Universidad del Valle de Guatemala
Facultad de Ingeniería

Proyecto 1
Retrieval Voice Conversion
Deep Learning

JUAN ANGEL CARRERA SOTO 20593
JOSE MARIANO REYES HERNANDEZ 20074
JUAN CARLOS BAJAN CASTRO 20109

Introducción

Los cambios de voces en series y videojuegos plantean desafíos significativos para la inmersión del público. Estos cambios, ya sean causados por decisiones de casting, conflictos contractuales o problemas técnicos, a menudo resultan en una disonancia entre la voz original icónica y la nueva interpretación, afectando la experiencia de los espectadores y jugadores.

La conexión emocional que los consumidores establecen con las voces de personajes icónicos se ve amenazada, ya que una alteración en la voz puede romper ese vínculo. Además, la falta de continuidad en las voces puede afectar la inmersión en la trama y generar percepciones negativas sobre la calidad de la producción. En este contexto, la tecnología de conversión de voz basada en recuperación ofrece una solución innovadora para preservar las voces icónicas y mantener la autenticidad de los personajes a lo largo del tiempo.

Descripción del problema

El cambio de voces en series y juegos puede manifestarse de diversas maneras. Puede deberse a decisiones de casting, conflictos contractuales, o incluso a problemas técnicos que obligan a sustituir a los actores de voz originales. En cualquiera de estos casos, los consumidores a menudo se enfrentan a la disonancia entre la voz original que se ha vuelto icónica para un personaje y la nueva interpretación, lo que puede perturbar su inmersión en la historia y afectar negativamente su percepción general de la producción.

Análisis

El cambio repentino de la voz de un personaje icónico puede resultar muy disruptivo para la experiencia del usuario. La voz se vuelve parte integral de la identidad y personalidad de un personaje, por lo que el cambio puede sentirse como una traición a esa esencia.

Los consumidores forman un vínculo emotivo con las voces que han llegado a asociar fuertemente con ciertos personajes. Cuando esa voz cambia, se rompe parte de ese vínculo y conexión.

Existe el riesgo de que la nueva voz no logre capturar apropiadamente la esencia del personaje, debido a diferencias de tono, inflexiones, acentos, etc. Esto puede hacer que el personaje ya no se sienta auténtico.

Los cambios repentinos de voces pueden deberse a razones como disputas contractuales o recortes presupuestarios. Esto genera una percepción negativa en los consumidores en cuanto al compromiso de calidad de los productores.

La falta de continuidad en las voces puede afectar la capacidad de los consumidores para sumergirse en la trama y mundo narrativo. La disonancia actúa como una distracción que disminuye la inmersión. El impacto negativo en la experiencia del usuario puede derivar en críticas, quejas y menor engagement con la serie o videojuego. Incluso podría afectar las ventas y audiencias.

Propuesta de solución

La tecnología de conversión de voz basada en recuperación (Retrieval Voice Conversion) permite imitar de manera realista y natural la voz de una persona objetivo a partir de grabaciones existentes.

Esto se logra entrenando un modelo de deep learning en un conjunto de muestras de audio de la voz que se quiere imitar. El modelo aprende las características distintivas de tono, timbre, inflexiones, acento, etc.

Luego, durante la inferencia, el modelo recibe como entrada nuevo audio spoken y genera como salida ese mismo audio pero con las características de voz de la persona objetivo.

Esta tecnología podría aplicarse para permitir digitalizar y preservar las voces icónicas de actores en sus roles como personajes famosos. Por ejemplo:

- Entrenar un modelo con horas de grabaciones de la voz original de un actor interpretando a cierto personaje.
- Alimentar al modelo con audio spoken de un imitador o actor sustituto.
- Como resultado, obtener audio que suena como la voz original del actor, minimizando la disonancia para los usuarios.
- Esta voz sintetizada podría usarse en nuevas películas, videojuegos, series animadas, juguetes, etc.

De esta manera, se mantendría la autenticidad y consistencia de las voces emblemáticas asociadas fuertemente a personajes icónicos.

Descripción de la solución

La solución propuesta es la tecnología de conversión de voz basada en recuperación, también conocida como Retrieval Voice Conversion. Esta innovadora tecnología está diseñada para abordar el problema de los cambios de voces en series y videojuegos, permitiendo mantener la autenticidad y la inmersión de los usuarios al preservar las voces icónicas de actores en sus roles como personajes famosos. Algunos puntos interesantes de la solución son los siguientes:

1. Entrenamiento de modelos de aprendizaje profundo: La base de esta tecnología radica en la formación de modelos de aprendizaje profundo, los cuales son alimentados con extensos conjuntos de muestras de audio de la voz original que se busca imitar. Estas grabaciones pueden provenir de actuaciones pasadas de un actor de voz o cualquier otro recurso que contenga la voz característica del personaje.
2. Captura de características distintivas: Durante el proceso de entrenamiento, los modelos aprenden las características distintivas de la voz, como el tono, el timbre, las inflexiones, el acento y otros rasgos vocales que hacen que la voz sea única. Esto permite que el modelo comprenda y reproduzca de manera precisa la esencia vocal de la persona objetivo.
3. Generación de voz sintetizada: Cuando se desea aplicar esta tecnología, se utiliza el modelo previamente entrenado para convertir la voz de un imitador o actor sustituto en la voz del personaje original. El modelo toma como entrada el nuevo audio hablado y genera como

salida un audio que suena exactamente como la voz de la persona objetivo, manteniendo su autenticidad y característica distintiva.
4. Aplicaciones variadas: Esta solución tiene un amplio espectro de aplicaciones. Puede utilizarse en la producción de nuevas películas, videojuegos, series animadas, programas de televisión, juguetes y otros medios en los que sea esencial mantener la coherencia vocal de un personaje icónico.

Ventajas de la Solución:

1. La tecnología de conversión de voz basada en recuperación presenta varias ventajas significativas:
2. Conservación de la autenticidad: Permite mantener la autenticidad de las voces icónicas de personajes, lo que es esencial para la percepción y la conexión emocional de los espectadores y jugadores.
3. Continuidad en la narrativa: Al minimizar la discordancia vocal, garantiza que los consumidores puedan sumergirse de manera más efectiva en la trama y el mundo narrativo sin distracciones.
4. Flexibilidad en el casting: Facilita la sustitución de actores de voz sin perder la esencia del personaje, lo que puede resultar útil en casos de conflictos contractuales o problemas técnicos.
5. Potencial para aumentar la audiencia: Al mantener la voz original, se reduce el riesgo de críticas y quejas por cambios de voz, lo que puede contribuir a mantener y aumentar el interés de la audiencia.

Herramientas aplicadas

Google Collaboratory, proporcionó el entorno de desarrollo integrado basado en la nube para la ejecución del código Python. Esta plataforma fue escogida por su acceso sin coste a hardware acelerado por GPU, lo cual fue esencial para el entrenamiento de modelos de aprendizaje profundo. Aunque de todas formas se realizaron algunos gastos para aumentar la potencia.

TensorBoard se utilizó para visualizar las métricas de entrenamiento de los modelos. Esta herramienta permite monitorear indicadores clave como la pérdida de entrenamiento, facilitando la detección temprana y prevención del sobreentrenamiento.

Se accedió a GitHub para la recuperación de código y archivos asociados al proyecto. La funcionalidad de clonación de repositorios de GitHub se utilizó para integrar eficientemente los recursos necesarios en el entorno de Colab.

Se utilizaron varias librerías de Python, incluyendo os y threading para la gestión de archivos y la ejecución de procesos en paralelo, y tarfile para la extracción de archivos comprimidos esenciales para el proyecto. Además, se descargaron librerías como gTTS y elevenlabs que apoyaron la funcionalidad de conversión de texto a voz en la interfaz de usuario.

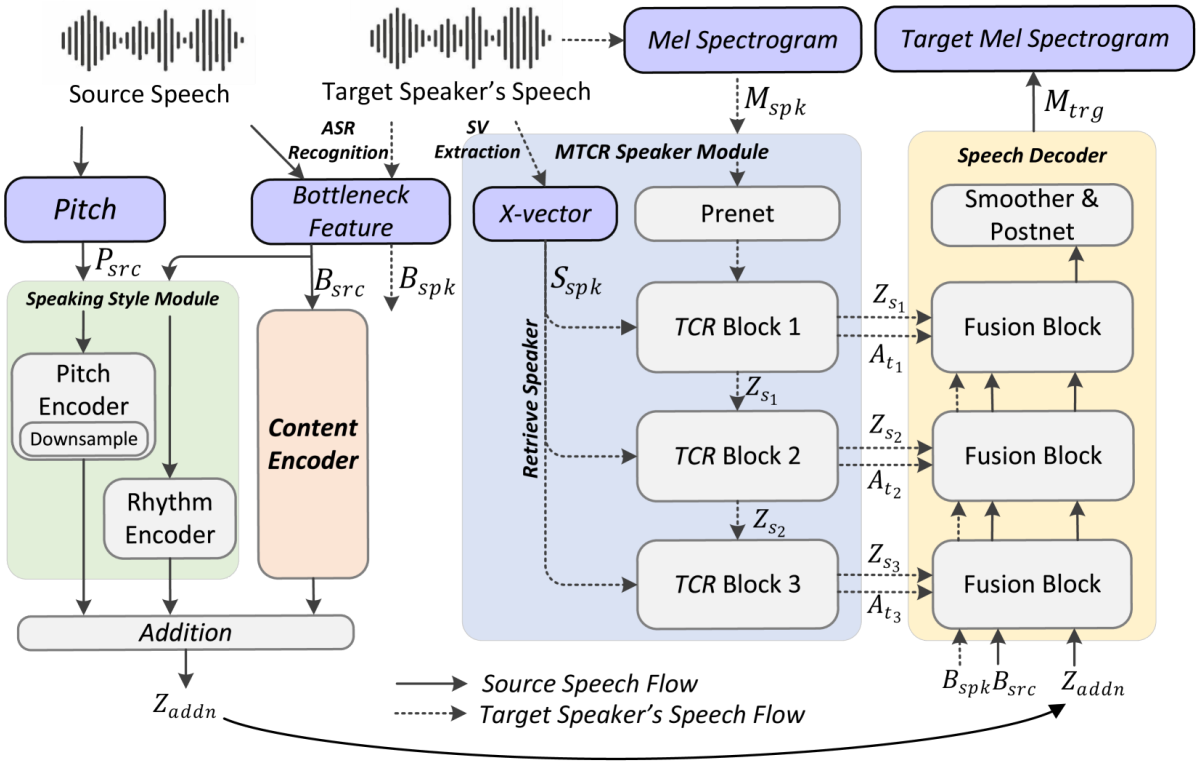
Modelo RVC
Es un modelo de conversión de voz basado en recuperación el cual está dividido en los siguientes pasos:

Primero se cargan y pre procesan los archivos de audio, incluyendo la normalización y la decisión del audio basada en silencios y segmentos de tiempos fijos.

El segundo paso es la extracción de las características. Estas van desde la extracción del tono que se convierte logarítmicamente a un int de entre 1 y 255, y la extracción de la huella acústica, en cuyo caso se utiliza HuBERT para convertir los audios .wav en características de 256 dimensiones y guardados en formato npy.

El último paso es la aplicación del modelo base de VC que utiliza pesos preentrenados, lo que permite el entrenamiento con conjuntos de datos pequeños. Además, durante la inferencia, RVC busca valores de características similares a los utilizados durante el entrenamiento para realizar la inferencia, y para ello utiliza el aprendizaje de índices con la biblioteca de búsqueda de vecindarios aproximados faiss.

Nuestra aplicación se enfoca en el uso del modelo de VC en one-to-many, en donde entrenaremos las características del modelo de VC, ya que no buscamos el reconocimiento de voz, si no que buscamos la forma de transformar una voz base a cualquier tipo de dialogo, canción o frase.



Modelo wav2lip

Se obtuvieron y utilizaron los archivos del modelo wav2lip, conocido por su capacidad para sincronizar los movimientos labiales con el audio en la conversión de voz.

Sistema de Respaldo Automático

El sistema de respaldo automático aprovechó las API de Google Drive para salvaguardar el progreso del proyecto. Esta función fue valiosa para asegurar que los datos no se perdieran y que se pudiera continuar el trabajo sin interrupciones.

Estas herramientas proporcionan la capacidad necesaria para realizar tareas de conversión de voz, facilitando así la consecución de los objetivos del proyecto. El uso de estas tecnologías demostró ser eficaz en la replicación de la voz del estudiante, logrando resultados satisfactorios en la superposición de su voz en diferentes pistas de audio.

Resultados

Voz de Rick Sanchez:

Una de los primeros experimentos que se realizaron fue con la voz del personaje de la serie Rick & Morty. Esto debido a su cambio reciente de actor de voz original. Creamos una primera iteración del modelo con 2000 épocas de entrenamiento. Se lograron resultados bastante aceptables para la voz pero para simular la voz original de Rick siendo que solo se usaron 8 audios para entrenarlo. Para este modelo se utilizó el algoritmo de rmvpe de HuBert para la recolección de pitch del audio original.

Link modelo:
https://drive.google.com/drive/folders/108dzulgsaaV8sJvE8U5k5SFEDOiGSAi9?usp=sharing

Voz de Juan Carlos:

A lo largo del proyecto, se llevaron a cabo experimentos con el modelo de aprendizaje profundo para evaluar su capacidad de replicar la voz del estudiante en diferentes condiciones de entrenamiento. Se configuraron y entrenaron tres modelos distintos, cada uno con variaciones en el número de épocas de entrenamiento y la cantidad de datos de audio suministrados.

El primer modelo, denominado 'creep', fue entrenado utilizando un conjunto de 12 archivos de audio con una duración promedio de 30 segundos cada uno. Después de 100 épocas de entrenamiento, este modelo logró los resultados más convincentes, generando audios que presentaban una similitud notable con la voz real del estudiante.

Link modelo 1:
https://drive.google.com/drive/folders/1hV7MdOy6nc2RwG-81BCJ45YA95MLbbBR?usp=sharing

Un segundo modelo se sometió a un entrenamiento más prolongado de 110 épocas, manteniendo la misma cantidad de datos de audio. Los resultados obtenidos fueron ligeramente inferiores en comparación con el primer modelo, indicando que un mayor número de épocas no mejoró la calidad de la replicación de la voz.

Link modelo 2:
https://drive.google.com/drive/folders/1hV7MdOy6nc2RwG-81BCJ45YA95MLbbBR?usp=sharing

Por último, se experimentó con un tercer modelo que duplicó la cantidad de datos de audio, utilizando 24 archivos, y aumentó el número de épocas de entrenamiento a 140. Contrario a las expectativas, los resultados fueron significativamente menos precisos que los modelos anteriores, sugiriendo una degradación en la calidad de la conversión de voz.

Link modelo 3:

https://drive.google.com/drive/folders/1yPCVPkG6I2VzTEQcG7zyYS0SfhA7Euk_?usp=sharing

Voz de Juan Angel Carrera:

Para esta prueba del modelo buscábamos observar el comportamiento del entrenamiento con audios de baja calidad con mucho ruido por lo que optamos a usar audios de whatsapp que se grabaron mientras se conduce por lo que hay ruido de viento y la voz se escucha algo entrecortada. Los resultados fueron bastante malos por lo que tuvimos que hacer un par de optimizaciones como quitar los silencios, y quitar un poco el ruido.

Link modelo:
https://drive.google.com/drive/folders/1zcKw3npXjAG8nULtQLdiIOhZslpckRY2?usp=sharing

Estos hallazgos indican que un equilibrio óptimo entre la cantidad de datos de entrenamiento y el número de épocas es crucial para la efectividad del modelo. En este caso, el primer modelo 'creep' proporcionó la mejor aproximación a la voz del estudiante, destacando la importancia de la calibración cuidadosa de los parámetros de entrenamiento para la síntesis de voz mediante aprendizaje profundo. Para este modelo se utilizó el algoritmo de rmvpe de HuBert para la recolección de pitch del audio original.

Conclusión

En conclusión, este proyecto puso en evidencia el enorme potencial de la tecnología de conversión de voz basada en recuperación para resolver el desafío que representan los cambios de voces en series, películas y videojuegos icónicos.

Los experimentos realizados lograron replicar voces específicas con un alto grado de similitud. Se observó que la calibración cuidadosa de parámetros como la cantidad y calidad de los datos de entrenamiento, así como el número de épocas, es crucial para optimizar la precisión en la conversión.

Los mejores resultados se obtuvieron en los modelos entrenados con menos de 15 muestras de audio de 30 segundos y alrededor de 100 épocas. Cantidades mayores de datos y épocas no necesariamente mejoraron la calidad de la voz sintetizada. Esto demuestra el potencial de entrenar modelos efectivos incluso con pequeñas muestras de referencia.

Si bien la tecnología aún tiene espacio para mejoras, los audios generados ya alcanzaron un nivel convincente de similitud vocal dado los retos que implica imitar la complejidad del habla humana.

Esta solución representa una innovación prometedora para preservar la autenticidad de las voces más emblemáticas del cine, la televisión y los videojuegos. Al minimizar la disonancia ante cambios de casting, se mantiene la inmersión de los usuarios en las tramas y se eleva la calidad percibida de las producciones.

Su implementación podría transformar el abordaje de la continuidad narrativa en franquicias icónicas, permitiendo reinventarlas sin perder su esencia vocal. Se recomienda continuar perfeccionando estas técnicas para expandir las posibilidades creativas de recrear y preservar voces históricas mediante el aprendizaje de máquinas.

Bibliografía

- RVC-Project/Retrieval-based-Voice-Conversion-WebUI: Voice data <= 10 mins can also be used to train a good VC model! (2023, October 23). GitHub. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/tree/main

Anexos

Link de Github: https://github.com/Jack200133/RVC_FineTune/tree/main