Édition visuelle pilotée par l'audio à l'aide d'outils de synthèse vocale

Rémi Decelle[1] Serge Miguet[1] Thibault Jaillon[2]

[1] Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205

[2] Mon Petit Placement

{remi.decelle, serge.miguet}@univ-lyon2.fr thibault@monpetitplacement.fr

Résumé

L'édition visuelle ou Facial Reenactment est une tâche complexe qui nécessite une compréhension approfondie de divers outils pour obtenir de bons résultats. Malgré les progrès récents, plusieurs défis subsistent, tels que la synchronisation des lèvres, l'absence de mouvements labiaux pendant les silences et la préservation de l'identité. L'utilisation des spectrogrammes de Mel pour représenter l'audio est limitée pour capturer les nuances et les expressions faciales. Dans notre approche, nous utilisons des outils de synthèse vocale tels que le réseau EnCodec pour fournir des caractéristiques audio et textuelles, extraites à l'aide du modèle CLIP. Ces caractéristiques devraient permettre une meilleure qualité visuelle et une compréhension plus nuancée des mots prononcés et des expressions faciales. Nous avons également construit un jeu de données francophones. Les expériences nous encouragent à approfondir cette approche, qui donne des résultats équivalents à l'état de l'art.

Mots clefs

Deepfake, Animation Faciale, Synthèse Vidéo, Génération conditionnée.

1 INTRODUCTION

L'apprentissage profond permet désormais de faire du doublage visuel automatique (ou édition visuelle ou bien encore *facial reenactment*) en synchronisant précisément les mouvements de lèvres avec l'audio. Cette technologie a de nombreuses applications, notamment dans le montage vidéo, la post-production et la traduction en temps réel, utiles dans divers secteurs tels que le cinéma, les jeux vidéo ou l'éducation.

Bien que des progrès aient été réalisés dans le domaine du doublage visuel automatique, il reste des défis à relever pour obtenir un rendu naturel. Les méthodes précédentes basées sur les repères faciaux et les réseaux neuronaux en 2D génèrent des visages déformés car elles ne parviennent pas à séparer correctement le mouvement de la tête et l'expression faciale. Les méthodes utilisant une représentation de l'animation du visage basée sur des données ont également des difficultés à produire des vidéos de haute qualité. Cependant, l'utilisation d'un modèle facial 3D comme extracteur de caractéristiques dans les méthodes d'apprentissage profond semble donner des résultats plus précis et naturels.

Sur la base des observations précédentes, une nouvelle architecture est proposée pour le doublage visuel automatique, utilisant des méthodes d'édition visuelle basées sur l'*inpainting*. Le fait d'utiliser uniquement des données audio comme données d'entraînement pour la génération peut conduire à des résultats désynchronisés. Pour limiter ce problème, nous proposons d'utiliser des couches telles que la normalisation adaptative des instances (AdaIn) et l'utilisation d'outils Speech-To-Text [1], qui permet d'extraire le texte à partir de l'audio, pour obtenir des mouvements labiaux plus précis et plus naturels. Le schéma de l'architecture de notre méthode s'inspire de [2] et est présenté dans la figure 1. Nous considérons tout d'abord les coefficients de mouvement de la 3DMM (3D Morphable Model) comme une représentation latente du mouvement et de l'expression du visage.

Notre architecture pour le doublage visuel automatique intègre les caractéristiques audio et textuelles extraites à partir des réseaux EnCodec [3] et CLIP [4] respectivement. Elle se compose de trois modules : extraction des caractéristiques 3D de la personne, prédiction du mouvement labial, et augmentation de la résolution de l'image. Les résultats préliminaires montrent que notre méthode est au moins aussi performante que l'état de l'art. Nous proposons également un nouveau jeu de données francophone pour une utilisation concrète et un déploiement en entreprise.

Les principales contributions sont :

- l'utilisation du texte et de l'audio ;
- un nouveau jeu de données francophone ;
- des résultats encourageant à approfondir.

2 État de l'art

Génération vidéo conditionné par l'audio Dans les méthodes qui n'utilisent que l'entrée audio pour la génération, la collecte de données audio et vidéo pour l'entraînement et le ré-entraînement sont généralement nécessaires. RAD-NeRF [5] décompose la représentation du visage dans un espace de grande dimension en trois grilles de caractéristiques à faible dimension, ce qui permet de générer le visage en temps réel. En raison du manque d'in-

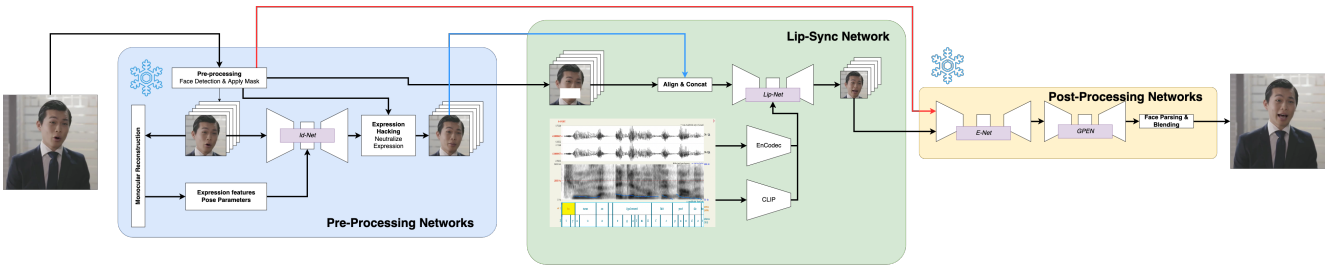

FIGURE 1 – *Architecture générale de la solution proposée.*

formations préalables, ces tâches peinent encore à rendre des expressions réalistes et des mouvements naturels. Plus récemment, SadTalker [6] propose un nouveau système conditionné par l'audio. À partir d'une seule image, en utilisant les coefficients 3D de la personne issus d'une réconstruction monoculaire 3D, il améliore la synchronisation des différents mouvements et la qualité vidéo. VideoReTalking [2] propose d'insérer une modèle 3D neutre pour préserver au mieux l'identité de la personne. Les coefficients 3D obtenus à partir d'une image sont modifiés pour faire paraître la personne sans expressions faciales marquées. Enfin, Hyperlips [7] utilise des hypergraphes pour mettre à jour les poids dans un réseau GAN. Cette méthode est proche des modèles de diffusion conditionnée par de l'audio. La méthode consistant à piloter l'audio avec la source vidéo est la plus poussée, car elle permet d'obtenir des expressions faciales suffisamment réalistes. Cependant, des défauts restent perceptibles visuellement.

Génération vidéo conditionnée par le texte Dans ce domaine, il y a eu plusieurs études sur la synthèse de vidéo à partir d'un texte. *Text-based Mouth Editing* [8] est une technique d'édition d'une vidéo existante avec une nouvelle entrée de texte. Cette méthode effectue une recherche de visèmes pour localiser les segments vidéo dont les mouvements de bouche correspondent au texte édité. Zhang et al. [9] ont proposé *Text2Video* pour animer des vidéos. Cette méthode est basée sur un dictionnaire de positions de phonèmes et ils ont entraîné un GAN pour générer des vidéos à partir de positions de phonèmes interpolées.

3 Notre approche

Notre approche multimodale utilise l'audio et le texte extrait pour prédire la partie inférieure du visage. Pour éviter les résultats indésirables causés par la corrélation entre le mouvement des lèvres et le discours dans les émissions TV, nous avons créé un réseau *lip-sync* qui prend en entrée une image masquée, l'image de la personne neutre, le nouvel audio souhaité et le texte extrait.

3.1 Caractéristiques faciales : 3DMM

Nous effectuons une détection du visage et des *landmarks* pour extraire uniquement le visage à modifier et mieux réintégrer la prédiction dans la vidéo d'origine. Nous extrayons les coefficients 3D de la personne à partir d'une image, que nous modifions pour obtenir un modèle 3D neutre reconverti en une image passée dans le réseau.

3.2 Caractéristiques audio

Plusieurs méthodes utilisent le spectogramme de Mel comme entrée audio. À partir de là, un réseau entraîné extrait des caractéristiques. Cependant, ces caractéristiques audio sont corrélées à la vidéo, limitant l'expressivité.
Pour réduire l'impact de cet entraînement, nous proposons d'utiliser un réseau pré-entraîné pour extraire ces caractéristiques. Pour cela, nous utilisons le réseau pré-entraîné EnCodec [3].
Le réseau encodeur prend le signal audio d'une durée d qui peut-être décrit par une séquence $x \in [-1, 1]^{C \times T}$ avec C le nombre de canaux et $T = d \times f_{sr}$ le nombre d'échantillons pour un taux d'échantillonnage f_{sr}.
Nous prenons 200 ms d'audio pour une frame. L'encodeur produit alors un vecteur caractéristique de taille 15 pour un signal à 24 kHz. Si l'audio d'entrée n'est pas à 24 kHz, il est ré-échantillonné. La configuration choisie correspond à l'encodeur à un bitrate de 24kbps pour un audio reconstruit à 24 kHz. Nous avons choisi un *bitrate* grand afin d'avoir le plus de caractéristiques audio. Dans cette configuration, en donnant un signal de 200 ms, les caractéristiques extraites sont une matrice $M_a \in \mathbb{R}^{15 \times 32}$. En effet, $15 = \frac{24000 * 0.2}{320}$ est le pas de temps sous-échantillonné et 32 est le nombre de quantificateurs du réseau. La matrice de sorti est convertie en un vecteur audio $f_a \in \mathbb{R}^{420 \times 1}$.

3.3 Caractéristiques textuelles

L'un des points clé de notre méthode est d'incorporer du texte, pouvant être extrait à partir de réseau *Speech-To-Text* [1]. Nous utilisons *Montreal-Forcer-Aligner* pour aligner le texte et l'audio [10]. Cela permet d'être sûr que chaque mot correspond bien au segment audio sélectionné. Nous extrayons les caractéristiques textuelles à l'aide du modèle CLIP [4]. Les mots prononcés pendant la séquence sélectionnée sont convertis en un vecteur puis le modèle CLIP génère un vecteur de caractéristiques textuelles $f_t \in \mathbb{R}^{512 \times 1}$.

3.4 Lip-Sync Network

Notre réseau est basé sur un cadre conditionnel reposant sur l'*inpainting*. Nous utilisons les images originales masquées et les caractéristiques audio et textuelles comme condition. La figure 1, montre l'architecture générale de

notre solution. La partie pré-traitement, qui consiste à l'extraction des caractéristiques 3D et leur neutralisation, est gelée. De même, pour les deux réseaux de post-traitement pour augmenter la résolution de l'image générée. Le réseau de génération de mouvement labiale est un réseau auto-encoder proche de celui de [2].

3.5 Fonctions de pertes

Nous avons entraîné le réseau sur une combinaison linéaire des fonctions suivantes : *perceptual loss*, *lip-sync discriminator* pour la qualité visuelle, *audio-visual synchronization* [11], la *L2-loss* sur le logarithme du spectre de Fourier réduit [12] et *L1-loss* dans le domaine spatial.
Soit $\mathcal{I}^{gt}$ l'images de référence, $\mathcal{I}^{lr}$ l'image générée. La *L1-loss* est définie comme :

$$\mathcal{L}_1 = ||\mathcal{I}^{gt} - \mathcal{I}^{lr}||_1$$

La *perceptual loss* est définie par

$$\mathcal{L}_{perc} = \sum_{l \in layers} ||f^l_{vgg}(\mathcal{I}^{gt}) - f^l_{vgg}(\mathcal{I}^{lr})||_2^2$$

où indique f^l_{vgg} le vecteur caractéristiques obtenus à la l-ième couche du réseau VGG19. La *audio-visual synchronization loss* est quant à elle définie par :

$$\mathcal{L}_{sync} = \frac{1}{N} \sum_{i=1}^{N} -\log P_{sync}$$

avec :

$$P_{sync} = \frac{v \cdot a}{\max(||v||_2, ||a||_2)}$$

où v et a sont respectivement les caractéristiques latentes du réseau SyncNet [11] de la vidéo et de l'audio.
Le discriminateur pour la qualité visuelle est donnée par la formule usuelle des GANs.
Enfin, la *L2-loss* du logarithme de la transformée de Fourier réduite est :

$$\mathcal{L}_{spec} = \frac{1}{\hat{H}} \sum_{k=0}^{\hat{H}-1} || \log\left(\tilde{S}(\mathcal{I}^{lr})\right)[k] - \log\left(\tilde{S}(\mathcal{I}^{gt})\right)[k]||_2^2$$

avec $\hat{H} = \frac{H}{\sqrt{2}}$, S est le carré de la magnitudes des composantes de Fourier et

$$\tilde{S}(r) = \frac{1}{2\pi} \int_0^{2\pi} S(r, \theta) \mathrm{d}\theta$$

La fonction de perte est une combinaison linéaire des fonctions introduites :

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{sync} + \lambda_3 \mathcal{L}_{spec} + \lambda_4 \mathcal{L}_{gan}$$

3.6 Post-traitement pour une haute résolution

La sortie du réseau est de taille 96×96 pixels. Afin d'augmenter la qualité et pour l'intégrer au mieux dans la vidéo d'origine, nous procédons à deux augmentations de résolution. Une première qui permet de préserver au mieux l'identité de la personne et une deuxième pour passer à la taille 512×512 pixels.

FIGURE 2 – *Exemples de visage du jeu de données French HDTF que nous avons constitué.*

TABLEAU 1 – *Détail du jeu de données francophone*

# Vidéos	# Personnes	Durée Totale
3529	382+	16H 23min 27s

4 Nouveau jeu de données

Le manque de jeux de données francophones limite la mise en production de ce type d'outils dans le monde francophone. C'est pourquoi nous avons créé un jeu de données francophones, nommé *French HDTF dataset*, en collectant des vidéos publiques de l'administration française des dernières années. Quelques exemples de ce jeu de données sont montrés dans la figure 2. Le jeu de données est résumé dans le tableau 1. Dans le cadre de l'édition visuelle, la détection de *landmarks* et du visage produisent une fenêtre d'au minimum 256×256 pixels. Avec la grande résolution des vidéos d'origine, il est possible d'aller jusqu'à des vidéos de visage de taille 512×512 pixels.

5 Expériences

Nous présentons le jeu de données d'entraînement, fournissons des détails d'implémentation et comparons les résultats quantitatifs avec des méthodes de l'état de l'art.

5.1 Jeu de données

Nous avons utilisé le jeu de données LRS2 [13] avec des vidéos de résolution 160p de différents programmes de la BBC. L'ensemble est traité en utilisant la détection des visages et en redimensionnant l'image d'entrée à 96×96 pixels. Les caractéristiques audio et textuelles sont déjà extraites.

5.2 Critères d'évaluation

Pour évaluer la méthode nous avons retenu les critères largement utilisés dans le domaine, à savoir : *Frechet Inception Distance* (FID), PNSR, SSIM, CPBD et *landmarks metric distance* (LMD). Pour la synchronisation labiale, nous l'évaluons avec le score LSE-D et LSE-C [11].

5.3 Résultats

Le tableau tableau 2 montre les résultats quantitatifs sur le jeu LRS2. Les résultats indiquent que notre approche est similaire à l'état de l'art, notamment pour le FID, CPBD

TABLEAU 2 – *Comparaison avec des méthodes de l'état de l'art sur le jeu LRS2 [13].*

Méthode	FID ↓	CPBD↑	PNSR ↑	SSIM ↑	LMD ↓	LSE-C ↑	LSE-D ↓
Wav2Lip	21.911	0.271	31.794	0.894	1.471	9.641	7.202
MakeItTalk	26.829	0.206	-	-	-	4.937	10.231
ATVGNet	-	-	32.812	0.871	1.984	4.610	8.445
PC-AVS	25.602	0.208	-	-	-	8.959	6.435
VideoReTalking	5.193	0.283	-	-	-	6.519	7.089
IP LAP	-	-	33.281	0.891	1.494	3.435	9.398
SadTalker	22.057	0.335	-	-	-	7.290	7.772
Our	22.432	0.290	32.914	0.820	1.203	5.939	8.193

et PSNR. Nous obtenons le meilleur résultat pour le LMD. Une explication serait la fonction de perte lié au domaine fréquentielle qui n'avait jamais été utilisée jusqu'alors. On peut noter que la méthode SadTalker [6] n'a pas fourni de score pour ce critère, mais il est probable qu'en le calculant les résultats soient meilleurs car prédiction des *landmarks* est incluse dans leur méthode. Le LSE-C est le meilleur pour Wav2Lip [11] car le réseau est entraîné pour ça. Les autres méthodes utilisent ce réseau dans leur fonction de perte de synchronisation audio-vidéo sans ré-entraînement. De même pour LSE-D, Wav2Lip est la meilleure méthode, à l'exception de deux méthodes VideoReTalking et PC-AVS. Cela s'explique par un ré-entraînement du réseau discriminateur permettant de baisser davantage ce score.

6 Conclusions

Ces résultats préliminaires nous encouragent à effectuer des comparaisons avec d'autres jeux de données tels que celui que nous avons constitué et HDTF [14]. Des études approfondies doivent également être menées, notamment une étude qualitative impliquant des personnes et des études dans lesquelles le module audio ou textuel est retiré pour évaluer l'impact respectif sur chaque module.

Références

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, et Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Dans *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[2] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, et Nannan Wang. Videoretalking : Audio-based lip synchronization for talking head video editing in the wild. Dans *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[3] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, et Yossi Adi. High fidelity neural audio compression.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. Dans *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[5] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, et Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. Dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.

[6] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, et Fei Wang. SadTalker : Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. type : article.

[7] Yaosen Chen, Yu Yao, Zhiqiang Li, Wei Wang, Yanru Zhang, Han Yang, et Xuming Wen. Hyperlips : Hyper control lips with high resolution decoder for talking face generation. *arXiv preprint arXiv :2310.05720*, 2023.

[8] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, et Maneesh Agrawala. Text-based editing of talking-head video. 38(4) :1–14.

[9] Sibo Zhang, Jiahong Yuan, Miao Liao, et Liangjun Zhang. Text2video : Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. Dans *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2659–2663. IEEE, 2022.

[10] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, et Morgan Sonderegger. Montreal forced aligner : Trainable text-speech alignment using kaldi. Dans *Interspeech*, volume 2017, pages 498–502, 2017.

[11] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, et C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. Dans *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492. ACM.

[12] Katja Schwarz, Yiyi Liao, et Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34 :18126–18136, 2021.

[13] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, et Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12) :8717–8727, 2018.

[14] Zhimeng Zhang, Lincheng Li, Yu Ding, et Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.