I-003

映像と音声のマルチモーダル学習による ライブ感のある歌声生成のための検討

A Study on Live-Feeling Singing Voice Generation through Multimodal Learning of Audio and Visual Information

畠山 智之[1)]　藤井 一貴[1)]　伊原 滉也[1)]
Tomoyuki Hatakeyama　Kazuki Fujii　Koya Ihara

概要

実際のアイドルのライブ現場では，ダンスに伴う息遣いの変化や，会場の盛り上がりに応じたアドリブ表現など，スタジオ収録の歌唱には含まれない要素が存在する．仮想キャラクターによるライブ表現においても，こうした要素が歌声に自然に反映されることが望ましい．本研究では，ライブ映像を条件付けとした，スタジオ収録歌唱からライブ歌唱への歌声スタイル変換により，ライブ感のある歌唱表現を再現する．これにより，スタジオ収録歌唱に対して仮想キャラクターによるダンス映像を条件としたスタイル変換を適用し，キャラクターのライブパフォーマンスを実際のライブ歌唱に近づけることを目指す．

1　はじめに

ビデオゲームにおけるキャラクターの振る舞いを現実の人間に近づけることはプレイヤーの没入感を高め，ゲーム体験の向上につながる．近年のハードウェア性能や 3 次元コンピュータグラフィックス（Three-Dimentional Conputer Grachics; 3DCG）技術の向上を受け，消費者向け端末上でも高品質な映像表現が可能になったことから，キャラクターの人間らしさを追求した表現や [1]、造形や表情変化へのこだわり [2] といったよりリアルな表現が探求されている．またこのような表現の精緻化だけでなく，プレイの達成度に応じてキャラクターのダンスや歌唱のパフォーマンスの巧拙を変化させたり，あらかじめ用意されたアドリブパターンをライブパフォーマンスに導入する試みがなされている [3, 4]．

これらのゲーム内ライブパフォーマンスは決定的な制御，つまり事前収録されたパフォーマンスを再生したり，複数パターンの組み合わせといった方式によって実現されている．一方，実際のアイドルグループによるライブパフォーマンスでは，演者のコンディションや会場の違いによる表現のゆらぎや，聴衆の盛り上がりに応じたその場限りアドリブといった要素により，全く同一のパフォーマンスが繰り返されることはない．高道らはテキスト音声合成において，この「一期一会」性に着目し，音声特徴量の生成を確率分布からのサンプリングによって，同じ文章でもゆらぎ方の異なる発話の合成を実現している [5]．歌声においても，ライブパフォーマンスならではの一期一会性をゲーム内の映像や音声に取り入れることによって，飽きられにくい新鮮なゲーム体験をプレイヤーに提供できる可能性がある．

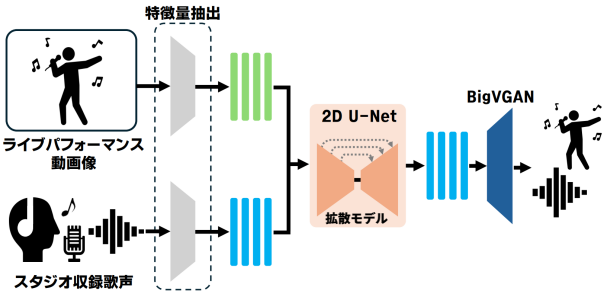

図 1: 提案する歌声変換システムの概要図．ライブ映像とスタジオ収録歌声は特徴量に変換され，2D U-Net で構成される音響モデルに入力される．音響モデルはライブパフォーマンスのスタイルに歌声の音響特徴量を変換し，BigVGAN によって音響信号波形が生成される．

本研究は，その初期検討として，スタジオ収録された歌声から，ゆらぎやアドリブを含むライブパフォーマンス中の歌声へ変換するアプローチを提案する．提案システムは，スタジオ収録とライブパフォーマンスとの差異を歌唱スタイルの違いとして解釈する．具体的には，スタジオ収録歌声と同一の楽曲のライブパフォーマンスを収録した動画像の特徴量を入力とし，ライブパフォーマンス歌唱スタイルへ変換された歌声を出力する Deep Neural Network; DNN を構築する．スタジオ収録音源に加えてライブパフォーマンスの映像特徴量を用いて条件付けすることで，聴衆の盛り上がりや歌唱者の表情に合わせた歌唱スタイルの変化を期待する．変換のための音響モデルには，拡散ベースの 2 次元 U-Net モデルを利用する．音響モデルへの入力には，スタジオ収録歌声から信号処理的に抽出したメルスペクトログラムに加え，動画像から抽出された特徴量を用いた．この構成により，推論時の初期ノイズのランダム生成結果に応じて，アドリブやビブラートなど，ライブ感に影響するさまざまな非決定的要素を確率的に生成することを目指す．実験的評価では，提案システムによる合成歌声が実際にライブ感をもたらすかを評価する．

2　関連研究

2.1　サンプリングに基づいた音声合成

同一のテキストや楽譜情報に対する話者および歌唱者内変動を考慮する試みとして，Conditional Generative Moment Matching Network; CGMMN [6] に基づくポストフィルタを用いたものがある [5, 7]．CGMMN はターゲットとなる条件付き確率分布に対し，モデルの出力のモーメントが一致するように学習する DNN である．あ

1)　サイバーエージェント

らかじめ推論された音響特徴量に加えて雑音を共に入力することで，CGMMN は生成するたびに異なる変動を特徴量に付与するポストフィルタとして機能する．このようにパフォーマンスのゆらぎは，データから統計的に学習された確率分布からのランダムサンプリングとしてモデル化できる．

2.2 Diffusion モデルによる歌声変換

本研究では，スタジオ収録の歌声からライブ収録の歌声への変換を行う上で，ステージ音響や観客の反応，アドリブなど，非決定的で多様な要素を自然に再現する必要がある．そのため，変換を行う音響モデルには，確率的なサンプリングにより多様性を担保できる拡散モデル（diffusion model） [8] のような生成モデルが適している．

歌声から歌声への変換に拡散モデルを応用した先行研究として，Liu らによる DiffSVC[9] がある．DiffSVC では，ASR 音響モデルによって抽出された音素後確率（Phonetic Posteriorgrams; PPGs）を内容情報として用い，それに対してピッチ（対数基本周波数）やラウドネスといった歌唱スタイルに関わる特徴を条件付けに用いることで，メルスペクトログラムのノイズ除去型復元を行い，歌声のスタイル変換を実現している．DiffSVC は，変換対象となる人物が異なる場合でも，内容情報を固定することで，話者固有のスタイルのみを変換可能としている．

一方で本研究では，同一人物における歌唱スタイルの変換，すなわち，スタジオ録音とライブパフォーマンスのように人物は同一だが状況が異なる音声間の変換を目的としている．このような変換では，PPG に含まれるような人物非依存の情報を固定すると，変換したいスタイルやアドリブなどの情報が変換できなくなってしまう．そこで本研究では，内容とスタイルの切り分けを厳密には行わず，メルスペクトログラムそのものの変換を行う拡散モデルをベースに音響モデルを構築することで，歌声のタイミングや強弱，音質の変化といった非決定的要素を含めたスタイル変換を可能とした．

2.3 動画像による条件付けを用いた音合成手法

本研究では，スタジオ収録の歌声からライブの歌声への変換を行う際に，ステージの音響や演者のアクションなどを反映するために，ライブ映像をの条件付けとして使用する．動画像を入力としそれに付随する音響信号を出力する Video-to-Audio; V2A には，これまで敵対的生成ネットワークや拡散モデルを用いたものなど，様々なアプローチが提案されている．本研究では，動画を条件付けとして使用するための特徴量抽出の手法として，Wang らの Frieren [10] を参考にした．Frieren では，事前学習済みの視覚エンコーダを用いて各フレームから特徴量を抽出し，映像中の時間的な情報を保ったまま活用できるようにしている．具体的には，Contrastive Audio-Visual Pretraining; CAVP [11] 特徴を用いて，映像と音声との整合性が高い視覚特徴を取得する．さらに，抽出されたフレームレベルの特徴列は，音声のメルスペクトログラム潜在表現と時間的に整合させるために，Length Regulator; LR [12] と呼ばれる，エンコードされたベクトル系列を複製する処理によって系列長を一致させる．これは，映像のフレームレートと音声表現の時間分解能の差を埋めるために，各映像フレームの特徴を一定回数繰り返すことで実現される．このようにして時間軸が整合された視覚特徴列は，音声生成に用いる条件情報として利用され，高い時間的同期性を持った音声生成を可能にしている．

3 提案手法

3.1 歌声変換システム

提案システムの概要図を図 1 に示す．まず，スタジオ収録された歌声の音響信号と同一楽曲を歌唱中のライブパフォーマンス動画像は，それぞれ音響特徴量（メルスペクトログラム）および映像特徴量（CAVP 特徴量）に変換される．これらは 2 次元 U-Net の音響モデルへ入力され，ライブパフォーマンススタイルの歌声の音響特徴量へと変換される．その後，ニューラルボコーダによって特徴量から時間波形へと変換されることで，目的の変換音響信号が得られる．

3.2 特徴量抽出

まず，ライブパフォーマンスの動画フレーム列を $\boldsymbol{v}$，対応するスタジオ収録歌声の時間波形を $\boldsymbol{x}^{\text{studio}}$ とする．

音声側の処理では，短時間フーリエ変換（STFT）を用いて $\boldsymbol{x}^{\text{studio}}$ から振幅スペクトル系列を算出し，これに対してメルスケールに基づいたフィルタバンクを適用することで，音高知覚に整合した低次元のメルスペクトログラムを得る．この一連の処理 $\mathcal{M}$ によって得られるメルスペクトログラムは以下のように定義される：

$$\boldsymbol{m}^{\text{studio}} = \mathcal{M}(x^{\text{studio}}) \in \mathbb{R}^{F\times N},$$

ここで，F はメルバンド数，N は時間フレーム数である．

映像側の処理では，事前学習済みの CAVP 視覚特徴抽出器を用いて，各動画フレームからセマンティックかつ時系列的に整合性のある特徴量を抽出する：

$$\boldsymbol{f} = f_v(\boldsymbol{v}) \in \mathbb{R}^{D\times N_v},$$

ここで，D は各フレームに対応する特徴ベクトルの次元数，N_v は映像フレームの数（すなわち視覚特徴系列の長さ）である．

音声と映像の特徴列は，通常フレームレートが一致しないため，$\boldsymbol{f}$ の時系列長 N_v を音声のフレーム長 N に整合させる必要がある．このため，本研究では LR を用いて $\boldsymbol{f}$ の各フレームを適切な回数繰り返すことで，$\boldsymbol{f}$ を音声特徴系列と同一長にリサンプリングする：

$$\tilde{\boldsymbol{f}} = \text{LengthRegulator}(\boldsymbol{f}) \in \mathbb{R}^{D\times N}.$$

最終的に得られた $\tilde{\boldsymbol{f}}$ は，音響モデルへの条件情報として，スタジオ音声のメルスペクトログラムと共に利用される．

3.3 音響モデル

本研究では，スタジオ収録音声 $\boldsymbol{m}^{\text{studio}}$ からライブ収録音声 $\hat{\boldsymbol{m}}^{\text{live}}$ への変換を実現するために，拡散モデルを生

成器として採用する．ネットワークは一般的に画像生成などに用いられる，2 次元の U-Net 構造でパラメータ化された ε_θ によって定義される．

3.2 で述べたように，視覚特徴量 f は，事前学習済みの CAVP 視覚エンコーダを用いて，ライブ映像からフレーム単位で抽出される．その後，音声のメルスペクトログラムと時間的に整合させるため，LR によって各フレーム特徴が複製され，音響特徴系列と同一の時間長にリサンプリングされる．こうして得られた視覚特徴列と，変換元のメルスペクトログラム m^{studio} は，時間軸・チャネル軸を揃えた上で連結され，拡散モデルの入力条件とする．

順方向拡散過程は，真のメルスペクトログラム $x_0 = m^{\text{live}}$ に段階的にガウス雑音を加えることで，完全なノイズ x_T を得るものであり，以下のように定義される：

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \quad t = 1, \ldots, T. \tag{1}$$

逆拡散過程では，スタジオ音声 m^{studio} と対応する映像から抽出された視覚特徴量 f を条件として，x_T から段階的に復元を行い，ターゲットのライブスタイル音声を生成する．このとき，各ステップにおける遷移は次のようにモデル化される：

$$p_\theta(x_{t-1} \mid x_t, f, m^{\text{studio}}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, f, m^{\text{studio}}, t), \Sigma_\theta(t)), \tag{2}$$

ここで μ_θ はネットワーク ε_θ を用いて構成される平均項であり，t は拡散ステップである．

拡散モデルの学習には，以下のような予測誤差に基づく損失関数を用いる：

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t, f, m^{\text{studio}}, t)\|_2^2\right], \tag{3}$$

ここで $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ はガウス雑音であり，$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$ により計算されるノイズ付与メルスペクトログラムである．

推論時には，完全なノイズから $x_T \sim \mathcal{N}(0, \mathbf{I})$ を初期値として (2) の逆拡散過程を経て，スタジオ音声 m^{studio} と映像特徴 f に条件付けられたライブスタイルのメルスペクトログラム $\hat{m}^{\text{live}}$ を生成する．

3.4 BigVGAN による波形生成

音響モデルから得られた $\hat{m}^{\text{live}}$ は，ニューラルボコーダと呼ばれる波形生成器 G_ϕ へと入力され時間波形 $\hat{x}^{\text{live}}$ へと変換される．

$$\hat{x}^{\text{live}} = G_\phi(\hat{m}^{\text{live}}) \tag{4}$$

本研究では G_ϕ に，GAN に基づくニューラルボコーダである BigVGAN [13] を用いた．Snake [14] と呼ばれる出力の周期性を誘導する活性化関数と，アップサンプリング時のエイリアシングを防ぐ機構を備えた大規模なモデルであり，学習データに含まれないドメインのメルスペクトログラムを条件付けした場合でも高い品質の波形を合成できることが報告されている．

4 実験的評価

本研究では，スタジオ収録音源にライブ感を付与する音源生成手法の有効性を検証するため，Mean Opinion Score; MOS [15] と AB テストに基づく主観評価実験を実施した．本節では，その実施条件および評価手法について述べ，最後にその結果と考察を行う．

4.1 実験条件

モデルの学習に用いる歌声および映像データは非公開の内部データを用いた．前処理として，混合音源から歌声を抽出するために Demucs [16] による音源分離を行った．検証に用いたモデルの重みとして，CAVP エンコーダおよび BigVGAN は Frieren プロジェクトで提供されている学習済み重み[1)]をそのまま用いた．音響モデル（パラメータ数約 300 M）は，学習率 1e-5，バッチサイズ 4，学習ステップ約 300 k の条件で学習を行った．

聴取実験には，3 種類の動画を使用した．映像は全てにおいて共通のライブ映像から一部を切り抜いたものを使用し，1 つ目は原音となるスタジオ収録音源（以下，**Studio**）と映像を組み合わせた動画，2 つ目は本手法により生成されたライブ感付与音源（**Generated**）と映像を組み合わせた動画，3 つ目は実際にライブ環境で収録された音源（**Live**）と映像を組み合わせた動画である．これらの動画を PC モニターとイヤホンを通じて提示した．被験者は音楽に対する専門知識の有無を問わず，成人 12 名を対象とした．

主観評価では，被験者は以下の 2 種類の基準において，1 問目は 3 つの動画それぞれに対して自然性 MOS 評価を，2 つ目は全ての組み合わせに対する AB テストを実施した．

1. **映像に対して，歌声がどれくらい自然に感じましたか？**：1（全く感じない）〜5（非常に強い）で評価.
2. **動画１〜３から抽出したペアを比較した場合，どちらがより歌声と映像がマッチしていると感じましたか？**：1 つを選択して評価

表 1: 歌声の自然性に関する MOS とその標準偏差.

Method	MOS (↑)
Studio	3.41 ± 0.99
Generated	3.08 ± 0.90
Live	3.08 ± 1.50

表 2: 歌声の映像に対するマッチ度合いに基づく AB スコア.

A	Scores	B
Studio	**0.67** vs 0.33	**Generated**
Studio	**0.58** vs 0.42	**Live**
Generated	**0.50** vs **0.50**	**Live**

1） https://github.com/cyanbx/Frieren-V2A

4.2 結果と考察

結果を表 1，2 に示す．MOS 評価の結果から，**Studio** とライブ映像を組み合わせた動画が，映像に対する歌声の自然さにおいて最も高評価であった．また，AB テストにおいても，**Studio** が **Generated**，**Live** と比較したどちらの場合においても高評価であった．実験前の仮説としては，どちらのスコアにおいてもライブ音源が最も高い評価になることを想定していたため，想定外の結果となった．

MOS 評価において，提案手法が自然性において **Live** と同等のスコアが得られたことから，ライブ音声における空間の残響や歓声をスタジオ収録歌声に重畳しても，大きな劣化が生じなかったことが示唆される．特に BigVGAN + CAVP による音響合成処理は，人工的な残響や観客ノイズを音楽体験を損なわない程度に抑制できたと解釈できる．以上の結果から，本研究で提案したライブ感付与手法は，**Generated** が **Live** と同等の主観評価を獲得した一方で，**Studio** を上回るには至らなかったことが分かる．想定外にスタジオ収録音源が最も高い評価を得た要因として，ライブ音源と生成音源には会場残響や観客音が含まれる一方，イヤホン聴取という提示条件が空間的臨場感を弱め，これらの要素を「不自然な付帯音」と捉えた被験者が一定数存在したと推察される．特に **Live** の標準偏差が大きい（± 1.50）ことは，残響や環境雑音に対する好みが被験者間で大きく揺らいだことを示唆している．

5 おわりに

本研究では，ライブパフォーマンスにおけるゆらぎやアドリブを歌声において再現することを目的に，スタジオ収録された歌声とライブパフォーマンスの動画像を入力とする歌唱スタイルへ変換システムを提案した．聴取実験に基づく主観評価結果は，提案手法がライブ感のある歌声へ変換できることを示唆するものであった．今後の展望として，変換の精度の向上に取り組みつつ，ライブ感をより具体的に評価できるような指標についても検討し，手法の優位性をより明確にしたいと考えている．

謝辞

原稿の添削などに協力いただいた同社の小口 純矢氏に感謝する．

参考文献

[1] 杉村 貴之, 見原 朋也, “「IDOLY PRIDE」の 3D 美少女キャラクターを魅力的かつ効率的に制作する手法”，https://cedec.cesa.or.jp/2021/session/detail/s606450b1dccf2.html, 2021, Computer Entertainment Developers Conference (CEDEC) 2021.

[2] ——，“神は細部に宿る！「学園アイドルマスター」のこだわり抜いた 3d キャラクター・背景制作”，https://cedec.cesa.or.jp/2024/session/detail/s660138bbdf4c1/, 2024, Computer Entertainment Developers Conference (CEDEC) 2024.

[3] バンダイナムコエンターテインメント，“学園アイドルマスター”，https://gakuen.idolmaster-official.jp/，2024.

[4] サイバーエージェント，“『プロジェクトセカイ カラフルステージ！ feat. 初音ミク』〜3D バーチャルライブの体験を支えるクリエイティブ〜”，https://cagc.cyberagent.co.jp/2024/session/index.html?id=M3xKrmDu，2024，CyberAgent Game Conference 2024.

[5] S. Takamichi, T. Koriyama, and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching networks,” in *Proceedings of INTERSPEECH 2017*, 2017, pp. 3961–3965.

[6] Y. Ren, J. Zhu, J. Li, and Y. Luo, “Conditional generative moment-matching networks,” in *Proceedings of Advances in Neural Information Processing Systems 29 (NIPS 2016)*, vol. 29, 2016.

[7] H. Tamaru, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, “Generative moment matching network-based neural double-tracking for synthesized and natural singing voices,” *IEICE Transactions on Information and Systems*, vol. E103-D, no. 3, pp. 508–516, 2020.

[8] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, vol. 33, 2020, pp. 6840–6851.

[9] S. Liu, Y. Cao, D. Su, and H. Meng, “Diffsvc: A diffusion probabilistic model for singing voice conversion,” in *Proceedings of INTERSPEECH*, 2022, equal contribution and work done during internship at Tencent AI Lab.

[10] Y. Wang, W. Guo, R. Huang, J. Huang, Z. Wang, F. You, R. Li, and Z. Zhao, “Frieren: Efficient video-to-audio generation network with rectified flow matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, 2024, equal contribution.

[11] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, “Contrastive audio-visual masked autoencoder,” in *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

[12] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech: Fast, Robust and Controllable Text to Speech,” in *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, vol. 32, 2019, pp. 1–10.

[13] K. Lee, J. Kim, J. Kong, J. Kim, S. Kim, K. Byun, S. Park, Y. Na, B. Kim, Y. Lee, S. hwan Baek, I.-H. Bae, J.-H. Kim, H.-S. Choi, I.-S. Sim, S.-H. Oh, S. Lee, and J.-W. Ha, “Bigvgan: A universal neural vocoder with large-scale training,” in *Proceedings of the International Conference on Learning Representations (ICLR) 2023*, 2023.

[14] L. Ziyin, T. Hartwig, and M. Ueda, “Neural networks fail to learn periodic functions and how to fix it,” in *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 1583–1594.

[15] International Telecommunication Union, “Methods of subjective determination of transmission quality,” *ITU-T Recommendation P.800*, 1998.

[16] S. Rouard, F. Massa, and A. Défossez, “Hybrid transformers for music source separation,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, 2023, pp. 1–5.