B4 の兵藤です。
2022 年 12 月 17 日に名古屋大学で行われた“JST CREST「共創型音メディア機能拡張」中間シンポジウム 2022”に参加しました。
講演やポスター発表、他の参加者の方との議論を通じて様々な知見を得ることができました。また、自分の研究で使用しているモデルの開発者の方と意見交換をしたり、最新の音声変換技術のデモを体験させていただくことができ、とても貴重な体験をすることができました。

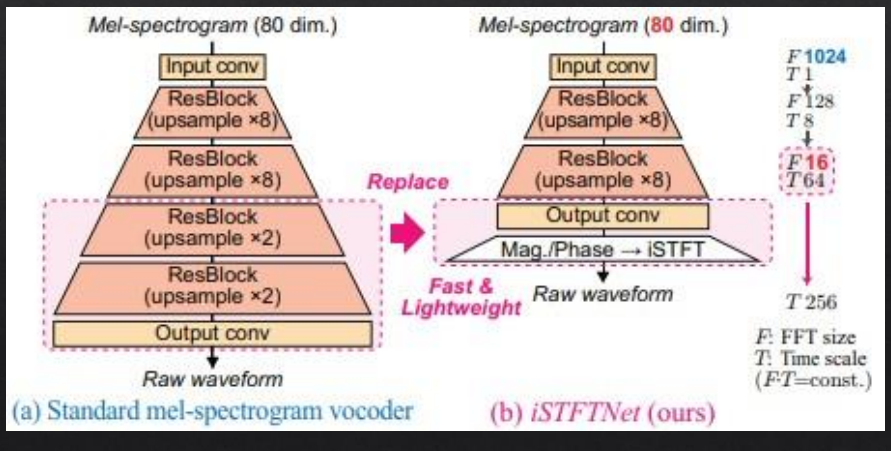
シンポジウムで発表されていた研究から、特に参考になったものをスライド形式で紹介します。

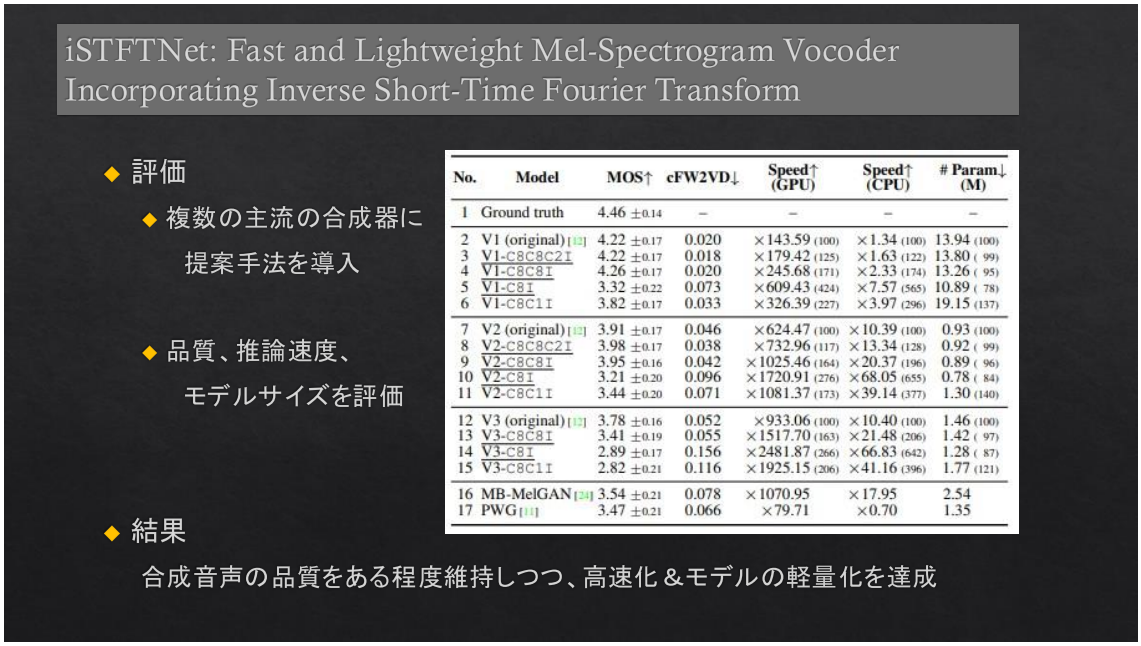
iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform

- 背景：DNNを用いた音声合成手法の流行
  - 高品質で、肉声に近い音声を合成可能
  - 信号処理ベースの手法と比べ、合成速度が遅い
  - モデルサイズが大きく、小型計算機上での合成が困難
- 提案手法
  - DNNの出力側の層の一部を信号処理で置換

iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform

- 提案手法





No.	Model	MOS↑	cFW2VD↓	Speed↑ (GPU)	Speed↑ (CPU)	# Param↓ (M)
1	Ground truth	4.46 ±0.14	–	–	–	–
2	V1 (original) [12]	4.22 ±0.17	0.020	×143.59 (100)	×1.34 (100)	13.94 (100)
3	V1-C8C8C2I	4.22 ±0.17	0.018	×179.42 (125)	×1.63 (122)	13.80 (99)
4	V1-C8C8I	4.26 ±0.17	0.020	×245.68 (171)	×2.33 (174)	13.26 (95)
5	V1-C8I	3.32 ±0.22	0.073	×609.43 (424)	×7.57 (565)	10.89 (78)
6	V1-C8C1I	3.82 ±0.17	0.033	×326.39 (227)	×3.97 (296)	19.15 (137)
7	V2 (original) [12]	3.91 ±0.17	0.046	×624.47 (100)	×10.39 (100)	0.93 (100)
8	V2-C8C8C2I	3.98 ±0.17	0.038	×732.96 (117)	×13.34 (128)	0.92 (99)
9	V2-C8C8I	3.95 ±0.16	0.042	×1025.46 (164)	×20.37 (196)	0.89 (96)
10	V2-C8I	3.21 ±0.20	0.096	×1720.91 (276)	×68.05 (655)	0.78 (84)
11	V2-C8C1I	3.44 ±0.20	0.071	×1081.37 (173)	×39.14 (377)	1.30 (140)
12	V3 (original) [12]	3.78 ±0.16	0.052	×933.06 (100)	×10.40 (100)	1.46 (100)
13	V3-C8C8I	3.41 ±0.19	0.055	×1517.70 (163)	×21.48 (206)	1.42 (97)
14	V3-C8I	2.89 ±0.17	0.156	×2481.87 (266)	×66.83 (642)	1.28 (87)
15	V3-C8C1I	2.82 ±0.21	0.116	×1925.15 (206)	×41.16 (396)	1.77 (121)
16	MB-MelGAN [24]	3.54 ±0.21	0.078	×1070.95	×17.95	2.54
17	PWG [11]	3.47 ±0.21	0.066	×79.71	×0.70	1.35

シンポジウムを企画いただいた方々、またシンポジウムへの参加を支援してくださった方々にこの場を借りて感謝申し上げます。ありがとうございました。