-fitting problem
- Common methods of addressing under-fitting yield improvements.

Thanks!

Any questions ?

You can find me at

- pavel.fakanov@gmail.com
- https://www.linkedin.com/in/pavel-fakanov