computers

MDPI

*Article*

# Seeing the Sound: Multilingual Lip Sync for Real-Time Face-to-Face Translation †

**Amirkia Rafiei Oskooei [1,*], Mehmet S. Aktaş [1,*] and Mustafa Keleş [2]**

[1] Computer Engineering Department, Yildiz Technical University, Istanbul 34320, Turkey
[2] Research and Development Center, Aktif Bank, Istanbul 34394, Turkey; mustafa.keles@aktifbank.com.tr
* Correspondence: amirkia.oskooei@std.yildiz.edu.tr (A.R.O.); aktas@yildiz.edu.tr (M.S.A.)

† This article is a revised and expanded version of a paper entitled "Can One Model Fit All? An Exploration of Wav2Lip's Lip-Syncing Generalizability Across Culturally Distinct Languages" which was presented at The 24th International Conference on Computational Science and Its Applications, Hanoi, Vietnam, July 2024.

**Abstract:** Imagine a future where language is no longer a barrier to real-time conversations, enabling instant and lifelike communication across the globe. As cultural boundaries blur, the demand for seamless multilingual communication has become a critical technological challenge. This paper addresses the lack of robust solutions for real-time face-to-face translation, particularly for low-resource languages, by introducing a comprehensive framework that not only translates language but also replicates voice nuances and synchronized facial expressions. Our research tackles the primary challenge of achieving accurate lip synchronization across culturally diverse languages, filling a significant gap in the literature by evaluating the generalizability of lip sync models beyond English. Specifically, we develop a novel evaluation framework combining quantitative lip sync error metrics and qualitative assessments by human observers. This framework is applied to assess two state-of-the-art lip sync models with different architectures for Turkish, Persian, and Arabic languages, using a newly collected dataset. Based on these findings, we propose and implement a modular system that integrates language-agnostic lip sync models with neural networks to deliver a fully functional face-to-face translation experience. Inference Time Analysis shows this system achieves highly realistic, face-translated talking heads in real time, with a throughput as low as 0.381 s. This transformative framework is primed for deployment in immersive environments such as VR/AR, Metaverse ecosystems, and advanced video conferencing platforms. It offers substantial benefits to developers and businesses aiming to build next-generation multilingual communication systems for diverse applications. While this work focuses on three languages, its modular design allows scalability to additional languages. However, further testing in broader linguistic and cultural contexts is required to confirm its universal applicability, paving the way for a more interconnected and inclusive world where language ceases to hinder human connection.

**Keywords:** talking head generation; lip synchronization; face-to-face translation; computer vision; deep learning; generative AI; human–computer interaction



Academic Editor: Paolo Bellavista

Received: 11 November 2024
Revised: 14 December 2024
Accepted: 26 December 2024
Published: 28 December 2024

**Citation:** Rafiei Oskooei, A.; Aktaş, M.S.; Keleş, M. Seeing the Sound: Multilingual Lip Sync for Real-Time Face-to-Face Translation. *Computers* **2025**, *14*, 7. https://doi.org/10.3390/computers14010007



## 1. Introduction

As we witness the rapid evolution of Artificial Intelligence (AI) technologies, new frontiers are being explored that promise to revolutionize how humans interact with machines. Recent advancements in Machine Learning (ML) and deep neural networks (DNNs) have fundamentally transformed the technological landscape [1–4]. The remarkable ability of deep neural models to learn and generate diverse and human-like content, including text,

maps, codes, tables, images, and music, has propelled Generative AI into a highly active research and application domain. This surge has attracted significant investments and opened up promising avenues for exploration [5–8]. One area that has significantly benefited from these advancements is talking head generation [9]. **Audio-driven talking head generation** involves creating a realistic talking head of a real person or avatar based on given audio. The growing interest in this field is driven by its diverse and valuable applications.

With the rise of virtual worlds, exemplified by the rebranding of Facebook to Meta and the introduction of tools like Apple Vision Pro and Meta Quest, the demand for lifelike talking avatars has become increasingly essential [10–13]. Generating a talking head or face from an audio input (audio-driven talking head generation, see Figure 1) requires the precise synchronization of various facial movements, including the head, eyes, and, most notably, the lips. **Lip synchronization** (lip sync) is a critical component of this process, as the mouth region exhibits the most pronounced motion during facial expressions and is essential for effective communication.

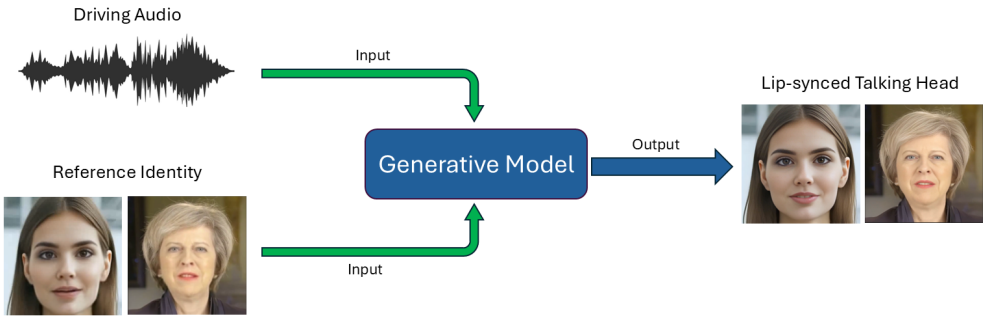


**Figure 1.** A high-level overview of the audio-driven talking head generation concept, in which a generative model takes audio and a Reference Identity as input and generates a talking head as output.

Before the advent of deep learning, lip sync was primarily achieved using techniques such as Visemes [14] and Timecode-based Systems [15], as well as rule-based approaches. However, the significant advancements in deep learning, driven by factors like the dramatic increase in neural network computational efficiency enabled by powerful GPUs and technologies like Nvidia CUDA, coupled with the introduction of diverse deep neural models and architectures, have shifted the research focus towards DNN-based lip sync models [16].

In recent years, a variety of models and architectures for lip synchronization, or talking head generation (which includes lip sync), have been explored. Many of these models claim to be language-agnostic, implying that they can generate highly realistic lip movements for any given audio input, accurately capturing the nuances of lip movements in different languages. With the proliferation of these models, the need for comprehensive performance evaluation becomes increasingly crucial [17]. These models are typically assessed using two primary categories of metrics: (1) image quality metrics, which evaluate the visual fidelity of the generated lip and mouth area, and (2) lip sync error metrics, which measure the temporal accuracy of lip movements in relation to the audio.

Many existing evaluation datasets and methodologies used in research primarily focus on assessing lip sync models in English. Some studies and observations suggest that these models may struggle to generalize to certain input speech languages [18], raising concerns about their robustness and highlighting the need for comprehensive evaluation across diverse linguistic contexts [19]. This research gap motivates our current investigation. We aim to evaluate several state-of-the-art lip sync models with different architectures to assess their performance on Turkish, Persian, and Arabic, languages from distinct origins and

linguistic families that have received limited research attention in this area. In summary, this paper addresses the challenge of evaluating the performance and generalizability of state-of-the-art lip sync models across linguistically diverse languages, a critical gap in existing research that primarily focuses on English.

Beyond the existing research gap, another promising area that warrants further exploration is the application of lip sync for **face translation** or **face-to-face translation**. While the concept of face translation is not entirely new, as first mentioned by [20], a face-to-face translation system can not only translate a spoken utterance into another language but also generate an audio-visual output that includes the speaker's face and synchronized lip movements. In essence, face-to-face translation involves translating input audio into a target language, voice cloning the translated speech, and generating a talking head that is synchronized with the translated speech. Lip sync has historically been a significant challenge in achieving this functionality. However, with recent advancements in lip sync technology, it can now be integrated into this process. We believe such a workflow has immense potential to revolutionize communication across various domains, including education, business, video conferencing, and emerging environments like the Metaverse.

To illustrate the transformative nature of this technology, consider a few potential applications. First, imagine a multilingual video conferencing system where participants with different languages can choose to hear the conversation in their preferred language while maintaining the original audio tone. Face translation not only enables audio translation but also ensures that the speaker's lips move in a way that aligns with the translated speech, creating a more immersive and natural experience. Another example is the use of such a system in VR/AR headsets [21–23]. Imagine a native English speaker from New York exploring the streets of Tokyo while wearing a VR/AR headset equipped with face-to-face translation. When interacting with locals, the system would translate the Japanese speaker's voice into English, preserving the original audio tone and style, and synchronizing the lip movements to match the translated speech. Such systems have the potential to foster a world of barrier-free communication across various domains such as education, healthcare, entertainments, and customer services. In education, lip sync and face-to-face translation could significantly enhance multilingual education, enabling real-time, seamless communication between instructors and students in diverse languages while maintaining the authenticity of the original speaker's expression and tone. In healthcare and medical environments, particularly in global health systems, these technologies could facilitate cross-lingual doctor–patient communication, improving patient care and access to medical services. In entertainment, the film, gaming, and virtual event industries could use lip sync models for more realistic and immersive experiences, allowing for the localization of content while preserving cultural nuances. In customer services, businesses operating in global markets could leverage this technology in customer service applications, offering personalized, real-time interactions in customers' native languages without compromising on tone, body language, or facial expressions. An exciting application recently explored in works such as [24,25] lies at the intersection of large language models (LLMs) and lip sync models. While LLMs are capable of generating diverse outputs, including text [26], code [27], images [28], and even music [29], combining them with lip sync models enables the creation of human-like talking avatars. These avatars not only resemble real humans visually but can also simulate human-like reasoning and behavior. This represents a compelling advancement in the field of human–computer interaction, offering significant opportunities for more natural and intuitive interactions between humans and AI. With the pervasive use of LLMs across various domains, this integration has the potential to redefine how we engage with AI systems.

Given the recent advancements in lip sync models and their potential applications in face-to-face translation systems, along with the gaps in literature that we identified above, several research questions emerge:

**RQ1**: Are lip sync models effective in synchronizing lip movements for languages other than English? How can we assess the generalizability of lip sync models to different languages? What metrics can be employed to evaluate the accuracy of lip sync?

**RQ2**: What is the optimal approach to developing a face-to-face translation system that utilizes lip sync technology? How well would such a face-to-face translation system perform in real-time and multilingual environments?

To address these questions, our paper offers the following contributions. First, we compile a dataset for three languages from distinct origins (Turkish, Persian, and Arabic) and evaluate the effectiveness of two state-of-the-art language-independent lip sync models in generating accurate and natural lip movements, both quantitatively and qualitatively. Second, building upon the advancements in face translation since its initial introduction, we propose a workflow that leverages these lip sync models to achieve this task. Finally, we implement a prototype face translation system based on our proposed workflow and evaluate its performance. In summary, the main objectives of the paper are as follows:

**Objective 1**: To evaluate the effectiveness of two state-of-the-art lip sync models in generating accurate and natural lip movements for Turkish, Persian, and Arabic, across both quantitative and qualitative measures.

**Objective 2**: To propose a novel face-to-face translation workflow that integrates lip sync technology to achieve synchronized audio-visual output.

**Objective 3**: To implement and assess a prototype face-to-face translation system based on the proposed workflow and evaluate its performance in real-time, multilingual settings.

To the best of our knowledge, this study is the first to systematically evaluate lip sync models in Turkish, Persian, and Arabic, providing a multilingual perspective absent from prior works. Additionally, we propose and prototype a novel face-to-face translation workflow, expanding the practical applications of lip sync technology beyond traditional domains.

The structure of this paper is organized to align with the research questions and objectives outlined earlier. In the Related Works section, we explore the evolution of lip sync research, discussing early approaches and the more recent advancements in deep learning architectures such as CNNs, RNNs, GANs, NeRFs, and diffusion models. The Methodology section provides an in-depth explanation of the datasets we have compiled, the lip sync models selected for our study, our performance evaluation approach, and our proposed face-to-face translation workflow that integrates these models. In the Experiments and Results section, we present the outcomes of our evaluations and analyze the performance of our proposed system across multiple languages. Finally, in the Discussion section, we interpret the findings, discuss key conclusions, and highlight the broader implications of our work.

## 2. Related Works

Lip synchronization (lip sync) is an important topic in computer vision and multimedia, gaining increased attention because of its growing impact in various new applications. Research investigates various applications, including audio-to-video generation, live video conferencing, accessibility tools, animation, dubbing, and subtitling, as well as entertainment, satire, deep fakes, and artistic storytelling [30,31]. Before the advent of deep learning, lip sync techniques depended on methods like rule-based approaches and viseme-based techniques [32]. Nonetheless, the rise of deep neural networks (DNNs) represented a significant shift, leading to an increase in methods based on deep learning. In this section, we examine related works on talking head generation and lip synchronization, investigating different architectural methods.

**CNNs:** One significant model is **Speech2Vid** [33], aimed at generating talking face videos from still images and speech segments. It is an encoder–decoder convolutional neural network (CNN) model that uses a joint embedding of the face and audio to generate synthesized talking face video frames. The model is trained with cross-modal self-supervision on unlabeled video data, and incorporates a multi-stream approach to visually re-dub videos by integrating the generated face into original video frames. **X2Face** [34] is a CNN-based neural network model that can control the pose and expression of a given face using another face or modality such as audio or pose codes, and the network is trained in a fully self-supervised manner using a large collection of video data.

**LSTMs:** One of the initial efforts was **ObamaNet** [35], an innovative architecture that employed a time-delayed Long Short-Term Memory (LSTM) network to produce synchronized lip sync videos from text input. However, ObamaNet faced a notable limitation—speaker dependency. This indicated that it could only successfully produce lip sync for one speaker, specifically former US President Barack Obama. Ref. [36] presents an LSTM model that is trained on numerous hours of footage from President Obama's weekly addresses. This model learns to map raw audio features to corresponding mouth shapes. The learned mapping is subsequently utilized to synthesize high-quality mouth textures, which are then composited with 3D pose matching. This process alters the appearance of Obama's speech in a target video to align with the provided audio track. This method is restricted to generating video of a particular individual (Barack Obama) using their audio, and may not apply to other persons. Another approach is **MakeItTalk** [37], which provides a straightforward and realistic audio-driven solution utilizing an LSTM network with a self-attention encoder. This model creates talking head videos that are expressive, starting from just one facial image and using audio as the only input. It achieves this by separating content and speaker information from the audio signal. It effectively manages lip movements and facial dynamics while predicting facial landmarks. This capability enables it to function with different portrait styles, such as artistic paintings and cartoon characters, and it generalizes well to unfamiliar faces and characters.

**GANs, Wav2Lip and Derivatives:** The quest for effective lip synchronization in videos has led to the creation of advanced models, with **Wav2Lip** standing out as a significant advancement. This framework, which is independent of the speaker, is an advancement of **LipGAN** [38]. It addresses the limitations of the previous model by integrating temporal context and utilizing a pre-trained discriminator. LipGAN employs audio and face encoders, a decoder, and a discriminator; however, it faces challenges such as speaker dependency, limited accuracy, and an emphasis on visual artifacts. Wav2Lip tackles these limitations by utilizing five consecutive frames to provide temporal context and using a SyncNet-based pre-trained discriminator [39]. This change from focusing on lip region reconstruction loss to emphasizing audio–lip correspondence leads to a notable enhancement in accuracy, ensuring that lip movements align precisely with the audio, irrespective of the speaker. The success of Wav2Lip led to the creation of many derivative models aimed at enhancing different aspects. Two models based on Wav2Lip were developed to improve the quality of lip sync videos through super-resolution image processing techniques. **Wav2Lip-HD** [40] employs **Real-ESRGAN** [41] to enhance frame quality post-inference, resulting in impressive outcomes. Alternatively, **Wav2Lip-HR** [42] tackles these challenges by emphasizing the development of a robust model instead of enhancements in post-processing. It acknowledges the significance of high-quality training data and utilizes **GFP-GAN** [43], an additional super-resolution model, to improve the training dataset itself. This enables the model to generate higher-quality lip sync videos with precise synchronization, removing the necessity for post-processing and ensuring real-time usability. In contrast to models that aim to improve lip sync quality, **compressed-Wav2Lip** [44] tackles the computational

constraints that impede real-world applications, given that model compression for deep neural networks has gained popularity [45]. The pursuit of more advanced lip synthesis has resulted in the incorporation of attention mechanisms. Ref. [46] exemplifies this with **AttnWav2Lip**, which incorporates an attention module into Wav2Lip, resulting in superior accuracy. This enhanced model guides the network to "pay attention" to specific features and suppress irrelevant ones, further optimizing lip sync quality. Another noteworthy variation is **LPIPS-AttnWav2Lip** [47], which builds upon AttnWav2Lip by incorporating the LPIPS loss function. Similarly, the **CA-Wav2Lip** [48] model deployed attention mechanisms in the speech-to-lip synthesis system by embedding CBAM and Coordinate Attention into the convolution layers. They also utilized the Structural Similarity Index measurement (SSIM) in the loss function of the Visual Quality Discriminator.

**NeRFs:** Recently, researchers have focused on neural radiance fields (NeRFs) as a potential method for lip synchronization. In contrast to GAN-based methods, NeRFs utilize a 3D volume to represent scenes, which allows for more precise modeling of facial geometry. **AD-NeRF** [49] is an audio-driven NeRF model that tackles cross-modal mapping challenges without the need for extra intermediate representations. This enables the smooth editing of talking head videos, such as pose manipulation and background replacement, which is essential for virtual reality applications. AD-NeRF faced challenges with dynamic scenes and motion blur, which resulted in the creation of **DFA-NeRF** [50]. DFA-NeRF incorporates deformation features to manage dynamic changes, employing a hierarchical structure for the effective representation of intricate scenes, which makes it suitable for generating dynamic objects and scenes. Although there have been advancements, NeRF-based methods, such as DFA-NeRF, encounter challenges in achieving lip synchronization when compared to GAN-based techniques. To address this challenge, **LipNeRF** [51] was developed, demonstrating the ability to perform high-quality lip syncing using cinematic content, even with limited video data, showing significant potential for lip synchronization. **NeRF-AD** [52] outperforms LipNeRF in accurately representing the complete range of facial expressions. Although LipNeRF is proficient in lip syncing, it might overlook other elements. NeRF-AD tackles this issue by employing attention-based disentanglement, which divides the face into regions for audio-driven lip movements while maintaining the speaker's identity. **GeneFace++** [53] employs a neural radiance field (NeRF)-based architecture, enhanced for real-time, audio-driven 3D talking face generation. By leveraging an efficient combination of audio-driven lip sync modules and 3D facial rendering, GeneFace++ achieves high fidelity and temporal consistency. This architecture enables real-time processing with generalized audio-lip synchronization, creating stable, lifelike video outputs that are computationally optimized. The **REAL3D-PORTRAIT** [54] model is designed for one-shot 3D talking portrait synthesis, generating a realistic 3D avatar from a single input image. It transforms a single 2D image into a dynamic 3D talking avatar by using neural networks to interpret facial geometry, landmarks, and textures from the image, creating a 3D model. This model then applies neural radiance fields (NeRF) for realistic view synthesis and texture mapping to produce lifelike details from multiple angles. Finally, an animation layer driven by audio or reference video synchronizes lip movements and expressions, enabling the avatar to perform natural, expressive speech animations in a 3D environment.

**Diffs:** One of the earliest models of this category is **DiffTalk** [55], a generative diffusion model that generates talking head videos by using audio signals, reference face images, and facial landmarks as input conditions. The model is based on Latent Diffusion Models, which gradually add noise to an image and then learn to reverse the process to generate new images. The incorporation of the reference face and landmarks allows the model to capture the personality and identity of the target face, enabling high-quality, synchronized

talking head videos that can be generalized to new identities without further training. **Diff2Lip** [56] is an audio-conditioned diffusion-based model for lip synchronization that is trained on the Voxceleb2 dataset of in-the-wild talking face videos. The model outperforms other popular lip synchronization methods like Wav2Lip according to the Fréchet inception distance (FID) metric. The **Diffused Heads** [57] model utilizes an autoregressive diffusion process, allowing it to generate realistic talking face videos from a single identity image and an audio sequence, with added natural head movements and expressions.

**Transformers:** Recent advancements in talking head generation have demonstrated significant progress, with transformer-based architectures playing a pivotal role in enhancing quality, synchronization, and expressiveness. Ref. [58] introduced a **Cross-Attention Transformer** for high-quality talking face generation, emphasizing improved temporal coherence and facial realism. Ref. [59] leveraged a **3D Morphable Model** combined with transformers to achieve audio-driven talking head generation, offering robust facial movement capture while maintaining adaptability to diverse audio inputs. Similarly, Ref. [60] proposed **Styletalk**, a one-shot talking head generation framework that incorporates controllable speaking styles, highlighting its versatility in style manipulation for personalized applications. Ref. [61] contributed with **Vocalist**, focusing on precise audio-visual synchronization for lip and voice alignment, showcasing its strength in dynamic scenarios. Finally, Ref. [62] presented **Faceformer**, a speech-driven 3D facial animation model that effectively balances naturalistic articulation with computational efficiency, pushing the boundaries of real-time performance. These models collectively showcase the breadth of innovation, from stylistic customization to real-time application and robust synchronization, marking significant strides in talking head generation research.

Given the rapid advancements in lip synchronization and talking head generation, interdisciplinary research has emphasized the role of real-time data processing and predictive analytics, particularly in complex systems like IoT and industrial applications. Studies on provenance-aware runtime verification and predictive maintenance architectures for IoT-based systems [63] have shown the importance of reliable, real-time data integration, which can support synchronization tasks by enhancing model adaptability and resilience. In multimedia applications, especially those involving lip synchronization, similar real-time processing needs can be addressed by leveraging predictive analytics to handle varying input conditions, much like in IoT systems where timely, context-aware responses are crucial. Furthermore, work on anomaly detection in business processes [64] highlights the potential for anomaly detection techniques to improve real-time multimedia synchronizations. Applying these methods to lip sync models could enhance the precision of face-to-face translation by identifying inconsistencies in generated outputs.

High-performance computational environments are crucial for the development and deployment of advanced lip synchronization software. These systems require significant computational resources to process audio and visual data in real time while maintaining high accuracy and synchronization quality. Previous studies, such as Pierce et al. (2008) [65] and Aktas et al. (2007) [66], highlighted the integration of web services with computational grids for geophysical and seismic applications, showcasing the ability of these environments to handle data-intensive tasks. Similarly, Fox et al. (2009) [67] explored algorithms optimized for grid environments, emphasizing the potential of distributed systems to enhance computational efficiency. Nacar et al. (2007) [68] and Aydin et al. (2008) [69] demonstrated the scalability and flexibility of grid-based platforms for material science and geographical information systems, respectively. These studies underscore the role of computational frameworks in enabling complex data processing workflows and real-time applications. In the context of this study, the integration of advanced computational models with efficient computational environments is essential to ensure real-time performance and

scalability across multiple languages. While prior works focus on general-purpose grid computing and distributed data processing frameworks, our research diverges by optimizing lip sync models for low-latency inference in real-time face-to-face translation systems. By leveraging cutting-edge deep learning architectures, such as Wav2Lip and GeneFace++, this study evaluates their performance in Turkish, Persian, and Arabic, demonstrating their applicability in multilingual and culturally diverse settings.

Furthermore, multimodal learning plays a crucial role in improving audio-visual synchronization by integrating additional input types, such as emotion cues, gestures, or contextual information, alongside facial landmarks and audio. By leveraging multiple modalities, models can better capture the complexity of human communication, including non-verbal cues like facial expressions, body language, and emotional tone. This integration enhances the accuracy and naturalness of synchronized outputs, enabling more lifelike and expressive talking head generation. Recent research suggests that combining these diverse inputs leads to more robust and contextually aware synchronization models, with potential applications in virtual reality, telecommunication, and human–computer interaction [70,71].

Effective data representation and embedding are crucial for lip synchronization studies, as they enable models to capture and process the complex temporal, linguistic, and structural patterns essential for accurate and realistic audio-visual alignment. Recent advancements in data representation and embedding methods have significantly enhanced the ability to process and analyze complex datasets. Uygun et al. (2020) [72] explored scalable graph data processing techniques for user interface testing, addressing the challenges of handling large-scale graph datasets. Olmezogullari and Aktas (2020, 2022) [73,74] developed embedding-based methodologies, such as Pattern2Vec, to represent sequential clickstream data for understanding user behavior. These works highlight the critical role of embedding techniques in capturing structural and sequential patterns, enabling efficient data analysis and predictive modeling. In the context of this study, the use of embeddings aligns with the need to represent complex temporal and linguistic features, which are integral to achieving accurate lip synchronization in multilingual applications. While these studies focus on effective data representation and the scalability of processing frameworks, our research diverges by integrating lip sync models with neural embeddings to address linguistic diversity in real-time face-to-face translation. Specifically, we evaluate the performance of state-of-the-art lip sync models across Turkish, Persian, and Arabic, showcasing their applicability beyond traditional embedding contexts and filling a critical gap in multilingual generative AI.

Ensuring software quality and thorough testing are critical for lip synchronization studies, as these factors directly impact the reliability, accuracy, and usability of the developed systems. High-quality lip sync software must consistently produce natural and synchronized results across diverse scenarios, which requires rigorous testing to identify and address potential issues related to functionality, performance, and user experience. Kapdan et al. (2014) [75] explored software quality by focusing on code clone detection, demonstrating the role of metrics in improving software maintainability and reliability. Sahinoglu et al. (2015) [76] conducted a systematic mapping study on mobile application verification, emphasizing the need for structured and comprehensive testing approaches to ensure robust and dependable software systems. These studies underscore the importance of quality assurance practices in the software development lifecycle, particularly for applications with complex requirements. In the context of this study, robust testing methodologies and quality assurance are vital to evaluate and enhance the performance of lip sync models, particularly when applied to multilingual real-time face-to-face translation systems. While prior works address general software testing and quality metrics, our research differentiates itself by integrating quantitative metrics such as lip sync error

(LSE-D, LSE-C) and qualitative assessments, specifically tailored for evaluating lip sync models in Turkish, Persian, and Arabic. This ensures not only the technical accuracy of the models but also their cultural and linguistic appropriateness in real-world applications.

In summary, the rapid advancements in deep learning techniques have significantly transformed the field of lip synchronization and talking head generation, leading to more realistic and expressive video outputs. The diversity of approaches—ranging from CNNs and LSTMs to GANs, NeRFs, and diffusion models—demonstrates the rich landscape of research aimed at overcoming challenges in this domain. While considerable progress has been made in improving the quality, efficiency, and generalizability of lip sync models, ongoing research must address remaining limitations, such as computational demands and the ability to effectively handle diverse and multilingual contexts.

## 3. Methodology

This section details the methodology employed in our research, structured into four key contributions. The methodology begins with an overview and selection of two state-of-the-art lip sync models with different architectures. Next, we curate a multilingual dataset comprising speech (audio) in three different languages other than English, ensuring diverse linguistic representation. Following this, we design a comprehensive evaluation framework, incorporating both quantitative metrics (e.g., LSE) and qualitative assessments to analyze the performance of the selected lip sync models in generating realistic and synchronized talking heads using our dataset. Building on these findings, we propose a face-to-face translation workflow that leverages recent advancements in deep learning. Finally, we implement and validate this workflow through the development of a functional prototype for a face-to-face translation system.

Our methodology is guided by the hypothesis that the selected models, due to their claimed language-independent nature, can generate lip-synchronized talking heads across various languages. This capability is critical for enabling their effective use in multilingual settings, such as a face-to-face translation system.

### *3.1. Model Selection*

As outlined in Section 2, various models have been proposed for lip synchronization. In this study, we focus on evaluating the performance of these models in terms of lip sync accuracy across multiple languages and system efficiency (inference time), which are key factors influencing the practical applications of this technology. To ensure a comprehensive and fair evaluation, we select one state-of-the-art model known for its lip sync accuracy (a GAN model) and one top model renowned for its stable real-time performance (a NeRF model).

Our first selected model is Wav2Lip [77], a GAN-based model. The underlying architecture of the Wav2Lip model revolves around the concept of generating accurate lip sync by learning from a well-trained lip sync expert, as precisely outlined in the original paper. The model comprises a generator responsible for generating video frames, which utilizes two encoders and a decoder. The video encoder (V) processes visual information, extracting a latent representation of the face and mouth region, while the audio encoder (A) analyzes the mel-spectrogram, capturing audio features. These latent representations from both encoders (V and A) are combined to generate new video frames with lip movements synchronized to the audio. During training, the generator minimizes the **L1** reconstruction loss between the generated frames and ground-truth frames. To further refine the model and penalize the generator architecture, a pre-trained lip sync expert, derived from a modified **SyncNet** with different loss function trained on the **LRS2** dataset, is employed. Finally, to enhance overall visual quality and mitigate issues such as blurry generated lips, a visual quality discriminator with binary cross-entropy loss is utilized.

The second selected model is GeneFace++ [53], a NeRF-based model. The underlying architecture of the GeneFace++ model focuses on achieving real-time, generalized audio-lip synchronization through a multi-stage approach. The model consists of three main components: a pitch-aware audio-to-motion module, a landmark refinement method, and a motion-to-video generator. The pitch-aware audio-to-motion module uses two encoders—HuBERT and a pitch encoder—to extract audio features and pitch contours from the input audio. These features are combined to predict the facial landmarks, which represent the movement of the mouth and facial expressions over time. To further refine the predicted facial landmarks and ensure consistency, GeneFace++ employs a Landmark Locally Linear Embedding (LLE) method. This method adjusts the landmarks by projecting them into the target domain, making them compatible with the training data, thus improving robustness against out-of-domain (OOD) inputs. The final component is the motion-to-video generator, which utilizes a grid-based neural radiance field (NeRF) to render high-quality 3D-aware video frames. This generator is conditioned on the refined landmarks and can render realistic facial movements. The generator optimizes the L2 loss between the rendered video frames and the ground-truth frames during training. Additionally, the lightweight design of the grid-based NeRF allows for faster training and real-time inference, making GeneFace++ suitable for practical applications.

By selecting two language-independent models with distinct architectures, each possessing its own strengths, we can conduct a comprehensive evaluation of their performance on our dataset and effectively utilize them in a multilingual setup, such as a face-to-face translation system.

### *3.2. Data Collection*

For our evaluation, we collected a dataset that encompasses three languages: **Turkish**, **Persian**, and **Arabic**. The selection of languages for this study was carefully guided by the need for diversity and relevance in evaluating the language independence of lip sync models. Each chosen language belongs to a distinct language family—Turkic, Indo-Iranian, and Semitic—ensuring linguistic diversity and a robust testing ground for the models. These languages, despite being widely spoken globally, have received limited attention in existing lip sync research. Addressing this gap is critical, as the Middle East region, where these languages are predominantly spoken, represents a rapidly expanding market characterized by increasing technological adoption and digital content creation. Additionally, the geographical proximity of this study to the region ensures access to native speakers for evaluation, providing culturally and linguistically accurate insights into the models' performance. This combination of linguistic diversity, regional relevance, and the ability to involve native speakers strengthens the rationale for selecting these languages and underscores the broader applicability of our research in real-world multilingual settings. Evaluating Wav2Lip and GeneFace++ on these languages can facilitate the development of localized applications, enhance accessibility for speakers of these languages, and drive innovation in communication technologies.

We chose the **MediaSpeech** [78] dataset for Turkish and Arabic, which consists of brief speech segments that were automatically extracted from media videos found on YouTube. We selected the **Persian Speech Corpus** [79] for the Persian language, which was recorded in a Tehrani accent with a professional studio setup. The evaluation dataset includes approximately one hour of audio for each language, which has been randomly chosen from the original datasets, ensuring a diverse and representative sample for evaluation. It is important to highlight that these audio samples are absent from the training datasets of the models we have chosen, which makes them appropriate for evaluating the generalizability of these models. Table 1 provides a summary of our dataset.

**Table 1.** An overview of our dataset, collected for three culturally distinct languages.

| Language | Origin | Source | Length (min) |
|---|---|---|---|
| Turkish | Turkic | MediaSpeech | 67 |
| Arabic | Semitic | MediaSpeech | 54 |
| Persian | Indo-Iranian | Persian Speech Corpus | 59 |

While this dataset is well suited for our study, it is important to acknowledge the limited availability of digital resources for these languages compared to widely researched languages such as English. This scarcity poses challenges in assembling large-scale datasets. However, we are confident that the quality and diversity of our dataset adequately address the objectives of this research, particularly in evaluating the language-independence claims of the selected lip sync models. The curated dataset not only represents linguistic diversity but also aligns with the specific requirements of this study, ensuring its relevance and robustness for testing. (All datasets and media content used in this study were sourced from publicly available repositories with appropriate citations and in accordance with their licensing terms. Ethical considerations, including copyright and privacy compliance, were strictly followed to ensure the responsible use of resources).

### *3.3. Video Generation*

After selecting the suitable models and preparing a dataset in multiple languages, we have to generate the lip synced videos (talking heads) using these models. Audio-driven talking head generation requires two inputs for video generation: **(1) Driving Audio:** This audio track determines the speech patterns for the talking face. In our experiment, we utilized approximately one hour of speech for each language, sourced from the dataset described in Section 3.2. **(2) Target Image/Video:** This input provides the visual foundation for which the lip movements will be generated. For this, we utilized a sample 3-min video of Theresa May's speech with the size of 512 × 512 pixels, provided by the authors of GeneFace++.

For both models, we configured them to generate videos with 25 fps and we employed an Nvidia A100 (40 GB) GPU. This powerful GPU was chosen due to its ability to handle the computationally intensive nature of talking head generation. Tables 2 and 3 detail the hyperparameters used to generate the videos during inference time of each model.

**Table 2.** Hyperparameters used to generate the videos using Wav2Lip.

| Hyperparameter | Value |
|---|---|
| - -static | False |
| - -fps | 25 |
| - -pads | [0, 10, 0, 0] |
| - -face_det_batch_size | 16 |
| - -wav2lip_batch_size | 128 |
| - -nosmooth | False |

It is important to note that Wav2Lip is an identity-agnostic model, meaning it can generate lip synced videos for any identity during inference without requiring training on the target video beforehand. In contrast, GeneFace++, like other NeRF-based methods, is identity-aware, requiring training on the target face before inference. Wav2Lip is more versatile for scenarios where quick adaptation to new identities is needed, such as video dubbing, virtual communication, and voice-over synchronization across different faces. Its ability to rapidly adapt to new subjects makes it suitable for applications where personalization to a single subject is less important. NeRF-based models like GeneFace++ are ideal for applications demanding high-fidelity 3D rendering of a specific person, such as digital humans, Metaverse

avatars, or personalized virtual characters. They excel in generating high-quality, photorealistic videos but are less flexible in generalizing across different subjects.

**Table 3.** Hyperparameters used to generate the videos using GeneFace++.

| Hyperparameter | Value |
|---|---|
| - -lle_percent | 0.2 |
| - -temperature | 0.2 |
| - -mouth_amp | 0.4 |
| - -raymarching_end_threshold | 0.01 |
| - -blink_mode | None |
| - -low_memory_usage | True |
| - -fast | False |

### *3.4. Lip Sync Evaluation*

In this subsection, we propose our evaluation methodology (both quantitative for objective evaluation and qualitative for subjective evaluation) designed for assessing the generalizablity of the lip sync models to culturally distinct languages.

#### 3.4.1. Quantitative (Objective) Evaluation

As mentioned previously, lip sync evaluation metrics can be divided into **lip sync accuracy** (audio–lip synchronization) and **visual quality** metrics. Advancements in image and video processing technologies, such as super-resolution and image quality enhancement, along with the pre-processing techniques discussed in the Related Work section of this paper, make it increasingly important to focus on lip sync accuracy (error) metrics. Metrics such as **SSIM**, **PSNR**, and **FID** are not specifically designed for lip sync evaluation and mainly focus on the general quality of images. The **SSIM** (Structural Similarity Index Metric) measures the similarity between two images by analyzing their luminance, contrast, and structure. It serves as an indicator of perceptual quality and helps in identifying artifacts or distortions. **PSNR** (Peak Signal-to-Noise Ratio) evaluates the highest potential power of a signal (such as the original image) against the power of noise caused by distortion (for instance, synthesis). This provides a metric for assessing the overall fidelity and noise level in synthesized content. The **FID** (Fréchet inception distance) quantifies the similarity between the distributions of feature representations for real and generated images, serving as a metric to assess the overall visual quality and diversity of synthesized content.

The **LMD** (Landmark Distance) metric is often utilized to assess audio-visual synchronization, specifically focusing on the alignment of lip movements. LMD calculates the Euclidean distance between landmarks in the lip region across synthesized video frames and their corresponding ground-truth frames, as described in [80]. This metric provides an estimate of the accuracy of lip shapes, which is indicative of the synchronization between audio signals and synthesized video frames. However, LMD has limitations as it does not directly measure lip synchronization from a human perceptual standpoint. Additionally, it fails to capture finer details of lip movements due to its reliance on only 20 sparse points in the lip region during calculation, as noted in [17].

The authors of [77] introduced **Lip Sync Error Distance (LSE-D)** and **Lip Sync Error Confidence (LSE-C)** as more suitable metrics for assessing the accuracy of generated lip movements in relation to the input audio. LSE-D represents the average error measure calculated based on the Euclidean distance between the lip and audio representations (embeddings). A lower LSE-D value signifies a higher audio-visual match. It is calculated as the negative value of the similarity score obtained from SyncNet. The LSE-D score is calculated by Equation (1):

$$LSE_D = -S(A(t), V(t)) \tag{1}$$

where $A(t)$: audio feature vector at time step $t$; $V(t)$: lip movement feature vector at time step $t$; and $S(A(t), V(t))$: similarity score between $A(t)$ and $V(t)$.

Conversely, LSE-C reflects the average confidence score, where a higher value indicates better audio-video correlation. Lower confidence scores suggest significant portions of the video exhibit out-of-sync lip movements. LSE-C derived from the variability of the similarity scores obtained across different time steps in the video, calculated by Equation (2):

$$LSE_C = \sigma\{(S(A(t)),\ t = 1 \text{ to } T)\} \tag{2}$$

where $\sigma$: standard deviation; $A(t)$: audio feature vector at time step $t$; $V(t)$: lip movement feature vector at time step $t$; and $S(A(t))$: represents the similarity score between $A(t)$ and $V(t)$ at each time step $t$.

We leverage a pre-trained **SyncNet** model to calculate these two metrics for our evaluation dataset. SyncNet boasts an accuracy of around 99% [39], making it a powerful tool for measuring lip sync error. It has also been shown that this model works across different languages, such as Korean or Japanese or any other language [39]. It is worth noting that SyncNet models are pre-trained, eliminating the need for ground-truth videos in our evaluation methodology. We can calculate LSE-D and LSE-C solely by providing the generated lip synced videos, without the need for ground truth. Our evaluation process involves utilizing Wav2Lip and GeneFace++ to generate talking faces using the evaluation dataset. Subsequently, we employ the provided method to calculate LSE-D and LSE-C for the generated videos.

#### 3.4.2. Qualitative (Subjective) Evaluation

Alongside this quantitative evaluation method, we will include qualitative assessment by a human observer in our research. SyncNet and the LSE metrics provide a strong framework for evaluating lip sync accuracy; however, there may be biases towards certain languages due to SyncNet's possible training on mainly English or a restricted range of languages. Thus, performing human evaluation acts as an additional method.

In qualitative evaluation, it is essential to choose suitable metrics or criteria for human observers, much like in quantitative evaluation. In the evaluation of lip synced videos, human observers may take into account several criteria. For this study, which assesses model performance across various languages, it is crucial to select criteria that are sensitive to the language of the input audio. For example, a criterion such as "Artifacts", which pertains to visual distortions or irregularities in the video, is not appropriate. Artifacts are influenced by the architecture of the model rather than the spoken language. Thus, choosing criteria that depend on language for qualitative evaluation is essential.

We identified and selected two appropriate criteria for this task: adequacy and naturalness. In this context, **adequacy** denotes the degree to which the lip movements correspond to the spoken words and sounds present in the audio. A lip synced video is deemed sufficient when the lip movements closely match the phonetic elements of the speech, accurately representing the intended words and sounds. Adequacy evaluates how accurately the lip syncing process visually represents the spoken content. **Naturalness** denotes the extent to which the lip movements in the synthesized video appear realistic and lifelike, mirroring the natural movements seen in actual speech. In lip syncing, naturalness assesses whether the lip movements are fluid, smooth, and contextually suitable, thereby improving the overall credibility and authenticity of the video. A lip synced video is deemed natural

when the lip movements correspond with the rhythm, pacing, and intonation of the spoken words, resulting in a smooth and credible visual depiction of speech. We adapt the **Mean Opinion Score (MOS)** method to survey the human evaluators about the adequacy and naturalness of videos, who are the native speakers of these languages. This involves showing samples of the lip synced videos and asking participants to rate them on a scale of 1 (Poor) to 5 (Excellent).

By considering the strengths and limitations of both quantitative and qualitative evaluation methods, and by selecting appropriate metrics and criteria, we ensured a comprehensive, robust, and reliable evaluation methodology.

### *3.5. Face-to-Face Translation*

The concept of face-to-face translation involves translating all elements of a speaker's communication, such as language, voice, facial expressions, and lip movements, into a different language. The generation of AI Avatars that can produce multilingual speech, commonly known as Video Translation, is supported by tools like HeyGen [81], D-ID [82], and Synthesia [83]. The modular workflow presented in Figure 2 demonstrates how this can be accomplished. Initially, the speaker's voice is recorded and identified by an **Automatic Speech Recognition (ASR)** tool. The output text is subsequently translated into the target language through the use of **Machine Translation (MT)**. Subsequently, a Speech Synthesis module is utilized that includes both a **Text-to-Speech (TTS)** model and a **Voice Cloning (VC)** model. This process translates the speaker's voice into the target language, forming the **Speech-to-Speech (S2S)** translation step. After translating the input speech into the target language, the generated audio and the target face (which may belong to the speaker or another person or avatar) are fed into the **Talking Face Generation** model, where lip sync plays a vital role. Finally, a talking head is generated and, if necessary, an **Image/Video Enhancer** framework (Super Resolution) can be utilized to enhance the visual quality of the generated video. This approach addresses potential issues that may arise from low-quality input images or videos, limitations in the talking face generation model, or inadequate computational resources.

A very similar system was implemented in 1999 using rule-based and statistical models [20]. Now, 25 years later, the progress in neural networks and deep learning allows us to develop such a system with much more efficiency and accuracy. In this study, we implemented the workflow in Figure 2 in a modular way with the cutting-edge and most advanced deep neural models. For the ASR stage, we utilized **distil-whisper/distil-small.en** model. This is a distilled version of the Whisper model that is 6 times faster, 49% smaller, and performs within 1% WER (Word Error Rate) on out-of-distribution evaluation sets [84]. After transcribing the input audio, for translating the text from source to target language, we used **Helsinki-NLP/opus-mt-{src}-{trg}** model [85]. This is trained on the OPUS parallel corpus, covering over a thousand {src}-{trg} language pairs, and is designed for efficient and high-quality translation tasks, making it a widely used choice for multilingual translation in research and practical applications. In order to synthesize translated speech from the translated text, we selected **facebook/mms-tts** as our TTS module [86]. This leverages the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) architecture. VITS is a conditional variational autoencoder (VAE) that can generate natural-sounding speech from text input, focusing on generalizing across diverse languages and dialects to improve accessibility and inclusion in speech synthesis. Finally, the translated speech is given to the Talking Face Generation module (Wav2Lip and GeneFace++ models) to finalize the face-to-face translation task. Optionally, in order to improve the visual quality of the videos, we applied **gfpgan** Super Resolution with upscale factor of 2 to each frame of the video, resulting in an enhanced video [43].

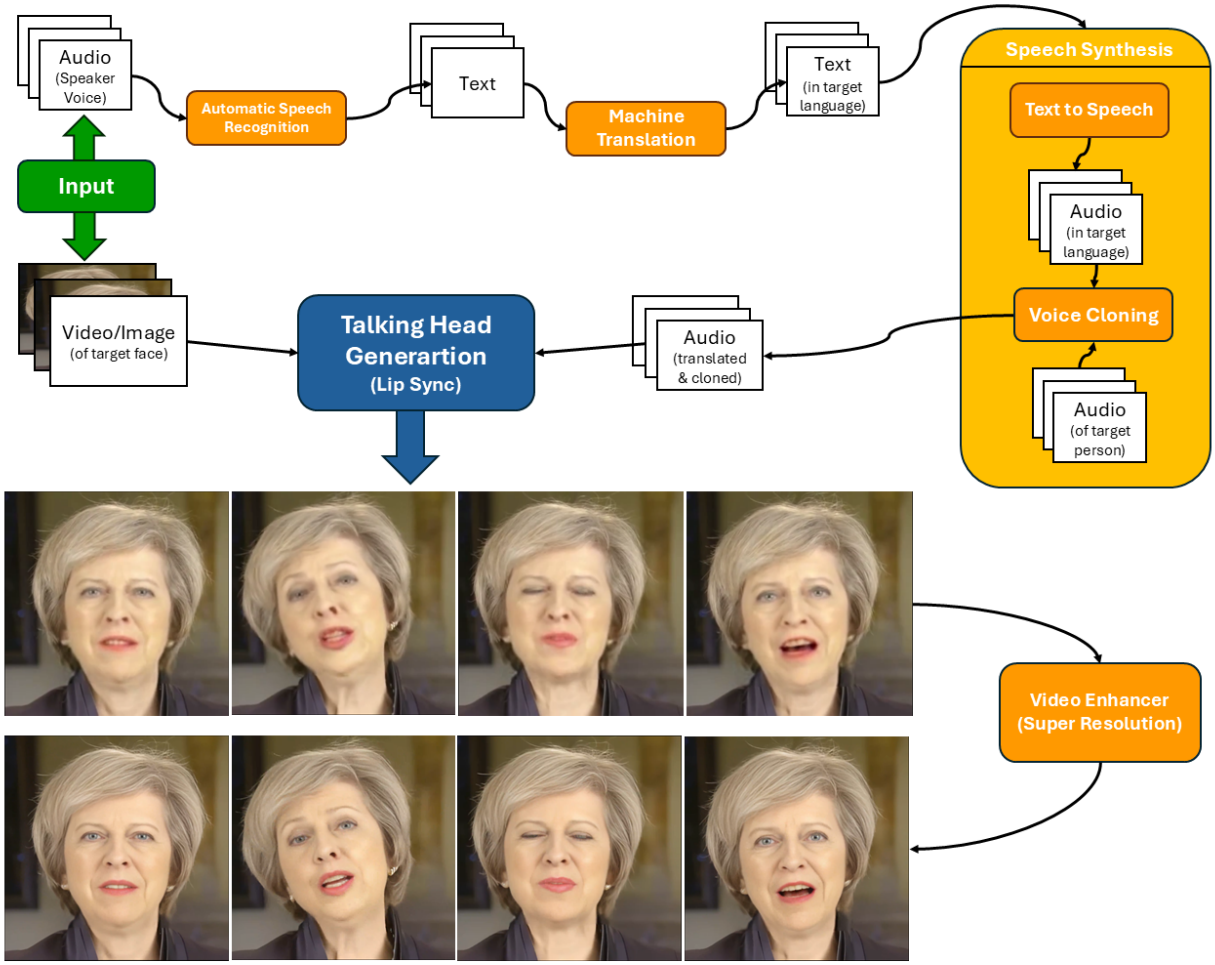


**Figure 2.** A modular face-to-face translation workflow.

Furthermore, in Section 4.3, we assess the efficiency of the two versions of our prototype (utilizing Wav2Lip or GeneFace++) objectively by measuring the inference time of each module and calculating the throughput of the whole system. This provides insights into the efficiency of our proposed workflow in various settings, such as a real-time video conferencing system or a VR/AR headset equipped with face-to-face translation.

## 4. Experiments and Results

In this section, we evaluate the performance of Wav2Lip and GeneFace++ on generating lip synced videos for languages beyond English. Initially, we generate approximately one hour of talking head videos for each language (Persian, Arabic, and Turkish) using these models and our dataset and use them for quantitative evaluation. From the same collection, we sample 3 min long videos and use them for qualitative evaluation. Lastly, we analyze the performance of the implemented face-to-face translation system.

### *4.1. Quantitative Evaluation Results*

To assess the quality of the generated lip synced videos, we employ the pre-trained SyncNet model and calculate two metrics: LSE-D (distance) and LSE-C (confidence). It is important to note that this pre-trained SyncNet instance differs from the Lip Sync Expert used within Wav2Lip itself, which was trained on the LRS2 [87] dataset. The pre-trained SyncNet achieves a reported accuracy of 99%, suggesting reliable evaluation results. Table 4 presents the quantitative evaluation results.

Language-wise, Persian emerges as the language with the best scores in LSE-D (distance) for both models by a significant margin. Arabic follows closely in second place, while Turkish exhibits the worst scores for LSE-D. For the LSE-C metric, a similar pattern to the other metric is observed in Wav2Lip, while in GeneFace++ Persian achieves the worst confidence score.

Model-wise, Wav2Lip scores much better than GeneFace++ in terms of the accuracy of lip synchronization in all languages.

**Table 4.** Quantitative evaluation results.

| Language | Length (min) | LSE_D ↓ | | LSE_C ↑ | |
|---|---|---|---|---|---|
| | | Wav2Lip | GeneFace++ | Wav2Lip | GeneFace++ |
| Turkish | 67 | 7.088657544 | 8.633640551 | 7.265054778 | 5.018934065 |
| Arabic | 54 | 7.011181367 | 8.833320711 | 7.410812804 | 5.238331367 |
| Persian | 59 | 6.235052688 | 7.58077689 | 8.389725332 | 4.936709722 |

We acknowledge that previous studies and models have utilized various datasets, primarily focused on English (LRW, LRS2, LRS3, VoxCeleb2) [88,89]. However, a comparison of their results with ours suggests promising generalizability by both models for Persian, Turkish, and Arabic languages. The selected models, especially Wav2Lip, demonstrate remarkable performance and accuracy across these different languages, validating the claim of their language independence, at least within the context of our selected languages.

### *4.2. Qualitative Evaluation Results*

Although our quantitative assessment utilizing SyncNet offers significant insights, it is essential to recognize its constraints. Initially, the potential bias of SyncNet towards English or related languages may result in an inadequate assessment of lip syncing quality for languages such as Persian, Turkish, and Arabic, thereby neglecting their unique pronunciation characteristics. Furthermore, numerical assessments may encounter difficulties in capturing the nuances of natural lip movements, often favoring elevated similarity scores despite the potential for the generated movements to seem unnatural in the context of a specific language. Ultimately, the environment in which recordings are made can influence the outcomes. The enhanced performance observed in the Persian dataset can be ascribed to its high-quality studio recording, characterized by minimal noise and background interference. The datasets in Turkish and Arabic are sourced from news broadcasts, characterized by less controlled environments in comparison to studio settings. The significance of human evaluation in conjunction with quantitative measures is underscored by these factors.

Human evaluators, preferably native speakers or individuals well acquainted with the languages, can evaluate how well the generated lip movements correspond to natural pronunciation and the mouth shapes characteristic of each language. For example, Persian includes specific sounds such as “qaf” and “khe” that necessitate particular lip movements that are not typically found in English. Turkish vowel harmony affects how words are pronounced and may also influence the shapes of lips when speaking. Finally, emphatic consonants in Arabic require strong pronunciation, which results in noticeable lip movements. Human evaluation confirms that the produced lip movements meet technical standards while also looking natural, being culturally suitable, and reflecting the subtleties of the intended languages.

In our study, native speakers of each language evaluate a subset of generated videos (3-minute long samples) for the **adequacy** and **naturalness** criteria.

In a comparison of **Wav2Lip** regarding **adequacy**, Persian evaluators report that the synchronization is excellent for most of the video, including distinct sounds that are not present in English. Evaluations of Turkish suggest that lip movements are generally well generated and accurate, although there are some subtle errors. Users observed cases of quicker lip movement transitions, which may be attributed to the vowel harmony feature found in Turkic languages. Arabic reviewers indicated that lip sync accuracy was high, even with strong pronunciations, though some errors were noted. A participant who is fluent in all three languages also assessed the videos. Their evaluation indicated that Persian had the highest adequacy with few errors, followed by Turkish and then Arabic.

Although the human evaluation regarding language for **Wav2Lip** showed good adequacy, the **naturalness** scores were typically lower. The evaluated videos included an artificial talking avatar, which is noteworthy, as users probably did not anticipate perfectly realistic lip movements. Observers noted weaknesses in motion and pacing. Many observers pointed out that, in certain cases, even when lip movements closely aligned with the timing and accuracy of the audio, they seemed excessively reactive to the speech. This indicates that the lips responded to even slight sounds that may not need a noticeable alteration in lip position. As a result, the videos occasionally showed quicker and unexpected transitions, affecting the overall smoothness.

A language-wise comparison of **GeneFace++** using the **adequacy** metric reveals minimal differences in its ability to generate lip synced videos across various languages. While a similar pattern is observed in GeneFace++ and Wav2Lip, GeneFace++ exhibits smaller language-based performance variations. Users report that GeneFace++ generally produces less prominent and bold lip movements compared to Wav2Lip. Although minor errors exist, GeneFace++ does not fully mimic lip movements for certain words.

Regarding **naturalness**, **GeneFace++** excels in generating highly natural lip motion. Users report that, despite potential lip sync inaccuracies, the generated videos appear very natural and lifelike, often indistinguishable from real videos at first glance. Notably, the language-based differences in naturalness ratings for GeneFace++ are minimal.

For a model-wise comparison, the results show that Wav2Lip significantly outperforms GeneFace++ in terms of adequacy. Similarly, GeneFace++ exhibits a higher level of naturalness compared to Wav2Lip.

Overall, user evaluations suggest that GeneFace++ is a more **stable** model, demonstrating consistent performance across different languages, albeit with occasional inaccuracies and highly lifelike lip synchronization. In contrast, Wav2Lip, while producing **highly accurate** lip movements, often lacks the naturalness and realism associated with real human speech.

The results of our qualitative evaluation using the Mean Opinion Score (MOS) are presented in Table 5. These human evaluations complement the quantitative results, highlighting the strengths and weaknesses of these models' performance across different languages. They emphasize the importance of considering language-specific nuances beyond solely relying on metrics potentially biased towards certain languages.

**Table 5.** Qualitative evaluation results.

| Language | No. of Participants | Length (min) | Adequacy | | Naturalness | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | | | Wav2Lip | GeneFace++ | Wav2Lip | GeneFace++ | Wav2Lip | GeneFace++ |
| Turkish | 12 | 3 | 4.2/5.0 | 3.8/5.0 | 3.4/5.0 | 4.4/5.0 | 3.8/5.0 | 4.1/5.0 |
| Arabic | 6 | 3 | 4.1/5.0 | 3.7/5.0 | 3.3/5.0 | 4.3/5.0 | 3.7/5.0 | 4.0/5.0 |
| Persian | 7 | 3 | 4.6/5.0 | 3.8/5.0 | 3.6/5.0 | 4.4/5.0 | 4.1/5.0 | 4.1/5.0 |

### *4.3. Face-to-Face Translation Results*

In this subsection, we evaluate the performance of the proposed face-to-face translation system outlined in Section 3.5. While individual modules and lip sync models have been extensively evaluated in the literature, we focus on assessing their integrated performance within our system. We previously evaluated the generalizability of lip sync models to culturally distinct languages. Here, we delve into the inference time of each module and the entire system. Additionally, we conduct an ablation study to analyze the impact of the optional Video Enhancing (Super Resolution) module on system performance. This module, while enhancing video quality, may implicitly affect lip sync accuracy by manipulating pixel-level details.

For **Inference Time Analysis**, we implemented the system using the models described in Section 3.5 with two variations: one incorporating Wav2Lip as the talking head generation module and the other using GeneFace++. For testing, we collected a 60-second English voice recording containing 156 tokens. By generating face-translated videos for this 60-second audio using the same identity (Theresa May's video), we measured the system's throughput, a critical factor for real-world applications of face-to-face translation systems. The system is deployed on a machine with A100 40GB GPU. Table 6 presents the inference time of each module and the overall system throughput without the Video Enhancing module.

**Table 6.** Performance analysis of two different implementations of our face-to-face translation system. Input audio of length 60 s in English is face-translated to Turkish, Arabic and Persian. The table displays the inference time of the Automatic Speech Recognition, Machine Translation, Speech Synthesis, and Lip Synchronization modules, as well as the total inference time which is equal to the sum of the inference time of these modules. Output refers to the length of the face-translated output video.

| Sys. Variation | Source | Target | Input (s) | ASR (s) | MT (s) | SS (s) | LipSync (s) | Output (s) | Total Inference (s) | Throughput (s) |
|---|---|---|---|---|---|---|---|---|---|---|
| GeneFace++ | English | Turkish | 60 | 0.68 | 1 | 8.3 | 39 | 47 | 48.3 | 0.805 |
| | | Arabic | | | 0.82 | 10 | 44 | 52 | 54.82 | 0.913 |
| | | Persian | | | 0.97 | 8.9 | 42 | 50 | 51.87 | 0.864 |
| Wav2Lip | English | Turkish | 60 | 0.68 | 1 | 8.3 | 13.6 | 47 | 22.9 | 0.381 |
| | | Arabic | | | 0.82 | 10 | 17.7 | 52 | 28.52 | 0.475 |
| | | Persian | | | 0.97 | 8.9 | 15.1 | 50 | 24.97 | 0.416 |

Super Resolution for video involves applying the algorithm to each frame of the video, making it computationally expensive and time-consuming. Table 7 displays the results when the Video Enhancing (Super Resolution) module is included, allowing us to analyze the performance of the system in more detail.

The results in Table 6 demonstrate that, on an A100 GPU, the Wav2Lip-based variation achieves real-time face-to-face translation, with a system throughput below 0.5 s. The GeneFace++ variation also exhibits real-time performance, with a throughput below 1 s. It is worth noting that the generated audio in the target language may be slightly shorter than the original source language audio due to factors such as language-specific conversational length differences or speech speed adjustments by the speech synthesis module. The results in Table 7 indicate that the Super Resolution module significantly impacts system throughput, increasing it substantially and moving the system away from real-time performance.

Next, we conduct an **Ablation Study** on our system by removing the Video Enhancing module and calculating the LSE-D and LSE-C metrics for the generated 1-minute-long videos (the same videos we used for Inference Time Analysis). We also compare the throughput of the systems with and without the module. Figure 3 illustrates the contribution of each component of the system to the total inference time, with and without the

Super Resolution component. (The values are approximate and the average of the whole system, illustrated for comparison purpose only.)

**Table 7.** Performance analysis of two different implementations of our face-to-face translation system when the Video Enhancing (Super Resolution) module is included.

| Sys. Variation | Language | Initial Inference (s) | Super Resolution (s) | Final Inference (s) | Final Throughput (s) |
|---|---|---|---|---|---|
| GeneFace++ | Turkish | 48.3 | 249 | 297.3 | 4.955 |
| | Arabic | 54.82 | 380 | 434.82 | 7.247 |
| | Persian | 51.7 | 309 | 360.7 | 6.011 |
| Wav2Lip | Turkish | 22.9 | 235 | 257.9 | 4.298 |
| | Arabic | 28.52 | 377 | 405.52 | 6.758 |
| | Persian | 24.97 | 303 | 327.97 | 5.466 |

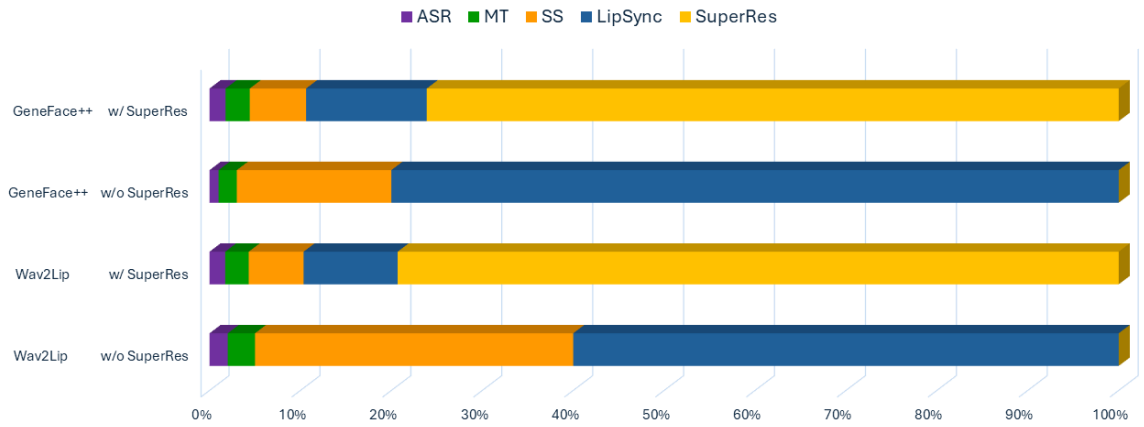


**Figure 3.** Stacked bar chart showing the approximate percentage contribution of each component to the total inference time, with and without Super Resolution component.

The stacked bar chart (Figure 3) reveals that, without Video Enhancing (Super Resolution), the Lip Sync module is the most time-consuming component. However, when Super Resolution is included, it becomes the dominant factor, significantly increasing the overall inference time. The contributions of the ASR, MT, and SS modules are minimal, especially in the presence of Super Resolution. This pattern holds true for both GeneFace++ and Wav2Lip-based system variations, with slight differences in the relative contributions of each component.

Table 8 compares the accuracy of lip synchronization as well as the system throughput, with and without the Video Enhancing (Super Resolution) component.

**Table 8.** A comparison of system throughput and LSE metrics, with and without Super Resolution component.

| Sys. Variation | Language | Throughput (s) | | LSE-D ↓ | | LSE-C ↑ | |
|---|---|---|---|---|---|---|---|
| | | w/o SuperRes | w/SuperRes | w/o SuperRes | w/SuperRes | w/o SuperRes | w/SuperRes |
| GeneFace++ | Turkish | 0.805 | 4.955 | 8.372032 | 8.355746 | 5.5287714 | 6.1372557 |
| | Arabic | 0.913 | 7.247 | 8.40481 | 8.485848 | 5.973344 | 6.240629 |
| | Persian | 0.8645 | 6.011 | 7.100972 | 7.192334 | 6.620623 | 6.940252 |
| Wav2Lip | Turkish | 0.346 | 4.298 | 7.7514715 | 6.7823787 | 7.354989 | 8.676653 |
| | Arabic | 0.475 | 6.758 | 7.9701333 | 6.9416428 | 7.313581 | 8.569199 |
| | Persian | 0.402 | 5.466 | 6.9580321 | 6.831578 | 7.554779 | 8.816901 |

While Super Resolution significantly enhances video quality (see Figure 4), the results in Table 8 demonstrates its negative impact on system efficiency and throughput. Interestingly, the ablation study revealed that, in most cases, adding Super Resolution improves lip synchronization accuracy (the only exception here is LSE-D for the GeneFace++ based

system which gets worse minimally), contrary to our initial expectations. We hypothesized that Super Resolution's manipulation of pixel-level details, including lip shapes, might negatively affect lip sync accuracy. However, our findings, limited to our dataset and selected models, suggest otherwise. This area warrants further investigation.


**Figure 4.** Illustration of the impact of Super Resolution on lip shape representation across both system variations, captured from our ablation study.

## 5. Discussion

Our evaluation methodology, detailed in Section 3.4, provides a robust framework for assessing the generalizability of talking head generation models to languages beyond English. Our experiments and results presented in Sections 4.1 and 4.2 demonstrate that both Wav2Lip (a GAN-based model) and GeneFace++ (a NeRF-based model) exhibit impressive lip sync accuracy across Turkish, Persian, and Arabic, supporting their claims of language independence. This addresses our first research question (RQ1) regarding the generalizability of these models. Still, the scarcity of research exploring the use of lip sync models for diverse languages is noteworthy. One study [18] focused on real-time multilingual talking face generation, employing a dataset encompassing Korean, Chinese, Japanese, and English. While their findings regarding real-time systems are valuable, they lacked quantitative evaluation. Another example [90] utilized a dataset containing French alongside English. For a comprehensive face-to-face translation system, a wider range of languages needs to be considered in future research. Unlike [18], which did not provide a quantitative evaluation, our study employs a comprehensive methodology that quantitatively assesses lip sync accuracy, providing detailed metrics such as LSE-D and LSE-C. Furthermore, while [18] primarily focuses on real-time processing, our approach integrates these models into a face-to-face translation system, evaluating their performance in terms of both accuracy and computational efficiency, thereby offering a more holistic and practical evaluation for multilingual applications. Our dataset, while limited to three languages, adds more depth to the literature, as we are the first authors to evaluate the models for these languages in the context of lip sync.

Our implementation of the face-to-face translation system, outlined in Section 3.5, and its performance analysis in Section 4.3 demonstrate that, by integrating state-of-the-art neural network models, we can design a system capable of real-time, multilingual face-to-face translation, particularly in scenarios demanding high-quality visuals. This addresses our second research question (RQ2) regarding the feasibility and performance of such a system.

The potential of a comprehensive face-to-face translation system that includes all the modules in our proposed workflow is significant because of its wide range of applications.

Recent developments in neural networks have produced established models for each module, addressing both real-time situations where efficiency is crucial and high-quality configurations that emphasize performance. At present, lip sync is the most important element in a face-to-face translation system. Many lip sync models focus on enhancing visual quality; however, this frequently results in reduced lip sync accuracy and increased latency [52,56,91]. Wav2Lip and its derivatives are effective in producing lips that are highly synchronized in real time. However, they may require post-processing using video enhancement models such as Real-ESRGAN or GFP-GAN, which demand considerable computational resources and time. Pre-processing techniques, like those used in [42], are also available. The researchers employ a modified Wav2Lip model that has been trained on a pre-processed 4K dataset, which was obtained using image enhancement models. Training Wav2Lip on high-resolution datasets produces high-resolution videos directly, thereby removing the extended inference times that are typically required for post-processing. Nonetheless, the computational expense associated with image enhancement models, even in the pre-processing phase, continues to be significant. Consequently, the creation of high-resolution speech datasets is essential.

In this study, we implemented two system variations with distinct configurations—one incorporating video Super Resolution and the other operating without it—resulting in a total of four experimental setups. The inclusion of Super Resolution significantly enhances the visual quality of generated videos, making it particularly suitable for non-streaming or offline applications where visual fidelity is prioritized over latency. Conversely, the configuration without Super Resolution achieves real-time performance, demonstrating the system's capability of catering to scenarios where efficiency and low latency are paramount. This dual setup effectively illustrates the inherent trade-off between performance and quality, shedding light on the operational flexibility required for diverse application scenarios. Our primary objective with this modular framework is to emphasize this trade-off and provide a foundation for researchers and practitioners to tailor the system to their specific requirements. The modularity of our approach not only supports adaptability but also paves the way for future advancements. For instance, optimizing system performance through techniques such as model compression, quantization, or the development of lightweight neural architectures could enable real-time operation even on resource-constrained platforms like mobile devices. By addressing the limitations imposed by computational overhead and integrating more efficient processing mechanisms, future research can extend the applicability of face-to-face translation systems to a wider array of practical environments.

## 6. Conclusions and Future Works

In conclusion, our study demonstrates the promising generalizability of existing lip sync models, particularly Wav2Lip and GeneFace++, in multilingual environments. Despite variations across languages, our evaluation in Turkish, Persian, and Arabic showcases their adequacy both quantitatively and qualitatively. In summary, our evaluation revealed LSE-D values as low as 6.235 and LSE-C values as high as 8.390, indicating robust lip sync accuracy across the tested languages. The worst-case LSE-D was 8.833, and the lowest LSE-C observed was 4.937, with other results distributed between these extremes. While direct comparisons across languages are limited due to language-specific inputs, our findings suggest no significant discrepancies in model performance across Turkish, Persian, and Arabic. By outlining a comprehensive workflow centered around lip sync, we have clarified the essential components required for effective face-to-face translation systems. Our implementation demonstrates the efficiency of this workflow in generating high-quality face-translated videos with acceptable latency, as the system achieved a throughput as low

as 0.381, showcasing the efficiency and scalability of the proposed modular workflow for multilingual settings. The language-agnostic nature of the lip sync models we employed highlights their potential for seamless integration into such systems. Lip sync, as a crucial component, completes the puzzle of face-to-face translation by enabling the full translation of both verbal and visual communication aspects.

Moving forward, addressing current limitations and challenges is paramount. This includes the necessity for high-resolution speech datasets and the development of real-time models. High-quality, annotated datasets for languages like Turkish, Persian, and Arabic are scarce compared to English, limiting the availability of training data for fine-tuning lip sync models. This scarcity necessitates the use of data augmentation techniques or transfer learning strategies, which may introduce their own biases or limitations. Variations in pronunciation, phoneme articulation, and lip movements across regions within the same language can impact the generalizability of lip sync models. For instance, dialectal variations in Arabic (e.g., Egyptian vs. Levantine) may require additional tuning or specialized datasets for effective adaptation. Balancing real-time performance with high-quality lip synchronization and visual fidelity remains a significant challenge. While neural architectures like GANs and NeRFs can produce visually compelling outputs, their computational demands may compromise latency, especially for high-resolution outputs in multilingual applications. Addressing these challenges requires a holistic approach, incorporating advancements in efficient neural architectures, robust dataset collection for diverse languages and accents, and optimization techniques to ensure scalability without sacrificing performance.

In future work, we aim to address key challenges in lip sync research. First, we plan to optimize computational efficiency for real-time applications, reducing processing time and resource demands while maintaining output quality. We will also focus on improving model generalization across diverse datasets, ensuring better performance for under-represented languages and speakers. Ethical concerns related to deepfakes will be another area of focus, where we will explore strategies to mitigate misuse and promote responsible application in fields like media, education, and virtual communication. Additionally, we aim to enhance avatar personalization, tailoring interactions to individual user preferences and behaviors. We also plan to integrate non-verbal cues, such as eye movements and facial expressions, into the lip sync process to create more natural and expressive avatars. These advancements will contribute to more immersive and personalized experiences in virtual environments, from gaming to virtual meetings. Future research should emphasize the importance of cross-disciplinary collaboration to improve lip sync and face translation models. Engaging with linguists, cognitive scientists, and human–computer interaction professionals will provide crucial insights into linguistic nuances, cognitive aspects of communication, and user experience. This collaboration can lead to the development of more culturally sensitive and contextually accurate models, ensuring that these technologies not only perform well technically but also resonate effectively across diverse languages and cultures. As with any advanced technology, the development of face-to-face translation systems raises important ethical considerations. Privacy and security are key concerns, particularly in sensitive or high-stakes interactions where the accuracy of the system could affect trust and confidentiality. There is also the potential for misuse, such as the generation of misleading or harmful content. To mitigate these risks, it is essential to implement strong data protection measures, informed consent protocols, and safeguards against malicious use. Future research should also explore the ethical implications of using such technologies in diverse contexts, ensuring that they are deployed responsibly and in a manner that respects individuals' rights and dignity. These insights can inform future research initiatives and enable researchers, developers, and companies to make contributions to this developing field.

In a broader context, our findings provide important insights that can facilitate the development of end-to-end, real-time solutions for face-to-face translation systems. These advancements have the potential to greatly change communication methods, representing an important shift in how people interact and connect with each other.

**Author Contributions:** Conceptualization, A.R.O. and M.S.A.; methodology, A.R.O. and M.S.A.; software, A.R.O.; resources, M.K.; data curation A.R.O. and M.S.A.; writing, reviewing, and editing, A.R.O., M.S.A. and M.K.; visualization, A.R.O.; supervision, M.S.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The authors agree to share datasets upon requests from readers.

**Acknowledgments:** The authors would like to extend their sincere thanks to **Aktif Bank** for their invaluable support in facilitating this research. Their provision of advanced computing resources and a conducive working environment was instrumental in the successful completion of this study. The video of Theresa May used in this paper is sourced from the UK Government and made available under the Open Government Licence (OGL). The video is used solely for non-commercial, educational purposes, in strict accordance with the terms of the OGL. We confirm that all usage of the content adheres to the requirements of the license, including the obligation to provide proper attribution to the UK Government as the original source. No modifications have been made to the video in a way that violates the OGL, and the content is used in full compliance with its stipulations. This material is not used for any commercial purpose. For full details of the OGL, please refer to the Open Government Licence.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ASR | Automatic Speech Recognition |
| MT | Machine Translation |
| SS | Speech Synthesis |
| TTS | Text to Speech |
| VC | Voice Cloning |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory |
| GAN | Generative Adversarial Network |
| NeRF | Neural Radiance Field |
| Diff | Diffusion |
| SSIM | Structural Similarity Index Metric |
| PSNR | Peak Signal-to-Noise Ratio |
| FID | Fréchet Inception Distance |
| LMD | Landmark Distance |
| MOE | Mean Opinion Score |
| VITS | Variational Inference for Text-to-Speech |
| LSE | Lip Synchronization Error |

## References

1. Montenegro-Rueda, M.; Fernández-Cerero, J.; Fernández-Batanero, J.M.; López-Meneses, E. Impact of the implementation of ChatGPT in education: A systematic review. *Computers* **2023**, *12*, 153. [CrossRef]
2. Vaccaro, L.; Sansonetti, G.; Micarelli, A. An empirical review of automated machine learning. *Computers* **2021**, *10*, 11. [CrossRef]

3. Mustafa, A.; Rahimi Azghadi, M. Automated machine learning for healthcare and clinical notes analysis. *Computers* **2021**, *10*, 24. [CrossRef]
4. Krichen, M. Convolutional neural networks: A survey. *Computers* **2023**, *12*, 151. [CrossRef]
5. de Winter, J.C.F.; Dodou, D.; Eisma, Y.B. System 2 Thinking in OpenAI's o1-Preview Model: Near-Perfect Performance on a Mathematics Exam. *Computers* **2024**, *13*, 278. [CrossRef]
6. Hannon, B.; Kumar, Y.; Li, J.J.; Morreale, P. Chef Dalle: Transforming Cooking with Multi-Model Multimodal AI. *Computers* **2024**, *13*, 156. [CrossRef]
7. Nichita, M.V.; Paun, M.A.; Paun, V.A.; Paun, V.P. The SARS-CoV-2 Virus Detection with the Help of Artificial Intelligence (AI) and Monitoring the Disease Using Fractal Analysis. *Computers* **2023**, *12*, 213. [CrossRef]
8. Rakhimova, D.; Karibayeva, A.; Karyukin, V.; Turarbek, A.; Duisenbekkyzy, Z.; Aliyev, R. Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools. *Computers* **2024**, *13*, 253. [CrossRef]
9. Toshpulatov, M.; Lee, W.; Lee, S. Talking human face generation: A survey. *Expert Syst. Appl.* **2023**, *219*, 119678. [CrossRef]
10. Kato, R.; Kikuchi, Y.; Yem, V.; Ikei, Y. Reality avatar for customer conversation in the metaverse. In Proceedings of the International Conference on Human-Computer Interaction, Virtual Conference, 26 June–1 July 2022; Springer: Cham, Switzerland, 2022; pp. 131–145.
11. Cruz, M.; Oliveira, A.; Pinheiro, A. Metaverse Unveiled: From the Lens of Science to Common People Perspective. *Computers* **2024**, *13*, 193. [CrossRef]
12. Schubert, M.; Endres, D. More Plausible Models of Body Ownership Could Benefit Virtual Reality Applications. *Computers* **2021**, *10*, 108. [CrossRef]
13. Abed, A.Z.M.; Abdelkader, T.; Hashem, M. SLACPSS: Secure Lightweight Authentication for Cyber–Physical–Social Systems. *Computers* **2024**, *13*, 225. [CrossRef]
14. Kolivand, H.; Ali, I.; Sulong, G. Realistic lip syncing for virtual character using common viseme set. *Comput. Inf. Sci.* **2015, *8*.** [CrossRef]
15. Terry, K.; Radhakrishnan, R. Detection and correction of lip-sync errors using audio and video fingerprints. *SMPTE Motion Imaging J.* **2010**, *119*, 42–52. [CrossRef]
16. Fenghour, S.; Chen, D.; Guo, K.; Li, B.; Xiao, P. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access* **2021**, *9*, 121184–121205. [CrossRef]
17. Chen, L.; Cui, G.; Kou, Z.; Zheng, H.; Xu, C. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv* **2020**, arXiv:2005.03201.
18. Song, H.K.; Woo, S.H.; Lee, J.; Yang, S.; Cho, H.; Lee, Y.; Choi, D.; Kim, K.w. Talking face generation with multilingual tts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 21425–21430.
19. Rafiei Oskooei, A.; Yahsi, E.; Sungur, M.S.; Aktas, M. Can One Model Fit All? An Exploration of Wav2Lip's Lip-Syncing Generalizability Across Culturally Distinct Languages. In Proceedings of the International Conference on Computational Science and Its Applications, Hanoi, Vietnam, 1–4 July 2024; Springer: Cham, Switzerland, 2024; pp. 149–164.
20. Ritter, M.; Meier, U.; Yang, J.; Waibel, A. Face translation: A multimodal translation agent. In Proceedings of the AVSP'99-International Conference on Auditory-Visual Speech Processing, Santa Cruz, CA, USA, 7–10 August 1999; Citeseer: Princeton, NJ, USA, 1999.
21. Arena, F.; Collotta, M.; Pau, G.; Termine, F. An overview of augmented reality. *Computers* **2022**, *11*, 28. [CrossRef]
22. Xue, H.; Sharma, P.; Wild, F. User Satisfaction in Augmented Reality-Based Training Using Microsoft HoloLens. *Computers* **2019**, *8*, 9. [CrossRef]
23. Dirin, A.; Laine, T.H. User Experience in Mobile Augmented Reality: Emotions, Challenges, Opportunities and Best Practices. *Computers* **2018**, *7*, 33. [CrossRef]
24. Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–28 Feburary 2024; Volume 38, pp. 23802–23804.
25. Zhao, Y.; Yuan, X.; Gao, S.; Lin, Z.; Hou, Q.; Feng, J.; Zhou, D. ChatAnything: Facetime Chat with LLM-Enhanced Personas. *arXiv* **2023**, arXiv:2311.06772.
26. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; Hu, X. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–32. [CrossRef]
27. Oskooei, A.R.; Babacan, M.S.; Yağcı, E.; Alptekin, Ç.; Buğday, A. Beyond Synthetic Benchmarks: Assessing Recent LLMs for Code Generation. In Proceedings of the International Workshop on Computer Science and Engineering (WCSE), Phuket Island, Thailand, 19–21 June 2024; pp. 290–296. [CrossRef]
28. Koh, J.Y.; Fried, D.; Salakhutdinov, R.R. Generating images with multimodal language models. *Adv. Neural Inf. Process. Syst.* **2024**, *36*. [CrossRef]

29. Ma, Y.; Øland, A.; Ragni, A.; Del Sette, B.M.; Saitis, C.; Donahue, C.; Lin, C.; Plachouras, C.; Benetos, E.; Shatri, E.; et al. Foundation models for music: A survey. *arXiv* **2024**, arXiv:2408.14340.
30. Kadam, A.; Rane, S.; Mishra, A.K.; Sahu, S.K.; Singh, S.; Pathak, S.K. A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation. *EAI Endorsed Trans. Creat. Technol.* **2021**, *8*, e2. [CrossRef]
31. Naitali, A.; Ridouani, M.; Salahdine, F.; Kaabouch, N. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers* **2023**, *12*, 216. [CrossRef]
32. Llorach, G.; Evans, A.; Blat, J.; Grimm, G.; Hohmann, V. Web-based live speech-driven lip-sync. In Proceedings of the 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), Skövde, Sweden, 7–9 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–4.
33. Jamaludin, A.; Chung, J.S.; Zisserman, A. You said that?: Synthesising talking faces from audio. *Int. J. Comput. Vis.* **2019**, *127*, 1767–1779. [CrossRef]
34. Wiles, O.; Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–686.
35. Kumar, R.; Sotelo, J.; Kumar, K.; De Brebisson, A.; Bengio, Y. Obamanet: Photo-realistic lip-sync from text. *arXiv* **2017**, arXiv:1801.01442.
36. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–13. [CrossRef]
37. Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; Li, D. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph. (TOG)* **2020**, *39*, 1–15. [CrossRef]
38. KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; Jawahar, C. Towards automatic face-to-face translation. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1428–1436.
39. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Proceedings of the Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part II 13; Springer: Cham, Switzerland, 2017; pp. 251–263.
40. GitHub-Saifhassan/Wav2Lip-HD: High-Fidelity Lip-Syncing with Wav2Lip and Real-ESRGAN—github.com. Available online: https://github.com/saifhassan/Wav2Lip-HD (accessed on 30 October 2024).
41. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 1905–1914.
42. Liang, C.; Wang, Q.; Chen, Y.; Tang, M. Wav2Lip-HR: Synthesising clear high-resolution talking head in the wild. *Comput. Animat. Virtual Worlds* **2024**, *35*, e2226. [CrossRef]
43. Wang, X.; Li, Y.; Zhang, H.; Shan, Y. Towards real-world blind face restoration with generative facial prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9168–9178.
44. Kim, B.K.; Kang, J.; Seo, D.; Park, H.; Choi, S.; Song, H.K.; Kim, H.; Lim, S. A Unified Compression Framework for Efficient Speech-Driven Talking-Face Generation. *arXiv* **2023**, arXiv:2304.00471.
45. Li, Z.; Li, H.; Meng, L. Model compression for deep neural networks: A survey. *Computers* **2023**, *12*, 60. [CrossRef]
46. Wang, G.; Zhang, P.; Xie, L.; Huang, W.; Zha, Y. Attention-based lip audio-visual synthesis for talking face generation in the wild. *arXiv* **2022**, arXiv:2203.03984.
47. Chen, Z.; Wang, X.; Xie, L.; Yuan, H.; Pan, H. LPIPS-AttnWav2Lip: Generic audio-driven lip synchronization for talking head generation in the wild. *Speech Commun.* **2024**, *157*, 103028. [CrossRef]
48. Wang, K.C.; Zhang, J.; Huang, J.; Li, Q.; Sun, M.T.; Sakai, K.; Ku, W.S. Ca-wav2lip: Coordinate attention-based speech to lip synthesis in the wild. In Proceedings of the 2023 IEEE International Conference on Smart Computing (SMARTCOMP), Bangkok, Thailand, 22–25 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
49. Guo, Y.; Chen, K.; Liang, S.; Liu, Y.J.; Bao, H.; Zhang, J. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 5784–5794.
50. Yao, S.; Zhong, R.; Yan, Y.; Zhai, G.; Yang, X. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv* **2022**, arXiv:2201.00791.
51. Chatziagapi, A.; Athar, S.; Jain, A.; Rohith, M.; Bhat, V.; Samaras, D. LipNeRF: What is the right feature space to lip-sync a NeRF? In Proceedings of the 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), Gwangju, Republic of Korea, 17–20 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
52. Bi, C.; Liu, X.; Liu, Z. NERF-AD: Neural Radiance Field With Attention-Based Disentanglement For Talking Face Synthesis. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 12–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 3490–3494.

53. Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; Zhao, Z. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv* **2023**, arXiv:2305.00787.
54. Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv* **2024**, arXiv:2401.08503.
55. Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; Lu, J. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1982–1991.
56. Mukhopadhyay, S.; Suri, S.; Gadde, R.T.; Shrivastava, A. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 22–27 January 2024; pp. 5292–5302.
57. Stypułkowski, M.; Vougioukas, K.; He, S.; Zięba, M.; Petridis, S.; Pantic, M. Diffused heads: Diffusion models beat gans on talking-face generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 22–27 January 2024; pp. 5091–5100.
58. Hou, R.; Zhao, X. High-Quality Talking Face Generation via Cross-Attention Transformer. In Proceedings of the 2024 IEEE International Conference on Real-time Computing and Robotics (RCAR), Bengaluru, India, 21–24 February 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 194–199.
59. Huang, R.; Zhong, W.; Li, G. Audio-driven talking head generation with transformer and 3d morphable model. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 7035–7039.
60. Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; Yu, X. Styletalk: One-shot talking head generation with controllable speaking styles. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA , 7–14 February 2023; Volume 37, pp. 1896–1904.
61. Kadandale, V.S.; Montesinos, J.F.; Haro, G. Vocalist: An audio-visual synchronisation model for lips and voices. *arXiv* **2022**, arXiv:2204.02090.
62. Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; Komura, T. Faceformer: Speech-driven 3d facial animation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 18770–18780.
63. Gultekin, E.; Aktas, M.S. A Business Workflow Architecture for Predictive Maintenance using Real-Time Anomaly Prediction On Streaming IoT Data. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Kuala Lumpur, Malaysia, 10–13 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 4568–4575.
64. Gultekin, E.; Aktas, M.S. Real-Time Anomaly Detection Business Process for Industrial Equipment Using Internet of Things and Unsupervised Machine Learning Algorithms. In Proceedings of the Computational Science and Its Applications—ICCSA 2023 Workshops, Vienna, Austria, 1–4 July 2023; Springer: Cham, Switzerland, 2023; pp. 16–31.
65. Pierce, M.E.; Fox, G.C.; Aktas, M.S.; Aydin, G.; Gadgil, H.; Qi, Z.; Sayar, A. The QuakeSim project: Web services for managing geophysical data and applications. In *Earthquakes: Simulations, Sources and Tsunamis*; Birkhäuser: Basel, Switzerland, 2008; pp. 635–651.
66. Aktas, M.; Aydin, G.; Donnellan, A.; Fox, G.; Granat, R.; Grant, L.; Lyzenga, G.; McLeod, D. iSERVO: Implementing the International Solid Earth Research Virtual Observatory by integrating computational grid and geographical information web services. In *Computational Earthquake Physics: Simulations, Analysis and Infrastructure, Part II*; Birkhäuser: Basel, Switzerland, 2007; pp. 2281–2296.
67. Fox, G.C.; Aktas, M.S.; Aydin, G.; Gadgil, H.; Pallickara, S.; Pierce, M.E.; Sayar, A. Algorithms and the Grid. *Comput. Vis. Sci.* **2009**, *12*, 115–124. [CrossRef]
68. Nacar, M.A.; Aktas, M.S.; Pierce, M.; Lu, Z.; Erlebacher, G.; Kigelman, D.; Bollig, E.F.; da Silva, C.R.S.; Sowell, B.; Yuen, D.A. VLab: Collaborative Grid services and portals to support computational material science. *Concurr. Comput. Pract. Exp.* **2007**, *19*, 1717–1728. [CrossRef]
69. Aydin, G.; Sayar, A.; Gadgil, H.; Aktas, M.S.; Fox, G.C.; Ko, S.; Bulut, H.; Pierce, M.E. Building and applying geographical information system Grids. *Concurr. Comput. Pract. Exp.* **2008**, *20*, 1653–1695. [CrossRef]
70. Li, P.; Zhao, H.; Liu, Q.; Tang, P.; Zhang, L. TellMeTalk: Multimodal-driven talking face video generation. *Comput. Electr. Eng.* **2024**, *114*, 109049. [CrossRef]
71. Yu, L.; Xie, H.; Zhang, Y. Multimodal learning for temporally coherent talking face generation with articulator synergy. *IEEE Trans. Multimed.* **2021**, *24*, 2950–2962. [CrossRef]
72. Uygun, Y.; Oguz, R.F.; Olmezogullari, E.; Aktas, M.S. On the large-scale graph data processing for user interface testing in big data science projects. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Virtual Conference, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2049–2056.
73. Olmezogullari, E.; Aktas, M.S. Representation of click-stream data sequences for learning user navigational behavior by using embeddings. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data) Virtual Conference, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3173–3179.

74. Olmezogullari, E.; Aktas, M.S. Pattern2Vec: Representation of clickstream data sequences for learning user navigational behavior. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6546. [CrossRef]
75. Kapdan, M.; Aktas, M.; Yigit, M. On the structural code clone detection problem: A survey and software metric-based approach. In Proceedings of the Computational Science and Its Applications–ICCSA 2014: 14th International Conference, Guimarães, Portugal, 30 June–3 July 2014; Proceedings, Part V; pp. 492–507.
76. Sahinoglu, M.; Incki, K.; Aktas, M.S. Mobile application verification: A systematic mapping study. In Proceedings of the Computational Science and Its Applications–ICCSA 2015: 15th International Conference, Banff, AB, Canada, 22–25 June 2015; Proceedings, Part V; pp. 147–163.
77. Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 484–492.
78. Kolobov, R.; Okhapkina, O.; Omelchishina, O.; Platunov, A.; Bedyakin, R.; Moshkin, V.; Menshikov, D.; Mikhaylovskiy, N. Mediaspeech: Multilanguage asr benchmark and dataset. *arXiv* **2021**, arXiv:2103.16193.
79. Halabi, N. Persian Speech Corpus—fa.persianspeechcorpus.com. Available online: https://fa.persianspeechcorpus.com/ (accessed on 30 October 2024).
80. Chen, L.; Li, Z.; Maddox, R.K.; Duan, Z.; Xu, C. Lip movements generation at a glance. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 520–535.
81. HeyGen-AI Video Generator—heygen.com. Available online: https://www.heygen.com/ (accessed on 11 November 2024).
82. D-ID Creative Reality$^{TM}$—d-id.com. Available online: https://www.d-id.com/ (accessed on 11 November 2024).
83. Best AI Video Generator—Start Creating FREE AI Videos Now—synthesia.io. Available online: https://www.synthesia.io/ (accessed on 11 November 2024).
84. Gandhi, S.; von Platen, P.; Rush, A.M. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv* **2023**, arXiv:2311.00430.
85. Tiedemann, J.; Thottingal, S. OPUS-MT–building open translation services for the world. In Proceedings of the 22nd annual conference of the European Association for Machine Translation, Lisbon, Portugal, 11–14 May 2020; pp. 479–480.
86. Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; et al. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.* **2024**, *25*, 1–52.
87. Mroueh, Y.; Marcheret, E.; Goel, V. Deep multimodal learning for audio-visual speech recognition. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 2130–2134.
88. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv* **2018**, arXiv:1806.05622.
89. Afouras, T.; Chung, J.S.; Zisserman, A. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv* **2018**, arXiv:1809.00496.
90. Patel, D.; Zouaghi, H.; Mudur, S.; Paquette, E.; Laforest, S.; Rouillard, M.; Popa, T. Visual dubbing pipeline with localized lip-sync and two-pass identity transfer. *Comput. Graph.* **2023**, *110*, 19–27. [CrossRef]
91. Guan, J.; Zhang, Z.; Zhou, H.; Hu, T.; Wang, K.; He, D.; Feng, H.; Liu, J.; Ding, E.; Liu, Z.; et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1505–1515.