

Can One Model Fit All? An Exploration of Wav2Lip's Lip-Syncing Generalizability Across Culturally Distinct Languages

Amirkia Rafiei Oskooei[1(✉)], Ezgi Yahsi[2], Mehmet Sungur[2], and Mehmet S. Aktas[1]

[1] Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey
aktas@yildiz.edu.tr, amirkia.oskooei@std.yildiz.edu.tr
[2] Aktifbank Research and Development Center, Istanbul, Turkey
{ezgi.yahsi,mehmet.sungur}@aktifbank.com.tr

Abstract. This study explores the potential of Wav2Lip, a state-of-the-art lip-sync model, in multilingual environments. We assess its performance in generating lip-synchronized videos for Turkish, Persian, and Arabic languages. The evaluation results reveal promising language independence for Wav2Lip, achieving comparable accuracy to English. The research identifies the gap in research on lip-sync models for diverse languages and emphasizes the need for broader exploration. Additionally, we introduce a comprehensive Face-to-Face Translation workflow, outlining the fundamental elements for a seamless cross-lingual communication system. This work highlights the importance of Lip Sync models and the potential of Wav2Lip within such a system. By acknowledging current limitations and advocating for advancements in real-time models and high-resolution datasets, this study lays the groundwork for the development of revolutionary Face-to-Face Translation systems, fostering a future of barrier-free communication.

Keywords: Talking-face Generation · Lip Sync · Face-to-Face Translation · Generative AI · Deep Learning · Computer Vision

1 Introduction

The recent advancements in Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), have revolutionized the field of audio-driven talking-face generation. This surge in research and development is driven by the diverse and valuable applications of such technology. With the emergence of virtual worlds like the Metaverse, signified by the rebranding of Facebook to Meta, and the introduction of tools like Apple Vision Pro and Meta Quest to enhance virtual world experiences, the need for realistic talking avatars has become increasingly critical [1–3].

Generating a talking face from an audio input (audio-driven talking-face generation) involves synchronizing various facial movements, including the head, eyes, and most importantly, the lips. Lip synchronization (lip-sync) plays a vital

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14819, pp. 149–164, 2024.
https://doi.org/10.1007/978-3-031-65282-0_10

role in this process, as the mouth area exhibits the most significant motion during facial expression and is crucial for effective communication.

Prior to the rise of deep learning, techniques like Visiomes [4] and Timecode-based [5] Systems were employed for lip-sync. However, the advancements in deep learning, fueled by factors like the dramatic improvement in neural network computations enabled by powerful GPUs and technologies like Nvidia CUDA, have shifted the research focus towards DNN-based lip-sync models [6]. Various architectures have been devised and trained for this purpose, as extensively discussed in the Related Works section.

With the proliferation of these models, there arises a critical need for performance evaluation. These models are typically assessed using two main categories of metrics: 1) Image Quality metrics that evaluate the visual fidelity of the generated lip and mouth area 2) Lip-Sync Error metrics that measure the temporal accuracy of lip movements with respect to the audio.

Among the leading Lip-Sync models, Wav2Lip stands out for achieving state-of-the-art results in both categories [7]. Despite the introduction of numerous models with varying architectures since its release, Wav2Lip remains a benchmark for open-source models in terms of Lip-Sync accuracy metrics. A key strength of Wav2Lip is its unconstrained nature, allowing it to function with any arbitrary audio, language, or facial type.

While this model is language independent, it is important to note that Wav2Lip's training data, the LRS2 dataset [8], is entirely English-based. Similarly, many evaluation datasets used in research focus primarily on English. Emerging studies and observations indicate that these models may fail to generalize to certain input speech languages [9], raising concerns about the robustness of these models and emphasizing the need for comprehensive evaluation across diverse linguistic contexts. This gap in research motivates our current investigation. We aim to evaluate Wav2Lip's efficacy with Turkish, Persian, and Arabic, languages from varied origins and linguistic families, that have received less research attention in this context.

Beyond this gap, another intriguing area demanding further attention is the use of lip-sync for Multilingual Talking Face Generation and Face-to-Face Translation [10]. Face-to-Face translation involves translating input audio to a target language and generating a talking face synchronized with the translated speech. Lip-sync has traditionally presented a significant hurdle in achieving this functionality. However, with the ongoing advancements in lip-sync technology, it can now be integrated into this process. We believe such a workflow holds immense potential and could revolutionize communication across various domains, including education, business, video conferencing, and emerging environments like the Metaverse.

Our paper presents two primary contributions. Firstly, we evaluate the performance of Wav2Lip, a highly successful open source lip-sync model, in the context of Turkish, Persian, and Arabic languages. Secondly, we propose a Face-to-Face Translation workflow that leverages advancements in lip-sync technology.

2 Related Works

Lip synchronization (lip-sync) is a critical area within computer vision and multimedia, experiencing a surge in interest due to its expanding influence across diverse and often novel applications. Research explores a multitude of uses, encompassing audio-to-video generation, live video conferencing, accessibility tools, animation, dubbing and subtitling, and even entertainment, satire, and artistic storytelling [11].

Prior to the deep learning revolution, lip-sync relied on methods such as rule-based approaches and viseme-based techniques [12]. However, the emergence of deep neural networks (DNNs) marked a turning point, triggering a surge in deep learning-based methods. This has significantly advanced the field, unlocking exciting possibilities for speech-related applications.

Early Attempts: One of the early attempts was **ObamaNet** [13], a groundbreaking architecture that utilized a time-delayed Long Short-Term Memory (LSTM) network to generate synchronized lip-sync videos from text input. However, ObamaNet suffered from a significant limitation - speaker dependency. This meant it could only effectively generate lip-sync for a single speaker, in this case, former US President Barack Obama. Another significant model, **Speech2Vid** [14], aimed to generate talking face videos from still images and speech segments. While not inherently speaker-dependent, its reliance on a large audio-visual dataset introduced language dependency concerns.

Wav2Lip and its Derivatives: The pursuit of seamless lip synchronization in videos has fueled the development of sophisticated models, with **Wav2Lip** emerging as a game-changer. This speaker-independent framework, an evolution of **LipGAN** [10], overcomes its predecessor's limitations by incorporating temporal context and a pre-trained discriminator. LipGAN, while utilizing audio and face encoders, a decoder, and a discriminator, was hindered by speaker dependency, limited accuracy, and a focus on visual artifacts.

Wav2Lip addresses these shortcomings by leveraging five consecutive frames for temporal context and employing a SyncNet-based pre-trained discriminator [15]. This strategic shift from lip region reconstruction loss to audio-lip correspondence significantly improves accuracy, ensuring lips move in perfect harmony with the audio, regardless of the speaker.

The success of Wav2Lip spurred the development of numerous derivative models focused on various aspects of improvement. Two models built upon Wav2Lip aimed to enhance lip-sync video quality using super-resolution image processing frameworks. **Wav2Lip-HD** [16] utilizes **Real-ESRGAN** [17] to improve frame quality after inference, achieving excellent results. Alternatively, **Wav2Lip-HR** [18] addresses these issues by focusing on training a powerful model rather than post-processing improvements. It recognizes the importance of high-quality training data and leverages **GFP-GAN** [19], another super-resolution model, to enhance the training dataset itself. This allows the model to inherently generate higher-quality lip-sync videos with accurate synchronization, eliminating the need for post-processing and maintaining real-time usability. Unlike models focused on enhancing lip-sync quality, the **compressed-**

Wav2Lip [20] addresses computational limitations hindering real-world applications.

The quest for even more refined lip synthesis has led to the integration of attention mechanisms. [21] exemplify 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** [22], which builds upon AttnWav2Lip by incorporating the LPIPS loss function. Similarly, the **CA-Wav2Lip** [23] 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.

Other Approaches: Beyond attention mechanisms, researchers have explored a variety of techniques and architectures to enhance lip-sync performance. One study investigates a Transformer-based model called **AV-CAT** [24], which capitalizes on the Transformer's ability to capture long-range dependencies within both audio and visual features. **Diff2Lip** [26] utilizes diffusion models to address lip-sync tasks while aiming for improved visual quality compared to traditional Generative Adversarial Networks (GANs). Another approach is **MakeItTalk** [27], which offers a simple and realistic audio-driven option using an LSTM network with a self-attention encoder.

More recently, researchers have turned their attention to neural radiance fields (NeRFs) as a promising avenue for lip synchronization. Unlike GAN-based methods, NeRFs represent scenes using a 3D volume, enabling more accurate modeling of facial geometry. **AD-NeRF** [28], an audio-driven NeRF model, addresses cross-modal mapping issues without additional intermediate representations, facilitating seamless talking head video editing, including pose manipulation and background replacement, which is crucial for virtual reality applications. However, AD-NeRF initially struggled with dynamic scenes and motion blur, leading to the development of **DFA-NeRF** [29]. DFA-NeRF introduces deformation capabilities to handle dynamic changes, utilizing a hierarchical architecture for efficient representation of complex scenes, making it ideal for dynamic object and scene generation.

Despite these advancements, NeRF-based approaches, including DFA-NeRF, face limitations in lip synchronization compared to GAN-based methods. In response to this challenge, **LipNeRF** [30] was introduced, capable of high-quality lip-syncing trained on cinematic content, even with minimal video data, offering promising potential for lip synchronization. **NeRF-AD** [31], surpasses LipNeRF in capturing the full spectrum of facial expressions. While LipNeRF excels in lip-syncing, it may neglect other aspects. NeRF-AD addresses this through attention-based disentanglement, separating the face into regions for audio-driven lip movements and preserving speaker identity.

This paper delves into the universality of the Wav2Lip lip-sync model across various languages, contrasting with the focus of the referenced studies on specific

technological and computational methodologies. The authors of [32] primarily concentrates on optimizing user interface testing through graph data analysis within big data projects. This is different from the lip-sync evaluation across languages presented in this paper. The authors of [33] surveys the structural code clone detection problem, a study that is more aligned with software engineering and code analysis, whereas the Wav2Lip paper investigates the application of lip-sync models in language translation and communication, showcasing the diversity in applications of machine learning and artificial intelligence. The authors of [34] discuss a distinct area of data processing and management compared to the linguistic and communicative focus of this paper, illustrating the varied use cases of data technologies. The authors of [35] and [36] emphasize user behavior analysis and data representation, differing from the exploration of lip-sync models' adaptability across languages, thus highlighting the cross-disciplinary nature of contemporary computational research. The authors of [37,44–48] discuss grid and web services, focusing on infrastructure and computational support, which is a contrast to the linguistic and phonetic analysis undertaken in this study, underscoring the breadth of research within the field of computer science and technology.

In conclusion, advancements in lip synchronization driven by deep learning methods and innovative models like **Wav2Lip** have revolutionized the field. Despite challenges such as speaker dependency and computational constraints, these developments hold great promise for more accurate and seamless lip-syncing, benefiting various applications in computer vision and multimedia.

3 Methodology

This section details the methodology employed in our research, encompassing two key contributions. First, we evaluate the performance of the Wav2Lip model in generating lip-synchronized talking faces across multiple languages. Second, we propose a novel Face-to-Face translation workflow leveraging recent advancements. We hypothesize that the Wav2Lip model, despite being trained on an English dataset, can effectively generate lip-synchronized talking faces for various languages due to its claimed language-independent nature. We follow this methodology to evaluate the claim.

3.1 Wav2Lip Overview

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 & 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.

3.2 Data Collection

For our evaluation, we collected a dataset that encompasses three languages: **Turkish**, **Persian**, and **Arabic**. Each language belongs to a distinct language family (Turkic, Indo-Iranian, and Semitic, respectively), ensuring dataset diversity and providing a robust test for Wav2Lip's language independence. These languages, although widely spoken globally, haven't received significant attention in lip-sync research. Addressing this gap is crucial as the Middle East region, where these languages dominate, represents a rapidly growing market with increasing technological adoption and digital content creation. Evaluating Wav2Lip on these languages can facilitate the development of localized applications, enhance accessibility for speakers of these languages, and drive innovation in communication technologies.

For Turkish and Arabic, we selected the **MediaSpeech** [38] dataset, comprising short speech segments automatically extracted from media videos available on YouTube. For the Persian language, we opted for the **Persian Speech Corpus** [39], recorded in a Tehrani accent using a professional studio setup. Our evaluation dataset consists of one-hour audio for each language, randomly selected from the original datasets, providing a diverse and representative sample for evaluation purposes.

3.3 Video Generation

Wav2Lip 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 the "Data Collection" subsection.

2) Target Image/Video: This input provides the visual foundation for which the lip movements will be generated. Wav2Lip can accept either a static image (single frame) or a video (multiple frames). We opted to utilize static images for this study. This decision was primarily driven by two factors. Firstly, our primary focus was on analyzing the movements of the mouth region, specifically the lips. Static images provided a clear and isolated view of this area, facilitating a more focused evaluation. Secondly, generating lip movements for videos requires significantly more computational resources compared to static images.

To generate the videos, we employed an Nvidia A100 (40GB) GPU. This powerful GPU was chosen due to its ability to handle the computationally intensive nature of Wav2Lip processing.

3.4 Evaluation

As discussed earlier, lip-sync evaluation metrics can be broadly categorized into **Lip-Sync Accuracy** and **Visual Quality** metrics. With advancements in image and video processing technologies like super-resolution and image quality enhancement, coupled with pre-processing techniques mentioned in the Related Work section, it becomes even more critical to prioritize Lip-Sync Accuracy (Error) metrics. Established metrics like **SSIM**, **PSNR**, and **FID** are not lip-sync specific and primarily assess overall image quality. The **SSIM** (Structural Similarity Index Metric) quantifies the similarity between two images based on their luminance, contrast, and structure, providing a measure of perceptual quality and detecting artifacts or distortions. The **PSNR** (Peak Signal-to-Noise Ratio) compares the maximum possible power of a signal (e.g., original image) to the power of noise introduced by distortion (e.g., synthesis), offering a measure of overall fidelity and noise level in synthesized content. The **FID** (Fréchet Inception Distance) measures the similarity between distributions of feature representations of real and generated images to evaluate the overall visual quality and diversity of synthesized content.

The Wav2Lip authors 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 distance between the lip and audio representations. 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 the following Eq. 1:

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

where $\boldsymbol{A(t)}$ **:** Audio feature vector at time step $\boldsymbol{t}$ and $\boldsymbol{V(t)}$ **:** Lip movement feature vector at time step $\boldsymbol{t}$ and $\boldsymbol{S(A(t), V(t))}$ **:** Similarity score between $\boldsymbol{A(t)}$ and $\boldsymbol{V(t)}$.

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 the following Eq. 2:

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

where $\boldsymbol{\sigma}$ **:** represents standard deviation and $\boldsymbol{A(t)}$ **:** Audio feature vector at time step $\boldsymbol{t}$ and $\boldsymbol{V(t)}$ **:** Lip movement feature vector at time step $\boldsymbol{t}$ and $\boldsymbol{S(A(t))}$ **:** represents the similarity score between $\boldsymbol{A(t)}$ and $\boldsymbol{V(t)}$ at each time step $\boldsymbol{t}$.

We leverage a pre-trained **SyncNet** model to calculate these two metrics for our evaluation dataset. SyncNet boasts an accuracy of around 99% [15], making it a powerful tool for measuring lip-sync error. It's 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 to generate talking faces using the evaluation dataset and a randomly generated avatar from DALL-E. Subsequently, we employ the provided method to calculate LSE-D and LSE-C for the generated videos.

In addition to this **quantitative** evaluation approach, we will incorporate **qualitative** evaluation by human observer into our research. While SyncNet and the LSE metrics offer a robust system for assessing lip-sync accuracy, potential biases towards specific languages may exist, considering SyncNet's potential training on primarily English or a limited set of languages. Therefore, conducting human evaluation serves as a complementary approach.

Similar to quantitative evaluation, selecting appropriate metrics or criteria for human observers is crucial in qualitative evaluation. When assessing lip-synced videos, human observers can consider various criteria. However, for this study, which evaluates Wav2Lip's performance across different languages, choosing criteria sensitive to the input audio's language is essential. For instance, a criterion like "Artifacts", which refers to visual distortions or irregularities in the video, is not suitable. This is because artifacts depend on the model's architecture rather than the spoken language. Therefore, selecting language-dependent criteria for qualitative evaluation is critical.

We identified and selected two suitable criteria for this task: Adequacy and Naturalness. In this context, **Adequacy** refers to how well the lip movements match the spoken words and sounds in the audio. A lip-synced video is considered adequate if the lip movements closely align with the phonetic content of the speech, effectively conveying the intended words and sounds. Adequacy essentially assesses the accuracy of the lip-syncing process in visually reproducing the spoken content. **Naturalness** refers to the degree of realism and lifelikeness in the synthesized video's lip movements, resembling natural movements observed in real-life speech. For lip-syncing, naturalness evaluates whether the lip movements are fluid, smooth, and contextually appropriate, enhancing the overall believability and authenticity of the video. A lip-synced video is considered natural if the lip movements align with the rhythm, pacing, and intonation of the spoken words, creating a seamless and convincing visual representation of speech.

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 Workflow

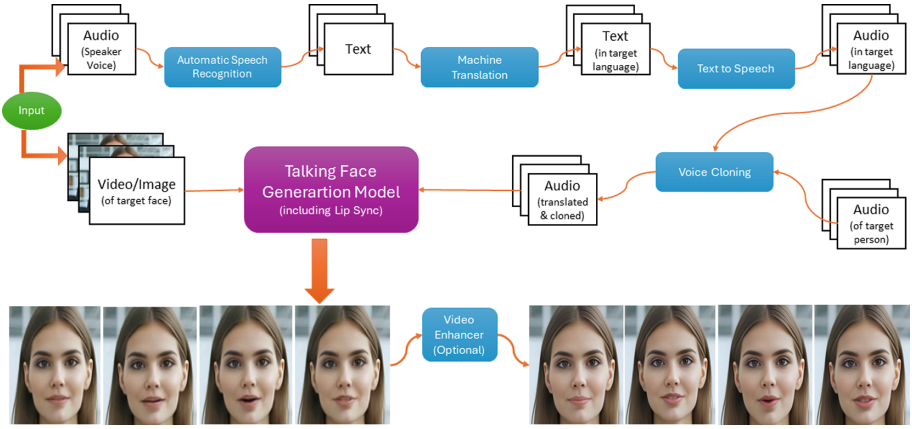

Fig. 1. Face-to-Face Translation Workflow

The concept of Face-to-Face translation entails translating all communication aspects of a speaker, including language, voice, mimics, and lip movements, into another language. The modular workflow in Fig. 1 illustrates how this can be achieved. Firstly, the speaker's voice is captured and recognized by an **Automatic Speech Recognition (ASR)** tool. The resulting text output is then translated to the target language using **Machine Translation (MT)**. Subsequently, employing a **Text-to-Speech (TTS)** model along with a **Voice Cloning (VC)** model, the speaker's voice is translated into the target language, constituting the **Speech-to-Speech (S2S)** translation step. After S2S, the generated audio and the target face (which may belong to the speaker or another person or avatar) are inputted into the **Talking Face Generation** model, with Lip Sync as a crucial component. Finally, a talking face is generated, and if necessary, an **Image/Video Enhancer** framework can be employed to improve the visual quality of the generated video, addressing potential issues arising from low-quality input images/videos, weaknesses in the talking face generation model, or insufficient computational resources.

While such a system incorporating all these modules is achievable considering current advancements, it's important to acknowledge the significant computational resources it would require. While the applications for such a system are vast, encompassing both real-time and non-real-time use cases, there is still a significant need for further improvements across all the modules. This study specifically focuses on the Talking Face Generation module within this larger workflow. We present this comprehensive workflow to provide a valuable outline for researchers, companies, and individuals interested in developing a Face-to-Face Translation system. We believe such a system has the potential to revolutionize the way we communicate

4 Experiments and Results

In this section, we evaluate the performance of Wav2Lip on generating lip-synced videos for languages beyond English. Initially, we generated approximately one hour of talking face videos for each language (Persian, Arabic, and Turkish) using Wav2Lip. The model configuration remained consistent, producing videos at 25 fps and a resolution of 96×96 pixels for the lips. Subsequently, we carried out a thorough assessment encompassing both quantitative and qualitative evaluation.

4.1 Quantitative Evaluation

To assess the quality of the generated lip-syncing, we employed the pre-trained SyncNet model and calculated two metrics: LSE-D (distance) and LSE-C (confidence). It's 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 [8] dataset. The pre-trained SyncNet achieves a reported accuracy of 99%, suggesting reliable evaluation results. Table 1 presents the quantitative evaluation results. Persian emerges as the language with the best scores in both LSE-D (distance) and LSE-C (confidence) by a significant margin. Arabic follows closely in second place, while Turkish exhibits the lowest scores.

Table 1. Quantitative evaluation results

Language	Length (minutes)	LSE-D ↓	LSE-C ↑
Turkish	67	7.088657544	7.265054778
Persian	59	6.235052688	8.389725332
Arabic	54	7.011181367	7.410812804

We acknowledge that previous studies and models have utilized various datasets, primarily focused on English (LRW, LRS2, LRS3, VoxCeleb2) [40] [41]. However, a comparison of their results with ours suggests promising performance by Wav2Lip for Persian, Turkish, and Arabic languages. Despite being trained on an English dataset, Wav2Lip demonstrates remarkable performance and accuracy across these different languages, validating the claim of its language-independence, at least within the context of our selected languages.

4.2 Qualitative Evaluation

While our quantitative evaluation using SyncNet provides valuable insights, it's crucial to acknowledge its limitations. Firstly, SyncNet's potential bias towards English or similar languages could lead to underestimating the quality of lip-syncing for languages like Persian, Turkish, and Arabic, overlooking their specific pronunciation nuances. Secondly, quantitative metrics might struggle with

subtleties in natural lip movements, prioritizing high similarity scores even if the generated movements appear unnatural for a particular language. Finally, the recording environment can impact results. In our case, the Persian dataset's superior performance might be attributed to its professional studio recording with minimal noise and background sounds. The Turkish and Arabic datasets come from news broadcasts, where environments are less controlled compared to studios. These factors highlight the importance of human evaluation alongside quantitative measures.

Human evaluators, ideally native speakers or those highly familiar with the languages, can assess the generated lip movements' alignment with natural pronunciation and mouth shapes specific to each language. For instance, Persian has distinct sounds like "qaf" and "khe"that require unique lip movements not commonly seen in English. Similarly, Turkish vowel harmony influences pronunciation and potentially lip shapes across words. Lastly, emphatic consonants in Arabic involve forceful pronunciation that translates to distinct lip movements. Human evaluation ensures the generated lip movements not only achieve technical accuracy but also appear natural, culturally appropriate, and capture the nuances of the target languages.

In our study, native speakers of each language evaluated a subset of generated videos for **Adequacy** and **Naturalness** criteria.

In terms of **Adequacy**, Persian evaluators reported excellent synchronization for most of the video, including the distinct sounds not found in English. Turkish evaluations indicated generally well-generated and accurate lip movements with some subtle errors. Interestingly, users noted instances of faster-than-usual lip movement transitions, potentially due to the vowel harmony characteristic of Turkic languages. Arabic reviewers reported high lip-sync accuracy even with forceful pronunciations, although some errors were present. Additionally, a participant fluent in all three languages evaluated the videos. Their assessment confirmed the best adequacy for Persian with minimal errors, followed by Turkish and then Arabic.

While the human evaluation indicated good adequacy, the **Naturalness** scores were generally lower. This is interesting because the evaluated videos featured an artificial talking avatar, and users likely didn't expect perfectly realistic lip movements. Despite this, observers reported weaknesses in motion and pacing. Most observers noted that in some instances, even though lip movements precisely matched the audio timing and accuracy (adequacy), they appeared overly responsive to the speech. This means the lips reacted to even subtle sounds that wouldn't necessarily require a visible change in lip position. Consequently, the videos sometimes exhibited faster-than-usual and unexpected transitions, impacting the overall naturalness.

Finally, participants rated the Lip Synced videos based on the selected criteria on a scale from 1 to 5. Results of the user rating is provided in the Table 2.

Table 2. Qualitative evaluation by human observer results

Language	No. of Participants	Length	Adequacy	Naturalness	Overall
Turkish	12	3 min	4.2/5.0	3.4/5.0	**3.8/5.0**
Persian	7	3 min	4.6/5.0	3.6/5.0	**4.1/5.0**
Arabic	6	3 min	4.1/5.0	3.3/5.0	**3.7/5.0**

These human evaluations complement the quantitative results, highlighting the strengths and weaknesses of Wav2Lip's performance across different languages. They emphasize the importance of considering language-specific nuances beyond solely relying on metrics potentially biased towards certain languages.

5 Discussion

Our experiments and results demonstrate that Wav2Lip, despite being trained on an English dataset, exhibits remarkable lip-sync accuracy across our chosen languages (Turkish, Persian, and Arabic). This reinforces the model's language-independence claims, as it achieves comparable performance to English within our language set. The scarcity of research exploring the use of lip-sync models for diverse languages is noteworthy. One study [9] 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 [25] 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.

The potential of an end-to-end Face-to-Face Translation system encompassing all the modules in our proposed workflow is immense due to its vast range of applications. Advancements in neural networks have yielded well-established models for each module, catering to both real-time scenarios where efficiency reigns supreme and high-quality setups prioritizing performance. Currently, lip-sync remains the most crucial component within a Face-to-Face translation system. While several models prioritize visual quality, this often comes at the expense of lip-sync accuracy [26,31,42]. Wav2Lip and its derivatives excel in generating highly synchronized lips, but may necessitate post-processing with video enhancement models like Real-ESRGAN or GFP-GAN, which require significant computational power and time. Pre-processing techniques, such as those employed in [18], also exist. They utilize a modified Wav2Lip model trained on a pre-processed 4K dataset achieved through image enhancement models. Training Wav2Lip on high-resolution datasets directly generates high-resolution videos, eliminating the lengthy inference times associated with post-processing. However, the computational cost of image enhancement models, even during pre-processing, remains high. Therefore, the development of high-resolution speech datasets is crucial.

To the best of our knowledge, there are no existing real-time, end-to-end models encompassing all the steps in our workflow. While a recent work by [43] incorporates intriguing modules like face swapping, creating a user-friendly, end-to-end system with all the proposed steps, its real-time performance falls short, potentially requiring hours to generate a 10-minute talking face even on high-end hardware.

An end-to-end Face-to-Face Translation system that effectively balances factors such as lip-sync accuracy, visual quality, computational cost, and inference time would be groundbreaking, revolutionizing communication methods. Imagine real-time systems within the Metaverse or video conferencing platforms, enabling seamless communication between individuals across languages, utilizing any desired face or avatar. Considering the current state of these technologies and advancements in hardware and processing units, achieving such a system seems like a foreseeable future reality. By overcoming the challenges and limitations discussed, we pave the way for the development of such a transformative system.

6 Conclusion and Future Works

In conclusion, our study demonstrates the promising performance of existing Lip Sync models, particularly Wav2Lip, in multilingual environments. Despite variations across languages, our evaluation of Wav2Lip in Turkish, Persian, and Arabic showcases its adequacy both quantitatively and qualitatively. By outlining a comprehensive workflow, we provide clarity on the essential components required for Face-to-Face Translation systems. Our experiments underscore the potential of language-agnostic Lip Sync models like Wav2Lip in integrating seamlessly within such workflows.

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. These insights can guide future research efforts and empower researchers, developers, and companies to contribute to this evolving field.

In the broader context, our findings offer valuable insights that can pave the way for the realization of end-to-end, real-time solutions for Face-to-Face translation systems. Such advancements hold immense potential to revolutionize communication methods significantly, marking a pivotal shift in how individuals interact and connect with one another.

Acknowledgement. The authors would like to thank Aktif Bank for supporting this study by providing the necessary working environment.

References

1. Toshpulatov, M., et al.: Talking human face generation: a survey. Expert Syst. Appl., **219**, 119678 (2023). https://doi.org/10.1016/j.eswa.2023.119678
2. A survey on the metaverse: the state-of-the-art, technologies, applications, and challenges(2023). https://doi.org/10.1109/jiot.2023.3278329
3. Kato, R., et al.: Reality avatar for customer conversation in the metaverse. https://doi.org/10.1007/978-3-031-06509-5_10
4. Ali, I.R., et al.: Realistic lip syncing for virtual character using common viseme set. Comput. Inf. Sci. **8**(3), (2015). https://doi.org/10.5539/CIS.V8N3P71
5. Bennett Terry, K., Radhakrishnan, R.: Detection and correction of lip-sync errors using audio and video fingerprints. SMPTE Motion Imaging J. **119**(3), 42–52 (2010). https://doi.org/10.5594/J11398
6. Fenghour, S., et al.: Deep learning-based automated lip-reading: a survey. IEEE Access. **9**, 121184–121205 (2021). https://doi.org/10.1109/ACCESS.2021.3107946
7. Prajwal, K.R., et al.: A lip sync expert is all you need for speech to lip generation in the wild. In: Presented at the August 23 (2020). https://doi.org/10.1145/3394171.3413532
8. Mroueh, Y., et al.: Deep multimodal learning for audio-visual speech recognition. In: Presented at the April 19 (2015). https://doi.org/10.1109/ICASSP.2015.7178347
9. Song, H.K., et al.: Talking face generation with multilingual TTS. In: Presented at the May 13 (2022). https://doi.org/10.1109/CVPR52688.2022.02074
10. KR, P., et al.: Towards automatic face-to-face translation. In: Presented at the March 1 (2020). https://doi.org/10.1145/3343031.3351066
11. Kadam, A., et al.: A survey of audio synthesis and lip-syncing for synthetic video generation. EAI Endorsed Trans. Creative Technol. **8**(28), e2–e2 (2021). https://doi.org/10.4108/EAI.14-4-2021.169187
12. Llorach, G., et al.: web-based live speech-driven lip-sync. In: Presented at the September 1 (2016). https://doi.org/10.1109/VS-GAMES.2016.7590381
13. ObamaNet: photo-realistic lip-sync from text (2018). https://doi.org/10.48550/arXiv.1801.01442
14. Jamaludin, A., et al.: You said that?: Synthesising talking faces from audio. Inter. J. Comput. Vision **127**(11), 1767–1779 (2019). https://doi.org/10.1007/S11263-019-01150-Y
15. Chung, J.S., Zisserman, A.: Out of Time: Automated Lip Sync in the Wild. https://doi.org/10.1007/978-3-319-54427-4
16. saifhassan/Wav2Lip-HD. https://github.com/saifhassan/Wav2Lip-HD. Accessed 26 Mar 2024
17. Wang, X., et al.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: Presented at the July 22 (2021). https://doi.org/10.1109/ICCVW54120.2021.00217
18. Wav2Lip-HR: synthesising clear high-resolution talking head in the wild. https://onlinelibrary.wiley.com/doi/10.1002/cav.2226. Accessed 26 Mar 2024
19. Wang, X., et al.: Towards real-world blind face restoration with generative facial prior. In: Presented at the June 20 (2021). https://doi.org/10.1109/CVPR46437.2021.00905
20. Kim, B.K., et al.: A unified compression framework for efficient speech-driven talking-face generation. abs/2304.00471 (2023). https://doi.org/10.48550/arXiv.2304.00471

21. Wang, G., et al.: Attention-based lip audio-visual synthesis for talking face generation in the wild. abs/2203.03984 (2022). https://doi.org/10.48550/arXiv.2203.03984
22. Chen, Z., et al.: LPIPS-AttnWav2Lip: generic audio-driven lip synchronization for talking head generation in the wild. **157**, 103028 (2024). https://doi.org/103028
23. Wang, K.C., et al.: CA-Wav2Lip: coordinate attention-based speech to lip synthesis in the wild. In: 2023 IEEE International Conference on Smart Computing (SMARTCOMP) (2023). https://doi.org/10.1109/SMARTCOMP58114.2023.00018
24. Sun, Y., et al.: Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In: Presented at the November 29 (2022). https://doi.org/10.1145/3550469.3555393
25. Patel, D., et al.: Visual dubbing pipeline with localized lip-sync and two-pass identity transfer. **110** (2022). https://doi.org/10.1016/j.cag.2022.11.005
26. Mukhopadhyay, S., et al.: Diff2lip: audio conditioned diffusion models for lip-synchronization, 5292–5302 (2024)
27. Zhou, Y., et al.: MakeltTalk: speaker-aware talking-head animation. ACM Trans. Graph. (TOG) **39**(6), 1–15 (2020). https://doi.org/10.1145/3414685.3417774
28. Guo, Y., et al.: AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In: Presented at the March 20 (2021)
29. Yao, S., et al.: DFA-NeRF: personalized talking head generation via disentangled face attributes neural rendering. abs/2201.00791 (2022)
30. Chatziagapi, A., et al.: LipNeRF: what is the right feature space to lip-sync a NeRF?. In: Presented at the January 5 (2023). https://doi.org/10.1109/FG57933.2023.10042567
31. Bi, C., et al.: NeRF-AD: neural radiance field with attention-based disentanglement for talking face synthesis (2024)
32. Uygun, Y., et al.: On the large-scale graph data processing for user interface testing in big data science projects. In: 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, pp. 2049–2056, (2020). https://doi.org/10.1109/BigData50022.2020.9378153.
33. Kapdan, M., Aktas, M., Yigit, M.: On the structural code clone detection problem: a survey and software metric based approach. In: Murgante, B., et al. (eds.) ICCSA 2014. LNCS, vol. 8583, pp. 492–507. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09156-3_35
34. Pierce, M.E., et al.: The QuakeSim project: web services for managing geophysical data and applications. In: Tiampo, K.F., Weatherley, D.K., Weinstein, S.A. (eds) Earthquakes: Simulations, Sources and Tsunamis . Pageoph Topical Volumes. Birkhäuser Basel. https://doi.org/10.1007/978-3-7643-8757-0_11
35. Olmezogullari, E., et al.: Pattern2Vec: representation of clickstream data sequences for learning user navigational behavior. Concurrency Computat Pract Exper. **34**(9), e6546 (2022). https://doi.org/10.1002/cpe.6546
36. Olmezogullari, E., et al.: Representation of click-stream datasequences for learning user navigational behavior by using embeddings In: 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, pp. 3173–3179 (2020). https://doi.org/10.1109/BigData50022.2020.9378437.
37. Nacar, M.A., et al.: VLab: collaborative grid services and portals to support computational material science. Concurrency Comput. Pract. Exper. **19**(12), 1717–1728 (2007). https://doi.org/10.1002/cpe.1199
38. MediaSpeech: Multilanguage ASR Benchmark and Dataset. Accessed 30 Mar 2021

39. Persian Speech Corpus. https://fa.persianspeechcorpus.com/. Accessed 26 Mar 2024
40. Chung, J.S., et al.: VoxCeleb2: deep speaker recognition. In: Presented at the June 14 (2018). https://doi.org/10.21437/INTERSPEECH.2018-1929
41. LRS3-TED: a large-scale dataset for visual speech recognition. Accessed 03 Sep 2018
42. Guan, J., et al.: StyleSync: high-fidelity generalized and personalized lip sync in style-based generator. abs/2305.05445 (2023). https://doi.org/10.48550/arXiv.2305.05445
43. Wav2Lip UHQ. https://github.com/numz/sd-wav2lip-uhq. Accessed 26 Mar 2024
44. Aktas, M.S., et al.: Fault tolerant high performance information services for dynamic collections of grid and web services. Future Gener. Comput. Syst. **23**(3), 317–337 (2007). https://doi.org/10.1016/j.future.2006.05.009
45. Aydin, G., et al.: Building and applying geographical information system grids. Concurrency Computat. Pract. Exper., **20**, 1653–1695 (2008). https://doi.org/10.1002/cpe.1312
46. Fox, G.C., et al.: Algorithms and the grid. Comput. Visual Sci. **12**, 115–124 (2009). https://doi.org/10.1007/s00791-007-0083-8
47. Aktas, M., et al.: iSERVO: implementing the international solid earth research virtual observatory by integrating computational grid and geographical information web services. Pure Appl. Geophys. **163**, 2281–2296 (2006). https://doi.org/10.1007/s00024-006-0137-8
48. Aydin, G., et al.: SERVOGrid complexity computational environments (CCE) integrated performance analysis. In: The 6th IEEE/ACM International Workshop on Grid Computing, 2005. Seattle, WA, USA, 6–pp (2005). https://doi.org/10.1109/GRID.2005.1542750.