# EAM: Emotional Avatar Generation for the Metaverse

Ander González-Docasal[1,2,*,†], Juan Camilo Vásquez-Correa[1,†], Aitor Álvarez[1], Aritz Lasarguren[3,†], Jone López[3,†] and Egoitz Rodríguez[3]

[1] *Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia – San Sebastián, Spain*
[2] *Aragon Institute for Engineering Research, University of Zaragoza, Zaragoza, Spain*
[3] *Baleuko S.L., Durango, Spain*

**Abstract**

This paper presents Project EAM, a framework for generating and managing avatars with emotionally expressive speech synthesis and automatic facial animation. The system supports two Iberian languages, Spanish and Basque, and synthesises speech in four distinct emotional tones. Its built-in Speech Synthesis system, trained with professionally recorded emotional voices, is complemented by an automatic forced aligner to generate viseme sequences. By incorporating real-time facial animation prediction, it enhances lip-sync accuracy and expands its applications to the scopes of augmented reality, video games, and film production. The system allows for a complete integration in the Unity 3-D engine, boosting the productivity of animators by reducing times on demanding tasks. In conclusion, Project EAM contributes to bilingual digital content creation and advances expressive avatar animation for the Metaverse and interactive media production.

**Keywords**

Voice cloning, Emotional Speech Synthesis, Speech to Text Alignment, Talking Face Generation

## 1. Introduction

The Metaverse represents a rapidly evolving digital ecosystem where users interact in immersive, interconnected virtual environments. Enabled by advancements in virtual reality (VR), augmented reality (AR) and artificial intelligence (AI), the Metaverse aims to reshape how people will socialise, work, and engage with digital content. As major industries invest in its development, the Metaverse is on the verge to become a fundamental extension of real-world experiences, fostering new economic models, communication methods, and creative expressions. Its relevance continues to grow as technological infrastructure improves, making it a critical domain for research and innovation in human-computer interaction and digital identity.

Nowadays, numerous options exist for developing avatars for the Metaverse, with many solutions available on the market. These options primarily focus on providing support for video games or social networking platforms, such as VRChat[1] [1]. However, due to the specific requirements of these environments, current solutions often fall into two extremes: either offering minimal personalisation – where users are limited to human-like avatars – or providing an overwhelming number of customisation options, which can result in inconsistent and incoherent virtual scenarios [2]. Furthermore, the animation of avatars in these digital spaces is typically constrained by a predefined set of motions or executed with a lack of precision, particularly evident in facial movements associated with speech and emotional expression, such as lip and eyebrow articulation [3].

*SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.*

[*] Corresponding author.

[†] These authors contributed equally.

agonzalezd@vicomtech.org (A. González-Docasal); jcvasquez@vicomtech.org (J. C. Vásquez-Correa); aalvarez@vicomtech.org (A. Álvarez); aritz@baleuko.eus (A. Lasarguren); jonelopez@baleuko.eus (J. López); egoitz@baleuko.eus (E. Rodríguez)

0000-0003-3585-6346 (A. González-Docasal); 0000-0003-4946-9232 (J. C. Vásquez-Correa); 0000-0002-7938-4486 (A. Álvarez); 0009-0006-7000-1916 (A. Lasarguren)

[1] https://hello.vrchat.com/

Therefore, when specifically considering the domain of animation, there are relatively few options available that effectively address the conveyance of emotional content – an essential factor in ensuring a minimum standard of quality, since facial and body movements must be as faithful and expressive as possible [4]. Additionally, many existing solutions rely on pre-recorded voice performances from actors to drive the corresponding animations. In this context, avatars that incorporate synthetic speech generation alongside synchronised lip movements present a valuable alternative, offering significant benefits for animators and studios [5]. Such technology can be employed to create proofs of concept or even short animated sequences before final in-studio recordings take place. Nevertheless, current applications providing these services often generate audio that lacks expressiveness, emotional nuance, or naturalness – essential components to be expressed in animation, particularly on content designed for a child audience. Moreover, the scarce tools that provide emotional speech synthesis often focus on major languages such as English or Mandarin, with no option for communities with a more modest number of speakers.

In this context, we introduce EAM: Emotional Avatar generation for the Metaverse, a project designed to facilitate the creation and management of avatars for both real-time interactions and pre-programmed events. The system is capable of generating emotionally expressive speech from a predefined script and automatically animating avatars based on textual or acoustic inputs. This solution implements speech technologies to the emerging field of the Metaverse, while also extending its applicability to well-established industries such as video game development and 3-D animation for films and television series. Furthermore, it is designed to support real-time scenarios, enabling avatars to produce highly accurate facial animations driven solely by an actor's voice, with minimal latency.

The solution is available in two Iberian languages, Spanish and Basque, thereby facilitating bilingual content creation and broadening audience reach, while simultaneously contributing to the preservation and promotion of regional minority languages. Moreover, the speech synthesis engine is capable of generating audio using two distinct speaker voices, each available in four different emotions, ensuring expressive and natural-sounding output in both supported languages.

This solution has been successfully employed to generate short commercial clips for social media engagement using predefined scripts. By automating the synchronisation of speech with mouth and lip movements, it significantly enhances the productivity of animators, reducing the time required for this otherwise labour-intensive task. These clips have been seamlessly produced in both Basque and Spanish, effectively reaching to a wider audience within a bilingual regional context.

The project EAM has been partially funded by the Spanish Government's call for year 2022 granting of aid for R&D projects in the audiovisual and video game fields, supported by European Union's Next Generation funds. The project started in September 2022 and finalised in December 2024, and was carried out by the consortium composed by the producer company Baleuko[2], the professional recording studio Sonora[3], and Vicomtech Foundation[4].

## 2. Project goals and challenges

The objective of EAM was to develop a system capable of managing avatars that autonomously generate emotionally expressive animations from textual or acoustic inputs. This encompassed the following key goals and challenges:

- **Avatars should automatically generate synthetic emotional speech from a script**. The system must incorporate a Speech Synthesis engine that leverages emotional embeddings to ensure accurate and expressive audio generation.
- **Two different characters should be available, both speaking Spanish and Basque, featuring four different emotions**. The system should include one or more models capable of

[2] https://baleuko.es
[3] https://www.sonoraestudios.es/
[4] https://www.vicomtech.org/

generating every possible combination of these characteristics, ensuring a seamless interpretation of any legible input string, particularly in terms of prosody and phoneme production.

- **The generated speech should be accompanied by synchronised facial animation**. The system produce a sequence of visemes (mouth shapes corresponding to each phoneme) that are precisely aligned with the output audio, including accurate start and end timestamps.
- **Live animation should be supported**. This system must generate real-time mouth and lip movements based on acoustic inputs, allowing for dynamic and responsive avatar animation.
- **The system should be integrated with Unity**[5]. This 3-D engine enables the development of a wide range of productions, including pre-rendered animations for films and series, as well as real-time applications such as AR experiences and video games.

## 3. Speech synthesis engine

The core component of the EAM solution is the Speech Synthesis engine. As previously stated, it must be capable of generating speech for two different characters, expressing four different emotions, in both Basque and Spanish.

The development of this system began with its design process. Given that the primary application of this project is the animation of children's cartoons, the selection of emotions was based on the characteristics typically found in such productions. To ensure clarity and expressiveness, two positive and two negative emotions were chosen, with each further distinguished by varying levels of energy. This classification not only enhances emotional depth but also prevents potential ambiguity between categories, ensuring precise and consistent vocal expression. The selected emotions and their corresponding features are shown in Table 1.

**Table 1**
Chosen emotions for the speakers available in the EAM platform.

| | Low energy | High energy |
|---|---|---|
| + | Neutral | Euphoria |
| − | Sadness | Fear |

### 3.1. Data generation

The next phase focused on generating the necessary data for training the Text-to-Speech (TTS) models. Initially, the contents of multiple scripts from the producer company Baleuko were extracted. These texts contained authentic examples from real shows in both Basque and Spanish, and therefore provided a valuable foundation for fine-tuning the models to better reflect real-world usage. To further expand the dataset and accommodate a greater volume of recordings, the corpus was supplemented with public domain texts from Common Voice [6].

Once the textual data was collected, a proprietary BERT-based [7] general-purpose sentiment detection module was employed to classify each sentence and map its results into one of the predefined emotional categories. As this emotion recognition tool was only available in Spanish, Basque texts were first translated into Spanish using Itzuli, a high-quality Neural Machine Translation (NMT) system accessible via API upon request[6] before undergoing emotional classification, an approach that has demonstrated its robustness for Basque-Spanish translation [8].

After categorisation, the script content was segmented into paragraphs to facilitate the reading task to the professional voice actors recording the database. Each paragraph was assigned a category based on the emotional classification of its constituent sentences, following heuristics designed to ensure

[5] https://unity.com/

[6] https://itzuli.vicomtech.org/api/

a balanced distribution across all classes. Finally, textual content belonging to Common Voice was designated a random emotion as these sentences are mostly neutral or informative.

The next step involved recording the textual corpus in a professional studio. Two characters from Baleuko-produced shows were selected as the primary voices for this project: Ezki (F) and Mithy (M). Their respective voice actors were provided with a pronunciation guide to minimise variability, particularly in the articulation of questions and foreign words. Moreover, in order to reduce studio costs, textual corrections were preferred over re-recording in cases where minor speech errors occurred. To optimise dataset balance, greater emphasis was placed on recordings for the *Neutral* emotion, as it was expected to be the most frequently used category. Conversely, the number of recordings for the *Fear* emotion was reduced, given that it was not anticipated to play a major role in the intended applications.

Once the recordings were completed, a forced-aligned model based on Kaldi [9] was employed to align phoneme-level timestamps for each language. Initially, session-level processing was conducted using a beam size of 40 and a retry beam of 70 to ensure correct alignment across all audio files. Following this, sentence-level alignment was performed with beam sizes of 1 and 2, discarding unaligned segments under the assumption that they were insufficiently literal. Nevertheless, filtering process resulted in the removal of only 0.67 % of sentences from the final corpus. The total amount of material comprising the final database can be found in Table 2.

**Table 2**
Total amount of audio (above) and sentences (below) comprising the processed database.

| | | Neutral | Euphoria | Sadness | Fear | Per language | Per speaker |
|---|---|---|---|---|---|---|---|
| Ezki | Spanish | 5:47:57 | 3:51:26 | 2:37:12 | 1:07:32 | 13:24:07 | 26:58:11 |
| | Basque | 6:31:52 | 2:47:49 | 3:09:30 | 1:04:53 | 13:34:04 | |
| Mithy | Spanish | 5:12:08 | 3:31:40 | 3:27:10 | 1:10:34 | 13:21:33 | 25:37:59 |
| | Basque | 4:11:18 | 3:36:04 | 3:24:09 | 1:04:56 | 12:16:27 | |
| Per emotion | | 21:43:15 | 13:46:58 | 12:38:02 | 4:27:56 | Total: | 52:36:10 |

| | | Neutral | Euphoria | Sadness | Fear | Per language | Per speaker |
|---|---|---|---|---|---|---|---|
| Ezki | Spanish | 6 723 | 4 718 | 2 690 | 1 164 | 15 295 | 28 081 |
| | Basque | 6 159 | 2 849 | 2 651 | 1 127 | 12 786 | |
| Mithy | Spanish | 5 234 | 4 341 | 3 270 | 1 156 | 14 001 | 26 955 |
| | Basque | 4 556 | 3 889 | 3 461 | 1 048 | 12 954 | |
| Per emotion | | 22 672 | 15 797 | 12 072 | 4 495 | Total: | 55 036 |

## 3.2. Speech synthesis models

The next phase involved training Speech Synthesis models with all the gathered data. Given the substantial amount of recorded acoustic material, the well-consolidated Tacotron 2 [10] was selected as the TTS framework. This architecture leverages an encoder-decoder with attention for generating Mel spectrograms given an input character sequence. To enhance its emotional expressiveness, the models were further improved by incorporating emotional embeddings, following an approach similar to that of [11]. While Cho et al. applied this technique in a multi-speaker setting, the approach taken in this project was a multi-emotional framework instead. Consequently, four separate models were trained, one for each combination of speaker and language, each incorporating an emotion-encoder.

Initially, a single model per speaker was trained, encompassing both languages and all four emotional classes, using a sampling rate of 22 050 Hz. These models were fine-tuned from a pre-existing model trained on a single Spanish speaker without prior knowledge of Basque or emotional nuances. Each model was trained for 100 k steps on a batch size of 6, starting with an initial learning rate of $10^{-3}$ exponentially decaying to $10^{-5}$ after the first 50 k steps. The emotional embeddings were represented as 16-dimensional vectors, trained alongside the network.

Following this, a separate model was trained for each language, using only audio samples with a maximum duration of 10 seconds. This restriction allowed for a higher batch size of 32. These models underwent an additional 50 k training steps with an initial learning rate of $3 \cdot 10^{-4}$.

Finally, a HiFi-GAN [12] vocoder was trained per speaker using the ground-truth aligned spectrograms generated by the final Tacotron 2 models, responsible for converting the outputs of the TTS module into audio waveforms. Each had an initial learning rate of $2 \cdot 10^{-4}$ and a batch-size of 16, maintaining the 22,050 Hz sampling rate for consistency with the synthesis models. The resulting four models, along their respective vocoders, were proven to successfully synthesise speech across both languages and all four different categories.

## 3.3. Inference pipeline

The final component of the Speech Synthesis engine corresponded to its inference pipeline. It consisted of the following components:

- **Text Preprocessor**: This module processes the input text, converting special characters such as numerical values into their fully readable form. It supports the two main languages from this project: Spanish and Basque.
- **Phonemiser**: It receives the normalised text from the Text Preprocessor and converts it into its phonetic representation. Given the relatively phonetic orthographies of the supported languages, this module operates using a rule-based approach.
- **TTS system**: This component loads the four trained models alongside their respective vocoders into RAM, prepared for generating an audio signal from any incoming textual input and its desired emotion.
- **Forced aligner**: Leveraging the aforementioned forced-align module based on Kaldi, this component aligns the phonetic transcription produced by the Phonemiser with the generated speech output from the TTS system. It determines precise phoneme timestamps and maps them to their corresponding visemes in the form of a JSON object.
- **Orchestrator**: This element serves as the central controller, managing the various pipelines for the multiple combination of languages and speakers. It ensures seamless integration of all components, handling user requests and processing inputs and outputs in a manner agnostic to the end user.

Each component is implemented as a RESTful API built in Go, facilitating modular communication via JSON objects. The whole system takes as input a text and a desired emotion, subsequently generating an audio file encoded as a base64 string or available for direct download. Additionally, a JSON object providing detailed timing information for each viseme is returned. A schematic drawing of the whole pipeline is shown in Figure 1.

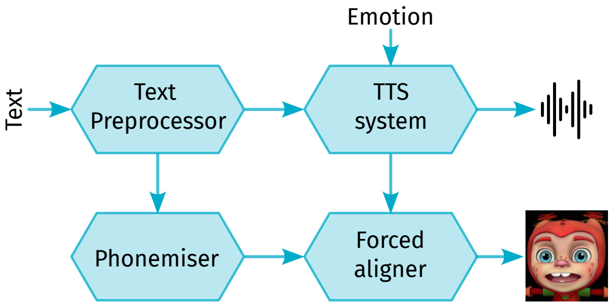
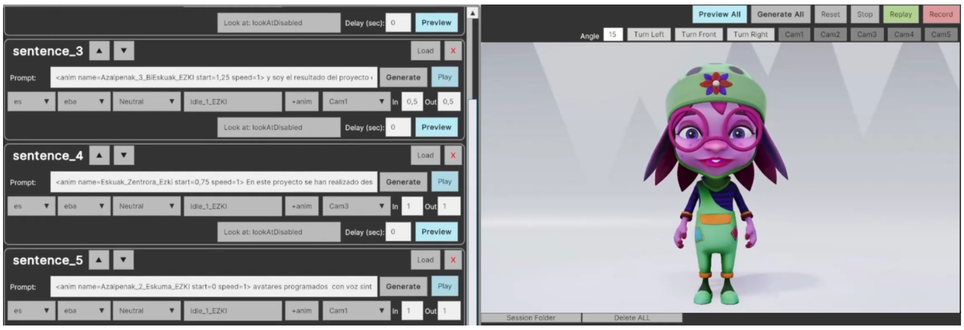


**Figure 1:** Schema of the modular Speech Synthesis engine with the integrated phonetic alignment.

This backend is then coupled with an automatic video sequence generation graphical user interface (GUI), which permits the creation of fully customisable animations using the Speech Synthesis engine. Using a configuration panel, the user is able to introduce text inputs along their corresponding emotional input and predefined body movements. The frontend incorporates fine grained camera control for constructing cinematic sequences, which are then exported to tracks and clips inside Unity's timeline. An image of the resulting GUI integrated in Unity is shown in Figure 2.



Look at: lookAtDisabled Delay (sec): 0

sentence_3

Prompt: <anim name=Azalpenak_3_BiEskuak_EZKI start=1,25 speed=1> y soy el resultado del proyecto

es eba Neutral Idle_1_EZKI Cam1 In 0,5 Out 0,5

Look at: lookAtDisabled Delay (sec): 0

sentence_4

Prompt: <anim name=Eskuak_Zentrora_Ezki start=0,75 speed=1> En este proyecto se han realizado des

es eba Neutral Idle_1_EZKI Cam3 In 1 Out 1

Look at: lookAtDisabled Delay (sec): 0

sentence_5

Prompt: <anim name=Azalpenak_2_Eskuma_EZKI start=0 speed=1> avatares programados con voz sint

es eba Neutral Idle_1_EZKI Cam1 In 1 Out 1

Angle 15

**Figure 2:** Sample image of the EAM solution integrated in Unity.

## 4. Speech-driven avatar animation

When generating facial animations from speech, it is important not only to ensure lip synchronisation, but also to transfer the emotions and intent from the audio to the avatar [13]. Humans are highly sensitive to facial expressions, making inconsistencies between speech and animation distracting and confusing [14]. Therefore, high-fidelity animation becomes essential to improve user experience in immersive environments.

Speech facial animation technologies fall into two broad categories based on complexity and expressiveness. Some engines leverage large-scale neural models for highly nuanced animations [15, 16]. However, these solutions demand significant computational resources, limiting their suitability for real-time processing during live events. On the contrary, simpler libraries based on viseme recognition [17] offer faster animation, but lack emotional expressiveness and intent transfer [18].

### 4.1. Facial representation

One of the core aspects of designing a speech-based facial animation engine is to find the most appropriate facial representation that can be predicted from speech signals and transferred to the avatar. Current systems are based on facial landmarks [18, 19] and 3D facial meshes [13, 20, 16], which are able to produce high fidelity and natural animations. However, they have limitations in computational efficiency and software compatibility.

For EAM, we considered a standard facial representation based on ARKit deformation blendshapes[7]. The main reason is the integration with animation engines like Unity, which already include some features for rendering these types of representations and transferring them to an avatar. ARKit's blendshapes consist of 52 coefficients that model the movement of different parts of the face, including the eyes, eyebrows, mouth, jaw, and nose.

#### 4.1.1. Training data

The facial animation model was trained using the CREMA-D corpus [21]. This is an emotional multimodal acted dataset, traditionally used for speech emotion recognition. Actors uttered a selection of 12 sentences in six emotions (Anger, Disgust, Fear, Happiness, Neutral, and Sadness) and three levels (Low, Medium, High), in English. This corpus has been used in similar studies, particularly in realistic talking face generation using facial landmark representations [22, 23]. Despite the language difference,

[7] ARkit blendshapes https://arkit-face-blendshapes.com/

as the principal objective of this model is to capture the main facial movements, it was supposed to be sufficient for this task due to the unavailability of existing data in the two target languages. Labels for blendshapes are obtained from the facial videos of the actors using the MediaPipe blendshape V2 model [24].

## 4.2. Facial animation model

The facial animation engine comprises a neural architecture trained to predict a continuous stream of blendshape coefficients from speech signals. The considered model relies on the combination of an upstream model, frozen during training and aimed to extract high quality representations of the input audio, and a downstream model trained to predict the 52 ARKit coefficients. A scheme of the trained model is observed in Figure 3.

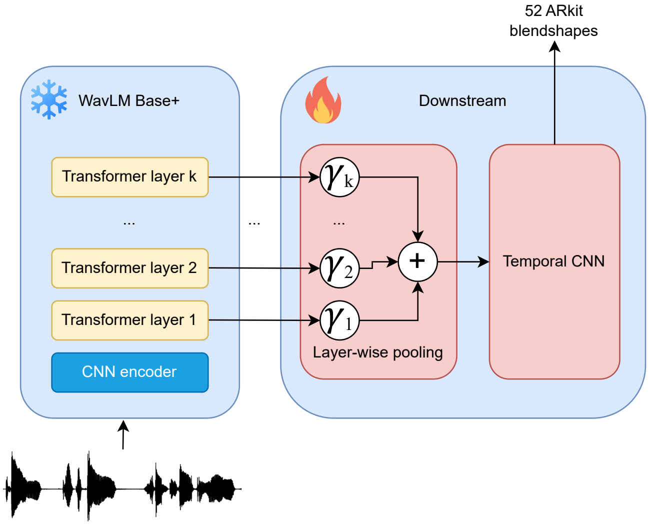


**Figure 3:** Neural architecture considered to predict ARkit blendshapes from speech signals.

The WavLM Base+ model [25] extracts high-level representations from the input audio while being light enough to be used in real-time scenarios. The hidden representations from each Transformer layer are combined with a layer-wise pooling mechanism in order to obtain a single feature embedding per frame. The layer's weights ($\gamma$) were implemented with a learnable vector followed by a Softmax layer. Finally, the sequence of feature embeddings is processed by a temporal convolution network to predict the 52 blendshapes per frame.

For training, 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 were trained using Adam optimiser, with a batch size of 32 audio samples, a learning rate of $10^{-5}$ and dropout of 0.1. The input audio corresponds to frames of 280 ms (7 frames at 25 fps), predicting the blendshapes for the central frame leveraging the rest as past and future context.

The quality of the predictions was evaluated using the mean absolute error (MAE) between the predicted and real blendshapes for different parts of the face. The average error in coefficient estimation is 13.3 %. Mouth, jaw, and cheek areas were observed to be more accurately animated than eyes or eyebrows, as expected considering our model being solely based on speech inputs. These results outperformed those obtained in previous studies [26], where a 1D-CNN model was considered for the same problem and using the same data. The quality of the predicted blendshapes were also perceptually evaluated when using synthetic speech generated with the previously described TTS models. In such cases, the avatar reflected accurate lip movements and emotion present in the input audio.

This model will continue to evolve once the project is finished, to ensure lower errors and smoother real-time interaction.

### 4.3. Real time processing

The facial animation model is integrated into a client-server application based on FFmpeg [27] and WebSocket for real-time audio processing. The client transmits continuous audio streams of 1024 bytes (32 ms of audio sampled at 16 kHz and 16-bit resolution) to the server, which receives and buffers the stream. Once 280 ms of audio is stored, the server predicts facial animations for the central frame and sends the results back to the client for rendering in Unity. After processing, the server releases the corresponding 40 ms audio segment and waits for new frames to arrive. An overview of the processing setup is shown in Figure 4.

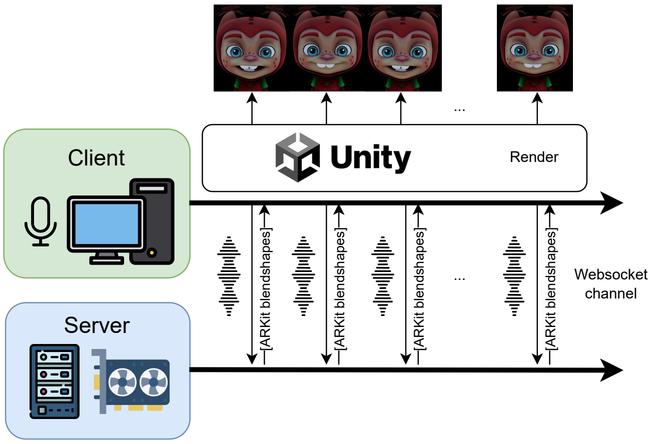


**Figure 4:** WebSocket communication for real time facial animation.

This engine is further enhanced with a scenario for real-time events that integrates multiple technologies for allowing the user a fine control of body animations of avatars. It includes the following components:

- **Rokoko motion capture**: to control body animations leveraging a motion capture suit.
- **ARKit face capture**: to further enhance facial animation of harder elements such as eyes and eyebrows.
- **Controlling of animations from a handheld device**: a video game controller is used to launch predefined animations to the avatar.

## 5. Conclusions

This work describes project EAM, a framework designed to facilitate the generation and management of two avatars, enabling speech synthesis in four distinct emotional tones while automatically animating corresponding mouth and lip movements. By automating and streamlining the traditionally labour-intensive process of vocal animation, EAM significantly enhances the productivity of animators and accelerates the creation of animated clips within the Unity 3-D engine.

Furthermore, the solution is available in two Iberian languages – Spanish and Basque – expanding the scope of the automatically generated content to a bilingual audience. Finally, the incorporation of real-time prediction of facial animations permits the application of this tool to many more scenarios, such as AR or video games.

As future work, research on further fine tuning of real-time blendshape detection algorithms will be performed. Additionally, efforts will be made to explore novel AI techniques in order to transfer the emotional capabilities of the TTS system to new unseen speakers or languages, reducing the costs of scaling the solution to new speakers.

## Acknowledgments

Project EAM was partially funded by the 2022 procedure for the granting of aid for business R&D projects applicable to the audiovisual and video game fields, and the transfer and adaptation of these developments to other application sectors, within the framework of the Plan to Boost the Audiovisual Sector Spain Audiovisual Hub of Europe; the "Digital Spain 2025" Agenda; the Recovery, Transformation and Resilience Plan; and the State Plan for Scientific, Technical and Innovation Research 2021–2023.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] H. Jehma, A. Akaraphattanawong, VRChat as a virtual learning platform for enhancing english listening skills, International Journal of Information and Education Technology 13 (2023) 813–817.

[2] S. Wu, et al., Factors affecting avatar customization behavior in virtual environments, Electronics 12 (2023) 2286.

[3] S. Noisri, et al., Designing avatar system and integrate to the metaverse, in: 2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), 2024, pp. 1–6.

[4] S.-M. Park, Y.-G. Kim, A metaverse: Taxonomy, components, applications, and open challenges, IEEE Access 10 (2022) 4209–4251.

[5] A. H. Abdelaziz, et al., Audiovisual speech synthesis using Tacotron2, in: Proc. International Conference on Multimodal Interaction, 2021, pp. 503–511. doi:10.1145/3462244.3479883.

[6] R. Ardila, et al., Common voice: A massively-multilingual speech corpus, in: Proceedings of LREC, European Language Resources Association, 2020, pp. 4218–4222. URL: https://aclanthology.org/2020.lrec-1.520/.

[7] J. Devlin, et al., BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. Conference of the North American Chapter of the ACL, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.

[8] T. Etchegoyhen, et al., Neural machine translation of Basque, in: Proceedings of the European Association for Machine Translation, 2018, pp. 159–168. URL: https://aclanthology.org/2018.eamt-main.14.

[9] D. Povey, et al., The Kaldi speech recognition toolkit, in: Proceedings of ASRU, IEEE Signal Processing Society, 2011. IEEE Catalog No.: CFP11SRW-USB.

[10] J. Shen, et al., Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions, in: Proceedings of ICASSP, 2018, pp. 4779–4783. doi:10.1109/ICASSP.2018.8461368.

[11] J. Cho, et al., Learning speaker embedding from text-to-speech, in: Interspeech 2020, 2020, pp. 3256–3260.

[12] J. Kong, et al., HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis, Advances in neural information processing systems 33 (2020) 17022–17033.

[13] Y. Chen, et al., Expressive speech-driven facial animation with controllable emotions, arXiv preprint arXiv:2301.02008 (2023). doi:10.1109/ICMEW59549.2023.00073.

[14] A. Alsius, et al., Forty years after hearing lips and seeing voices: The mcgurk effect revisited, Multisensory Research 31 (2018) 111–144. doi:https://doi.org/10.1163/22134808-00002565.

[15] K. D. Yang, et al., Probabilistic speech-driven 3D facial motion synthesis: New benchmarks, methods, and applications, arXiv preprint arXiv:2311.18168 (2023). doi:https://doi.org/10.48550/arXiv.2311.18168.

[16] Q. Zhao, et al., Media2face: Co-speech facial animation generation with multi-modality guidance, arXiv preprint arXiv:2401.15687 (2024). doi:https://doi.org/10.48550/arXiv.2401.15687.

[17] P. Edwards, et al., JALI: an animator-centric viseme model for expressive lip synchronization, ACM Transactions on Graphics (TOG) 35 (2016) 1–11. doi:10.1145/2897824.2925984.

[18] S. Taylor, et al., A deep learning approach for generalized speech animation, ACM Transactions On Graphics (TOG) 36 (2017) 1–11. doi:10.1145/3072959.3073699.

[19] A. Vidal, C. Busso, Multimodal attention for lip synthesis using conditional generative adversarial networks, Speech Communication 153 (2023) 102959. doi:https://doi.org/10.1016/j.specom.2023.102959.

[20] B. Thambiraja, et al., 3DiFACE: Diffusion-based speech-driven 3D facial animation and editing, arXiv preprint arXiv:2312.00870 (2023). doi:https://doi.org/10.48550/arXiv.2312.00870.

[21] H. Cao, et al., CREMA-D: Crowd-sourced emotional multimodal actors dataset, IEEE Transactions on Affective Computing 5 (2014) 377–390. doi:10.1109/TAFFC.2014.2336244.

[22] K. Vougioukas, et al., Realistic speech-driven facial animation with GANs, International Journal of Computer Vision 128 (2020) 1398–1413.

[23] T. Kefalas, et al., Speech-driven facial animation using polynomial fusion of features, in: Proc. ICASSP, IEEE, 2020, pp. 3487–3491. doi:10.1109/ICASSP40776.2020.9054469.

[24] I. Grishchenko, et al., Mediapipe blendshape v2 model card, 2022. URL: https://storage.googleapis.com/mediapipe-assets/Model%20Card%20Blendshape%20V2.pdf.

[25] S. Chen, et al., WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[26] J. C. Vásquez-Correa, et al., Real-time speech-driven avatar animation by predicting facial landmarks and deformation blendshapes, in: Proceedings of ICNLSP, 2024, pp. 109–118.

[27] S. Tomar, Converting video formats with FFmpeg, Linux Journal 2006 (2006) 10.