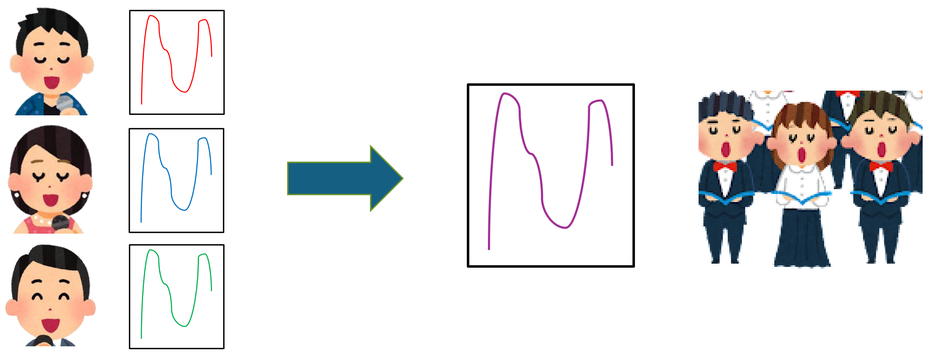
歌声合成を用いた斉唱の自然性に関する要因調査

3-P-31

西澤佳飛[1]，山本龍一[1,2]，Wen-Chin Huang[1]，戸田智基[1]([1]名古屋大学，[2] LINEヤフー)

1. 研究背景・目的

● **斉唱**：複数人が同一のパートを歌唱



● **斉唱情報処理実現への課題**

斉唱がどのような音響特徴を持つのか不明

● **仮説**

発声タイミングや音高、スペクトル特徴の**揺らぎ**が本質的に重要

● **本研究**

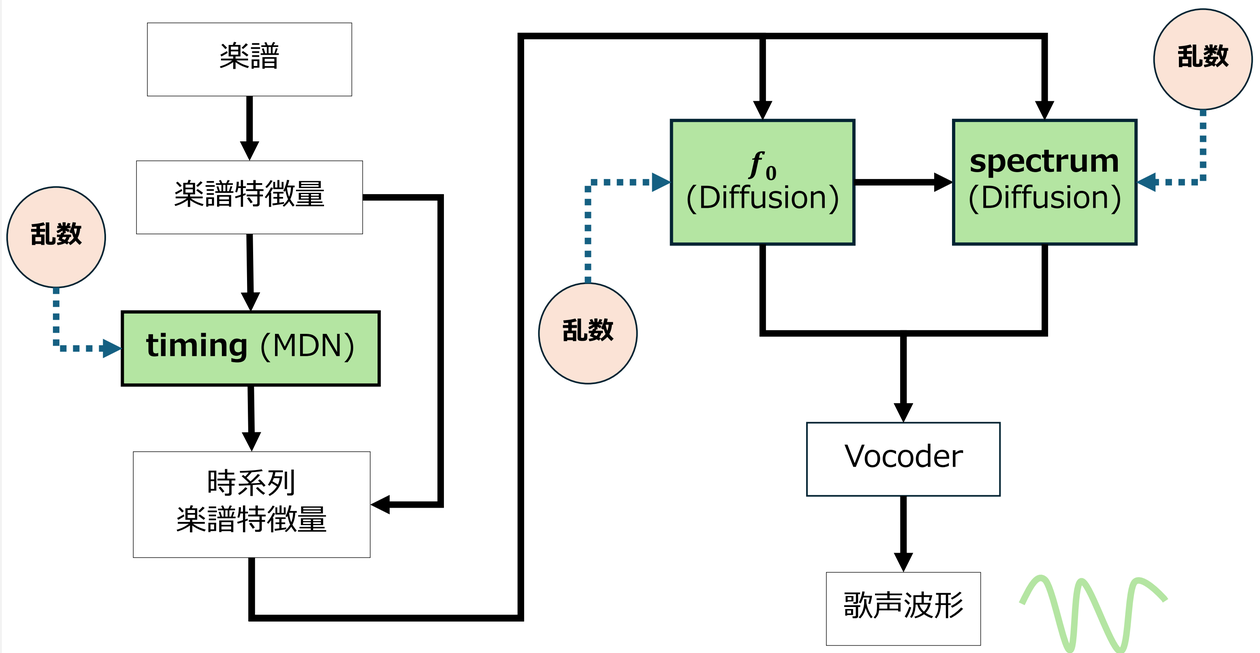
歌声合成器を用いて各種揺らぎを制御した疑似斉唱音声を生成し，
それらを知覚的に評価することで斉唱の自然性に関わる要素を調査

2. 歌声生成器

● NNSVS [1]を使用 (Yamamoto+2023)

● 生成される歌声の**パラメータ**を乱数で制御可能にした

timing	発声のタイミング・音の長さ
f_0	音高
spectrum	スペクトル特徴



● timing (MDN:混合密度ネットワーク)

■既存のNNSVS

◆ 楽譜特徴量に対するタイミング値を混合正規分布から最尤推定

■ **本研究のNNSVS**

◆ 混合正規分布からタイミング値をランダムサンプリング

● f_0・spectrum (Diffusion:拡散確率モデル)

■既存のNNSVS

◆ 固定の雑音系列(シード固定)を各特徴系列に変換

■ **本研究のNNSVS**

◆ 都度ランダム生成される雑音系列を各特徴量系列に変換

3. 評価実験

● データセット

①「波音リツ」歌声データベースVer.2[2] (110曲，4時間21分8秒)
訓練/検証/評価=100/5/5としてNNSVSを学習

② JVS-MuSiC[3](日本語話者100人，共通:1曲，個別:100曲)

● 実験1　揺らぎの有無による斉唱の自然性への影響

■ NNSVSで歌声を16回生成して混合することで16人による疑似斉唱歌声を作成

■ データセット①から20フレーズを使用

● 結果

timing	f_0	spectrum	MOS↑
			1.56±0.15
✓			**3.94±0.11
	✓		**3.80±0.11
		✓	1.66±0.16
✓	✓		*4.10±0.10
	✓	✓	*4.04±0.11
✓		✓	*4.06±0.11
✓	✓	✓	***4.14±0.11**

*，**は同じ記号同士で有意差が見られないことを表す($P < 0.05$)

■ 単一の揺らぎ
timingとf_0の揺らぎが斉唱の自然性を上げる

■ 複数の揺らぎ
単一の揺らぎよりも斉唱の自然性のスコアが高い
揺らぎの組み合わせによる自然性の差は見られない

● 実験2　歌唱人数による斉唱への影響の調査

■ NNSVSですべてのパラメータに揺らぎを与えて歌声を作成、混合

■ 合成歌唱音声：データセット①から15フレーズを使用

■ 実歌唱音声：データセット②からfemale_highの「かたつむり」16フレーズを使用

● 結果

合成歌唱音声

正解の人数＼推測した人数	2人	4人	8人	16人	32人	64人	128人
2人	0.908	0.083	0.008	0.000	0.000	0.000	0.000
4人	0.388	0.542	0.067	0.004	0.000	0.000	0.000
8人	0.088	0.496	0.267	0.133	0.017	0.000	0.000
16人	0.033	0.200	0.383	0.258	0.121	0.004	0.000
32人	0.025	0.117	0.329	0.313	0.171	0.029	0.017
64人	0.008	0.067	0.313	0.325	0.179	0.058	0.050
128人	0.004	0.092	0.250	0.208	0.275	0.125	0.046

実歌唱音声

正解の人数＼推測した人数	2人	4人	8人	16人
2人	0.813	0.125	0.063	0.000
4人	0.500	0.313	0.125	0.063
8人	0.188	0.563	0.188	0.063
16人	0.000	0.438	0.313	0.250

正解率

揺らぎを加えることで、実歌唱音声と同じ複数人による
歌唱感を出すことができた

4. 展望

● 実斉唱音声との比較

■ 実斉唱音声がもつ音響特徴の特定

■ 実斉唱音声がもつ音響特徴との比較

参考文献

1. R. Yamamoto, R. Yoneyama and T. Toda, "NNSVS: A Neural Network-Based Singing Voice Synthesis Toolkit," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5
2. https://www.canon-voice.com/voicebanks/
3. https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs_music