ISSN 1870-4069

Red neuronal convolucional ramificada con atención para la mejora de voz

Noel Zacarias-Morales, José Adán Hernández-Nolasco,
Pablo Pancardo

Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información,
México

{adan.hernandez,pablo.pancardo}@ujat.mx
201H18002@alumno.ujat.mx

Resumen. La mejora de la voz es un proceso que implica eliminar o atenuar el ruido presente en una señal de voz. En este sentido, muchos autores han empleado las redes neuronales para extraer la voz de manera inteligible y de calidad. A diferencia de los artículos que se revisaron, este trabajo propone una red convolucional ramificada con atención de múltiples encabezados que hace posible reducir el número de parámetros entrenables, así como incrementar la precisión en la extracción de la voz. Los resultados obtenidos demostraron que la incorporación del mecanismo de atención basado en múltiples encabezados mejoró en términos generales la capacidad del modelo convolucional ramificado, conforme a los valores de las métricas de calidad PESQ, STOI y SI-SDR. Los valores logrados confirman que incorporar un mecanismo de atención permite la mejora de los modelos de redes neuronales convolucionales ramificadas para extraer voz inteligible y de calidad.

Palabras clave: Red neuronal, convolución, atención, voz, ruido.

Branched Convolutional Neural Network with Attention for Voice Enhancement

Abstract. Speech enhancement is a process that involves removing or attenuating noise present in a speech signal. In this sense, many authors have used neural networks to extract the voice in an intelligible and quality way. Unlike the articles that were reviewed, this work proposes a branched convolutional network with attention to multiple headers that makes it possible to reduce the number of trainable parameters, as well as increase the precision in voice extraction. The results obtained showed that the incorporation of the attention mechanism based on multiple headers generally improved the capacity of the branched convolutional model, according to the values of the PESQ, STOI and SI-SDR quality

metrics. The values obtained confirm that incorporating an attention mechanism allows the improvement of branched convolutional neural network models to extract intelligible and quality voice.

Keywords: Neural network, convolution, attention, voice, noise.

1. Introducción

Un reto fundamental en la audición es escuchar selectivamente diferentes sonidos en una mezcla de señales acústicas. Es decir, la extracción de parámetros de una sola fuente de sonido es especialmente difícil en las grabaciones de un solo canal. El mejoramiento de la voz es la tarea de eliminar o atenuar el ruido añadido en una señal de voz, y generalmente se ocupa en mejorar la inteligibilidad y la calidad de la voz que sufre degradación por incluir ruido. El mejoramiento de la voz se emplea como procesado previo en aplicaciones como en el reconocimiento automático de la voz.

El propósito de la mejora de voz monocanal es proporcionar una solución al problema en el que se utilizan grabaciones hechas con un único micrófono. La mejora del habla monocanal se considera un problema muy difícil, ya que no se tienen pistas direccionales del origen de las distintas señales de audio que componen los ruidos presentes. En el mundo real, las señales de voz se ven fácilmente corrompidas por ruido. Los ruidos pueden agruparse en ruidos estacionarios (que no cambian en función del tiempo) y ruidos no estacionarios (que cambian cuando transcurre el tiempo).

Algunos ruidos que pertenecen a la categoría de no estacionarios son los ruidos de la calle, el ruido de un tren, el ruido de balbuceo (la voz de otras personas) y los sonidos de instrumentos musicales. Algunos que pertenecen a la categoría de estacionarios son los procedentes de acondicionadores de aire, ventiladores, compresores o bombas impulsoras. La relación entre la voz y el ruido en el dominio del tiempo puede escribirse como (1):

$$y(t) = x(t) + n(t), \tag{1}$$

donde *x(t)* es la señal de voz limpia y *n(t)* es el ruido añadido, dando como resultado que *y(t)* sea la señal de voz con ruido. Ahora bien, sea t el índice de tiempo, la señal puede representarse como $y = [y(1), \ldots, y(T)]$, donde T es la longitud del fragmento de audio. Al aplicar la transformada de Fourier de tiempo corto (STFT), podemos representar la señal acústica de (1) en el dominio de tiempo-frecuencia (TF) como (2):

$$Y(k,l) = X(k,l) + N(k,l), \tag{2}$$

donde k es el índice de la banda de frecuencias, l denota el índice de la trama temporal, *Y(k,l)*, *X(k,l)*, y *N(k,l)* son los coeficientes STFT de la señal de voz

ruidosa, la señal objetivo y la señal de ruido, respectivamente. Las definiciones anteriores son válidas únicamente para un micrófono de un solo canal.

En este caso, la tarea de mejora de voz tiene como objetivo recuperar la señal de voz objetivo x de la señal de voz ruidosa y [20]. Nuestra propuesta se basa en el mapeo del espectrograma de magnitud, y en este método basado en el mapeo, el objetivo de entrenamiento del modelo es mapear una función no lineal F desde la señal de voz con ruido *y(t)*, a una señal de voz limpia mejorada *x(t)*, como se escribe en (3):

$$y(t) \rightarrow^{F} x(t). \tag{3}$$

Debido a que existen problemas de variación rápida cuando se usa la señal de voz sin procesar (en el dominio del tiempo), el método basado en el mapeo se aplica habitualmente al espectrograma de magnitud de la señal de voz (dominio de la frecuencia), que se crea aplicando la transformada de Fourier de tiempo corto en una ventana temporal de un banco de filtros. Posteriormente, se realiza la operación inversa de la transformada de Fourier de tiempo corto para reconstruir el espectrograma de vuelta a la señal en el dominio del tiempo utilizando la información de fase de la señal de voz original con ruido.

Las redes neuronales basadas en el método de mapeo se entrenan para reconstruir los datos de salida a partir de los datos de entrada. Los datos de salida se obtienen de la señal de voz limpia *x(t)*, mientras que los datos de entrada se extraen de la señal de voz mezclada con ruido *y(t)*. En concreto, la red neuronal aprende una función F minimizando la pérdida del error cuadrático medio (MSE) entre el espectrograma de entrada y su entrada reconstruida, como en (4):

$$L_{MSE} = \|Y - F(X)\|^2, \tag{4}$$

o la pérdida de error medio absoluto (MAE) entre la entrada de la señal de voz y su entrada reconstruida, como en (5). Recientemente, se ha avanzado en la resolución de problemas de mejora de la voz en mezclas acústicas monocanal en escenarios cada vez más difíciles, gracias a los métodos de aprendizaje profundo:

$$L_{MAE} = \|Y - F(x)\|. \tag{5}$$

Un tipo de red neuronal muy utilizado en el problema de mejora de la voz es la red neuronal convolucional (CNN, por sus siglas en inglés), que tienen la capacidad de capturar patrones en los fotogramas vecinos mediante un conjunto de conexiones locales.

Se ha reportado de que las redes neuronales convolucionales son más eficaces que las redes neuronales perceptrón multicapa [1] y más eficiente que las redes neuronales recurrentes [11]. De los trabajos donde ha sido relevante el uso de las redes neuronales convolucionales destacan los siguientes.

En [11], Park & Lee demuestran que una red neuronal convolucional puede lograr un mejor rendimiento con una red 12 veces más pequeña que una red neuronal recurrente. La red neuronal convolucional es capaz de tratar las estructuras temporales y espectrales locales de la voz, por lo que es eficaz para separar los elementos de la voz y del ruido de las señales ruidosas.

Las redes neuronales convolucionales ha demostrado su eficacia para mejorar la voz tanto en el dominio de la frecuencia como en el del tiempo (forma de onda). Kinoshita et al. [5] emplearon la eliminación de ruido de la voz basada en la estimación del enmascaramiento utilizando una red neuronal convolucional. Este trabajo fue motivado por el éxito de las redes de convolución temporal para la separación de voz (Conv-TasNet) [19].

Ellos adaptaron la arquitectura de la red para la tarea de reducción de ruido de una señal, que se realiza tanto en el dominio temporal como en el de la frecuencia. En este trabajo los autores también investigaron la pérdida multitarea que predice dos salidas, la voz y el ruido. Además, se propuso una versión ampliada de la red neuronal convolucional usando una red residual (ResNet) [12], con la que se puede conseguir un mejor resultado, ya que la arquitectura de ResNet se ajusta a la tarea de la mejora de voz, que es reconstruir la señal de entrada eliminando la señal ruidosa residual.

La atención es un mecanismo cognitivo de procesamiento de señales de nuestro cerebro. Permite a nuestros cerebros captar eficazmente varias características informativas de los distintos estímulos sensoriales. La fusión de los modelos basados en el aprendizaje profundo y el mecanismo de atención ha ayudado a los modelos a enfatizar las características más informativas y suprimir las menos útiles.

Uno de los mecanismos de atención más utilizados recientemente es la atención de múltiples encabezados (del inglés Multi-Head Attention), que es un módulo que ejecuta varios mecanismos de atención en paralelo [18]. Las salidas de atención independientes se concatenan y se transforman linealmente en la dimensión esperada. Intuitivamente, los encabezados de atención múltiples permiten atender a partes de la secuencia de forma diferente, y se puede expresar como (6):

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, ..., \text{head}_h]W_0, \tag{6}$$

donde (7):

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{7}$$

donde W son todas las matrices de parámetros entrenables. La atención de múltiples encabezados es un módulo que utiliza la atención de producto punto escalado, que es un mecanismo de atención en el que los productos de puntos se escalan de forma $\sqrt{d_k}$.

Formalmente tenemos una consulta $\boldsymbol{Q}$, una clave $\mathbf{K}$ y un valor $\mathbf{V}$, y calculamos la atención como (8):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}. \tag{8}$$

La figura 1 muestra la representación gráfica de la atención de producto-punto escalado y atención de múltiples encabezados, y los detalles se pueden consultar en [18]. Existen trabajos en los que se implementaron las redes neuronales convolucionales con mecanismos de atención exitosamente; por ejemplo, Sun

et al. [15] proponen una red neuronal convolucional recurrente que combine las ventajas de ambas, y optimizar aún más el rendimiento de la separación de una señal de voz con ruido mediante el uso de un mecanismo de atención.

Lan et al. [7] introducen un mecanismo de atención en un modelo con la arquitectura codificador-decodificador convolucional para enfatizar explícitamente la información útil, sus resultados experimentales mostraron que los mecanismos de atención que propusieron pueden emplear una pequeña fracción de parámetros para mejorar eficazmente el rendimiento de los modelos basados en redes neuronales convolucionales en comparación con sus versiones normales, además que su modelo logro generalizar bien a los ruidos no vistos durante el proceso de entrenamiento.

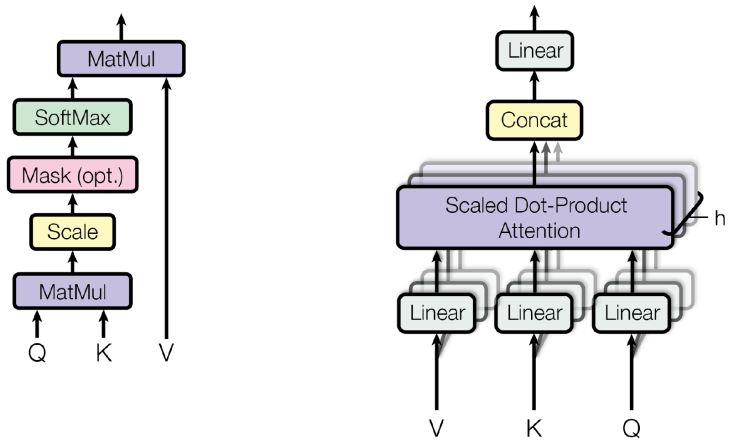

Fig. 1. (Izquierda) Atención de producto-punto escalado. (Derecha) La atención de múltiples encabezados consiste en varias capas de atención que funcionan en paralelo [18].

Motivados por estos trabajos, se propone una red convolucional ramificada con atención para solucionar el problema de la mejora de la voz en el dominio de la frecuencia. La propuesta se basa en una arquitectura convolucional con dos ramificaciones, un módulo de atención y capas densas.

El resto de este artículo se organiza como sigue. En la sección 2 se describe el modelo propuesto. La configuración experimental de los datos y el entrenamiento se presentan en la Sección 3. La sección 4 presenta los resultados obtenidos. Y la sección 5 concluye este trabajo.

2. Descripción del modelo

Se utilizó una red neuronal que se compone de una serie de capas convolucionales y densas, así como un módulo de atención. A continuación, primero se

Tabla 1. Hiperparámetros utilizados en el modelo implementado.

Tipo	Filtros	Nodos	Kernel	Activación	Strides	Dropout	Conexión
1D-Conv_01	64	-	16	PReLu	1	0.1	-
1D-Conv_02	64	-	16	PReLu	1	0.1	-
1D-Conv_03	32	-	16	PReLu	1	0.1	-
1D-Conv_04	2	-	16	PReLu	1	-	-
Rama_01							
1D-Conv_05	64	-	16	PReLu	1	0.1	1D-Conv_04
1D-Conv_06	64	-	16	PReLu	1	0.1	-
1D-Conv_07	32	-	16	PReLu	1	0.1	-
1D-Conv_08	1	-	16	PReLu	1	-	-
Rama_02							
1D-Conv_09	64	-	16	PReLu	1	0.1	1D-Conv_04
1D-Conv_10	64	-	16	PReLu	1	0.1	-
1D-Conv_11	32	-	16	PReLu	1	0.1	-
1D-Conv_12	1	-	16	PReLu	1	-	-
Modulo Atención	-	-	-	-	-	0.1	1D-Conv_08 1D-Conv_12
Dense	-	512	-	ReLu	-	-	-
Dense	-	512	-	ReLu	-	-	-
Lineal	-	256	-	-	-	-	-

describen las operaciones de convolución en la arquitectura ramificada y luego se describe el módulo de atención.

2.1. Red neuronal convolucional ramificada

La arquitectura general de la red neuronal convolucional ramificada (B-CNN, por sus siglas en inglés) se muestra en la figura 2, y los detalles de los hiperparametros utilizados se pueden consultar en la tabla 1. Se construyó un modelo convolucional ramificado que se puede dividir en cinco componentes:

1. Primeramente, la capa de entrada que alimenta este modelo con vectores de tamaño (256, 1).
2. A continuación, un primer bloque compuesto de 4 capas convolucionales con 64, 64, 32 y 2 filtros respectivamente, utilizando un tamaño de kernel = 16, stride = 1, padding = same, PReLu como función de activación. Se empleó dropout = 0.1 únicamente en las 3 primeras capas.
3. Posteriormente se encuentran las dos ramificaciones con configuraciones similares al primer bloque convolucional mencionado; cada una con 64, 64, 32 y 1 filtros respectivamente, usando un tamaño de kernel = 16, stride = 1, padding = same y PReLu como función de activación, y con dropout = 0.1, únicamente en las tres primeras capas.

4. Seguidamente se encuentra el módulo de atención, el cual recibe como datos de entrada dos vectores de tamaño (256,1) provenientes de las dos ramificaciones, los cuales son concatenados antes de ingresar en el módulo de atención.
5. Por último, se encuentran dos capas densas con 512 nodos cada una que emplean ReLu como función de activación, más una capa final de tipo lineal con 256 nodos que genera datos de salida de dimensión (256,).

La tabla 2 resume las dimensiones de los datos de salida de las capas sucesivas en la red propuesta, así como la cantidad de parámetros entrenables de cada capa. Cabe hacer mención que la última capa convolucional del primer bloque (1D-Conv_04) genera datos de salida de dimensión (256, 2).

Esto se debe a que estos datos son divididos para generar dos vectores de dimensiones (256, 1) cada uno, que sirven como datos de entrada para cada una de las dos ramas del modelo propuesto. El inicializador de la matriz de pesos del kernel de todas las capas convolucionales es el inicializador uniforme Glorot (también llamado inicializador uniforme Xavier).

Tabla 2. Dimensiones y cantidad de parametros entrenables del modelo implementado.

Capa (tipo)	Dimencion de salida	Cantidad de Parametros
Input	(None, 256, 1)	-
1D-Conv_01	(None, 256, 64)	17,472
1D-Conv_02	(None, 256, 64)	81,984
1D-Conv_03	(None, 256, 32)	40,992
1D-Conv_04	(None, 256, 2)	1,538
Rama_01		
1D-Conv_05	(None, 256, 64)	17,472
1D-Conv_06	(None, 256, 64)	81,984
1D-Conv_07	(None, 256, 32)	40,992
1D-Conv_08	(None, 256, 1)	769
Rama_02		
1D-Conv_09	(None, 256, 64)	17,472
1D-Conv_10	(None, 256, 64)	81,984
1D-Conv_11	(None, 256, 32)	40,992
1D-Conv_12	(None, 256, 1)	769
Modulo Atención	(None, 512)	273,420
Dense	(None, 512)	262,656
Dense	(None, 512)	262,656
Lineal	(None, 256)	131,328
Total de parámetros entrenables		1,354,480

El código del modelo utiliza internamente capas auxiliares de la librería Keras que no generan parámetros entrenables adicionales para el modelo, sino que

ayudan con la modificación de las dimensiones de los datos, estas capas auxiliares son:

1. Reshape: capa para separar la matriz de la capa Conv1D_04 de dimensión (None, 256, 2) en dos vectores de dimensión (None, 256, 1) utilizadas en cada una de las dos ramificaciones del modelo.
2. Concatenate: capa para unir los datos de salida de las dos ramificaciones con dimensión (None, 256, 1), en un vector con dimensión (None, 512, 1).
3. Flatten: capa para aplanar los datos de entrada de dimensión (None, 512, 1) a (None, 512).

2.2. Modulo de atención

Los métodos de aprendizaje profundo basados en mecanismos de atención alcanzaron el éxito en muchas tareas como la traducción automática [21], el reconocimiento de voz [6] y el procesamiento de imágenes [3]. Los mecanismos de atención son eficaces, ya que pueden ayudar al modelo a obtener mejores resultados mediante la identificación de características importantes. Considerando esto, se construyó e incorporó un módulo de atención para ayudar a identificar las características más importantes para mejorar el rendimiento del modelo convolucional ramificado.

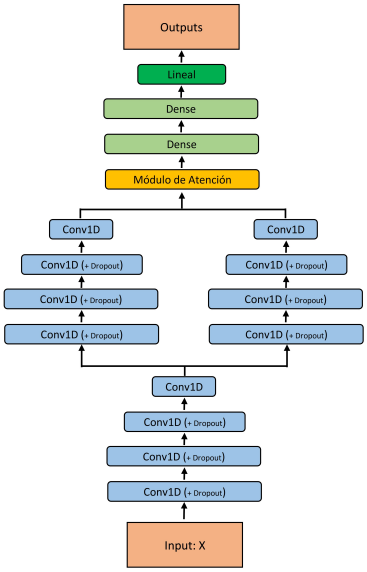

Fig. 2. Diagrama del modelo convolucional ramificado propuesto.

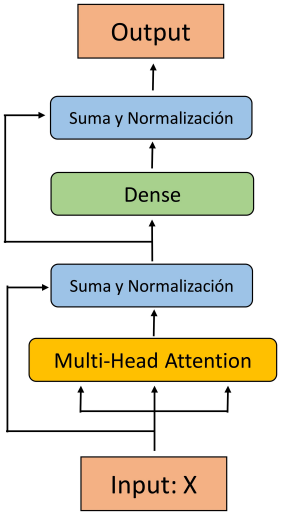

Fig. 3. Diagrama del modulo de atención.

El módulo de atención se basa en el uso de la atención de múltiples encabezados (multi-head attention), que ha resultado más eficaz que otros tipos

de atención. Por lo tanto, se integró la atención de múltiples encabezados en el modelo convolucional ramificado para identificar las características más relevantes. El módulo de atención está compuesto de las siguientes cuatro capas:

1. Una capa de atención de múltiples encabezados, con heads = 4 y dropout = 0,1.
2. Una capa de normalización, la cual suma los datos de salida de la capa de atención con los datos de entrada del módulo de atención, para despues aplicar la normalización con epsilon = 1e-6.
3. Una capa densa con 512 nodos y ReLu como función de activación.
4. Una capa de normalización, la cual suma los datos de salida de la capa densa con los datos de entrada de la capa densa, para despues aplicar la normalización con epsilon = 1e-6.

3. Configuración experimental

3.1. Datos

Se entrenó y evaluó el modelo propuesto realizando mezclas de audio con tres conjuntos de datos; como conjunto de datos de voz se utilizó el TIMIT [2], y como conjunto de datos de ruido se combinó NoiseX-92 [17] y DEMAND [16]. El TIMIT es un conjunto de datos que contiene grabaciones de enunciados de 630 hablantes que representan 8 divisiones dialectales del inglés americano, cada uno de ellos hablando 10 frases con diferente fonética de hablantes masculinos y femeninos.

El material del TIMIT está subdividido originalmente en porciones equilibradas para el entrenamiento y las pruebas (los criterios de subdivisión se describen en [2]). NoiseX-92 es un conjunto de datos compuesto de grabaciones de varios tipos de ruidos acústicos; entre ellos: ruidos de equipos de corte y soldadura eléctrica, ruido blanco, ruidos militares y de vehículos.

DEMAND contiene grabaciones de varios tipos de ruidos acústicos en entornos interiores (domésticos, oficina, público y transporte), y entornos al aire libre (calle y naturaleza). Para el conjunto de entrenamiento y validación se crearon cinco horas de mezclas de audio en clips de un minuto con SNRs uniformemente muestreados entre -10 dB y 10 dB (con lo que la señal de voz se corrompió con los diferentes ruidos).

Tanto los clips de voz como los de ruido fueron elegidos al azar. Posteriormente, se muestrearon los audios a 8 kHz para alimentar el modelo con las bandas de frecuencias más relevantes; se calculó el espectro de potencia de la magnitud de la señal utilizando la transformada de Fourier de corto tiempo (STFT) con un tamaño de 256 FFT, una ventana de longitud de trama de 32 ms (256 muestras), y con solapamiento del 50 % (128 muestras).

Por último, los datos de entrenamiento y validación se normalizaron a media cero y varianza unitaria para facilitar el proceso de entrenamiento. Para evaluar el modelo se crearon mezclas de audio con SNRs uniformemente muestreados de -10 dB, -5 dB, 0 dB, 5 dB y 10 dB. La fase de la señal únicamente se conservó

durante el proceso de predicción del modelo, para luego añadirla a la señal limpia estimada, de forma similar a lo mostrado en [9] y [8].

3.2. Métricas de evaluación

Para evaluar el desempeño del modelo propuesto se utilizaron como métricas de evaluación: la evaluación perceptiva de la calidad de la voz (PESQ) [13]; la inteligibilidad objetiva a corto plazo (STOI) [4]; y la relación señal-distorsión invariable en escala (SI-SNR) [14], que son las métricas estándares más utilizadas para evaluar el desempeño de las propuestas para el problema de mejora de voz. Los valores de PESQ oscilan entre -0.5 y 4.5 (cuanto más alto sea el valor, mejor será la calidad de voz); los valores de STOI suelen oscilar entre 0 y 1 (por lo general se convierte como un porcentaje de inteligibilidad).

3.3. Estrategia de entrenamiento

La estrategia de entrenamiento consistió en el mapeo del espectrograma de magnitud como el objetivo de entrenamiento. Se implementó el modelo convolucional ramificado utilizando la librería Keras con Tensorflow. La función de pérdida usada durante el proceso de entrenamiento fue el Error Cuadrático Medio (MSE), ya que el objetivo fue mejorar todas las métricas de evaluación, no una específica.

Se empleó Adam como optimizador con tasa de aprendizaje = 0.0001, b1 = 0.9, b2 = 0.999 y epsilon=1e-08. Se empleó un tamaño de lote de 64, y se utilizó el 10 % de los datos de entrenamiento para la validación, con el propósito de monitorear y controlar el rendimiento de la red y evitar el sobreajuste. Se eligió la precisión como métrica a monitorizar.

Tabla 3. Resultados de STOI (%), PESQ y SI-SDR de los modelos bajo diferentes ruidos.

Modelo	SNR	STOI (%)	PESQ	SI-SDR
B-CNN (sin atención)	-10	69.62	2.44	6.26
	-5	81.59	2.79	11.47
	0	89.85	3.14	15.96
	5	94.48	3.53	22.17
	10	97.38	3.79	22.90
B-CNN (con atención)	-10	69.71	2.45	6.30
	-5	81.76	2.78	11.51
	0	90.06	3.15	16.09
	5	95.43	3.54	24.20
	10	97.48	3.79	23.34

La duración del entrenamiento se estableció en 50 épocas; y se implementaron dos estrategias: una estrategia de detención anticipada (con lo que

el entrenamiento se detenía si después de 6 épocas consecutivas la métrica monitorizada dejaba de mejorar); así como una estrategia de reducción de la tasa de aprendizaje, con factor = 0.8 y una tasa de aprendizaje mínima = 0.00001 (aplicada cada época que la métrica monitorizada dejaba de mejorar).

4. Análisis y resultados

La tabla 3 muestra los resultados de las tres metricas estándar de mejora de la voz utilizadas habitualmente: la evaluación perceptiva de la calidad de la voz (PESQ), la inteligibilidad objetiva a corto plazo (STOI), y relación señal-distorsión invariable en escala (SI-SDR). Los resultados se basan en cinco niveles de SNR: -10 dB, -5 dB, 0 dB, 5 dB y 10 dB. La evaluación se realizó en el modelo convolucional ramificado con y sin el módulo de atención para contrastar el resultado del impacto del módulo de atención.

En los resultados de la tabla 3 se aprecia que la inclusión del módulo de atención basado en la atención de múltiples encabezados mejoró en términos generales la capacidad del modelo convolucional ramificado, con excepción del valor obtenido con PESQ en SNR de -5 dB. El audio con SNR de 5 dB fue el que mejor resultado mostró al incorporar el módulo de atención, pasando de 94.48 a 95.43 en STOI y de 22.17 a 24.20 en SI-SDR.

Respecto al entrenamiento de los modelos, las figuras (4) y (5) muestran las curvas de pérdida de los datos de entrenamiento y validación para los dos modelos (el modelo sin el módulo de atención y con el módulo de atención) con el fin de mostrar cómo la complejidad afectó al proceso de entrenamiento.

Aunque se estableció la duración del entrenamiento en 50 épocas, la estrategia de detención anticipada detuvo el entrenamiento en la época 38 en ambos modelos, ya que la fue en la época 32 en donde se alcanzó el valor más alto de la métrica monitorizada.

La estrategia de reducción de la tasa de aprendizaje también contribuyó en la rapidez de la convergencia de ambos modelos; en el caso del modelo sin el módulo de atención, la tasa de aprendizaje se modificó en 11 ocasiones, y en el caso del modelo con el módulo de atención, en 12 ocasiones.

Durante el entrenamiento se pudo observar que la incorporación de Dropout en ambos modelos (con y sin el módulo de atención) contribuyó significativamente a evitar el sobreajuste de ambos. Se identificó que, aunque ReLU es la función de activación más comúnmente utilizada; en este modelo, ReLU refleja un mejor rendimiento en las capas densas del modelo, mientras que PReLU es la función de activación con mejor rendimiento para las capas convolucionales, similar a lo mencionado por [10].

También se encontró que la normalización a media cero y varianza uno mejoró el proceso de entrenamiento de ambos modelos (con y sin el módulo de atención), esto es, la precisión mejoró alrededor de 10 % cuando se aplicó la normalización.

En cuanto a los datos utilizados en el proceso de entrenamiento de los modelos convolucionales ramificados, se identificó que se requiere de una gran cantidad de datos para predecir una mejor señal de voz limpia; y en este caso

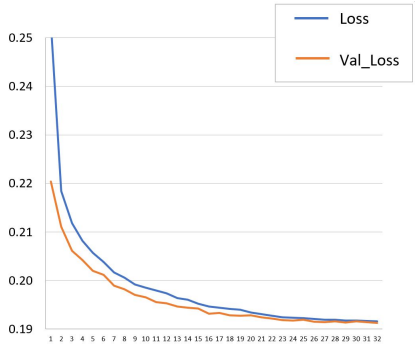

Fig. 4. Gráfica de la curva de pérdida de entrenamiento del modelo sin el módulo de atención para los datos de entrenamiento y validación.

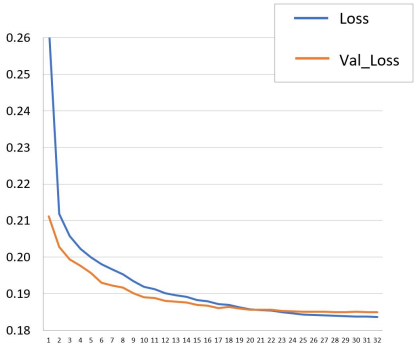

Fig. 5. Gráfica de la curva de pérdida de entrenamiento del modelo con el módulo de atención para los datos de entrenamiento y validación.

se identificó que aproximadamente cinco horas de audio fueron suficientes para que los modelos convergieran.

Usar más de cinco horas no mostró mejora en la velocidad de convergencia del entrenamiento ni en el incremento de su precisión, por lo menos no en estos modelos convolucionales ramificados con menos de 1.36 millones de parámetros entrenables.

5. Conclusiones

Aunque la mejora de la voz basada en el aprendizaje profundo ha demostrado ser muy eficiente al generar una señal de voz limpia con una calidad e inteligibilidad relativamente alta, todavía se considera que algunos entornos de ruido son muy difíciles de tratar para una red neuronal. En este trabajo se incorporó exitosamente un módulo de atención basado en la atención de múltiples encabezados en un modelo de red neuronal convolucional ramificado, el cual mostró mejoras en la tarea de atenuar y eliminar ruido de la voz, tal como se demuestra cuando se evalúa con las métricas STOI, PESQ, SI-SDR.

En un futuro se planea continuar con la investigación en mejora de voz, incorporando otras arquitecturas convolucionales como las codificador-decodificador, o modificar los datos para procesarlos como matrices en vez de vectores, e incluso incorporar una mayor variedad de señales de ruidos no estacionarios para permitir que el modelo se pueda evaluar en ambientes más complejos.

Agradecimientos. Los autores agradecen al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONACYT, por los recursos computacionales, el apoyo y la asistencia técnica brindados, a través del proyecto No. 202103086N.

Referencias

1. Fu, S., Tsao, Y., Lu, X., Kawai, H.: Raw waveform-based speech enhancement by fully convolutional networks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 6–12 (2021)
2. Garofolo, J. S., Lamel, L .F., Fisher, W .M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., Zue. V.: TIMIT acoustic-phonetic continuous speech corpus (1993)
3. Glaser, T., Ben-Baruch, E., Sharir, G., Zamir, N., Noy, A., Zelnik-Manor, L.: PETA: Photo albums event recognition using transformers attention (2021)
4. Jensen, J., Taal, C. H.: An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 24, pp. 2009–2022, IEEE Press (2016)
5. Kinoshita, K., Ochiai, T., Delcroix, M., Nakatani, T.: Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7009–7013, IEEE (2020)
6. Kipyatkova, I.: End-to-end russian speech recognition models with multi-head attention. Speech and Computer, pp. 327–335, Springer International Publishing (2021)
7. Lan, T., Lyu, Y., Ye, W., Hui, G., Xu, Z., Liu, Q.: Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement. IEEE Access, vol. 8, pp. 78979–78991 (2020)
8. Li, L., Lu, Z., Watzel, T., Kürzinger, L., Rigoll, G.: Light-weight self-attention augmented generative adversarial networks for speech enhancement. Electronics, vol. 10 (2021)
9. Nossier, S. A., Wall, J., Moniri, M., Glackin, C., Cannings, N.: An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics, vol. 10 (2021)
10. Nossier, S., Wall, J., Moniri, M., Glackin, C., Cannings, N.: An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics, vol. 10 (2021)
11. Park, S., Lee, J.: A fully convolutional neural network for speech enhancement. Proc. Interspeech, pp. 1993–1997 (2017)
12. Plantinga, P., Bagchi, D., Fosler-Lussier, E.: An exploration of mimic architectures for residual network based spectral mapping. IEEE Workshop on Spoken Language Technology (2018)

13. Rix, A. W., Beerends, J. G., Hollier, M. P., Hekstra, A. P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 749–752 (2001)
14. Roux, J., Wisdom, S., Erdogan, H., Hershey, J. R.: SDR - half-baked or well done?. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 626–630 (2019)
15. Sun, C., Zhang, M., Wu, R., Lu, J., Xian, G., Yu, Q., Gong, X., Luo, R.: A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. Scientific Reports, vol. 11, pp. 1–14 (2021)
16. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In: Proceedings of Meetings on Acoustics ICA2013, vol. 19 (2013)
17. Varga, A., Steeneken, H. J .M.: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. speech communication, vol. 12, pp. 247–251 (1993)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
19. Yi L., Nima M.: Conv-TasNet: Surpassing ideal time frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, pp. 1256–1266 (2019)
20. Yuliani, A. R., Faizal, M., Suryawati, E., Ramdan, A., Ferdinandus, H.: Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, vol. 21, pp. 19–26 (2021)
21. Zhang, T., Huang, H., Feng, C., Cao, L.: Enlivening redundant heads in multi-head self-attention for machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3238–3248, Association for Computational Linguistics (2021)