IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 December 2022, accepted 14 December 2022, date of publication 20 December 2022,
date of current version 28 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3231137

RESEARCH ARTICLE

Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing

DAN BIGIOI[1], (Graduate Student Member, IEEE), HUGH JORDAN[2], RISHABH JAIN[1], (Member, IEEE), RACHEL MCDONNELL[2], AND PETER CORCORAN[1], (Fellow, IEEE)
[1]School of Electrical and Electronics Engineering, National University of Ireland, University of Galway, Galway, H91 TK33 Ireland
[2]Trinity College Dublin, University of Dublin, Dublin 2, D02 PN40 Ireland

Corresponding author: Dan Bigioi (d.bigioi1@nuigalway.ie)

This work was supported by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant 18/CRT/6224.

ABSTRACT A novel neural pipeline allowing one to generate pose aware 3D animated facial landmarks synchronised to a target speech signal is proposed for the task of automatic dubbing. The goal is to automatically synchronize a target actors' lips and facial motion to an unseen speech sequence, while maintaining the quality of the original performance. Given a 3D facial key point sequence extracted from any reference video, and a target audio clip, the neural pipeline learns how to generate head pose aware, identity aware landmarks and outputs accurate 3D lip motion directly at the inference stage. These generated landmarks can be used to render a photo-realistic video via an additional image to image conversion stage. In this paper, a novel data augmentation technique is introduced that increases the size of the training dataset from N audio/visual pairs up to NxN unique pairs for the task of automatic dubbing. The trained inference pipeline employs a LSTM-based network that takes Mel-coefficients as input from an unseen speech sequence, combined with head pose, and identity parameters extracted from a reference video to generate a new set of pose aware 3D landmarks that are synchronized with the unseen speech.

INDEX TERMS Machine learning, computer vision, lip synchronization, talking head generation, automatic dubbing, audio driven deep fakes, artificial intelligence.

I. INTRODUCTION

Automatic speech dubbing is an area of great interest to the entertainment sector as not only is it relevant to the task of automatic dubbing for movies, television, and videos in general, it is also applicable to speech-based animation pipelines for video game characters, CG animated movies, and increasingly, personal avatars within the realm of virtual reality.

Automatic audio-visual speech dubbing is a topic which falls under the broader field of talking head generation, or talking heads for short. A talking head video is a video which contains one subject talking directly to the camera. The goal of talking head generation is either to generate a photo-realistic talking head video from a static reference image and target audio source (image-based methods), or in the case of this paper and the task of automatic dubbing / speech driven video editing, to modify an existing video based on a new target audio clip (video-based methods).

The meteoric rise in popularity of deep learning over the last decade has in turn lead to a surge in interest towards talking head generation and its associated sub tasks such as dubbing, video editing, and video generation. Numerous approaches have been suggested over the last five years, each one looking to advance the state of the art within the field of talking heads. For the vast majority of image-based methods (where a video is generated from a single reference image + audio), a neural network is trained to generate the lip movements and facial expressions from audio, while a second network is trained to generate the head pose information. Likewise for most video-based methods (where the content of an already existing video is modified based off the audio), a single network is used to generate the lip movements onto a static face mesh, which then gets fitted on top of landmarks

The associate editor coordinating the review of this manuscript and approving it for publication was Ángel F. García-Fernández.

extracted from each frame of the video before rendering. For both cases, these pipelines are quite complex, and there is a need for simpler, more intuitive approaches such that artists can make better use of these technologies.

Speech dubbing itself is a highly complex task, as not only does one need to generate accurate lip and jaw motion to match the target speech signal, special care must be taken to not diminish from the actors visual performance. Factors such as the actors facial expressions, head movements, and mannerisms, must be kept as close to the original performance as possible such that the only difference between the dubbed video, and the original is the motion of the lips and jaw in response to the new target audio.

The aim of this paper is to test the feasibility of a novel 3D landmark pipeline that outputs pose, and identity aware talking head landmarks directly in one forward pass given an unseen target audio speech signal, and video to be modified. This differs from other approaches in the literature which typically generate moving lips onto an identity removing, static fixed head before aligning the lips with the desired head pose in a later step. In these approaches the static mesh then must be given identity specific information such as head pose and general head movement by either a separate network that generates artificial head pose sequences (when generating a video from a static image), or by extracting head pose information from the reference video and refitting the static mesh to match it. More commonly when modifying video, an intermediate 3D model is used to generate the desired facial animations, before rendering back to photorealistic frames like in [1]. Typically these methods and techniques are a lot more complex to implement and run than landmark-based solutions. An aim of this work therefore is to take the first steps towards a landmark based video modifying pipeline that may serve as a lighter, simpler, and more practical tool for animation. To this end, two main contributions are made as part of this work:

- A novel lightweight LSTM-Based Model capable of generating pose and identity aware 3D landmark sequences driven by a target audio speech signal and source video clip.
- A novel data augmentation technique for de-correlating lip, jaw, and head motion, making the generation of pose-aware landmarks possible directly at inference.

The rest of this paper is organized as follows. In section 2 a review of recent relevant works in the literature is provided to give context for this paper. To this end a concise taxonomy of papers and methods within the field of audio driven talking head generation is presented. In section 3 the methodology of the approach is reviewed, discussing the contribution of the paper in depth, and detailing the data processing methods, the network architecture, the training set up, and experiments. In section 4 the results are presented and discussed. In section 5 societal impact and ethical considerations of the work are discussed before the conclusion of the paper.

II. RELATED WORKS

Talking head generation is a topic which falls under the wider umbrella of ''Deep Fakes'', where the goal is to generate realistic fake content of a target person. There are many different approaches for generating ''Deep Fakes'', making a detailed literature review of the topic challenging. Here the scope of the related works section is limited to research with a focus on facial animation and motion driven directly from a speech sequence - audio driven talking heads. For a more thorough review of the literature surrounding the topic of ''Deep Fakes'' the reader is directed to [2] as it provides a comprehensive overview of the field and the main methods for generating fake content.

Following a thorough review of the literature surrounding audio-driven talking heads, several interesting pipelines were identified that could be applied to the task of automatic speech dubbing. These pipelines can be broadly classified into two over-arching approaches: Structural approaches [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30] transform the input image/video and audio into an intermediate structural representation (typically, 2D facial landmarks, or a 3D mesh) that is used as input to a neural renderer to generate a photo-realistic talking head sequence. Image reconstruction approaches [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] leverage direct image reconstruction techniques and latent feature learning to generate a photo-realistic sequence from a target audio and reference image/video.

While there are many approaches out there that generate lip movements from audio such as [19], [22], [26], and [42], none of these approaches generate pose-aware landmarks in a single forward pass, instead generating the lip movements onto a static face shape, introducing head movement at a later step. In this paper it is argued that this is inefficient, and can be done directly at inference time through a simple data augmentation. Creating a faster pipeline, with less moving parts, that lends itself better to real time usage.

This work is inspired by and extends the methodology presented in [19] and [20], which are approaches that take in a target audio clip as input, and generate fixed (no head pose, just lip movement) 2D talking face landmarks as output. The approach presented in this paper allows one to generate 3D talking face landmarks that maintain the head pose and identity of the original speaker, while accurately driving the lips from the target audio.

This work is also comparable to [22], which is an approach used to generate talking head animations given a single target image and audio clip. Specifically, one can compare the model in this work to their landmark prediction network which disentangles the audio into content and speaker identity embeddings. These embeddings are used to predict the landmark displacements, which are then rendered into either photo-realistic or animated frames.The approach presented in this paper works on modifying an existing video rather than generating a new one from a single image, and modifies the landmarks based on Mel Coefficients extracted from the audio sequence that are fed into the network.

The aim of talking head generation is to generate a photo-realistic audio driven talking head video in which the facial movements of the talking head are naturally synchronized with the target speech. Note that ''audio driven talking head video'' is used as a blanket term to encompass all works related to generating facial motions and animations driven by audio, regardless of whether it is modifying a preexisting video or animating a static image.

Most of the works referenced in this section, can be classified into one of two fundamental approaches for the task of audio driven talking head generation: structural based methods [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], and image reconstruction-based methods [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41].

Structural Based Methods: These are approaches where the input image, video, or audio are transformed into an intermediate structural representation of some sort such as a 3D model / mesh [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] a sequence of facial landmarks [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], and more recently a sequence of dense motion fields [18], [30]. These are then used as a training feature for an underlying network that takes these structural sequences as input to render a photo-realistic video. These methods are the most relevant to this paper, specifically landmark based ones, as this work introduces a novel way of generating pose aware 3D facial landmark sequences from a preexisting video sequence and target audio clip.

Image Reconstruction Methods: These are the approaches which use pure image reconstruction techniques and latent feature learning [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41]. One could consider these as true ''End to End'' approaches, essentially passing the target image/video and audio through a generative neural network, outputting the synchronized talking head video directly. Other Methods: These are approaches which do not strictly fall within the two classes above, that are still highly relevant to this field and worth mentioning. Approaches such as [42] and [43] which are audio driven models trained to animate face rigs through visemes. Or [44] that can generate dynamic neural radiance fields from audio and using them to synthesize photorealistic talking head videos.

It is also worth noting that each of the categories mentioned above can be further broken down into whether they are image or video-based methods.

- *Image Based Methods:* The goal is to animate a cropped facial image given an input image/limited number of frames as a reference, and an audio clip.
- *Video Based Methods:* Where the goal is to alter the lip movements and facial expressions of an already existing video so that they are synchronized with a new audio clip. Generally, the videos are full frame, containing the background, face, neck, and torso regions, not just the cropped face, unlike the image-based methods. The work presented in this paper is a video-based approach, as it seeks to modify existing video based on a new target speech audio clip.

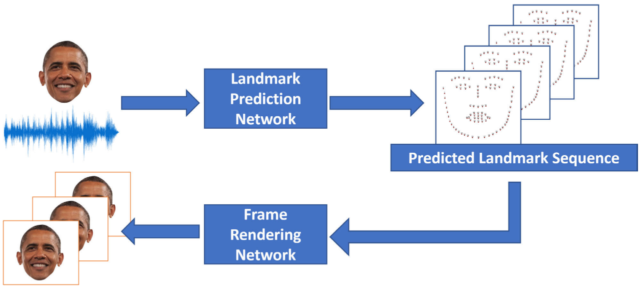


FIGURE 1. Typical landmark based pipeline.

A. LANDMARK BASED METHODS

Typically, with landmark-based methods [16], [17], [18], [19], [20], [21], [22], [23], [24], [26], [27], [28], [29], the goal across all approaches is to generate frame by frame a set of predicted facial landmarks based on a reference image/video, driven by an audio clip. The predicted landmark sequence is then passed through a separate rendering network to generate the photorealistic video frames required for the final output. Figure 1 is a simplified example of what a typical pipeline looks like, note the two main components, the landmark prediction, and the frame rendering modules. As the contribution of this paper is a novel landmark generation technique, this section focuses discussion on the various landmark prediction modules across the literature, with less emphasis on the rendering side of things.

It stands to reason that there is a lot of variation across approaches regarding the most effective method of construction for the landmark prediction module. Most modules in the literature can be grouped according to the following design choices:

- *Audio input pre-processing:* Some approaches take in phoneme labels extracted from audio like in [16]. Others extract Mel spectrograms or MFCCs from the audio first which are then fed into the predictor such as the approaches taken by [18], [19], [20], [23], [24], [25], [26], [28], and [29]. Audio embeddings obtained from trained speech to text modules such as the approaches employed [21] and [27] have also been tried, along with methods that take in custom audio embeddings such as [17] and [22]. For the approach within this paper, mel-coefficients are extracted from the audio and fed in as input features to the network. They were chosen as they are quite easy to extract compared to other audio features used by some of the approaches mentioned above, and they are immensely popular in classical speech related tasks such as text to speech, speaker recognition, and automatic speech recognition.
- *The underlying network architecture:* Some approaches such as [17], [19], [20], [21], [22], [23], [25], [26], and [28]

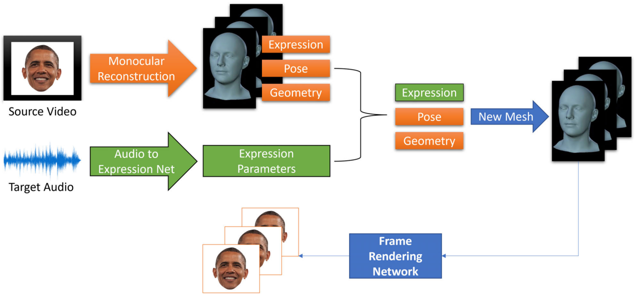


FIGURE 2. Typical 3D Model based pipeline.

employ a recurrent neural architecture and others such as [16], [18], [24], [27], and [29] use feed-forward designs. While feed-forward architectures are typically faster, a recurrent, lstm-based approach very similar to [19] and [22] was chosen for this paper. The idea was that by using a recurrent architecture, the network would learn the temporal dependence associated with audio and its output lip movements, generating a higher quality of lip movements.

- *Generating the Output Landmarks:* Some approaches in the literature generate the output landmarks using a static face mesh with moving lips that needs to then be fitted to a target video such as [16], [17], [19], [20], [23], [26], [27], and [28] while others such as [18], [21], [22], [24], [25], and [29] generate the head pose information using one network, and a second network generates the lip movements, combining both to have a pose inclusive face mesh. This paper's approach differs to these as it uses a single network trained to generate 3D pose aware landmarks synchronized to audio as described by Figure 3. This is done to simplify the overall landmark generation pipeline for faster inference speeds, and doing so allows for the generation of more accurate landmarks as less information is lost through extensive normalisation of the ground truth.

Often, the rendering modules are variations of either CycleGAN [45] or Pix2Pix [46], which are approaches for training a neural network for the task of image 2 image translation. Recently however, denoising diffusion models are becoming more and more popular for the task of image 2 image translation, and it would not be a surprise to see future renderers incorporate the power of these generative models. As the main contribution of this paper is a novel landmark generation module for the task of overdubbing, no further analysis is carried out on these modules as they fall outside the scope of this work.

B. 3D MODEL BASED METHODS

Even though the approach presented in this paper is a landmark based one, it is worth briefly discussing 3D model-based ones [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. Most of these approaches follow the high-level pipeline denoted by Figure 2 above.

Monocular reconstruction is carried out on each frame of the target video, generating a 3D mesh for every frame. From these meshes, pose, facial expression, and geometry parameters are extracted. The target audio is passed through a specially designed ''audio to expression'' network, that can generate blend shape expression parameters from the audio directly. Finally, the newly generated expression parameters are combined with the pose and geometry parameters from the original video, in order to generate a new set of meshes. These are then rendered back into photo-realistic frames with the help of a neural rendering network.

C. IMAGE RECONSTRUCTION METHODS

As mentioned earlier, these are the approaches which use pure image reconstruction techniques and latent feature learning [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], where one can feed in a reference image and target audio clip into the network, and output a photo-realistic talking head video. Because these are typically ''End-to-End'' systems, they have several advantages over structural methods: all parameters can be trained under one loss, they are typically faster, and can be deployed with more ease on neural inference chips. The quality of the videos they produce however are not as good as structural methods, and produce a lot more artifacts, especially when dealing with more extreme head poses. While these methods are exciting and show promising results, it was chosen to use a structural-based method for this paper, as the quality of the final rendered video is significantly better, and more control over the contents of the generated video can be exerted.

III. METHODOLOGY

Following a detailed review of the current literature for automatic dubbing, a gap was identified that provides the basis for this work. Typically, in image-based networks, where the goal is to generate a photo-realistic talking head video from a single reference image and target audio, the common approach is to have one network focus on generating the lip movements, and another network for generating the rest of the facial movements such as head pose, jaw movements, etc. The outputs of both networks are then combined to generate the fully animated facial landmark sequence like in [22]. For video-based methods, the goal is to modify a reference video given a target audio. The approach generally involves generating the lip movements first, before refitting them onto a landmark sequence extracted from the reference video.

As this approach is a video modifying task for the purposes of automatic dubbing, it is proposed to discard the intermediate processing steps mentioned above entirely and train a network to generate audio driven moving lips that are in alignment with 3D head pose extracted from the reference video directly. An advantage to this is that the overall animation pipeline is faster, simpler, saves on compute, and lends itself better to real time applications. Secondly, due to the unique pre-processing approach employed before training, classic normalisation techniques for this task such as removing speaker identity and head motion are not used, allowing the training data to maintain its structural integrity, and therefore the network can learn to generate more accurate and expressive lip movements. This is evidenced by the strong results obtained by the approach in this paper from the subjective user study carried out as part of this work, comparing the method presented in this paper, against other relevant landmark-based techniques from the literature.

Therefore, the contributions in this paper are twofold:

1) A novel LSTM-based pipeline is introduced, that takes as input a target speech clip along with pose and identity parameters extracted from a reference video. The network outputs a pose, and identity aware 3D facial landmark sequence with the lips synchronised to the target speech clip. This approach works in the 3D space and does not use a static face model to first generate the lip movements before retargeting them to a moving one, separating this work from other similar approaches such as [19], [22], and [20]. The model directly outputs lips synchronized to audio, that also follow the head pose and movement of the speaker, simplifying the overall pipeline.

2) A novel data augmentation method is introduced for the pipeline training task, increasing the number of usable audio/visual pairs during training from N pairs up to NxN pairs, allowing the network to better learn the relationship between audio, lip expression, and pose. More precisely, Procrustes alignment is used to take the lip movements corresponding to a given audio signal and apply them to N additional landmark sequences, essentially ending up with a dataset where every audio sequence has N associated landmark sequences, each with unique head pose and movement, but with the lips being synchronized to that respective audio sequence. This augmentation helps when training the network as not only does it provide additional unique data, it decorrelates the lip movement from the rest of the face. During early experiments it was noticed that prior to adding this augmentation that lip movements become strongly correlated with global facial motion and head pose. Extensive details are provided in the data augmentation section on how to implement this and why it is important to do so.

An objective study evaluating the accuracy of the landmarks generated by this method against its ground truth was carried out and compared to other approaches in the literature. Additionally, a subjective user study was also carried out testing the quality of pose-aware landmarks versus other approaches by asking a series of carefully thought questions for each landmark sequence tested. The results of these experiments show it is possible to generate accurate, pose-aware landmarks at inference that are superior than other relevant approaches which use a static face shape and that by simply using the Procrustes lip augmentation at train time, one can generate accurate pose-aware landmarks using any existing method or architecture. Details on these experiments are provided in the results section.

To summarize, this work presents an automatic facial dubbing network that takes in a target speech audio and a reference facial landmark sequence as input. The network modifies the lip displacements of the reference landmark sequence in order to produce a new sequence whose mouth movements are correctly aligned with the speech audio while maintaining the original head movements and poses of the reference video. This is done to keep the actor's performance as close to the original as possible, and not to take away from its quality in any way. See Figure 3 below as it depicts a high-level overview of the network architecture.

A. DATA PROCESSING

1) DATA-SET SELECTION

While the end of goal of any automatic dubbing pipeline is for it to be subject / speaker independent, for the purposes of this paper a single speaker dataset was chosen to establish a proof of concept and determine the different elements of the training pipeline. Therefore, the Obama Weekly Address [47] data-set was chosen. A collection of nearly 300 frontal full-face videos of President Barrack Obama, consisting of over 18 hours of audio-visual content. This dataset was selected for the following reasons:

1) It contains high quality audio, available at a frequency of 48KHz to go with video available at several different resolutions. For the task at hand, a video resolution of 720p was chosen.

2) In most of the videos, President Obama is the only speaker on video, making it very easy to isolate his facial region using an off the shelf face detector, and extract his 3D facial landmark co-ordinates.

3) President Obama is an ideal subject, as in his weekly address speeches he always faces the camera, speaks clearly, and while there is a large amount of variation in the head pose, there are not many extremes.

The native frame rate of the dataset is 29.97 FPS. For the experiments in this paper, the videos were down sampled to 25FPS as it made aligning each frame of audio with its corresponding video frame a much easier task and ensured that no audio information would be lost, i.e., with a frame rate of 25fps, each frame in the video would have an associated audio sequence of 40ms. For training of the network, most of the videos in the dataset were used, with a train/validation/test split of 85/10/5 percent maintained. Lists of the names and indexes of the videos, as well as pre-processing code are available on the project GitHub page, which will be made openly available to the public with the paper.

2) LANDMARK EXTRACTION

Initially, an off-the-shelf facial landmark extractor provided by [48] was employed to extract 68 3D facial landmarks from the individual frames from the videos in the dataset. Unfortunately the quality of the predicted landmarks from this library was found to be highly inconsistent, and to contain a lot of global jitter that had to be eliminated using smoothing techniques. It was found however that even small amounts of smoothing caused the landmarks to lose fine details in the lip motion, reducing the overall quality of the ground truth which ultimately affected the network's ability to generate accurate lip motions. Due to this, it was decided to use the 468 key

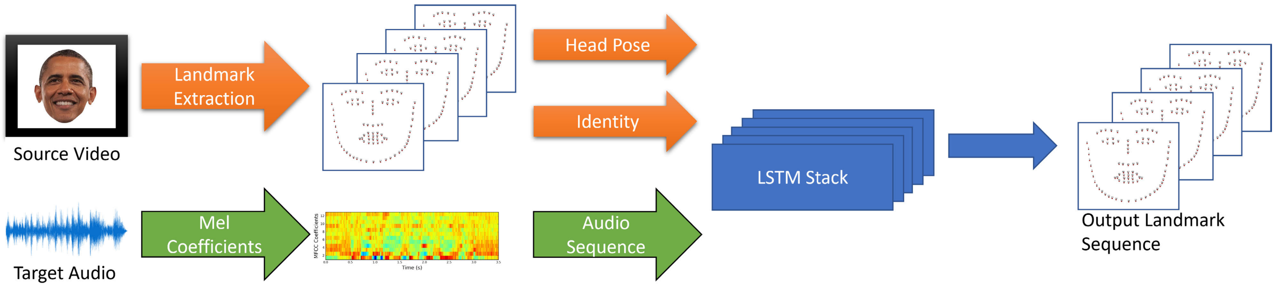


FIGURE 3. High level overview of architecture.

point face mesh extraction algorithm provided by Google's MediaPipe library [49]. Compared to [48], the 3D landmarks were near perfect, and more importantly, had virtually no global jitter present. For the sake of simplicity, 68 of the 468 keypoints that best resembled the landmarks returned by the traditional 68 keypoint dlib extractor were chosen to use as ground truth for training.

The landmark extraction process is very simple. First, the videos are processed with FFMPEG to get rid of any thumbnails at the start of the video or blank frames at the end. Second, individual frames are extracted from the training videos using the Open CV python library. Finally, the 468 3D landmark coordinates are extracted using the media pipe face mesh extractor, before selecting 68 custom keypoints that best resemble the traditional DLIB extractor to use for training the network. This process is done for every video in the dataset, with the code to do so available on the project GitHub page. To prepare the data for training, the landmark frames extracted from each video are combined such that a matrix of shape [N, 68, 3] is created for each video, where N is the total number of frames in that particular video.

Once the 3D facial landmarks have been extracted from every frame in every video, the next step is to normalise all the landmarks, and then apply a smoothing filter to get rid of any remaining jitter present. Normalisation is done by scaling the width of the face and centering the landmarks at the zero point like in [22]. The Savitzky-Golay filter is then used to smooth out the remaining jitter.

3) AUDIO FEATURE EXTRACTION

Once the videos are processed and the landmarks are extracted, the next step is to prepare the audio for training. The audio being used as part of the training set is single channel, has a sampling rate of 48000 Hz and is stored as a WAV file. Remember that since the framerate in the training videos is 25FPS, each frame covers 0.04 seconds of audio information.

The chosen audio features which are to be fed into the neural network are known as Mel Coefficients. These are state of the art features used in many related applications, most commonly in automatic speaker/speech recognition tasks. Reference [50] provide an in-depth explanation of what they are and how they are computed.

The audio signal is framed into 40ms frames, and various experiments were carried out training the network with a range of hop lengths starting from a hop length of size 1920 (no overlapping frames), to 960, to 480, ensuring various degrees of overlap between audio frames. It was decided to not use overlapping frames as no visible difference was noticed in the accuracy of the predicted landmarks against the original. A mel-filterbank of size 80 was also chosen. Therefore, for a 1 second audio sequence, the resulting feature matrix would have shape (25,80).

4) ALIGNING AUDIO WITH LANDMARKS

Now that the audio and landmark features are ready, the next step is to pair them together in preparation for training. For a given video V that contains T number of frames is depicted as V_T. Additionally, for the corresponding audio sequence A, which contains T-1 audio frames, is depicted as $A_{(T-1)}$. Notice that there is one extra video frame at the start of every sequence which is discarded from the audio/landmark pair. This is done as it is assumed that the audio preceding the frame influences it, therefore there is no need to keep the first frame in the sequence as it has no audio associated with it. Note that this assumption is made as the data is being fed into an lstm as a sequence, therefore the network has knowledge of past and future frames. Had we been using an architecture that would generate the output frame by frame, we would need to expand the audio window to cover future frames too. This is to ensure that facial movements caused by plosive sounds would be correctly learnt. The first frame in the video is instead saved as a separate entity from which the identity parameter is extracted for its associated sequence.

The final step is to combine these audio/visual frames into a sequence of 100 pairs for training. 100 is chosen as it is equivalent to 4 seconds worth of audio/visual content (25 fps × 4). This was a simple design choice influenced by the memory constraints of the available GPU. Please see Figure 4 below for a visual description of the alignment process. Note how the first frames in each of the 4 second sequences are discarded as explained above.

B. DATA AUGMENTATION

1) PROCRUSTES LIP AUGMENTATION

In this section "Procrustes Lip Augmentation" is introduced, a novel augmentation technique designed to increase the

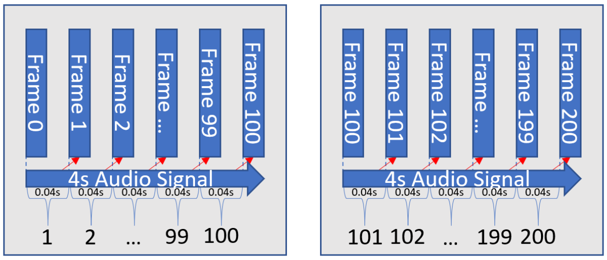


FIGURE 4. Landmark frame / audio sequence pairing process.

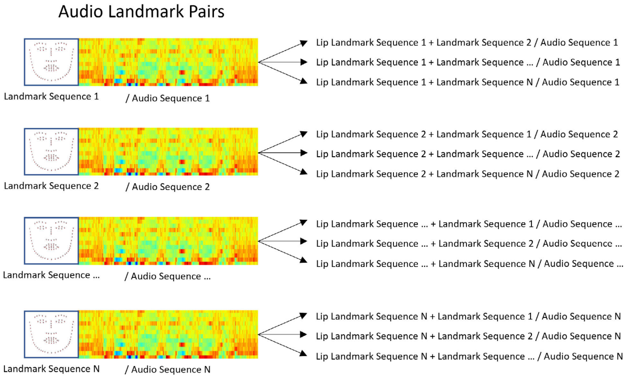

Audio Landmark Pairs
Landmark Sequence 1
/ Audio Sequence 1
Lip Landmark Sequence 1 + Landmark Sequence 2 / Audio Sequence 1
Lip Landmark Sequence 1 + Landmark Sequence ... / Audio Sequence 1
Lip Landmark Sequence 1 + Landmark Sequence N / Audio Sequence 1
Landmark Sequence 2
/ Audio Sequence 2
Lip Landmark Sequence 2 + Landmark Sequence 1 / Audio Sequence 2
Lip Landmark Sequence 2 + Landmark Sequence ... / Audio Sequence 2
Lip Landmark Sequence 2 + Landmark Sequence N / Audio Sequence 2
Landmark Sequence ...
/ Audio Sequence ...
Lip Landmark Sequence ... + Landmark Sequence 1 / Audio Sequence ...
Lip Landmark Sequence ... + Landmark Sequence 2 / Audio Sequence ...
Lip Landmark Sequence ... + Landmark Sequence N / Audio Sequence ...
Landmark Sequence N
/ Audio Sequence N
Lip Landmark Sequence N + Landmark Sequence 1 / Audio Sequence N
Lip Landmark Sequence N + Landmark Sequence 2 / Audio Sequence N
Lip Landmark Sequence N + Landmark Sequence ... / Audio Sequence N

FIGURE 5. Visualisation of the Procrustes lip augmentation process.

number of usable audio-visual pairs available during training from N pairs up to NxN pairs, as well as decouple the relationship between the movement of a person's lips, and the direction, pose, and movement of their face.

Assuming one has a number of aligned audio/landmark sequences denoted as $A_0/L_0 \rightarrow A_N/L_N$ where N denotes the total number of sequences. For a given audio sequence A_0, it's associated lip landmarks are extracted from the overall landmark sequence L_0, and inserted into every other landmark sequence in the dataset using Procrustes Analysis. Through this, one can obtain N sequences of landmarks, where the lip movements are synchronized to the speech from A_0, while the head poses, and head movement are all unique. By doing this, one can successfully de-correlate the relationship between head pose, and lip movement. The following steps depict the process:

2) RATIONALE FOR LIP AUGMENTATION

The technique evolved from some initial experiments, where the model was being trained on speech from a single speaker, and having it output the aligned animated facial landmarks. In this initial training experiment, 4 second long sequences of audio combined with a head position vector at each frame were fed to the network, where the role of the position vector was to provide information about the head pose to the network. The intuition was that the network would take these inputs and use them to output the new pose aware facial landmark co-ordinates, with the lips being synchronised to the audio. The idea was that the audio would drive the movement of the lips, while the position sequence would tell the network the direction in which the head was facing, and generate the position of the lips on the face accordingly.

Rather than having the desired effect of outputting pose aware facial landmarks, the network ended up treating the audio portion of the input as noise, and completely ignoring it. Instead, the network learned how to generate accurate lip movements from the head position sequence alone. A number of tests were carried out to confirm this, specifically silence was fed into the network, along with a variety of head position sequences to test whether the network would still generate lip motion. The tests indicated that the network was ignoring the audio portion of the input entirely, as in each of the tests with silence, the generated lips would still be moving. Therefore it was concluded that there was a strong correlation between the movement the speakers lips in the dataset, and their head pose at any given frame.

This phenomenon led to the realization that in order to train a network to generate audio driven lip landmarks, it is crucial to de-correlate the relationship between the motion of the lips, and the head pose / general movement of the face. It is for this very reason that most approaches in the literature employ a static face mesh during training as it allows their models to learn the movement of the lips with respect to the audio, without having to worry about other aspects like head pose and facial movement. As the purpose of this work is to output pose aware moving lips, a workaround for this issue was necessary.

Initially it was believed that the model was overfitting on the single speakaer dataset, and introduced a multispeaker dataset during training to try and alleviate this issue. Despite this, the network continued to treat the audio portion of the input as noise, learning the lip movement from the head pose sequence alone. It was at this point that the idea to use the "Procrustes lip augmentation" came about. The augmentation had the desired effect, successfully de-correlating the relationship between the head pose, and lip movement in the training data set. This allowed the network to learn to output 3D facial landmark sequences, with the head pose controlled by the pose sequence extracted from a reference video, and the lip movement synchronised to and driven by the target audio. To replicate this augmentation, please see the steps below:

3) STEPS FOR PROCRUSTES LIP AUGMENTATION

1) Process the whole dataset as described in section 3.1, such that you have Audio/Landmark Sequence pairs ready.
2) "Procrustes analysis determines a linear transformation (translation, reflection, orthogonal rotation and scaling) of the points in Y to best conform them to the points in matrix X, using the sum of squared errors as the goodness of fit criterion" [51]. For a given audio/landmark pair, A_0/L_0, take L_0 and run Procrustes analysis against sequence L_1.

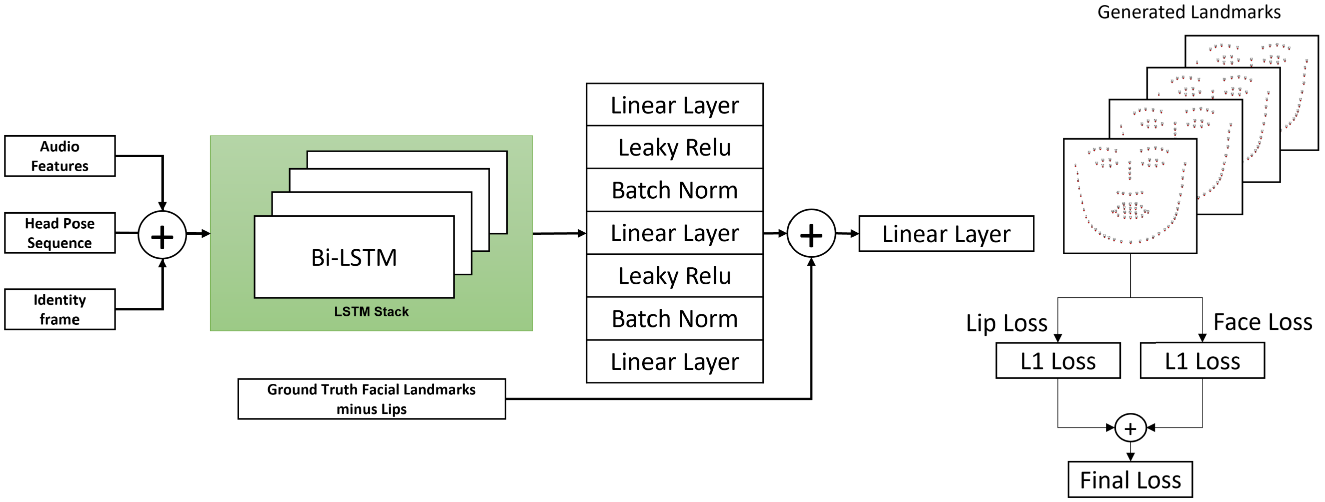


FIGURE 6. High level overview of model architecture.

TABLE 1. Detailed breakdown of the model layers displayed in figure 6 above. The input and output shapes, along with any relevant hyperparameters are included.

Layer Name	Input Shape	Output Shape	Number of Layers	Other Hyperparams
Bidirectional LSTM	89	256	4	Dropout = 0.5
Fully Connected Block:	-	-	-	-
Linear Layer	256	256	1	Bias = True
Batch Normalisation Layer	256	256	1	-
Leaky ReLu Layer	256	256	1	Negative Slope = 0.2
Linear Layer	256	128	1	Bias = True
Dropout	-	-	-	p = 0.5
Batch Normalisation Layer	128	128	1	-
Leaky ReLu Layer	128	128	1	Negative Slope = 0.2
Linear Layer	128	60	1	Bias = True

3) You will have to do it frame by frame. L_0 is Y, and L_1 is X. What you obtain is the conformed sequence $\hat{L}_0$
4) Isolate the lip landmark positions from $\hat{L}_0$ and use them to replace the lip landmark positions in L_1.
5) Repeat steps 2 and 3 for the rest of the landmark sequences in your dataset such that you end up with L_N modified sequences that are synchronized to A_0.
6) Repeat the steps above with the rest of the audio sequences in your dataset $A_{1\rightarrow N}$

Realistically though, one cannot do this for every single audio sample in the dataset as the processing time would take too long. Instead, for every audio sample, 10 random landmark sequences were chosen to do the Procrustes lip augmentation for, increasing the size of the training dataset by 10 times. The number 10 was chosen as no noticeable improvements in the accuracy of the network were discovered by increasing this number further. In fact, even applying the augmentation to 5 landmark sequences for every audio clip was found to be more than enough to de-correlate the audio from the head pose. Please see figure 5 for a visual representation of this process.

C. NETWORK ARCHITECTURE AND TRAINING SET UP

In this section the network architecture, and training set up of the work in this paper is discussed. With a focus on the choice of model, hyper-parameters, and rationale behind certain design choices.

1) NETWORK ARCHITECTURE

The network is a very simple LSTM-based neural network, that takes speech audio features as input combined with a head pose sequence and identity embedding. The network is trained to output the pose-aware facial landmark co-ordinates. This is depicted by the architecture diagram presented in Figure 6.

The audio features are sequences of Mel Coefficients spanning 4 seconds of audio each, as described in section 3.1. They have a shape of (99,80). The head pose sequence is extracted from each frame of the corresponding landmark sequence. For each frame, the "pose" is computed from 3 co-ordinates associated with the nose on the face. In total for a 4-second-long sequence, 99 such head pose embeddings are obtained, having a shape of (99,3,3). The head pose sequence array is then flattened, and concatenated with the mel coefficients array, ending up with a new training feature of shape (99,89). Recall that the first landmark frame in the sequence is removed as it has no equivalent audio information. This frame is saved, and from it the identity parameter is extracted by passing the landmarks extracted from the frame through 3 linear layers, reshaping it to be of size (1,89). This feature is then inserted at the beginning of the training array, ending up with a final feature shape of (100,89).

The input is then passed through a stack of 4 bidirectional LSTMs with an input size of 89, and hidden size of 128. The output of the LSTMs then passes through a linear layer of in size 256, and out size 256. This then passes through a batch normalization layer followed by a leaky ReLU layer with a 0.2 slope coefficient. Another linear layer then takes the embedding of size 256 as input, and outputs one of size 128. Followed by a dropout layer, and another batch norm and leaky ReLU. Finally, one last linear layer of in size 128, outputs an embedding of size 60. This is a flattened set of 20 lip co-ordinates. Please see table 1 for a summary of all layer parameters.

Next, the original landmark sequence minus the lips is concatenated with the newly generated lip co-ordinates. This passes through a final linear layer of in size 204, out size 204 to smooth out any jitter. The output is our new set of generated landmarks for the given audio sequence.

For training the network, L1 loss is chosen as the loss function combined with the ADAM optimizer. Two losses are calculated, a lip loss, and a face loss. The lip loss simply takes the generated lips and compares them to the original lips, while the face loss takes the entire set of landmarks and compares them against the original. The lip loss is weighted 90 percent, while the face loss is given a weight of 10 percent.

2) TRAINING SET UP

The network is trained using a 3070-laptop edition GPU. The training data is prepared and extracted from 200 videos of the Obama Weekly Address dataset. The data is augmented as described in section 3.1, for every audio sequence, 10 random landmark sequences were chosen, and modified their lips such that they would be synchronised for the given audio. Ending up with 10 sequences of unique head motion per audio sequence. The network is trained on the augmented dataset for approximately 12 hours with a learning rate of 0.001 and the ADAM optimizer.The batch size is set to 512, and the Audio/Landmark sequences are shuffled for training.

IV. EXPERIMENTS AND RESULTS

In this section, the experiments and results of this paper are presented and discussed. The results in this work are subjectively compared to the results obtained by works presented in [22] and [20] as these are the methods most relevant to the one in this paper. It was attempted to also compare the model to the approach taken by [17] however the authors have not made the code necessary for this available. Additionally, an objective comparison is also provided between the generated landmarks of this paper versus the ground truth, and those of [19], [22], and [20] and their respective ground truth data. Note that both [19] and [20] use the same approach for generating landmarks. Because this work focuses on the landmark generation aspect of the automatic dubbing pipeline, the evaluation is carried out on the generated landmarks. Sample video renderings that are generated using landmarks extracted from the approach presented in this paper are provided as a proof of concept however a dedicated renderer to transform the 3D landmarks back to 2D RGB frames has not been trained, as that falls outside the scope of this paper.

TABLE 2. Mean opinion score per question.

Models	Q1↑	Q2↑	Q3↑	Q4↑	Q5↓
Ground Truth	3.975	3.839	3.961	3.836	2.554
ATVG Net [20]	2.607	2.982	2.454	2.475	3.514
MakeItTalk [22]	2.65	2.814	2.954	2.564	3.843
Proposed Approach	4.018	3.986	4.029	3.929	2.554

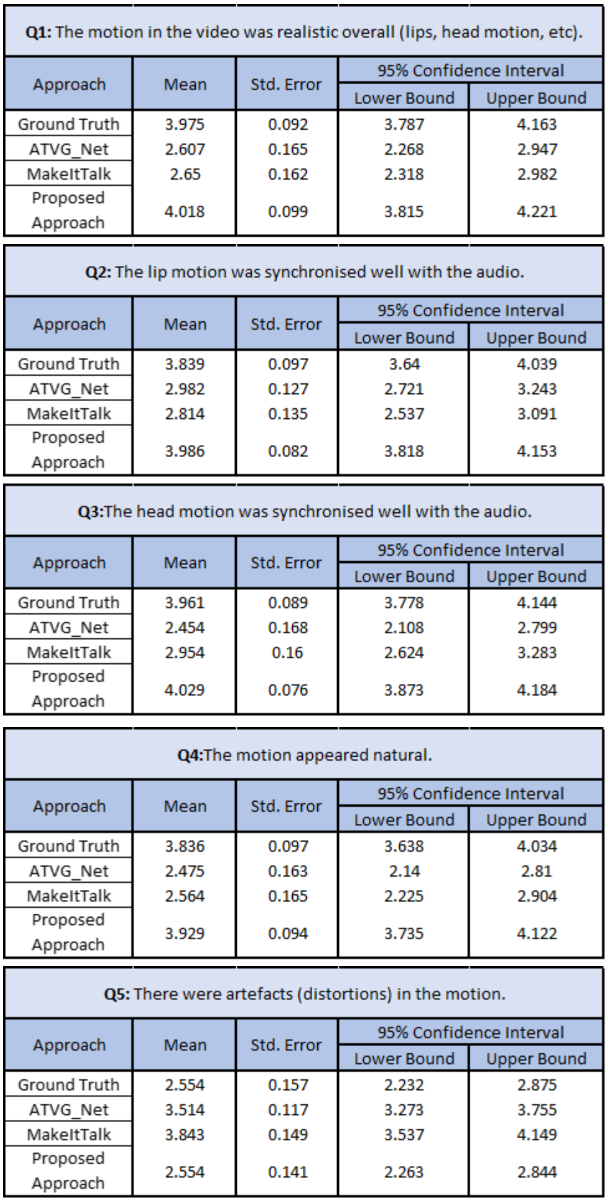
A. SUBJECTIVE USER STUDY

A subjective user study was carried out, evaluating the quality of the landmarks generated by the work in this paper, the work in [22], and the work in [20]. Ten different videos of President Barrack Obama speaking were evaluated per model, 30 (3 models × 10) videos in total. Each video had length 16 seconds. Additionally, ten ground truth videos were also evaluated as part of the study to be used as a baseline. In total, 28 subjects participated, evaluating 40 videos each. Note that the subjects were asked to evaluate videos produced using landmarks, and not the RGB frames. The scale of the study was kept small as the goal was to show that generating highly accurate, pose-aware landmarks at inference is possible, and that the accuracy of the generated lip movements using the method outlined in this paper is comparable to other "static" face based methods.

Subjects were asked to watch each of the 40 videos in random order, and to answer 5 questions per video to evaluate it. The subjects had a choice of 5 answers per question, which were "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree". The subjects were not told which approaches were used to generate the particular video they were evaluating, nor were they told whether the video came from the generated or ground truth set. Table 2 contains a summary of the results, showing the mean score each model obtained per question while figure 7 contains a more detailed breakdown for each individual question. Note that the questions asked are listed above their respective tables. From these results it is clear that the approach presented in this paper produces a model capable of generating audio driven pose-aware landmarks that are near indistinguishable from the ground truth landmarks extracted directly from video. Readers are encouraged to view the generated videos provided in the supplementary materials section to see the accuracy of the model.

B. OBJECTIVE STUDY

Evaluating the the predicted landmarks in an objective manner is a non-trivial task. Distance based metrics are by far the most popular method of evaluating the predicted landmarks against their ground truth, and some type of a distance metric (usually L1/L2 distance) is often used as the loss function during the training phase. As part of this work, an objective study is carried out using the distance-based



Q1: The motion in the video was realistic overall (lips, head motion, etc).

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.975	0.092	3.787	4.163
ATVG_Net	2.607	0.165	2.268	2.947
MakeItTalk	2.65	0.162	2.318	2.982
Proposed Approach	4.018	0.099	3.815	4.221

Q2: The lip motion was synchronised well with the audio.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.839	0.097	3.64	4.039
ATVG_Net	2.982	0.127	2.721	3.243
MakeItTalk	2.814	0.135	2.537	3.091
Proposed Approach	3.986	0.082	3.818	4.153

Q3: The head motion was synchronised well with the audio.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.961	0.089	3.778	4.144
ATVG_Net	2.454	0.168	2.108	2.799
MakeItTalk	2.954	0.16	2.624	3.283
Proposed Approach	4.029	0.076	3.873	4.184

Q4: The motion appeared natural.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.836	0.097	3.638	4.034
ATVG_Net	2.475	0.163	2.14	2.81
MakeItTalk	2.564	0.165	2.225	2.904
Proposed Approach	3.929	0.094	3.735	4.122

Q5: There were artefacts (distortions) in the motion.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	2.554	0.157	2.232	2.875
ATVG_Net	3.514	0.117	3.273	3.755
MakeItTalk	3.843	0.149	3.537	4.149
Proposed Approach	2.554	0.141	2.263	2.844

FIGURE 7. Estimated marginal means calculated for each question the subjects answered.

metrics described by [17], comparing the accuracy in the predicted landmarks from a range of models against their respective ground truths. The ground truth landmarks associated with each of the models were extracted from their respective test sets, and pre-processed in accordance with the instructions provided by their respective GitHub pages, and papers. The landmark distance, and landmark velocity difference [20], [22] functions are used to evaluate the predicted landmarks against their ground truths. The results of these evaluations are provided for in figure 9. Like in [17], the LD and LVD functions are used on the mouth and face area separately. This is denoted by M-LVD, M-LD, F-LVD, and F-LD respectively. Note that F-LD and F-LVD are very low for this paper compared to other approaches because the model is trained with knowledge of what the face looks like, and the direction in which the head is facing.

TABLE 3. Objective evaluation results against GT.

Models	M-LD↓	M-LVD↓	F-LD↓	F-LVD↓
ATVG Net [20]	7.111%	0.947%	7.149%	0.719%
MakeItTalk [22]	8.492%	0.391%	8.896%	0.507%
Proposed Approach	3.042%	0.332%	0.178%	0.001%

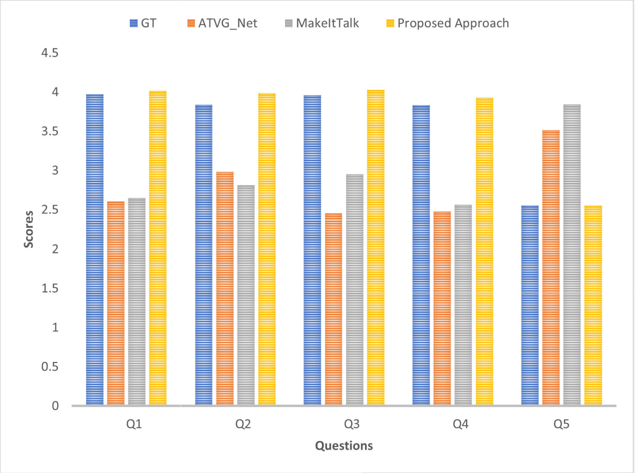


FIGURE 8. Plot of mean scores each model obtained per question.

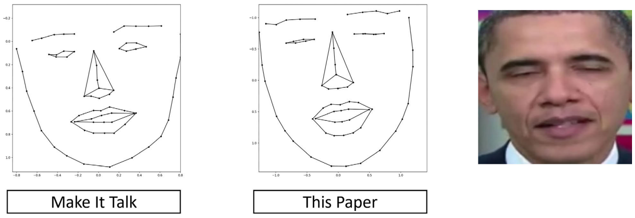


FIGURE 9. Comparison of ground truth landmarks extracted from the same frame.

C. INTERPRETING RESULTS

Both the subjective and objective studies carried out as part of this work show that the approach presented in this paper for generating 3D pose-aware landmarks is a feasible one for generating accurate, and expressive talking head landmarks. It is indeed possible to generate high quality, pose-aware landmarks at inference, without a suffering losses in the quality of the lip synchronization. Based on the subjective results it is clear that subjects preferred talking heads that had identity, and pose information. While the approach presented in this paper slightly outperforms the ground truth in most of the question categories, this can be simply attributed to the very high similarity between the ground truth landmarks and generated ones. A common piece of feedback from subjects who did the study was that they were confused why they were shown two of the same video (recall that the information that one was ground truth and one was generated was not revealed).

Additionally, it can be seen that the approach presented in [22] outperforms [20] in categories related to overall motion, motion naturalness, and head motion, while scoring slightly worse on lip/audio synchronization and motion artefacts. This tracks well as [22] approach is capable of generating realistic head motion for the landmarks, and it can be seen from this study that subjects noticed and preferred that over [20].

V. CONCLUSION

The goal of this paper was to introduce an approach for generating 3D pose and identity aware talking head landmarks given a source video and driving speech signal for the task of automatic video dubbing. It is shown throughout the paper that this is quite feasible to do via a novel data augmentation technique, and that subjects preferred landmarks generated by this approach over other, existing approaches such as [20], [22], and [19]. A number of key insights were gained by conducting this work:

1) Generating 3D pose aware landmarks is possible, and can be easily achieved by de-correlating the relationship between the lips, jaw, and global head movement through the Procrustes lip augmentation that is proposed in this work.

2) The quality of the ground truth data is much more important over the choice of model for learning the relationship between audio and lip movement. Oftentimes noisy data needs smoothing, and smoothing leads to losing valuable lip motion, therefore the model wont be able to learn anything meaningful from audio. This can best be seen from the results of the subjective study where models trained with inferior ground truth scored poorly on metrics such as naturalness and motion distortions.

3) By carrying out a subjective study, and surveying 28 users, it was shown how important the inclusion of head movement information is when evaluating the quality of talking head landmarks.

As part of this work, all data-sets, code, and trained model weights will be made available to the community.

A. FUTURE WORK

This research has opened up a number of potential avenues for future work. At the forefront of these, is the idea to develop a generalised pose-aware model with the capability to few-shot learn individual speaking styles. Over the course of this work, it was discovered that when training a landmark prediction network on a single speaker, the network was robust to generating landmarks from a wide variety of speakers. Regardless of what speech was being input to the network, it was observed that the network would always generate accurate landmarks but in the speaking *style* of President Obama. This indicates that it may be possible to train a generalised model and teach it via few-shot learning techniques to output landmarks in a specific speaker style given a very small amount of data of that speaker.

A dedicated neural renderer for the task of landmark based automatic dubbing is also in the works. Sample renderings were generated using the pretrained model provided by [22] as a proof of concept, however it does not handle extreme variations in the head pose very well as it is an image-based renderer. These can be seen in the supplemental videos section. Training of a video-based renderer is necessary to generate the best possible results. Recent advances in generative neural networks related to diffusion models seem like a promising avenue to explore.

Additionally there is still room to improve the lip landmark generation, increasing its robustness to unseen speakers via deep-learning based audio augmentation techniques such as voice cloning, and synthetic speech generation, as well as more classical approaches like pitch variation, time warping, or noise addition.

B. LIMITATIONS

There are several limitations when generating pose aware landmarks using the method presented in this paper.

1) The network does not generate realistic jaw movements from audio. Due to the nature of the data augmentation (de-correlating lip motion from jaw/head movement), the network is not able to learn to also generate the corresponding jaw movement from the audio. This limitation can be overcome by computing the distance between the upper lip, and lower lip, and raising/lowering the position of the jaw by this amount via a simple linear equation. Alternatively, a very simple network can be trained to solve this, consisting of just a couple of LSTM layers as there is a very direct correlation between lip and jaw movement that can be learnt.

2) Throughout this work it is shown that head pose is related to audio, and a method is demonstrated to decouple this relationship. Due to this, the approach in this paper is not a suitable one for audio-driven video generation. Rather than generating talking head videos from scratch, the proposed network learns to modify an existing video, keeping the original headpose but changing the lip content in response to a new audio signal. This is ideal for the task of dubbing, as it is assumed that the speech content and emotion of the dubbed speech is similar to that of the original. Therefore it is desired that the performance of the actor in the generated video is kept as close to the original performance as possible, including the head movements. However, this is a limitation, because when inputting new speech content that doesn't necessarily match the original headpose, such as silence, the resulting output will contain the original head motion, but with the lips firmly shut. This may lead to the user perceiving the resulting video as being ''unnatural'', however more study in this direction is needed.

3) The approach presented in this paper is a single speaker approach. Because the network was trained using videos and audio from a single speaker (President Barrack Obama), it should not perform as well when exposed to audio from different speakers. That being said, the network is very robust, generating accurate and realistic landmarks from speech coming from a wide variety of speakers who were unseen to the network. Instead, it was observed that the speaking ''style'' of the output landmarks was very similar to that of President Obama regardless of the identity of the input

speech. It is possible to extend this work to be a multi-speaker network by training the network with data processed using the same techniques as employed by the works presented by [22] and [20] combined with the Procrustes lip augmentation.

4) Distance-based metrics are not that useful when attempting to judge the quality of landmarks generated across different models in the literature, especially when one tries to make a direct comparison between said models. For example, consider model A, trained using landmarks extracted, and normalised using pre-processing method A. The quality of the landmarks generated by model A are only as good as the ground truth model A was trained on. Furthermore, any distance metric used for evaluating the landmarks, will be calculated using the predicted landmarks, and it's associated ground truth, therefore one cannot directly compare the landmarks from model A and model B with distance based metrics as they are both likely to have different methods for extracting their ground truths. See figure 8 to see just how different the landmarks extracted from the same frame can be. Due to the reasons outlined above, it is entirely possible that in a comparison between two models, A, and B, where model A has inferior ground truth to B due to variations in the landmark extraction process, model A could report better scores than B even though B may look visually better. Despite this, it is still very useful to provide distance based comparisons between other similar models in the literature, and their respective ground truths, as it helps one gain a rough idea regarding the quality of their generated landmarks with respect to other approaches.

5) As this is an approach towards generating audio driven pose-aware landmarks, rendering the landmarks falls outside the scope of this work. That being said, example renderings of the landmarks generated by this approach using the pretrained renderer from [22] are provided in the supplementary videos section. These videos are there as a proof of concept, showing that one can render high quality videos from the 3D pose-aware landmarks presented in this paper.

C. FINAL REMARKS

While automated dubbing has implications for deep-fakes, it is becoming a reality and the benefits for making entertainment more readily available to a wider and more global audience is important - this doesn't just mean English to other languages - it can also mean content in low-resource languages dubbed back into more realistic English! Major streaming companies already have a lot of non-English content so this is important for the further democratisation of content.

REFERENCES

[1] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's Talkin': Let me talk as you want," 2020, *arXiv:2001.05201*.

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[3] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo, "3D talking face with personalized pose dynamics," *IEEE Trans. Vis. Comput. Graph.*, early access, Oct. 4, 2021, doi: 10.1109/TVCG.2021.3117484.

[4] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.

[5] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," 2020, *arXiv:2002.10137*.

[6] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," 2019, *arXiv:1905.03079*.

[7] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3867–3876.

[8] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 3660–3669.

[9] H. Wu, J. Jia, H. Wang, Y. Dou, C. Duan, and Q. Deng, "Imitating arbitrary talking style for realistic audio-driven talking face synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1478–1486.

[10] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "LipSync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 2754–2763.

[11] A. Richard, M. Zollhofer, and Y. Wen, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1173–1182.

[12] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," 2019, *arXiv:1912.05566*.

[13] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 12, pp. 3457–3466, Dec. 2020.

[14] L. Song, B. Liu, G. Yin, X. Dong, Y. Zhang, and J.-X. Bai, "TACR-Net: Editing on deep video and voice portraits," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 478–486.

[15] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," 2020, *arXiv:2007.08547*.

[16] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[17] X. Ji, H. Zhou, K. Wang, W. Wu, C. Change Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," 2021, *arXiv:2104.07452*.

[18] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2Head: Audio-driven one-shot talking-head generation with natural head motion," 2021, *arXiv:2107.09293*.

[19] S. E. Eskimez, R. K Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," 2018, *arXiv:1803.09803*.

[20] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7824–7833.

[21] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: Real-time photorealistic talking-head animation," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–17, Dec. 2021.

[22] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," 2020, *arXiv:2004.12992*.

[23] D. Aneja and W. Li, "Real-time lip sync for live 2D animation," 2019, *arXiv:1910.08685*.

[24] S. Biswas, S. Sinha, D. Das, and B. Bhowmick, "Realistic talking face animation with speech-induced head motion," in *Proc. 12th Indian Conf. Comput. Vis., Graph. Image Process.*, Jodhpur, India, Dec. 2021, pp. 1–9.

[25] W. Wang, Y. Wang, J. Sun, Q. Liu, J. Liang, and T. Li, "Speech driven talking head generation via attentional landmarks based representation," in *Proc. Interspeech*, Oct. 2020, pp. 1326–1330.

[26] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1031–1044, Oct. 2021.

[27] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded GANs for learning of motion and texture," in *Computer Vision—ECCV 2020*, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 408–424.

[28] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[29] T. Xie, L. Liao, C. Bi, B. Tang, X. Yin, J. Yang, M. Wang, J. Yao, Y. Zhang, and Z. Ma, "Towards realistic visual dubbing with heterogeneous sources," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1739–1747.

[30] R. Zhao, T. Wu, and G. Guo, "Sparse to dense motion transfer for face image animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1991–2000.

[31] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," 2020, *arXiv:2008.10010*.

[32] G. Mittal and B. Wang, "Animating face using disentangled audio representations," 2019, *arXiv:1910.00726*.

[33] H. Zhu, H. Huang, Y. Li, A. Zheng, and R. He, "Arbitrary talking face generation via attentional audio-visual coherence learning," 2018, *arXiv:1812.06589*.

[34] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "End-to-end generation of talking faces from noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1948–1952.

[35] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," 2018, *arXiv:1803.10404*.

[36] H. Zhou, Y. Sun, W. Wu, C. Change Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," 2021, *arXiv:2104.11116*.

[37] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," 2019, *arXiv:1906.06337*.

[38] N. Kumar, S. Goel, A. Narang, and M. Hasan, "Robust one shot audio to video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 3334–3343.

[39] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," 2018, *arXiv:1807.07860*.

[40] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," 2018, *arXiv:1804.04786*.

[41] J. Son Chung, A. Jamaludin, and A. Zisserman, "You said that?" 2017, *arXiv:1705.02966*.

[42] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[43] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "VisemeNet: Audio-driven animator-centric speech animation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–10, Aug. 2018.

[44] Y. Guo, K. Chen, S. Liang, and Y.-J. Liu, "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5784–5794.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*.

[46] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*.

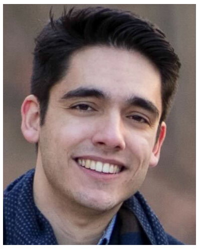
[47] *Your Weekly Address*, The White House, Washington, DC, USA, May 2015.

[48] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.

[49] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[50] L. Roberts, "Understanding the Mel spectrogram," Tech. Rep., Mar. 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

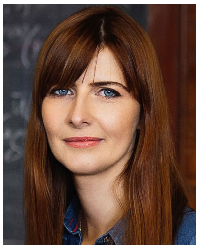
[51] *Procrustes Analysis—MATLAB Procrustes*. [Online]. Available: https://www.mathworks.com/help/stats/procrustes.html



DAN BIGIOI (Graduate Student Member, IEEE) received the bachelor's degree in electronic and computer engineering from the University of Galway, in 2020, where he is currently pursuing the Ph.D. degree, sponsored by D-REAL, SFI Centre for Research Training in Digitally Enhanced Reality. Upon graduating, he worked as a Research Assistant at the University of Galway studying text to speech and speaker recognition methods under the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project. His research interests include studying and implementing novel deep learning-based techniques for automatic speech dubbing and discovering new ways to process multi-modal audio/visual data.

HUGH JORDAN received the B.A.I. and M.A.I. degrees in computer engineering from Trinity College Dublin, in 2021. He is currently a Ph.D. Researcher with Trinity College Dublin and the SFI Centre for Research Training in Digitally-Enhanced Reality. His research interests include audio-driven facial animation for automatic dubbing and virtual humans.

RISHABH JAIN (Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the National University of Ireland Galway (NUIG), in 2020, where he is currently pursuing the Ph.D. degree. He is also working as a Research Assistant at NUIG under Data-Center Audio/Visual Intelligence on-Device (DAVID) project. His research interests include machine learning and artificial intelligence, specifically in the domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.



RACHEL MCDONNELL is an Associate Professor of creative technologies at Trinity College Dublin, Ireland. Her research focuses on animation of virtual characters, using perception to both deepen our understanding of how virtual characters are perceived, and directly provide new algorithms and guidelines for industry developers on where to focus their efforts. She has published over 100 papers in conferences and journals in her field, including many top-tier publications at venues such as SIGGRAPH, Eurographics, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. She is a regular member of many international program committees (including ACM SIGGRAPH and Eurographics). She serves as an Associate Editor for journals, such as *ACM Transactions on Applied Perception*, *Computers and Graphics*, and *Computer Graphics Forum*.

PETER CORCORAN (Fellow, IEEE) currently holds the Personal Chair in electronic engineering with the College of Science and Engineering, National University of Ireland Galway (NUIG). He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of *IEEE Consumer Electronics Magazine*.