Abstract

本論文では、新しいトーキングヘッドビデオ編集のフレームワークを提案します。このフレームワークにより、ユーザーは頭部の向き、感情、および瞬きを自由に編集しつつ、視聴覚の同期を維持することができます。従来の手法が主にトーキングヘッドビデオの生成に焦点を当てていたのに対し、我々の提案モデルは入力ビデオのトーキングヘッドを編集し、フルフレームに戻すことができ、より広範なアプリケーションをサポートします。我々の提案フレームワークは2つの部分で構成されています。a) 頭部の向き、感情、瞬きなど自由に制御可能な属性に対応しながら、元のフレームに適合するトーキングヘッドを生成する再構築ベースのジェネレーター。b) 属性と視覚の同期を強制する複数属性のディスクリミネーター。さらに、全体の生成品質を向上させるために注意モジュールと知覚損失を導入します。定量的指標および定性的比較により、既存のアプローチと比較します。

1.

オーディオ駆動のトーキングフェイス生成は近年大きく進歩しました。以前の研究者たちは、ソースオーディオからトーキングヘッドビデオを生成する問題に取り組んでいました【5, 6, 7, 8】。この分野では最近、重要な進展が見られるものの、多くは唇の同期を改善することに焦点を当てており、頭部の向き、表情、瞬きといった全体的な自然さに関わる他の重要な要素を無視していました。Wav2Lip【1】のように、生成されたビデオで元の頭部の向きや瞬きを保持するものもあります。他の研究【9】では、ある程度の自然さを示す頭部の動きをオーディオソースから推測して生成できるものの、自由に制御することは困難です。最近では、PC-AVS【3】が参照ビデオから抽出された低次元の姿勢コードを用いて生成ビデオに頭部の向きを適用することに成功しました。EAMM【2】やGC-AVT【4】は感情を持つトーキングフェイス生成を可能にし、生成された顔の品質と多様性を大いに向上させました。しかし、これらのアプローチはソースフレームやソースビデオからトーキングヘッドビデオを生成するのみで、生成されたビデオをフルフレームに戻すことはできず、実用性が制限されます。それに対して、我々の提案する手法はトーキングヘッド領域を編集し、元のビデオに戻すことが可能であり、ビデオコンテンツ編集、視覚的ダビング、リモート教室やオンライン講義など、より広範な用途に対応します。その他の研究【10, 11】でもある程度のトーキングヘッドの編集が可能ですが、顔の属性の制御が限られており、トレーニングと推論のために複雑な3Dモデルが必要です。

本論文では、元のフレームに適合するトーキングヘッドを生成できる再構築ベースのジェネレーターを設計します。エンコーダー側では、頭部領域がマスクされた入力フレームがフェイスエンコーダーによって処理され、顔の特徴が埋め込まれます。オーディオエンコーダーはターゲットオーディオから抽出されたメルスペクトログラムを処理し、追加のエンコーダーはターゲットビデオから抽出された頭部の向きコード、感情コード、および瞬きコードを埋め込み特徴に変換します。これらは連結され、残差スキップ接続を持つデコーダーネットワークに供給されます。このパイプラインでは、提案するジェネレーターは、マスクされたフレームからトーキングヘッドをインペイントしながら、オーディオおよび追加の属性をビデオに同期させ続けます。

さらに、Wav2Lip【1】のリップシンクディスクリミネーターのアイデアを拡張し、唇だけでなく頭部の向きや表情もビデオに同期させる複数属性のディスクリミネーターを導入しました。これにより、提案モデルでは自由視点の表情豊かなトーキングヘッド生成が可能となり、生成結果の精度と品質が大幅に向上します。

我々の提案するフレームワークでは、頭部の向き、感情、瞬きといった複数の属性を持つトーキングヘッドビデオを合成することができます。さらに重要なのは、生成されたトーキングヘッドを元の位置に戻すことができるため、任意のサイズやスケールのビデオを編集することが可能になる点です。多くのシーンでは、ユーザーは話す人物、特に手や上半身の動きを含むものを好むと想定しています。類似のアプローチとの高レベルな比較を表1にまとめます。

まとめると、本研究では以下を提示します：

- 元のフレームに適合しつつ、頭部の向き、顔の感情、瞬きといった自由に制御可能な属性に対応するトーキングヘッドを生成できる再構築ベースのジェネレーター。

- 属性と視覚の同期を強制する複数属性のディスクリミネーター。

- これら2つのモジュールを組み合わせることで、参照の有無にかかわらず合成中に複数の属性を持つフルフレームのトーキングヘッドビデオを編集できる提案モデル。このタスクにおいて初期の試みの1つです。

2.提案するフレームワーク

我々のモデルは、以下の3つのサブコンポーネントで構成されています。

a) 複数の入力を埋め込み特徴に処理するエンコーダー。

b) 入力に対応するフレームを生成するジェネレーター。

c) 属性と視覚の同期を強制する、エンコーダーと似た構造を持つ事前学習済みの複数属性ディスクリミネーター。

モデルの概要は図2に示されています。

2.1. Model Input

我々の提案するモデルの入力は以下の通りです：

- 入力フレーム：マスクされた頭部領域を持つ5つのソースフレームと5つのターゲットフレーム。画像サイズは256 × 256。

- オーディオ入力：ターゲットフレームに対応するオーディオから計算されたメルスペクトログラム。

- 頭部の向きコード：オイラー角（ヨー、ピッチ、ロール）を表すb × 3 × tテンソル。事前学習済みのHopenet【12】によって予測されます。

- 感情コード：7つの感情カテゴリを表すb × 7 × tテンソル。事前学習済みのRMN【13】から計算されます。

- 瞬きコード：目の開閉度を表すb × 1 × tテンソル。dlib【14】を使用して顔のランドマークから計算されます。

ここで、bはバッチサイズ、tはタイムステップ（5に設定）を表します。推論段階では、ターゲットフレームは入力フレーム自体であり、他の属性は手動で定義するか参照フレームから抽出できます。

2.2. Model Architecture

\*\*Encoders:\*\* 5つの異なるエンコーダーを使用し、左側に示されているように、2D畳み込みのスタックとして異なる入力データを処理します。全ての出力はb × D × tであり、連結されてジェネレーターに供給されます。

\*\*Generator:\*\* Wav2Lipと同様のジェネレーターを使用し、元々はLipGAN【15】から借用されたもので、アップサンプリングのための転置畳み込みを含む2D畳み込みのスタックです。さらに、画像生成品質を向上させるためにデュアルアテンションモジュール【16】をジェネレーターに採用します。ジェネレーターは生成されたフレームとターゲットフレーム間で計算されるL1損失および知覚損失【17】を使用して訓練します。また、複数属性ディスクリミネーターから同期損失を採用します。生成されたフレームとターゲットフレーム間の目のランドマーク距離に基づいて計算される瞬き精度を改善する瞬き損失も使用します。

### Multi-Attribute Discriminator

Wav2Lip【1】が他の既存の手法に比べてリップシンクロナイゼーションにおいて顕著な改善を示したことを踏まえ、事前学習済みのディスクリミネーターは同期タスクに非常に有効であると仮定します。そこで、Wav2Lipのリップシンクディスクリミネーターを拡張し、より強力なマルチアトリビュートディスクリミネーターにしました。これにより、音声データと視覚データだけでなく、頭部の向きコードや顔の感情コードなどの追加の特徴も処理します。顔のエンコーダーから生成された埋め込み特徴と他のエンコーダーからの埋め込み特徴との間のL2距離を最小化します。

## 3. RESULTS

### 3.1. Datasets

我々は、最初にSynthesizing Obama【18】によって収集されたバラク・オバマのYouTubeビデオ5時間分を使用しました。それに加えて、ジョー・バイデン大統領のYouTubeビデオを手動で5時間収集し、Merkel Podcast Corpus【19】で提案されたデータセットの一部も借用しました（これも約5時間です）。各データセットの90%を訓練セットとして、10%を検証セットとして使用します。現在のところ、モデルは精度向上のために単一人物で訓練されています。

### 3.2. Experiments

トーキングヘッド生成における2つの主要なモデルを選択しました。Wav2Lip【1】はフルフレームで編集できますが、追加の属性を適用することはできません。一方、PC-AVS【3】は生成されたビデオに頭の向きを変更できますが、フルフレームに戻すことはできません。Wav2Lipは提案モデルと同じデータセットで訓練しました。PC-AVSモデルについては、PC-AVSの訓練コードが公開されていないため、我々のデータセットで訓練することはできませんでした。したがって、著者によって提供されたVoxCeleb2【20】で事前学習されたモデルを使用しました。

我々の問題設定は、入力ビデオの属性を編集することです。つまり、別のビデオから属性を参照し、元のビデオに頭部の向き、感情、瞬きを適用しながら同期を保つことです。セクション3.1で説明した検証データセットを使用します。具体的には、各データセットからランダムに100組の短いビデオペアを選択します。それぞれのペアについて、1つのビデオは入力フレームとオーディオとして使用し、もう1つは属性参照用に使用します。

EAMM【2】モデルおよびGC-AVT【4】モデルは、非常に最近発表されたため、評価のためのコードが公開されていません。そのため、これら2つの方法と我々の結果を比較することはできません。しかし、表1で議論されているように、EAMMは感情のみを適用でき、GC-AVTは瞬きを変更できず、両方の方法はフルフレームでの編集ができません。

### 3.3. Quantitative Comparison

ビデオ生成の品質を評価するためにSSIM【21】を使用し、口の動きを評価するためにSyncNet【22】スコアとLMDm【23】を使用し、表情や瞬きの適合性を評価するためにLMDfを使用しました。表2に示されているように、提案モデルは生成品質とランドマークの精度においてベースラインモデルを上回っています。なお、Wav2Lipは同じSyncNetディスクリミネーターで訓練されているため、最高のSyncNet信頼スコアを持っていますが、我々のモデルは唇の同期においても同等の性能を持ち、他の多くの面でより柔軟です。

### 3.4. Visual Quality Comparison

結果の定性的な比較を図3に示します。明らかに、Wav2Lipはオーディオ入力のみを受け付けるため、何の変更も適用できません。一方、PC-AVSは頭部の向きを適切に変更できますが、それを元のフレームに戻すことができないワンショット生成モデルです。

### 3.5. Attribute Editing

提案するフレームワークは、頭部の向き、表情、瞬きといった3種類の顔の属性を完全に制御することができます。異なる顔の属性を編集する例を図4に示します。

### 3.6. Ablation Study

我々はマルチアトリビュートディスクリミネーターの有効性を検証するためにアブレーションスタディも実施しました。各属性は頭部の向き（P）、感情（E）、瞬き（B）として記述されています。結果は表3に示されており、対応する要素が追加されるとパフォーマンスが向上することがわかります。

## 4. CONCLUSION

本論文では、音声と視覚の同期を維持しながら、トーキングヘッドビデオの特定の属性を編集できる新しいフレームワークを提案しました。提案するフレームワークは、この目標を最も早く実現しようとする挑戦者の1つであり、評価の結果、既存のモデルを上回ることが示されました。我々の提案モデルの主な制限は、複数のアイデンティティで訓練した場合に満足のいく結果を生成できない可能性があることです。将来的な研究では、アイデンティティ学習を含む一般的なモデルの開発を期待しています。