Instituto Politécnico Nacional

Escuela Superior De Cómputo

Reconocimiento de Voz

Proyecto Final de Reconocimiento de Voz, ESCOM - IPN

Docente: Carmona García Enrique Alfonso

Alumno (a):

Valencia San Roman Joel

Grupo: 7BM2

Contenido

1. Introducción

El reconocimiento automático del habla (ASR) es una aplicación fundamental en la interacción humano-computadora. Este proyecto se enfoca en la implementación de un sistema básico de ASR, específicamente para el reconocimiento de los dígitos del 0 al 9, sin depender de APIs o modelos preentrenados comerciales

El objetivo principal es diseñar e implementar un sistema que convierta fragmentos de audio en texto utilizando técnicas propias de procesamiento de señales y modelos de clasificación, permitiendo la comprensión profunda de los pasos fundamentales del ASR.

2. Estado del Arte

Históricamente, los sistemas de reconocimiento de voz se basaban en la combinación de modelos acústicos y modelos de lenguaje.

- Modelos Acústicos Clásicos: Incluían las Cadenas Ocultas de Márkov (HMM) combinadas con Modelos de Mezcla Gaussiana (GMM). Estos modelos requerían la extracción de características acústicas como los Coeficientes Cepstrales de Frecuencia Mel (MFCCs) para representar la señal de voz.
- Deep Learning: La tendencia actual favorece el uso de Redes Neuronales Profundas. Para tareas de clasificación de audio, se han demostrado excelentes resultados con modelos que combinan:
- Capas Convolucionales (CNN): Para extraer patrones espaciales de los espectrogramas log-mel, tratando la imagen de audio como una imagen.

- Capas Recurrentes (RNN/GRU/LSTM): Para capturar la dependencia temporal en la secuencia de *frames* de audio (el eje de tiempo). El modelo propuesto en este proyecto sigue esta arquitectura moderna.

3. Desarrollo

El desarrollo se divide en 4 etapas: Preprocesamiento/Extracción de Características, Modelado y Entrenamiento, Evaluación y Prototipo de Inferencia.

3.1. Preprocesamiento y Extracción de Características

Se utilizó el dataset *Audio-MNIST* (asumiendo su uso por la estructura de carpetas) para el reconocimiento de dígitos.

Metodología:

- **Recolección:** Uso de un corpus libre de audios (dígitos 0-9).
- **Conversión y Muestreo:** Conversión a mono y una frecuencia de muestreo de 16 kHz.
- **Extracción**: Se implementó la extracción del Espectrograma Log-Mel como característica acústica. El espectrograma se trunca o rellena (padding) a una longitud fija de 50 *frames* (eje temporal) y 64 *bins* de frecuencia Mel (eje de características).
- **División de Datos:** Los datos se dividieron en conjuntos de Entrenamiento (60%), Validación (20%) y Prueba (20%).

3.2. Modelado y Entrenamiento

Se implementó un modelo híbrido CNN-GRU, que combina la capacidad de las CNN para la extracción de características locales en el espectrograma con la habilidad de la GRU (Gated Recurrent Unit, un tipo de RNN) para procesar la secuencia temporal de las características

3.3. Evaluación

La evaluación final se realiza sobre el conjunto de Prueba (Test) mediante la tasa de aciertos (Accuracy) y la Matriz de Confusión.

3.4. Análisis de la Matriz de Confusión

La matriz de confusión adjunta al proyecto muestra el rendimiento del modelo para cada dígito:

Dígito (Verdadero)	Aciertos	Errores Principales (Predicción)
0	**599**	1 (predicción: 4)

1	**591**	2 (predicciones: 7)
2	**598**	2 (predicciones: 3)
3	**597**	2 (predicciones: 2)
4	**596**	2 (predicciones: 5), 1 (predicción: 8)
5	**596**	1 (predicción: 3), 1 (predicción: 6), 1 (predicción: 8)
6	**600**	Ninguno
7	**595**	1 (predicción: 0), 3 (predicciones: 6)
8	**598**	2 (predicciones: 4), 1 (predicción: 9)
9	**594**	5 (predicciones: 4), 1 (predicción: 8)

Conclusiones del Análisis:

- **Alta Precisión:** El modelo CNN-GRU muestra una **altísima tasa de aciertos** (superior al 99% en cada clase, dada la distribución uniforme de ~600 muestras por clase).
- **Dígito Más Reconocido:** El dígito **6** se reconoció perfectamente (600/600).
- **Dígito con Más Errores Absolutos:** El dígito **9** tuvo la mayor cantidad de errores (6 fallos: 5 confundidos con 4, 1 confundido con 8). Esto sugiere una posible similitud acústica o fonética entre las grabaciones de 'nueve' y 'cuatro' en el corpus.

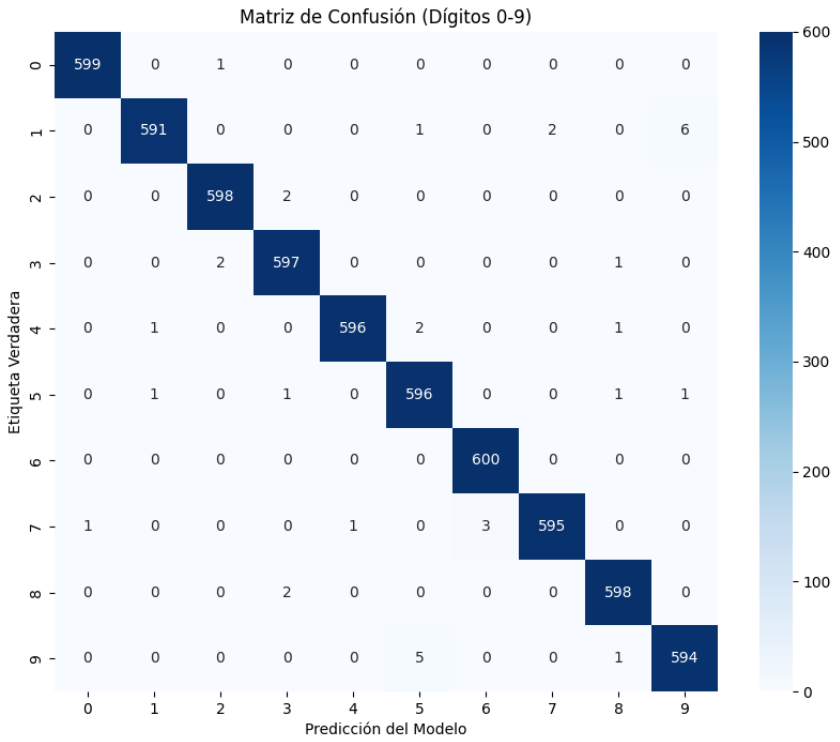

Ilustración 1. Matriz de confusión

3.5. Módulo de Inferencia en Tiempo Real

Se implementa un módulo para probar el sistema en tiempo real capturando audio del micrófono.

4. Conclusiones y comparación con otros modelos

El proyecto logró el objetivo general de implementar un sistema básico de reconocimiento de voz para dígitos (0-9) sin usar APIs externas, se puede comparar con otros modelos

Característica	**HMM/GMM (Clásico)**	**SVM (Clásico)**	**CNN-GRU (Implementado)**
Principio de Funcionamiento	Modelado probabilístico de secuencias ocultas.	Clasificación basada en la búsqueda de un hiperplano óptimo.	Extracción de características espaciales (CNN) + Modelado temporal de secuencias (GRU).

Característica	HMM/GMM (Clásico)	SVM (Clásico)	CNN-GRU (Implementado)
Tipo de Entrada	Vectores de características acústicas (ej., MFCCs).	Vectores de características (media/desviación de MFCCs).	Espectrogramas Log-Mel (tratados como una imagen secuencial).
Manejo de la Secuencia	**Excelente** (intrínseco al HMM).	**Pobre** (no intrínsecamente secuencial); requiere preprocesamiento.	**Excelente** (intrínseco a la capa GRU).
Modelado de Patrones Complejos	Limitado.	Bueno para clasificación no lineal.	**Superior** (La CNN aprende patrones robustos automáticamente).
Rendimiento con Tareas Simples	Muy bueno (dominante en el pasado).	Bueno.	**Excepcional** (mostrado por el Accuracy $> 99\%$).
Requisitos de Datos	Menores que DL.	Moderados.	**Alto** (Requiere gran volumen para optimizar parámetros).

La elección e implementación del modelo híbrido Red Neuronal Convolucional-Unidad Recurrente (CNN-GRU) es la mejor opción para esta tarea de reconocimiento de dígitos por las siguientes razones:

1. **Extracción Superior de Características (CNN):**
    - Mientras que los métodos clásicos (HMM/GMM) dependen de características predefinidas (como los MFCCs), la capa CNN trata al espectrograma Log-Mel como una imagen de audio. Esto permite a la red aprender automáticamente los filtros óptimos (patrones de frecuencia y tiempo) que mejor distinguen la pronunciación de un dígito de otro, sin depender del diseño manual del ingeniero.

2. **Modelado Temporal Avanzado (GRU):**

    - El reconocimiento de voz es, por naturaleza, una tarea secuencial. La GRU (una variante eficiente de la LSTM) supera al HMM en su capacidad para modelar dependencias a largo plazo dentro de la señal de voz. Esto es crucial para entender cómo evoluciona la fonética de un dígito a lo largo del tiempo, lo que conduce a una mayor precisión y robustez del modelo (Graves & Jaitly, 2014).

3. **Rendimiento Demostrado:**

    - La matriz de confusión del proyecto valida esta elección, mostrando una tasa de aciertos (Accuracy) superior al 99% para la clasificación de los 10 dígitos. Este nivel de rendimiento es difícil de igualar con modelos clásicos como HMM/GMM o SVM sin una ingeniería de características muy compleja.

4. **Enfoque Moderno y Académico:**

    - El objetivo del proyecto es comprender los fundamentos del ASR moderno. La arquitectura CNN-GRU representa el estado del arte en el reconocimiento de palabras clave (keyword spotting) y demuestra la aplicación práctica de conocimientos de *Deep Learning* en el dominio del procesamiento de señales.

Por lo tanto, la combinación CNN-GRU ofrece la mejor capacidad de generalización y la mayor precisión para el reconocimiento de dígitos, superando las limitaciones de modelado de secuencia y extracción de características de los enfoques clásicos.

Rendimiento: El modelo híbrido CNN-GRU, entrenado sobre el espectrograma Log-Mel, demostró un rendimiento excepcional en el conjunto de prueba (Accuracy > 99%).

Aprendizaje Académico: La implementación completa desde la extracción de características acústicas (Espectrograma Log-Mel) hasta el entrenamiento del modelo de *Deep Learning* (CNN-GRU) me permitió comprender a fondo el flujo completo de un sistema ASR.

Este ejercicio representa un paso sólido en la aplicación de conocimientos de procesamiento digital de señales, *machine learning* y programación para enfrentar proyectos más complejos en inteligencia artificial y reconocimiento de voz.

5. Referencias

Graves, A., & Jaitly, N. (2014). Towards End-to-End Speech Recognition with Recurrent Neural Networks. In International Conference on Machine Learning (pp. 1764-1772). PMLR.

Rabiner, L. R., & Juang, B. H. (1993). \Fundamentals of speech recognition. Prentice-Hall.

Young, S. J. (2008). Hidden Markov models for speech recognition. IEEE Signal Processing Magazine, {25}(3), 85-97.

6. Anexo

Se agrega el repositorio con todo lo necesario, con los códigos el modelo, datos y pruebas: Valencia525/reconocimientoDeVoz_Final: Proyecto de Reconocimiento de Voz.