УДК 004.89

МЕТОДИ ОЦІНКИ ЯКОСТІ СИНХРОНІЗАЦІЇ АУДІО ТА ВІДЕО НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Мирошник Ю.Ю., Рябова Н.В.
e-mail: yurii.myroshnyk@nure.ua, nataliya.ryabova@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This thesis analyses the impact of audiovisual synchronisation on the realism of digital avatars used in film dubbing, video conferencing and digital assistants. Accurate lip-audio alignment is highlighted as a key factor in achieving natural, Comparative attention is given to specialised lip-sync models (e.g. SyncNet) and advanced multimodal approaches (e.g. AV-HuBERT). The results show that stable synchronisation significantly improves audience engagement and reduces visual inconsistencies. Practical applications in multilingual face-dubbing illustrate the potential and limitations of current methods.

Аудіовізуальна синхронізація – це точне співвідношення між рухами губ цифрового аватара та відповідним аудіозаписом, що є ключовою умовою для створення реалістичного та правдоподібного мовлення. Візуальне сприйняття людиною навіть незначних розбіжностей між звуком та зображенням суттєво впливає на сприйняття аватара як «живого» та викликає дискомфорт у глядача [1].

Дана робота присвячена дослідженню та аналізу сучасних підходів щодо вирішення проблеми аудіовізуальної синхронізації з метою забезпечення природності та комфортного сприйняття цифрового аватара при його взаємодії з глядачем. Високоточна синхронізація аудіо та відповідних візуальних рухів губ має дві головні цілі:

– досягнення природності мовлення цифрового аватара;

– забезпечення узгодженості та стабільності ідентичності аватара, що передбачає якісний контроль візуальних деталей обличчя та якість фінального зображення [1].

На сьогодні розроблено низку підходів, що дозволяють об'єктивно визначити якість синхронізації. Одним із таких підходів є спеціалізовані моделі, які допомагають кількісно оцінити правильність співвідношення візуальних рухів губ з аудіо доріжкою. В даному аналізі увагу сфокусовано на двох важливих моделях: SyncNet та AV-HuBERT.

SyncNet є вузькоспеціалізованою нейромережею, яка визначає ступінь синхронізації губ цифрового аватара з аудіодоріжкою. Ця модель ефективна завдяки своїй здатності швидко надавати об'єктивну числову оцінку синхронізації. Саме на основі SyncNet було вперше запропоновано підхід, де оцінки синхронізації використовувались у вигляді зворотного зв'язку під час навчання моделей генерації цифрових аватарів. Проте

результати аналізу показують, що SyncNet має певні слабкі сторони, зокрема нестабільність, що призводить до помилок у готових відеоаватарах [2].

Модель AV-HuBERT (Audio-Visual Hidden Unit BERT) використовує інший підхід: замість прямої оцінки синхронізації вона формує спільні аудіовізуальні ознаки з великих наборів даних методом самонавчання. При цьому AV-HuBERT не вимагає явного забезпечення міток синхронності, а опосередковано визначає відповідність рухів губ аудіозаписам. Завдяки цьому модель здатна визначати помилки синхронізації на рівні природних особливостей мовлення. Таким чином, AV-HuBERT потенційно має можливість точніше та ефективніше визначати узгодженість рухів губ і мови цифрових аватарів [3].

В ході аналізу було окремо розглянуто нещодавні результати дослідження [1], де наведено вплив стабілізованої SyncNet моделі, AVSyncNet, на якість генерації цифрових аватарів. Згідно з цими результатами, стабільне та точне визначення синхронності аудіо та відео дозволяє значно покращувати фінальну реалістичність аватарів, при цьому всі інші компоненти залишаються без змін.

Аналіз фактичних прикладів застосування аудіовізуальної синхронізації показав, що дана технологія має значний потенціал у системах мультимовного цифрового перекладу («Face-dubbing») [1]. Створення аватарів, які здатні реалістично вимовляти мову, суттєво збільшує залученість аудиторії та підвищує ефективність передачі інформації, порівняно з традиційними способами перекладу, такими як озвучення чи субтитрування. Розвиток стабільності моделей аудіовізуальної синхронізації дозволить легко покращити фінальну точність генерації, дозволити використання в реальних продуктах за рахунок обробки більш складних сценаріїв, таких як повернення голови під різними ракурсами і умовами освітлення, а в комбінації з моделями синхронного перекладу, що перспективніше у порівнянні з вузькоспеціалізованими моделями.

Список використаних джерел:

1. Yaman D., Eyiokur F. I., Bärmann L., Ekenel H. K., Waibel A. Audio-driven talking face generation with stabilized synchronization loss // European Conference on Computer Vision (ECCV). 2024.

2. Prajwal K. R., Mukhopadhyay R., Namboodiri V. P., Jawahar C. V. A lip sync expert is all you need for speech to lip generation in the wild // Proceedings of the 28th ACM International Conference on Multimedia. 2020. P. 484–492.

3. Anwar M., Shi B., Goswami V., Hsu W.-N., Pino J., Wang C. MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation: arXiv preprint, arXiv:2303.00628. 2023. URL: https://arxiv.org/abs/2303.00628 (дата звернення: 04.03.2025).