差分メルケプストラムを用いた声質変換による喉締め歌唱音声改善方式の検討

Study of choking singing voice improvement method based on voice-conversion using mel-cepstral differential

植田 遥人

Haruto Ueda

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本稿では喉締め歌唱しかできない歌唱者の歌唱音声を話者性を維持したまま腹式歌唱音声らしく変換することを目的とする．方式として，同一歌唱者の腹式歌唱音声から喉締め歌唱音声の差分メルケプストラムに，変換したい任意の歌唱者の喉締め歌唱音声メルケプストラムを足し合わせることによる変換方式を検討する．

1 はじめに

複数人で行う合唱では声質をそろえて歌唱することが重要である．声質の個人差の大きさは歌唱方法によって異なり，声帯に負担がかかる歌唱方法である喉締め歌唱 [1] は，声質の個人差が大きい歌唱方法である．反対に，声帯に負担がかかっていない歌唱方法である腹式歌唱は，声質の個人差が小さく，合唱に適した歌唱方法である．しかしながら，腹式歌唱をすることができない人にとって，腹式歌唱を習得するのは非常に困難である．本研究は，複数人の歌唱を腹式歌唱に変換し，合成することによって，個人で収録した音声から上質な合唱音声を生成することを目的とする．

歌声の声質変換として，音色変化を制御する声質変換の研究がある [2]．本報告では，声質変換を応用した新たな歌唱変換方法について検討を行う．提案方式として，複式歌唱と喉締め歌唱との差分メルケプストラムを用いて変換する方式を２つ検討する．差分メルケプストラムは，変換したい任意歌唱話者の喉締め歌唱メルケプストラムに足し合わせる．方式１は変換したい任意歌唱者の喉締め歌唱メルケプストラムから差分メルケプストラムを推定する方式である．方式２は変換したい任意歌唱者の喉締め歌唱を，差分メルケプストラム推定の学習に用いた歌唱者の喉締め歌唱メルケプストラムに変換した後に，学習時に用いた歌唱者の喉締め歌唱メルケプストラムから差分メルケプストラムを推定する方式である．評価実験では２つの方式について，メルケプストラム歪みによる客観評価実験を行った．さらに，喉締め歌唱改善度，話者性という観点から主観評価実験を行った．

2 提案方式

本方式は差分スペクトル補正 [3] をベースにした方式である．差分スペクトル補正は目標話者音声のメルケプストラムと入力話者音声のメルケプストラムとの差分を用いて目標話者音声に変換する方式である．本方式では，腹式歌唱から喉締め歌唱の差分メルケプストラムを用いることで任意歌唱者の喉締め歌唱を腹式歌唱に変換することを目指す．

提案方式では，音声の分析・合成を行うために WORLD[4] を用いている．また，学習には Long Short-Term Memory(LSTM)[5] を用いる．

2.1 腹式歌唱変換法

本方式では，式 1 で歌唱メルケプストラムがあらわされると仮定している．

$$c(t;spk,sty) = \tilde{c}(t;spk) + \tilde{c}(t;sty) + \tilde{c}_{base}(t) \quad (1)$$

$c(t;spk,sty)$ は spk の話者性をもつ，sty の歌唱方法での歌唱メルケプストラムである．$\tilde{c}(t;spk)$ は歌唱者の話者性をもつメルケプストラムであり，$\tilde{c}(t;sty)$ は歌唱者の喉締め歌唱か腹式歌唱かの歌唱方法を表すメルケプストラムである．$\tilde{c}_{base}(t)$ はベース音声のメルケプストラムを表している．

式 1 から，歌唱者 A の腹式歌唱から喉締め歌唱の差分メルケプストラム $c_{diff}(t)$ は次のように表される．

$$\begin{aligned} c_{diff}(t) &= c(t;A,\text{腹式}) - c(t;A,\text{喉締め}) \\ &= \tilde{c}(t;\text{腹式}) - \tilde{c}(t;\text{喉締め}) \end{aligned} \quad (2)$$

ここで，歌唱者 B の喉締め歌唱メルケプストラムを差分メルケプストラムに足し合わせることで，式 3 のように腹式歌唱メルケプストラムに変換できる．

$$c_{diff}(t) + c(t;B,\text{喉締め}) = c(t;B,\text{腹式}) \quad (3)$$

2.2 方式１

方式 1 の概要図を図 1 に示す．方式 1 は式 1 の $\tilde{c}(t;sty)$ が話者によって影響されないことを仮定したモデルであり，差分メルケプストラムにも話者による影響が包含されない．そのため，喉締め to 差分メルケプストラム変換モデルが任意歌唱者の喉締め歌唱メルケプストラムから差分メルケプストラムを推定できるとして提案する方式である．

2.3 方式２

方式 2 の概要図を図 2 に示す．方式 2 は式 1 の $\tilde{c}(t;sty)$ が話者によって影響されると仮定したモデ

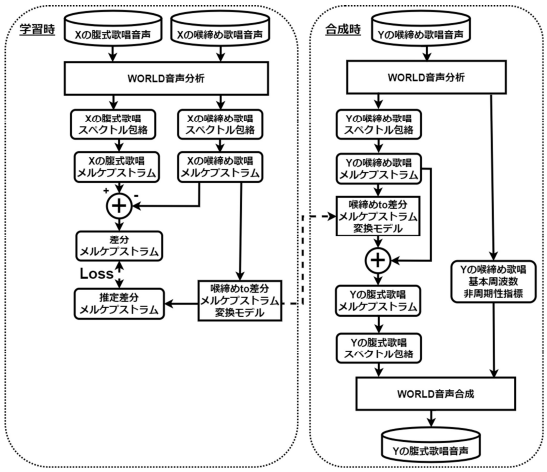

図 1: 方式 1 の概要図．この図ではモデルの学習に用いた歌唱者を X とし，変換を行う任意の歌唱者を Y としている．

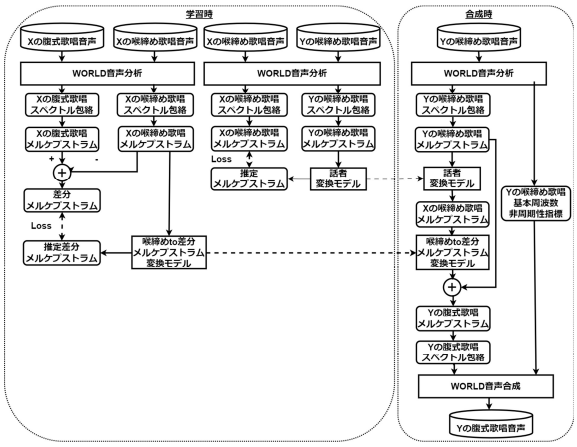

図 2: 方式 2 の概要図．この図では喉締め to 差分メルケプストラム変換モデルの学習に用いた歌唱者を X とし，変換を行う任意の歌唱者を Y としている

ルであり，差分メルケプストラムにも話者による影響が包含される．そのため，喉締め to 差分メルケプストラム変換モデルが学習した歌唱者の喉締め歌唱メルケプストラムからでしか差分メルケプストラムを推定できないとして提案する方式である．

3　歌唱音声データベース

評価実験には表 1 に示す歌唱音声データベースを用いる．喉締め歌唱と腹式歌唱は対となるように，同じ曲，同じ歌唱母音の歌唱をパラレルコーパスとして収録した．収録では歌唱の音高が楽譜データと大きくずれることがないように，ヘッドホンから自身の歌唱と歌唱する楽譜の MIDI 音声を聴いて収録を行った．前半に収録した曲と後半に収録した曲で歌唱の質が変化しないように収録前に 1 分間の曲を歌うことで練習を行い，また，連続で収録することは疲労により歌唱の質が低下すると考えられるため，5 曲ごとに 30 分の休憩をとって収録を行った．収録後，変換時の音量による

表 1: 歌唱音声データベース

サンプリング周波数	96 kHz
量子化ビット	24 bitPCM
歌唱条件	童謡曲 １曲あたり約１分
歌唱方法	歌詞の代わりに 単一母音で歌唱 /a/，/i/，/u/，/e/，/o/
歌唱者 A の 歌唱データコーパス	喉締め歌唱 150 曲 （各母音 30 曲） 腹式歌唱 150 曲 （各母音 30 曲） 約 120 分
歌唱者 B の 歌唱データコーパス	喉締め歌唱 25 曲 （各母音 5 曲） 腹式歌唱 25 曲 （各母音 5 曲） 約 20 分

影響を小さくするために，音声信号の二乗平均値が等しくなるように正規化した．

4　評価実験

評価実験では方式 1 と方式 2 について，メルケプストラム歪みを用いた客観評価実験を行った．さらに，喉締め歌唱音声の改善度と話者性という観点から主観評価実験を行った．

実験では，喉締め to 差分メルケプストラム変換モデルの学習に用いる歌唱者 X の歌唱として表 1 の歌唱者 A の歌唱を用いた．また，喉締め to 差分メルケプストラム変換モデルの学習に用いない歌唱者 Y の歌唱として表 1 の歌唱者 B の歌唱を用いた．

4.1　実験条件

WORLD を用いた音声分析を行うためのパラメータと，LSTM モデルを学習するためのパラメータを表 2 に示し，モデルを学習するためのデータを表 3 に示す．検証データはモデルの学習には用いず，モデルの過学習を防ぐために用いる．

評価実験では，検証データを含めた学習を行っていない喉締め歌唱，腹式歌唱各 25 曲（各母音 5 曲）のテストデータで評価した．

4.2　客観評価実験

客観評価の指標としてメルケプストラム歪みを用いる．メルケプストラム歪みを用いることで，メルケプストラム間の距離を計算する．本実験では，評価する音声のメルケプストラムと腹式歌唱メルケプストラムとのメルケプストラム歪みを算出した．

歌唱者 A の喉締め歌唱音声から歌唱者 A の腹式歌唱

表 2: 実験条件

音声分析パラメータ	
メルケプストラム	49 次元
fft 長	8192 点 （85.3 ms）
フレームシフト	5 ms
LSTM モデルパラメータ	
バッチサイズ	5
中間層	3
中間層ユニット数	256
学習率	0.001
最適化手法	Adam[6]

表 3: モデル学習データ

話者変換モデル	
学習データ	歌唱者 A 喉締め歌唱 20 曲 歌唱者 B 喉締め歌唱 20 曲
検証データ	歌唱者 A 喉締め歌唱 5 曲 歌唱者 B 喉締め歌唱 5 曲
喉締め to 差分メルケプストラム変換モデル	
学習データ	歌唱者 A 喉締め歌唱 125 曲 歌唱者 A 腹式歌唱 125 曲
検証データ	歌唱者 A 喉締め歌唱 10 曲 歌唱者 A 腹式歌唱 10 曲

音声への変換後音声のメルケプストラム歪みを表 4 に示し，歌唱者 B の喉締め歌唱音声から歌唱者 B の腹式歌唱音声への変換後音声のメルケプストラム歪みを表 5 に示す．

歌唱者 A では，全母音で変換前に比べ，変換後のメルケプストラム歪みが大きく減少した．歌唱者 B では，方式 1 の平均メルケプストラム歪みが変換前に比べ変換後が減少しており，変換後歌唱メルケプストラムが腹式歌唱メルケプストラムに近づいていることを表している．しかしながら，/u/では変換前よりもメルケプストラム歪みが増加した．方式 2 の平均メルケプストラム歪みは，変換前に比べ変換後が増加している．方式 1 の歌唱者 A の変換後歌唱メルケプストラムが減少していることから，話者変換モデルの推定制度が低いと考えられる．

4.3 主観評価実験

主観評価実験では方式 1 の評価を行った．喉締め歌唱音声の改善度という観点から，歌唱者 A と歌唱者 B のそれぞれについて，変換後音声を各母音ごとに評価した．さらに，話者性の観点から，歌唱者 B の変換後音声について各母音ごとに評価を行った．

表 4: 歌唱者 A の変換後歌唱と腹式歌唱とのメルケプストラム歪み

母音	変換前 [dB]	変換後 [dB]
		方式 1
/a/	7.424	4.309
/i/	9.541	5.587
/u/	8.448	5.674
/e/	7.231	4.841
/o/	8.448	4.359
平均	7.847	4.954

表 5: 歌唱者 B の変換後歌唱と腹式歌唱とのメルケプストラム歪み

母音	変換前 [dB]	変換後 [dB]	
		方式 1	方式 2
/a/	7.799	7.229	7.603
/i/	8.483	7.021	8.589
/u/	8.136	8.374	9.773
/e/	7.060	6.739	7.885
/o/	8.061	7.247	8.441
平均	7.908	7.322	8.458

4.3.1 実験方法

喉締め歌唱改善度の評価は，ある歌唱者について，喉締め歌唱音声，腹式歌唱音声，変換後音声の 3 つによる ABX テストにより評価した．A，B には喉締め歌唱音声と腹式歌唱音声をランダムで再生し，X として喉締め歌唱音声，腹式歌唱音声，変換後音声について評価した．実験結果では，X が腹式歌唱音声と判断された割合によって評価した．

話者性の評価は，学習した歌唱者 A の腹式歌唱音声，歌唱者 B の腹式歌唱音声，歌唱者 B の変換後音声の 3 つによる ABX テストにより評価した．A，B には歌唱者 A の腹式歌唱音声と歌唱者 B の腹式歌唱音声をランダムで再生し，X として歌唱者の腹式歌唱音声，学習していない歌唱者の腹式歌唱音声，学習していない歌唱者の変換後音声について評価した．実験結果では，X が歌唱者 B の腹式歌唱音声と判断された割合によって評価した．

実験参加者は 7 名であり，1 名につき 1 つの評価実験で 30 個の音声サンプルについて評価した．いづれの ABX テストにおいても，実験参加者は X の音声が 1 つ目の音声の声質と 2 つ目の声質のどちらに近いかという 2 択のテストを行った．

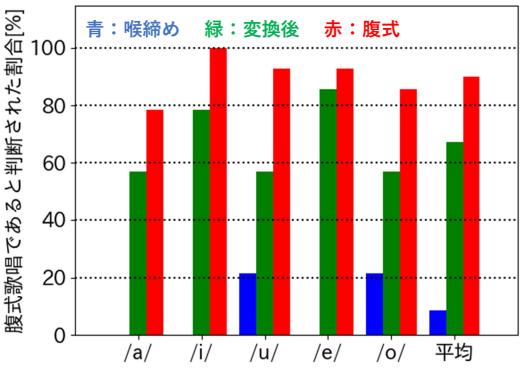

図 3: 歌唱者 A の喉締め歌唱改善度に関する主観評価実験結果

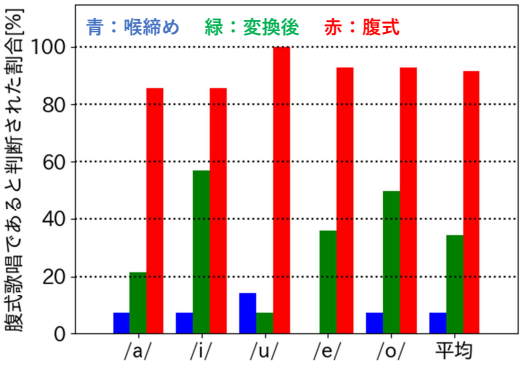

図 4: 歌唱者 B の喉締め歌唱改善度に関する主観評価実験結果

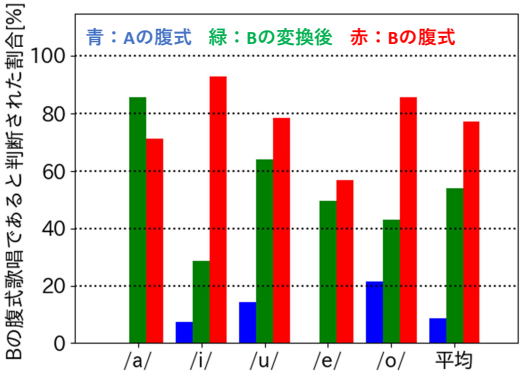

図 5: 歌唱者 B の話者性に関する主観評価実験結果

4.3.2　実験結果

喉締め歌唱音声改善度の主観評価実験結果を図 3，4 に示す．歌唱者 A の変換後歌唱の母音平均は腹式歌唱の母音平均に近く，喉締め歌唱が改善されていると言える．一方歌唱者 B の変換後歌唱の母音平均は半分程度の割合で腹式歌唱と判断されており，歌唱者 A と比べると腹式歌唱への変換性能は低下している．これは，歌唱方法は歌唱者による影響があることを示している．

話者性における主観評価実験結果を図 5 に示す．B の変換後歌唱の母音平均は B の腹式歌唱の母音平均に近く，話者性を維持できたと言える．しかしながら，/i/，/o/では変換後歌唱が学習に用いた歌唱者の結果に近く，母音による影響を考慮する必要がある．

5　まとめ

本報告では，差分メルケプストラムを用いて喉締め歌唱音声を改善する方式について提案した．客観評価実験では，方式 1 の変換後メルケプストラムが変換前の喉締め歌唱メルケプストラムに比べ，腹式歌唱メルケプストラムとのメルケプストラム歪みが減少していることを示した．一方，方式 2 の変換後歌唱メルケプストラムは腹式歌唱メルケプストラムとのメルケプストラム歪みが変換前に比べ上昇した．主観評価実験では，方式 1 について喉締め歌唱改善度と話者性について評価した．喉締め歌唱改善度では，学習に用いていない歌唱者の変換後歌唱は改善されているとは言えない結果であった．話者性は維持できていたが，母音による影響を考慮する必要がある．

今後の課題として，方式 2 の話者変換モデルの推定精度を上げるために，LSTM 以外を学習に用いた方式について検討したい．

参考文献

[1] 平山ら “ポピュラー歌唱における高音域の声区と発声状態の判別手法,”. 情処研報 vol. 2012-MUS-94, no. 16, pp. 1–6, 2012.
[2] 金井ら “複合正弦波モデルと動的計画法を用いた喉頭挙上音声への声質変換システムの提案,”. 情処研報 vol. 2020-SLP-132, no. 17, pp. 1–6, 2020.
[3] K. Kobayashi, *et al.*, “Statistical singing voice conversion with direct waveform modification based on the spectrum differential,” Proc. Interspeech 2014 Association, pp. 2514–2518, 2014.
[4] M. Morise, *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” Proc. of IEICE Trans. on Inf. and Sys., vol. 99, no. 7, pp. 1877–1884, 2016
[5] S. Hochreiter, *et al.*, “Long short-term memory,” Proc. of Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[6] Kingma and Ba, “Adam: A method for stochastic optimization,” Proc. of arXiv preprint arXiv:1412.6980, pp. 1–15, 2014