深層学習を用いた歌声音声の帯域強調の検討 *

☆北村毅 (神戸大), 足立優司 (メック株式会社),
田井清登 (メック株式会社), 滝口哲也 (神戸大)

1　はじめに

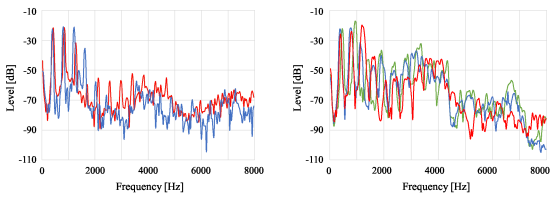
本研究では，歌声の音声を対象として，ある人物の歌声が持つ音声の帯域成分を別の人物の歌声へ移動，もしくは強調するシステムを提案する．歌唱は，カラオケなどを通じて多くの人々が娯楽行為として行なっているが，近年ではプロかアマチュアかを問わず，CD の販売やインターネットへの動画投稿を通じて，表現の場を増やしている．ある人物の歌声は，ピッチやテンポの正確さ，強弱や感情などの表現力に加えて発声方法などにより習熟度を分類することができる．本稿では，プロのオペラ歌手の発声方法による歌声音声とアマチュアの歌声音声を周波数スケールで比較し，プロが持つ強い周波数帯域のエネルギーをアマチュアの音声に自動的に付与，もしくは強調する手法を示す．

コンサートなどで音楽を演奏する場合とは異なり，CD を作成したりインターネットなどに投稿する場合は，ボーカルや楽器などのパートごとに収録を行い，各収録信号に対して修正処理や加工を行なった後，時間同期を取ることで一つの波形を得る場合がある．ボーカルの音声を処理する際，現在の商用ソフトやツールでは基本周波数や振幅に修正を行う場合や，周波数スケールで強調処理を行うことでエフェクトをかける場合がある．しかし，これらの処理はユーザが手動で修正する必要があるため，多くの手間がかかる．そのため，歌声信号の修正や加工を自動化する研究が広く行われている．Robel [1] らは，スペクトル包絡の平滑化を用いることで，ビブラートを除去する手法を提案した．Sangeon [2] らは，ある人物の歌声のテンポ，ピッチと振幅を，プロの歌声音声を用いて修正する手法を示した．これは，ターゲットの歌声長にあうように DTW(Dynamic Time Warping) を用いてアライメントを取り，その後 PSOLA を用いてピッチなどの修正を行なっている．

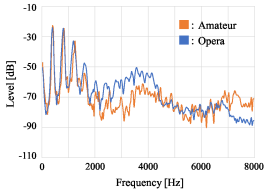
本研究では，プロのオペラ歌手が持つ声の艶となる周波数帯域のエネルギー成分を，アマチュアの歌声音声に付与することで，より艶があり通る声への変換を目指す．これにより，変換した音声を歌唱の訓練に用いる，もしくは歌声の表現の変換ソフトとして収録した歌声をより習熟度が高い人物が歌うような歌声に変換する用途が考えられる．また，ニューラルネットワークを用いることで，オペラ歌手が歌唱していないオープンな音声に対しても変換が可能であることを示す．

2　オペラ歌手及びアマチュアの歌唱音声

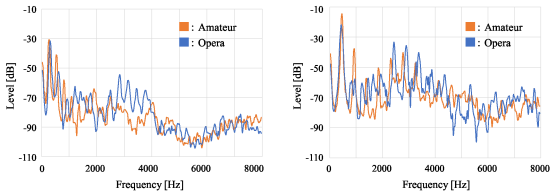
本研究では，女性のプロのオペラ歌手と，女性でオペラを学んでいないアマチュアの歌唱音声を用いる．Fig. 1 に，オペラ歌手とアマチュアが歌唱中に発声した'a'，'u' 及び'i' の音素のスペクトラムを示す．



(a) 'a' of two amateur singers.　(b) 'a' of three professional opera singers.



(c) 'a' of amateur and opera singers.



(d) 'u' of amateur and opera singers.　(e) 'i' of amateur and opera singers.

Fig. 1　Sample Spectrums.

Fig. 1(a) はアマチュア 2 人が，Fig. 1(b) はオペラ歌手 3 名が歌唱中に発話した'a' 音素のスペクトラムをそれぞれ示している．また，Fig. 1(c)(d)(e) はそれぞれ'a'，'u' 及び'i' の音素について，2 名のアマチュアの平均スペクトラムと 3 名のオペラ歌手の平均スペクトラムを示している．スペクトラムの解析の際に

* Vocal Sound Band Emphasis Using Deep Learning.　by Tsuyoshi Kitamura (Kobe University), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Tetsuya Takiguchi (Kobe University/ JST PRESTO)

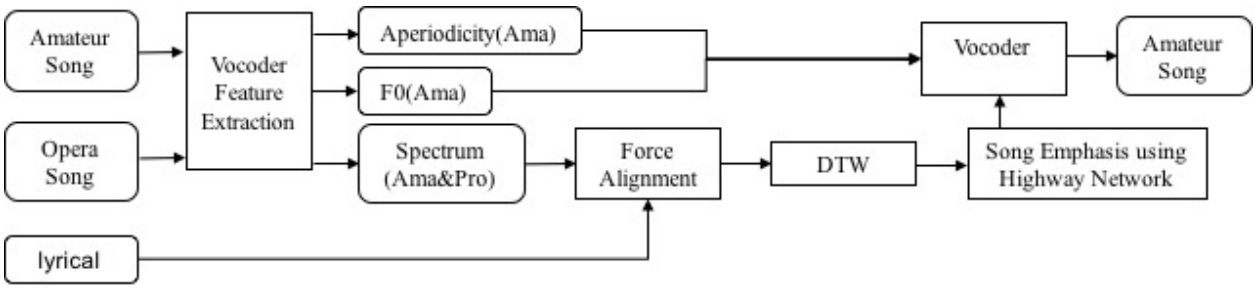


Fig. 2　Proposed method.

は，楽譜上の同一の部分から，0.5 秒の区間を歌唱音声から切り出して解析を行なった．Fig. 1 から，低い周波数の帯域についてオペラ歌手とアマチュアの音声に差は見られない．3000Hz から 4000Hz 帯の中高音域のエネルギーは，プロの音声がアマチュアの音声と比較して強く出ており，音声の艶となる成分が多いことが示されている．

3　提案手法

２節から，オペラ歌手のように歌唱音声に艶を持った通る音声へと変換を行うため，オペラ歌手の歌唱音声が持つ 3000Hz から 4000Hz の中高音域のエネルギー成分を，アマチュアの音声に付与する，もしくは強調することを検討する．また，プロが歌っていないオープンなデータの変換を行うため，モデルとしてニューラルネットワークの枠組みを用いる．本研究では，ニューラルネットワークの入力としてアマチュアの音声特徴量，教師としてオペラ歌手の音声特徴量を用いる．入出力の特徴量の構造が同一であり，3000Hz から 4000Hz の帯域の差分を推定を目的とするため，内部にスペクトルの差分を計算する構造を持つ Highway Networks[3] を用いる．

Fig. 2 に提案手法の概要を示す．学習に用いるパラレルデータを作る処理と，アマチュアの音声を強調する処理に分けられる．音声強調を行なった後，ボコーダを用いて音声を再合成する際，基本周波数 F0 及び非周期性指標はアマチュアの値をそのまま用いる．

3.1　パラレルデータの作成

本研究で用いる歌唱音声について，話者毎に歌うテンポは揃っておらず，パラレルデータを作成する必要がある．Sangeon [2] らの論文内で，歌唱音声についてパラレルデータを作成する際，ビブラートやピッチの変動が大きい箇所でアライメントのエラーが起きることが示されている．そのため，本研究ではアマチュア及びプロの歌唱音声に対して，歌詞情報と歌唱音声を用いて強制アライメントを行うことで音素の時間長を求め，音素ごとに DTW [4] を行うことでパラレルデータを作成した．本研究では，歌詞を

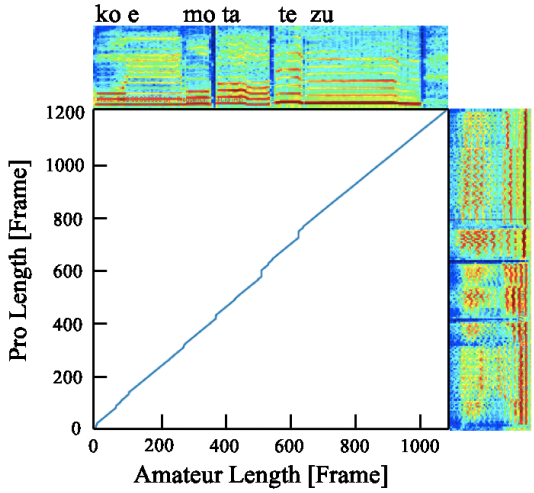


Fig. 3　Example of DTW path between two singing voices.

44 音素からなるラベルデータに変換し，強制アライメントを行なった．Fig. 3 に計算された DTW のパスを示す．Fig. 3 中のスペクトログラムは，0Hz から 5000Hz までのスペクトログラムを表示している．Fig. 3 から，ビブラートによるアライメントのエラーが回避されている．

3.2　深層学習を用いた帯域強調

Highway Networks は，音声合成 [5] や声質変換 [6] などの音声信号処理に広く用いられている．本研究では，パラレルデータの作成後，Highway Networks を用いて，プロのオペラ歌手が強く持つ周波数帯域のエネルギーを差分として，アマチュアの音声に付与することを試みる．Highway Networks の学習には，入力としてアマチュアの歌唱音声を解析して得られるスペクトルと，教師としてオペラ歌手の歌唱音声から得られるスペクトルを用いる．Fig. 4 に Highway Networks の順方向の伝播の概要を示す．Highway Networks の順伝播は，以下の式を用いて計算される．

$$y = H(x) \circ T(x) + x \tag{1}$$

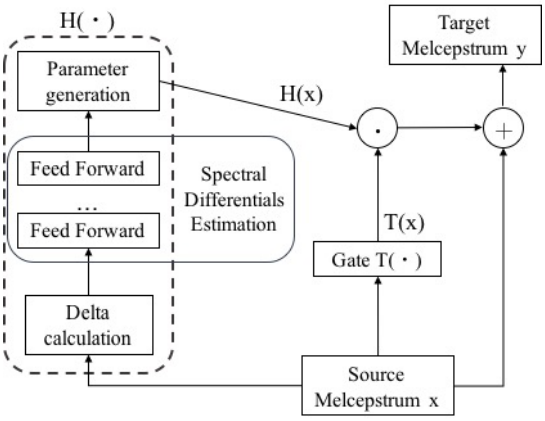


Fig. 4 Speech emphasis using input-to-output highway networks.

$\circ$ はアダマール積を示す．$H(\cdot)$ は，入力スペクトルに対して 1 次の動的特徴量を計算して静的特徴量と結合し，順方向ニューラルネットワークによって差分を推定した後に，MLPG アルゴリズム [7] によってトラジェクトリを求める処理を含んでいる．$T(\cdot)$ は Highway Networks のゲート関数である．ゲート出力 $T(x)$ は [0.0-1.0] の間の値を持ち，推定されたスペクトル差分である $H(x)$ に対して重み付けを行う．ゲート出力 $T(x)$ の値を 0 にすることで，入力がそのまま推定値となる．また，ゲート出力 $T(x)$ の値を 1 にすることで，Residual Networks と同様の振る舞いを行う．

4 実験評価

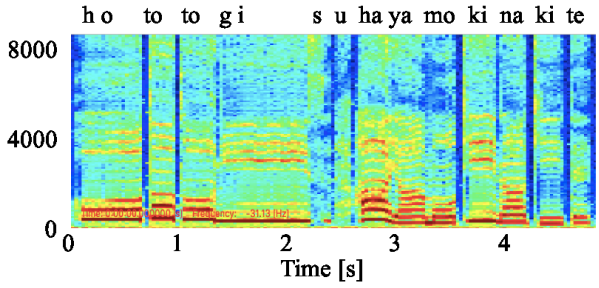
4.1 実験条件

実験データとして，ソプラノのオペラ歌手 1 名とオペラの経験のないアマチュア 1 名の歌声音声を用いた．童謡を中心とした 1 分半程度（無音区間を含む）の曲を 1 5 曲学習データとして用いた．サンプリング周波数は 16KHz，フレームシフトは 5ms とした．また，学習に用いる歌唱音声について，パラレルデータ中から休符を含む無音区間を取り除いた．スペクトル，基本周波数及び非周期性指標の抽出には WORLD[8] を用いた．

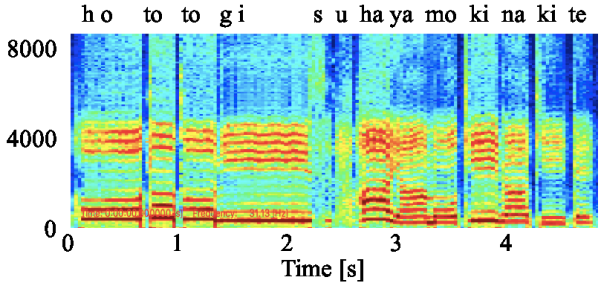
Highway Networks の特徴量には，入出力ともに WORLD を用いて抽出したスペクトルにメルフィルタバンクをかけて得られたの 59 次元のメルケプストラム [9] を用いた．また，入出力はともに平均 0 分散 1 となるように正規化している．差分スペクトルの推定を行うネットワークは，2 層の隠れ層を持つフィードフォワードネットワークを用いた．ネットワークの学習アルゴリズムには Adagrad[10] を用い，学習率は 0.015 とした．隠れ層はそれぞれ 256 のユニットを持ち，隠れ層の活性化関数は ReLU 関数を用いた．また，出力層の活性化関数は，シグモイド関数を用いた．ゲート関数には入力層 59 ユニット，出力層 59 ユニットから構成されるネットワークを用いた．ネットワークの学習アルゴリズムには Adagrad を用い，学習率は 0.01 とした．出力層の活性化関数は，シグモイド関数を用いた．

4.2 実験結果

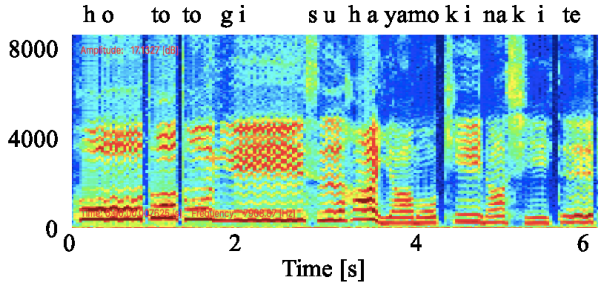
Fig. 5 に，学習に用いていないオープンな歌声音声について，アマチュアの音声，提案手法による変換音声，プロの音声のスペクトログラムを示す．Fig. 5



(a) An amateur singer.



(b) Proposed method.

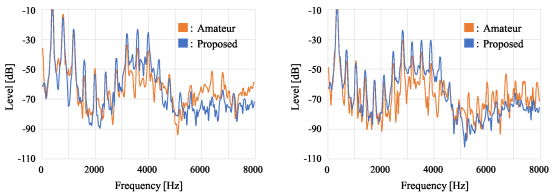


(c) A professional opera singer.

Fig. 5 Examples of Spectrograms.

より，提案手法によってアマチュアの音声の母音について，プロが持つ 3000Hz から 4000Hz の帯域のエネルギーが強調されている．また，話者性を多く含む低い周波数の帯域についてはアマチュアのスペクトルを維持しており，教師にプロの音声を用いたが，話者性の変化が起きていない．

Fig. 6 に，変換前後の'o' と'i' の音素についてスペクトラムを示す．Fig. 6 より，3000Hz から 4000Hz の



(a) spectrum of 'o'. (b) spectrum of 'i'.

Fig. 6 Sample Spectrums.

周波数帯域の成分が強調されている．しかし，ニューラルネットワークを用いたことによる平滑化が原因として，細かい変動が無視されており，高周波の成分が減衰することで音質が劣化していると考えられる．

5 おわりに

本研究では，歌声の音声を対象として，プロのオペラ歌手が持つ音声の帯域成分をアマチュアの歌声へ自動的に移動するシステムを提案した．対象としたアマチュアとプロの歌声を比較すると，プロは 3000Hz から 4000Hz の帯域のエネルギーが強く，音声の艶となる周波数成分が多く含まれていた．前処理として，学習データとなる音声に楽譜情報を用いて強制アライメントをかけ，その後音素毎に DTW を行いアライメントを取ることでパラレルデータを作成した．学習時には，プロの音声が持つ帯域エネルギーを付与するため，Higyway Networks を用いることで内部にアマチュアの音声との差分を推定する手法を用いた．実験の結果，話者性を維持したまま，音声の中高音域の周波数のエネルギー成分が強調された．しかし，ニューラルネットワークによる平滑化により音質が劣化したため，今後は音質の劣化を防ぐとともに，歌声信号の帯域強調やスタイル変換を検討する．

謝辞 本研究の一部は，メック株式会社の支援を受けたものである．

参考文献

[1] S. Bock and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Conference on Digital Audio Effects*, 2013.

[2] S. Yong and J. Nam, "Singing expression transfer from one voice to another for a given song," in *IEEE ICASSP*, 2018.

[3] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[4] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, pp. 1–23, 2008.

[5] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," *SSW-9 (accepted)*, 2016.

[6] Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Transactions on Information and Systems*, vol. 100, no. 8, pp. 1925–1928, 2017.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE ICASSP*, vol. 3, 2000, pp. 1315–1318.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE ICASSP*, vol. 1, 1992, pp. 137–140.

[10] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.