東北大学
TOHOKU UNIVERSITY

Press Release

2026 年 11 月 11 日

報道機関　各位

国立大学法人東北大学

音声に合わせて自然に動く 3D デジタルヒューマン AI を開発
― 表情・手指・全身動作をなめらかに生成、姿勢や動作も柔軟 ―

【発表のポイント】

- 分身のアバタやキャラクタ等のデジタルヒューマンの顔の表情、手指、全身の 3D 動作を音声発話から一体的に生成し、長い発話でも動きをなめらかにつなぐ AI「DynaConTalk」（東北大が命名）を開発しました。
- 人の動きを、ゆっくりした姿勢変化、身振り、素早い手指・表情の動きに分けて学習するとともに、話すリズム、声の特徴、言葉の情報を場面ごとに使い分けることで、画一的な身振りや不自然な揺れを抑えました。
- 一般の人を含む 30 名が映像を見比べた評価では、生成したデジタルヒューマンの動きが、従来の技術よりも自然で、話の内容や声のタイミングに合っていると評価されました。学習済みの一つのモデルに、姿勢を変更したり参考となる動作を与えたりすることで、長時間の多様な動作を柔軟に生成することができます。

【概要】

デジタルヒューマンの活用が進む中、より豊かで自然な動作をもっと手軽に表現したいというニーズが高まっています。

こうした中、東北大学電気通信研究所の朱 逸飛 大学院生と香港大学の幸村 琢 教授らによる共同研究グループは、音声から表情・手指・全身の動きを長時間生成し、姿勢や参考動作で制御できる AI「DynaConTalk（ダイナコントーク）」を開発しました。動きを速さごとに分けて学習し、発話のリズム、声、言葉の情報を動的に使い分け、単調な身振りや不自然な揺れを抑えました。

AI 分野で注目される最新の公開データセット「BEAT2」と「SHOW」で、生成動作と実際の動作の分布の近さを測る指標 FGD（フレシェ開始距離）(注 1) を、従来の最良値からそれぞれ約 51%、約 29%低減しました。30 名による評価でも、自然さ、発話内容との一致、同期性、多様性で最も高い平均評価を得ました。

本研究成果は、2026 年 11 月 9 日に出版された ACM Transactions on Graphics に掲載され、12 月 1 日～4 日にマレーシア・クアラルンプールで開催される CG とインタラクティブ技術の国際会議・展示会 SIGGRAPH Asia 2026 の Technical Papers で発表する予定です。

【詳細な説明】

研究の背景

人は言葉だけで話しているわけではありません。体を傾ける、手を広げる、指を動かす、眉や口元を変化させるといった動きが、発話のリズムや内容、感情、話者らしさを補っています。こうした「発話に伴う動作」を自動生成できれば、ゲームや映画のキャラクタ制作、バーチャルヒューマン、オンライン学習、遠隔コミュニケーションなどで、多数の 3D デジタルヒューマンを、音声に合わせて動作させることができます。

一方、全身動作は一様な信号ではありません。胴体の姿勢は比較的ゆっくり変化し、腕や手は言葉を強調する身振りを作り、手指や表情には短く細かな変化が現れます。従来の座標空間での AI による動作生成では、振幅の大きいゆっくりした動きが全体の学習を支配し、短く小さな動きが埋もれて現れにくくなります。また、一つの時点に生じた小さな予測誤差が、再生時には高周波の不自然な動きの揺れとして現れることがあります。

音声側にも同様の問題があります。リズムや音量の変化は頻繁に現れる一方、特定の手振りや身振りにつながる言葉は稀にしか現れません。これら全ての情報を常に同じ比重で混合すると、分かりやすいリズム情報が優先され、発話内容に対応した身振りが弱くなる可能性があります。さらに、長い動作を一定長の区間ごとに独立して生成してから接続すると、姿勢が連続していても、速度や加速度の差が「継ぎ目」として見えてしまいます。

今回の取り組み

DynaConTalk は、動きを速さごとに分け、発話情報を場面ごとに使い分け、過去の動きを生成途中から引き継ぐという三つの考え方を一つの生成過程に統合しました（図 1）。

研究グループは、音響信号の分析などに用いられる定常ウェーブレット変換（注 2）を 3D デジタルヒューマンの動作に適用しました。これにより、各動作を、低速の姿勢変化、中速の身振り、高速の手指・顔表情の詳細、のそれぞれに対応する、時間的に位置の揃った複数の成分に分けます。各成分をそれぞれ正規化して学習することで、小さく短い動きも大きな姿勢変化に埋もれにくくなります。予測誤差も各成分の時間・周波数範囲に限定されるため、フレーム単位の誤差が広い周波数の揺れとして現れることを抑えることができます。

DynaConTalk は、発話のリズム、音響的特徴、音声表現、文字情報、話者・体型情報を利用します。動的ゲーティングネットワーク（注 3）が、リズム・音響・文字情報をどの程度用いるかを、動きの状態や生成段階に応じて調整します。さらに、音声から

提案された動きが現在の姿勢や直前の動きと矛盾しないかを確認する「合意」処理を行い、言葉への反応性と動作の連続性を両立させます。

長い動作を生成する際、DynaConTalk は直前に確定した動きを、その時点での拡散過程（注4）と同じノイズ量に変換して次の区間へ入力します。このため、独立に生成した区間を最後に貼り合わせるのではなく、一つ前の動きを条件としてその次の動きを生成できます。研究グループは、同じ仕組みを特定フレームの姿勢指定（注5）、参考動作のスタイル（注6）、一定姿勢の保持、参考動作への追従にも拡張しました。制御していない部分は、音声と周囲の動きに合わせて AI が補完します。

研究グループは、表情、手指、身体、移動を含む公開3D発話動作データセットBEAT2と SHOW を用いて、DynaConTalk を評価しました。動作の分布の差を測る FGD（値が小さいほど実際の動作に近い）では、BEAT2 の全話者設定で従来最良の 0.356 から 0.175 へ約 51%、SHOW で 2.018 から 1.425 へ約 29%低減し、より自然に近い動きになりました。顔の形状誤差も両データセットで最小となりました。

30 名が 15 種類の同一音声発話の事例を評価したユーザ実験では、DynaConTalk は比較した三つのシステムの中で、動作の自然さ、発話内容との一致、音声との同期、多様性の全てで最も高い平均評価を得ました。身体、手指、表情を別々に良くするだけでなく、それらが一体となった発話動作として改善されたことを示しています。

同一音声発話を用いた比較では、DynaConTalk は例えば、「not considering（検討していない）」に対応した外向きの否定動作を生成しました。従来手法より発話内容に対応した動きが明確で、顔、手、身体が協調しています（図 2）。

動作の長時間生成では、従来の重複区間の置き換えと比べて、区間境界の位置誤差を約 34%、速度の飛びを約 58%、加速度の飛びを 50%低減しました（図 3）。指定姿勢の編集では、座標空間で再生成する方法で起こりやすい急激な加速度変化を抑え、編集の影響が指定範囲外へ広がる度合いも低減しました。

DynaConTalk では、参考動作のスタイルの反映、指定姿勢の保持、参考動作への追従を、一つの学習済みモデルで実現しました（図 4）。

今後の展開

今後、ゲームや映像制作のキャラクタ、バーチャルプレゼンター、オンライン学習、遠隔コミュニケーションなど、音声に合わせて長時間動くデジタルヒューマンへの応用が期待されます。

一方、学習に用いる現在の公開データセットには、大きく歩き回るような動作が少ないため、全身の移動経路は音声だけから十分に決められないという限界があります。研究グループは今後、より大規模な動作データや移動に関する事前知識を取り入れ、表現豊かな全身移動への拡張を目指します。



図 1. 生成動作のウェーブレット分解の例。左は 3D デジタルヒューマンの動作、右上は低速から高速までの 4 つの時間周波数成分（姿勢・身振り・ストローク・細部）、右下は部位ごとの帯域エネルギーを示す。（GIF 動画: https://www.cr-ict.riec.tohoku.ac.jp/wp-content/uploads/2026/11/DynaConTalkPRfig1_wavelet.gif）

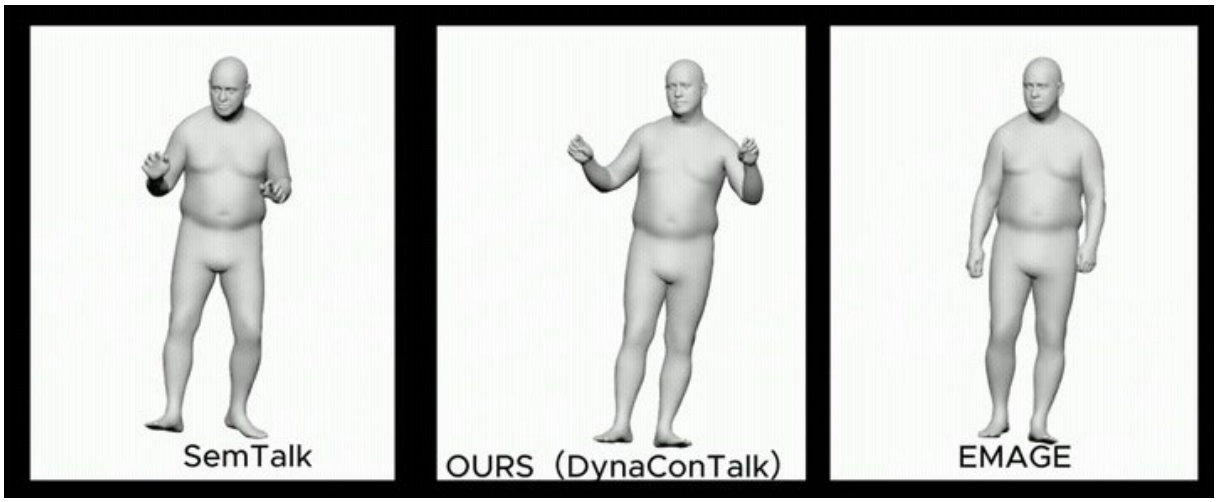


図 2. 同一音声発話に対する全身動作の比較。左から SemTalk、提案手法（DynaConTalk）、EMAGE。提案手法では発話内容に対応した身振りがより明確に現れている。（GIF 動画: https://www.cr-ict.riec.tohoku.ac.jp/wp-content/uploads/2026/11/DynaConTalkPRfig2_compare.gif）

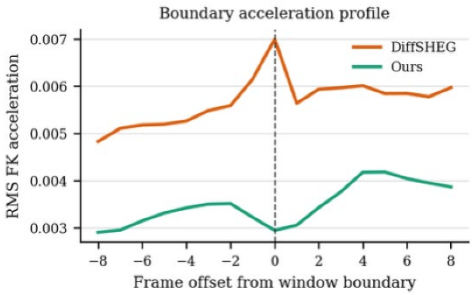


図 3. 長時間生成における区間境界前後の加速度。従来方式（橙）は境界で加速度が急激に増加しているのに対し、DynaConTalk（緑）は変化を抑制できている。

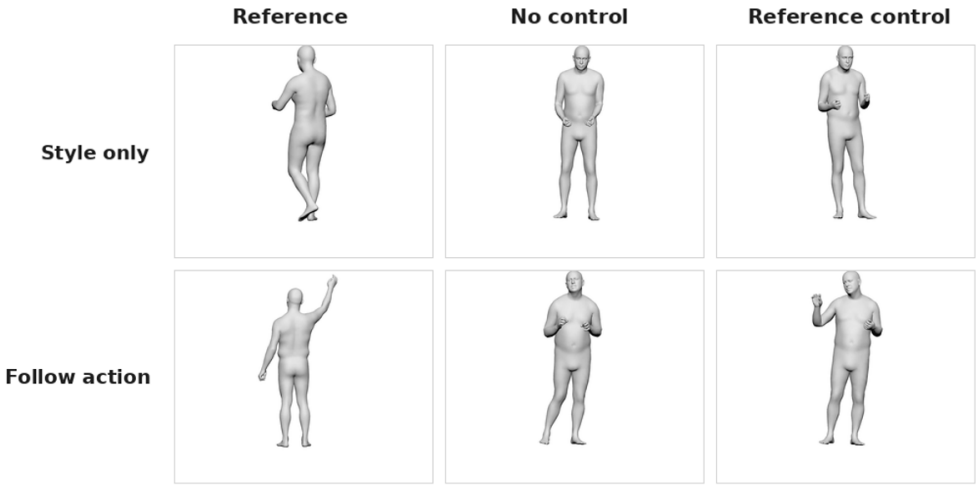


図 4. 参考動作を用いた制御例。左が参考動作、中央が制御なし、右が参考動作で制御した結果。上段はスタイル、下段は動作追従の例を示す。

【用語説明】

注1. FGD（Fréchet Gesture Distance）：生成動作と実際の動作の特徴分布の距離を測る指標。値が小さいほど、生成動作全体の分布が実際の動作に近いことを示す。

注2. 定常ウェーブレット変換：時間上の位置を保ったまま、信号を低速から高速までの複数の成分に分ける解析方法。

注3. 動的ゲーティングネットワーク：人工知能（AI）や深層学習において、入力データに合わせて処理を動的に振り分け・制御する仕組み

注4. 拡散過程：動作データに段階的にノイズを加える過程。生成時には逆にノイズを取り除き、音声などの条件に合う動作を得る。

注5.　姿勢指定：生成 AI で、キャラクタの体勢や手足の向きを指定することを指す。立っているか座っているか、正面向きか斜め向きか、手足や目線の方向など。ポーズ指定。

注6.　参考動作のスタイル：参考動作に含まれる、動きの大きさや速さなどの特徴。本研究では、参考動作のこうした特徴を生成動作に反映する。

【論文情報】

タイトル: DynaConTalk: Wavelet-Constrained Diffusion for Long-Form and Controllable Holistic Co-Speech 3D Motion

著者: Yifei Zhu*, Yangyang Cai, Mingyi Shi, Miao Cheng, Lin Gu, Taku Komura, Yoshifumi Kitamura

*責任著者: 東北大学 電気通信研究所 博士課程学生 朱 逸飛（Yifei Zhu）

掲載誌: ACM Transactions on Graphics, Vol. 45, No. 6, Article 194

DOI: 10.1145/3842553

URL: https://doi.org/10.1145/3842553

口頭発表する国際会議: The 19th ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia 2026) Technical Papers

開催期間・場所: 2026 年 12 月 1 日〜4 日、Kuala Lumpur Convention Centre（マレーシア）

【参考】

デモ動画: https://youtu.be/PUwR3NZxi3s?si=ZIbrtusn0TEAZacR

プロジェクトページ: https://zhuyifeiabcd1.github.io/DynaConTalk/

【問い合わせ先】
（研究に関すること）
東北大学 電気通信研究所 インタラクティブコンテンツ研究室
博士課程学生　朱 逸飛（Yifei Zhu）
Email: zhuyifei.q4@dc.tohoku.ac.jp
教授　北村 喜文（Yoshifumi Kitamura）
Email: kitamura@riec.tohoku.ac.jp
TEL:022−217−5540

（報道に関すること）
東北大学 電気通信研究所 広報室
TEL: 022−217−5427
Email:riec-kohoshitsu@grp.tohoku.ac.jp