Catch Me If You Can: Blackbox Adversarial Attacks on Automatic Speech Recognition using Frequency Masking-Extension File

1 Results

1.1 RQ1: Comparison of Frame Selection Techniques

We present one-way Anova and Tukey's Honest Significant Difference (HSD) test(at 5% significance level) on `WER` and `Similarity` to compare our frame selection techniques.

1.1.1 `WER`: P-values for pairwise comparisons of `WER`s between frame selection techniques.

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
All vs Random	**0.001**	**0.001**	**0.001**	**0.011**	0.07	**0.31**
All vs Important	**0.043**	**0.001**	0.06	0.40	0.9	0.43
Important vs Random	**0.001**	**0.001**	**0.006**	0.35	0.23	0.9

Table 1: P-values for pairwise comparison of `WER` achieved by frame selection methods (using `GL` attack generation).

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
All vs Random	**0.001**	**0.001**	0.23	0.58	**0.001**	0.9
All vs Important	**0.036**	**0.001**	0.28	0.51	**0.001**	0.9
Important vs Random	**0.03**	**0.032**	0.9	0.07	0.9	0.9

Table 2: P-values for pairwise comparison of `WER` achieved by frame selection methods (using `DE` attack generation).

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
All vs Random	**0.001**	**0.001**	0.59	0.58	**0.03**	0.9
All vs Important	**0.001**	**0.001**	0.80	0.9	**0.04**	0.87
Important vs Random	0.228	**0.01**	0.85	0.76	0.8	0.9

Table 3: P-values for pairwise comparison of `WER` achieved by frame selection methods (using `OP` attack generation).

1.1.2 Similarity: P-values for pairwise comparisons of Similarity between frame selection techniques.

	Librispeech	Commonvoice
Random VS All	**0.01**	**0.014**
Important VS All	0.57	0.9
Random VS Important	0.09	0.06

Table 4: P-values for pairwise comparison of Similarity achieved by frame selection methods (using GL attack generation).

	Librispeech	Commonvoice
Random VS All	**0.001**	**0.001**
Important VS All	**0.001**	**0.001**
Random VS Important	0.34	0.9

Table 5: P-values for pairwise comparison of Similarity achieved by frame selection methods (using DE attack generation).

	Librispeech	Commonvoice
Random VS All	**0.001**	**0.001**
Important VS All	**0.001**	**0.001**
Random VS Important	0.09	0.11

Table 6: P-values for pairwise comparison of Similarity achieved by frame selection methods (using OP attack generation).

Tables 6, 5 and 4 in Section 1.1.2 do not show different ASRs as the adversarial attacks are agnostic to the ASR used.

1.1.3 Pareto Front: Number of non-dominated samples for three frame selection techniques

Table 7 and 8 compares frame selection configurations for a fixed attack generation in terms of number of non-dominated samples on two datasets. Column heading in the table shows the fixed parameter; we fix one attack generation at a time and compare frame selection configurations.

	Deepspeech			Sphinx			Google		
	GL	OP	DE	GL	OP	DE	GL	OP	DE
All	4	3	1	7	5	3	5	2	5
Random	3	7	12	5	7	9	6	5	8
Important	4	9	17	7	9	13	6	9	9

Table 7: Number of non-dominated samples for frame selection techniques using different attack and ASRs on **Commonvoice**

	Deepspeech			Sphinx			Google		
	GL	OP	DE	GL	OP	DE	GL	OP	DE
All	3	3	3	7	6	4	1	2	2
Random	8	4	5	6	3	6	7	6	8
Important	9	5	7	7	6	6	8	7	13

Table 8: Number of non-dominated samples for frame selection techniques using different attack and on ASRs on **librispeech**

1.2 RQ2: Comparison of Attack Generation Techniques

We present one-way Anova and Tukey's Honest Significant Difference (HSD) test(at 5% significance level) on `WER` and `Similarity` to compare our attack generation techniques.

1.2.1 `WER`: P-values for pairwise comparisons of `WER`s between frame selection techniques.

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
GL vs OP	**0.001**	**0.001**	**0.001**	**0.009**	**0.001**	0.81
GL vs DE	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.79
OP vs DE	0.55	0.66	0.9	0.75	0.63	0.81

Table 9: P-values for pairwise comparison of `WER` achieved by attack generation methods (using `Important` frames).

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
GL vs OP	**0.001**	**0.001**	**0.007**	**0.009**	**0.001**	0.9
GL vs DE	**0.001**	**0.001**	**0.001**	**0.006**	**0.001**	0.9
OP vs DE	0.9	0.66	0.9	0.9	0.21	0.9

Table 10: P-values for pairwise comparison of `WER` achieved by attack generation methods (using `Random` frames).

	Librispeech			Commonvoice		
	Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google
GL vs OP	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.189
GL vs DE	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	0.05
OP vs DE	**0.04**	**0.03**	0.60	0.9	0.58	0.818

Table 11: P-values for pairwise comparison of `WER` achieved by attack generation methods (using `All` frames).

1.2.2 Similarity: P-values for pairwise comparisons of Similarity between attack generation techniques.

	Librispeech	Commonvoice
OP VS GL	**0.001**	**0.001**
DE VS GL	**0.001**	**0.001**
DE VS OP	0.1	**0.001**

Table 12: P-values for pairwise comparison of Similarity achieved by attack generation methods (using Important frames).

	Librispeech	Commonvoice
OP VS GL	**0.001**	**0.001**
DE VS GL	**0.001**	**0.001**
DE VS OP	0.38	**0.001**

Table 13: P-values for pairwise comparison of Similarity achieved by attack generation methods (using Random frames).

	Librispeech	Commonvoice
OP VS GL	**0.001**	**0.001**
DE VS GL	**0.001**	**0.001**
OP VS DE	0.06	0.56

Table 14: P-values for pairwise comparison of Similarity achieved by attack generation methods (using All frames).

1.2.3 Pareto Front: Number of non-dominated samples for three attack generation techniques

Table 15 and 16 compares attack generation configurations for a fixed frame selection in terms of number of non-dominated samples on two datasets.

	Deepspeech			Sphinx			Google		
	All	Random	Important	All	Random	Important	All	Random	Important
GL	2	1	2	7	7	10	1	1	5
OP	10	0	5	8	1	2	11	3	5
DE	6	17	16	9	23	25	8	20	12

Table 15: Number of non-dominated samples for attack generation techniques using different frame selection techniques and ASRs on **Commonvoice**

	Deepspeech			Sphinx			Google		
	All	Random	Important	All	Random	Important	All	Random	Important
GL	4	2	4	7	7	6	3	3	2
OP	8	3	7	8	1	6	8	1	5
DE	1	7	9	4	7	8	2	4	8

Table 16: Number of non-dominated samples for attack generation techniques using different frame selection techniques and ASRs on **Librispeech**

1.3 RQ4: Comparison with Abdullah et al.

Results comparing our attack with Abdullah et al. on Librispeech dataset is shown in Table 18. We present one-way Anova and Tukey's Honest Significant Difference (HSD) test (at 5% significance level) on `WER` and `Similarity` in Tables 17 and 19 for Commonvoice and Librispeech datasets, respectively.

1.3.1 P-values for the comparison of `WER` and `Similarity` between our approach and Abdullah et al. on Commonvoice dataset.

	Similarity	WER on Deepspeech	WER on Sphinx	WER on Google
`OP+All` vs Abdullah's work	**0.041**	**0.026**	**0.037**	**0.001**
`OP+Important` vs Abdullah's work	**0.001**	0.66	0.08	**0.003**
`DE+All` vs Abdullah's work	0.79	**0.027**	**0.013**	**0.001**
`DE+Important` vs Abdullah's work	**0.001**	0.88	0.06	**0.001**

Table 17: P-values for pairwise comparison of `Similarity` and `WER` achieved by Abdullah et al, against `OP+All, OP+Important, DE+All, DE+Important` on Commonvoice dataset.

1.3.2 Comparison with Abdullah et al. on Librispeech Dataset

Technique	Time	Similarity	Success rate			WER			Detection score
			Deepspeech	Sphinx	Google	Deepspeech	Sphinx	Google	
Abdullah	22 seconds	2.6	76%	86%	50%	0.10	0.18	0.06	0.40
OP	3.5 seconds	3.65	95%	96.5%	97.5%	0.17	0.28	0.20	0.14
DE	2.5 seconds	3.72	91%	94%	95.5%	0.11	0.19	0.20	0.11

Table 18: Comparison of `OP, DE` with Abdullah et al. with respect to generation time for per adversarial audio sample, `Similarity` to original audio samples,`WER`, `Success Rate` and `Detection score` against defense system in attacking all three ASRs on Librispeech dataset

1.3.3 P-values for the comparison of `WER` and `Similarity` between our approach and Abdullah et al. on Librispeech dataset.

	Similarity	WER on Deepspeech	WER on Sphinx	WER on Google
`OP+All` vs Abdullah's work	**0.001**	**0.001**	**0.012**	**0.001**
`OP+Important` vs Abdullah's work	**0.001**	0.38	0.56	**0.001**
`DE+All` vs Abdullah's work	**0.009**	0.077	0.072	**0.001**
`DE+Important` vs Abdullah's work	**0.009**	0.13	0.9	**0.001**

Table 19: P-values for comparison of `Similarity` and `WER` achieved by Abdullah et al. against `OP+All, OP+Important, DE+All, DE+Important` on Librispeech dataset.