Comparison of German DeepSpeech models

Researcher	Training-Dataset	Duration	Noise	WER (↓)
	TD+VF	162h	+	15,1
[AZ19]	TD+VG+Mz	302h	+	21,5
	TD	127h	-	26,8
	TD+Mz	267h	+	57,3
	VF	35h	+	72,1
	Mz	140h	+	79,7
Comparison by Agarwal and Zesch (2019)				
[sil19]	multi	494h	+	9,0
[Xu+20]	TD+VF+Mz	247h	+	12,3
[sil19]	TO+VF+N	230h	+	66,0
Comparison by Xu et al. (2020)				

Comparison of different end-to-end solutions

		Dataset	
Researcher	Architecture	Clean (†)	Other (↓)
Google	LAS [Par+19]	2,80	6,80
Facebook	Transformer [Syn+20]	2,90	7,00
	Transformer (groß (‡)) [Wan+20]	3,50	7,80
	Transformer (mittel (‡)) [Wan+20]	3,70	8,10
	Transformer (klein (‡)) [Wan+20]	4,40	9,20
	RNN (groß (‡‡)) [Wan+20]	3,90	11,50
Mozilla	Menschliche Transkription [Amo+15]	5,83	12,69
	DeepSpeech 2 (RNN) [Amo+15]	5,33	13,25
	DeepSpeech 1 (RNN) [Amo+15]	7,89	21,74

†: Libri-Speech divides the test dataset into Clean and Other, where Clean contains the speech files with the lowest WER and Other contains the remaining
‡: 12 encoder/6 decoder layers each.
With 1024 (large), 512 (medium), 256 (small) units per layer
‡‡: 5 encoder/3 decoder layers with 512 units each.