Résumé de la recherche

Cette étude évalue la capacité de l'IA à traduire la voix en gestes humains, dans des contextes immersifs (robotique sociale, réalité virtuelle, animation virtuelle). Trois modèles audio-to-motion ont été testés pour générer des mouvements naturels, synchronisés avec la parole. Les résultats sont encourageants, mais révèlent des défis persistants: latence, ambiguïté émotionnelle, et adaptation interculturelle. L'enjeu: créer des interactions homme-machine plus sensibles et crédibles.

Contexte et hypothèse

Dans les interfaces homme-machine, la parole seule ne suffit plus.
Les gestes — subtils, illustratifs, expressifs — façonnent notre manière de communiquer.

Aujourd'hui, on s'attend à ce que les avatars ou les agents virtuels réagissent comme nous: avec la voix, mais aussi avec le corps. Générer ces gestes à partir du son, de manière naturelle et en temps réel, reste un défi: les machines doivent apprendre à interpréter non seulement le quoi, mais aussi le comment on parle. Cette recherche explore comment améliorer cette synchronisation voix-geste à l'aide de modèles d'apprentissage profond.

Objectifs

Évaluer les modèles audio-to-motion pour des gestes naturels en temps réel.

Identifier les limites:
latence, émotions, diversité culturelle.

Méthode

Collecte de corpus audio-mouvement

Enregistrement synchronisé de locuteur·rice·s produisant des discours émotionnellement marqués avec capture de mouvement.

Évaluation comparative

Trois architectures de modèles de type diffusion et GAN ont été évaluées.

Analyse experimentale

Quantitative : mesure de la fluidité, de la précision temporelle et de la cohérence gestuelle

Qualitative : retours des utilisateurs et utilisatrices (N=10)

Contexte expérimental

Environnement de simulation de conversation en temps réel dans Unity

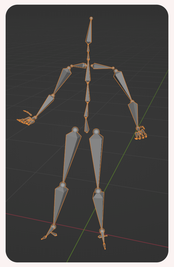
Résultats

Les modèles DiffSHEG, Audio2Gesture et Speech2Gesture génèrent des gestes majoritairement alignés avec l'émotion vocale.

Les gestes produits ne sont pas perçus de la même manière selon les cultures représentées dans les données.

Une latence moyenne de 350 à 500 ms limite la fluidité des interactions en temps réel.

L'ironie, l'humour et d'autres nuances implicites sont difficilement codés de façon fiable.



Perceptions recueillies

85 % des participant·es perçoivent une amélioration notable de l'interaction non verbale surtout pour les gestes illustratifs.

75 % des participant·es trouvent les avatars plus crédibles et expressifs.

40 % des participant·es relèvent des moments d'incohérence ou de désynchronisation, liés à la latence et à un manque de contexte sémantique.

Perspectives et limitations

Perspectives d'amélioration des modèles audio-to-motion

- Réduire la latence par anticipation audio-contextuelle.
- Adapter les gestes aux variations culturelles pour une meilleure expressivité.
- Intégrer le système à des environnements de réalité augmentée.
- Évaluer l'effet à long terme dans des interactions prolongées.

Limitation actuelle

- La taille limitée de l'échantillon (N = 10) restreint la portée statistique des retours des participant·es.

Conclusion

Nos résultats démontrent le potentiel des modèles audio-to-motion à enrichir l'interaction virtuelle par une gestuelle naturelle et adaptée. Toutefois, la latence, l'expressivité émotionnelle et l'adaptation interculturelle limitent encore leur déploiement en temps réel. Mieux maîtrisés, ces systèmes pourraient transformer les usages en formation immersive, robotique sociale ou création numérique.

Colloque de l'ARC dans le cadre du 92e Congrès de l'Acfas, École de technologie supérieure à Montréal, 5 mai 2025,

Learning Individual Styles of Conversational Gestures: https://github.com/JeremyCJM/DiffSHEG
Rhythmic Gesticulator: A Simplified Version: https://github.com/DGeneAI/Audio2Gesture
DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation: https://github.com/amirbar/speech2gesture

CDRIN

Transformer la voix en gestes par l'IA ouvre la voie à des interfaces plus immersives.

Azzahrae El Khiati

Des sons aux gestes : valider l'IA pour des mouvements humains naturels

PARTENAIRE OVA