NANYANG TECHNOLOGICAL UNIVERSITY

NTU-NPU System for Voice Privacy 2024 Challenge

Nikita Kuzmin, Hieu-Thi Luong, Jixun Yao, Lei Xie, Kong Aik Lee, Eng-Siong Chng

Motivation

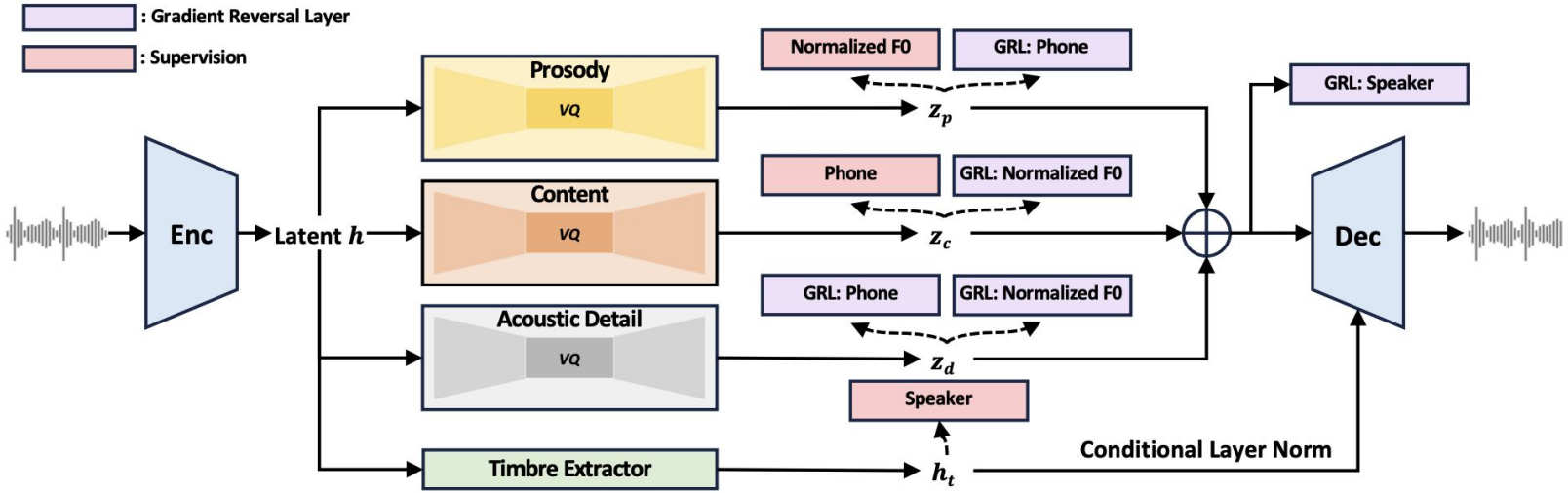
1. Analysis of baseline systems and push baseline performance for privacy and utility metrics.

2. Baseline models are lacking of Emotion recognition performance.

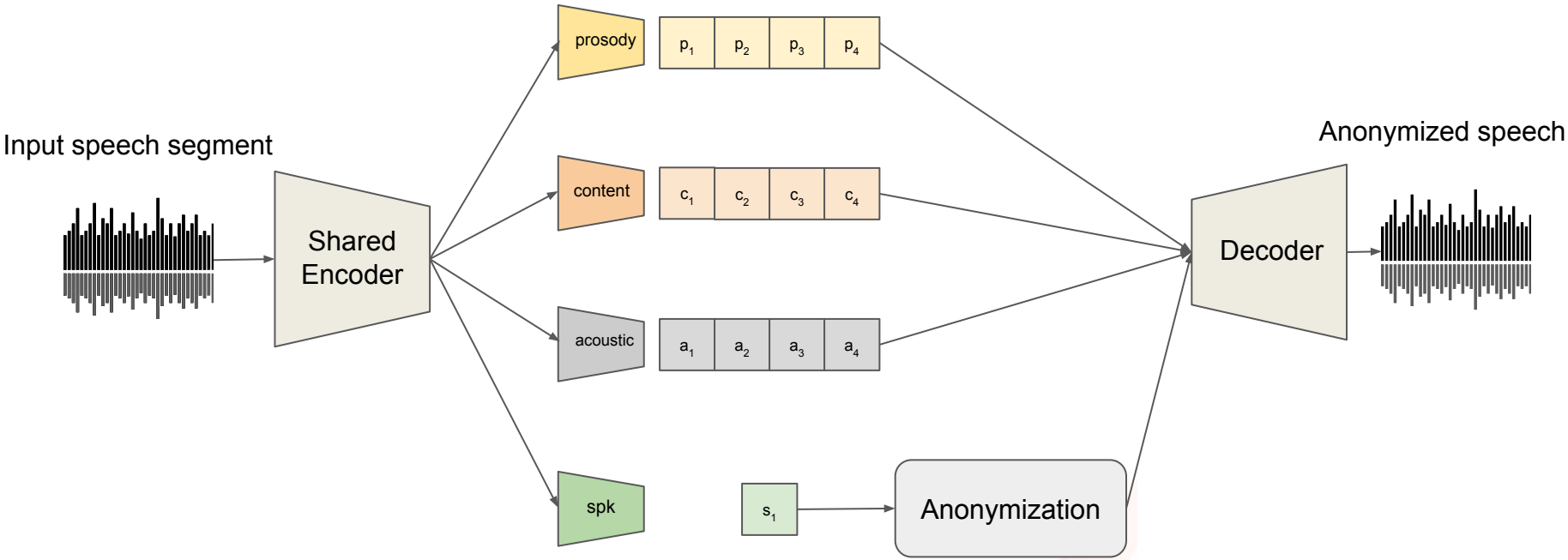
3. Interest in disentanglement-based models such as NaturalSpeech3 FACodec.

NaturalSpeech3 FACodec [1]

NaturalSpeech3 FACodec [1]



NaturalSpeech3 FACodec [1]



NANYANG TECHNOLOGICAL UNIVERSITY

NaturalSpeech3 FACodec [1]

#	Module	Description	Output features	Data
1	Encoder [2]	4 Downsampling Convolution-based Layers with Snake activation function Input: speech waveform	Output vector^{256}	Librilight train[3]
2	Prosody extractor	Factorized Vector Quantization with 1 quantizer, codebook size: 1024	Prosody vector^{256}	Librilight train
3	Content extractor	Factorized Vector Quantization with 2 quantizers, codebook size: 1024	Content vector^{256}	Librilight train
4	Speaker embedding extractor	Several Conformer blocks	Speaker embedding^{256}	Librilight train
5	Speaker anonymization module	Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring AWGN with scale= 0.075 Cross-gender	Anonymized speaker embedding^{256}	LibriTTS: [4] train-clean-100
6	Decoder [2]	Upsampling Convolution-based Layers with Snake activation function	speech waveform	Librilight train

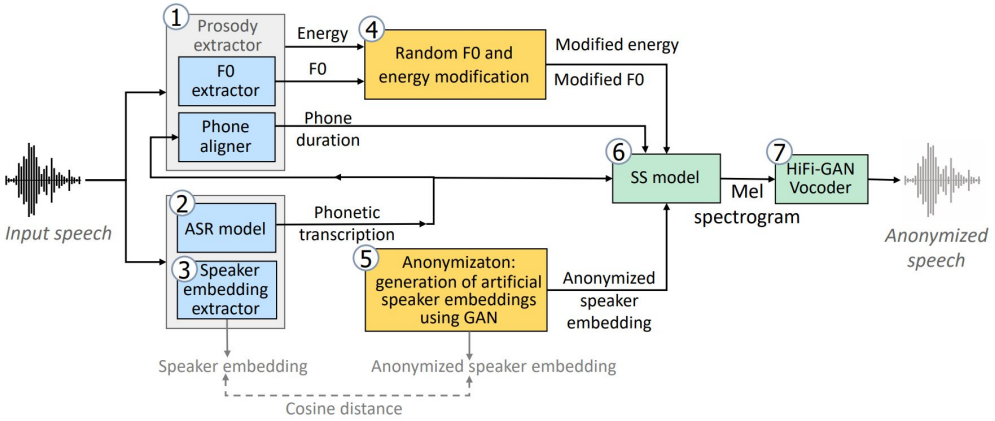
Experiments. NS3 + AWGN to Speaker Embedding + Cross Gender

Speaker Anon	AWGN	Cross Gender	EER		UAR		WER	
			dev	test	dev	test	dev	test
–	–	–	7.40	6.25	63.36	62.46	2.69	2.51
+	–	–	9.29	8.78	51.64	52.89	2.97	2.77
+	+	–	12.25	9.14	48.00	48.09	4.66	4.63
+	+	+	12.09	10.46	49.20	49.12	4.97	4.60

Baseline B3 [5]

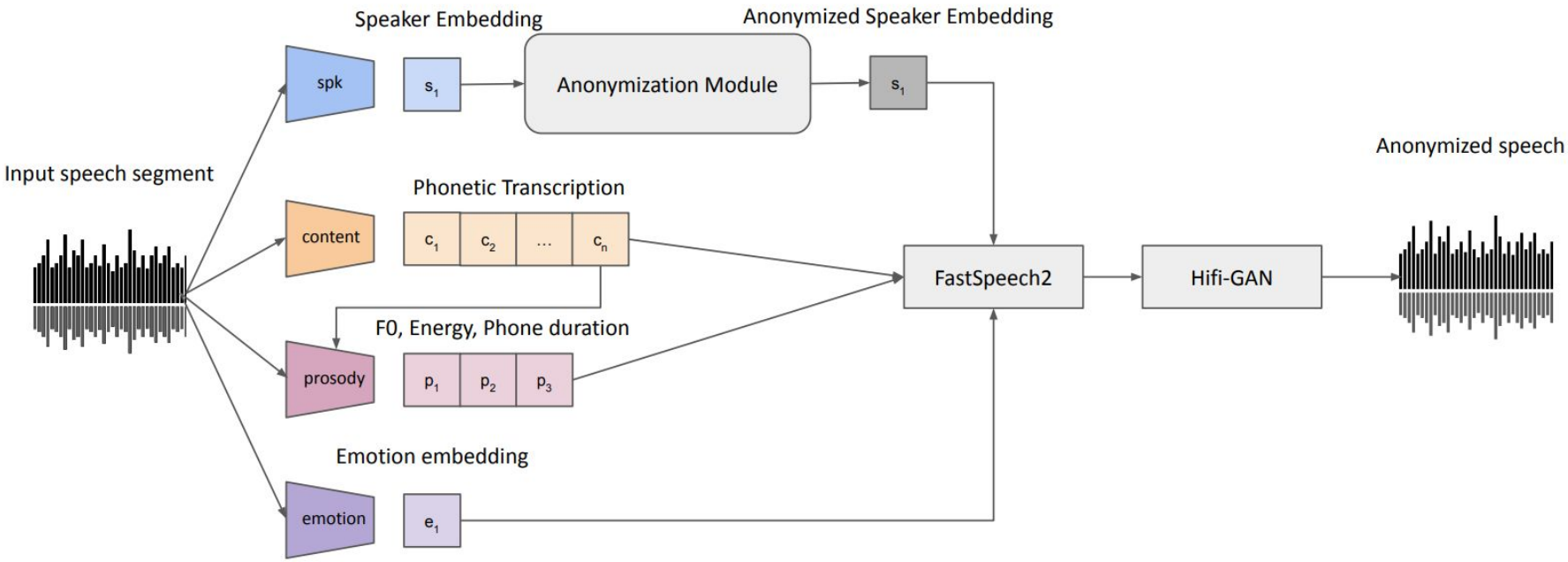
Baseline B3

#	Module	Description	Output features	Data
1	Prosody extractor	Phone aligner: 6-layer CNN + LSTM with CTC loss F0 estimation using Praat F0, energy, durations normalized by each vector's mean	F0[1], energy[1] phone durations[1]	LibriTTS: train-clean-100
2	ASR	End-to-end with hybrid CTC-attention Input: log mel Fbank[80] Encoder: Branchformer Decoder: Transformer Output: phone sequences CTC and attention criteria	phonetic transcript with pauses and punctuation	LibriTTS: train-clean-100 train-other-500
3	Speaker embedding extractor	GST trained jointly with SS model Input: mel spectrogram[80] 6 hidden layers + 4-head attention Output: GST speaker embedding[128]	GST speaker embedding[128]	LibriTTS: train-clean-100
4	Prosody modification module	Value-wise multiplication of F0 and energy with random values in $[0.6, 1.4)$	F0[1], energy[1]	-
5	Speaker anonymization module	Wasserstein GAN Input: Random noise[16] from normal distribution Generator: ResNet with three residual blocks, 150k params Critic: ResNet with three residual blocks, 150k params Output: MSE and Quadratic Transport Cost criteria	pseudo-speaker GST embeddings[128]	LibriTTS: train-clean-100 RAVDESS ESD
6	SS model	*IMS Toucan* implementation of *FastSpeech2* Input: F0[1] + energy[1] + phone duration[1] + phonetic transcript + GST embeddings[128] Training criterion defined in *FastSpeech2*	mel spectrogram[80]	LibriTTS: train-clean-100
7	Vocoder	HiFi-GAN vocoder Input: mel spectrogram[80] Training criterion defined in Hifi-GAN	speech waveform	LibriTTS: train-clean-100



NANYANG TECHNOLOGICAL UNIVERSITY

Modified B3


Speaker Embedding
Anonymized Speaker Embedding
spk
s_1
Anonymization Module
s_1
Input speech segment
Anonymized speech
Phonetic Transcription
content
c_1 c_2 … c_n
FastSpeech2
Hifi-GAN
F0, Energy, Phone duration
prosody
p_1 p_2 p_3
Emotion embedding
emotion
e_1

Modified B3

#	Module	Description	Output features	Data
1	Prosody extractor	Phone aligner: 6-layer CNN + LSTM with CTC loss F0 estimation using Praat F0, energy, durations normalized by each vector's mean	$F0^{1}$, $energy^{1}$ phone $durations^{1}$	LibriTTS: train-clean-100
2	ASR	End-to-end with hybrid CTC-attention Input: log mel $Fbank^{80}$ Encoder: Branchformer Decoder: Transformer CTC and attention criteria	phonetic transcript with pauses and punctuation	LibriTTS: train-clean-100 train-other-500
3	Speaker embedding extractor	GST, trained jointly with SS model Input: mel $spectrogram^{80}$ 6 hidden layers + 4-head attention	GST speaker $embedding^{128}$	LibriTTS: train-clean-100
4	Emotion embedding extractor	**1b, 2a:** Dimensional Speech Emotion Recognition Model based on Wav2vec 2.0 Input: Wav2vec 2.0 Large features	emotion $embedding^{1024}$	MSP-Podcast (v1.7)
		2b: –	–	–
5	Prosody modification module	**1b, 2b:** –	–	–
		2a: Value-wise multiplication of F0 and energy with random values in [0.7, 1.3)	$F0^{1}$, $energy^{1}$	LibriTTS: train-clean-100
6	Speaker anonymization module	**1b:** Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring + cross-gender **2a, 2b:** Random Speaker selection per each source utterance + cross-gender	Anonymized speaker $embedding^{128}$	LibriTTS: train-clean-100
7	SS model	IMS Toucan implementation of FastSpeech2 Input: $F0^{1}$ + $energy^{1}$ + phone $durations^{1}$ + phonetic transcript + GST $embeddings^{128}$ (**1b, 2a:** + emotion $embeddings^{1024}$) Training criterion defined in FastSpeech2	mel $spectrogram^{80}$	LibriTTS: train-clean-100
8	Vocoder	HiFi-GAN vocoder Input: mel $spectrogram^{80}$ Training criterion defined in HiFi-GAN	speech waveform	LibriTTS: train-clean-100

NANYANG TECHNOLOGICAL UNIVERSITY

Experiments. B3 + Emotion embedding

Speaker Anonymization	Speaker Embedder	Prosody Anonymization	Emotion Embedding	EER		UAR		WER	
				dev	test	dev	test	dev	test
–	–	–	–	5.72	4.59	69.08	71.06	1.80	1.85
+	GST	+	–	25.76	28.42	37.97	37.39	4.33	4.33
+	GST	+	+	22.59	24.09	42.52	41.74	4.39	4.40

NANYANG TECHNOLOGICAL UNIVERSITY

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938

Experiments. B3 + Prosody Modification

Multiplier Range	EER		UAR		WER	
	dev	test	dev	test	dev	test
[0.6, 1.4]	25.76	28.42	37.97	37.39	4.33	4.33
[0.7, 1.3]	23.93	25.62	37.49	37.59	4.07	4.05
[0.8, 1.2]	22.70	25.92	38.01	37.96	3.89	3.91
[0.9, 1.1]	19.88	22.62	39.03	37.17	3.80	3.77
–	19.47	21.82	38.91	38.11	3.70	3.75

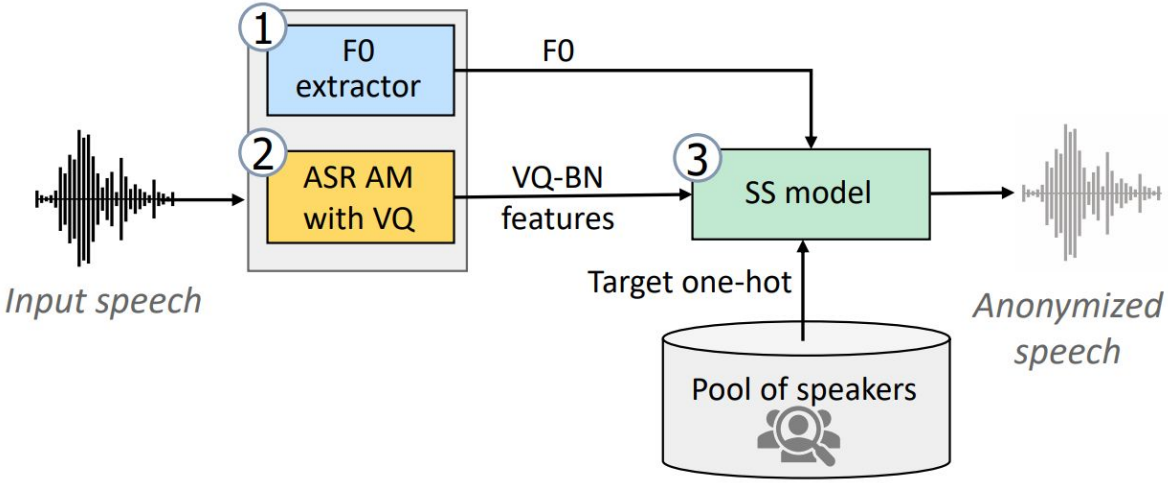
NANYANG TECHNOLOGICAL UNIVERSITY

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938

Baseline B5 [5]

Baseline B5

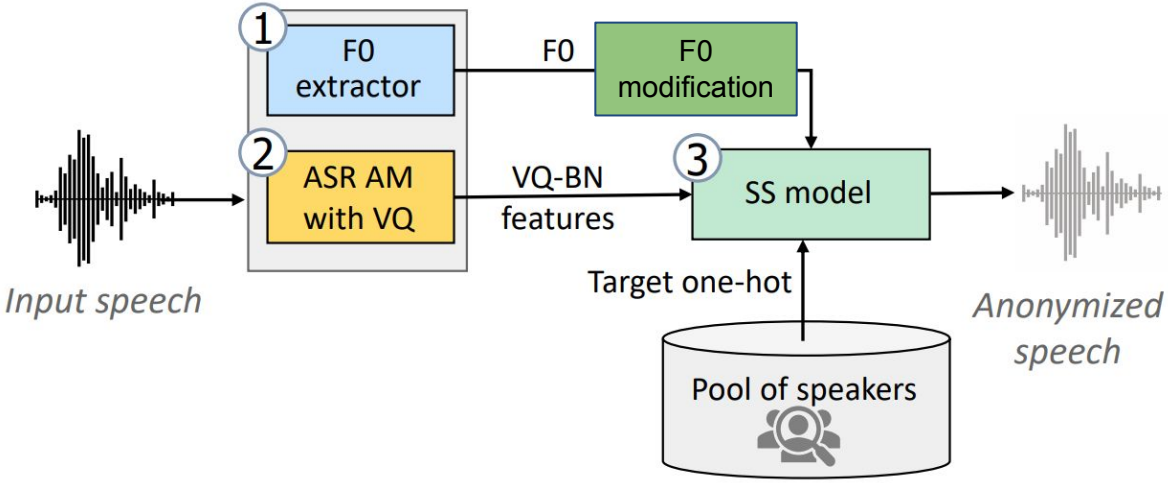
#	Module	Description	Output features	Data
1	F0 extractor	F0 extracted with s pytorch implementation of YAAPT	F0	N/A
2	ASR AM with VQ	Acoustic Model trained to identify left bi-phones and a VQ bottleneck layer	Linguistic representation	VoxPopuli Librispeech: train-clean-100
3	Speaker embedding	One-hot vector represented speaker in training set	Speaker embedding	LibriTTS: train-clean-100
4	Speech Synthesis	HiFi-GAN vocoder Input: F0 + lingusitic representation + speaker embedding	Speech waveform	LibriTTS: train-clean-100



NANYANG TECHNOLOGICAL UNIVERSITY

Modifications of B5

#	Module	Description	Output features	Data
1	F0 extractor	F0 extracted with s pytorch implementation of YAAPT **3:** Using Mean Reversion F0 ($\alpha = 0.75$) in inference **4:** Using Mean Reversion F0 ($\alpha = 0.75$) and 10-db AWGN	F0	N/A
2	ASR AM with VQ	Acoustic Model trained to identify left bi-phones and a VQ bottleneck layer	Linguistic representation	VoxPopuli Librispeech: train-clean-100
3	Speaker embedding	One-hot vector represented speaker in training set	Speaker embedding	LibriTTS: train-clean-100
4	Speech Synthesis	HiFi-GAN vocoder Input: F0 + lingusitic representation + speaker embedding	Speech waveform	LibriTTS: train-clean-100



NANYANG TECHNOLOGICAL UNIVERSITY

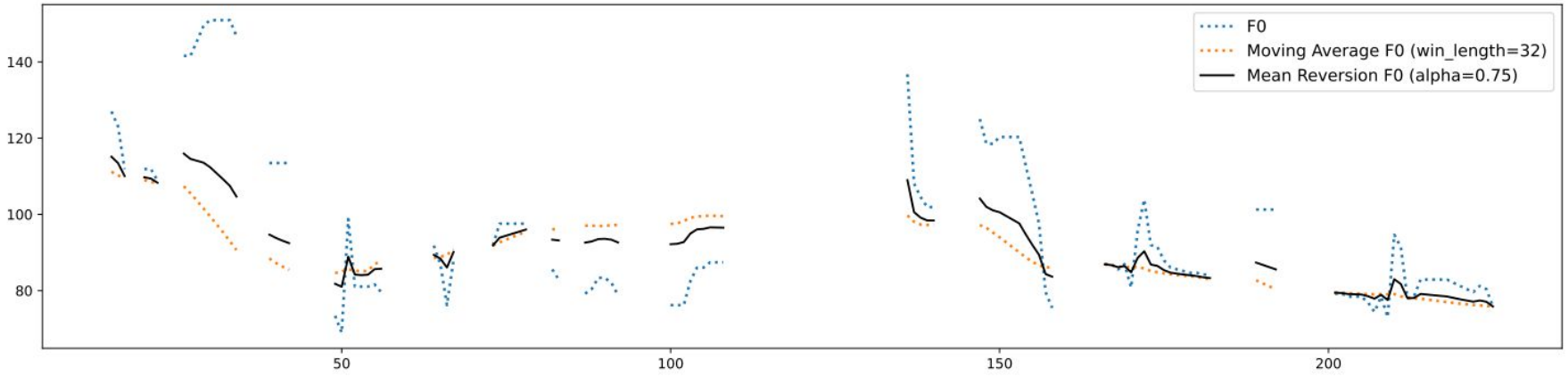
Mean Reversion of F_0

F_0 – original prosody

$\overline{F_0}$ – n-frame moving average (n=32)

$$\hat{F_0} = (1 - \alpha)F_0 + \alpha\overline{F_0}$$

Experiments. B5 + Mean Reversion F0

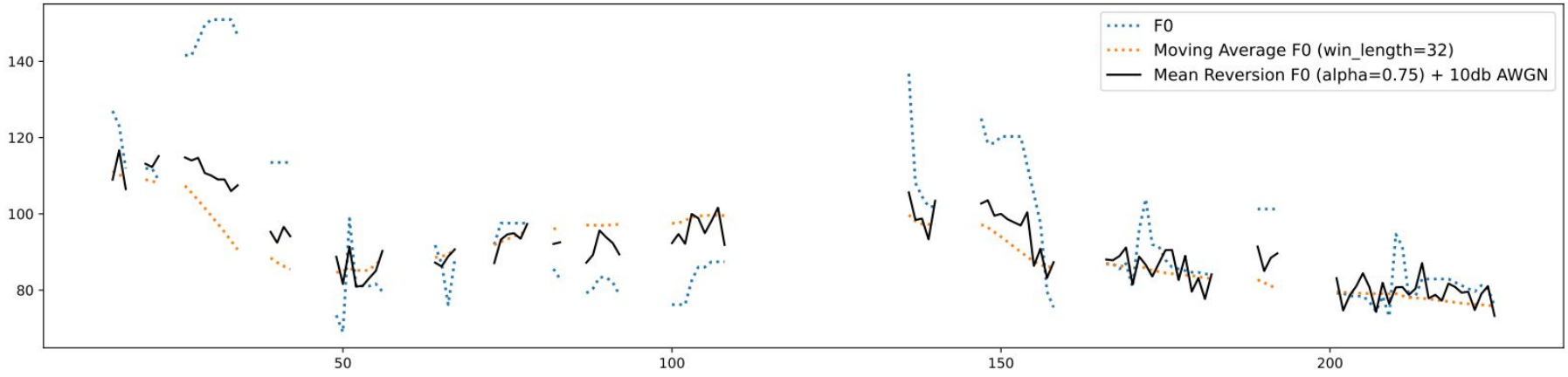


(a) Mean Reversion F0

Experiments. B5 + Mean Reversion F0

α	EER		UAR		WER	
	dev	test	dev	test	dev	test
0.00	31.64	31.36	39.18	38.24	4.79	4.44
0.25	32.13	32.03	**39.61**	38.38	4.74	4.54
0.50	33.48	34.08	38.60	37.34	**4.62**	4.54
0.75	**38.56**	37.48	38.06	37.60	4.70	4.47
1.00	37.91	**37.93**	38.50	**38.78**	4.79	**4.43**

Experiments. B5 + Mean Reversion F0 + AWGN



Experiments. B5 + Mean Reversion F0 + AWGN

dB	EER		UAR		WER	
	dev	test	dev	test	dev	test
0	38.56	37.48	38.06	37.60	4.70	4.47
5	39.58	40.00	38.91	37.12	4.67	4.49
10	42.46	**43.15**	**39.41**	**38.47**	**4.63**	**4.40**
15	**42.97**	40.36	38.50	37.49	4.66	4.50
30	41.43	39.62	38.41	37.88	4.77	4.64

Conclusion

Conclusion

Condition	Model	EER (avg-dev)	EER (avg-test)	UAR (dev)	UAR (test)	WER (dev)	WER (test)
–	Orig.	5.72	4.59	69.08	71.06	1.80	1.85
>20% EER	B3	28.43	22.04	37.57	38.09	4.29	4.35
>30% EER	B4	32.71	30.26	41.97	42.78	6.15	5.90
>30% EER	B5	34.37	34.34	38.08	38.17	4.73	4.37
>30% EER	B6	23.05	21.14	36.39	36.13	9.69	9.09
>10% EER	1a (NS3)	12.09	10.46	**49.20**	**49.12**	4.97	4.60
>10% EER	1b (B3)	16.88	17.45	**42.76**	**43.21**	**3.81**	**3.83**
>20% EER	2a (B3)	21.47	24.13	**44.67**	**42.78**	4.21	4.29
>20% EER	2b (B3)	20.07	22.85	39.18	37.67	**3.61**	**3.68**

NANYANG TECHNOLOGICAL UNIVERSITY

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938

Conclusion

Condition	Model	EER (avg-dev)	EER (avg-test)	UAR (dev)	UAR (test)	WER (dev)	WER (test)
–	Orig.	5.72	4.59	69.08	71.06	1.80	1.85
>20% EER	B3	28.43	22.04	37.57	38.09	4.29	4.35
>30% EER	B4	32.71	30.26	41.97	42.78	6.15	5.90
>30% EER	B5	34.37	34.34	38.08	38.17	4.73	4.37
>30% EER	B6	23.05	21.14	36.39	36.13	9.69	9.09
>30% EER	3 (B5)	**38.56**	**37.48**	38.06	37.60	4.70	4.47
>40% EER	4 (B5)	**42.46**	**43.15**	39.41	38.47	4.63	4.40

NANYANG TECHNOLOGICAL UNIVERSITY

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938

Key Takeaways

1. NaturalSpeech3 FACodec:
   - Promising results for ER and ASR
   - But there may be leakage of speaker identity in other branches (content/acoustic)

2. Emotion Embeddings:
   - Helps to improve ER performance
   - But leads to speaker identity leakage

3. Cross-Gender anonymization:
   - Improves privacy protection and ER metrics
   - But reduces ASR performance

4. Mean-reversion of F_0 and AWGN:
   - Improves privacy protection while keeping ASR and ER
   -

Thank you

NANYANG TECHNOLOGICAL UNIVERSITY

NORTHWESTERN POLYTECHNICAL UNIVERSITY 1938

References

[1] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” ArXiv, vol. abs/2403.03100, 2024.

[2] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved rvqgan,” Advances in Neural Information Processing Systems, vol. 36, 2024.

[3] J. Kahn, M. Rivi`ere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazar´e, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-Light: A Benchmark for ASR with Limited or No Supervision,” arXiv e-prints, p. arXiv:1912.07875, Dec. 2019

[4] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in Interspeech, 2019, pp. 1526–1530.

[5] https://www.voiceprivacychallenge.org/