Real-Time Speech-Driven Avatar Animation by Predicting Facial landmarks and Deformation Blendshapes

Juan Camilo Vásquez-Correa[1], **Santiago Moreno-Acevedo**[1], **Ander Gonzalez-Docasal**[1], **Aritz Lasarguren**[2], **Jone López**[2], **Egoitz Rodriguez**[2] and **Aitor Álvarez**[1]
[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia – San Sebastian, Spain
[2]Baleuko S L, Durango, Spain
jcvasquez@vicomtech.org

Abstract

The evolution of virtual spaces and live events demands sophisticated methods for avatar animation. While existing techniques offer diverse approaches, limitations persist in achieving real-time responsiveness and natural communication. This paper proposes a novel approach for real-time speech-driven avatar animation, covering the prediction of 2D and 3D facial landmarks, and deformation blendshapes from ARkit. Specific models were trained to generate both emotional and neutral animated faces, and using convolutional neural networks able to deal with low latency requirements. The quality of the generated animations was addressed both objectively and subjectively. Both evaluations suggest that our approach is accurate to generate high-fidelity and expressive animations. In addition, we create a client-server application that achieved real time performance, enabling frame rates and latencies suitable for live interactions, fostering a seamless and immersive experience.

1 Introduction

Modern animated movies and games rely on expressive facial animation to convey emotions and enhance storytelling. While vision-based technology plays a vital role in capturing real actors' performances and translating them onto animated characters, it often comes at a significant cost (Karras et al., 2017). Elaborated hardware setups are frequently required for computer vision systems, and re-shoots necessitate the actors' physical presence and consistent appearance. Conversely, speech-driven algorithms are a compelling alternative by significantly reducing costs. For instance, animating vast amounts of in-game dialogue becomes significantly cheaper through audio processing instead of costly video capture setups (Karras et al., 2017). Additionally, speech-driven systems can leverage natural animations even from Text-to-Speech models, opening up new possibilities for character creation.

When generating facial animations from speech, it is important not only to ensure lip-sync, but also to transfer the emotions of the speaker into the avatar to guarantee a more natural communication (Chen et al., 2023). Humans are experts in facial reading, making inconsistencies between speech and facial expression to be potentially distracting, unpleasant, and even confusing. This is evident in the McGurk effect, where mismatched visual and auditory speech can alter perceived words (Alsius et al., 2018). Therefore, high-fidelity speech animation becomes essential for conveying emotions, intentions, and creating truly immersive experiences.

Speech facial animation technologies fall into two broad categories based on complexity and expressiveness. Some engines leverage large-scale neural models for highly nuanced animation, as described by Yang et al. (2023); Zhao et al. (2024). However, these solutions often demand significant computational resources, limiting their suitability for resource-constrained projects. On the contrary, simpler libraries based on viseme recognition (Edwards et al., 2016) offer faster animation, but are often criticized for lacking emotional expressiveness and intent transfer (Taylor et al., 2017).

Despite advancements in speech-driven animation, achieving real-time performance and seamless integration with animation software remains a challenge. Current systems are based on facial landmark predictions (Taylor et al., 2017; Eskimez et al., 2019; Vidal and Busso, 2023) and 3D facial meshes (Chen et al., 2023; Thambiraja et al., 2023; Zhao et al., 2024), which are able to produce high fidelity and natural animations. However, they have limitations in computational efficiency and software compatibility. Additionally, approaches directly mapping speech to video animations (R. et al., 2023; Zhang et al., 2024) often prioritize

expressiveness over real-time performance, hindering practical applications. Existing models rely on large-scale architectures like Recurrent Neural Networks (RNNs) (Pham et al., 2017; Eskimez et al., 2018; Y. et al., 2020; Zhou et al., 2020; Villanueva et al., 2022), Transformer models (Chen et al., 2023; Yang et al., 2023; Xing et al., 2023; Zhang et al., 2023), diffusion models (Thambiraja et al., 2023), and Generative Adversarial Networks (GANs) (K. and E., 2021; Zhang et al., 2021; Vougioukas et al., 2020). RNNs effectively model temporal dependencies in speech, transformers excel at long-range context analysis, and GANs can generate highly natural animations. Although current techniques offer high quality animations, they struggle to achieve the real-time responsiveness and natural communication cues necessary for truly immersive experiences.

The growing sophistication of virtual spaces and interactive live events demands new methods for avatar animation that go beyond high fidelity. This paper addresses current limitations by proposing a novel, real-time speech-driven avatar animation engine to bridge the gap between high-fidelity visuals and smooth interaction during interactive live events. We considered deep architectures to generate several animation representation types in real time, including 2D/3D facial landmarks, and ARkit deformation blendshapes[1]. Several studies have addressed the prediction of facial landmarks based on speech (Taylor et al., 2017; Zhou et al., 2020) as part of their pipelines. Studies relying on ARkit blendhapes have focused on performing audiovisual speech synthesis, using adaptations of Tacotron2 (Hussen Abdelaziz et al., 2021). However, such approaches limits both the emotional expressiveness that real actors can transmit to the generated faces. To the best of our knowledge, this is one of the first studies focused on predicting ARkit blendshapes directly from speech, and the first one aiming to generate them in real time, paving the way for expressive and interactive avatars during live animation events.

The performance of the proposed models is evaluated both objectively and subjectively in order to check not only the accuracy of the predicted landmarks and blendshapes, but also perceptual indicators about expressiveness, coherence, quality, and lip-sync. In particular, subjective tests are conducted by a group of 3D animation experts, increasing the novelty of the proposed approach with respect to related studies that have perform subjective tests only with naive users (Y. et al., 2020). Furthermore, we performed an extensive evaluation of the run-time capabilities that are essential for real-time animation production in live events. Unlike previous studies focusing only on limited audio samples and single frame prediction times (Tian et al., 2019; Lu et al., 2021), our work provides a more comprehensive assessment in production-ready environments typically found in live events.

2 Methods

2.1 Facial Animation Representations

We incorporated three animation representation types to address different application scenarios when animating avatars: (1) 2D facial landmarks, (2) 3D facial landmarks, and (3) deformation blendshapes. These animation types are intended to be transmitted in real time to animation engines like Unity[2], Blender[3], or Maya[4] to animate cartoon-type avatars that follow the facial expressions of an actor. Each representation is considered depending on the type and realism of the avatar to be animated.

Facial landmarks are key reference points on a face, used to track movement, expression, and individual facial structures on a coordinate system. We considered both 2D and 3D facial landmark representations that are automatically extracted from video frames. The 2D landmark points correspond to 68 x-y coordinates extracted using the DLib library (King, 2009), and which have been used in similar studies to map the general facial structure (R. et al., 2023; Eskimez et al., 2018) (see Figure 1a). The 3D landmark representation consists of 478 x-y-z coordinates extracted using the MediaPipe Facemesh model from Google (Grishchenko et al., 2020; Yan, 2022), and which is able to extract more fine-grained information from the facial structure and map it into more realistic 3D avatars (see Figure 1b).

Complementary to facial landmarks, blendshapes are pre-sculpted variations of an object e.g., the face, used to smoothly animate complex deformations of its geometry. Blendshapes are standard animation mechanisms widely used in professional animation engines. We considered a stan-

[1] https://developer.apple.com/documentation/arkit/

[2] https://unity.com/

[3] https://www.blender.org/

[4] https://www.autodesk.es/products/maya

dard set of 52 ARKit blendshapes[5] that allow to animate the eyebrows, mouth, jaw, and lips in different ways, and adapt the coefficients to a specific avatar (Figures 1c and 1d). The set of blendshapes was extracted using the MediaPipe Blendshape V2 model (Grishchenko et al., 2022).

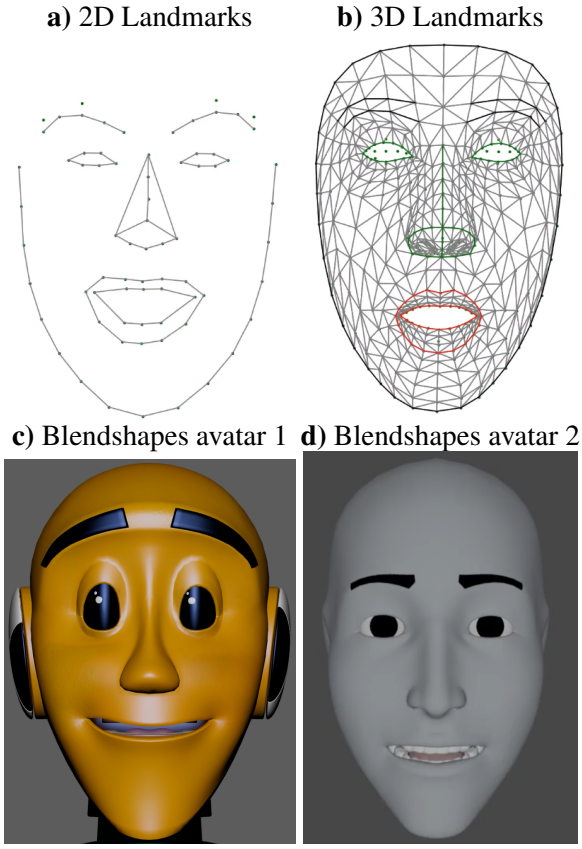

Figure 1: Facial representations for speech-driven avatar animation, covering facial landmarks and deformation blendshapes.

2.2 Deep Architectures

Several models for the three animation representations were trained using combinations of convolutional and recurrent networks. In particular, we considered the Long Short-Term Memory (LSTM) network proposed in (Eskimez et al., 2018) as a baseline, an adapted version of the 1D-CNN used in (Eskimez et al., 2019), and a CNN built on top of SincNet filters (Ravanelli and Bengio, 2018). These architectures were selected with the purpose of producing speech-driven animations in real time. As a consequence, more complex and bigger models like those based on diffusion (Zhao et al., 2024; Zhang et al., 2024; K. and H., 2023) or Transformers (Chen et al., 2023; Xing et al., 2023) were not considered.

The baseline model from (Eskimez et al., 2018) uses the first and second order temporal difference of log-Mel spectrograms as input of a four-layer LSTM network. This network was trained to generate 2D landmark points with a temporal resolution of 40 ms.

The second considered model is a 4-layer 1D-CNN (kernel size of 21 and number of channels=$\{64, 128, 256, 512\}$, respectively), adapted from (Eskimez et al., 2019), and which is trained to predict Point Distributed Models (PDMs) for 2D/3D landmarks, and the 52 ArKit blendshapes. PDMs reduce variability in landmark predictions due to face shape, scale, and orientation (Cootes et al., 1995). These PDMs are shape models that represent the high-dimensional landmark space with a set of coefficients obtained after PCA decomposition. The output of the last convolutional layer is finally processed by a linear layer to make the final predictions of the landmarks and blendshape coefficients. The CNN receives as input 280 ms of the raw speech waveform (7 frames of 40 ms) and predicts the PDM coefficients of the central frame, using the remaining frames as past and future context.

Finally, we propose the use of a SincNet model (Ravanelli and Bengio, 2018) trained also to predict the PDM coefficients for 2D/3D landmarks, and the blendshapes. Our model consists of a SincNet layer fed by 280 ms of the raw waveform and which generates speech tokens with a 40 ms resolution. The output of the SincNet layer is then processed by two convolutional layers and two linear layers to make the final prediction of the PDMs or the blendshapes.

For training all considered models, we employed the Smooth-L1 loss function and implemented a 5-fold speaker independent cross-validation strategy, using four folds for training and development, and the remaining one for independent testing. The models are trained sing Adam, with a batch size of 32 audio samples, a learning rate of 10^{-5} and dropout of 0.1. The dimension of the PDMs was set to 20 when predicting the landmarks, keeping 0.99 of the cumulative variance when computing PCA. The models were trained during 20 epochs.

Finally, to reduce high-frequency noise, particularly visible as tremors in the eyebrows and eyes, the predicted blendshapes undergo post-processing with a Savitzky-Golay filter (Schafer, 2011). This

[5] https://arkit-face-blendshapes.com/

filter smoothes the data while preserving underlying trends, resulting in more natural and visually appealing facial animations.

2.3 Real Time Processing

We developed a client-server application utilizing FFmpeg[6] and Websockets for real-time audio stream processing. The client transmits continuous audio streams of 1024 bytes (corresponding to 32 ms of audio sampled at 16 kHz and 16-bit resolution) to the server. The server continuously receives and buffers the stream, maintaining a processing buffer. Once the buffer reaches 280 ms (7 frames of 40 ms), the server predicts facial animations for the central frame and sends the results back to the client for visualization and integration with animation engines. After processing, the server releases the corresponding 40 ms audio segment from the buffer and waits for new frames to arrive. An overview of the processing setup is shown in Figure 2.

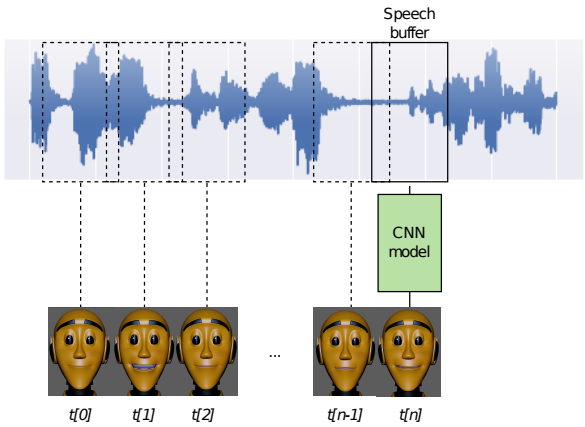

Figure 2: Overview of the proposed system for real time avatar animation. An animation is generated every 40 ms from a buffer size of 280 ms. This distribution guarantees a continuous stream of blendshapes and landmarks at 25 FPS, with a delay of 280 ms.

This configuration ensures a continuous stream of 2D/3D landmarks or blendshapes from the server to the client with an average rate of 25 frames per second (FPS) and a processing delay of 280 ms (reflecting the queue length used for context during prediction). Crucially, single frame processing time must be guaranteed to be less than 32 ms (duration of the received audio stream) to avoid queue build-up and maintain uninterrupted streaming. Sending larger audio chunks leads to faster queue filling and requires consecutive frame processing, potentially causing server response delays and packet loss due to queue overflow. Finally, with the aim to generate more natural animations, artificial blinks were introduced in the server predictions every 5 seconds (with a certain probability) by modifying the corresponding blendshape coefficients or the 2D/3D landmarks.

[6] https://ffmpeg.org/

3 Data Description

The animation models were trained using the CREMA-D (Cao et al., 2014) and the Grid (Cooke et al., 2006) corpora. These datasets were selected with the aim to have individual models for emotional and neutral speech-driven facial animations (see Table 1). Both corpora have been used in similar studies, particularly in realistic talking face generation (Vougioukas et al., 2020; Kefalas et al., 2020). Labels for 2D/3D landmarks, and blendshape coefficients were extracted from videos using the methods described in Section 2.1.

	CREMA-D	**Grid**
Emotions	Six emotions	Neutral
# Utterances	7,442	34,000
Duration (hours)	6.2	28.3
# Sentences	12	1000
# Speakers	91	34
Camera	Panasonic AG-HPX170	Canon XM2
Video	Flash at 30 FPS 480x360	MPEG at 25 FPS 360x288

Table 1: Information of source corpora used to train the speech-driven facial animation models.

CREMA-D (Cao et al., 2014) is an emotional multimodal acted dataset, used traditionally for speech emotion recognition. Actors spoke a selection of 12 sentences in six emotions (Anger, Disgust, Fear, Happiness, Neutral, and Sadness) and three emotion levels (Low, Medium, High), in English. Models trained with this dataset will generate more expressive and emotional animations.

The Grid Corpus (Cooke et al., 2006) was designed for audiovisual speech recognition, in English language. The dataset includes high-quality audio and facial video recordings of 1,000 sentences spoken by 34 subjects (18 male, 16 female). The sentences spoken by each actor are composed of six words randomly chosen from a limited dictionary. Although this corpus has a restricted vocabulary, it was selected to facilitate the development of models capable of generating accurate animations with high lip-synchronization quality (Vougioukas et al., 2020) and to serve as a benchmark for measuring the potential performance limits of the trained models.

4 Experiments and Results

4.1 Objective Evaluation

The 2D and 3D landmark prediction models were evaluated using different metrics from the literature (Zhou et al., 2020). In particular, we included the landmark distance (L-D), the landmark velocity difference (L-VD), the L-D for jaw-lips, and the difference in the open mouth area (OMA-D). We introduced additional metrics to specifically evaluate lip-sync quality such as the L-D and L-VD for mouth-specific landmarks.

Table 2 shows the quality evaluation of the models to predict 2D landmarks. Both 1D-CNN and SincNet networks achieved significantly lower errors than the baseline (for L-D related metrics), and especially for the landmarks defining the jaw-lips and the mouth, as demonstrated by one-way ANOVA with pairwise Tukey post hoc tests (p-value $\ll 0.005$). These results were observed for both neutral and emotional models trained with respective datasets (Grid and CREMA-D). For velocity related metrics that evaluate the temporal dynamics of the facial animations, the baseline models exhibited lower error rates. This is expected due to the nature of the LSTM-based model from the baseline, which is better to model temporal dependencies. However, note that the recurrent nature of such model make it not being able for real time processing, which is a key objective of this work. In addition, no differences were found between the 1D-CNN and SincNet predictions (p-value $= 0.652$). Finally, the models performed better at predicting facial landmarks for neutral faces than the emotional ones from the CREMA-D corpus (p-value $\ll 0.005$). A separate analysis showed that within the CREMA-D corpus, facial landmarks of low-arousal emotions (sadness and disgust) had lower prediction errors than high-arousal ones.

	CREMA-D			Grid		
Metric	**Baseline**	**1D-CNN**	**SincNet**	**Baseline**	**1D-CNN**	**SincNet**
L-D	1.28	1.06	1.20	0.81	0.52	0.52
L-VD	5.00	5.41	5.38	4.21	4.37	4.55
L-D jaw-lips	0.73	0.56	0.59	0.77	0.37	0.36
L-VD jaw-lips	4.81	5.55	5.58	4.53	5.11	5.09
OMA-D	0.51	0.31	0.72	0.31	0.16	0.17
L-D mouth	1.54	1.21	1.56	1.00	0.54	0.55
L-VD mouth	5.79	6.11	6.36	4.49	4.63	4.99

Table 2: Error metrics (%) for the prediction of 2D landmarks.

Table 3 presents the results predicting 3D facial landmarks. The baseline models was not considered here considering again that our ultimate goal is to perform real time predictions, which the baseline model is not able to achieve. The errors were higher than those reported for 2D landmarks, which is expected because the significantly larger number of points to predict (over 10 times more). However, the errors remained below 6 % for the entire set of landmarks, and below 4 % for the ones related to the mouth movement. In this case, the 1D-CNN model surpassed SincNet, with statistically significant lower errors (p-value $\ll$ 0.005) for both neutral and emotional datasets.

	CREMA-D		Grid	
Metric	**1D-CNN**	**SincNet**	**1D-CNN**	**SincNet**
L-D	3.45	3.64	5.85	6.00
L-VD	9.92	13.1	6.32	7.47
L-D jaw-lips	1.54	1.83	5.97	5.92
L-VD jaw-lips	9.78	14.87	5.28	6.41
OMA-D	0.23	0.23	0.20	0.20
L-D mouth	3.51	3.69	3.75	3.79
L-VD mouth	10.74	14.73	8.03	9.88

Table 3: Error metrics (%) for the prediction of 3D landmarks.

Finally, the quality of the blendshape predictions is evaluated with the average Mean Absolute Error (MAE) of the 52 blendshapes, and subsets related to specific facial areas such as the mouth, cheeks, jaw, eyes, and eyebrows. The results are shown in Table 4. Similar to the 2D landmarks case, there were no significant differences between the predictions obtained with the 1D-CNN and the SincNet models. Moreover, the neutral blendshapes from the Grid corpus were more accurately predicted, similar also to the 2D-landmark scenario. Regarding the blendshape generation of specific parts of the face, the cheek and jaw areas were the most accurately modeled, while the eyes and eyebrows were the most challenging to predict.

	CREMA-D		Grid	
Face area	**1D-CNN**	**SincNet**	**1D-CNN**	**SincNet**
All	14.67	14.33	9.65	9.74
Mouth	13.78	13.55	8.14	8.12
Cheeks	1.21	0.61	1.15	0.43
Jaw	6.55	6.44	6.08	5.94
Eyes	19.26	19.10	15.01	15.44
Eyebrows	28.57	27.58	14.69	15.52

Table 4: MAE (%) for the prediction of Blendshape coefficients.

4.2 Subjective Evaluation

Even though the previous results can evaluate the deviation of the reconstructed landmarks and blendshapes from the ground truth values, they are not able to measure the subjective aspects that come naturally to human viewers. We considered the emotional 1D-CNN model trained to predict ARkit blendshapes, and generated 40 videos featuring two emotions (euphoria and fear) on two different avatars (Figures 1c and 1d). The videos were produced using 20 independent audio samples (10 per emotion) recorded by an actress. 23 participants, divided into two groups (12 naive users and 11 3D animation experts), rated the videos. Each video received scores from 1 to 5 (with higher scores indicating better quality) across four criteria: (1) expressiveness, (2) coherence between the emotions conveyed by speech and facial expressions, (3) quality, which refers to the global quality of the animation in terms of realism, fluency and precision, and (4) lip-synchronization to measure how well the lip movements of the speaker matches the corresponding audio. The results are shown in Figure 3.

Our results are consistent with those reported in similar studies when animation and lip-sync quality are subjectively rated (Y. et al., 2020). There were no significant differences in the scores between the two rater groups (Mann Whitney U test, p-value > 0.05), although we observed that 3D animation experts usually assign higher scores than naive users. This can be likely explained because they are more aware of the difficulties of creating high-quality animations. In terms of emotions, the scores assigned for fear were slightly higher. However, they did not differ significantly from the obtained for euphoria. Finally, we observed that the perceived quality depends on the selected avatar. Users rated the avatar 1 (yellow avatar in Figure 1c) significantly higher, mainly because it has less human-like features. Therefore, the importance of correctly producing visemes was less important, contrary to the avatar 2 (gray avatar in Figure 1d).

4.3 Real Time Performance

The application was tested in an experimental setting consisting of separate client and server machines connected via WiFi through a VPN. The hardware specifications for both machines are shown in Table 5.

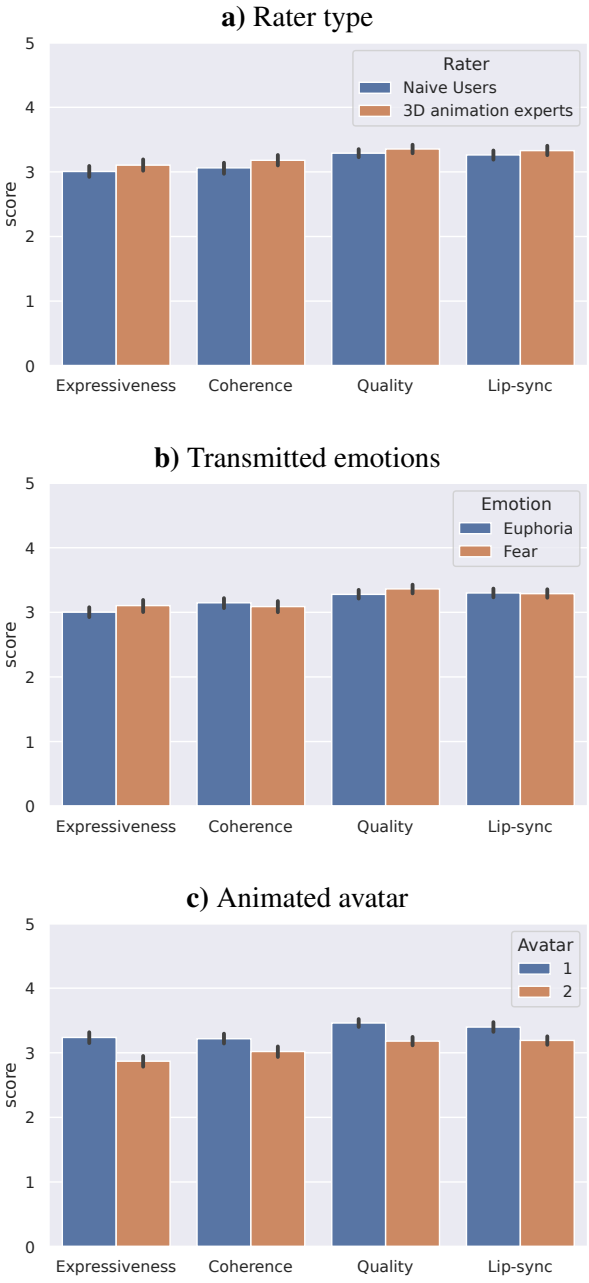

Figure 3: Subjective evaluations performed on the predicted ARkit blendshapes of 20 independent audio utterances recorded by an actress. The evaluations are discriminated in terms of the type of rater, the transmitted emotion, and the type of avatar.

	Client	**Server**
CPU	13th Gen Intel(R) Core i7-1355U 10 cores	Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10GHz 16 cores (x2 Threads)
RAM	16 GB	128 GB
GPU	-	NVIDIA TITAN X (Pascal) 12GB

Table 5: Hardware specifications of the client and server machines for the real time evaluation

We evaluated the Real Time Factor (RTF) when predicting 2D landmarks from the CREMA-D corpus using the baseline, the 1D-CNN, and the SincNet models. The results are shown in Figure 4. Both the 1D-CNN and the SincNet models are suitable for real-time predictions as they achieved RTF $\ll 1$. Conversely, the recurrent nature of the LSTM model from the baseline resulted in an RTF > 1, making it unreliable for real-time predictions. Considering also that the 1D-CNN is the most accurate model for predicting landmarks and blendshapes, this model was used to test the reliability of a real application for performing avatar animations during continuous audio streams.

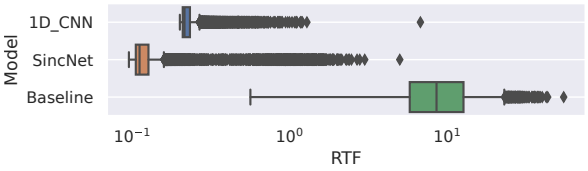

Figure 4: RTF when predicting 2D landmarks from the CREMA-D corpus using the baseline, the 1D-CNN, and the SincNet models.

To further evaluate the run-time performance, a one-hour speech stream was transmitted from client to server for real-time prediction of 2D/3D landmarks and blendshapes. Table 6 and Figure 5 summarize the performance in terms of several resources and quality metrics.

Run-time Metric	2D Land.	3D Land.	Blendshapes
RTF model prediction	0.07	0.14	0.07
Processed packages (%)	99.9	99.9	99.9
Maximum latency (ms)	21.2	22.6	19.7
Average FPS	24.9	25	24.9
Single frame processing time (ms)	2.77	5.84	2.66
Client RAM (MB)	182.1	183	181
Server RAM (MB)	480	477	489
Server GPU VRAM (MB)	2117	2139	2090
Queue time (ms)	285	260	283

Table 6: Runtime-performance of a continuous one-hour audio stream for real-time speech-driven facial avatar animation in terms of 2D/3D landmarks and ARkit blendshapes.

The system achieved real-time animation at 25 FPS with minimal latency (maximum of 22.6 ms) and no packet loss for all three scenarios. Individual frame processing consistently met the 32 ms requirement, ensuring uninterrupted streaming. Differences in processing time and RTF between 3D landmarks and the other animation modes arose from transforming predicted PDM coefficients into 3D landmarks (478×3 coordinates) and transmitting them back. While 3D landmarks required more computational and network resources than 2D landmarks and blendshapes, they did not hinder continuous transmission. Throughout the process, memory consumption remained low and stable across client, server RAM, and GPU memory. Notably, only 1/6th of GPU capacity was utilized, indicating potential cost reduction in future deployments.

The results obtained offer a more comprehensive overview of the requirements and run-time performance of a real application. Related studies that reported run-time performance have focused solely on generating predictions for a limited number of pre-existing audio samples, basing their conclusions exclusively on the time the model takes to predict a single animation frame (Tian et al., 2019; Lu et al., 2021). These studies did not consider critical factors during live events, such as connectivity issues, audio queuing, and memory overflow, which can occur during extended live transmissions.

5 Conclusion

We introduced a novel approach to produce facial animations in real time, specifically designed for interactive live events and virtual spaces. Different configurations of facial representations were considered, including 2D and 3D landmarks, and ARkit blendshapes, the latter one being a standard in professional animation engines. The modeling and prediction of the facial representations was performed using different configurations of CNNs due to the low latency requirements of the addressed application. The quality of the considered methods was evaluated both objectively using metrics from the state-of-the-art, and subjectively, where naive and expert raters estimated the quality of the generated animations. Finally, the best performing models were used to create a client-server application able to produce facial animations in real time.

The results indicated that both the 1D-CNN and the SincNet models were accurate enough to predict the three types of considered facial animations. The results also confirmed that it is more challenging to generate emotional facial animations than neutral ones. Additionally, the models demonstrated greater accuracy in predicting landmarks and blendshapes associated with mouth and jaw movements, compared to other facial regions like the eyebrows. Finally, the conducted runtime evaluations offer a broader understanding of real-time

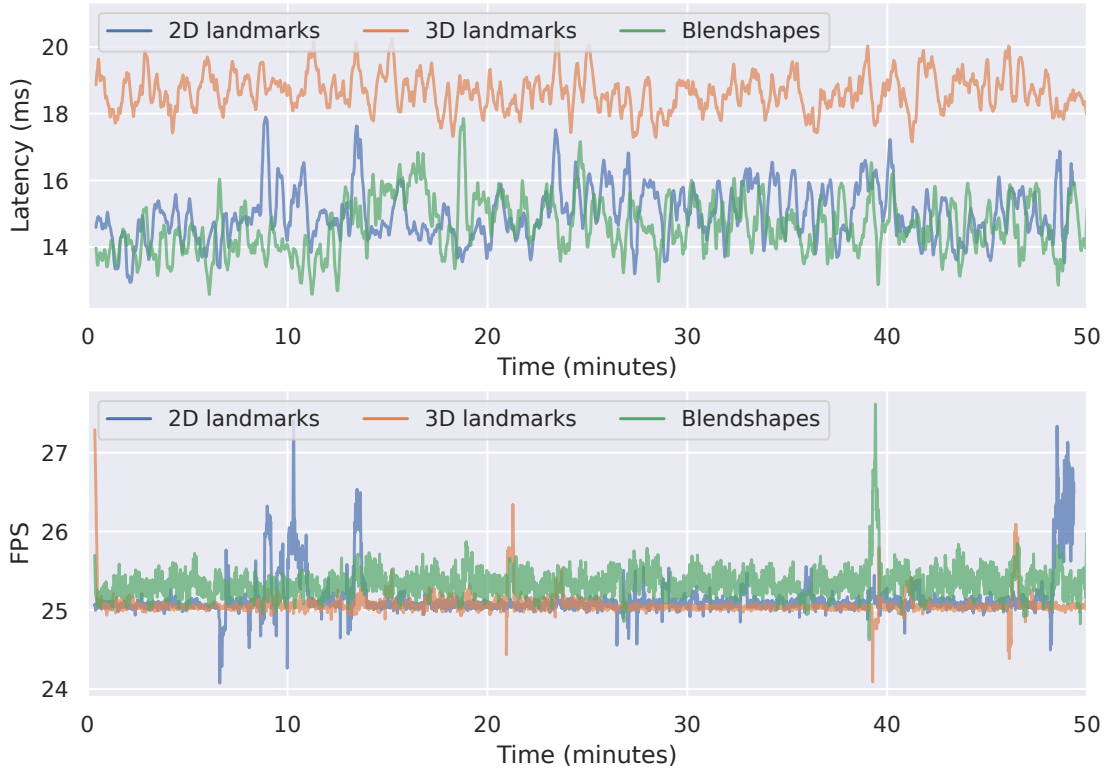

Figure 5: Latency and FPS of a continuous one hour audio streaming for the three different animation representation types.

application requirements and performance. Our real-time application showed that it is possible to generate facial landmarks and blendshapes in real-time at a constant rate of 25 FPS with a relative low latency and delay, and with low requirements of memory and GPU computation. Future work will be related to improve the quality of the generated animations in order to make them more natural and expressive. In this sense, novel architectures that also consider emotional classification can be proposed and evaluated. Exploring the integration of emotional intelligence into the system could be a promising direction for enhancing the expressiveness of the avatar animations.

Limitations

Despite the advancements and promising results from this paper, there are inherent limitations that should be considered: The first one is related to data availability. Although efforts were made to curate diverse datasets, the availability of comprehensive corpora covering a wide range of emotional expressions, linguistic diversity, and demographic variability might have been limited. This could potentially introduce issues in the generalization of the model to broader scenarios. Capturing the full spectrum of human emotions with high fidelity remains a challenge. Therefore, the current models may oversimplify the representation of certain emotional cues, leading to potential discrepancies between the intended and perceived expressions.

The second limitation relies on the latency and performance trade-offs when dealing with real-time applications. Achieving real-time responsiveness often requires optimizing for low latency, which may come at the expense of animation quality or computational resources. The study may have made certain compromises in this regard, and further optimizations could be explored to enhance the overall user experience.

Finally, while subjective evaluations provide valuable insights into the perceived quality of animations, they are inherently subjective and susceptible to biases. Factors such as individual preferences, cultural background, or expertise in animation could influence the raters' judgments. Employing diverse and representative rater groups, along with structured evaluation methodologies, can help mitigate bias to some extent but may not entirely eliminate it.

Ethics Statement

This study was conducted in accordance with the ACL Ethics Policy, ensuring that all research practices adhered to ethical standards in the development and evaluation of real-time speech-driven avatar animation.

In this study, all data used for training and evaluation were sourced from publicly available datasets, ensuring compliance with relevant data protection regulations. No personally identifiable information was used, and all data were anonymized to protect the privacy of individuals. Additionally, any data involving human participants was used in accordance with informed consent protocols.

While the technology developed in this study has positive applications, there is a potential for misuse, such as in the creation of deepfakes or unauthorized use of avatars. We emphasize the importance of deploying this technology responsibly, with safeguards to prevent misuse.

Finally, regarding transparency and accountability, we have provided detailed descriptions of our methodologies and evaluation metrics to ensure replicability and accountability. We encourage the research community to engage with and scrutinize our work to foster improvements and address any ethical concerns.

Acknowledgements

This study has received funding from the Basque Government Hazitek 2023 programme under grant agreement No ZL-2023/00497, Project ANIMEX.

References

A. Alsius, M. Paré, and K. G. Munhall. 2018. Forty years after hearing lips and seeing voices: The mcgurk effect revisited. *Multisensory Research*, 31(1-2):111–144.

H. Cao et al. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Y. Chen, J. Zhao, and W. Zhang. 2023. Expressive speech-driven facial animation with controllable emotions. *arXiv preprint arXiv:2301.02008*.

M. Cooke et al. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

T. F Cootes et al. 1995. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

P. Edwards et al. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)*, 35(4):1–11.

S. E. Eskimez et al. 2018. Generating talking face landmarks from speech. In *Latent Variable Analysis and Signal Separation*, pages 372–381. Springer.

S. E. Eskimez et al. 2019. Noise-resilient training method for face landmark generation from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:27–38.

I. Grishchenko et al. 2020. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*.

I. Grishchenko et al. 2022. Mediapipe blendshape v2 model card.

A. Hussen Abdelaziz et al. 2021. Audiovisual speech synthesis using tacotron2. In *Proc. International Conference on Multimodal Interaction*, pages 503–511.

Ege K. and Engin E. 2021. Investigating Contributions of Speech and Facial Landmarks for Talking Head Generation. In *Proc. Interspeech 2021*, pages 1624–1628.

Shuhei K. and Taiichi H. 2023. Speech-to-Face Conversion Using Denoising Diffusion Probabilistic Models. In *Proc. Interspeech*, pages 2188–2192.

T. Karras et al. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

T. Kefalas et al. 2020. Speech-driven facial animation using polynomial fusion of features. In *Proc. ICASSP*, pages 3487–3491. IEEE.

D. E. King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

Y. Lu et al. 2021. Live speech portraits: real-time photo-realistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17.

H. X. Pham et al. 2017. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88.

Xin R. et al. 2023. Emotion-Aware Audio-Driven Face Animation via Contrastive Feature Disentanglement. In *Proc. INTERSPEECH 2023*, pages 2728–2732.

M. Ravanelli and Y. Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *Proc. SLT*, pages 1021–1028. IEEE.

R. W. Schafer. 2011. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117.

S. Taylor et al. 2017. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)*, 36(4):1–11.

B. Thambiraja et al. 2023. 3diface: Diffusion-based speech-driven 3d facial animation and editing. *arXiv preprint arXiv:2312.00870*.

G. Tian et al. 2019. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *Proc. international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE.

A. Vidal and C. Busso. 2023. Multimodal attention for lip synthesis using conditional generative adversarial networks. *Speech Communication*, 153:102959.

A. Villanueva et al. 2022. Voice2face: Audio-driven facial and tongue rig animations with cvaes. In *Computer Graphics Forum*, volume 41, pages 255–265. Wiley Online Library.

K. Vougioukas et al. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413.

J. Xing et al. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790.

Ravindra Y. et al. 2020. Stochastic Talking Face Generation Using Latent Distribution Matching. In *Proc. Interspeech 2020*, pages 1311–1315.

G. Yan. 2022. Mediapipe facemesh model card.

K. D. Yang et al. 2023. Probabilistic speech-driven 3d facial motion synthesis: New benchmarks, methods, and applications. *arXiv preprint arXiv:2311.18168*.

B. Zhang et al. 2024. Emotalker: Emotionally editable talking face generation via diffusion model. *arXiv preprint arXiv:2401.08049*.

C. Zhang et al. 2021. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*.

C. Zhang et al. 2023. Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*.

Q. Zhao et al. 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. *arXiv preprint arXiv:2401.15687*.

Y. Zhou et al. 2020. MakeItTalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15.