Paper ID: 2515

NUS National University of Singapore

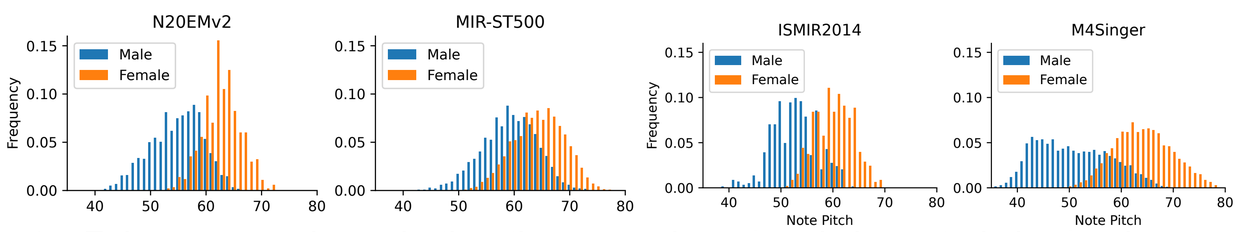
ACM MM 2023
Ottawa, Canada
October 29 – November 3, 2023

Elucidate Gender Fairness in Singing Voice Transcription

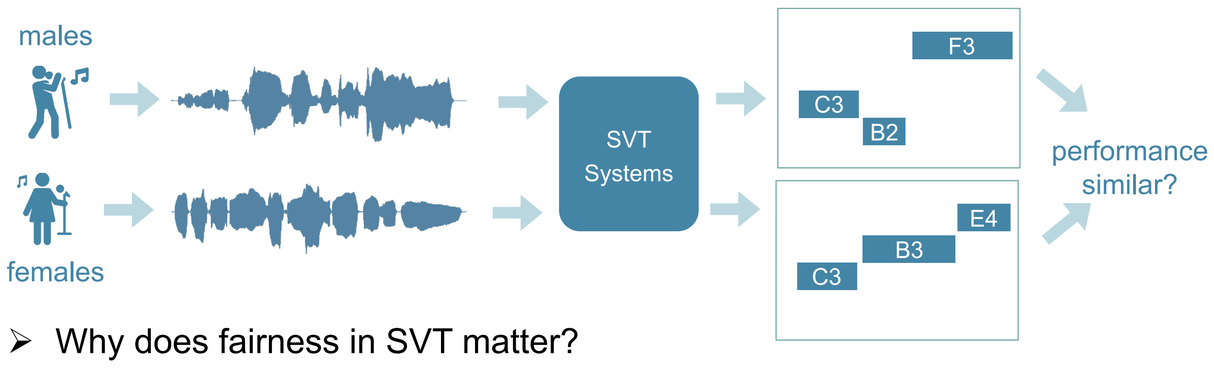
Xiangming Gu, Wei Zeng, Ye Wang

• Motivation:

➢ Females and males sing differently. They have different sound characteristics, e.g. timbre and pitch distributions:



➢ Fairness question: whether these gender-related characteristics lead to a performance disparity in singing voice transcription.



➢ Why does fairness in SVT matter?

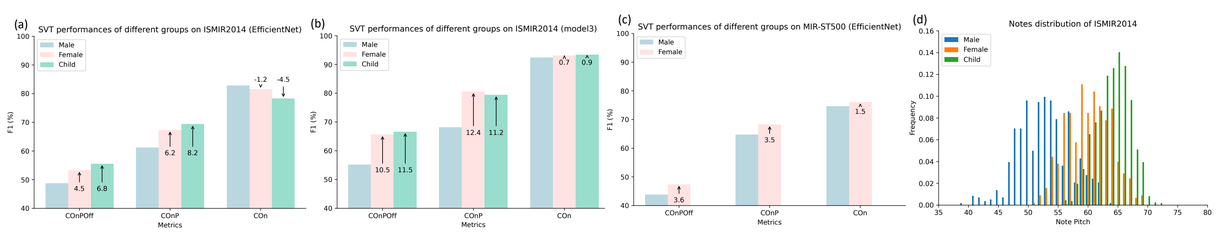
1. Biased machine learning systems can lead to discriminatory treatment of certain groups.
2. Biased SVT systems can lead to user inconvenience
3. Biased SVT systems negatively affect downstream applications, e.g. singing voice synthesis

• Fairness Analysis in SVT:

➢ Female superiority in SVT: evidence from four benchmark singing datasets.

Dataset	Model	COnPOff (%)				COnP (%)				COn (%)			
		total ↑	female ↑	male ↑	gap	total ↑	female ↑	male ↑	gap	total ↑	female ↑	male ↑	gap
MIR-ST500	model1	52.39	**53.11**	51.47	-1.64	70.73	**72.54**	68.42	-4.12	78.32	**79.36**	77.00	-2.36
	model2	34.55	**35.29**	33.60	-2.55	51.64	**52.76**	50.21	-2.55	71.33	**72.07**	70.39	-1.68
	model3	52.84	**54.02**	51.33	-2.69	70.00	**71.17**	67.85	-3.32	78.05	**78.84**	77.05	-1.79
N20EMv2	model1	55.20	**60.96**	51.55	**-9.41**	72.03	**79.76**	67.11	**-12.65**	88.51	**90.63**	87.16	-3.47
	model2	68.62	**74.82**	64.68	**-10.14**	75.69	**81.27**	72.14	**-9.12**	92.83	**94.33**	91.88	-2.44
	model3	73.06	**78.34**	69.69	**-8.65**	79.56	**84.38**	76.49	**-7.89**	93.66	**94.96**	92.83	-2.13
ISMIR2014	model1	52.58	**61.22**	45.28	**-15.94**	67.75	**74.90**	61.70	**-13.20**	92.13	91.93	**92.30**	+0.37
	model2	57.35	**62.42**	53.06	**-9.36**	72.15	**79.35**	66.06	**-13.29**	91.53	**92.25**	90.92	-1.33
	model3	59.95	**65.61**	55.16	**-10.45**	73.85	**80.55**	68.19	**-12.36**	92.80	**93.18**	92.49	-0.69
M4Singer	wav2vec2	53.66	**57.27**	49.93	**-7.34**	61.95	**66.36**	57.38	**-8.98**	82.60	81.66	**83.58**	+1.91
	Hubert	55.11	**58.39**	51.73	**-6.66**	64.17	**68.10**	60.10	**-8.00**	82.13	81.53	**82.76**	+1.23
	wavLM	57.06	**60.33**	53.68	**-6.66**	65.40	**68.87**	61.82	**-7.05**	82.73	82.29	**83.19**	+0.90
	data2vec	53.98	**57.45**	50.40	**-7.05**	63.04	**67.09**	58.86	**-8.23**	82.30	82.13	**82.48**	+0.35

➢ Female superiority in SVT: evidence from SSL-based SVT models and EfficientNet-based SVT models



➢ Potential reasons:

1. Rule out the gender data imbalance
2. Hypothesis: pitch distributions make substantial contributions.

Dataset	Songs Num			Duration (h)		
	total	female	male	total	female	male
MIR-ST500	400	**221**	179	27.62	**15.13**	12.49
N20EMv2	123	52	**71**	6.44	2.74	**3.70**
M4Singer	521	246	**275**	22.71	10.27	**12.44**

• Fairness formulation in SVT:

➢ Training data — Test data

$$\mathcal{D}_{\mathcal{S}} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}, A^{(n)})\}_{n=1}^{N_{\mathcal{S}}} \qquad \mathcal{D}_{\mathcal{T}} = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}, A^{(n)})\}_{n=1}^{N_{\mathcal{T}}}$$

➢ SVT metric
COnPOff: correct onset, offset, pitch — COnP: correct onset, pitch

➢ Utility metric

$$U = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_{\mathcal{T}}}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right]$$

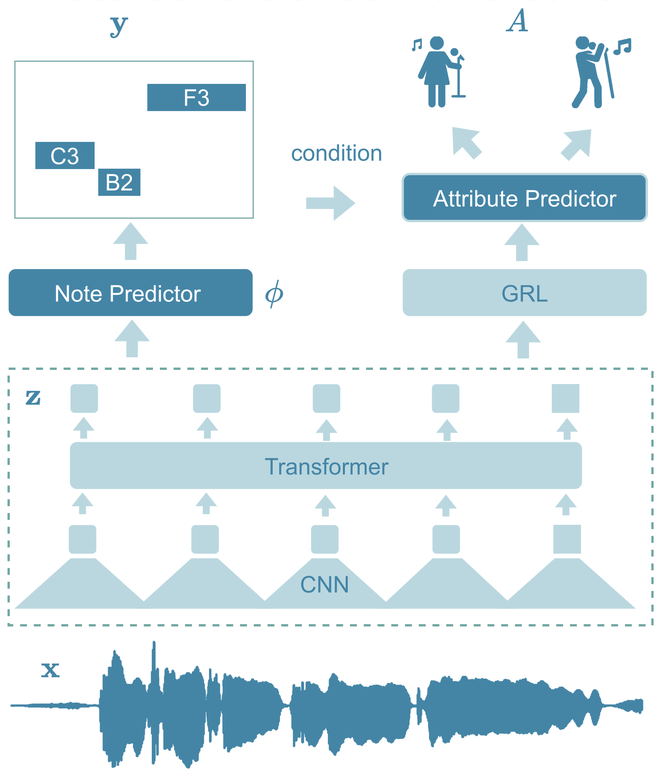
➢ Fairness metric ($A = 1$: males $A = 0$: females)

$$F = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y},A=1)\sim\mathcal{D}_{\mathcal{T}}}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y},A=0)\sim\mathcal{D}_{\mathcal{T}}}\left[\text{metric}(\hat{\boldsymbol{y}}, \boldsymbol{y})\right]$$

➢ Fairness-Utility tradeoff

$$\max \min\{F, 0\}, \quad \max U$$

• Note-conditioned Adversarial Learning for Fairness:



➢ Introduce attribute predictor to predict gender labels

$$\mathcal{L}_A = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})\sim\mathcal{D}_{\mathcal{S}}}\left[\frac{1}{T}\sum_{t=1}^{T} l_{\text{BCE}}(\hat{A}_t, A)\right]$$

$$\hat{A} = \psi(\boldsymbol{z}) = \psi \circ \theta(\boldsymbol{x})$$

➢ Baseline Adversarial Learning (no condition)

$$\hat{\boldsymbol{y}} \perp A$$

➢ NCAL variant 1 (condition on ground truth notes)

$$\boldsymbol{z} \perp A|\boldsymbol{y}$$

➢ NCAL variant 2 (condition on note predictions): align different pitches simultaneously

$$\boldsymbol{z} \perp A|\hat{\boldsymbol{y}}$$

Algorithm 1 Note-conditioned adversarial learning for SVT

Require: Acoustic encoder $\theta^{(0)}$ pre-trained on speech data under SSL objective, randomly initialized label predictor $\phi^{(0)}$ and attribute predictor $\psi^{(0)}$, learning rates η_1, η_2, η_3 for θ, ϕ, ψ, training steps K_1, K_2 for linear probing and full finetuning.

for $k = 1$ **to** $K_1 + K_2$ **do**
- $\boldsymbol{z} = \theta^{(k-1)}(\boldsymbol{x})$, $\hat{\boldsymbol{y}} = \phi^{(k-1)}(\boldsymbol{z})$
- $\hat{A}_t = \psi^{(k-1)}(\boldsymbol{z}.\text{detach}(), \hat{\boldsymbol{y}}.\text{detach}())$, compute $\mathcal{L}_A$ in Eq. 6
- $\psi^{(k)} = \psi^{(k-1)} - \eta_3 \frac{\partial \mathcal{L}_A}{\partial \psi^{(k-1)}}$ ▷ Update ψ
- $\hat{A}_t = \psi^{(k)}(\boldsymbol{z}, \hat{\boldsymbol{y}})$, compute $\mathcal{L}_y, \mathcal{L}_A$ in Eq. 2 and 6
- $\phi^{(k)} = \phi^{(k-1)} - \eta_2 \frac{\partial \mathcal{L}_y}{\partial \phi^{(k-1)}}$ ▷ Update ϕ
- **if** $k \le K_1$ **then** ▷ Update θ
  - $\theta^{(k)} = \theta^{(k-1)}$
- **else**
  - $\theta^{(k)} = \theta^{(k-1)} - \eta_1\left(\frac{\partial \mathcal{L}_y}{\partial \theta^{(k-1)}} - \lambda \frac{\partial \mathcal{L}_A}{\partial \theta^{(k-1)}}\right)$
- **end if**

end for

• Enforce Fairness in SVT:

➢ Gender Bias Mitigation Results of NCAL.

Train set	Test set	Method	COnPOff (%)		COnP (%)	
			Utility (U) ↑	Fairness (F) ↑	Utility (U) ↑	Fairness (F) ↑
M4Singer	M4singer	ERM	53.66	- 7.34	61.95	- 8.98
		Ours	52.48 (-1.18)	- 3.61 (+3.73)	60.67 (-1.28)	- 4.21 (+4.77)
MIR-ST500	N20EMv2	ERM	55.20	- 9.41	72.03	-12.65
		Ours	53.29 (-1.91)	- 4.14 (+5.27)	72.71 (+0.68)	-10.82 (+1.83)
	ISMIR2014	ERM	52.58	-15.94	67.75	-13.20
		Ours	48.18 (-4.40)	- 8.29 (+7.65)	65.40 (-2.35)	- 9.08 (+4.12)
MIR-ST500 N20EMv2	N20EMv2	ERM	73.06	- 8.65	79.56	- 7.89
		Ours	72.43 (-0.63)	- 5.78 (+2.87)	78.47 (-1.09)	- 6.82 (+1.07)
	ISMIR2014	ERM	59.95	-10.45	73.85	-12.36
		Ours	59.57 (-0.38)	- 7.49 (+2.96)	73.33 (-0.52)	- 7.75 (+4.61)

➢ Why do we need fairness-utility tradeoff?

Train set	Test set	Method	η_3	λ	COnPOff (%)		COnP (%)	
					Utility (U) ↑	Fairness (F) ↑	Utility (U) ↑	Fairness (F) ↑
M4Singer	M4singer	ERM	-	-	53.66	- 7.34	61.95	- 8.98
		AL	1.0	3.0	42.56 (-11.10)	+ 4.65	49.60 (-12.35)	+ 6.35
		NCAL	0.1	5.0	49.97 (- 3.69)	+ 0.39	57.66 (- 4.29)	+ 1.57
		NCAL	0.1	3.0	49.87 (- 3.79)	+ 1.29	58.53 (- 3.42)	+ 0.60
MIR-ST500	N20EMv2	ERM	-	-	55.20	- 9.41	72.03	-12.65
		NCAL	0.0001	5.0	45.78 (- 9.42)	+ 2.45	69.54 (- 2.49)	- 3.77
	ISMIR2014	ERM	-	-	52.58	-15.94	67.75	-13.20
		NCAL	0.0001	5.0	42.24 (-10.34)	+ 3.01	62.44 (- 5.31)	+ 1.22
MIR-ST500 N20EMv2	N20EMv2	ERM	-	-	73.06	- 8.65	79.56	- 7.89
		NACL	0.001	5.0	72.28 (- 0.78)	- 7.78	78.70 (- 0.86)	- 8.23
	ISMIR2014	ERM	-	-	59.95	-10.45	73.85	-12.36
		NACL	0.001	5.0	50.95 (- 9.00)	+ 2.49	65.87 (- 7.98)	+ 0.64

➢ Compared to baseline adversarial learning and domain independent learning