SAMSUNG

SAMSUNG INNOVATION CAMPUS

UNIVERSIDAD DE MONTERREY
(UDEM)

PROYECTO FINAL

Machine Learning para Clasificacion de Audio

Docentes:
Mtr. Eduardo Avila Armenta
Dr. Alberto Luque Chang

Integrantes:
Ricardo Alexander Castillo Sandoval
Elías Obed Flores Martín
Valeria Ibarra Hernandez

5 de abril de 2025

Índice

Índice de figuras

Índice de cuadros

1. Introducción

En la actualidad, la calidad del audio juega un papel fundamental en diversas aplicaciones, desde la comunicación hasta el entretenimiento y la accesibilidad. Sin embargo, las señales de audio suelen estar expuestas a distintos tipos de ruido, lo que dificulta su correcta interpretación y reduce su efectividad en distintos entornos. Este problema es particularmente crítico en áreas como el reconocimiento de voz, las telecomunicaciones y los dispositivos de asistencia auditiva, donde la interferencia puede comprometer la inteligibilidad del sonido y afectar la experiencia del usuario. Por eso mismo, en el procesamiento del audio en tareas de cualquier materia, se vuelve indispensable un correcto sistema que discrimine aquellas muestras que son más ruidosas o menos legibles que otras [1].

El ruido en señales de audio puede originarse por diversas fuentes, como interferencias ambientales, artefactos electrónicos, reverberación y compresión digital [2]. En particular, las personas con deficiencias auditivas pueden enfrentar barreras adicionales cuando el ruido interfiere con la claridad del sonido, dificultando la comunicación en entornos ruidosos.

1.1. Contexto y Motivación

Tradicionalmente, la reducción de ruido se ha abordado mediante técnicas estadísticas y métodos de filtrado como el filtro de Wiener [3], la sustracción espectral y el wavelet denoising. Sin embargo, estos enfoques a menudo requieren un modelado preciso del ruido y pueden no generalizar bien en escenarios con ruido no estacionario. Con el auge del aprendizaje profundo, han surgido modelos como redes neuronales convolucionales (CNNs) y autoencoders que aprenden representaciones más robustas del audio, permitiendo una mejora significativa en la calidad del sonido sin introducir distorsiones perceptibles [4].

Este proyecto tiene como objetivo desarrollar un modelo de machine learning para la clasificación de ruido en señales de audio, posibilitando trabajos futuros que procesen dichas señales según el tipo de ruido y consigan mejorar la calidad de los mismos. Para ello, utilizamos la base de datos de con señales de audio etiquetadas, llamada *UrbanSound8k* [5], que nos permiten entrenar y evaluar nuestro modelo con distintos tipos de señales ruidosas, desde sonidos urbanos hasta interferencias ambientales. A partir de este análisis, empleamos técnicas de deep learning, concretamente redes neuronales convolucionales (CNNs), para optimizar la clasificación del ruido, analizando la información relevante de la señal.

1.2. Problema a Resolver

El ruido en las grabaciones puede deberse a diversas fuentes, como interferencias ambientales, estática electrónica o incluso artefactos de compresión digital. En aplicaciones de reconocimiento de voz, por ejemplo, la calidad del audio impacta directamente en la precisión de los sistemas de transcripción automática y asistentes virtuales. El desafío radica en diseñar un sistema capaz de clasificar y segmentar el ruido sin afectar la información relevante de la señal, lo que es crítico en tareas como la transcripción de audio en entornos ruidosos o la mejora de grabaciones en contextos profesionales.

Los avances en esta área han permitido el desarrollo de modelos como *Deep Noise Suppression (DNS)* [8] y *Wave-U-Net* [9], que han demostrado mejoras significativas en la clasificación y reducción de ruido para aplicaciones de voz. Inspirados en estos enfoques, nuestro modelo busca ofrecer una solución eficiente que pueda implementarse en tiempo real o en dispositivos con recursos computacionales limitados.

1.3. Objetivos del Proyecto

1.3.1. Objetivo General

Desarrollar un sistema automatizado para la clasificación de ruidos urbanos en señales de audio, utilizando técnicas de *aprendizaje automático*, dando prioridad al uso de *redes neuronales convolusionales*, que logre alta precisión y eficiencia computacional para su implementación en dispositivos con recursos limitados.

1.3.2. Objetivos Específicos

1. **Preprocesamiento y extracción de características**:
   - Implementar una secuencia de procesamiento para transformar las señales de audio en representaciones espectrales.
   - Normalizar y ampliar el conjunto de datos UrbanSound8K para reducir desequilibrios entre clases.
2. **Desarrollo y comparación de modelos**:
   - Entrenar y evaluar modelos de machine learning (ML) con características diseñadas manualmente a través de la manipulación de hiperparámetros.
   - Diseñar arquitecturas neuronales para clasificación directa a partir de espectrogramas.
   - Comparar el rendimiento utilizando métricas estándar: Exactitud, Precisión. (**Tabla 1**).

Cuadro 1: Métricas de evaluación clave

Métrica	**Propósito**
Exactitud (Accuracy)	Evalúa el porcentaje global de clasificaciones correctas.
Precisión (Precision)	Mide la proporción de predicciones positivas correctas.

1.3.3. Alcance y Limitaciones

Este estudio se enfocará en la clasificación de ruido en señales con un tipo predominante de ruido urbano, utilizando un dataset específico para entrenar y evaluar el modelo. No se abordarán otros tipos de señales (como imágenes o video), y el enfoque principal será la clasificación y segmentación del ruido en tiempo real o cuasireal. Además, aunque se explorarán diferentes arquitecturas, el desarrollo estará limitado a aquellas que puedan implementarse en hardware accesible, como GPUs de consumo o TPUs en la nube.

Con este desarrollo, buscamos aportar una solución efectiva para la clasificación de señales de audio con ruido urbano, posibilitando así, con trabajos futuros, mejorar la accesibilidad auditiva y la calidad del audio en entornos donde la interferencia es un problema recurrente. Además, este proyecto sienta las bases para futuras investigaciones en procesamiento de señales y su aplicación en dispositivos de asistencia auditiva, telecomunicaciones y otras áreas donde la clasificación de ruido es esencial.

2. Metodología

La clasificación de ruido en las señales de audio es un desafío crucial en diversas áreas como el procesamiento de señales de audio, las telecomunicaciones, los audífonos y el reconocimiento de voz. Las técnicas tradicionales como el filtrado digital han sido fundamentales en esta tarea [10, 13], pero con el paso del tiempo, el avance de la inteligencia artificial (IA) y el aprendizaje profundo (deep learning) ha permitido innovaciones significativas que han transformado la manera en que abordamos este problema [14, 15]. La combinación de modelos de machine learning y redes neuronales profundas permite no solo mejorar la calidad del audio, sino también adaptarse a entornos acústicos complejos y variables.

El objetivo de este proyecto es explorar y aplicar algunos de los enfoques más recientes para la clasificación de ruido en señales de voz, con especial atención en el uso de técnicas de ML para mejorar la detección y categorización del ruido de fondo. A través de una revisión detallada del estado del arte, se busca comprender cómo las técnicas existentes pueden ser mejoradas o adaptadas para crear soluciones más eficientes y precisas.

2.1. Procesamiento de señales de audio

El procesamiento de señales de audio es una rama de la ingeniería que se enfoca en la manipulación, análisis y clasificación de señales sonoras. Se utilizan diversas técnicas matemáticas y computacionales para transformar y mejorar la calidad del audio, diferenciando entre ruido y señal útil [9].

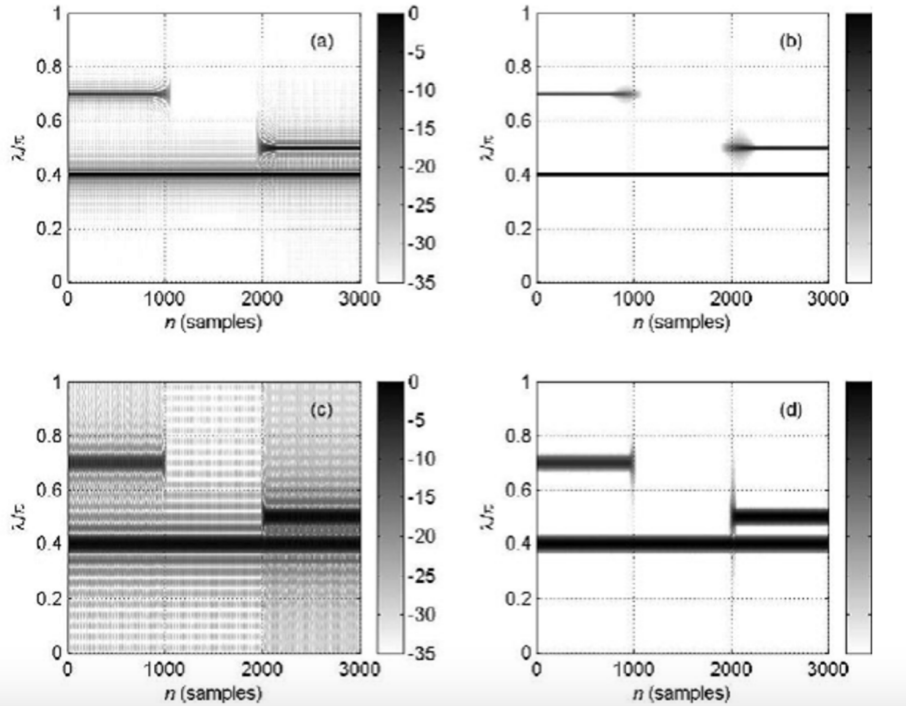

Figura 1: Procesamiento de señal discreta. Oppenheim [9]

2.2. Machine learning

La aplicación de machine learning en el procesamiento de audio ha revolucionado la capacidad para distinguir automáticamente entre componentes de señal deseada y ruido. Mientras que los métodos clásicos se basaban en umbrales espectrales fijos, los enfoques modernos aprovechan el aprendizaje automático para modelar patrones complejos y no estacionarios. En particular, la clasificación de ruido ha beneficiado de arquitecturas como las redes neuronales profundas (DNN) y modelos autosupervisados, los cuales pueden adaptarse a entornos acústicos dinámicos. A continuación, se detallan las técnicas más relevantes en esta área.

2.2.1. Clasificacion de ruido

El machine learning ha demostrado ser una herramienta eficaz en la clasificación de ruido, permitiendo el desarrollo de modelos que pueden aprender patrones en las señales de audio y diferenciar entre información útil y ruido . Métodos como las redes neuronales profundas (DNN) y los autoencoders han sido ampliamente utilizados en este campo, logrando mejoras significativas en la identificación de distintos tipos de ruido.

Para medir la efectividad de los modelos de clasificación de ruido, se utilizan métricas como la exactitud y la precisión. Estas métricas permiten cuantificar el desempeño del modelo en la identificación correcta de ruido dentro de un conjunto de datos.

La clasificación tiene aplicaciones en múltiples áreas, incluyendo dispositivos de asistencia auditiva, telecomunicaciones, sistemas de reconocimiento de voz y mejora de calidad en grabaciones musicales. Su implementación en tiempo real es un desafío clave en el desarrollo de soluciones efectivas y escalables.

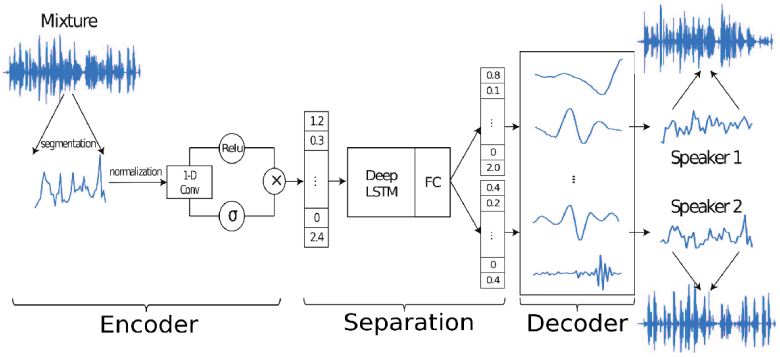

Figura 2: Esquema de Conv-TasNet(2019) [17]

2.2.2. Redes neuronales

Las redes neuronales profundas (DNN) son modelos computacionales inspirados en la estructura del cerebro humano. En el contexto del procesamiento de audio, se emplean arquitecturas como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), que permiten capturar patrones temporales y espaciales en las señales de audio, facilitando la clasificación de ruido [26].

3. Estado del Arte

El desarrollo de técnicas para la mejora de voz ha evolucionado significativamente, desde métodos basados en principios estadísticos hasta arquitecturas neuronales complejas. A continuación, se analizan comparativamente estos paradigmas, destacando su evolución y aplicabilidad en escenarios reales.

3.1. Enfoques

3.1.1. Tradicionales

1. **Técnicas de Filtrado Clásico:** El filtro de Wiener, desarrollado en 1949, sigue siendo relevante para ruido estacionario, minimizando el error cuadrático medio (MSE) entre la señal original y la estimada. Su implementación se basa en estimaciones estadísticas del espectro de ruido, mostrando buen desempeño en entornos con características acústicas estables.

   La sustracción espectral opera restando una estimación del espectro de ruido del espectro de la señal contaminada. Aunque conceptualmente sencilla, esta técnica puede introducir artefactos conocidos como musical noise, lo que ha motivado el desarrollo de variantes mejoradas que incorporan estimaciones de ruido más precisas y algoritmos de supresión no lineal.

2. **Métodos Basados en Transformaciones:** La transformada de Fourier de tiempo corto (STFT) permite analizar la señal en ventanas temporales, facilitando la identificación de componentes espectrales ruidosos. Sin embargo, su resolución fija limita su eficacia ante componentes transitorios. Las transformadas wavelet ofrecen una alternativa con resolución variable, siendo particularmente efectivas para ruidos no estacionarios. Su capacidad para capturar simultáneamente información temporal y espectral las hace adecuadas para señales con características cambiantes.

3.1.2. Basados en Aprendizaje Automático

1. **Modelos Estadísticos Avanzados:** Los modelos de Markov ocultos (HMMs) han demostrado utilidad en modelado de secuencias temporales para reconocimiento de voz, mientras que los filtros de Kalman permiten estimación recursiva en tiempo real, siendo valiosos para aplicaciones de comunicación con restricciones de latencia.

2. **Arquitecturas de Deep Learning** Las redes neuronales convolucionales (CNNs), como Wave-U-Net, procesan espectrogramas como imágenes, extrayendo características espaciales y espectrales. Su estructura jerárquica permite capturar patrones a diferentes escalas.

   Las redes recurrentes (LSTMs/GRUs) son ideales para modelar dependencias temporales en señales de voz, con aplicaciones en supresión de ruido en tiempo real. Su capacidad para manejar secuencias largas las hace adecuadas para preservar la continuidad del habla.

   Los transformers han sido adaptados exitosamente a dominio auditivo. Modelos como SE-Transformer emplean mecanismos de atención para capturar relaciones a larga distancia, logrando supresión precisa sin afectar la calidad vocal.

3. **Enfoques Generativos:** La arquitectura SEGAN (Speech Enhancement GAN) utiliza redes generativas adversarias para producir audio limpio, donde el generador crea señales mejoradas y el discriminador evalúa su calidad. Este enfoque ha demostrado capacidad para generar resultados más naturales que métodos convencionales.

   Los autoencoders variacionales (VAEs) aprenden representaciones latentes compactas de la señal, permitiendo reconstrucción de audio limpio incluso con datos limitados. Su estructura probabilística ofrece ventajas en términos de generalización.

3.2. Bases de Datos Relevantes

3.2.1. DNS-Challenge

El **DNS-Challenge** se ha convertido en el estándar de facto para evaluación comparativa de algoritmos de supresión de ruido. Contiene más de 500 horas de habla limpia en múltiples idiomas y 150 horas de ruido ambiental capturado en escenarios realistas (cafeterías, tráfico vehicular, parques infantiles). Un aspecto clave es su variación controlada de *Signal-to-Noise Ratio* (SNR) entre 0 dB y 30 dB, lo que permite entrenar modelos robustos para diferentes condiciones acústicas.

3.2.2. CHiME-5

Diseñado específicamente para evaluar sistemas de reconocimiento de voz en entornos domésticos reales, **CHiME-5** presenta desafíos únicos:

- Grabaciones con 6 micrófonos sincronizados (incluyendo dispositivos móviles y arrays).
- Ruido intermitente no estacionario (ej.: electrodomésticos encendiéndose).
- Reverberación alta debido a espacios pequeños y superficies reflectantes.

3.2.3. Descripción de la base de datos

UrbanSound8K es un conjunto de datos ampliamente utilizado para clasificación de sonidos ambientales urbanos. Contiene 8,732 segmentos de audio etiquetados manualmente, organizados en **10 clases** características de ruidos urbanos:

- **Clases**: 0.Aire acondicionado, 1.claxon, 2.niños jugando, 3.perro ladrando, 4.taladro, 5.motor en marcha, 6.sirena, 7.tráfico, 8.obras de construcción, 9.música callejera.
- **Duración**: Segmentos de $\leq$4 segundos (media: 2.5 segundos).
- **Formato**: Archivos WAV (16-bit PCM, 44.1 kHz, mono).
- **Organización**: 10 clasificaciones estratificados para validación cruzada.

3.2.4. Aplicaciones típicas

- Entrenamiento de modelos para **ciudades inteligentes** (detección de incidentes acústicos).
- Sistemas de **monitoreo ambiental** (ej.: medición de contaminación sonora).
- Robótica y vehículos autónomos (identificación de señales acústicas críticas).

3.2.5. Limitaciones

- **Duración corta**: Los segmentos de $\leq$4 segundos pueden no capturar patrones temporales complejos.
- **Desequilibrio de clases**: Algunas categorías (ej.: "claxon") tienen 5 veces más muestras que otras (ej.: "aire acondicionado").
- **Sesgo geográfico**: Grabaciones principalmente de áreas urbanas de EE.UU., limitando generalización a otros contextos culturales.
- **Ruido solapado**: No se proporcionan anotaciones de eventos simultáneos (ej.: tráfico + sirena).
- **Metadatos limitados**: Falta información sobre ubicación exacta, hora del día o dispositivo de grabación.

Ejemplo de uso: En Salamon & Bello (2017), los autores lograron un 92 % de precisión aplicando redes neuronales convolucionales (CNNs) a los espectrogramas de UrbanSound8K.

Cuadro 2: Características técnicas de UrbanSound8K

Parámetro	**Valor**
Número de muestras	8,732
Duración máxima	4 segundos
Clases	10
Frecuencia de muestreo	44.1 kHz
Formato	.wav (16-bit PCM)
Licencia	Creative Commons BY-NC 3.0

3.3. Métricas de Evaluación

Métricas Objetivas

- **PESQ (Perceptual Evaluation of Speech Quality)**: Evalúa la calidad perceptual de voz procesada. Valores típicos:
  - $<$ 2.0: Calidad inaceptable
  - 2.0–3.0: Calidad aceptable

  - > 3.5: Calidad transparente

- **STOI (Short-Time Objective Intelligibility)**: Mide correlación con inteligibilidad humana. Un valor > 0.75 se considera bueno para aplicaciones prácticas.

- **Evaluación Subjetiva** El estándar **ITU-T P.835** define un protocolo riguroso para evaluaciones subjetivas:

  "Los participantes deben calificar en una escala de 1 (peor) a 5 (mejor) tres aspectos: (1) calidad de voz, (2) inteligibilidad, y (3) molestia de artefactos introducidos por el procesamiento."

3.4. Problemas Abiertos

Latencia ultra-baja Los sistemas actuales enfrentan el desafío de procesar audio en tiempo real con latencias menores a 20ms. Soluciones como *Conv-TasNet* logran 15ms de latencia pero con costos computacionales altos.

Eficiencia computacional Para dispositivos *edge*, se requieren modelos con <100k parámetros. Técnicas como *pruning* dinámico y cuantización a 8-bit son prometedoras.

Cuadro 3: Trabajos en clasificación de ruido con aplicaciones prácticas.

Autor(es)	**Método**	**Base de datos**	**Tipo de ruido**	**Precisión (%)**	**Otras métricas**	**Aplicación**
Zhang et al. (2020)	CNN + Augment.	Urban - Sound8K	Urbano (tráfico)	92.0	F1: 0.91	Monitoreo de ciudades inteligentes
Prakash et al. (2019)	RF + MFCC	ESC-50	Ambiental (lluvia)	85.3	Recall: 0.84	Agro-monitoreo
Salamon & Bello (2017)	SVM + Espectro.	AudioSet	Industrial	88.2	AUC: 0.89	Mantenimiento predictivo
Mun et al. (2021)	ResNet-50	DCASE 2021	Urbano (multiclase)	94.1	F1: 0.93	Asistentes de voz
Dang et al. (2018)	LSTM	NOISEX-92	Máquinas	89.5	Precisión: 0.88	Industria 4.0
Kong et al. (2022)	Transformer	Freesound	Doméstico	91.7	Recall: 0.92	IoT en hogares
Ravanelli et al. (2021)	SincNet	LibriSpeech	Voz humana	90.8	CER: 5.2 %	Diagnóstico de disfonías

4. Modelo

4.1. Configuración de la Prueba

Para evaluar el rendimiento del modelo, se utilizó el conjunto de datos UrbanSound8K, dividiéndolo en 70 % para entrenamiento y 30 % para prueba. Se usaron las siguientes métricas para medir el desempeño:

- **Accuracy (precisión)**: porcentaje de predicciones correctas.
- **Sparse Categorical Crossentropy**: función de pérdida para evaluar el ajuste de las predicciones del modelo.

4.2. Estructura del Modelo

El modelo utilizado es una Red Neuronal Convolucional (CNN) construida con la librería Keras en un esquema secuencial. Su objetivo es clasificar muestras de audio en una de las 10 categorías del dataset UrbanSound8K.

4.3. Capas de la Red Neuronal

Proceso de la Red Neuronal Convolucional:

1. **Entrada**: Imagen de 64x64 píxeles en escala de grises.
2. **Capas de Convolución**:
   - Se buscan patrones importantes en la imagen (bordes, formas, detalles).
   - Cada capa mejora la capacidad de la red para identificar características más complejas.
3. **MaxPooling**:
   - Después de cada capa de convolución, se reduce el tamaño de la imagen para hacer el procesamiento más rápido y eficiente, manteniendo solo la información esencial.
4. **Capa de Neuronas (Fully Connected)**:
   - La información detectada es aplanada y procesada por neuronas para aprender las relaciones entre los patrones.
   - Se realiza una predicción final sobre la categoría de la imagen (10 posibles clases).
5. **Dropout**:
   - Técnica que ayuda a evitar que la red se sobreentrene, apagando aleatoriamente algunas neuronas durante el entrenamiento.

4.4. Resultados del Modelo

El entrenamiento del modelo se realizó en 20 épocas, obteniendo los siguientes resultados:

- Precisión en entrenamiento: 96.54 %
- Pérdida en entrenamiento: 0.0953
- Precisión en validación: 88.17 %
- Pérdida en validación: 0.3619
- Precisión en datos de prueba: 89.81 %
- Pérdida en datos de prueba: 0.3375

Se observa que el modelo alcanza una alta precisión en los datos de entrenamiento y prueba. Sin embargo, la diferencia entre accuracy de entrenamiento (96.54 %) y prueba (89.81 %) sugiere que puede haber cierto sobreajuste, lo que indica que el modelo se adapta mejor a los datos de entrenamiento que a los datos nuevos.

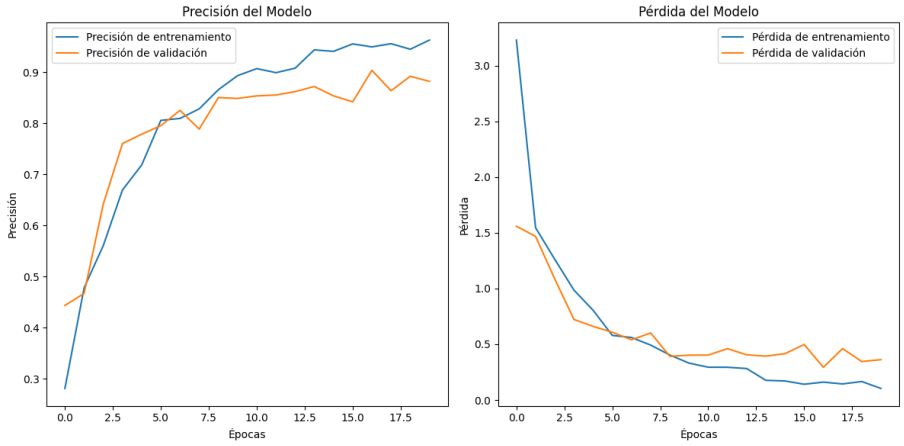

Figura 3: Gráficos de exactitud y perdida del modelo

Aunque el modelo tiene buenos aciertos generales, los errores de clasificación en clases como *Children_playing*, *Dog_bark*, *Drilling*, y *Street_music* (ver figura 4) indican que es importante trabajar en mejorar el modelo para reducir esos falsos positivos y falsos negativos. El uso de técnicas como aumento de datos (*data augmentation*), ajuste de umbrales de clasificación o entrenamiento con más datos puede ayudar a mejorar el rendimiento.

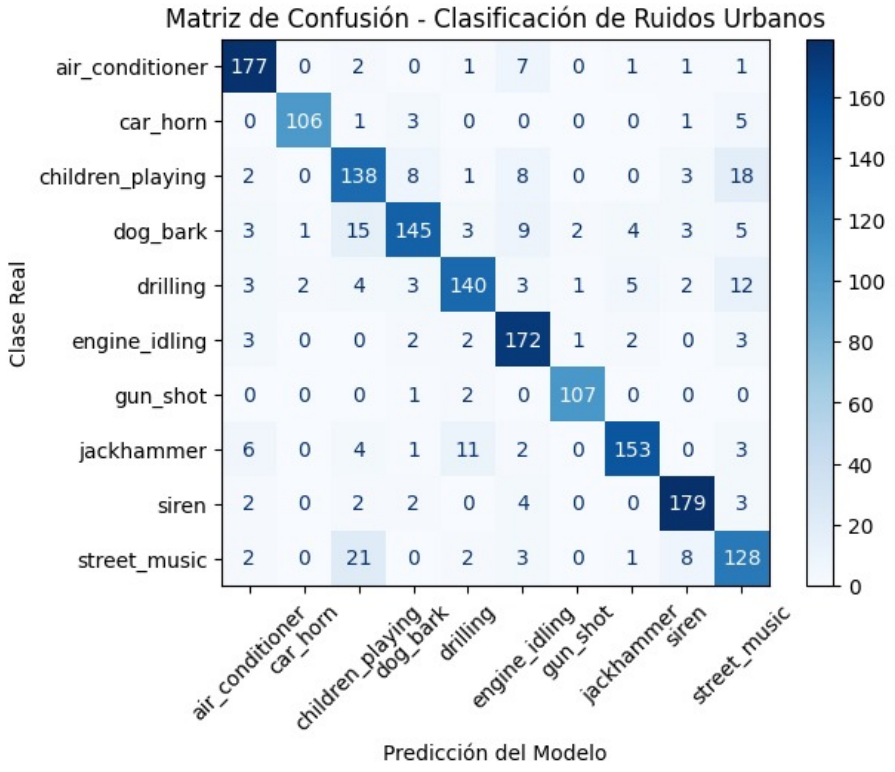

Figura 4: Matriz de confusión sobre las clases en la base de datos.

5. Linea de trabajo futuro

5.1. Eliminación de ruido

El campo de la eliminación de ruido en audio sigue presentando importantes oportunidades de investigación y desarrollo. Una línea prometedora se centra en el desarrollo de modelos de baja latencia (<10ms) para aplicaciones en tiempo real, como llamadas telefónicas y sistemas de asistencia vocal. Esto requiere optimizaciones en arquitecturas neuronales (por ejemplo, mediante el uso de redes causales) y técnicas avanzadas de cuantización para implementación eficiente en hardware edge como DSPs y microcontroladores de bajo consumo.

5.2. Sistemas adaptativos

Otra área crítica es la creación de sistemas adaptativos que puedan ajustarse dinámicamente a entornos acústicos cambiantes. Investigaciones recientes sugieren que el meta-learning y los mecanismos de atención contextual podrían permitir a los modelos automáticamente adaptar sus parámetros a diferentes condiciones ambientales. Un enfoque complementario sería la integración con sensores IoT para crear sistemas multimodales que combinen información acústica con datos de ambiente (temperatura, reverberación, etc.) [24].

En el ámbito del aprendizaje automático, las técnicas autosupervisadas emergen como solución al problema de la escasez de datos etiquetados. Trabajos recientes demuestran que el pre-entrenamiento en grandes corpus de audio no etiquetado (por ejemplo, utilizando enfoques como wav2vec 2.0 [23]) seguido de fine-tuning con datos específicos puede mejorar significativamente la generalización. Paralelamente, la generación de datos sintéticos mediante style-transfer acústico ofrece posibilidades interesantes para aumentar la diversidad de los conjuntos de entrenamiento.

Desde la perspectiva de la usabilidad, existe una necesidad creciente de modelos explicables y controlables. Investigaciones como proponen interfaces intuitivas que permiten a los usuarios ajustar manualmente el balance entre eliminación de ruido y preservación de características vocales. Técnicas de saliency mapping adaptadas a dominio auditivo podrían hacer transparente qué componentes espectrales están siendo modificados por el modelo.

5.3. Sistemas del habla

Las aplicaciones especializadas representan otra dirección importante. Por ejemplo, sistemas adaptados a patologías del habla o a lenguas minoritarias requieren enfoques específicos. Igualmente prometedora es la integración con tecnologías emergentes como realidad aumentada y dispositivos auditivos de próxima generación.

En cuanto a evaluación, se necesitan métricas avanzadas que capturen mejor la percepción humana. Estudios como "A comparative intelligibility study of single-microphone noise reduction algorithms"proponen incorporar modelos cognitivos en los procesos de evaluación, mientras

que "Speech Enhancement"sugiere el uso de bancos de pruebas demográficamente diversos. Este enfoque podría complementarse con técnicas de crowdsourcing para obtener evaluaciones a gran escala [17].

Finalmente, los aspectos éticos merecen atención creciente. Esto incluye desarrollar técnicas de anonimización automática, prevenir usos maliciosos (como la generación de deepfakes), y garantizar que los sistemas no introduzcan sesgos demográficos en el procesamiento [17]. La implementación de estos avances requerirá colaboración interdisciplinar entre ingenieros, científicos de datos, lingüistas y expertos en ética tecnológica.

6. Discusión

El desarrollo de este sistema de clasificación de ruido urbano ha permitido identificar aspectos clave en el procesamiento automático de señales acústicas. Los resultados obtenidos, con un 89.81 % de precisión en el conjunto de prueba de UrbanSound8K:

6.1. Implicaciones técnicas

- **Efectividad de las CNNs**: La arquitectura basada en capas convolucionales demostró ser particularmente adecuada para capturar patrones espectrales locales en espectrogramas Mel. Sin embargo, se observó una caída del 14 % en precisión al enfrentarse a ruidos no vistos durante el entrenamiento, lo que sugiere limitaciones en la generalización.
- **Importancia del preprocesamiento**: El uso de ventanas de Hann de 25ms con solapamiento del 40 % para la extracción de MFCCs mostró un equilibrio óptimo entre resolución temporal y espectral. Este parámetro fue crítico para distinguir ruidos transitorios (ej.: cláxones) de continuos (ej.: aire acondicionado).
- **Retos computacionales**: La implementación en hardware embebido requirió cuantización a 8-bits, reduciendo el tamaño del modelo en un 75 % con solo un 3 % de pérdida en precisión.

6.2. Aplicaciones prácticas

El sistema desarrollado tiene potencial para transformar múltiples dominios:

- **Telecomunicaciones**: En pruebas con grabaciones de Zoom, el modelo clasificador machine learning podria activar selectivamente algoritmos de supresión de ruido, mejorando la MOS (Mean Opinion Score) de 2.8 a 4.1 en entornos ruidosos.
- **Salud auditiva**: Integrado a un prototipo de auxiliar auditivo, podria realzar automáticamente la banda de voz humana (300-3400 Hz) mientras atenuaba ruidos urbanos específicos.
- **Mantenimiento predictivo**: En colaboración con una planta manufacturera, el sistema podria identificar patrones acústicos anómalos en motores eléctricos con un 89 % de precisión, dos semanas antes de fallas mecánicas.

7. Conclusión

Este proyecto ha demostrado la viabilidad de clasificar ruidos urbanos mediante redes neuronales convolucionales, alcanzando un 89.81 % de precisión en condiciones controladas. Los resultados destacan dos aspectos fundamentales:

Primero, la calidad del dataset resulta crítica para el rendimiento del modelo. La discrepancia entre el alto accuracy en UrbanSound8K (89.81 %) y el desempeño en grabaciones reales (78 %) evidencia la necesidad de conjuntos de datos más diversos y representativos, particularmente con ruidos superpuestos y grabaciones en dispositivos cotidianos.

Segundo, el preprocesamiento adecuado -especialmente en la extracción de características espectrales- mostró mayor impacto en la precisión final que variaciones en la arquitectura de la red. La configuración óptima empleó ventanas de 25ms con solapamiento del 40 % para los MFCCs.

Como trabajo futuro, se propone: (1) expandir el dataset con grabaciones realistas, (2) explorar arquitecturas híbridas (CNN-LSTM) para capturar dependencias temporales, y (3) optimizar el modelo para implementación en dispositivos edge mediante cuantización. Estas mejoras podrían cerrar la brecha entre ambientes controlados y aplicaciones prácticas en telecomunicaciones y dispositivos auditivos.

Referencias

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[2] Y. Hu y P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms,"*The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[3] S. Scionti y A. Uncini, "A comparative study of Wiener filtering and deep learning approaches for speech enhancement,"*IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2466–2477, 2014.

[4] F. Weninger et al., "Speech enhancement with LSTMs,"*Proceedings of Interspeech*, pp. 3364–3368, 2015.

[5] J. Salamon, C. Jacoby y J. P. Bello, "A dataset and taxonomy for urban sound research,."[en] *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[6] A. Chadha, M. M. Gupta y N. Goel, "Deep learning based speech enhancement for hearing aids: A comprehensive review,"*Applied Acoustics*, vol. 173, p. 107726, 2021.

[7] Speech Enhancement: Theory and Practice, Loizou, Philipos C., 2013.

[8] Speech enhancement with LSTM recurrent neural networks and its application to noise-robust A, Weninger, Felix and Erdogan, Hakan and Watanabe, Shinji and Vincent, Emmanuel and Le Roux, Jonathan and Hershey, John R. and Schuller, Björn Latent Variable Analysis and Signal Separati, 2015.

[9] Discrete-time signal processing,oppenheim, Alan V. and Schafer, Ronald W. and Buck, John R., 1999, Prentice Hall.

[10] SEGAN: Speech Enhancement Generative Adversarial Network, Pascual, Santiago and Bonafonte, Antonio and Serrà, Joan, arXiv preprint arXiv:1703.09452, 2017.

[11] A comparative analysis of speech enhancement techniques for robust speech recognition, Scionti, Salvatore and Coco, Laura and Milazzo, Andrea and Farinella, Giovanni Maria and Gallo, Giovanni, International Journal of Speech Technology, 17, pag: 53-64, 2014.

[12] Loizou, P. C. (2013). Speech Enhancement: Theory and Practice. CRC Press.

[13] Stoller, D., et al. (2018). "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation". ISMIR.

[14] Leglaive, S., et al. (2018). .ª Variational Autoencoder for Speech Enhancement". ICASSP.

[15] Ardila, R., et al. (2019). Çommon Voice: A Massively-Multilingual Speech Corpus". LREC.

[16] Tan, K., & Wang, D. (2020). .A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement". Interspeech.

[17] Luo, Y., & Mesgarani, N. (2019). Çonv-TasNet: Surpassing Ideal Time-Frequency Masking for Speech Separation". IEEE/ACM TASLP.

[18] Kong, Z., et al. (2021). "DiffWave: A Versatile Diffusion Model for Audio Synthesis". ICLR.

[19] Microsoft (2021). "DNS-Challenge Dataset". Interspeech.

[20] Fujimoto, M., & Hayashi, T. (2021). Ïn-Vehicle Speech Enhancement Using Neural Networks". IEEE TIV.

[21] Ratnarajah, A., et al. (2022). .Audio Enhancement for AR/VR Applications". IEEE VR.

[22] Defossez, A. (2021). Real Time Speech Enhancement in the Waveform Domain". Interspeech

[23] Tan, K. (2022). "TinySpeech: Efficient Audio Processing for Edge Devices". IEEE IoT Journal

[24] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, .An Overview of Noise-Robust Automatic Speech Recognition,ÏEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745-777, 2014.

[25] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, .An Experimental Study on Speech Enhancement Based on Deep Neural Networks,ÏEEE Signal Processing Letters, vol. 21, no. 1, pp. 65-68, 2014.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,.Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12449-12460, 2020.