UNIVERSIDAD DEL AZUAY

Facultad de Ciencias de la Administración

Carrera de Ingeniería en Ciencias de la Computación

Arquitectura de red neuronal tipo GAN para la reducción y eliminación de ruido en audios de llamadas de emergencia

Trabajo de titulación previo a la obtención del grado de Ingeniero en Ciencias de la Computación

Autor:

Roberto Geraldo Velásquez Cedeño

Director:

Marcos Patricio Orellana Cordero

Cuenca – Ecuador

Año

2026

DEDICATORIA

A mis padres, por su ejemplo y apoyo incondicional.
A mi hermana, por siempre ser mi inspiración y motivación.
A mi pareja, cuyo amor ha sido mi fortaleza constante.

AGRADECIMIENTO

Expreso mi sincero agradecimiento a mi tutor, el Ing. Marcos Orellana, por su guía durante el desarrollo de esta investigación.

A la Ing. Catalina Astudillo por su valioso apoyo y orientación durante este proceso.

Al Laboratorio de Investigación y Desarrollo en Informática (LIDI), por facilitar sus equipos, los cuales fueron indispensables para el desarrollo de la red.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

RESUMEN

Esta investigación presenta AME (ASR-Aware Multibranch Enhancer), un modelo de mejora de habla diseñado para optimizar el rendimiento de sistemas de reconocimiento automático de voz (ASR) en condiciones de datos limitados y supervisión débil. A diferencia de los enfoques tradicionales basados en GAN, centrados en la calidad perceptual, la propuesta se enfoca en los factores que degradan el desempeño del ASR, incluyendo ruido y distorsiones en escenarios de llamadas de emergencia, sin requerir datos pareados limpio-ruidoso ni la optimización explícita de métricas no diferenciables como el Word Error Rate (WER). El modelo opera sobre la señal en el dominio del tiempo e incorpora un componente adversarial, D_WER, que aprende una representación proxy de la calidad relevante para ASR mediante la aproximación y el ordenamiento relativo de señales derivadas del WER, permitiendo guiar el entrenamiento en función del desempeño en reconocimiento. El método se evalúa sobre 1000 audios no pareados en español provenientes de llamadas de emergencia (ECU911), con ruido leve (MOS ≈ 4.25), donde el margen de mejora es limitado. Los resultados muestran que el modelo no mejora consistentemente el desempeño del ASR, con un aumento del WER promedio de 0.2765 a 0.2941 y solo un 36.5% de muestras con mejora. Sin embargo, el discriminador muestra una fuerte capacidad de ordenamiento, con una correlación de Spearman de 0.53 y NDCG@20 superior a 0.85. Estos resultados indican que la señal aprendida es informativa, pero carece de calibración absoluta, limitando su efectividad en la optimización adversarial.

Palabras clave: aprendizaje débilmente supervisado, aprendizaje orientado a ASR, mejora de habla, mejora de habla en bajo recurso, redes generativas adversariales

ABSTRACT

This paper introduces AME (ASR-Aware Multibranch Enhancer), a speech enhancement model designed to optimize the performance of automatic speech recognition (ASR) systems under low-resource and weakly supervised conditions. Unlike conventional GAN-based approaches focused on perceptual quality, the proposed method targets factors that degrade ASR performance, including noise and distortions in emergency call scenarios, without requiring paired clean-noisy data or explicit optimization of non-differentiable metrics such as Word Error Rate (WER). The model operates on raw waveforms and incorporates an adversarial component, D_WER, which learns a proxy representation of ASR-relevant quality by approximating and ranking WER-derived signals, enabling training guided by recognition performance. The approach is evaluated on 1000 unpaired Spanish emergency call recordings (ECU911) with mild noise (MOS ≈ 4.25), where improvement margins are limited. Results show the model does not consistently improve ASR performance, with average WER increasing from 0.2765 to 0.2941 and only 36.5% of samples improving. Nevertheless, the discriminator shows strong ranking capability, achieving a Spearman correlation of 0.53 and NDCG@20 above 0.85. These findings indicate the learned signal is informative but lacks absolute calibration, limiting its effectiveness for stable adversarial optimization.

Keywords: ASR-aware learning, generative adversarial networks, low-resource speech enhancement, speech enhancement, weakly supervised learning