Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Synthèse de gestes communicatifs via STARGATE

Louis ABEL[1] Vincent COLOTTE[1] Slim OUNI[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
louis.abel@loria.fr, vincent.colotte@loria.fr, slim.ouni@loria.fr

RÉSUMÉ

La synthèse de gestes liés à la parole est un domaine de recherche en pleine expansion. Cependant, les nouveaux systèmes utilisent souvent des architectures complexes, les rendant souvent inadaptés à leur utilisation dans des agents conversationnels incarnés ou dans d'autres domaines de recherche comme la linguistique, où le lien entre la parole et les gestes est difficile à étudier manuellement. Cet article présente STARGATE, une nouvelle architecture tirant parti de l'autorégression pour fournir des capacités en temps réel, mais aussi des convolutions de graphe couplées à l'attention pour incorporer des connaissances structurelles explicites et permettre une forte compréhension spatiale et temporelle du geste. Nous avons démontré que notre modèle est capable de générer des gestes convaincants en surpassant l'état de l'art dans une étude quantitative, tout en obtenant des scores légèrement meilleurs en termes de cohérence et de crédibilité des gestes générés liés à la parole sur une étude perceptive.

ABSTRACT

Co-Speech gestures synthesis using STARGATE

Co-speech gestures synthesis is a growing field of research. However, new systems often use complex or heavy architecture, making them unsuitable for incorporation into Embodied Conversational Agents (ECAs) or for interpretation in other research fields such as linguistics, where the link between speech and gestures is difficult to research manually. This paper presents STARGATE, a novel architecture for Spatio-Temporal Autoregressive Graph from Audio-Text Embeddings. The model takes advantage of autoregression to provide real-time capabilities, but also graph convolutions coupled with attention to incorporate explicit structural prior knowledge and enable efficient spatial and temporal processing. The model was evaluated against state-of-the-art models in both perceptive and quantitative studies. We demonstrated that our model is capable of generating convincing gestures by outperforming the state-of-the-art in a quantitative study, while achieving slightly better scores in terms of consistency and credibility of the generated gestures related to speech.

MOTS-CLÉS : Apprentissage profond, Synthèse de gestes, Synthèse audiovisuel de la parole.

KEYWORDS: Deep learning, Gestures synthesis, Audiovisual speech synthesis.

1 Introduction

La synthèse de gestes à partir de la parole est un domaine émergent qui a fait l'objet d'une attention particulière au cours des dernières années. Bien que les mécanismes mettant en œuvre la relation entre la production des gestes et la parole sont encore peu connus, des progrès considérables ont été faits dans le développement de techniques de génération de gestes à partir de la parole. L'omniprésence des gestes dans la communication humaine souligne leur importance dans les interactions humaines naturelles. Afin de capturer l'essence des gestes humains et de les intégrer dans des systèmes de communication artificiels, plusieurs chercheurs ont analysé et tenté de classifier les gestes. Au départ, des systèmes basés sur des règles ont été utilisés pour développer des agents conversationnels incarnés

(ECA) (Cassell, 2001), en s'appuyant sur les connaissances des neurosciences et de la linguistique. Les premiers systèmes étaient rudimentaires et souvent incompatibles avec les conclusions émises dans la littérature. L'absence d'un système de classification unifié pour les gestes (par exemple (McNeill, 1992; Kendon, 2004; Boutet, 2008)) et les conclusions disparates concernant la relation entre les gestes et la parole ((So *et al.*, 2009; de Ruiter *et al.*, 2012; Krauss & Hadar, 1999)) au sein de ces cadres ont entravé l'élaboration de règles cohérentes et fiables. Ces dernières années, les approches basées sur les données sont apparues comme une voie prometteuse pour extraire implicitement les modèles et les règles complexes qui régissent la relation entre la parole et le geste. Ces approches utilisent une variété d'architectures, allant des simples autoencodeurs (Takeuchi *et al.*, 2017; Kucherenko *et al.*, 2019) aux autoencodeurs variationnels (VAE) et aux VAE conditionnels (Li *et al.*, 2021; Lu *et al.*, 2023), pour couvrir une plus large distribution de gestes et un meilleur conditionnement à partir de la parole. Les systèmes basés sur la diffusion (Alexanderson *et al.*, 2023; Zhao *et al.*, 2023; Zhang *et al.*, 2023; Deichler *et al.*, 2023) ont également fait l'objet d'une attention particulière, produisant des séquences de gestes de haute qualité. Dans la littérature, le travail de (Alexanderson *et al.*, 2020), StyleGestures, se distingue par son architecture autorégressive innovante utilisant les flux de normalisation (Henter *et al.*, 2020). Les flux de normalisation sont un type spécialisé de réseau de neurones qui permet de modéliser efficacement des distributions complexes. Cette structure de réseau particulière a été largement reconnue comme une référence pour l'évaluation des performances des systèmes de synthèse gestuelle, comme en témoigne son adoption massive dans les recherches ultérieures (Li *et al.*, 2021; Alexanderson *et al.*, 2023; Ao *et al.*, 2022). Il a notamment été sélectionné comme référence pour le GENEA Challenge 2020 (Kucherenko *et al.*, 2021), un défi visant à faire progresser l'état de l'art en matière de synthèse gestuelle. Compte tenu de son architecture autorégressive et de son utilisation intensive en tant que référence, nous avons adopté StyleGestures comme référence de l'état de l'art pour nos comparaisons.

Malgré les recherches approfondies sur la synthèse de gestes liés à la parole, il n'existe que peu d'études sur l'explicabilité et l'interprétabilité de ces méthodes de génération des gestes, indépendamment de leur cohérence ou de leur complexité. Ce manque d'explication pose un défi important dans un domaine qui recherche de nouveaux cadres théoriques pour approfondir la relation complexe entre la parole et le geste. Pour relever ce défi, nous explorons des mécanismes plus simples et plus interprétables, tels que les convolutions de graphe (Kipf & Welling, 2016). Inspirées par leur application réussie dans le domaine de la synthèse de mouvement, sans l'implication de la parole (par exemple, marcher, danser, se battre), les convolutions de graphe sont prometteuses pour améliorer notre compréhension du comportement des réseaux d'apprentissage profond et pour permettre la création de représentations latentes des gestes plus intéressantes.

Inspirés par ces avancées, nous proposons une nouvelle architecture de réseau qui vise à remédier aux limites susmentionnées de la synthèse de gestes. L'architecture que nous proposons vise à atteindre trois objectifs clés :

- Générer des gestes convaincants ;
- Intégrer les convolutions de graphes pour obtenir une représentation latente plus explicite ;
- S'adapter aux applications en temps réel, comme les ECA, notamment en utilisation une architecture autorégressive ;

Dans les sections suivantes, nous décrivons notre architecture et les mécanismes utilisés, suivis d'une évaluation complète de notre modèle face au modèle StyleGestures. Enfin nous discutons des résultats obtenus et concluons en explorant les directions futures potentielles qu'ouvre notre modèle.

2 Méthodes

Nous proposons une nouvelle architecture appelée STARGATE (pour Spatio-Temporal Auto-Regressive Graph from Audio-Text Embeddings), illustré dans la figure 1. Nous suivons une structure encodeur-décodeur, avec une approche autorégressive par blocs. Cela se traduit par un réseau qui prend 3 modalités différentes en entrée :

— **Audio** : Une fenêtre de 1s de parole passée et de 1s de parole future ;
— **Texte** : Une fenêtre de 1s de texte passé et de 1s de texte futur ;
— **Gestes** : Un historique de 1s de gestes passés ;

L'existence d'une fenêtre contextuelle aussi longue est motivée par le fait que les gestes sont une modalité lente, avec une durée moyenne de 1 à 2 secondes selon que le geste se réfère à un seul mot ou à une phrase complète (Ferré, 2010). Chaque modalité dispose d'un encodeur dédié pour produire un espace latent particulier à celle-ci, qui est ensuite fusionné, via une simple concaténation, pour créer une représentation multimodale de la parole/des gestes. Cette représentation est finalement décodée en un bloc de nouvelles poses. Dans ce contexte, une pose définit les rotations de chaque point du corps à un instant donné, et ainsi une série temporelle de poses définit un mouvement. Nous avons choisi d'utiliser une sortie par bloc, c'est-à-dire un groupe de 16 poses. Contrairement à un réseau où chaque étape produit une pose, notre réseau produit bloc par bloc. Outre une implémentation plus efficace en parallélisant partiellement les calculs du bloc courant, cela laisse plus de liberté au réseau pour générer des gestes, l'historique autorégressif ne portant que sur le bloc courant.

2.1 Encodeurs de parole

La parole peut être séparée en deux composantes principales : le contenu acoustique et le contenu linguistique. Le signal acoustique produit pendant la parole contient de nombreuses informations telles que la prosodie (composée de l'énergie, de la hauteur, des rythmes) ou l'état émotionnel. Quant à lui, le contenu linguistique fait partie du signal acoustique, mais avec une représentation phonétique de ce qui a été prononcé, donnant une représentation plus explicite des informations sémantiques contenues dans le texte. Le texte est en effet une source d'information cruciale pour modéliser les gestes iconiques, déictiques et métaphoriques, qui sont tous directement liés au contenu sémantique, tandis que la dernière catégorie de gestes, les gestes dits rythmiques ou de battement, sont quant à eux plutôt liés au signal acoustique (McNeill, 1992). Les deux modalités (acoustique et textuelle) sont donc nécessaires pour générer des gestes dynamiques et significatifs. Dans notre architecture, nous utilisons ces deux modalités par le biais de deux encodeurs convolutionnels (CNN) similaires, mais distincts. Celui dédié à l'audio utilise les MFCCs comme entrées, tandis que celui du texte utilise des *embeddings* BERT (Devlin *et al.*, 2018).

2.2 Encodeur de gestes

Grâce à l'approche autorégressive, les gestes en tant que sorties peuvent être utilisés comme entrée pour la prédiction suivante. Ainsi la troisième modalité d'entrée est le mouvement qui permet de conserver une bonne cohérence pour les trajectoires des gestes, mais aussi de créer par la suite une représentation multimodale parole-gestes. Notre encodeur de mouvement est basé sur les travaux de (Zhou *et al.*, 2021). Le mouvement d'entrée est représenté sous forme d'*exponential map* (Grassia, 1998) pour 17 nœuds, ayant l'avantage d'être une représentation continue de la rotation par rapport aux angles d'Euler, et sont plus compactes que les quaternions (3 valeurs vs 4).

2.2.1 Réseau de neurones de graphe

Pour pouvoir à la fois générer des gestes convaincants et expliquer comment le réseau produit sa représentation latente des gestes, nous avons utilisé plusieurs mécanismes à l'intérieur de notre

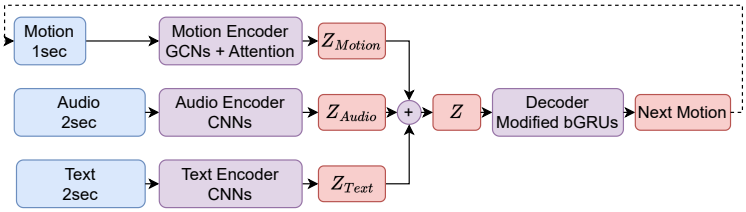

FIGURE 1 – Vue d'ensemble du réseau STARGATE avec sa structure encodeur-décodeur. Notre réseau utilise trois encodeurs distincts pour traiter les trois modalités séparément et un décodeur unique pour générer le mouvement à partir d'une représentation multimodale latente.

encodeur de geste. Le premier est l'utilisation de réseaux de convolution de graphe (GCN) (Kipf & Welling, 2016) au lieu des CNN classiques. Dans le cas d'un graphe, le calcul de la convolution est régi par une matrice d'adjacence, indiquant quels nœuds sont voisins et quel est le poids de chaque connexion. Dans notre cas, nous avons utilisé le bloc ST-GCN (pour Spatio-Temporal Graph Convolution Network) de (Zhou *et al.*, 2021), qui effectue à la fois une transformation spatiale à l'aide d'une convolution de graphe et un traitement temporel à l'aide d'un réseau de convolution temporel (TCN). Celui-ci utilise 3 matrices d'adjacence, l'une définissant des liens de boucle (l'information reste sur le même nœud), une définissant des liens de voisinages directs (lien hanche - colonne par exemple), la dernière définissant des liens de voisinages symétriques (lien main gauche - main droite par exemple). Motivés par le fait que nous voulons créer une nouvelle représentation des gestes, les matrices d'adjacence dans le réseau sont initialisées au début de l'entraînement, comme une connaissance initiale, mais sont modifiables et servent de paramètres pour que le réseau modifie les matrices pour la tâche de synthèse des gestes, en créant potentiellement de nouveaux liens intéressants entre les nœuds.

À notre connaissance, il s'agit du premier travail dans le domaine de la synthèse des gestes lié à la parole qui utilise des convolutions de graphe pour injecter une représentation plus explicite du mouvement.

2.2.2 Mécanisme d'attention

Dans l'implémentation de ST-GCN de (Zhou *et al.*, 2021) et la nôtre, il existe un mécanisme d'auto-attention sur les matrices d'adjacence avant la convolution de graphe. Les données d'entrée passent par un bloc d'auto-attention, afin de créer une "matrice d'attention", une pour chaque matrice d'adjacence. Ces matrices d'attention sont ajoutées aux matrices d'adjacence initiales pour produire ce que nous appelons des "matrices d'adjacence dynamiques". Ceci est motivé par le fait que même si nous laissons le réseau apporter de petites modifications aux matrices d'adjacence pendant l'apprentissage, au moment de l'inférence, elles resteront statiques. Ce mécanisme d'attention permet d'introduire des modifications dynamiques au moment de l'inférence, afin que le réseau accorde plus d'attention à certaines parties du corps pour chaque groupe d'images générées. Ainsi le réseau traite le geste de façon temporelle (TCN) et spatiale (GCN) avec un focus sur les parties du corps fourni par le mécanisme d'attention.

2.3 Décodeur de gestes

Les espaces latents audio, textuels et gestuels sont ensuite combinés pour produire un espace latent multimodal qui est transmis au décodeur de mouvement. Le décodeur, qui se compose de réseaux de neurones récurrents (RNNs) empilés (dans notre cas, les GRUs (Cho *et al.*, 2014)), produira le prochain bloc de pose, ces poses sont ensuite utilisées comme entrée de l'encodeur de mouvement. L'inconvénient majeur de l'autorégression est de ne travailler qu'avec des informations antérieures

FIGURE 2 – Un exemple d'ECA utilisant des mouvements générés par STARGATE. Cet exemple illustre sa capacité à générer différents types de gestes, tels que des gestes iconiques.

et de ne pas pouvoir analyser une séquence complète de gestes. Par conséquent, nous n'avons pas pu utiliser les GRU bidirectionnels pour obtenir une compréhension approfondie de l'ensemble des séquences de gestes. Cependant, motivés par les avantages que cela pourrait apporter et comme nous générons des séquences de poses, nous pouvons utiliser les GRUs bidirectionnelles sur ces séquences partielles. L'introduction de ce mécanisme de bidirectionnalité locale a pour but de permettre au réseau d'apprendre la relation entre les informations passées et futures présentes dans la représentation multimodale.

3 Entrainement

3.1 Corpus : BEAT

Nous avons entraîné tous nos modèles sur le corpus de données BEAT (Liu *et al.*, 2022). Il fournit à la fois des données de grande qualité et en grand volume. Dans notre cas, nous n'avons utilisé qu'un seul locuteur, car nous pensons que l'utilisation de plusieurs locuteurs sans fournir leur identité dans le modèle pourrait entraîner une confusion entre les styles de gestes. Nous disposons ainsi de 4 heures de données réparties entre les ensembles d'entraînement, de validation et de test avec un ratio 90/5/5. Nous n'avons pas utilisé les mouvements des doigts, car ils représentent une énorme quantité de données à traiter et à comprendre par le réseau. Le prétraitement des données audio et de mouvement suit le protocole et le code proposés par StyleGestures (Alexanderson *et al.*, 2020). Nous avons également augmenté les données en utilisant une stratégie miroir (symétrie axiale au niveau de la colonne vertébrale), permettant de doubler les données utilisables. Notre entraînement a duré environ 25 époques, soit 5h sur une machine équipée d'une carte NVIDIA RTX A5000.

3.2 Fonction objectif (*Loss*)

Notre modèle est entraîné pour minimiser deux termes : une fonction de Huber (Huber, 1992) (mixant une distance L1 et L2) sur les *exponential map* et une fonction de Huber sur les positions (dérivé des *exponential map*). Ceci est motivé par le fait que la seule minimisation de l'erreur sur les *exponential map* comme objectif du réseau donnera une importance égale à chaque articulation du squelette, cependant, comme le squelette est intrinsèquement une hiérarchie, nous voulons un contrôle le plus optimal possible des hanches et de la colonne vertébrale, car ils auront un impact sur tous les effecteurs finaux, ce qui nous est fourni en minimisant l'erreur sur les positions, qui sont calculées en traversant toute la hiérarchie, propageant les erreurs potentielles des *exponential map*. La fonction objectif est donc définie comme suit :

$$Loss = \mathcal{H}(r, \hat{r}) + \mathcal{H}(p, \hat{p})$$

avec p et r respectivement, les positions et les exponentielles de la référence, $\hat{p}$ et $\hat{r}$ respectivement les positions et les exponentielles de la génération et $\mathcal{H}$ la fonction de Huber.

	G ?	A ?	T ?	FGD ↓	Performance ↓ [Temps par image ↓] 5s	10s	30s	80s
StyleGestures	✗	✓	✗	14,15	7,76s [90ms]	12,90s [70ms]	31,05s [50ms]	80,07s [26ms]
STARGATE	✓	✓	✓	10,58	6,51s [37ms]	8,31s [17ms]	13,40s [8ms]	23,78s [5ms]
STARGATE Audio uniquement	✓	✓	✗	**8,61**	**3,49s [19ms]**	3,98s [8ms]	6,13s [3ms]	10,68s [2ms]

TABLE 1 – Résultats de la comparaison quantitative utilisant le FGD pour mesurer la qualité des gestes et des temps de traitements de chaque modèle (pour des générations de différentes durées). G, A, T représentent respectivement l'utilisation de graphe, audio et texte. Les valeurs en gras représentent le meilleur modèle. Calculé sur une machine équipé d'une carte NVIDIA RTX A6000 Laptop.

4 Évaluation

4.1 Métriques quantitatives

Dans cette partie, nous présentons l'évaluation de notre modèle ainsi qu'une variante sans texte (et son encodeur) nommée "Audio uniquement". Ces deux modèles sont comparés à StyleGestures.

Distance de gestes de Frechet (FGD). La meilleure tentative d'obtenir une métrique objective pour la synthèse de gestes est inspirée par la "Frechet Inception Distance" (FID) dans la synthèse d'images (Heusel *et al.*, 2017), adaptée dans (Yoon *et al.*, 2020) pour créer la FGD. Cette métrique est donc une distance de Frechet calculée sur un espace latent produit par un réseau d'inception. Nous avons réentraîné le réseau d'inception proposé sur nos données, car notre sortie différait significativement du réseau disponible. Ce réseau est un autoencodeur entraîné à transformer un mouvement d'entrée en une représentation latente compressée, puis recréer le mouvement initial. La métrique est donc basée sur une distance de Frechet calculée entre l'espace latent issu du mouvement de référence et l'espace latent issu du mouvement prédit. L'utilisation d'un réseau d'inception comme évaluateur permet d'obtenir une mesure plus proche de la perception humaine (Yoon *et al.*, 2020).

Comme nous pouvons le voir dans le tableau 1, les deux variantes de STARGATE sont plus performantes que StyleGestures, la variante "Audio uniquement" étant le meilleur modèle en ce qui concerne la FGD.

Performance. Bien que la performance n'est généralement pas une préoccupation majeure lors de la conception d'un modèle de synthèse de gestes, nous avons cherché à développer un réseau capable de fonctionner dans des scénarios en temps réel (par exemple pour les ECAs), en générant des gestes convaincants aussi rapidement que possible. Pour évaluer les performances dans ce contexte, nous avons effectué des tests qui prennent en compte les étapes de prétraitement, étant donné qu'elles peuvent avoir un impact significatif sur la charge de calcul (comme les calculs de BERT). Par conséquent, toutes les durées indiquées sont basées sur une entrée brute audio/texte, avec une taille de batch de 1. Nous indiquons la durée par image, les modèles ayant une sortie de longueur différente.

De même que pour la FGD, les deux variantes de STARGATE sont systématiquement plus rapides que StyleGestures. De plus, tous nos modèles sont capables de fonctionner en temps réel. Nous pouvons également observer que les performances de StyleGestures ne s'améliorent pas avec l'augmentation de la longueur de l'entrée, alors que les modèles STARGATE sont meilleurs pour le traitement de séquences plus longues, en prenant avantage du fait que les entrées audio et texte ne font pas parties de la boucle autorégressive, permettant le calcul en parallèle des 2 latent space sur toute la séquence.

Modèle	Naturel ↑ (Sans son)	Crédibilité ↑	Cohérence ↑
Référence	6,19 ± 0,28	5,27 ± 0,23	5,16 ± 0,23
Permuté	N/A	4,92 ± 0,20	4,77 ± 0,22
StyleGestures	5,97 ± 0,25	4,87 ± 0,22	4,70 ± 0,23
STARGATE	5,89 ± 0,28	5,0 ± 0,20	4,85 ± 0,22

TABLE 2 – Résultats de notre évaluation MOS, nous indiquons la moyenne et un intervalle de confiance à 95% pour chaque aspect.

4.2 Évaluation subjective

Pour mieux évaluer notre modèle, nous avons procédé à une évaluation subjective via un score d'opinion moyen (MOS) afin d'évaluer la qualité globale des gestes générés par notre nouvelle architecture. Nous avons adapté le protocole d'évaluation du GENEA Challenge 2020 (Kucherenko *et al.*, 2021) sur les énoncés pour avoir une compréhension plus claire des aspects évalués pour chacune des questions.

L'évaluation a été divisée en deux parties. La première partie consistait à visionner des vidéos sans audio et à répondre à la question suivante : *"How human-like does the gesture motion appear ?"* La seconde partie consistait à regarder des vidéos avec le son et à répondre à deux questions : *"How credible are the gestures with respect to the speech ?"* et *"How consistent are the gestures with respect to the speech ?"*, ces 2 aspects jugent les aspects sémantiques, le premier sa réalisation (si un geste est bien réalisé), le second sa cohérence (un geste peut ne pas correspondre à ce qui est dit). Les participants devaient évaluer chaque question sur une échelle de 1 à 7. Nous avons évalué quatre systèmes de génération de gestes dans cette étude : Référence (vérité terrain), Permuté (mouvement synthétique avec un audio différent), StyleGestures et STARGATE. Nous avons présenté 30 vidéos pour chaque système, chacune durant 9 secondes et nous avions un total de 25 participants issus de la plateforme Prolific (12 femmes et 13 hommes) parlant un anglais natif. Chaque vidéo est générée en utilisant le modèle 3D du GENEA Challenge 2020 (Kucherenko *et al.*, 2021) sur lequel les poses prédites (rotations) ont été transférées. Les résultats sont présentés dans le tableau 2.

Comme on peut le voir, StyleGestures a obtenu un score légèrement supérieur à celui de notre modèle STARGATE en ce qui concerne l'aspect "naturel" des gestes, mais notre modèle est légèrement meilleur en termes de cohérence et de crédibilité lorsque l'audio est disponible.

4.3 Discussions

Les résultats quantitatifs présentés dans le tableau 1 démontrent que notre modèle surpasse l'état de l'art en termes de FGD. Cela correspond à notre évaluation MOS, où les scores de cohérence et de crédibilité sont légèrement supérieurs à ceux du modèle StyleGestures. Il est intéressant de noter que la variante "Audio uniquement" de notre système présente des valeurs FGD inférieures à celles de notre modèle de base. Ce comportement peut être attribué au fait que notre modèle est capable de générer des gestes non rythmiques convaincants (tels que ceux décrits dans la figure 2), bien qu'en petit nombre. Ces gestes sont nettement plus difficiles à maîtriser et s'écartent souvent de manière significative des gestes de référence, ce qui contribue à des scores FGD plus élevés. Ce n'est pas le cas du modèle StyleGestures où la production de gestes iconiques est absente. Cela peut suggérer que la structure de graphe améliore la compréhension des mouvements et permet au réseau d'établir des liens plus forts entre le texte et le mouvement.

Les performances de notre modèle démontrent des capacités temps réel, l'incorporation du texte dans

le modèle nécessite cependant la génération de séquence supérieure à 10s pour être temps réel. Dans les deux cas, cela implique que l'intégration dans des ECAs peut être effectuée, permettant ainsi une interaction naturelle avec des interfaces homme-machine (les avatars pouvant produire des gestes en temps réel).

Dans le tableau 2, nous avons observé que le modèle StyleGestures recevait des notes plus élevées pour l'aspect "naturel" des gestes lorsqu'il était évalué sans audio. Nous attribuons cette différence à la présence de gestes non rythmiques dans notre modèle, qui ne sont pas toujours produits clairement (comme le montre la figure 2, où le geste "walking in" est "avorté"). Cette incohérence se traduit parfois par un mélange de gestes iconiques et de gestes rythmiques, ce qui donne l'impression d'un manque de naturel lorsque les gestes ne sont pas accompagnés de la parole. Le tableau 2 corrobore également les résultats de recherches antérieures (Kucherenko *et al.*, 2021), où le modèle Permuté présente des évaluations plus élevées que StyleGestures et notre modèle. Nous attribuons cette observation à la forte prévalence des gestes de battement dans l'ensemble de données et dans les gestes générés. Les gestes de battement sont intrinsèquement cohérents et crédibles lorsqu'ils s'alignent sur le rythme de parole. Cela est vrai pour les mouvements avec ou sans permutations de l'audio, car ils proviennent tous deux du même modèle, qui est capable de s'aligner sur le rythme global de l'ensemble de données. Par conséquent, les gestes dans les deux scénarios ont pu convaincre les utilisateurs. En revanche, le mouvement de référence présente des gestes plus cohérents et plus crédibles en raison de la présence de gestes hautement sémantiques. Les résultats de cette évaluation perceptive sont relativement proches. Ce point est certainement lié à la difficulté de la tâche par des évaluateurs novices. Les évaluateurs sont peu sensibles à l'apparition de gestes non synchronisés ou, à la présence ou l'absence de gestes iconiques lors de la parole. Nous comptons évaluer notre modèle grâce à des experts en linguistiques et gestualité, notamment à travers des annotations précises des gestes prédits, attestant la qualité de la synthèse de façon plus objective.

5 Conclusion

Nous avons développé STARGATE, une nouvelle architecture autorégressive par blocs qui utilise trois modalités d'entrée pour construire une représentation latente unifiée de la parole et des gestes, et synthétise des gestes liés à la parole. À notre connaissance, cette architecture est la première à utiliser des convolutions de graphe au lieu de convolutions traditionnelles, en incorporant explicitement une connaissance de la structure du squelette humain. Nous avons évalué cette architecture à l'aide de mesures quantitatives et d'études subjectives, démontrant sa capacité à générer des gestes convaincants, non seulement des gestes de battement, mais aussi des gestes plus complexes tels que des gestes iconiques et métaphoriques. Ces gestes exigent une compréhension approfondie du contenu linguistique, de la parole et de la coordination des mouvements du corps. En outre, nos tests de performance indiquent que notre architecture peut générer des gestes en temps réel, même lorsqu'un calcul d'*embeddings* BERT est incorporé.

Dans nos travaux futurs, nous visons à améliorer l'explicabilité et l'interprétabilité de nos résultats. Nous avons l'intention d'analyser en profondeur le comportement du réseau pour comprendre comment la génération de gestes complexes est déclenchée au sein du réseau et comment la structure du graphe favorise les connexions entre les gestes et la parole. Les matrices d'adjacence dynamiques peuvent nous aider à interpréter visuellement le comportement du réseau. Une telle analyse pourrait fournir une perspective nouvelle et unique sur la corrélation entre la parole et la production de gestes.

Références

ALEXANDERSON S., HENTER G. E., KUCHERENKO T. & BESKOW J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, p. 487–496 : Wiley Online Library.

ALEXANDERSON S., NAGY R., BESKOW J. & HENTER G. E. (2023). Listen, denoise, action ! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, **42**(4), 1–20.

AO T., GAO Q., LOU Y., CHEN B. & LIU L. (2022). Rhythmic gesticulator : Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, **41**(6), 1–19.

BOUTET D. (2008). Une morphologie de la gestualité : structuration articulaire. *Cahiers de linguistique analogique*, (5), 81–115. HAL : hal-00607593.

CASSELL J. (2001). Embodied conversational agents : representation and intelligence in user interfaces. *AI magazine*, **22**(4), 67–67.

CHO K., MERRIENBOER B., GULCEHRE C., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

DE RUITER J. P., BANGERTER A. & DINGS P. (2012). The interplay between gesture and speech in the production of referring expressions : Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, **4**(2), 232–248. DOI : 10.1111/j.1756-8765.2012.01183.x.

DEICHLER A., MEHTA S., ALEXANDERSON S. & BESKOW J. (2023). Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, p. 755–762.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

FERRÉ G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous french. In *Language Resources and Evaluation, Workshop on Multimodal Corpora*, volume 6, p. 86–91.

GRASSIA F. S. (1998). Practical parameterization of rotations using the exponential map. *Journal of graphics tools*, **3**(3), 29–48.

HENTER G. E., ALEXANDERSON S. & BESKOW J. (2020). Moglow : Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, **39**(6), 1–14.

HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B. & HOCHREITER S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, **30**.

HUBER P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics : Methodology and distribution*, p. 492–518. Springer.

KENDON A. (2004). *Gesture : Visible Action as Utterance*. Cambridge ; New York : Cambridge University Press.

KIPF T. N. & WELLING M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv :1609.02907*.

KRAUSS R. M. & HADAR U. (1999). The role of speech-related arm/hand gestures in word retrieval. In *Gesture, Speech, and Sign*, p. 93–116. Oxford University Press. DOI : 10.1093/acprof :oso/9780198524519.003.0006.

KUCHERENKO T., HASEGAWA D., HENTER G. E., KANEKO N. & KJELLSTRÖM H. (2019). Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, p. 97–104.

KUCHERENKO T., JONELL P., YOON Y., WOLFERT P. & HENTER G. E. (2021). A large, crowdsourced evaluation of gesture generation systems on common data : The genea challenge 2020. In *26th international conference on intelligent user interfaces*, p. 11–21.

LI J., KANG D., PEI W., ZHE X., ZHANG Y., HE Z. & BAO L. (2021). Audio2gestures : Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 11293–11302.

LIU H., ZHU Z., IWAMOTO N., PENG Y., LI Z., ZHOU Y., BOZKURT E. & ZHENG B. (2022). Beat : A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, p. 612–630 : Springer.

LU S., YOON Y. & FENG A. (2023). Co-speech gesture synthesis using discrete gesture token learning. *arXiv preprint arXiv :2303.12822*.

MCNEILL D. (1992). *Hand and Mind : What Gestures Reveal about Thought*. Chicago and London : The University of Chicago Press.

SO W. C., KITA S. & GOLDIN-MEADOW S. (2009). Using the hands to identify who does what to whom : Gesture and speech go hand-in-hand. *Cognitive Science*, **33**(1), 115–125. DOI : 10.1111/j.1551-6709.2008.01006.x.

TAKEUCHI K., KUBOTA S., SUZUKI K., HASEGAWA D. & SAKUTA H. (2017). Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*, p. 198–202 : Springer.

YOON Y., CHA B., LEE J.-H., JANG M., LEE J., KIM J. & LEE G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, **39**(6), 1–16.

ZHANG F., JI N., GAO F. & LI Y. (2023). Diffmotion : Speech-driven gesture synthesis using denoising diffusion model. In *International Conference on Multimedia Modeling*, p. 231–242 : Springer.

ZHAO W., HU L. & ZHANG S. (2023). Diffugesture : Generating human gesture from two-person dialogue with diffusion models. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, p. 179–185.

ZHOU K., CHENG Z., SHUM H. P., LI F. W. & LIANG X. (2021). Stgae : Spatial-temporal graph auto-encoder for hand motion denoising. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, p. 41–49 : IEEE.