口真似音声からの効果音合成技術による音響制作手法の評価 - 音響制作未経験者を対象とした映像への効果音付与 -

滝沢 力[1,3,a)]　平井 重行[2]　須田 仁志[3]

概要：人の調音器官（発音能力）を活用し，口真似音声の表現力を用いた効果音合成手法 PronounSE は、コンテンツ制作のための新たな音響合成手法として位置付けできる．この手法は，頭の中の音のイメージをボイスパーカッションのように言語非依存な表現として口で真似て発声し，その音響信号を元に効果音を合成するものである．これまで，爆発音に焦点を当てたデータセットを元にしたモデルを元に，口真似の微細表現の追従性や合成品質の観点で有効性について評価を行ってきたが、コンテンツ制作タスクにおける有用性に関するユーザ評価は未実施であった．そこで，本研究では，無音動画に対して音を付与するタスクにおいて，効果音ライブラリからの音素材選択と PronounSE を用いた音素材制作とで比較実験を行った．音響制作未経験者を対象としたこの評価実験により，PronounSE による音素材制作の良し悪しが制作者の性別や口真似の巧拙で左右することが確認できた一方で，生成系深層学習技術を用いた音素材を得る行為において，創作性や偶然性が音響制作に寄与することも明らかになった．本論文ではそれらのユーザ評価内容について述べる．

1. はじめに

ゲームやアニメ，映画などの映像作品では，場面に応じて非現実的な音や，日常での環境音，ユーザインタフェース（UI）の操作音など様々な効果音が使用される．プロの制作現場においては，音響監督・ディレクターによる方針の元，サウンドデザイナーによって膨大な効果音ライブラリからの音素材選定，フォーリー録音 [1]，シンセサイザを駆使した制作，プロシージャルオーディオ専用ソフトウェアでの合成など，様々な手法で効果音制作が行われる [2], [3]．それらいずれの方法においても、専門的な知識や経験、アイデアが求められる．

他方，大規模データに基づく深層学習技術の進展により，Dream Machine[*1]や，Veo 3[*2]，Sora 2[*3]などで高品質な映像を容易に制作可能となってきており，それに伴い，人の声や音楽以外の音として環境音を合成する技術も盛んに研究されている．それら多くの環境音合成技術では，動画像 [4], [5], [6] やテキストプロンプト [7], [8], [9]，環境音ラベル [10], [11]，音素列 [10]，音響信号 [8], [9], [11], [12] など様々なモダリティからの合成手法として提案されているものの，音の細かな特徴表現が困難であったり，動画のフレーム外で起こっている音響イベントまでは考慮できない手法が多い．

そのような中，環境音の合成音品質を評価するための客観評価手法の提案や主観評価も実施されており，合成音の自動評価に関する研究 [14] やコンペティションも行われ始めている [15], [16]．ただ，音響制作における有用性や効率性，新規性評価に向けた，人を対象としたユーザ評価を扱った研究は少ないのが現状である．

我々は，プロのサウンド制作の場面で効果音を非言語的な口真似音声で微細な表現をしつつ音のイメージ伝達が行われていることに着目し，口真似音声から効果音を合成する新たな手法 PronounSE を提案し，合成音の品質評価等を行ってきた [13], [18]．この手法は，前述の様々なモダリティの中では音響信号から音響信号を合成するものだが，微細変化を含め人の非言語的発音能力を最大限活用する新たな音響合成手法と言える．この PronounSE は，効果音と非言語的口真似音声との対応関係を深層学習することにより，音の微細なニュアンス制御を行うものである．現時点では爆発音に焦点をあてたデータセットで口真似ニュアンスを反映した合成が実現されている．

本研究では，人の口真似による新たな音響合成手法 PronounSE を用いて，映像に音響効果を付与する場面を設定し

[1] 京都産業大学大学院
[2] 京都産業大学
[3] 産業技術総合研究所
[a)] takiriki1216@gmail.com
[*1] https://lumalabs.ai/dream-machine
[*2] https://gemini.google/us/overview/video-generation
[*3] https://sora.chatgpt.com/

© 2026 Information Processing Society of Japan.

て，音響制作初心者を対象としたユーザ評価を行った．本論文は，まず口真似音声から効果音合成を行う PronounSE について概説する．そして，口真似による効果音合成を用いた映像への音響効果付与と，効果音ライブラリから音素材を選択する従来の音響効果付与する方法とでユーザ比較評価実験を行った結果について述べる．

環境音を含む効果音合成手法の研究分野における，本研究論文の貢献を以下に示す．

- 人の口真似能力を活用する新たな音響合成手法の制作効率に関して，音響制作未経験者を対象としたユーザ評価を実施し，その効果や傾向について明らかにしたこと
- 口真似からの効果音合成の操作量について：多くの実験参加者が 3〜8 回程度の合成回数でタスクを終えることを確認し，その間に動画への付与に対する迷いの軽減など有効性が確認できたこと
- 創作性の観点：口真似音声での繰り返しのインタラクティブな合成が音素材のイメージに近づく点がタスク後のインタビューから確認でき，人と AI との共創による音響制作の有効性が示唆されたこと
- 口真似音声による音響制作手法の位置付け：AI 支援デザインプロセスや不確実性と曖昧さへの対応，AI 技術への期待をタスク後インタビューから得られたこと

2. 関連研究

2.1 ユーザ評価を実施した音響合成技術

AutoSFX [5]: AutoSFX は，入力映像に対する物体セグメンテーションを行い，視覚情報と音響情報を融合することで，映像内のどのオブジェクトがどのタイミングでどのような音を持つべきかを考慮可能な音響合成技術である．また，オブジェクトの距離情報を基に合成音をミキシングするモジュールも備わっている．ユーザスタディとして効果音付与された動画に対する主観評価と AutoSFX の使用感に対するアンケート調査が行われた．主観評価では，30 名（クリエイター 10 名，視聴者 20 名）に映像と音の真実味と時間的一致性，音質，全体的な体験の 4 観点について AutoSFX の合成音付き動画を評価させた．アンケート調査では，同じ 30 名に AutoSFX を用いて動画からの効果音合成を行わせ，役に立たない・やや役に立つ・非常に役に立つの三段階で評価させた．

MultiFoley [8]: MultiFoley は，映像を基本条件としテキストや参照音の複数のモダリティを統合して扱え，各モダリティの潜在特徴を結合した特徴量を条件付けて拡散モデルを学習することで，マルチモーダルな音響合成を可能としている．人による評価として，動画とテキストで条件付けて得られた合成音を対象に，テキストとの意味一致性，動画との同期性，明瞭性・高解像度性，および総合的な音質の観点で，20 名の参加者に評価させた．

2.2 音響制作における効果音合成 AI のユーザスタディ

生成系深層学習による音響合成技術がプロのサウンドクリエイターの制作実践をどのように支援できるかを調査した研究が報告されている [17]．ユーザスタディでは，9 名のプロサウンドクリエイターに対し，音の潜在特徴を制御可能な効果音生成モデルを基盤とする 2 種類の制作支援ツールを各自の創造的タスクに利用させ，その使用過程に関して半構造化インタビューを行った．その結果，次の知見が報告されている．

- **AI 支援サウンドデザインプロセス**：生成 AI は制作作業の自動化ではなく，フィールド録音やライブラリ検索の代替・補完として，素材探索や候補生成の補助に適している．
- **不確実性と曖昧さへの対応**：生成 AI は予測不能な結果や曖昧なパラメータを伴うため新たなアイデア創発に寄与する一方，締め切り付き作業では効率性が求められるため，予測可能性と偶然性を切り替え可能なインターフェースが有効である．
- **AI に期待するサウンドデザイン**：生成 AI は「共同制作者」ではなく「ツール」として位置付けられるべきであり，ユーザ主体で操作できる仕組みが必要である．

3. 口真似音声からの効果音合成手法 PronounSE

3.1 PronounSE の概要

図 1a に示す通り，PronounSE[13] は所望の音の音韻・韻律特徴を模した口真似のみを入力とし，その特徴に基づいて効果音合成を行うため，音の説明文や音響パラメータの制御の専門知識が不要であり，口真似のみから直感的に合成を試みることができる．そして，GPU 上で高速に動作し，瞬時に口真似から効果音を合成可能なことから，発声後に合成された音を聴いて，発声ニュアンスを繰り返し修正しながらインタラクティブに合成を繰り返すことで狙った音を作り出す制作スタイル（図 1b）が可能となる．

現状の PronounSE は，多種多様な爆発音とそれらの口真似音声によるデータセットで学習されており，図 2 に示す口真似ニュアンスを反映した爆発音合成を実現している．

3.2 口真似音声を基にした合成音に関する他手法との定量評価

口真似音声を入力して音響合成が可能な Stable Audio 2.0 [9] と T-Foley [12] を PronounSE の比較手法とし，爆

© 2026 Information Processing Society of Japan.

(a) 口真似音声を入力しての音響合成　　(b) 繰り返し合成によるインタラクティブな制作スタイル

図 1: PronounSE の利用概念図

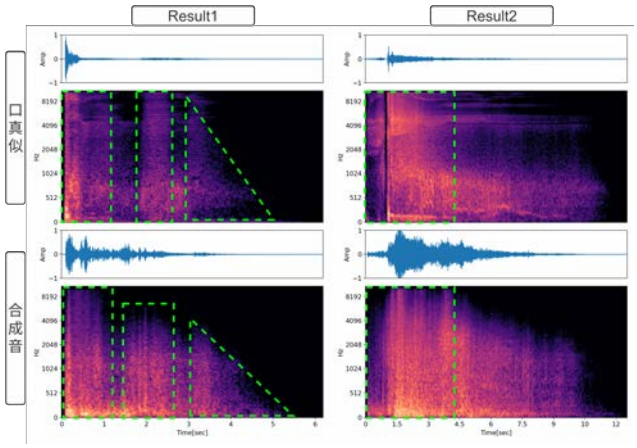


図 2: PronounSE による爆発音合成例．上段は口真似，下段は合成音．Result 1 は，口真似の爆発と共に風が靡く特徴変化（点線で囲われた箇所）が合成音に反映．Result 2 は，口真似の風のフェードイン直後に爆発する表現（点線で囲われた箇所）が合成音に反映．

発音を題材に客観評価と主観評価を行った [18]．客観評価では，参照音（口真似の元になった音源）と合成音の音響埋め込み特徴から，Fréchet Audio Distance（FAD） [19] とコサイン類似度による再現性評価を行い，PronounSE が T-Foley と Stable Audio 2.0 よりも高い性能を示した．主観評価は，参照音への再現性，口真似ニュアンスの合成音への反映性，自然性の観点で，350 人のクラウドワーカーによる聴取評価を行った．その結果，再現性と自然性で PronounSE は T-Foley より高い評価が得られ，製品である Stable Audio 2.0 と同等の評価を得た．

4. 動画への効果音付与タスクによる効果音制作ユーザ評価

2.1 節の AutoSFX では，音源に対する主観評価に加え，AutoSFX のインターフェースを用いたユーザスタディを実施しているが，その使用過程に関する質的な調査は報告されておらず，MultiFoley では，合成性能に関する主観評価しか実施されていない．2.2 節で述べた研究報告では，実際の制作に生成 AI による音響合成技術を使用させた際の質的調査を行っているが，プロのサウンドクリエイターのみで評価を行い，専門知識が必要とされる音響合成技術を評価対象としていた．一方で，非プロや音響制作未経験者による評価も行えば，別視点での有用性や応用性について広く検討・考察が行えると言える．これは，入力に用いるモダリティや，様々な技術・手法によらず確認すべきことであり，口真似音声のみを用いる本研究の手法においても該当する．そこで本研究では，まずは音響制作経験未経験者を対象とし，手法の有用性や応用性について検討すべく，具体的なタスクを設定して従来手法と比較しながら質的調査を行うユーザ評価を実施した．以下，本論文においては，音響制作未経験者によるユーザ評価の内容と結果について述べていく．プロや玄人による評価は別途行う予定である．また，本ユーザ評価では，3.2 節で述べた通り，口真似音声からの音響合成手法について，PronounSE が T-Foley よりも高い性能を，製品である Stable Audio 2.0 と同等の性能が示されたことから，PronounSE をユーザ評価に用いる．音響制作場面での新規性や有用性，効率性を検討する具体的なタスクには，数秒の無音動画に対し効果音を付与するタスクを設定し，以下の 2 手法を実験協力者に課した．

表 1: ユーザ評価タスクで使用する動画

クリップ	内容	長さ [秒]
CLIP 1	砂漠で地面が発破され土埃が舞う映像.	4.0
CLIP 2	室内の射撃場でライフルを発砲する映像.	4.6
CLIP 3	炎が燃え上がり，煙が舞う映像.	6.3
CLIP 4	手りゅう弾が爆発する映像.	4.6
練習用	戦車砲が発射され地面が爆発する映像.	3.7

- **従来手法**： 音素材選択による効果音付与
- **新手法**： 生成系深層学習技術に基づく口真似音声からの音素材制作（PronounSE）による効果音付与

4.1 実験条件

4.1.1 実験で使用する動画

新手法としての PronounSE のモデルは爆発音データセットで学習したものであるため，動画素材においても爆発や銃撃に関するフリー動画素材 5 本（付録 A.1）を用意した．各動画内容を表 1 に示す．

© 2026 Information Processing Society of Japan.

(a) 従来手法のユーザインタフェース．

(b) 新手法のユーザインタフェース．

図 3: （a）従来手法のユーザインタフェースでは，A と B で素材の試聴と配置ができ，C で素材に目印をつけることができる．指定タイミングに自動配置された音源は D の箇所で動画のタイムラインと同期して確認でき，配置オブジェクトをドラッグすることで位置調整が可能である．（b）新手法のユーザインターフェースでは，E で動画を再生しながら口真似音声を録音できる．録音した口真似音声は、F で音声波形として表示され，PronounSE へ入力する箇所をトリミングできる．G で口真似音声を入力し合成処理を実行する．合成音の試聴・配置・目印・位置調整は（a）と同様である．

4.1.2 使用する手法と音源

新手法の PronounSE は，3.2 節で性能評価されたモデルを使用する．従来手法では，（付録 A.2 の）有料素材から各動画に対して，ファイル名に記載されたメタ情報を基に 8 または 9 カテゴリの爆発音を 20 サンプルずつ割り当てた．参加者が従来手法で作業を実施する際，ファイル名から素材探索することが考えられるため，未経験者でも分かり易いようにファイル名を変更した．例えば，「`EXP_FX_Explosion_Destruction_1_ST.wav`」が「爆発音（破壊）A.wav」，「`EXP_FX_Fattener_Body_Blast_Bomb_1_ST.wav`」が「爆発音（爆弾）A.wav」となるようにした．

新手法で用いる PronounSE は，22.05 kHz サンプリングのモノラル信号を合成するため，従来手法で用いる音源も 22.05 kHz サンプリングのモノラル信号に統一した．また，実験で使用する全ての音源は，音量差による影響を減らすためラウドネス正規化により音圧を −16 LUFS に調整した．

4.1.3 制作時の制約

今回の実験では，動画へ付与する効果音を「探索」する行為と「合成」する行為を比較することで PronounSE の有用性・効率性を評価するものであり，それ以外の（音源の加工編集や音響イベント位置への配置調整の）作業工程で参加者に負荷がかかることは目的に沿わない．従って，音源の音量調整や加工編集機能は設けず，予め設定した音響イベント位置に自動配置するように設定し，多少のタイミング修正のみ可能とした．動画への効果音付与に関して参加者が行うタスクは以下の通りである．

- **従来手法**：動画を見て，イメージに合う音を素材一覧から選定し，最終的に音源付与した動画を提出
- **新手法**：動画を見て，口真似からイメージする音を試行錯誤して合成し，最終的に音源付与した動画を提出

実験参加者が最も納得できる音源で動画を提出してもらうために明確な制限時間は設けないこととし，参加者には制作時間の目安として各動画 5 分程度と伝えた．

4.2 実験用 UI（ユーザインタフェース）

図 3a に示す従来手法の UI では、素材一覧に動画毎に割り当てたファイル名が表示され，自由に試聴・配置できるようにした．図 3b に示す新手法の UI では，録音機能と合成機能を設けており，合成音は従来手法と同様，自由に試聴・配置が可能になっている．また，どちらの UI でも画面右上に作業の経過時間を表示している．

4.3 評価

本研究では，両手法に関してインタビューによる定性評価と作業負荷に関する定量評価，制作効率に関する定量評価を行う．

4.3.1 各手法に関する調査

各手法で練習含め動画 5 本分のタスク終了後に，タスク遂行時の精神的労働負荷に関するアンケートと手法に関するインタビューを実施し，両手法終了後に総括インタビューを実施する．

精神的労働負荷に関するアンケートでは，参加者の負担軽減を考慮し，NASA-TLX [20] の簡易版（Raw TLX[*4]）

[*4] 各質問項目の重要度を加味した加重平均を算出しないため，重み

© 2026 Information Processing Society of Japan.

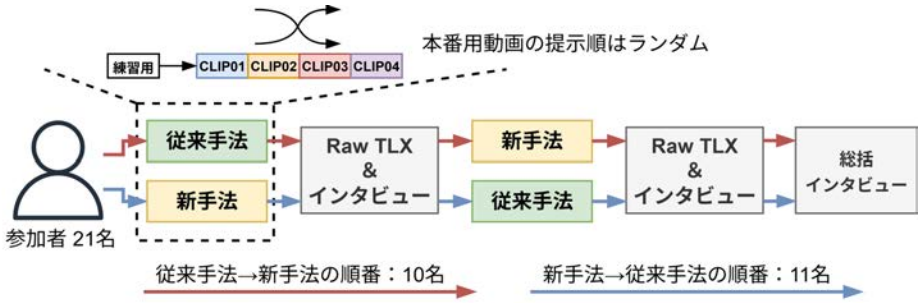


図 4: 実験手順．手法毎に効果音付与タスク，Raw TLX，インタビューを実施し，両手法終了後に総括インタビューを行って実験を終了する．参加者の内 10 名は従来手法から，残りの 11 名は新手法から実施させ，動画の順番は無作為にする．

表 2: Raw TLX の質問項目と内容

項目	質問内容
精神的要求	どの程度頭を使ったか／考える必要があったか（1. ほとんど考える必要がなかった〜7. 非常に強く頭を使った）
身体的要求	どの程度身体的な操作が大変だったか（1. 全く大変でなかった〜7. 非常に強く大変だった）
時間的要求	時間的に追われる感覚がどの程度あったか（1. ほとんどなかった〜7. 常に時間に追われていた）
作業成績	自分の成果はどの程度良かったと思うか（1. 非常に悪い成果〜7. 非常に良い成果）
努力	成果を得るためにどの程度頑張る必要があったか（1. まったくなかった〜7. 最大限頑張った）
不満	苛立ちやストレスはどの程度あったか（1. まったくなかった〜7. 非常に強く感じた）

表 3: 各手法のタスク実施後に行うインタビューの質問文

従来手法

（1） 良かった点，困った点.
（2） どの観点を優先して音を探したか？
（3） 候補をどのように絞り込んだか？
（4） プレビューをやめた判断基準.

新手法

（1） 良かった点，困った点.
（2） どの観点を優先して音を合成したか？
（3） 再合成・差し替えの決め手になった要因.
（4） 合成結果への予測可能性・再現性の印象.

を採用し，表 2 に示す 6 項目を 7 段階リッカート尺度で評価させる．作業成績に関しては，高い値ほど良い成果を意味し，それ以外の項目では，低い値であるほど負荷が少ない（良い）ことを意味するように設定した．

手法に関するインタビューでは，手法毎に少し内容を変え，表 3 の質問を用意し，参加者の回答に応じて追加で質問を行う半構造化インタビューを実施する．

総括インタビューでは，表 4 の質問内容を基に半構造化インタビューを実施する．

付けのための一対比較が不要である．

表 4: 総括インタビューで用いる質問文

（1） 2 つの手法はどのような場面に適していると感じたか？
（2） 実験全体の満足度や印象に残った点.
（3） 制限時間・進行が作業に与えた影響.

4.3.2 制作時間と探索・合成の効率性

各手法の制作効率を定量的に評価するために，タスク実施時の制作コストを動画毎に記録する．両手法で共通する記録情報は，動画提出までにかかった時間（T_{total}）と配置音源の置換回数（N_{replace}）である．これらに加え，従来手法では素材の試聴回数（N_{audition}），新手法では口真似からの爆発音合成回数（N_{gen}）を記録する．

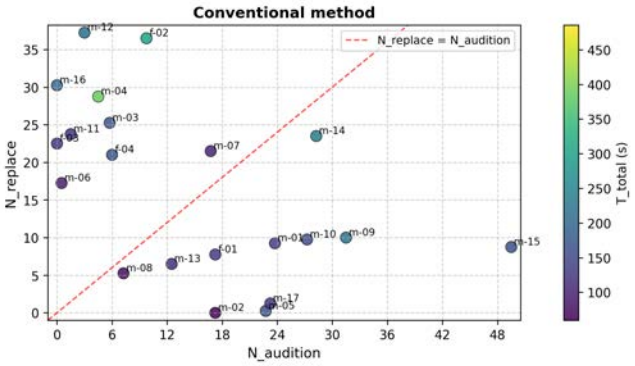
4.4 実験手順

本実験の流れを図 4 に示す．参加者は，本実験の説明を受け，同意書に署名した後，実験前のアンケート（年齢帯，聴覚に関する特記事項，動画編集・オーディオ編集・効果音制作・シンセサイザーの音作りに関する経験の有無，楽器経験年数）に回答した．そして，防音室にて各手法のタスクを実施した．手法毎に UI の説明を行った後に練習用動画で制作の練習を行い，練習完了後に本番用動画 4 つに対して（ランダムな順番で）効果音付与タスクを行った．各手法のタスク実施後に実施した手法に関する Raw TLX と半構造化インタビューを実施し，両手法終了後にまとめのインタビューを実施して実験を終了した．

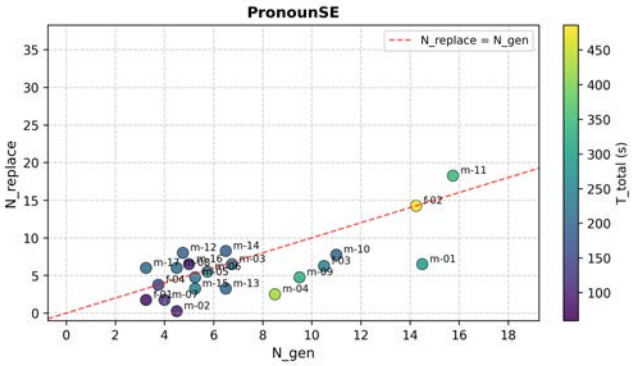
4.5 実験参加者

21 名（女性 4 名：f-01〜f-04，男性 17 名：m-01〜m-17）が実験に参加した．m-16 は 10 年以上の楽器経験者で，（レコーディング，ミキシング，マスタリング含む）制作や音楽ライブでのパブリックアドレッサーの経験者であるが効果音制作に関しては初心者である．m-16 以外の参加者は，動画やオーディオ編集，効果音制作経験，シンセサイザーの音作り経験において初心者である．

© 2026 Information Processing Society of Japan.



(a) 従来手法での音源置換回数と音源試聴回数の散布図.



(b) 新手法での音源置換回数と音源合成回数の散布図.

図 5: 両手法での 4 本の動画に対する操作量平均の散布図.（a）は，従来手法での各参加者の音源試聴回数（N_{audition}）と音源置換回数（N_{replace}）の平均値をそれぞれ横軸，縦軸の座標とした散布図である.（b）は，新手法での各参加者の音源合成回数（N_{gen}）と音源置換回数（N_{replace}）の平均値をそれぞれ横軸，縦軸の座標とした散布図である．参加者の点は，各々の制作時間（T_{total}）の平均を基に色の濃度を変化させている.

表 5: 従来手法と新手法における制作時間（平均 ± SD）

参加者 ID	従来法の制作時間 [秒]	新手法の制作時間 [秒]
f-01	135.00 ± 20.97	79.09 ± 12.42
f-02	303.65 ± 65.78	485.65 ± 213.59
f-03	122.90 ± 30.05	246.30 ± 76.10
f-04	165.73 ± 64.61	146.88 ± 49.30
m-01	137.25 ± 32.28	295.48 ± 57.60
m-02	59.56 ± 26.95	89.31 ± 40.49
m-03	170.00 ± 40.74	233.62 ± 92.88
m-04	386.25 ± 82.51	427.23 ± 71.05
m-05	166.26 ± 48.44	223.37 ± 44.46
m-06	95.10 ± 35.86	247.56 ± 16.83
m-07	100.35 ± 37.30	118.23 ± 25.59
m-08	75.33 ± 7.46	208.14 ± 39.50
m-09	220.49 ± 29.72	321.51 ± 11.66
m-10	157.71 ± 31.20	196.33 ± 59.72
m-11	124.24 ± 31.13	345.93 ± 45.48
m-12	215.07 ± 36.11	196.48 ± 33.77
m-13	115.06 ± 24.46	205.34 ± 49.68
m-14	239.97 ± 48.38	187.10 ± 34.13
m-15	171.00 ± 09.66	244.81 ± 34.23
m-16	196.61 ± 30.40	95.55 ± 42.78
m-17	126.90 ± 16.30	207.61 ± 18.57

5. 実験結果

実験参加者の提出動画は，付録 A.3 から視聴可能である.

5.1 素材選択と素材合成の操作量

図 5 に，両手法での参加者毎の制作時間平均を色の濃淡で表現した操作量平均の散布図を示し，表 5 に参加者毎の両手法の制作時間の平均と標準偏差を示す．図 5a から，従来手法では音源試聴回数が極端に少なく音源置換回数が多い場合とその逆の（音源試聴回数が多く，音源置換回数が極端に少ない）場合，両方の回数が同程度の場合の三つの傾向が確認できる．音源試聴回数が極端に少なく音源置換回数が多い場合は，音源のみを聴取して吟味するのではなく，動画と音源を同期させて音源の候補を絞っていくアプローチであると考えられる．その逆の場合では，音源リストから試聴のみで素材の取捨選択を行った後に動画に配置して候補を絞っていったと考えられる．また，制作アプローチに関わらず大半の参加者が 100〜200 秒程度の制作時間に収まっていたが，m-04 は 400 秒程度かかっていた.

図 5b の新手法の操作量の散布図では，合成回数が 3〜8 回付近で 15 名ほどの参加者がまとまっており，音源置換回数が従来手法に比べて少ない．参加者は全ての合成音を一度は試聴もしくは配置して動画と共に確認すると仮定すると，合成する度にその音を候補にするか否かを決定できるため，吟味する必要がある音源の数が従来手法に比べて少なくなり，従来手法より音源置換回数が大幅に減少したと示唆される．制作時間に関しては，従来手法よりばらつきがあり，ほとんどの参加者が従来手法より長い時間になっていたが，f-01，f-04，m-12，m-14，m-16 の 5 名は従来手法より制作時間が短くなっていた．従来手法より大幅に制作時間が長くなった f-02 は，（5.3 節で詳細を述べる）インタビューにて，「低音の爆発音を作る際に低い声の口真似ができず，合成に苦戦した」，m-01 は「減衰時のニュアンス反映性が低かった」，m-11 は「繰り返して合成することでより良い音が出来上がった印象」と回答しており，3 名とも 15 回程度の合成回数であった．また，従来手法に比べ大幅に制作時間が短くなった m-16 は，「直感的で，素材選択より負荷が少ない」と回答していた．これら操作量とインタビュー内容から，口真似の巧拙・PronounSE の合成品質やニュアンス反映性・音制作に対するこだわりが制作時間の長さや参加者間のばらつきに影響していると示唆される．従って，効果音素材から求める音を探索するのに

© 2026 Information Processing Society of Japan.

表 6: 手法毎の Raw TLX の結果（平均値 ± SD）．作業成績に関しては，値を反転した（低いほど良い）スコアとして示す．

手法	精神的要求	身体的要求	時間的要求	作業成績（スコア反転）	努力	不満
従来手法	**4.14 ± 1.39**	**2.19 ± 1.54**	**1.90 ± 1.14**	2.90 ± 1.18	**4.05 ± 1.53**	**1.81 ± 1.36**
新手法	4.33 ± 1.68	2.57 ± 1.69	2.24 ± 1.34	**2.76 ± 1.18**	4.57 ± 1.57	2.05 ± 1.56

比べ，合成回数や音源置換回数は軽減されると示唆されるが，ユーザ自身が素材を制作する特性上，制作時間はユーザに依存すると考えられる．

5.2 Raw TLX

表 6 に各手法の Raw TLX の結果を示す．従来手法の Raw TLX は作業成績以外の 5 項目で新手法の値より低くなっており，作業成績のみ新手法が上回っていた．

5.3 インタビュー

手法毎のインタビューと総括インタビューの結果に関しては，質問項目毎に回答結果をカテゴリ分けした．付録の表 A·1 と表 A·2 に手法毎のインタビューと総括インタビューの結果のまとめを示す．

5.3.1 従来手法に関するインタビュー

良かった点と困った点で，同じカテゴリの対立意見が得られた．例えば，「探すだけなので手軽」に対して「似た音が多く違いを判断しずらい」や，「ファイル名や尺を目安にできた」に対して「ファイル名への先入観／ファイル名と実際の音に齟齬がある」という意見があげられ，従来手法に対する感想が参加者毎に異なっていた．

音源探索時の優先観点では，映像内の音響イベントとファイル名の整合性を考慮する参加者が 2 名いたが，多くの参加者が映像への没入感を意識していた．

音源候補の絞り込み方法でも，ファイル名のカテゴリや尺を基に選別している参加者がいたが，ほとんどの参加者が映像と音源の一致度や理想とする音との類似度を基準に音源を聴いて候補にするか否かを判断していた．また，20 種類の素材全て聞いて選別する参加者も半数程度いた．

提出に至った経緯に関しては，参加者全員が候補の中から納得できる音源を選択したと回答しており，制限時間を設けていないことが影響していると示唆される．

5.3.2 新手法に関するインタビュー

新手法に関しても，良かった点と困った点で相反する意見が得られた．例えば，オノマトペ的な表現に関して，「表現した口真似よりも尤もらしい音が合成できた」に対し「オノマトペ的な発話だとイメージした音の合成が困難」という意見や，制御性に関して，「繰り返し合成で修正可能／口真似の長さで尺を制御可能」に対し「爆発のリリース部分の制御，連続で音響イベントが発生する音の合成が困難」という意見が挙げられた．また，ユーザが主体となって制作するという特性から創造性に関する利点や，深層学習技術という特性から偶然性に関する利点が挙げられた．

合成時の優先観点では，全ての参加者が映像への没入感を高めるための観点を意識していた．特に声の大小や余韻，アタックの母音子音，リリースを意識して口真似を試行錯誤していた参加者が多かった．

再合成の要因では，合成音がイメージと異なる場合が主な理由であるが，合成数に限りが無いことによる向上心から何度も合成を試みる参加者もいた．また，困った点で挙げられたモデル性能も要因の一つであった．

合成音に対する予測可能性に関しては，「所望する音に近い音が合成できた／尺や声のトーンに応じて雰囲気が変わった」という好印象が大半であった一方で，「口真似ニュアンスを変化させても合成音はあまり変化しなかった」等の悪印象も少し得られた．また，参加者がモデルの結果を基に制御の要領を学習していくような意見や，偶然性に関して，予想通りに合成できる割合が 50%，偶然が 50%との意見も得られた．

5.3.3 総括インタビュー

適している場面の回答で，玄人向きと初心者向きの相反する意見が両手法で得られた．従来手法では，「音素材のメタ情報含め特徴を理解して取捨選択できる経験者であれば効率良く制作可能」という意見に対し，「映像へのイメージが困難な場合，素材を動画に挿入することで強制的に同期性をイメージできる点／候補を選ぶだけ」という意見が得られた．新手法では，ユーザの口真似を用いるということから「制作者の感性に沿って作るという側面でクリエイター向き」という意見に対し，「知識経験問わずイメージした音を作ることが可能なため初心者向き」という意見が得られた．

総合的な感想では，従来手法より新手法の方が効率的という意見が得られた一方で，ニュアンス反映性や口真似表現の難しさなどのネガティブな意見も得られた．

作業時間に関する感想では，新手法に関して，合成回数に限りがないことから制作に区切りをつけるために目安時間の必要性が挙げられた．また，参加者によっては，上手く合成できず焦りを感じていた．一方で，従来手法だと素材を全て聞いてしまうため，口真似からの合成の方が効率が良いという感想も得られた．

5.4 結果のまとめ

操作量に関して従来手法では，参加者によって素材のみを試聴して候補を絞る方法と動画に素材を入れて取捨選

© 2026 Information Processing Society of Japan.

択する方法に偏っており，新手法では合成回数と試聴回数の相関が高くなっていた．新手法の制作時間は全体的に従来手法に比べて増加傾向にあった一方で，動画への音源置換回数が著しく減少しており，候補に入れる音源を吟味する頻度が減ったと示唆される．また，制作時間とインタビュー結果から，所望の音を真似る能力（性別による発音能力や口真似の巧拙）や合成モデルの性能，音制作に対するこだわりに応じて制作時間がばらつくことが示唆された．

それぞれの手法の作業負荷に関しては，Raw TLX の結果から今回の（動画 1 本につき音素材が 20 種類という）条件では従来手法の作業負荷が新手法よりも少ないことが明らかになった．しかし，従来手法の困った点で「素材が増えれば探索コストも増す」という意見も得られており，従来手法の作業負荷は素材数に影響されるため，更に素材の数が増える状況においては，探索コストより新手法の合成による制作コストの方が少なくなるとも考えられる．また，新手法の作業成績が従来手法と比べて良い成果になっており，新手法のインタビューで「創作性」や「音作りとしての満足度」が意見として得られたように，参加者自らが制作したという実感が起因すると考えられる．

制作後のインタビューでは，新手法に関して，繰り返し合成を行うことでより良いものに修正可能な点や，再合成の要因として向上心が挙げられたことから，口真似からの音響合成技術において即座に合成結果が聴けるインタラクティブに制作する AI との共創スタイルが有効であることが示唆された．総括インタビューでは，新手法が素材探索の代替として適しているという意見が半数程度得られた．また，深層学習技術特有の偶然性を活かしたアイデア創発が利点として挙げられた一方で，目安時間の必要性や，時間が限られた場面には適さないとの意見が得られた．これらのことから，口真似からの音響合成技術による効果音制作に対する意見が，2.2 節で述べた関連研究 [17] で報告された三つの知見に従っており，ユーザ主体で操作できる仕組みとして PronounSE が位置づけされるとも考えられる．

6. 議論と課題

実験結果から PronounSE に対して，2.2 節で述べた AI 支援サウンドデザインプロセス，不確実性と曖昧さへの対応，AI に期待するサウンドデザインの 3 つの観点に関する意見が得られた一方で，予測可能性と偶然性を切り替え可能なインタフェースについては課題がある．現状の PronounSE では，口真似ニュアンスが上手く反映される場合とそうでない場合があり，それらの割合や偶然上手く合成できる頻度，要因は未知数である．従って，PronounSE の性能改善に向けて，話者数や参照音の追加を含めたデータセット拡充やモデル構造の検討が必要である．一方，ユーザの経験や性別，口真似の巧拙に依存せず効率的に効果音制作を行える枠組みとして，入力口真似に対するイコライジング処理の有効性も検証する．本実験用に作成した PronounSE のツールは，1 つの入力に対して 1 つの爆発音が合成されるものであった．そこで，口真似 1 つに複数種類のフィルタリング処理を施し，それらを元に合成することで，一回のインタラクションで複数の合成音が得られ，予測可能性と偶然性の両方を考慮できると期待できる．さらに，口真似では説明が困難な音の（種類や素材，含まれる要素などの）メタ的説明変数を加えることで音韻・韻律以外のより詳細な制御が可能になり，口真似による表現が難しい効果音を対象とする際の補助情報になると考える．メタ的説明変数としては，AudioLDM [7] で用いられる音とそのキャプションの対照学習による埋め込み表現が考えられるが，（音の距離，音の発生源，発生源の場所などの）高次元な情報を用いる場合は，学習に使用するデータについて検討する必要がある．

また，PronounSE の精度改善以外に，他の効果音への適用性に関しても課題があり，現状，爆発音のみの検証となっている．従って，新たに爆発音以外の効果音（レーザビームの音や魔法の効果音など）も対象にしたデータセット構築を行い，PronounSE の技術が適用できる効果音の種類を明らかにしていく予定である．一方，非現実な効果音の中には，複数の音が重なった（複数の基本周波数を含む）音も多く存在するため，そのような音に対する口真似は困難であると考えられる．そのような場合は，先述したように，口真似以外の入力としてテキストなどメタ的説明変数で補填する方向性も検討する．

本研究では，口真似からの音響合成技術における音響制作未経験者を対象としたユーザ評価実験を実施したが，2.2 節で述べた関連研究 [17] と同様にプロフェッショナルのクリエイターも対象として，実制作における有効性や新規性を検証するユーザ評価も実施する予定である．

7. おわりに

本研究では，イメージした効果音を真似た音声から効果音を合成する手法 PronounSE を用いて，効果音制作の未経験者を対象としたユーザ評価実験を設計し，21 名の参加者に無音動画への効果音付与タスクを行わせた．そこでは，素材選択による効果音付与と，口真似音声からの音素材合成による効果音付与の 2 手法を課した．その結果からは，口真似音声からの音素材合成の方が，従来の素材選択による効果音付与に比べ，音源を動画に配置して吟味する回数が少なくなった．また，参加者が主体的に音制作できることから作業成績に関する主観評価で従来手法より良い成果を示した．インタビューの結果からは，サウンドデザインにおける生成系深層学習技術による音響合成技術の位置付けとして提唱された 3 観点（AI 支援サウンドデザインプロセス，不確実性と曖昧さへの対応，AI に期待するサウンドデザイン）に沿った意見が得られた．

© 2026 Information Processing Society of Japan.

今後は，口真似音声からの効果音合成技術 PronounSE に対し，ニュアンス反映性や合成音質の向上を目指したモデルの改善や，魔法やレーザービームなどの非現実な効果音合成の試みに取り組む．また，入力音声への前処理エフェクト適用による創発的偶然性を持たせた音響合成や制作効率向上などにも取り組む．さらに，音のメタ情報も学習に用いたモデル構築により，音のニュアンスに加え音の種類や残響感等の特徴も制御可能なモデルの実現を目指す．並行してプロのサウンドクリエイターを対象としたユーザ評価も実施する．

謝辞　本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2157，産業技術総合研究所政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の支援を受けて実施した．

本論文執筆にあたってご指導くださった，東京科学大学 准教授 金崎朝子先生に深謝致します．

参考文献

[1] Ament, V. T.: The Foley Grail: *The Art of Performing Sound for Film*, Routledge, New York, NY, USA (2021).
[2] Viers, R.: *The Sound Effects Bible: How to Create a Record Hollywood Style Sound Effects*, Michael Wiese Productions, Studio City, CA (2008).
[3] Sonnenschein, D.: *Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema*, Michael Wiese Productions, Studio City, CA (2001).
[4] Luo, S., et al.: Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models, Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS), Article 2121, pp. 48855-48876 (2023).
[5] Wang, Y. et al.: AutoSFX: Automatic Sound Effect Generation for Videos, Proceedings of the 32nd ACM International Conference on Multimedia (MM ’ 24), pp. 9923-9932 (2024). DOI:10.1145/3664647.3681109
[6] Zhang, Y., et al.: Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds, arXiv preprint arXiv:2407.01494 (2024).
[7] Liu, H., et al: AudioLDM: Text-to-Audio Generation with Latent Diffusion Models, Proceedings of the 40th International Conference on Machine Learning, pp. 21450-21474 (2023).
[8] Chen, Z., et al.: Video-Guided Foley Sound Generation with Multimodal Controls, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18770-18781 (2025).
[9] Z. Evans., et al.: Long-form music generation with latent diffusion, In: International Society for Music Information Retrieval Conference (2024).
[10] Okamoto, Y., et al.: Onoma-to-wave: Environmental Sound Synthesis from Onomatopoeic Words, APSIPA Transactions on Signal and Information Processing, Vol. 11, p. e13 (online), DOI: 10.1017/ATSIP.2022.13 (2022).
[11] Okamoto, Y., et al.: Environmental Sound Synthesis from Vocal Imitations and Sound Event Labels, ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 411-415, DOI: 10.1109/ICASSP48485.2024.10446906 (2024)
[12] Chung, Y., et al.: T-FOLEY: A Controllable Waveform-Domain Diffusion Model for TemporalEvent-Guided Foley Sound Synthesis, ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2024).
[13] Takizawa, R. and Hirai, S.: PronounSE: SFX Synthesizer from Language-Independent Vocal Mimic Representation, Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST Adjunct ’ 24, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3672539.3686748 (2024).
[14] Kanamori, Y., et al.: RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio, Interspeech 2025, pp. 3155-3159 (2025).
[15] Wen-Chin, H., et al.: The AudioMOS Challenge 2025, arxiv preprint arxiv:2509.01336 (2025).
[16] Okamoto, Y., et al.: Xacle challenge 2026: The first x-to-audio alignment challenge (2025).
[17] Kamath, P. et al.: Sound Designer – Generative AI Interactions: Towards Designing Creative Support Tools for Professional Sound Designers, Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), Article 730, pp. 1 – 17 (2024).
[18] 滝沢 力, 平井 重行, 金崎 朝子, 須田 仁志: 言語非依存な口真似による効果音合成手法 PronounSE の評価, 研究報告音楽情報科学（MUS）, 2025-MUS-143, 51 (2025).
[19] Kilgour, K. et al.: Fréchet Audio Distance: A Referenceree Metric for Evaluating Music Enhancement Algorithms, Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2350-2354 (online), DOI: 10.21437/Interspeech.2019-2219 (2019).
[20] Hart, S. G. et al.: Development of NASA-TLX (Task. Load Index): Results of Empirical and Theoretical Research, In Hancock PA, Meshkati M (Eds.): Human Mental Workload, Elsevier Science Publishers B. V., North-Holland, pp. 139-183 (1988).

付　　録

A.1　実験で使用する動画

練習用

- https://www.pexels.com/ja-jp/video/854281/

本番用

- https://www.pexels.com/ja-jp/video/1350924/
- https://www.pexels.com/ja-jp/video/5243090/
- https://www.pexels.com/ja-jp/video/17807763/
- https://pixabay.com/ja/videos/Γ-ΓΓΓ-ΓΓ-ΓΓ-ΓΓ-15562/

A.2　従来手法で用いた音源のライブラリ

https://sonicwire.com/product/B9642

A.3　実験で提出された動画

実験で得られた計 168 個の動画（= 21 名 × 4 種類の無音動画 × 2 手法）は以下のページから視聴可能である．

https://jinmaro.github.io/sound_video_demo/

© 2026 Information Processing Society of Japan.

表 A·1: 各手法に関するインタビューの結果の概要.

(a) 従来手法				(b) 新手法			
質問項目	カテゴリ	件数	代表的な意見（要約）	質問項目	カテゴリ	件数	代表的な意見（要約）
良かった点	直感性	3	探すだけなので手軽	良かった点	直感性	5	音は言語での指示が困難→口真似が直感的
	素材の豊富さ	13	素材が豊富／音源毎に違いがあった		オノマトペ対応性	5	表現した口真似よりも尤もらしい音が合成できた
	タスクのシンプルさ	5	作業として単純／数が少ないため作業負荷が少ない		タスクのシンプルさ	5	音は画像と異なり見えないので探索が省けて楽
	メタ情報の有用性	7	ファイル名や尺を目安にできた		制御性	14	繰り返し合成で修正可能／口真似の長さで尺を制御可能
	素材の音質	1	複雑な音の品質が高い		創作性	4	無から音を作れる／何度も試せる
					偶然性	2	予期せぬ合成音による思わぬ成果
困った点	迷い	8	似た音が多く違いを判断しずらい	困った点	オノマトペ対応性	1	オノマトペ的な発話だとイメージした音の合成が困難
	素材の少なさ	1	効果音の選択肢が少ない		タスクの難しさ	1	モデルの制御性を把握するのに時間がかかる
	タスクの難しさ	4	素材が増えれば探索コストも増す／イメージに合った音を探すことしかできない		制御の難しさ	14	爆発のリリース部分の制御，連続で音響イベントが発生する音の合成が困難
	メタ情報への先入観	4	ファイル名への先入観／ファイル名と実際の音に齟齬がある		話者性	5	低い音が出せずに低音の制御が困難，想像できても口真似方法が分からない
	柔軟性の欠如	6	音が良くても尺が合わない場合／不要な要素が含まれている場合		モデル性能	2	リリース時の電気的なノイズ／音質が素材に比べ低い
探索の優先観点	映像と音源の一致度	11	尺が合っているか／映像の場所や素材を考慮	合成の優先観点	映像と合成音の一致度	3	シーンへの没入感／映像と合成音の違和感
	ファイル名の妥当性	2	動画内の音響イベントとファイル名の整合／音源の尺		口真似のニュアンス	14	声の大小，余韻／アタックの母音子音／リリース
	爆発音のニュアンス	6	爆発のアタックとその後の余韻		爆発の内容	6	動画内の物，現象から生じそうな音かどうか／何がどのように爆発するか
	リアリティ	2	音が遅延するニュアンス／リアルな音かどうか				
候補の絞り込み方法	映像と音源の一致度	6	音と場面の一致度を基に取捨選択	再合成の要因	向上心	4	良い音が得られていても，しばらく素材作成を試す／より良くするため
	理想の音と類似度	5	優先した観点にマッチするか／直感的に想像する音と合っているか		イメージと異なる場合	12	聴感上爆発の系統が異なる場合／口真似ニュアンスが反映されていない場合
	メタ情報の妥当性	3	ファイル名のカテゴリと尺を基に絞り込み		品質	2	ノイズ等で再生成
	音源同士の比較	9	上から順に聞いて選別／ファイル名に頼らず全て聞く／全て聞いて動画に入れて確認				
判断基準	満足できる音源	全員	候補を全て動画に当てて聞き比べて決定したため／動画と最も合っていると確認できたため	予測可能性・再現性	好印象	15	所望する音に近い音が合成できた／尺や声のトーンに応じて雰囲気が変わった
					悪印象	5	口真似ニュアンスを変化させても合成音はあまり変化しなかった／リリース部分のニュアンス制御性が低い印象
					ユーザも共に学習	2	AI に導かれている感じがした／試行していくにつれ要領をつかんだ
					運要素	2	ランダム要素が強い／予想通りの場合とそうでない場合がある／予想通りの結果が 50 %，偶然が 50 %

表 A·2: 両手法終了後の総括インタビュー結果の概要.

質問項目	カテゴリ	件数	代表的な意見（要約）
従来手法に適している場面	品質を求める場合	4	品質の高い音を当てたい時／理想の音に近づけたい時
	複雑な音が欲しい場合	2	合成では再現が難しい多層的な効果音が欲しい場合
	時間が無い時	1	締め切りがあり，品質が求められる際，短時間で候補が得られる
	プロトタイプ制作時	2	簡単な音入れや下地の作成
	主観が一致するシーンへの音入れ	2	複数人で作業する際に制作者の主観が一致するシーンに対して有効
	玄人向き	3	音素材の特徴を理解して取捨選択できる経験者に向く
	初心者向き	5	映像へのイメージが困難な場合，素材を当てることで映像との同期性をイメージできる／候補を選ぶだけ
新手法に適している場面	素材探しの代替	8	素材を持っていない場合／素材探しが難航する場合
	制作にこだわりたいとき	6	時間に余裕がある場合，試行錯誤できる／納得できる音を得られるため
	新たなアイデア創発	2	ランダム性を活かして，答え（効果音）が定まっていないような映像作品の演出に有効
	玄人向き	2	制作者の感性に沿って作るという側面でクリエイター向き／イメージができ，それを口真似できる場合
	初心者向き	2	知識経験問わずイメージした音を作れるため初心者に向いている
総合的な感想・気づき（新手法に関して）	素材選択より効率的	4	AI による制作の方が楽に感じた／ AI による制作では複数の音での迷いがなかった
	合成音への印象	4	巧拙関わらず，言語的な発音からでも合成できた／リアリティのある爆発音合成ができた
	音作りとしての満足度	4	AI との共創の面白さ／探すだけの単調な作業ではないため，音作りを楽しめる
	ユーザの学習	2	試行錯誤を通して，理想の音に近づける工夫が備わる
	ネガティブな意見	4	ニュアンス反映性が低い／口真似表現が困難／動画によって上手く合成できず，ストレスを感じた／運要素
	気になる点	2	録音環境の影響や、口真似へのエフェクト適用時の挙動についての興味
作業時間に関する感想（新手法に関して）	目安時間の必要性	2	合成を無限に試せるため，区切りをつけるために目安の時間が必要
	効率性の向上	1	素材を全て聞いてしまうので，クリエイターとしては合成する方が効率が良い印象
	制作時の焦り	2	上手く合成できない時に焦りを感じた

© 2026 Information Processing Society of Japan.