NoiseSuppressNet: Supresión de Ruido en Audio en Tiempo Real mediante Deep Learning

Abstract

Palabras clave: *supresión de ruido, deep learning, procesamiento de audio en tiempo real, redes neuronales recurrentes, speech enhancement, habla en español.*

Desarrollo

Motivación. La calidad del audio en comunicaciones de voz es un problema vigente en entornos de trabajo modernos, especialmente en sistemas de contact center, videollamadas y bots de voz. Los algoritmos clásicos de procesamiento de señales digitales (DSP), basados en filtros de Wiener o sustracción espectral, ofrecen resultados aceptables frente a ruido estacionario —como el zumbido eléctrico o el ventilador de una PC— pero fallan notoriamente ante ruido no estacionario: voces de fondo, tráfico urbano, teclas, o el ruido ambiental típico de oficinas abiertas. El avance del deep learning ha demostrado que modelos relativamente livianos pueden superar ampliamente al DSP clásico en este dominio, con latencias algorítmicas por debajo de los 20 ms que los hacen viables en tiempo real.

A esto se suma un vacío específico en la literatura: la gran mayoría de los modelos de referencia actuales —RNNoise (Mozilla), DeepFilterNet (IIS Fraunhofer), NSNet2 (Microsoft)— son entrenados y evaluados sobre corpus de habla en inglés. El español, con sus particularidades fonéticas y prosódicas propias, carece de modelos de supresión de ruido optimizados para su fonología. Este proyecto propone cubrir dicha brecha.

Métodos. Se implementará una arquitectura CRN (Convolutional Recurrent Network) en PyTorch, compuesta por un encoder convolucional 2D, una capa GRU unidireccional causal para modelar dependencias temporales, y un decoder convolucional que predice una máscara de ganancia espectral por bin de frecuencia. La señal de entrada es audio PCM mono a 16 kHz; el procesamiento opera sobre espectrogramas STFT con frames de 20 ms y hop de 10 ms. El modelo resultante es causal: no requiere contexto futuro, garantizando latencia algorítmica real y medible.

El entrenamiento se realizará en dos etapas. En la primera, el modelo se entrena sobre el corpus estándar DNS Challenge (Microsoft) con voz en inglés mezclada con ruidos del dataset MUSAN y ESC-50, usando una función de pérdida SI-SDR como baseline. En la segunda etapa —la contribución diferenciadora del proyecto— se realizará un fine-tuning sobre Common Voice en español, y se incorporará una función de pérdida basada en PESQNet (red auxiliar que estima el score PESQ de forma diferenciable), siguiendo la técnica propuesta por Xu et al. (IEEE TASLP, 2022). Esto permite optimizar directamente la calidad perceptual de la voz en lugar de una métrica de señal pura.

Resultados esperados. Se espera obtener un modelo causal de baja complejidad (< 5M parámetros) capaz de operar en tiempo real sobre CPU de consumo. Como referencia orientativa, la arquitectura CRN de base reporta en la literatura mejoras

de PESQ de ~0.7 puntos y STOI de ~0.08 puntos sobre la señal sin procesar. La hipótesis central del proyecto no es alcanzar un umbral absoluto, sino demostrar que el fine-tuning con PESQNet loss mejora las métricas perceptuales respecto al baseline entrenado con SI-SDR, y que entrenar sobre habla española produce ventajas medibles al evaluar sobre hablantes hispanohablantes."

Conclusiones. *[Esta sección se completará una vez finalizado el desarrollo experimental, al término del período de implementación.] Se prevé documentar los hallazgos más relevantes respecto a la efectividad de la pérdida perceptual para habla en español, el trade-off entre tamaño del modelo y calidad de supresión, y las limitaciones prácticas encontradas en la inferencia en CPU de bajo costo.*

Tutores Externos

- **Diego Durante**
- **Ramiro Verrastro**