Multilingual video dubbing—a technology review and current challenges

frontiers | Frontiers in Signal Processing

TYPE Mini Review
PUBLISHED 25 September 2023
DOI 10.3389/frsip.2023.1230755

OPEN ACCESS

EDITED BY
Feng Yang,
Google, United States

REVIEWED BY
Xinwei Yao,
Google, United States
Keren Ye,
Google, United States
Seung Hyun Lee,
Korea University, Republic of Korea
Xi Chen,
Rutgers, The State University of New Jersey, United States

*CORRESPONDENCE
Dan Bigioi,
d.bigioi1@universityofgalway.ie
Peter Corcoran,
peter.corcoran@universityofgalway.ie

RECEIVED 29 May 2023
ACCEPTED 11 September 2023
PUBLISHED 25 September 2023

CITATION
Bigioi D and Corcoran P (2023), Multilingual video dubbing—a technology review and current challenges. *Front. Sig. Proc.* 3:1230755. doi: 10.3389/frsip.2023.1230755

Multilingual video dubbing—a technology review and current challenges

Dan Bigioi* and Peter Corcoran*

School of Engineering, University of Galway, Galway, Ireland

The proliferation of multi-lingual content on today's streaming services has created a need for automated multi-lingual dubbing tools. In this article, current state-of-the-art approaches are discussed with reference to recent works in automatic dubbing and the closely related field of talking head generation. A taxonomy of papers within both fields is presented, and the main challenges of both speech-driven automatic dubbing, and talking head generation are discussed and outlined, together with proposals for future research to tackle these issues.

KEYWORDS

talking head generation, dubbing, deep fakes, deep learning, artificial intelligence, video synthesis, audio video synchronisation

1 Introduction and background

The problem of Video dubbing is not a recent challenge. Looking back through the literature in Cao et al. (2005) the authors discuss the complexity of mimicking facial muscle movements and note that data-driven methods had yielded some of the most promising results at that time, almost two decades ago. More recently Mariooryad and Busso (2012) synthesized facial animations based on the MPEG-4 facial animation standard, using the audiovisual IEMOCAP database (Busso et al., 2008). While the Face Animation Parameters (FAP) defined in MPEG are useful, such model based approaches are no longer considered as state of the art (SotA) for photo-realistic speech dubbing or facial animation. Nevertheless, these earlier works attest to long-standing research on speech-driven facial re-enactment in the literature.

Today, there have been many new advances in facial rendering and acoustic and speech models. The requirements of video dubbing are mainly driven by the evolution of the video streaming industry (Hayes and Bolanos-Garcia-Escribano, 2022) and will be the focus of this review. The rapid growth of streaming services and the resulting competition has led to a proliferation of new content, with a significant growth in non-English language content and a global expansion of audiences to existing and new non-English speaking audiences and markets. Much of the success of the leading content streaming services lies in delivering improved quality of content to these new markets with a need for more sophisticated and semi-automated subtitle and dubbing services.

Subtitle services are well-developed and provide a useful bridge to the growing libraries of video content for non-English audiences. The leading services have also begun to release new content with dubbing in multiple languages and to annotate and dub legacy content as well (Roxborough, 2019; NILESH and DECK, 2023). Auto-translation algorithms can help here, but typically human input is also needed to refine the quality of the resulting translations.

When content is professionally dubbed a voice actor will carefully work to align the translated text with the original actors facial movements and expressions. This is a challenging and skilled task and it is difficult to find multi-lingual voice actors, so often only the lead actors in a movie will be professionally overdubbed. This creates an "uncanny valley" effect for most overdubbed content which detracts from the viewing experience and it is often preferable to view content in the original language with subtitles. Thus the overdubbing of digital content remains a significant challenge for the video streaming industry (Spiteri Miggiani, 2021).

For the best quality of experience in viewing multi-lingual content it is desirable not only to overdub the speech track for a character, but also to adjust their facial expressions, particularly the lip and jaw movements to match the speech dubbing. This requires a subtle adjustment of the original video content for each available language track, ensuring that while the lip and jaw movements change in response to the new language track, the overall performance of the original language actor is not diminished in any way. But achieving this seamless audio driven automatic dubbing is a non-trivial task, with many approaches proposed over the last half-decade tackling this problem. Deep learning techniques especially have proven popular in this domain (Yang et al., 2020; Vougioukas et al., 2020; Thies et al., 2020; Song et al., 2018; Wen et al., 2020), demonstrating compelling results on the tasks of automatic dubbing, and the lesser constrained, more well-known task of "talking head generation."

In this article, current state-of-the-art approaches are discussed with reference to the most recent and relevant works in automatic dubbing and the closely related field of talking head generation. A taxonomy of papers within both fields is presented, and current SotA for both audio-driven automatic dubbing, and talking head generation are discussed and outlined. Recent approaches can be broadly classified as falling within two main schools of thought: end-to-end, or structural-based generation (Liang et al., 2022). It is clear from this review that much of the foundation technology is now available to tackle photo-realistic multilingual dubbing, but there are still remaining challenges which we seek to define and clarify in our concluding discussion.

2 The high-level dubbing pipeline

Traditionally, dubbing is a costly post-production affair that consists of three primary steps:

- Translation: This is the process of taking the script of the original video, and translating it to the desired language(s). Traditionally, this is done by hiring multiple language experts, fluent in both the original, and target languages. With the emergence of large language models in recent years however, accurate automatic language to language translation is becoming a reality (Duquenne et al., 2023), and has been adopted into industry use as early as 2020 by the likes of Netflix (Alarcon, 2023). That being said, the models are not perfect and are susceptible to mistranslations, therefore to ensure quality an expert is still required to look over the translated script.
- Voice Acting: Once the scripts have been translated, the next step is to identify and hire suitable voice actors for each of the desired languages. For a high quality dub, care must be taken to ensure that the voice actors can accurately portray the range of emotions of the original recording, and that their voices suitably match the on-screen character. This is a costly, and time-consuming endeavour, and would benefit immensely from automation. Despite incredible advances in text-to-speech, and voice-cloning technologies in recent years, a lot of work still remains to be able to truly replicate the skill of a professional voice actor (Weitzman, 2023), however for projects where quality is not as important, text to speech is an attractive option due to its reduced cost.
- Audio Visual Mixing: As soon as the new language voice recordings are obtained, the final step is to combine them with the original video recording in as seamless a manner as possible. Traditionally this involves extensive manual editing work in order to properly align and synchronise the new audio to the original video performance. Even the most skilled of editors however cannot truly synchronise these two streams. High quality dubbing work is enjoyable to watch yet oftentimes it is still noticeable that the content is dubbed. Poor quality dubbing work detracts from the user experience, oftentimes inducing the "uncanny-valley" effect in viewers.

Due to the recent advancements in deep learning, there is scope for automation in each of the traditional dubbing steps. Manual language translation can be carried out automatically by large language models such as Duquenne et al. (2023). Traditional voice acting can be replaced by powerful text to speech models such as Łańcucki (2021); Liu et al. (2023); Wang et al. (2017). Audio-visual mixing, can then be carried out by talking head generation/video editing models such as Zhou et al. (2020). Given the original video and language streams, the following is an example of what such an automatic dubbing pipeline might look like for dubbing an English language video into German:

- Transcribing and Translating Source Audio: Using an off-the-shelf automatic speech recognition model, an accurate transcript can be produced from the speech audio. The English transcript can then be translated into German using a large language model such as BERT or GPT3 finetuned on the language to language translation task.
- Synthesizing Audio: Synthetic speech can be produced by leveraging a text to speech model, taking the translated transcript as input, and outputting realistic speech. Ideally the model would be finetuned on the original actors voice, and produce high quality speech that sounds just like the original actor but in a different language.
- 3D Character Face Extaction: From the video stream, detect and isolate the target character. Map the target characters face onto a 3D morphable model using monocular 3D reconstruction, and isolate the headpose/global head movement, obtaining a static 3D face. Remove the original lip/jaw movements, but retain the overall facial expressions and eye blinks on the character model.
- Facial Animation Generation: Generate the expression parameters corresponding to the lip and jaw movements on the 3D face model in response to the driving synthetic German audio speech signal via a recurrent neural network. Introduce

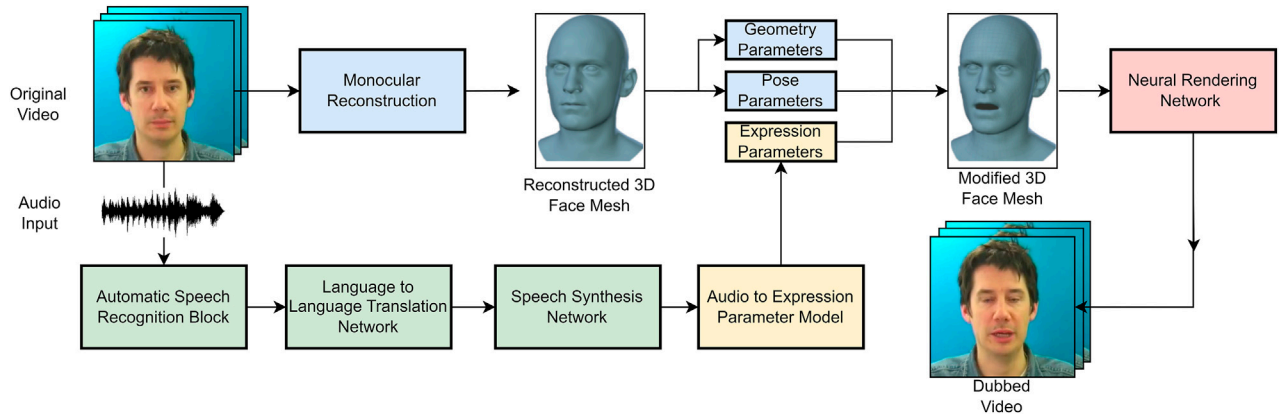

FIGURE 1
A high-level diagram depicting the automatic dubbing process described in this section. 3D model image is taken from the work of Cudeiro et al. (2019), while the subject displayed is part of the Cooke et al. (2006) dataset.

the global head movement information back to the 3D model to obtain a 3D head whose facial expressions and head pose correspond to the original performance, but with the lips and jaws modified in response to the new audio.

- Rendering: Mask out the facial region of the character in the original video, insert the newly generated 3D face model on top, and utilise an image-to-image translation network to generate the final photorealistic output frames.

The hypothetical pipeline described above is known as a structural-based approach, and is Figure 1. The next section shall go into more detail on popular structural-based approaches, as well as end-to-end methods for talking head generation, audio driven automatic dubbing/audio driven video editing.

The scope of this article is limited to discussions surrounding state of the art works tackling facial animation generation, namely, we explore the recent trends in talking head generation, and audio driven automatic dubbing/video editing. The rest of the papers is organised as follows: Section 3 provides a detailed discussion on methods that seek to tackle the talking head generation, and automatic dubbing, classifying them as either end-to-end or structural-based methods, and discussing their merits and pitfalls. Section 4 provides details on popular datasets used to train models for these tasks, as well as a list of common evaluation metrics used to quantify the performance of such models. Section 5 provides discussion on open challenges within the field, and how researchers have been tackling them, before concluding the paper in Section 6.

3 Taxonomy of talking head generation and automatic dubbing

Talking head generation can be defined as the creation of a new video from a single source image or handful of frames, and a driving speech audio input. There are many challenges associated with this (Chen et al., 2020a). Not only must the generated lip and jaw movements be correctly synchronised to the speech input, but the overall head movement must also be realistic, eye blinking consistent to the speaker should be present, and the expressions on the face should match the tone and content of the speech. While many talking head approaches have been proposed in recent years, each addressing some or all of the aforementioned issues to various degrees, there is plenty of scope for researchers to further the field, as this article will demonstrate.

As touched upon earlier, the task of audio driven automatic dubbing is a constrained version of the talking head generation problem. Instead of creating an entire video from scratch, the goal is to alter an existing video, resynchronizing the lip and jaw movements of the target actor in response to a new input audio signal. Unlike talking head generation, factors such as head motion, eye blinks, and facial expressions are already present in the original video. The challenge lies in seamlessly altering the lip and jaw content of the video, while keeping the performance of the actor as close to the original as possible, so as to not detract from it.

3.1 End-to-end vs. structural-based generation

At a high level, existing deep learning approaches to both tasks can be broken down into two main methods: end-to-end or structural-based generation. Each method has its own set of advantages and disadvantages, which we will now go over.

3.1.1 Pipeline complexity and model latency

End-to-end approaches offer the advantage of a simpler pipeline, enabling faster processing and reduced latency in generating the final output. With fewer components and streamlined computations, real-time synthesis becomes achievable. However, the actual performance relies on crucial factors like the chosen architecture, model size, and output frame size. For example, GAN-based end-to-end methods can achieve real-time results, but they are often limited to lower output resolutions, such as 128×128 or 256×256. Diffusion-based approaches are even slower, often taking seconds or even minutes per frame, even with

more efficient sampling methods, albeit at the cost of image quality. Striking the right balance between speed and output resolution is essential in optimizing end-to-end talking head synthesis. It is important to highlight that these same limitations are also present for structural-based methods, particularly within their rendering process. However, structural-based methods tend to be even slower than end-to-end approaches due to the additional computational steps involved in their pipeline. Structural-based methods often require multiple stages, such as face detection, facial landmark/3D model extraction, expression synthesis, photorealistic rendering and so on. Each of these stages introduces computational overhead, making the overall process more time-consuming.

3.1.2 Cascading errors

In structural-based methods, errors made in earlier stages of the pipeline can propagate and amplify throughout the process. For example, inaccuracies in face or landmark detection can significantly impact the quality of the final generated video. End-to-end approaches, on the other hand, bypass the need for such intermediate representations, reducing the risk of cascading errors. At the same time, however, when errors do occur in end-to-end approaches, it can be harder to identify the source of the error, as such methods do not explicitly produce intermediate facial representations. This lack of transparency in the generation process can make it challenging for researchers to diagnose and troubleshoot issues when the output is not as expected. It becomes essential to develop techniques for error analysis and debugging to improve the reliability and robustness of end-to-end systems.

3.1.3 Robustness to different data

Structural-based methods rely on carefully curated and annotated datasets for each stage of the pipeline, which can be time-consuming and labor-intensive to create. End-to-end approaches are often more adaptable and generalize better to various speaking styles, accents, and emotional expressions, as they can leverage large and diverse datasets for training. This flexibility is crucial in capturing the nuances and variations present in natural human speech and facial expressions.

3.1.4 Output quality

The quality of output is a critical aspect in talking head synthesis, as it directly impacts the realism and plausibility of the generated videos. Structural-based methods excel in this regard due to their ability to exert more fine-grained control over the intermediate representations of the face during the synthesis process. With such methods, the face is typically represented using a set of keypoints (or 3D model parameters), capturing essential facial features and expressions. These landmarks serve as a structured guide for the generation of facial movements, ensuring that the resulting video adheres to the anatomical constraints of a human face. By explicitly controlling these keypoints, the model can produce more accurate and realistic facial expressions that are consistent with human facial anatomy.

End-to-end approaches sacrifice some level of fine-grained control in favor of simplicity and direct audio-to-video mapping. While they offer the advantage of faster processing and reduced latency, they may struggle to capture the intricate details and nuances present in facial expressions, especially in more challenging or uncommon scenarios.

3.1.5 Training data requirements

End-to-end approaches typically require a large amount of training data to generalize well across various situations. While structural-based methods can benefit from targeted, carefully annotated datasets for specific tasks, end-to-end methods may need a more diverse and extensive dataset to achieve comparable performance. This, in turn, means longer training times as the model needs to process and learn from a vast amount of data, which can be computationally intensive and time-consuming. This can be a significant drawback for researchers and practitioners, as it hinders the rapid experimentation and development of new models. It may also require access to powerful hardware, such as high-performance GPUs or TPUs, to accelerate the training process.

3.1.6 Explicit output guidance

Structural-based methods allow researchers to incorporate explicit rules and constraints into different stages of the pipeline. This explicit guidance can lead to more accurate and controllable results, which can be lacking in end-to-end approaches where such guidane is more difficult to implement.

3.2 Structural based generation

Structural based deep learning approaches have been immensely popular in recent years, and are considered the dominant approach when it comes to both talking head generation and audio driven automatic dubbing. As mentioned above, this is due to the relative ease with which one can exert control over the final output video, high quality image frame fidelity, and relative speed with which animations can be driven for 3D character models.

Instead of training a single neural network to generate the desired video given an audio signal, the problem is typically broken up into two main steps: 1) Training a neural network to drive the facial motion from audio of an underlying structural representation of the face. The structural representation is typically either a 3D morphable model or 2D/3D keypoint representation of the face. 2) Rendering photorealistic video frames from the structural model of the face using a second neural rendering model. Please see Table 1 for a summary of relevant structural-based approaches in the literature.

3.2.1 2D/3D landmark based methods

In this section we discuss methods that rely on either 2D or 3D face landmarks as an intermediate structural representation for producing facial animations from audio. Some of the discussed methods use the generated landmarks to animate a 3D face model, these methods shall also be considered "landmark-based." Figure 2 depicts a high level overview of what a typical landmark-based approach could look like.

Suwajanakorn et al. (2017), Taylor et al. (2017) were among the first works to explore using deep learning techniques to generate speech animation. The former trained a recurrent network to generate sparse mouth key points from audio before compositing them onto an existing video, and the latter presenting an approach

TABLE 1 Table summarising some of the most relevant structural-based approaches in the literature.

Method	Animation network architecture	Audio input	Intermediate representation	Additional inputs	Head motion	Rendering network architecture
Suwajanakorn et al. (2017)	LSTM	MFCC	PCA mouth coefficients	None	No	AAM-based rendering
Taylor et al. (2017)	Feed forward	Phoneme transcript	Face model animation parameters	None	No	Video compositing approach
Eskimez et al. (2018)	LSTM	Mel spectrograms	2D landmarks	None	No	Not applicable
Chen et al. (2019)	LSTM	MFCC	2D landmarks	None	No	GAN
Das et al. (2020)	GAN	Deep speech features	2D landmarks	None	No	GAN
Zhou et al. (2020)	LSTM	Learned speech embeddings	2D landmarks	None	Yes	GAN
Lu et al. (2021)	LSTM	Learned speech embeddings	2D Landmarks	None	Yes	GAN
Wang et al. (2021)	LSTM	MFCC + FBANK features	Keypoints—dense motion field	None	Yes	CNN
Ji et al. (2021)	LSTM	Learned speech embeddings	2D landmarks + 3D face model	Driving video	From video	GAN
Bigioi et al. (2022)	Recurrent LSTM	Mel spectrogram	2D landmarks	None	Yes	Not applicable
Karras et al. (2017)	CNN	Autocorrelation features	3D vertex positions of face mesh	Emotional State	No	Not applicable
Cudeiro et al. (2019)	CNN Encoder-Decoder	DeepSpeech features	Flame face model	None	No	Not applicable
Thies et al. (2020)	CNN	DeepSpeech features	3D expression parameters	None	No	CNN
Chen et al. (2020b)	CNN	Raw Audio	3D keypoints	Reference frames	Yes	GAN
Yi et al. (2020)	LSTM	MFCC	3D expression parameters	Driving video	Yes	GAN
Wu et al. (2021)	Encoder-Decoder + Unet	DeepSpeech features	3D expression parameters	Driving video	Yes	GAN
Zhang et al. (2021b)	GAN	Learned speech embeddings	3D expression parameters	Reference image	Yes	GAN
Zhang et al. (2021a)	GAN	DeepSpeech features	3D expression parameters	Driving video	Yes	GAN
Song et al. (2022)	LSTM + Unet	MFCC	3D expression parameters	Driving video	No	UNet
Wen et al. (2020)	GAN	MFCC	3D expression parameters	Driving video	No	GAN
Lahiri et al. (2021)	CNN	Spectrograms	3D vertex positions	Driving video	No	CNN

for generalised speech animation by training a neural network model to predict animation parameters of a reference face model given phoneme labels as input. The field has come a long way since then, with Eskimez et al. (2018) presenting a method for generating static (no headpose) talking face landmarks from audio via a LSTM based model, and Chen et al. (2019) expanding the work by conditioning a GAN network on the landmarks to generate photorealistic frames. Similarly, Das et al. (2020) also employed a GAN based architecture to generate facial landmarks from deepspeech features extracted from audio, before using a second GAN conditioned on the landmarks to generate the photorealistic frames.

Zhou et al. (2020)'s approach was among the first to generate talking face landmarks with realistic head pose movement from audio. They did this by training two LSTM networks, one to handle the lip/jaw movements, and a second to generate the headpose, before combining the two outputs and passing them through an off-the-shelf image-to-image translation network for generating photorealistic frames.

Lu et al. (2021)'s approach also simulated headpose and upper body motion using a separate auto regressive model trained on deepspeech audio features before generating realistic frames using an image-to-image translation model conditioned on feature maps based on the generated landmarks. While also proposing an approach for the head pose problem, Wang et al. (2021) tackled the challenge of stabilising non-face (background) regions when generating talking head videos from a single image.

Unlike the previous methods which were all approaches at solving the talking head generation task, the following papers fall into the audio-driven automatic dubbing category that seek to modify existing videos. Ji et al. (2021) were among the first to

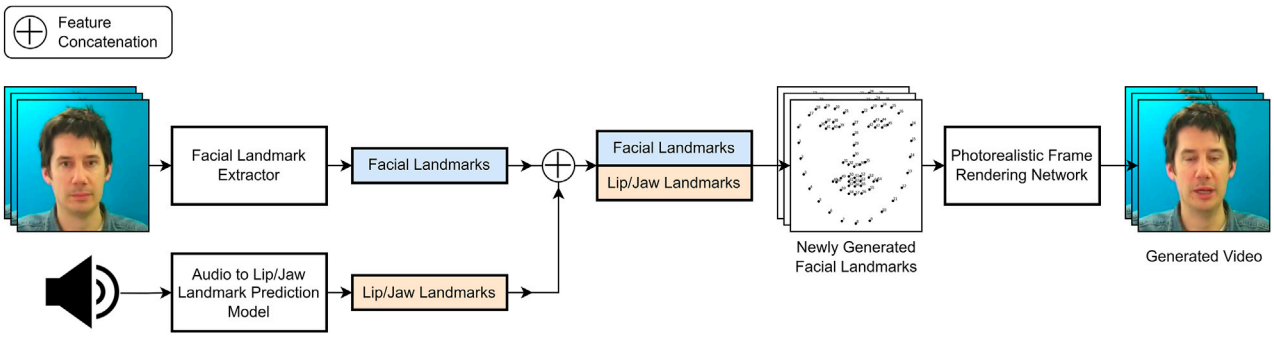

FIGURE 2
A high-level landmark based pipeline, where headpose and face structure from an existing video is combined with predicted lip/jaw displacements from a target audio clip to generate a modified video.

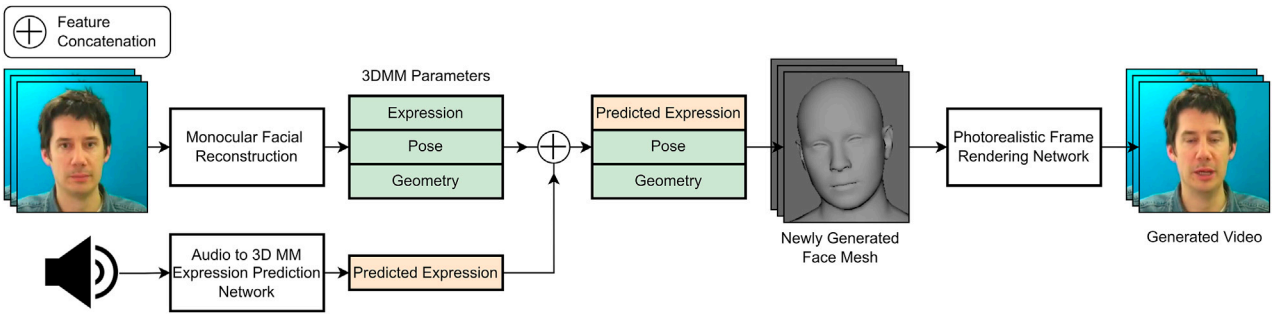

FIGURE 3
A high-level 3d model based pipeline, where monocular facial reconstruction is performed on a source video to extract expression, pose, and geometry parameters. A separate audio to expression parameter prediction network is then trained. The predicted expression parameters are then used to replace the original ones, to generate a new 3D facial mesh, which is then rendered into a photorealistic video via a neural rendering model.

tackle the problem of generating emotionally aware video portraits by disentangling speech into two representations, a content-aware time dependent stream, and an emotion-aware time independent stream, and training a model to generate 2D facial landmarks. It may be considered a "hybrid" structural approach, as from both the predicted and ground truth landmarks they perform monocular 3D reconstruction to obtain two 3D face models. They then combine the pose parameters from the ground truth with the expression and geometry parameters of the predicted to create the final 3D face model before extracting edge maps and generating the output frames via image-to-image translation. Bigioi et al. (2022) extracted ground truth 3D landmarks from video, and trained a network to alter them directly given an input audio sequence without the need to first retarget them to a static fixed face model before animating it and then returning the original headpose.

3.2.2 3D model based methods

In this section we discuss methods that use 3D face models as intermediate representations when generating facial animations. In other words, we talk about methods that train models to produce blendshape face parameters from audio signals as input. Figure 3 above depicts a high-level overview of one such model.

Karras et al. (2017) were among the first to use deep learning to learn facial animation for a 3D face model from limited audio data. Cudeiro et al. (2019) introduced a 4D audiovisual face dataset (talking 3D models), as well as a network trained to generate 3D facial animations from deepspeech audio features. Thies et al. (2020) also utilised deepspeech audio features to train a network to output speaker independent facial expression parameters that drive an intermediate 3D face model before generating the photorealistic frames using a neural rendering model. Chen et al. (2020b)'s approach involved learning head motion from a collection of reference frames, and then combining that information with learned PCA components denoting facial expression in a 3D aware frame generation network. Their approach is interesting because their pipeline addresses various known problems within talking head generation such as maintaining the identity/appearance of the head consistent, maintaining a consistent background, and generating realistic speaker aware head motion. Yi et al. (2020) presented an approach to generate talking head videos using a driving audio signal by training a neural network to predict pose and expression parameters for a 3D face model from audio, and combining them with shape, texture, and lighting parameters extracted from a set of reference frames. They then

render the 3D face model to photo realism via a neural renderer, before fine tuning the rendered frames with a memory augmented GAN. Wu et al. (2021) presented an approach to generate talking head faces of a target portrait given a driving speech signal, and "Style Reference Video." They train their model such that the output video mimics the speaking style of the reference video but whose identity corresponds to the target portrait. Zhang et al. (2021b) presented a method for one shot talking head animation. Given a reference frame and driving audio source they generate eyebrow, head pose, and mouth motion parameters of a 3D morphable model using an encoder-decoder architecture. A flow-guided video generator is then used to create the final output frames. Zhang et al. (2021a) synthesize talking head videos given a driving speech input and reference video clip. They design a GAN based module that can output expression, eyeblink, and headpose parameters of a 3D MM given deepspeech audio features.

While the previously referenced methods are all examples of pure talking head generation approaches, the following are in the automatic dubbing category. Both Song et al. (2022) and Wen et al. (2020) presented approaches to modify an existing video using a driving audio signal by training a neural network to extract 3D face model expression parameters from audio, and combining them with pose and geometry parameters extracted from the original video before applying neural rendering to generate the modified photorealistic video. To generate the facial animations, Song et al. (2021) employ a similar pipeline to the methods referenced above, however they go one step further, and transfer the acoustic properties of the original video's speaker onto the driving speech via an encoder-decoder mechanism, essentially dubbing the video. Richard et al. (2021) provided a generalised framework for generating accurate 3D facial animations given speech, by learning a categorical latent space that disentangles audio-correlated (lips/jaw motion), and audio un-correlated (eyeblinks, upper facial expression) information at inference time. Doing so, they built a framework that can be applied to both automatic dubbing, and talking head generation tasks. Lahiri et al. (2021) introduced an encoder-decoder architecture trained to decode 3D vertex positions [similar to Karras et al. (2017)], and 2D texture maps of the lip region from audio and the previously generated frame. They combine these to form a textured 3D face mesh which they then render and blend with the original video to generate the dubbed video clip.

We would also like to draw attention to the works of Fried et al. (2019) and Yao et al. (2021). These are video editing approaches which utilise text, in addition to audio, to modify existing talking head videos. The former approach works by aligning phoneme labels to the input audio, and constructing a 3D face model for each input frame. Then, when modifying the text transcript (e.g., dog to god), they search for segments of the input video where the visemes are similar, blending the 3D model parameters from the corresponding video frames to generate a new frame which is then rendered via their neural renderer. The latter approach builds off this work, by improving the efficiency of the phoneme matching algorithm, and developing a self-supervised neural retargeting technique for transferring the mouth motions of the source actor to the target actor.

3.3 End-to-end generation

Though less popular in recent times than their structural based counterparts, the potential to generate or modify a video directly given an input audio signal is one of the key factors that make end-to-end approaches an attractive proposition to talking head researchers. These methods aim to learn the complex mapping between audio, facial expressions and lip movements using a single unified model that combines the traditional stages of talking head generation into a single step. By doing so, they eliminate the need for explicit intermediate representations, such as facial landmarks, or 3D models, which can be computationally expensive and prone to error. This ability to directly connect the audio input to the video output streamlines the synthesis process and can enable real-time or near-real-time generation. Please see Table 2 for a summary of relevant end-to-end based approaches in the literature.

Chung et al. (2017) proposed one of the first end-to-end talking head generation techniques. Given a reference identity frame and driving speech audio signal, they succeeded in training an encoder-decoder based architecture to generate talking head videos, additionally demonstrating how their approach could be applied to the dubbing problem. Their approach was limited however as it only generated the cropped region around the face, discarding any background.

Chen et al. (2018) presented a GAN based method of generating lip movement from a driving speech source and reference lip frame. Similar to the above method, theirs was limited to generating just the cropped region of the face surrounding the lips. Song et al. (2018) presented a more generalised GAN-based approach for talking head generation that also took the temporal consistency between frames into account by introducing a recurrent unit in their pipeline, generating smoother videos. Zhou et al. (2019) proposed a model that could generate videos based on learned disentangled representations of speech and video. The approach is interesting because it allowed authors to generate a talking head video from a reference identity frame, and driving speech signal or video. Mittal and Wang (2020) disentangled the audio signal into various factors such as phonetic content, and emotional tone, and conditioned a talking head generative model on these representations instead of the raw audio, demonstrating compelling results. Vougioukas et al. (2020) proposed an approach to generate temporally consistent talking head videos from a reference frame and audio using a GAN-based approach. Their method generated realistic eyeblinks in addition to synchronised lip movements in an end-to-end manner. Prajwal et al. (2020) introduced a "lip-sync discriminator" for generating more accurate lip movements on talking head videos, as well as proposing new metrics to evaluate lip synchronization on generated videos. Eskimez et al. (2020) proposed a robust GAN based model that could generate talking head videos from noisy speech. Kumar et al. (2020) proposed a GAN-based approach for one shot talking head generation. Zhou et al. (2021) proposed an interesting approach to exert control over the pose of an audio-driven talking head. Using a target "pose" video, and speech signal, they condition a model to generate talking head videos from a single reference identity image whose pose is dictated by the target video.

TABLE 2 Table summarising some of the most relevant end-to-end approaches in the literature.

Method	Architecture	Audio input	Additional inputs	Head motion	Photorealistic frame rendering
Chung et al. (2017)	Encoder-Decoder	MFCC	Reference identity	No	Yes
Chen et al. (2018)	GAN	Mel spectrogram	Reference lip image	No	Limited to lip region only
Song et al. (2018)	GAN	MFCC	Reference image	No	Yes
Mittal and Wang (2020)	LSTM + GAN	Learned speech embeddings	Reference image	No	Yes
Zhou et al. (2019)	GAN	MFCC	Reference frames	No	Yes
Vougioukas et al. (2020)	GAN	Raw audio	Reference image	No	Yes
Prajwal et al. (2020)	GAN	Mel spectrogram	Driving video	Yes	Yes
Kumar et al. (2020)	GAN	DeepSpeech features	None	Yes	Yes
Eskimez et al. (2020)	LSTM + GAN	Raw audio	Reference image	No	Yes
Zhou et al. (2021)	GAN	Spectrograms	Driving video + reference frame	Yes	Yes
Stypułkowski et al. (2023)	Diffusion Unet	Learned speech embeddings	Reference image	Yes	Yes
Shen et al. (2023)	Diffusion Unet	Learned speech embeddings	Reference image + face landmarks	Yes	Yes
Bigioi et al. (2023)	Diffusion Unet	Mel spectrograms	Reference image	Yes	Yes

While GAN-based Goodfellow et al. (2014) methods such as the approaches referenced above have been immensely popular in recent years, they have been shown to have a number of limitations by practitioners in the field. Due to the presence of multiple losses and discriminators their optimization process is complex and quite unstable. This can lead to difficulties in finding a balance between the generator and discriminator, resulting in issues like mode collapse, where the generator fails to capture the full diversity of the target distribution. Vanishing gradients is another issue, which occurs when gradients become too small during back propagation, preventing the model from learning effectively, especially in deeper layers. This can significantly slow down the training process and limit the overall performance of the model. With that in mind, we would like to draw special attention to diffusion models (Sohl-Dickstein et al., 2015, Ho et al., 2020, Dhariwal and Nichol, 2021, Nichol and Dhariwal, 2021), a new class of generative model that has gained prominence in the last couple of years due to strong performance on a myriad of tasks such as text based image generation, speech synthesis, colourisation, body animation prediction, and more.

3.4 Diffusion-based generation

We dedicate a short section of this paper towards diffusion based approaches, due to their recent rise in use and popularity. Note that within this section, we describe methods found from both the end-to-end, and structural-based schools of thought as at this time, there are only a handful of diffusion-based talking head works.

For a deeper understanding of the diffusion architecture, we direct readers to works of Sohl-Dickstein et al. (2015); Ho et al. (2020); Dhariwal and Nichol (2021); Nichol and Dhariwal (2021), as these are the pioneering works that contributed to their recent popularity and wide-spread adoption. In short however the diffusion process can be summarised as consisting of two stages 1) the forward diffusion process, and 2) the reverse diffusion process.

In the forward diffusion process, the desired output data is gradually "destroyed" over a series of time steps by adding Gaussian noise at each step until the data becomes just another sample from a standard Gaussian distribution. Conversely, in the reverse diffusion process, a model is trained gradually denoise the data by removing the noise at each time step, with the loss typically being computed as a distance function between the predicted noise vs. the actual noise that was added at that particular time step. The combination of these two stages enables diffusion models to model complex data distributions without suffering from mode collapse unlike GANs, and to generate high-quality samples without the need for adversarial training or complex loss functions.

Within the context of talking head generation, and video editing there are a number of recent works that have explored using diffusion models. Specifically, Stypułkowski et al. (2023), Shen et al. (2023), and Bigioi et al. (2023) being among the first to explore their use for end-to-end talking head generation and audio driven video editing. All three methods follow a similar auto-regressive frame-based approach where the previously generated frame is fed back into the model along with the audio signal and a reference identity frame to generate the next frame in the sequence. Notably, Shen et al. (2023) condition their model with landmarks, and perform their training within the latent space to save on computational resources, unlike that of Stypułkowski et al. (2023) and Bigioi et al. (2023). Stypułkowski et al. (2023) approach can be considered a true talking head generation method, as their method does not rely on any frames from the

original video to guide their model (except for the initial seed/identity frame), and their resultant video is completely synthetic. Bigioi et al. (2023) perform video editing by modifying an existing video sequence by teaching their model to inpaint on a masked out facial region of the video in response to an input speech signal. Shen et al. (2023)'s approach is similar, where they perform video editing rather than talking head generation by modifying an existing video with the use of a face mask designed to cover the facial region of the source video.

While the above approaches are currently the only end-to-end diffusion based methods, a number of structural based approaches, that leverage diffusion models have also been proposed in recent months. Zhang et al. (2022) proposed an approach that used audio to predict landmarks, before using a diffusion based renderer to output the final frame. Zhua et al. (2023) also utilised a diffusion model similarly, using it to take the source image and the predicted motion features as input to generate the high-resolution frames. Du et al. (2023) introduced an interesting two stage approach for talking head generation. The first stage consisted of training a diffusion autoencoder on video frames, to extract latent representations of the frames. The second stage involved training a speech to latent representation model, with the idea being that the latents predicted by the speech, could be decoded by the pretrained diffusion autoencoder to image frames. The method achieves impressive results, outperforming other relevant structural-based methods in the field. Xu et al. (2023) use a diffusion-based renderer conditioned on multi-model inputs to drive the emotion, and pose of the generated talking head videos. Notably their approach is also applicable to the face swapping problem.

Within the realm of talking heads, diffusion models have shown incredibly promising results, often producing videos with demonstratively higher visual quality, and similar lip sync performance compared to more traditional GAN-based methods. One major limitation, however, lies in their inability to model long sequences of frames without the output degrading in quality over time due to their autoregressive nature. It will be exciting to see what the future holds for further research in this area.

3.5 Other approaches

There are certain approaches that do not necessarily fit into the aforementioned subcategories, that are still relevant and worth discussing.

Viseme based methods such as Zhou et al. (2018) are early approaches at driving 3D character models. The authors presented an LSTM based network capable of producing viseme curves that could drive JALI based character models as described by Edwards et al. (2016).

Guo et al. (2021) is a unique method for talking head generation that instead of relying on traditional intermediate structural representations such as landmarks or 3DMMs, instead generates a neural radiance field from audio from which a realistic video is synthesised using volume rendering.

4 Popular datasets and evaluation metrics

In this section we describe the most popular metrics for measuring the quality of videos generated by audio-driven talking head, and automatic dubbing models.

4.1 Evaluation metrics

Quantitatively evaluating both talking head, and dubbed videos is a non-straight forward task. Traditional perceptual metrics such as SSIM, or distance-based metrics such as the L2 Norm, or PSNR, which seek to quantify the similarity between two images, are inadequate. Such metrics do not take into account the temporal nature of video, with the quality of a video being affected not only by the individual quality of frames, but also by the smoothness and synchronisation of the frames as they are played back in the video.

Although these metrics may not provide a perfect evaluation of video quality, they are still important for bench marking purposes as they provide a good indication of what to expect from the model. As such, when there is access to ground truth samples to compare a model's output with, the following metrics are commonly used:

PSNR (Peak Signal to Noise Ratio): The peak signal to noise ratio between the ground truth and the generated image is computed. The higher the PSNR value, the better the quality of the reconstructed image.

Facial Action Units (AU) Ekman and Friesen (1978) Recognition: Song et al. (2018) and Chen et al. (2020b) popularised a method for evaluating reconstructed images with respect to ground truth samples using five facial action units.

ACD (Average Content Distance) (Tulyakov et al., 2018): As used by Vougioukas et al. (2020), the Cosine (ACD-C) and Euclidean (ACD-E) distance between the generated frame and ground truth image can be calculated. The smaller the distance between two images the more similar the images.

SSIM (Structural Similarity Index) (Wang et al., 2004): This is a metric designed to measure the similarity between two images by looking at the luminance, contrast, and structure of the pixels in the images.

Landmark Distance Metric (LMD): Proposed by Chen et al. (2018), Landmark Distance (LMD) is a popular metric used to evaluate the lip synchronisation of a synthetic video. It works by extracting facial landmark lip coordinates for each frame of both the generated, and ground truth videos using an off-the-shelf facial landmark extractor, calculating the euclidean distance between them, and normalising based on the length of video and number of frames.

Unfortunately, when generating talking head or dubbed videos, oftentimes it is impossible to use the metrics discussed above as there is no corresponding ground truth data with which to compare the generated samples. Therefore, a number of perceptual metrics (metrics which seek to emulate how humans perceive things) have been proposed to address this problem. These include:

CPBD (Cumulative Probability Blur Detection) (Narvekar and Karam, 2011): This is a perceptual based metric used to detect blur in imageness and measure image sharpness. Used by Kumar et al. (2020); Vougioukas et al. (2020); Chung et al. (2017) to evaluate their talking head videos.

WER (Word Error Rate): A pretrained lip reading model is used to predict the words spoken by the generated face. Works such as Kumar et al. (2020) and Vougioukas et al. (2020) use the LipNet Assael et al. (2016) model which is pre-trained on the GRID data set and achieves 95.2 percent lip reading accuracy.

SyncNet Based Metrics: These are perceptual metrics based on the SyncNet model introduced by Chung and Zisserman (2017b)

that evaluate lip synchronisation in unconstrained videos. Prajwal et al. (2020) introduced two such metrics: 1) LSE-D which is the average error measure calculated in terms of the distance between the lip and audio representations, and 2) LSE-C which is the average confidence score. These metrics have proven popular since their introduction, with a vast majority of recent papers in the field using them for evaluating their videos.

4.2 Benchmark Datasets

There are a number of benchmark datasets used to evaluate talking head and video dubbing models. They can be broadly categorised as being either "in-the-wild," or "lab conditions" style datasets. In this section we list some of the most popular ones, and briefly describe them.

- VoxCeleb 1 and 2 (Nagrani et al., 2017; Chung et al., 2018): This dataset contains audio and video recordings of celebrities speaking in the wild. It is often used for training and evaluating talking head generation, lip reading, and dubbing models. The former contains over 150,000 utterances from 1,251 celebrities, and the latter over 1,000,000 utterances from 6,112 celebrities.
- GRID (Cooke et al., 2006): The GRID dataset consists of audio and video recordings of 34 speakers reading 1,000 sentences in lab conditions. It is commonly used for evaluating lip-reading algorithms but has also been used for talking head generation and video dubbing models.
- LRS3-TED (Afouras et al., 2018): This dataset contains audio and video recordings of over 400 h of TED talks, which are speeches given by experts in various fields.
- LRW (Chung and Zisserman, 2017a): The LRW (Lip Reading in the Wild) dataset consists of up to 1,000 utterances of 500 different words, spoken by hundreds of different speakers in the wild.
- CREMA-D (Cao et al., 2014): This dataset contains audio and video recordings of people speaking in various emotional states (happy, sad, anger, fear, disgust, and neutral). In total it contains 7,442 clips of 91 different actors recorded in lab conditions.
- TCD-TIMIT (Harte and Gillen, 2015): The Trinity College Dublin Talking Heads dataset (TCD-TIMIT) contains video recordings of 62 actors speaking in a controlled environment.
- MEAD Dataset (Wang et al., 2020): This dataset contains videos featuring 60 actors talking with eight different emotions at three different intensity levels (except for neutral). The videos are simultaneously recorded at seven different perspectives with roughly 40 h of speech recorded for each person.
- RAVDESS Dataset (Wang et al., 2020): The Ryerson Audio-Visual Database of Emotional Speech and Song is a corpus consisting of 24 actors speaking with calm, happy, sad, angry, fearful, surprise, and disgust expressions, and singing with calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. It contains 7,356 recordings in total.
- CelebV-HQ (Zhu et al., 2022): CelebV-HQ is a dataset containing 35,666 video clips involving 15,653 identities and 83 manually labeled facial attributes covering aspects such as appearance, action, and emotion

5 Open challenges

Although significant progress has been made in the fields of talking head generation and automatic dubbing, these areas of research are constantly evolving, and several open challenges still need to be addressed, offering plenty of opportunities for future work.

5.1 Bridging the uncanny valley

Despite existing research, generating truly realistic talking heads remains an unsolved problem. There are various factors that come into play when discussing the topic of realism and how we can bridge the "Uncanny Valley" effect in video dubbing. These include:

- Visual quality: Realistic talking head videos should have high-quality visuals that accurately capture the colors, lighting, and textures of the scene. This requires attention to detail in the rendering process. Currently, most talking head and visual dubbing approaches are limited to generating videos at low output resolutions, and those that do work on higher resolutions are quite limited both in terms of model robustness, and generalisation (more on that later). This is due to several reasons: 1) the computational complexity of deep learning models rises significantly when generating high-resolution videos, both in terms of training time, and inference speed; this, in turn, has an adverse effect on real-time performance; 2) generating realistic talking head videos requires the model to capture intricate details of facial expressions, lip movements, and speech patterns; as the output resolution of the video increases, so too does the demand for more fine-grained details, making it more difficult for models to achieve high degrees of realism; 3) Storage and bandwidth limitations; high-resolution videos require both of these in abundance, limiting high resolution generation to researchers who have access to state of the art in hardware systems. Some approaches that have sought to tackle this issue are the works of Gao et al. (2023), Guo et al. (2021), and Shen et al. (2023), who's approaches are capable of outputting high resolution frames.
- Motion: Realistic talking head/dubbed videos should have realistic motion, including smooth and natural movements of the face in response to speech, and realistic head motion when generating videos from scratch. This is a continuous topic of interest, with many works exploring it such as Chen et al. (2020b), Wang et al. (2021), and more recently Zhang et al. (2023).
- Disembodied Voice: The phenomenon of a Disembodied Voice is characterized by a jarring mismatch between a speaker's voice and their physical appearance, which is a commonly encountered issue in movie dubbing. Despite its

significance, this issue remains relatively unexplored within the realm of talking head literature, thereby presenting a promising avenue for researchers to investigate further. The work conducted by Oh et al. (2019) demonstrated that there is an inherent link between a speaker's voice and their appearance that can be learned, thus lending credence to the idea that dubbing efforts should prioritize the synchronization of voice and appearance.

- Emotion: Realistic videos should evoke realistic emotions, including facial expressions, body language, and dialogue. Achieving realistic emotions requires careful attention to acting and performance, as well as attention to detail in the animation and sound design. Recent works seeking to incorporate emotion into their generated talking heads include Ma et al. (2023), Liang et al. (2022), Li et al. (2021).

5.2 The data problem: single vs. multispeaker approaches

As mentioned previously there are two primary approaches to video dubbing—structural and end-to-end. In order to train a model to generate highly photorealistic talking head videos with current end-to-end methods, many dozens of hours of single-speaker audiovisual content are required. The content should be of a high quality with factors such as good lighting, consistent framing of the face, and clear audio data. The quantity of data on an individual speaker may be reduced when methods are trained on a multi-speaker dataset, but sufficiently large datasets are only starting to become available. At this point in time it is not possible to estimate how well end-to-end methods might generalize to multiple speakers, or how much data may eventually be required to fine-tune a dubbing model for an individual actor in a movie to achieve a realistic mimicry of their facial actions. The goal should be of the order of tens of minutes of data, or less, to allow for the dubbing of the majority of characters with speaking roles.

5.3 Generalisation and robustness

Developing a model that can generalize across all faces, and audios, under any conditions such as poor lighting, partial occlusion, or incorrect framing, remains a challenging task yet to be fully resolved.

While supervised learning has proven to be a powerful approach for training models, it typically requires large amounts of labeled data that are representative of the target distribution. However, collecting diverse and balanced datasets that cover all possible scenarios and variations in facial appearance and conditions is a challenging and time-consuming task. Furthermore, it is difficult to anticipate all possible variations that the model may encounter during inference, such as changes in lighting conditions or facial expressions.

To address these challenges, researchers have explored alternative approaches such as self-supervised learning, which aims to learn from unlabelled data by creating supervisory signals from the data itself. In other words, self-labelling the data. Methods such as Baevski et al. (2020); Hsu et al. (2021), which fall under the self-supervised learning paradigm, have gained popularity in speech-related fields due to their promising results in improving the robustness and generalization of models. These methods may help overcome the limitations of traditional supervised learning methods that rely solely on labeled data for training. That being said, Radford et al. (2022) showed that while such methods can learn high-quality representations of the input they are being trained on,"they lack an equivalently performant decoder mapping those representations to useable outputs, necessitating a finetuning stage in order to actually perform a task such as speech recognition". The authors demonstrate that by training their model on a "weakly-supervised" dataset of 680,000 h of speech, their model performs well on unseen datasets without the need to finetune. What this means for talking head generation/dubbing is that a model trained on large amounts of "weakly-supervised," or in other words, imperfect data, may potentially acquire a higher level of generalization. This can be particularly valuable for tasks like talking head generation or dubbing, where a system needs to understand and replicate various speech patterns, accents, and linguistic nuances that might not be explicitly present in labeled data.

5.4 The multilingual aspect

In the realm of talking head generation, it is fascinating to observe the adaptability of models trained exclusively on English-language datasets when faced with speech from languages they have not encountered during training. This phenomenon can be attributed to the models' proficiency in learning universal acoustic and linguistic features. While language diversity entails a wide array of phonetic, prosodic, and syntactic intricacies, there exists an underpinning foundation of shared characteristics that traverse linguistic boundaries. These foundational aspects, intrinsic to human speech, include elements like phonetic structure and prosodic patterns, which exhibit commonalities across languages. Talking head generation models that excel in capturing these universal attributes inherently possess the ability to generate lip motions that align with a range of linguistic expressions, irrespective of language.

While the lip movements generated by models trained on English-language datasets may exhibit a remarkable degree of fidelity when applied to unseen languages, capturing cultural behaviors associated with those languages is a more intricate endeavor. Cultural gestures, expressions, and head movements often bear an intimate connection with language and its subtle intricacies. Unfortunately, these models, despite their linguistic adaptability, may lack the exposure needed to capture these culturally specific behaviors accurately. For instance, behaviors like the distinctive head movements indicative of agreement in certain cultures remain a challenge for these models. This underscores the connection between language and culture, highlighting the need for models to not only decipher linguistic components but also to appreciate and simulate the cultural nuances that accompany them. As such, we believe that further research is necessitated to ensure a unified representation of both linguistic and cultural dimensions in the realm of talking head generation and automatic dubbing, leaving this an open challenge to the field.

5.5 Ethical and legal challenges

Lastly we mention that the modification of original digital media content is subject to a wide range of ethical and data-protection considerations. While it is expected for most digital content that the work of paid actors is considered as "work for hire," there are broader considerations if auto-dubbing technology becomes broadly adopted. Even as we write there is a large-scale strike of actors in Hollywood, fighting for rights with respect to the use of AI generated acting sequences. A full discussion of the broad ethical and intellectual property implications arising as today's AI technologies mature into sophisticated end-products for digital content creation would require a separate article.

Ultimately there is a clear need for advanced IP rights management within the digital media creating industry. Past efforts have focused on media manipulation, such as fingerprinting or encryption (Kundur and Karthik, 2004) but were ultimately unsuccessful. More recently researchers have proposed techniques such as blockchain might be used in the context of subtitles (Orero and Torner, 2023), while legal researchers have provided a broader context for the challenge of digital copyright in the context of the evolution of the Metaverse (Jain and Srivastava, 2022). Clearly, multi-lingual video dubbing represents just one specific sub-context of this broader ethical and regulatory challenge.

Looking at ethical considerations for the focused topic of multi-lingual video-dubbing one practical approach is to adopt a methodology that can track pipeline usage. One technique adopted in the literature is to build traceability into the pipeline itself, as discussed by Pataranutaporn et al. (2021). These authors have included both human and machine traceability methods into their pipeline to ensure safe and ethical use thereof. Their human traceability technique was inspired by fabrication detection techniques drawn from other media paradigms (e.g., text, video) and incorporates perceivable traces like signatures of authorship, distinguishable appearance or small editing artefacts into the generated media. Machine traceability, on the other hand, involves incorporating traces imperceptible to humans, such as non-visible noise signals.

6 Concluding thoughts

In this paper we have attempted to capture the current state-of-art for automated, multi-lingual video dubbing. This is an emerging field of research, driven by the needs of the video streaming industry and there are may interesting synergies with a range of neural technologies, including auto-translation services, text-to-speech synthesis, and talking-head generators. In addition to a review and discussion of the recent literature we have also outlined some of the key challenges that remain to blend today's neural technologies into practical implementations of tomorrow's digital media services.

This work may serve both as an introduction and reference guide for researchers new to the fields of automatic dubbing, and talking head generation, but also seeks to draw attention to the latest techniques and new approaches and methodologies for those who already have some familiarity with the field. We hope it will encourage and inspire new research and innovation on this emerging research topic.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and the ADAPT Centre (Grant 13/RC/2106).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afouras, T., Chung, J. S., and Zisserman, A. (2018). Lrs3-ted: A large-scale dataset for visual speech recognition. https://arxiv.org/abs/1809.00496.

Alarcon, N. (2023). *Netflix builds proof-of-concept AI model to simplify subtitles for translation*.

Assael, Y. M., Shillingford, B., Whiteson, S., and De Freitas, N. (2016). Lipnet: end-to-end sentence-level lipreading. https://arxiv.org/abs/1611.01599.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. neural Inf. Process. Syst.* 33, 12449–12460. doi:10.48550/arXiv.2006.11477

Bigioi, D., Basak, S., Jordan, H., McDonnell, R., and Corcoran, P. (2023). Speech driven video editing via an audio-conditioned diffusion model. https://arxiv.org/abs/2301.04474. doi:10.1109/ACCESS.2022.3231137

Bigioi, D., Jordan, H., Jain, R., McDonnell, R., and Corcoran, P. (2022). Pose-aware speech driven facial landmark animation pipeline for automated dubbing. *IEEE Access* 10, 133357–133369.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi:10.1007/s10579-008-9076-6

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* 5, 377–390. doi:10.1109/TAFFC.2014.2336244

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Trans. Graph. (TOG)* 24, 1283–1302. doi:10.1145/1095878.1095881

Chen, L., Cui, G., Kou, Z., Zheng, H., and Xu, C. "What comprises a good talking-head video generation?," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Glasgow, UK 2020a.

Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., et al. "Talking-head generation with rhythmic head motion," in Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 2020b.

Chen, L., Li, Z., Maddox, R. K., Duan, Z., and Xu, C. (2018). Lip movements generation at a glance. https://arxiv.org/abs/1803.10404.

Chen, L., Maddox, R. K., Duan, Z., and Xu, C. (2019). Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. https://arxiv.org/abs/1905.03820.

Chung, J. S., Jamaludin, A., and Zisserman, A. (2017). You said that? https://arxiv.org/abs/1705.02966.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: deep speaker recognition. https://arxiv.org/abs/1806.05622.

Chung, J. S., and Zisserman, A. "Lip reading in the wild," in Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 2017a, 87–103.

Chung, J. S., and 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, November 2017b, 251–263.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* 120, 2421–2424. doi:10.1121/1.2229005

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019). Capture, learning, and synthesis of 3d speaking styles. https://arxiv.org/abs/1905.03079.

Das, D., Biswas, S., Sinha, S., and Bhowmick, B. "Speech-driven facial animation using cascaded gans for learning of motion and texture," in Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 2020, 408–424.

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Adv. neural Inf. Process. Syst.* 34, 8780–8794. doi:10.48550/arXiv.2105.05233

Du, C., Chen, Q., He, T., Tan, X., Chen, X., Yu, K., et al. (2023). Dae-talker: high fidelity speech-driven talking face generation with diffusion autoencoder. https://arxiv.org/abs/2303.17550.

Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., et al. (2023). *SeamlessM4t—massively multilingual and multimodal machine translation*. Menlo Park, California, United States: Meta.

Edwards, P., Landreth, C., Fiume, E., and Singh, K. (2016). Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph. (TOG)* 35, 1–11. doi:10.1145/2897824.2925984

Ekman, P., and Friesen, W. V. (1978). Facial action coding system. *Environ. Psychol. Nonverbal Behav*. doi:10.1037/t27734-000

Eskimez, S. E., Maddox, R. K., Xu, C., and Duan, Z. "End-to-end generation of talking faces from noisy speech," in Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona, Spain, May 2020, 1948–1952.

Eskimez, S. E., Maddox, R. K., Xu, C., and Duan, Z. "Generating talking face landmarks from speech," in Proceedings of the Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2018, 372–381.

Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., et al. (2019). Text-based editing of talking-head video. *ACM Trans. Graph. (TOG)* 38, 1–14. doi:10.1145/3306346.3323028

Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., and Lu, Y. (2023). High-fidelity and freely controllable talking head video generation. https://arxiv.org/abs/2304.10168.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. neural Inf. Process. Syst.* 27.

Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., and Zhang, J. (2021). Ad-nerf: audio driven neural radiance fields for talking head synthesis. https://arxiv.org/abs/2103.11078.

Harte, N., and Gillen, E. (2015). Tcd-timit: an audio-visual corpus of continuous speech. *IEEE Trans. Multimedia* 17, 603–615. doi:10.1109/TMM.2015.2407694

Hayes, L., and Bolanos-Garcia-Escribano, A. (2022). *Streaming English dubs: a snapshot of netflix's playbook: Conference: Transtextual and transcultural circumnavigations. 10th international conference of aieti (iberian association for translation and interpreting studies)*. Braga, portugal: universidade do minho.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. neural Inf. Process. Syst.* 33, 6840–6851.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 3451–3460. doi:10.1109/TASLP.2021.3122291

Jain, S., and Srivastava, A. (2022). Copyright infringement in the era of digital world. *Int'l JL Mgmt. Hum.* 5, 1333.

Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C., Cao, X., et al. (2021). Audio-driven emotional video portraits. https://arxiv.org/abs/2104.07452.

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. (TOG)* 36, 1–12. doi:10.1145/3072959.3073658

Kumar, N., Goel, S., Narang, A., and Hasan, M. (2020). Robust one shot audio to video generation. https://arxiv.org/abs/2012.07842.

Kundur, D., and Karthik, K. (2004). Video fingerprinting and encryption principles for digital rights management. *Proc. IEEE* 92, 918–932. doi:10.1109/JPROC.2004.827356

Lahiri, A., Kwatra, V., Frueh, C., Lewis, J., and Bregler, C. (2021). Lipsync3d: data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. https://arxiv.org/abs/2106.04185.

Łańcucki, A. "Fastpitch: parallel text-to-speech with pitch prediction," in Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, June 2021.

Li, L., Wang, S., Zhang, Z., Ding, Y., Zheng, Y., Yu, X., et al. (2021). Write-a-speaker: text-based emotional and rhythmic talking-head generation. *Proc. AAAI Conf. Artif. Intell.* 35, 1911–1920. doi:10.1609/aaai.v35i3.16286

Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., et al. "Expressive talking head generation with granular audio-visual control," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, June 2022, 3387–3396.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., et al. (2023). Audioldm: text-to-audio generation with latent diffusion models. https://arxiv.org/abs/2301.12503.

Lu, Y., Chai, J., and Cao, X. (2021). Live speech portraits: real-time photorealistic talking-head animation. *ACM Trans. Graph. (TOG)* 40, 1–17. doi:10.1145/3478513.3480484

Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., et al. (2023). Styletalk: one-shot talking head generation with controllable speaking styles. https://arxiv.org/abs/2301.01081.

Marioryad, S., and Busso, C. (2012). Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Trans. Audio, Speech, Lang. Process.* 20, 2329–2340. doi:10.1109/TASL.2012.2201476

Mittal, G., and Wang, B. "Animating face using disentangled audio representations," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, March 2020, 3290–3298.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. https://arxiv.org/abs/1706.08612.

Narvekar, N. D., and Karam, L. J. (2011). A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Trans. Image Process.* 20, 2678–2683. doi:10.1109/TIP.2011.2131660

Nichol, A. Q., and Dhariwal, P. "Improved denoising diffusion probabilistic models," in Proceedings of the International Conference on Machine Learning (PMLR), 2021, Glasgow, UK 8162–8171.

Nilesh, C., and Deck, A. (2023). Forget subtitles: youtube now dubs videos with AI-generated voices. https://restofworld.org/2023/youtube-ai-dubbing-automated-translation/.

Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., et al. (2019). Speech2face: learning the face behind a voice. https://arxiv.org/abs/1905.09773.

Orero, P., Torner, A. F., et al. (2023). The visible subtitler: blockchain technology towards right management and minting. *Open Res. Eur.* 3, 26. doi:10.12688/openreseurope.15166.1

Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., et al. (2021). Ai-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.* 3, 1013–1022. doi:10.1038/s42256-021-00417-9

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and 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, October, 2020, 484–492.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. https://arxiv.org/abs/2212.04356.

Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., and Sheikh, Y. (2021). Meshtalk: 3d face animation from speech using cross-modality disentanglement. https://arxiv.org/abs/2104.08223.

Roxborough, S. (2019). *Netflix's global reach sparks dubbing revolution: "the public demands it"*. Los Angeles, California, United States: The Hollywood Reporter.

Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., et al. (2023). Difftalk: crafting diffusion models for generalized audio-driven portraits animation. https://arxiv.org/abs/2301.03786.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. "Deep unsupervised learning using nonequilibrium thermodynamics," in Proceedings of the International conference on machine learning (PMLR), Lille, France, July 2015, 2256–2265.

Song, L., Liu, B., Yin, G., Dong, X., Zhang, Y., and Bai, J.-X. "Tacr-net: editing on deep video and voice portraits," in Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China, October 2021, 478–486. doi:10.1109/TIFS.2022.3146783

Song, L., Wu, W., Qian, C., He, R., and Loy, C. C. (2022). Everybody's talkin': let me talk as you want. *IEEE Trans. Inf. Forensics Secur.* 17, 585–598.

Song, Y., Zhu, J., Li, D., Wang, X., and Qi, H. (2018). Talking face generation by conditional recurrent adversarial network. https://arxiv.org/abs/1804.04786.

Spiteri Miggiani, G. (2021). English-Language dubbing: challenges and quality standards of an emerging localisation trend. *J. Specialised Transl.*

Stypułkowski, M., Vougioukas, K., He, S., Zieba, M., Petridis, S., and Pantic, M. (2023). Diffused heads: diffusion models beat gans on talking-face generation. https://arxiv.org/abs/2301.03396.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph. (ToG)* 36, 1–13. doi:10.1145/3072959.3073640

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., et al. (2017). A deep learning approach for generalized speech animation. *ACM Trans. Graph. (TOG)* 36, 1–11. doi:10.1145/3072959.3073699

Thies, J., Elgharib, M., Tewari, A., Theobalt, C., and Nießner, M. "Neural voice puppetry: audio-driven facial reenactment," in Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 2020, 716–731.

Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. "Mocogan: decomposing motion and content for video generation," in Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, June 2018, 1526–1535.

Vougioukas, K., Petridis, S., and Pantic, M. (2020). Realistic speech-driven facial animation with gans. *Int. J. Comput. Vis.* 128, 1398–1413. doi:10.1007/s11263-019-01251-8

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., et al. "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in Proceedings of the ECCV, Glasgow, UK, (August 2020.

Wang, S., Li, L., Ding, Y., Fan, C., and Yu, X. (2021). Audio2head: audio-driven one-shot talking-head generation with natural head motion. https://arxiv.org/abs/2107.09293.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., et al. (2017). Tacotron: towards end-to-end speech synthesis. https://arxiv.org/abs/1703.10135.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. image Process.* 13, 600–612. doi:10.1109/TIP.2003.819861

Weitzman, C. (2023). *Voice actor vs. AI voice: Pros and cons. Speechify. Section: VoiceOver*. St Petersburg, Florida, USA: Speechify.

Wen, X., Wang, M., Richardt, C., Chen, Z.-Y., and Hu, S.-M. (2020). Photorealistic audio-driven video portraits. *IEEE Trans. Vis. Comput. Graph.* 26, 3457–3466. doi:10.1109/TVCG.2020.3023573

Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., and Deng, Q. "Imitating arbitrary talking style for realistic audio-driven talking face synthesis," in Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China, October 2021, 1478–1486.

Xu, C., Zhu, S., Zhu, J., Huang, T., Zhang, J., Tai, Y., et al. (2023). Multimodal-driven talking face generation, face swapping, diffusion model. https://arxiv.org/abs/2305.02594.

Yang, Y., Shillingford, B., Assael, Y., Wang, M., Liu, W., Chen, Y., et al. (2020). Large-scale multilingual audio visual dubbing. https://arxiv.org/abs/2011.03530.

Yao, X., Fried, O., Fatahalian, K., and Agrawala, M. (2021). Iterative text-based editing of talking-heads using neural retargeting. *ACM Trans. Graph. (TOG)* 40, 1–14. doi:10.1145/3449063

Yi, R., Ye, Z., Zhang, J., Bao, H., and Liu, Y.-J. (2020). Audio-driven talking face video generation with learning-based personalized head pose. https://arxiv.org/abs/2002.10137.

Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., et al. (2021a). Facial: synthesizing dynamic talking face with implicit attribute learning. https://arxiv.org/abs/2108.07938.

Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., et al. (2023). Sadtalker: learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. https://arxiv.org/abs/2211.12194.

Zhang, X., Wang, J., Cheng, N., Xiao, E., and Xiao, J. (2022). "Shallow diffusion motion model for talking face generation from speech," in *Asia-pacific web (APWeb) and web-age information management (WAIM) joint international conference on web and big data* (Berlin, Germany: Springer), 144–157.

Zhang, Z., Li, L., Ding, Y., and Fan, C. (2021b). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Glasgow, UK 3661–3670.

Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. *Proc. AAAI Conf. Artif. Intell.* 33, 9299–9306. doi:10.48550/arXiv.1807.07860

Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X., and Liu, Z. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. https://arxiv.org/abs/2104.11116.

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., and Li, D. (2020). Makelttalk: speaker-aware talking-head animation. *ACM Trans. Graph. (TOG)* 39, 1–15. doi:10.1145/3414685.3417774

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018). Visemenet: audio-driven animator-centric speech animation. *ACM Trans. Graph. (TOG)* 37, 1–10. doi:10.1145/3197517.3201292

Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., et al. (2022). CelebV-HQ: A large-scale video facial attributes dataset. https://arxiv.org/abs/2207.12393.

Zhua, Y., Zhanga, C., Liub, Q., and Zhoub, X. "Audio-driven talking head video generation with diffusion model," in Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June 2023, 1–5.