ISSN 1870-4069

Sistema de aprendizaje del movimiento de labios utilizando Q-Learning y redes neuronales convolucionales

Leonardo Nevárez Porras, Hernán de la Garza Gutiérrez, Carlos Humberto Rubio Rascón, Arturo Legarda Sáenz, Marisela Ivette Caldera Franco

[1] Tecnológico Nacional de México/Campus Chihuahua II, Chihuahua, México

{mm19550853, hernan.gg, carlos.rr, arturo.ls, marisela.cf}@chihuahua2.tecnm.mx

Resumen. Se describe un sistema de aprendizaje del movimiento de los labios a partir de un audio sin haberle dado tratamiento previo, de una persona hablando en español. Como salida se genera una animación 2D a 30 cuadros por segundo de un personaje que muestra el movimiento de los labios generado a partir de la implementación de la estrategia de Aprendizaje por Refuerzo, que incluye el uso de una red neuronal convolucional profunda, entrenada con el algoritmo de Q- Learning.

Palabras clave: Aprendizaje por refuerzo, Q-Learning, animación, audio, red neuronal convolucional, sincronización de labios.

Lips Movements Learning System with Q-Learning and Convolutional Neural Network

Abstract. We describe a learning system for lips' movements, which takes raw human speech audio in Spanish. The system creates a 2D animation at 30 frames per second of a character that shows the movements of the lips, generated by the implementation of the Reinforcement Learning strategy, that includes a convolutional neural network trained using the Q-learning algorithm.

Keywords: Reinforcement learning, Q-Learning, animation, audio, convolutional neural network, lip synchronization.

1. Introducción

En la actualidad, la generación de material digital es de gran importancia y utilidad, en diferentes ámbitos como son en la enseñanza a distancia y en la creación de contenido de entretenimiento, además la producción de video con personajes animados requiere de herramientas especializadas y de personas que sepan usarlas. Dependiendo

de la naturaleza de la animación y los movimientos de los personajes se puede requerir de una refinación de dichos movimientos para que se asemejen a los movimientos naturales de las personas de manera que se ajusten al contenido [1]. A partir de este tipo de retos, surgen sistemas de automatización de la animación, algunos de ellos apuntados a generar el movimiento de los labios, ya sea a partir del movimiento original del actor de voz [2] o mediante otras técnicas [1, 3].

El sistema que se describe a continuación, pertenece a este grupo y busca generar los movimientos de labios a partir de un audio en español utilizando técnicas de Aprendizaje por Refuerzo. Recientemente han emergido técnicas de aprendizaje de máquina que permiten entrenar redes neuronales profundas utilizando el algoritmo de propagación hacia atrás [4], las cuales pueden aprender a partir de datos sin procesamiento previo, tales como imágenes, video y audio [5].

El proyecto que se presenta, busca aprender a imitar el movimiento de labios a partir de un video, del cual se hace la separación de las imágenes y del audio. El audio se toma como entrada al módulo de Aprendizaje por Refuerzo para obtener una posición de los labios de salida, la cual se compara con las posiciones reales obtenidas de las imágenes del video y como resultado de dicha comparación generar la recompensa que promueva el aprendizaje.

Principalmente se utiliza una Red Neuronal Convolucional (RNC) la cual describiremos a fondo más adelante, para estimar una función que entrega las expectativas de recompensa de un agente al mover unos labios basado en la señal de audio. En el contexto del trabajo actual, nos referimos a un agente como la parte del sistema que percibe un estado del ambiente y toma acciones. En la figura 1, dentro del módulo de aprendizaje se pueden apreciar estos elementos, más el intérprete, que juntos forman parte del aprendizaje por refuerzo.

2. Trabajos relacionados

Se describen algunos de los trabajos publicados y que tienen relación con las tres principales áreas de nuestro proyecto: Sincronización de labios, Aprendizaje de Máquina y Sistemas de visión.

2.1. Sincronización de labios

Existen trabajos que apuntan a crear avatares digitales con sincronización de labios tomando varios enfoques:

El trabajo realizado en [1], genera una animación en 2D basándose en un audio en español, sin embargo, la animación está estilizada como caricatura, de manera que el muestreo es bajo, genera poca credibilidad y la sencillez de la animación generada busca no causar distracciones al usuario. En [3] se utiliza un enfoque similar para audio en inglés. Además de generar el avatar 2D en tiempo real (con pequeño retraso).

En el trabajo desarrollado por [2] se creó un sistema capaz de generar videos realistas de Barack Obama a partir de un audio del mismo Obama. El sistema se entrena primero con hasta 15 horas de video tomado de los reportes presidenciales semanales de Obama. Este sistema solo es capaz de generar contenido con la voz de un solo personaje y

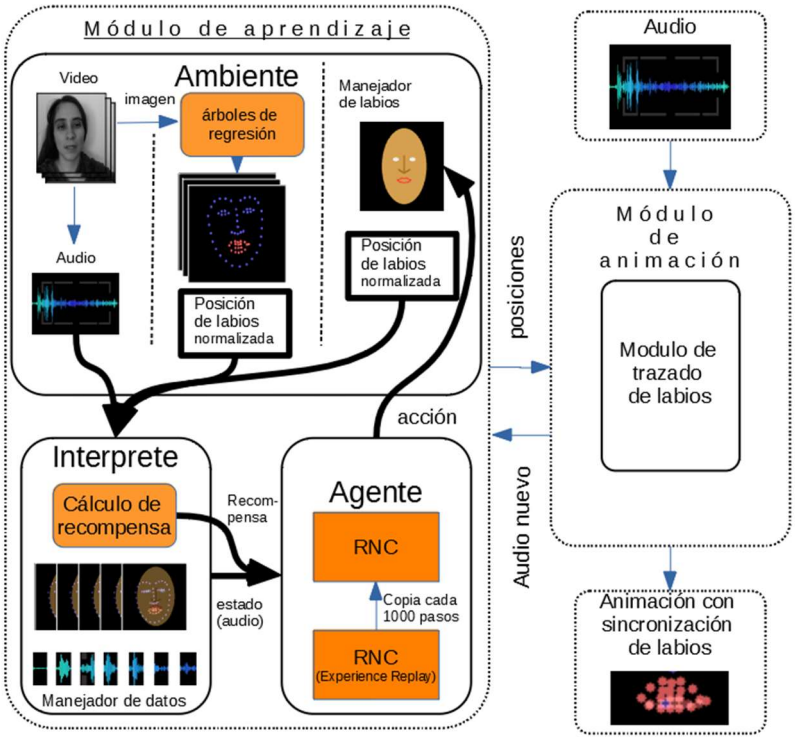

Fig. 1. Esquema general de las partes del sistema.

necesita alrededor de 15 horas de video de referencia además de procesamiento posterior adicional para producir resultados realistas.

Otros trabajos similares de síntesis de avatares a partir de un audio utilizando aprendizaje automático [6, 7, 8] hacen uso de material en otros idiomas para entrenar sus modelos, por lo que son aptos para generar material en otros idiomas distintos al español, además de utilizar técnicas de Aprendizaje Supervisado. Nuestra propuesta crea un avatar capaz de mover sus labios a partir de un audio en español utilizando técnicas de Aprendizaje por Refuerzo.

2.2. Aprendizaje de máquina

Avances recientes en Redes Neuronales Profundas permiten entrenar estas redes para procesar datos sin un tratamiento previo, tales como audio y video [4].

El algoritmo descrito en [2] para entrenar a un agente a jugar al Atari, permite entrenar una RNC que estima una función Q(s,a) que toma como entrada un estado *s* determinado por los pixeles de la pantalla actual y cuya salida representa la recompensa esperada al tomar la acción *a* dado el estado *s*. Al tomar la acción que maximice la recompensa y con suficiente exploración e iteraciones se logra obtener una aproximación lo suficientemente cercana a la función Q(s,a) real para superar retos en distintos juegos de Atari.

Para procesar audio existen enfoques como el de [9] y [10] donde se utiliza una RNC Profunda, es decir, de varias capas. Este tipo de arquitecturas permiten procesar audio en forma de onda directamente y crear a partir del entrenamiento, filtros que extraen de forma secuencial características del audio cada vez más abstractas.

Otros trabajos como [11] utilizan representaciones del audio en espectrogramas los cuales se alimentan a RNC similares a las utilizadas para procesar imágenes y video con convoluciones en 2D.

El sistema aquí presentado hace uso de estas técnicas para extraer características del audio sin ser procesado previamente, y generar animaciones a partir de éste.

2.3. Sistemas de visión

Para entrenar a un agente de Q-learning a que mueva los labios de manera correcta siguiendo un audio, se requiere de una señal de recompensa que es un número real, que funciona como una medida del rendimiento del agente al imitar el movimiento de los labios de manera similar al puntaje de un videojuego como en [5].

Como en otros trabajos de sincronización de labios [6, 7], el modelo se somete a un entrenamiento durante el cual se debe poder cuantificar la diferencia entre la posición de los labios del locutor original del audio y la posición de los labios generada por el agente a partir del mismo audio.

Por esta razón, se debe poder ubicar la posición de los labios tanto originales como los generados por el agente, ya sea ubicando puntos selectos en labios inferior y superior o de un trazado completo de estos. Existen diferentes métodos para ubicar los puntos de referencia en la cara.

Uno de ellos es el descrito en [12], el cual consta de un conjunto de árboles de regresión el cual se puede entrenar en una base de datos de caras etiquetadas con la ubicación de distintos puntos de referencia, incluyendo los labios.

3. Sistema de sincronización de labios

Después de un entrenamiento de preparación y teniendo un audio sin procesamiento previo de una persona hablando en español como entrada, el sistema de labios debe imitar el movimiento de labios del locutor y generar una animación de los labios. La estrategia mencionada se identifican las siguientes etapas:

1. Se captura el video de entrenamiento de una persona hablando de frente a la cámara, con toda la cara visible a 30 cuadros por segundo.
2. Para cada cuadro del video, se extraen las posiciones de puntos clave de los labios mediante árboles de regresión en forma de coordenadas de un plano cartesiano, y se acoplan con el segmento correspondiente de 300 milisegundos de audio.
3. Comenzando por el primer cuadro del video, uno a uno se alimentan las posiciones de labios y audio correspondiente. Con esto, el módulo de aprendizaje hace modificaciones internas a sus propios parámetros de forma que aprende a predecir las acciones que aproximan más cercanamente la posición de labios del locutor. La forma en que se lleva a cabo el aprendizaje se detalla en la sección 3.1.

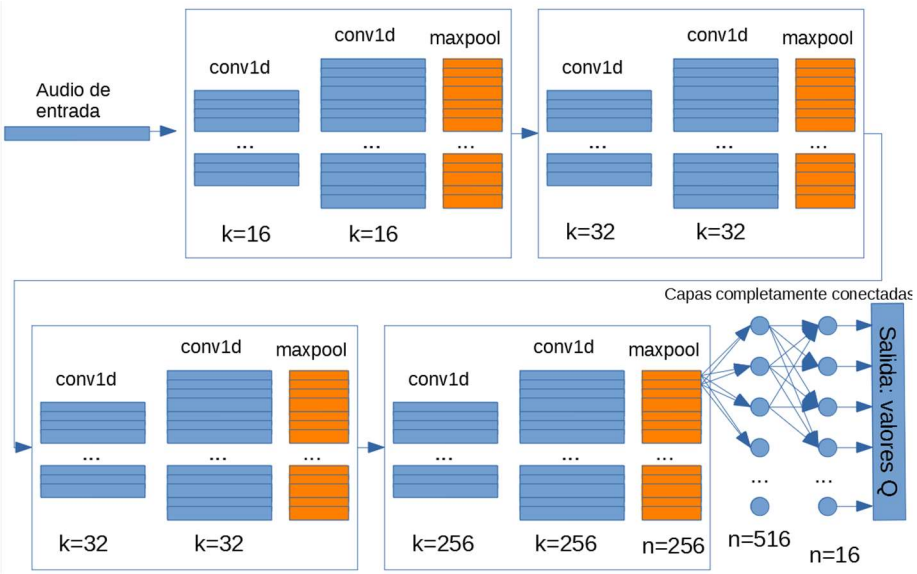

Fig. 2. Arquitectura de la Red Neuronal Convolucional.

4. Una vez llevado a cabo el aprendizaje, el módulo de animación hace uso de la experiencia del módulo de aprendizaje para generar una animación de labios con audio distinto al utilizado en el entrenamiento. La implementación del módulo de aprendizaje se explica a fondo en la sección 3.2.

El modelo general del sistema se puede ver en la Fig. 1. A continuación, se describen los módulos del sistema:

3.1 Módulo de aprendizaje

El Aprendizaje por Refuerzo se puede describir como un agente interactuando en un ambiente que cambia de estado a partir de las acciones tomadas por el agente y a su vez presenta una señal de recompensa la cual el agente busca maximizar [5].

Existen múltiples trabajos de sincronización de labios con personajes digitales basados en técnicas de aprendizaje supervisado tales como [2, 3, 6], sin embargo, buscamos explorar las técnicas del aprendizaje por refuerzo para posteriormente poder compararlas con los resultados obtenidos mediante aprendizaje supervisado.

El módulo de aprendizaje consta de los modelos y algoritmos necesarios para aprender el movimiento de labios después de un entrenamiento y se define en base a el marco de Aprendizaje por Refuerzo descrito en el párrafo anterior y también según [13] que consta de:

- Una política: Es una relación de estados a acciones que le dice al agente cual acción tomar en cada estado.
- Una señal de recompensa: La cual se busca maximizar para encontrar la política óptima.

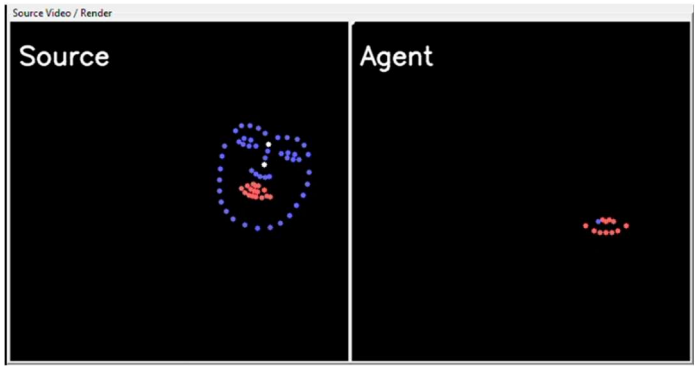

Fig. 3. Interfaz gráfica del módulo de animación.

– Una función de valor (de estado-acción): Que indica la recompensa que se espera recibir a partir de ese estado o estado-acción.

Al igual que en otros trabajos de Aprendizaje por Refuerzo, buscamos maximizar una señal de recompensa, la cual en este caso indica qué tan cercanamente los movimientos de los labios del agente, siguen a los del locutor original. Suponemos que las recompensas en el ambiente están dadas por la función Q(s,a) la cual podemos aproximar iterativamente mediante (1) como se describe en [13]:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[R_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)], \quad (1)$$

donde *s* es el estado actual del ambiente dado por la señal de sonido, *a* la acción tomada por el agente en el estado *s*, elegida entre 16 acciones diferentes. $Q(s_t,a_t)$, que representa la recompensa que el agente espera ganar a partir de un estado s_t tomando la acción a_t puede ser inicializada con un valor arbitrario para cada s_t,a_t.

El valor de α es el ritmo de aprendizaje y puede ser cualquier valor entre 0 y 1; R_{t+1} es la recompensa obtenida durante la transición de estado, γ es el valor de descuento que define la importancia que se le da a la recompensa futura mientras que (a)max Q (s_{t+1}, a) - $Q(s_t,a_t)$ es el valor máximo de recompensa a partir del estado siguiente tomando la acción *a*.

Esta actualización iterativa aproxima directamente q_* que es la función acción-valor óptima que maximiza la recompensa [13]. Sin embargo, ya que resulta impráctico el crear una función Q(s,a) debido a los recursos computacionales y el hecho de que buscamos poder generalizar, es práctica común en este tipo de problemas el utilizar funciones de aproximación estando entre ellas las Redes Neuronales de Convolución que dan muy buenos resultados para aproximar la función Q(s,a) como en [5].

Sin embargo, surgen dificultades a partir de usar una Red Neuronal para aproximar la función Q(s,a), principalmente el que la Red Neuronal tiende a no converger en la función óptima, al estar iterando sobre una política cambiante que depende en las salidas de la misma red.

Este problema se resuelve utilizando un mecanismo llamado *experience replay,* que toma muestras aleatorias de experiencias pasadas para entrenar la red, así como también limitando las actualizaciones a la Red Neuronal que representa a la función Q. Se utilizan dos redes, una para tomar las decisiones directamente y otra que se entrena en

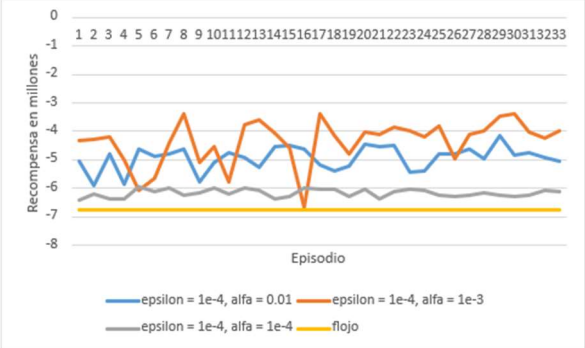

Fig. 4. Recompensa en millones por episodio, con ε = 1e-4 y diferentes valores de α, comparada con el agente flojo.

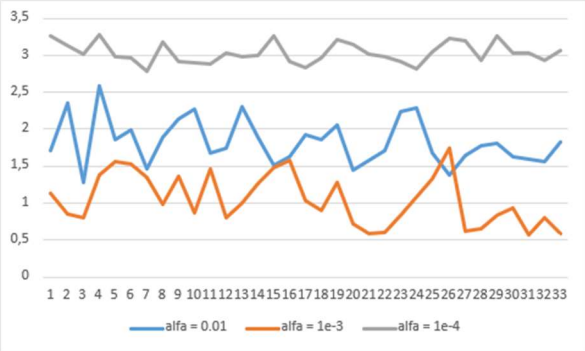

Fig. 5. Pérdida en millones por episodio, con ε = 1e-4 y diferentes valores de α.

cada transición de estado y cuyos cambios se copian a la primera red cada 1000 pasos [5], ilustrado en la Fig. 1., Agente Q-learning.

Como se menciona en [13], cabe destacar que el entrenamiento de la RNC que estima la función Q se puede ver como un problema de aprendizaje supervisado, ya que se cuenta con datos (el estado) y una etiqueta que es el valor de recompensa potencial hacia el cual queremos acercar las predicciones de la RNC.

La forma en que se implementa (1) es en el entrenamiento de la RNC, donde se calcula el valor de salida esperado, el cual se utiliza a su vez para obtener la pérdida con la cual entrenamos a la RNC.

Arquitectura de la Red Neuronal Convolucional. La RNC del sistema, ilustrada en la Fig. 2, está basada en la Red descrita en [14] y consta de convoluciones, agrupamientos y al final, capas completamente conectadas.

A cada dos capas de convolución le siguen una de agrupación tomando los mayores valores de cada segmento (*max-pooling*) y después se aplica una desactivación aleatoria (dropout) del 10% para prevenir sobreajuste (overfitting), todo esto, desde las convoluciones, agrupación y desactivación aleatoria se hace cuatro veces en cadena.

De esta forma tenemos dos capas de convolución con 16 filtros cada una, dos con 32 y dos capas más con 256 filtros, cada par seguido por agrupamiento y desactivación aleatoria.

La salida de las capas de convolución representa las características del audio extraídas a partir de los filtros de convolución y esta salida se alimenta a la segunda parte de la Red Neuronal la cual consiste en dos capas de neuronas completamente conectadas las cuales generan un valor numérico por cada acción disponible en el ambiente. El valor de cada salida representa el estimado de la función Q(s,a) donde *a* y *s* son los datos de entrada a la Red Neuronal.

Cálculo de la recompensa. La recompensa en el Aprendizaje por Refuerzo es un número real que le dice al agente qué tan bien se está desempeñando en el ambiente. Se busca que el agente siga los labios del locutor lo más cercanamente posible, por lo que durante el entrenamiento la recompensa debe ser una medida de qué tan cerca están siguiendo los labios controlados por el agente a los del locutor.

Los labios del locutor se representan como 4 pares de valores que ubican las coordenadas de 4 diferentes puntos de los labios en una imagen. Estos puntos son las comisuras izquierda y derecha, el centro del labio superior y centro del labio inferior. Los puntos de los labios del agente se representan de la misma manera.

La recompensa en cada paso se calcula con el siguiente algoritmo:

Entrada: *Posiciones de los labios del agente y el locutor.*

Para labios del agente y locutor, calcular:

distancia_labio_superior = centro_labio_superior_y – centro_boca_y

distancia _labio_inferior = centro_labio_inferior_y – centro_boca_y

distancia _comisuras = comisura_izq_x – centro_boca_x

a = agente.distancia_labio_superior - locutor.distancia_labio_superior

b = agente.distancia_labio_inferior - locutor.distancia_labio_inferior

c = agente.distancia_comisuras - locutor.distancia_ inferior

recompensa = - (a^2 + b^2 + c^2)

De esta forma, la recompensa es el negativo de la suma de los cuadrados de las distancias entre los puntos clave de los labios del locutor y el agente. Y sirve como una medida de qué tan cerca sigue a los labios del locutor conforme avanza el entrenamiento, especialmente al promediarlo para varios pasos. Se utiliza el cuadrado de las distancias para disminuir la penalización cuando las posiciones de la boca son ligeramente diferentes.

A través del entrenamiento el agente busca maximizar la recompensa, o en este caso, reducir la penalización, al aproximar la recompensa a cero.

Durante el entrenamiento, se alimenta la RNC con 300 milisegundos de audio, sin tratamiento previo, es decir en forma de onda y se busca que la RNC determine la acción que brinde un mayor valor esperado de recompensa. De esta forma se obtiene la política del agente.

Después de tomar la acción con mayor valor, se compara la recompensa real con el estimado generado por la red neuronal, la diferencia se conoce como error y se utiliza para cambiar ligeramente los pesos de la red e irla entrenando.

Uno de los principales problemas del Aprendizaje por Refuerzo está en encontrar un balance entre exploración (de las partes desconocidas) y explotación (del conocimiento) [13]. En este caso, la exploración está representada por ε (épsilon) cuyo valor indica el

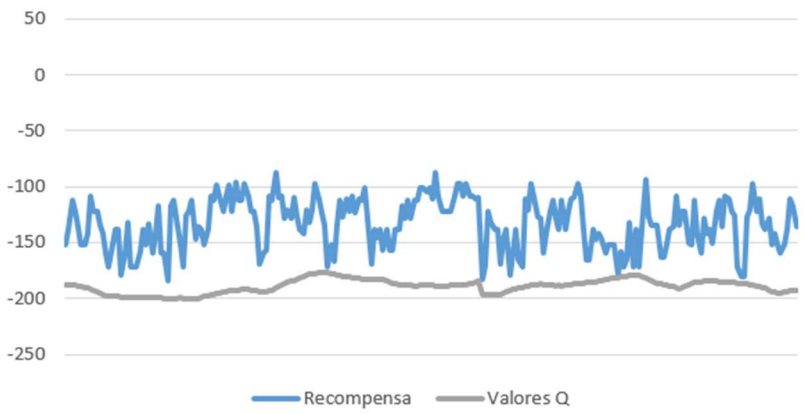

Fig. 6. Recompensa acumulada en millones por episodio, con ε = 1e-4 y diferentes valores de α.

grado de exploración deseado, es el porcentaje de acciones que se escogerán aleatoriamente del espacio de acciones sin importar que exista una acción conocida que entregue una recompensa máxima.

El sistema utiliza un valor de ε que disminuye conforme avanza el entrenamiento y también se experimentó con distintos valores de ε constantes. En la figura 3 podemos observar la recompensa promedio por episodio para 33 episodios de entrenamiento con diferentes valores de α (ritmo de aprendizaje) y un valor de ε fijo. Mientras que en la figura 4 se observa la pérdida correspondiente, la cual se busca disminuir con el entrenamiento.

3.2 Módulo de animación

Este módulo genera una animación de labios siguiendo un audio como entrada, a esto también se le conoce como sincronización de labios. Las entradas de este módulo son:

- Los pesos de la RNC, previamente entrenada.
- Un archivo de audio en formato WAV a 16khz con monólogo en español

La salida del módulo es una secuencia de imágenes (30 imágenes por cada segundo de audio de entrada) que representan el movimiento de los labios generado a partir de las acciones del agente.

A continuación, se describe la forma en que se generan las imágenes:

1. Inicializar el Manejador de Labios que funciona como una interfaz para que el agente manipule la posición de unos labios virtuales, la posición inicial de los labios es cerrada.
2. Cargar el audio en memoria. Inicializar Índice Audio para apuntar al elemento del audio en la posición 16000/30, redondeando hacia abajo.

3. Tomar el segmento de audio con inicio en [Índice Audio - 16000*0.30] y final en Índice Audio y alimentar la RNC con este segmento como entrada, para obtener la acción que maximiza la recompensa.
4. Tomar la acción que maximiza la recompensa, generar una imagen con la nueva posición de labios.
5. Índice Audio <= Índice Audio + 16000*0.30.
6. Ir a paso 3 hasta procesar todo el audio.

Al combinar las imágenes generadas en los pasos descritos anteriormente con el audio de entrada, podemos producir un video con sincronización de labios. Se puede observar la interfaz de usuario en la Fig. 3.

3.3 Experimentos y resultados

Los experimentos realizados consisten en el entrenamiento del modelo con un video de 32 minutos de una persona leyendo un texto de frente a la cámara. El entrenamiento consistió en 33 episodios o 1 millón de pasos.

Se parte de dos premisas relevantes: se alimentará la RNC con el audio sin procesar, buscando que la RNC sea capaz de extraer la información importante para el aprendizaje y la segunda premisa es que no se le proporcionarán las posiciones de los labios del agente, viendo si también la RNC es capaz de determinar la acción correspondiente sin esa información.

Como punto de comparación se toma la recompensa obtenida por un agente sin movimiento que mantiene una posición de boca cerrada, a este agente lo denominamos agente flojo. Con el agente flojo vamos a obtener un nivel base de recompensas cuando no se hacen acciones y permitirá hacer comparaciones con los modelos que sí realicen acciones. En la Fig. 4 podemos comparar la recompensa a través de 33 episodios con diferentes valores de α, el agente busca maximizar la recompensa a través del aprendizaje, explorando diferentes políticas.

Otra forma de medir el aprendizaje de la RNC es por la pérdida promedio durante un número de pasos. La pérdida se puede describir como una medida del error en las predicciones de la RNC por lo que buscamos que disminuya con el entrenamiento. En la Fig. 5 podemos observar la pérdida por episodio del sistema para diferentes valores de α. Una perdida que disminuye a través de los episodios indica que la red va mejorando en sus predicciones.

Al visualizar la recompensa para los últimos pasos del primer episodio, en la Fig. 6, podemos observar que a pesar de que el agente es capaz de estimar los cambios más burdos en la recompensa por episodio, no puede discriminar los cambios más pequeños, lo cual se ve reflejado en las acciones de movimiento que toma, es decir, al observar la animación de labios generada se observa que el agente da preferencia a los movimientos del labio inferior el cual tiene un mayor rango de movimiento y por ende mayor influencia en la recompensa resultante. Esto podría ser corregido utilizando una fracción de la distancia de los labios inferiores en lugar de la distancia total.

4. Conclusiones y trabajo a futuro

A partir de los experimentos realizados se obtiene un movimiento de labios que se asemejan a los del locutor original, demostrando con ello que funciona el Aprendizaje por Refuerzo. A partir de la recompensa total por episodio de varios agentes y comparándola con los resultados de un agente flojo, podemos ver que el sistema de Aprendizaje por Refuerzo, obtiene mayor recompensa con α = 1e-3 y ε = 1e-4. Se observa que la animación sigue de forma general los labios del locutor.

Se planea comparar el rendimiento del sistema con el de otros trabajos que utilizan distintas técnicas y plataformas, por el momento la métrica de comparación es la recompensa promedio por episodio, lo cual permite comparar el aprendizaje de distintos modelos que utilizan el mismo entorno. Otro plan es el de probar la capacidad de generalización de la técnica de aprendizaje por refuerzo aquí aplicada a la sincronización de labios en comparación con otras técnicas como las de aprendizaje supervisado.

Para este trabajo se limitó el entrenamiento a 33 episodios, pero se puede dar tanto tiempo como el agente requiera para quedar completamente entrenado. Una variante probada limitadamente hasta el momento es la duración del audio que recibe el agente en cada paso, para este trabajo fue de 300 milisegundos. Se experimentará con ventanas de audio con diferente duración buscando mejorar los resultados.

Otra forma de incrementar el aprendizaje del agente puede lograrse si a la salida de la última serie de capas de convolución y de agrupamiento de la Red, se le suman las posiciones de los labios del agente, y este nuevo conjunto de datos sea el que se alimente a las capas finales de la red. Esto tendría el potencial de acelerar el entrenamiento, ya que la recompensa depende en parte de la posición de labios del agente.

Adicionalmente, se sugiere crear un modelo similar que reciba el audio en forma de espectrograma, esto daría la facilidad de poder usar modelos adicionales probados en aplicaciones similares tal como en [11].

Referencias

1. Cabiedes F., Pelczer, I., Gamboa F., Bretón J., Rodríguez S.: Sincronización de labios: Método sin visemas. Revista Iberoamericana de Educación a Distancia, vol. 10, no. 1, pp. 37–50 (2007) doi: 10.5944/ried.1.10.1012
2. Suwajanakorn S., Seitz, S. M., Kemelmacher-Shlizerman, I.: Synthesizing Obama: Learning lip sync from audio. ACM Transactions on Graphics, vol. 36, no. 4, pp. 1–13 (2017) doi: 10.1145/3072959.3073640
3. Aneja, D., Li. W.: Real-time lip sync for live 2D animation. ArXiv 1910.08685v1 (2019) doi: 10.48550/arXiv.1910.08685
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature no. 521, pp. 436–444 (2015) doi: 10.1038/nature14539
5. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature 518, pp. 529–533 (2015) doi: 10.10 38/nature14236

6. Fan, B., Xie, L., Yang, S., Wang, L., Soong, F. K.: A deep bidirectional LSTM approach for video-realistic talking head. Multimedia Tools and Applications archive, vol. 75, no. 9, pp. 5287–5309 (2015) doi: 10.1007/s11042-015-2944-3
7. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics, vol. 36, no. 4, pp. 1–12 (2017) doi: 10.1145/3072959.3073658
8. Xu, Y., Feng, A. W., Marsella, S., Shapiro, A.: A practical and configurable lip sync method for games. MIG '13 In: Proceedings of Motion on Games, pp 131–140 (2013) doi: 10.1145/2522628.2522904
9. Dai, W., Dai, C., Qu, S., Li, J., Das, S.: Very deep convolutional neural networks for raw waveforms. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 421–425 (2017) doi: 10.1109/ICASSP.2017.7952190
10. Aytar, Y., Vondrick, C., Torralba, A.: SoundNet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems 29 NIPS (2016)
11. Wyse, L.: Audio spectrogram representations for processing with convolutional neural networks. In: Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN, arXiv:1706.09559 (2017) doi: 10.48550/arXiv.1706.09559
12. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
13. Sutton, R. S., Barto, A. G.: Reinforcement learning: An introduction. Francis Bach (2018)
14. Mansar, Y.: Audio classification: A convolutional neural network approach. https://medium.com (2018)