Exploring the Importance of F0 Trajectories for Speaker Anonymization Using X-vectors and Neural Waveform Models

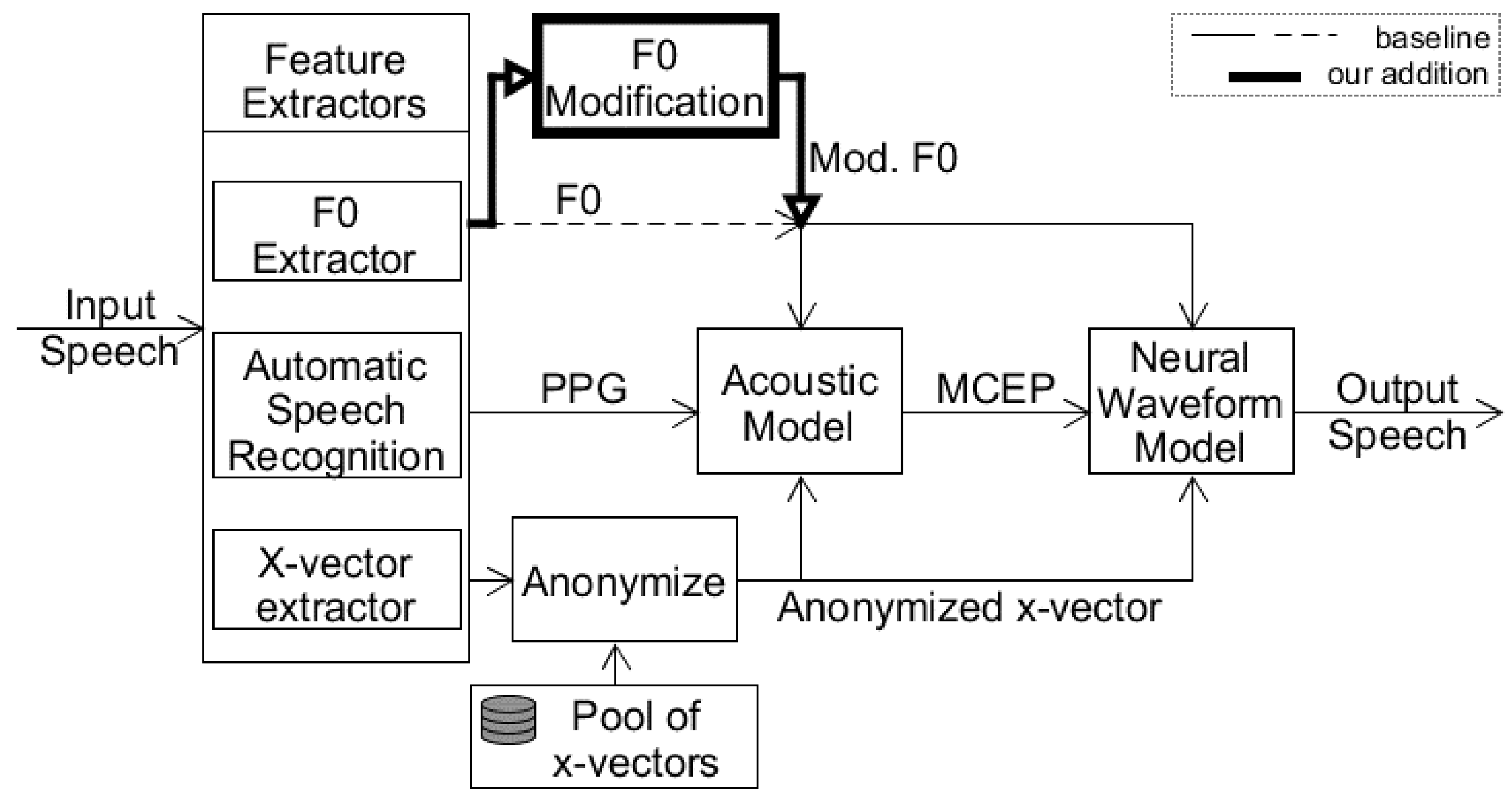
Ünal Ege Gaznepoğlu[1,2] and Nils Peters[1,2]

[1] University of Erlangen-Nuremberg, [2] International Audio Laboratories Erlangen
ege.gaznepoglu@fau.de, nils.peters@audiolabs-erlangen.de

Introduction

- Resynthesizing phoneme posteriorgrams (PPG), the pitch (F0) and modified X-vectors is a common basis for many state-of-the-art voice anonymization systems.
- Works on F0 are scarce, so we developed and evaluated eight low-complexity F0 modifications prior resynthesis, utilizing the VoicePrivacy Challenge 2020 framework.
- Altering F0 can improve equal-error rate by up to 8% with minor word-error rate degradation.

Baseline [1] / Our Contribution

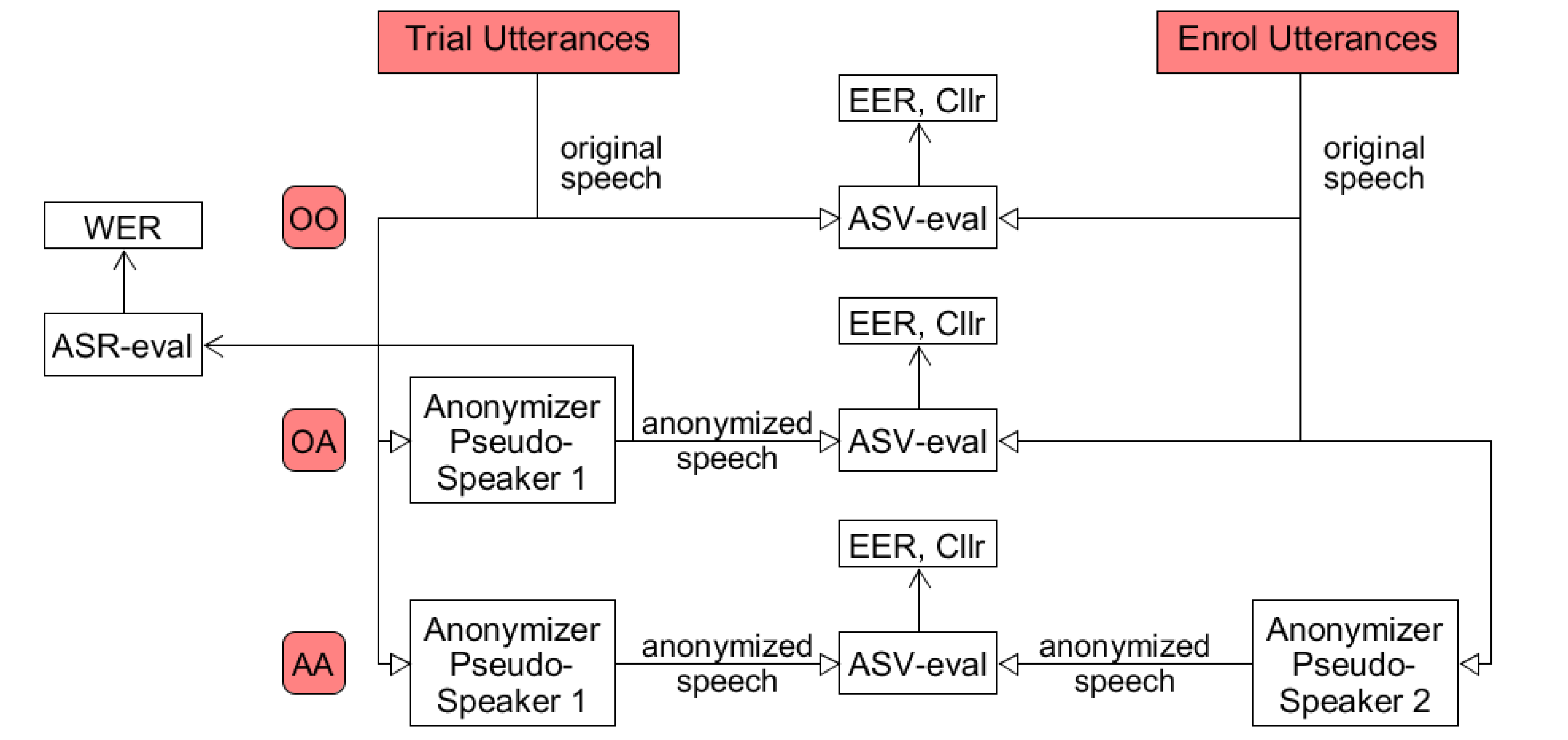


Feature Extractors
F0 Modification
F0 Extractor
Input Speech
Automatic Speech Recognition
X-vector extractor
F0
Mod. F0
PPG
Acoustic Model
MCEP
Neural Waveform Model
Output Speech
Anonymize
Anonymized x-vector
Pool of x-vectors
baseline
our addition

Examples

Scan me!

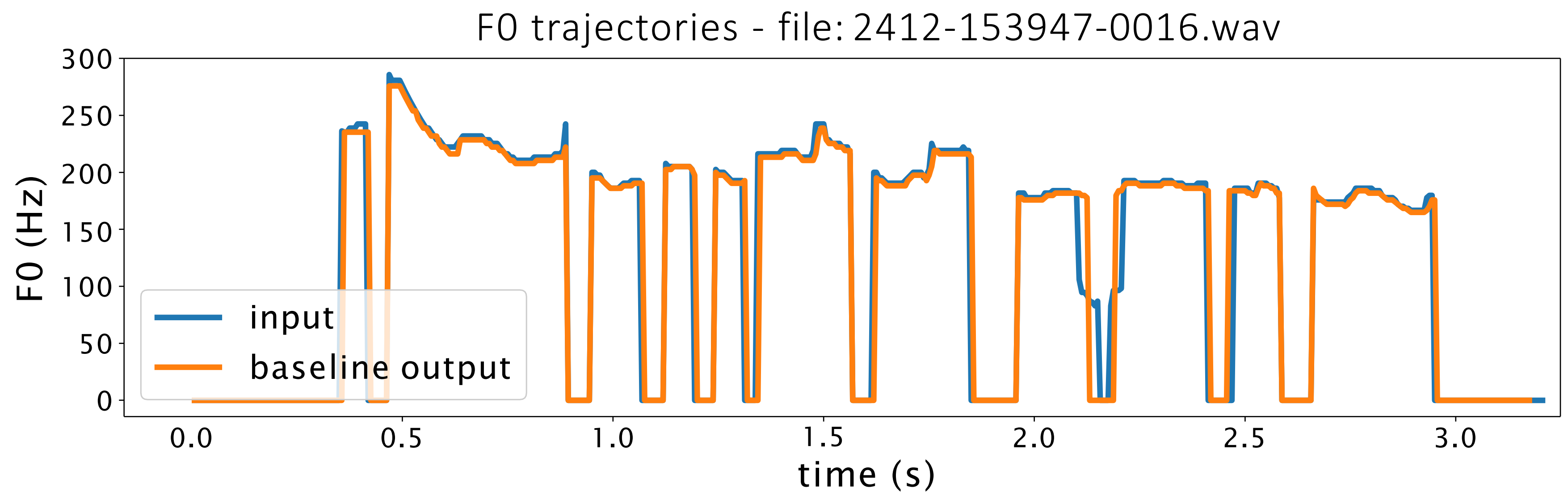
Evaluation [2]



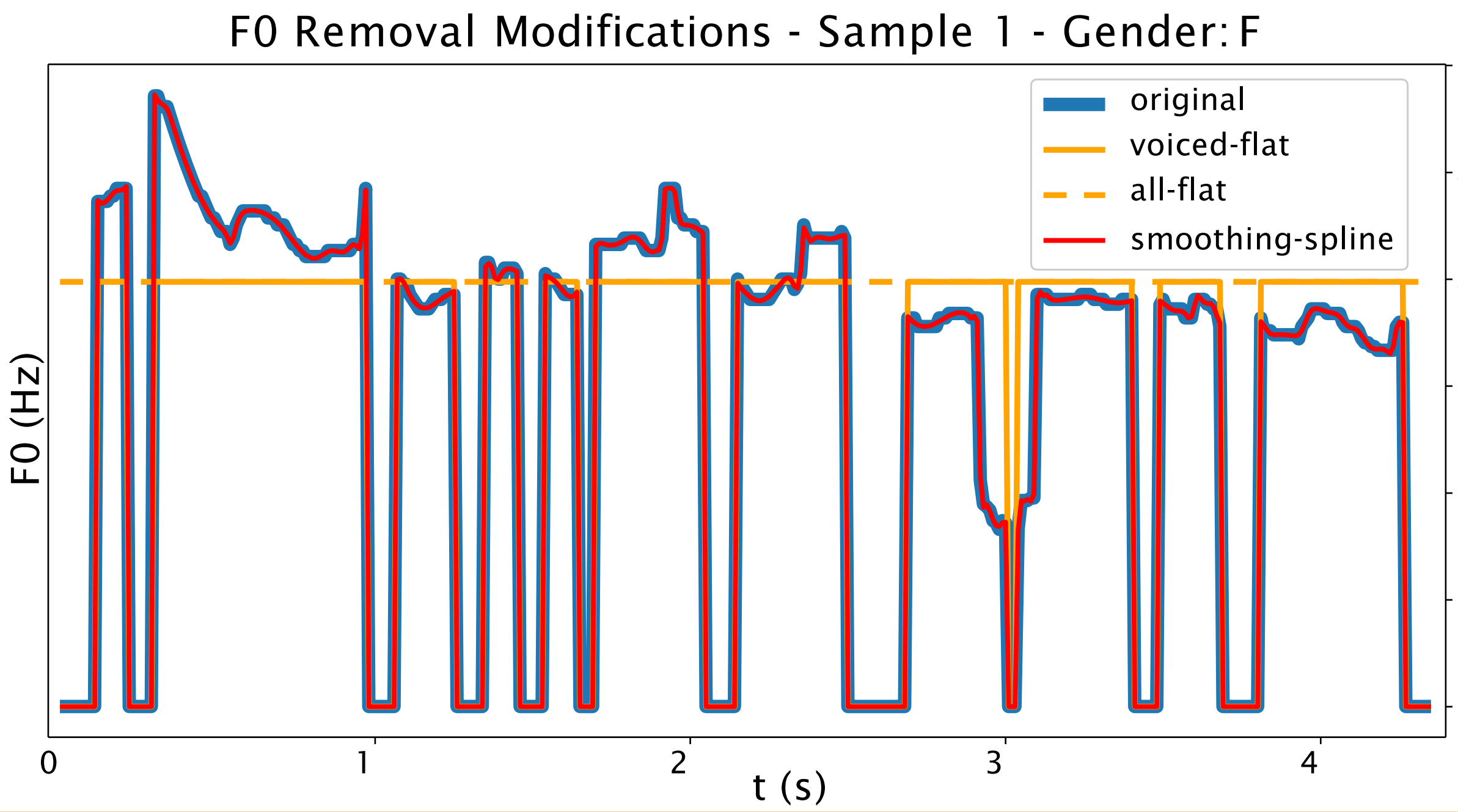
Trial Utterances
Enrol Utterances
original speech
original speech
EER, Cllr
ASV-eval
WER
ASR-eval
OO
EER, Cllr
ASV-eval
OA
Anonymizer Pseudo-Speaker 1
anonymized speech
EER, Cllr
AA
Anonymizer Pseudo-Speaker 1
anonymized speech
ASV-eval
anonymized speech
Anonymizer Pseudo-Speaker 2

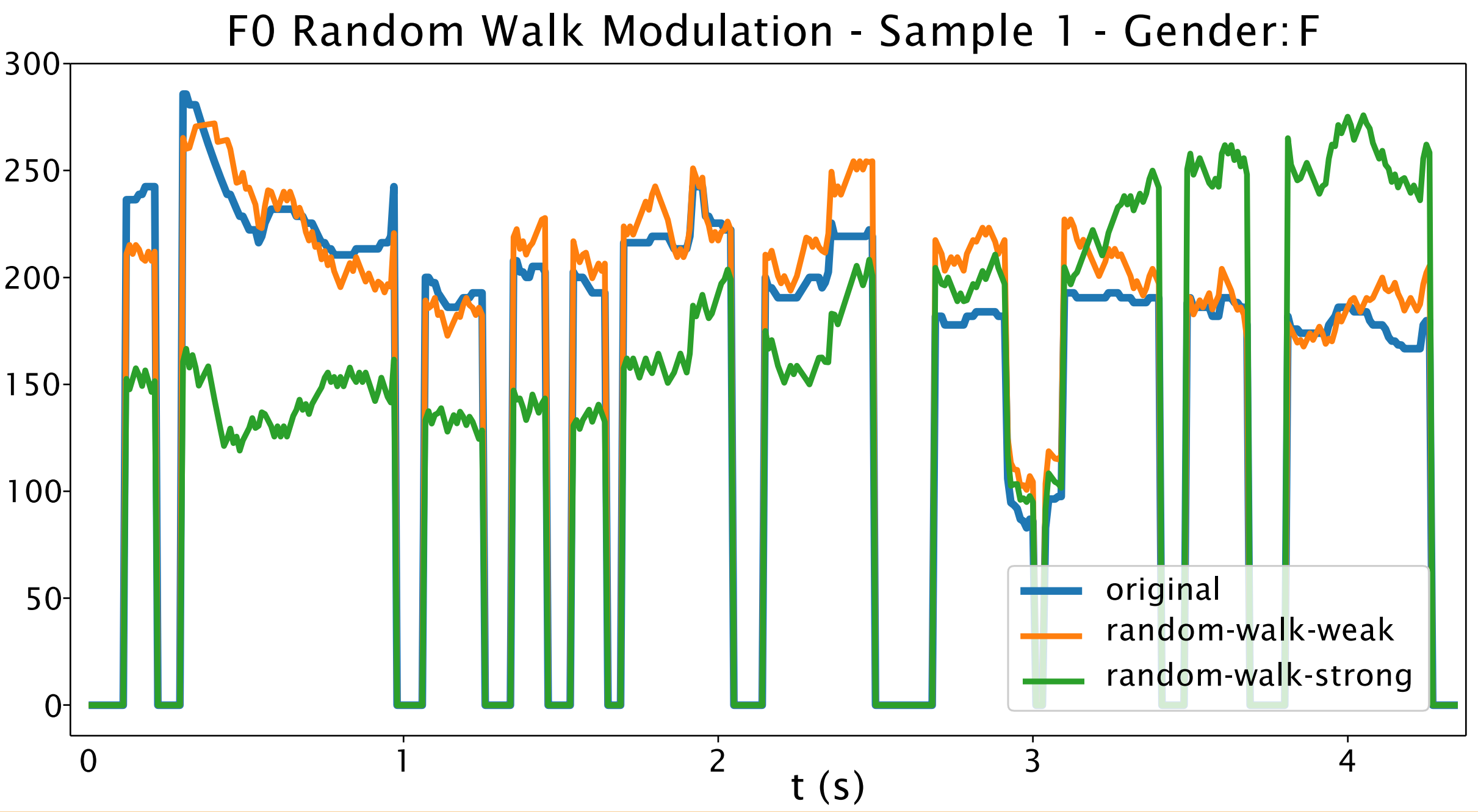
Input / Baseline Output F0

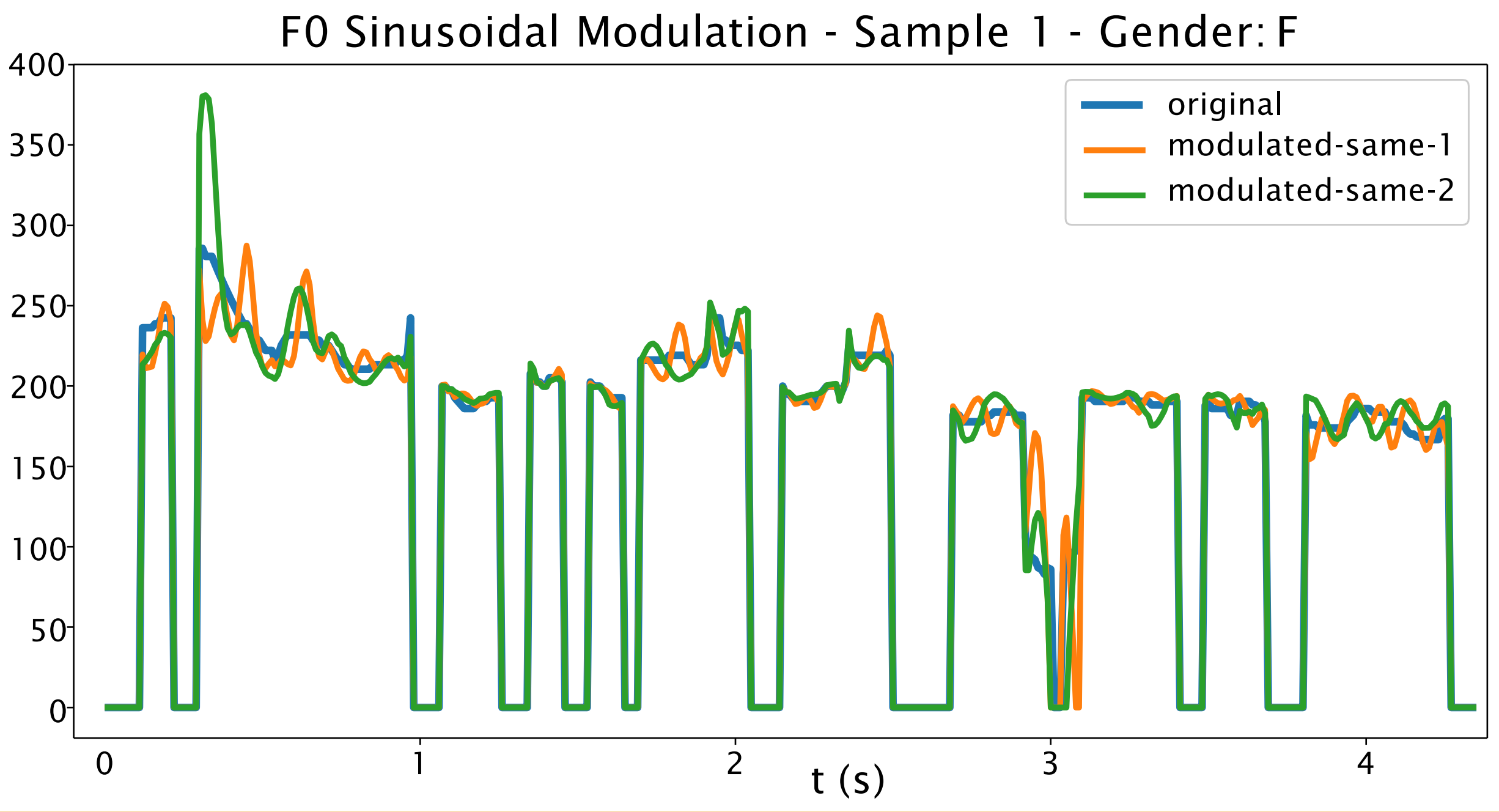
- Baseline output mostly follows input F0 (expected)
- F0 can be used to distinguish speakers [3] – anonymity hazard!



Our Proposed Modifications







ASR and ASV evaluation

	EER (ASV) – Higher better				WER (ASR) – Lower better			
	LibriDev		LibriTest		Libri-		VCTK-	
Method	OA	AA	OA	AA	Dev	Test	Dev	Test
Raw Data	4.95	N/A	4.39	N/A	10.79	12.79	3.83	4.15
Baseline	54.02	35.41	50.32	33.63	15.38	**15.22**	6.32	**6.71**
Voiced-flat	**54.73**	33.24	**51.15**	32.35	15.69	15.48	6.42	6.93
All-flat	54.61	30.98	50.83	30.17	16.22	15.80	6.81	7.25
Smoothing-spline	54.25	35.40	50.23	33.83	**15.34**	15.25	**6.29**	6.75
Modulated-same-1	54.15	35.33	50.14	33.69	15.74	15.57	6.65	6.97
Modulated-same-2	53.69	35.83	50.36	34.19	15.55	15.36	6.57	6.93
Modulated-different	53.69	35.62	50.36	34.24	15.55	15.36	6.57	6.93
Random-walk-weak	54.56	36.56	50.79	34.63	15.54	15.37	6.38	6.89
Random-walk-strong	54.59	**37.21**	50.08	**36.31**	15.96	15.88	6.74	7.12
F0-shift-scale [3]	55.14	36.61	50.78	38.68	15.50	15.29	6.43	6.92
X-vector-gmm-pca [4]	46.75	39.15	45.70	39.45	15.56	15.63	6.75	7.26
X-vector-domain-adv [5]	53.95	35.48	49.69	34.44	15.20	15.16	6.75	6.74

Conclusion

- Random walk noise addition is a viable option
  - Low amplitudes attain mostly similar scores to shift-and-scale [4], with a further irreversibility bonus
  - High amplitudes perform similar to x-vector based techniques [5,6], but with a price of possible intonation changes
- Other modifications have drawbacks
  - Smoothing splines do not alter any of the evaluation metrics
  - Modulation causes audible vibrato effect
  - Flattening decreases ‘AA’ score significantly

Our low-complexity modifications improve anonymization with minor WER. F0 manipulation requires further investigation to unleash its potential

References

[1] F. Fang et. al. “**Speaker Anonymization Using X-vector and Neural Waveform Models**” 10th ISCA Speech Synthesis Workshop, 2019
[2] N. Tomashenko et. al. “VoicePrivacy 2020 Challenge Evaluation Plan.”, https://www.voiceprivacychallenge.org/docs/VoicePrivacy2020EvalPlanv13.pdf
[3] P. Labutin et. al. “**Speaker identification based on the statistical analysis of f0**,” IAFPA, 2007
[4] P. Champion et. al. “**A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender**,” arXiv:2101.08478, 2021
[5] H. Turner et. al. “**Speaker Anonymization with Distribution-Preserving X-Vector Generation for the VoicePrivacy Challenge 2020**,” arXiv:2010.13457, 2020
[6] F. M. Espinoza-Cuadros et. al. “**Speaker De-identification System using Autoencoders and Adversarial Training**”, arXiv:2011.04696, 2020