UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Dereverberación del habla a partir de algoritmos de aprendizaje profundo†

Autor: Martin Bernardo Meza
Tutores: Ing. Leonardo Pepino

(†) **Tesis para optar por el titulo de ingeniero/a de Sonido**.

Octubre 2020

Índice

Índice de figuras

Índice de tablas

1. Introducción

1.1. Fundamentación

Las tecnologías que explotan el procesamiento digital de señales de voz mostraron grandes avances en las ultimas décadas, llegando a ocupar roles de primera importancia en nuestro día a día. Las investigaciones realizadas en este campo fueron impulsando diversas aplicaciones basadas en el análisis de la voz humana [1][2]. Estas tareas, en mayor o menor medida, deben lidiar con una característica intrínseca a cualquier emisión sonora dentro de un recinto: la reverberación. Las señales de voz que reciben las aplicaciones anteriormente nombradas por lo general se obtienen a partir de un transductor que no siempre se encuentra cercano a la fuente que desea registrar, provocando que la señal resultante capte la reverberación propia del entorno de origen de la señal. Esta reverberación interfiere en detrimento la señal de voz, produciendo una reducción en el rendimiento de aquellas aplicaciones que dependen de la integridad de dicha señal, como ser:

- Reconocimiento del habla [3]
- Verificación del hablante[1] [4]
- Localización del hablante [5]
- Inteligibilidad de la palabra

Si bien esta problemática fue abordada desde el enfoque de diversas técnicas de procesamiento de señales, en los últimos años este campo de estudio tuvo grandes avances producto de la implementación de una tecnología emergente de amplio crecimiento en el ambiente científico: los algoritmos de aprendizaje profundo. La capacidad y robustez que esta técnica demostró a la hora de resolver problemas pertinentes al procesamiento de imágenes y detección de patrones frente a los enfoques clásicos la pusieron al frente de las herramientas utilizadas para resolver problemas de este ámbito. Sin embargo, las tareas relacionadas al procesamiento de audio aun son un campo de estudio reciente para estas tecnologías, en donde todavía se presentan obstáculos para lograr una implementación plena de estas técnicas como por ejemplo: la falta de bases de datos masivas de señales acústicas, la selección de una manera de representación óptima de las señales que permita explotar sus características intrínsecas, las maneras de medir el rendimiento de los procesos, entre otros.

Por este motivo, esta investigación pretende realizar un análisis de esta problemática desde el punto de vista de la ingeniería de sonido, para comprender las limitaciones de los modelos actualmente utilizados en este campo de estudio, y poder aportar al progreso y mejora del rendimiento de dichos modelos.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de esta investigación es implementar un algoritmo de dereverberación de señales de voz a partir del uso de redes neuronales y algoritmos de aprendizaje profundo.

[1] Debe distinguirse entre reconocimiento del habla y verificación del hablante. Lo primero refiere a poder distinguir que palabras fueron dichas, y lo segundo refiere a identificar quien es el que esta pronunciando las palabras.

1.2.2. Objetivos específicos

Los objetivos específicos son

- Realizar una revisión de las técnicas utilizadas para resolver el problema de dereverberación.
- Diseñar e implementar una estructura de red neuronal para dereverberación de señales de voz en lenguaje Python.
- Analizar las técnicas de pre y post procesamiento de datos, estudiando el impacto que tienen en el rendimiento del algoritmo.
- Optimizar el sistema propuesto, y comparar los resultados obtenidos con los modelos actuales de manera objetiva.
- Diseñar e implementar una interfaz gráfica en donde se permita visualizar los efectos del proceso de dereverberación aplicados a una señal particular.

1.3. Estructura de la Investigación

2. Estado del Arte

En los últimos años se ha registrado un marcado desarrollo y progreso en el campo de el procesamiento de señales del habla. En este campo, la dereverberación ocupa un rol crucial debido a que es una característica que influye en la mayoría de las aplicaciones del procesamiento de señales del habla.

Los primeros enfoques que apuntaron a resolver el problema de la dereverberación fueron aquellos referidos a las respuestas al impulso y la estimación de filtros inversos a partir de estas [6]. Como el efecto de la reverberación en una señal se puede pensar como el resultado de una convolución entre una señal anecoica y una respuesta al impulso, este enfoque apunta a estimar la respuesta al impulso con el fin de poder generar un filtro inverso que permita realizar una deconvolución de la señal para poder revertir el efecto de la respuesta del recinto, recuperando la señal en su estado anecoico. Sin embargo esta metodología presenta varios inconvenientes, como el hecho de considerar que las respuestas al impulso son lineales e invariantes en el tiempo, lo cual no siempre se cumple en la práctica [7], o bien el hecho de que la respuesta no siempre pueda ser deducida de manera directa y deba ser estimada.

También surgieron enfoques basados en los modelos matemáticos de la generación del habla [8]. Se implementaron algoritmos que se basaban en el estudio de la señal de residuo obtenida luego de la predicción lineal del habla. Se detectó que esta señal residuo contenía información sobre los efectos de la reverberación, por lo cual se la utilizó para excitar filtros variantes en el tiempo que al aplicarse sobre la señal del habla mostraban una mejora respecto a la eliminación de los efectos reverberantes [9]. También se realizaron análisis de dereverberación a partir del uso de varios transductores, aplicando técnicas de descomposición sobre el conjunto de señales captadas [10]. Otras características propias de la señal del habla fueron explotadas en pos de eliminar los efectos de la reverberación, tales como la estructura armónica [11], y el espectro de modulación [12].

Posteriormente, se aplicó la idea de la sustracción espectral [13] [14] que básicamente consiste en la estimación del espectro de potencia generado por la reverberación a partir de modelos estadísticos. En 2006, Wang et. al. aplicaron este enfoque combinado con el de la estimación de filtros inversos logrando presentar avances importantes en la efectividad de los algoritmos [15].

A partir del año 2006 en el campo del estudio de la separación de fuentes se popularizó un enfoque al problema denominado Análisis Computacional de la Escena Auditiva [16] que está inspirado en la teoría perceptiva del Análisis de la Escena Auditiva [17] la cual intenta explicar la capacidad del sistema auditivo de descomponer una señal captada en varias señales correspondientes a diferentes fuentes de sonido. Este enfoque trajo consigo el uso de máscaras binarias ideales en el dominio temporal-frecuencial para extraer las señales buscadas [18]. Las máscaras se definen como ideales ya que su obtención requieren del conocimiento de la señal buscada y de la señal que interfiere. El uso de estas máscaras implica primero realizar una transformación de la señal de entrada de manera de trasladarla al dominio tiempo-frecuencia (por ejemplo un espectrograma, o un cocleograma) y luego asignarle a cada punto del espacio temporal-frecuencial un valor de 1 cuando su energía pertenece a la señal objetivo, y un valor de 0 en el caso contrario. Roman et. al. [19] aplicaron este concepto para tratar el problema de dereverberación, donde se busca estimar la máscara binaria ideal tomando como señal objetivo la señal del habla en condiciones anecoicas y como interferencia a la parte reverberante. Para conseguir la dereverberación, este método requiere seleccionar de manera correcta parámetros como el punto desde el cual se distingue la parte temprana y tardía de la reverberación,y el nivel del umbral en base al cual se identifica a un punto específico como parte de la señal deseada o de la interferencia [20]. Hazrati et al. [21] propusieron estimar la máscara binaria a partir de un parámetro dependiente de la varianza de la señal, la cual define un umbral adaptativo, obteniendo mejores resultados.

Los primeros antecedentes de la implementación de redes neuronales en la tarea de la dereverberación se encuentran desde el año 2007. Jin y Wang [22] aplicaron la estructura de perceptrón multicapa para estimar las mascaras binarias necesarias para la separación de la componente reverberante en una señal voz. La estructura de red neuronal debía realizar el mapeo entre características extraídas de la señal de entrada y cada unidad temporal-frecuencial de la señal de salida. Mas adelante, con el avance de los modelos de aprendizaje profundo, esta técnica se iría perfeccionando reflejándose en mejores resultados en la tarea de dereverberación. En 2014 Kun et al. [23] proponen el uso de redes neuronales profundas para aprender el mapeo espectral de señales reverberantes hacia señales anecoicas. Esto quiere decir, en otras palabras, que se entrena una red neuronal profunda para que sea capaz de estimar el espectro anecoico de una señal reverberante. Nuevamente se vuelve al planteo de la búsqueda del filtro inverso que permita la deconvolución de la señal reverberante para obtener su versión anecoica, pero en este caso será la red quien aprenda la forma de ese filtro inverso. Entonces, esto se logra entrenando una red que tiene como entrada el espectro de la señal reverberante y como salida (es decir, como objetivo) el espectro de la señal anecoica. Se implementan soluciones desde el post-procesamiento para lograr reconstruir la fase de la señal estimada. Por otro lado, Weninger et al. [24] implementaron redes neuronales recurrentes bidireccionales de larga memoria de corto término en la tarea de la dereverberación en pos de conservar la continuidad del habla. Luego, se pasan a utilizar redes neuronales convolucionales [25]. Estas ofrecen una mayor habilidad de modelado, permitiendo considerar patrones locales presentes en la representación

temporal-frecuencial. Además, utilizan menos parámetros y distribuyen de manera mas eficiente los pesos sinápticos, lo que se traduce en un menor costo computacional de procesamiento. Globalmente, los modelos de redes convolucionales logran ser mas eficientes y mas precisos en sus resultados. Por lo general se utilizan estructuras secuenciales, formando estructuras denominadas codificadores-decodificadores, en las cuales la información temporal-frecuencial de entrada sufre una compresión que disminuye su dimensión derivándose a un espacio latente, para luego a partir de este espacio poder estimar el espectro objetivo que corresponde a la señal dereverberada. Como este tipo de redes consideran cada punto de tiempo-frecuencia en un contexto de pocos puntos contiguos, el siguiente paso fue implementar redes completamente convolucionales [26], en las cuales se contempla el conjunto completo de puntos de tiempo-frecuencia. Este último presentó mejores resultados que los modelos mas acotados. Entre los estudios mas recientes, se encuentran enfoques que proponen la combinación de métodos utilizados previamente que demostraron un buen funcionamiento en algún aspecto para lograr que en conjunto logren minimizar el error que producen por separado [27].

3. Marco Teórico

3.1. Representación temporal y frecuencial de señales

Fundamentalmente, el audio esta compuesto por formas de ondas. Cuando un objeto vibra genera ondas de presión que cuando alcanzan nuestros oídos son percibidas como sonido [28]. Si bien entonces podemos definir a una señal de audio como una variación continua, para su análisis digital nos interesa estudiar este tipo de señales en un dominio discreto. Para lograrlo, se toman muestras equiespaciadas de la señal continua, en un proceso que se denomina muestreo. La distancia temporal entre dos muestras contiguas es determinado según la frecuencia máxima que se desea representar, acorde al teorema de Nyquist [29]. Entonces, una señal continua $x_a(t)$ que es muestreada a una frecuencia de $f_s = 1/T_s$ muestras por segundo, produce la señal discreta $x(n)$ que se puede definir a partir dela señal continua a partir de la ecuación 1, que equivale a la representación vectorial vista en la ecuación 2, siendo N el número de muestras tomadas.

$$x(n) = x_a(nT_s) \tag{1}$$

$$x_n = [x(0), x(1), ..., x(N-1)]^T \tag{2}$$

Partiendo de esta representación temporal de la señal, se puede obtener una representación frecuencial de la misma a partir de la Transformada Discreta de Fourier (DFT) [29] que matemáticamente equivale a la expresión 3, lo cual es útil para poder realizar un análisis mas profundo de la señal. La DFT permite entonces representar a la señal a partir de componentes frecuenciales complejas. Es decir que para cada punto se tiene un valor de amplitud y un valor de fase. A su vez, esta transformación supone un proceso reversible, por lo cual la señal temporal puede ser recuperada partiendo de la señal frecuencial, aplicando la expresión 4.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-jnk2\pi/N} \qquad 0 \leq k \leq N-1 \tag{3}$$

$$x[n] = 1/N \sum_{k=0}^{N-1} X[k] e^{jnk2\pi/N} \qquad 0 \leq n \leq N-1 \tag{4}$$

El cálculo de esta transformación es costoso en términos computacinoales, y hay un alto grado de redundancia en este proceso. Por esto, comúnmente se utiliza una implementación definida como transformada rápida de Fourier que permite optimizar el cómputo de esta transformación [30].

Cuando se trabaja con señales no estacionarias es de interés evaluar la variación del espectro de frecuencias en el tiempo. Para esto se utiliza una transformación denominada transformada de Fourier de corto plazo (STFT por sus siglas en inglés) la cual consiste en una representación tridimensional formada al calcular la transformada de Fourier para sub-intervalos temporales de la señal, y luego representarlos de manera contigua [28]. De esta manera se obtiene un gráfico con dimensiones de tiempo, frecuencia y amplitud en donde se puede ver la evolución del espectro en función del tiempo. Un ejemplo de un espectrograma donde se conserva solo la magnitud se puede ver en la Figura 1.

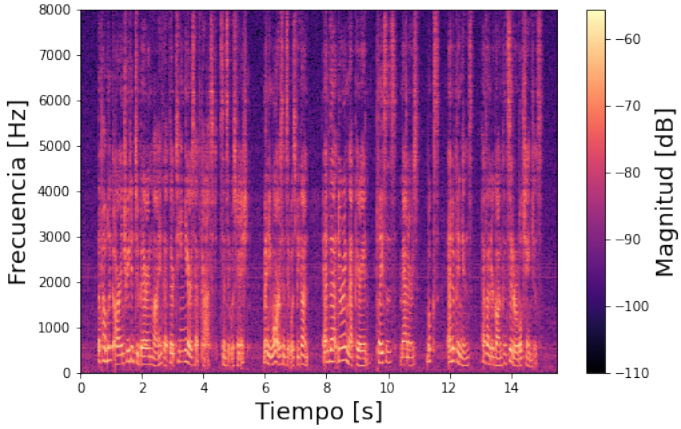

Figura 1: Espectrograma de una señal de audio

3.2. Respuesta al impulso y reverberación

Si en un recinto se tiene una fuente y un micrófono captando a una cierta distancia de la fuente, las ondas sonoras que emite la fuente se reflejarán en las paredes del recinto y alcanzarán el micrófono inmediatamente después que la onda sonora directa. Las reflexiones continúan ocurriendo, y cada instancia de reflexión supone una disminución de la energía sonora de la onda, principalmente causada por el efecto de absorción acústica de las superficies que producen las reflexiones. En un determinado tiempo, la energía sonora decaerá en todo el recinto hasta ubicarse por debajo del ruido de fondo. A este proceso se lo denomina reverberación. Al camino mas corto entre la fuente y el punto de captura se denomina camino directo, y a la relación de nivel entre la presión sonora que genera la onda propia del camino directo y la presión que genera el efecto de reverberación se lo conoce como relación directo-reverberado.

Si el micrófono se ubica cerca de la fuente va a captar en mayor medida la señal correspondiente al camino directo, y una pequeña porción del sonido reverberado. Es decir, una relación

directo-reverberado alta. A medida que el punto de captura se aleja de la fuente va a captar una menor cantidad del sonido correspondientemente al camino directo, mientras que el campo reverberado se mantendrá aproximadamente invariante. Esto se traduce en una disminución de la relación directo-reverberado.

De esta manera, habrá una distancia específica para la cual el nivel de presión sonora generado por la fuente sera igual al nivel de presión sonora generado por el efecto de la reverberación. Esta distancia se conoce como distancia crítica. Esta depende tanto de las condiciones del recinto como de las características del micrófono.

La función de transferencia entre la fuente emisora y el micrófono se define como la respuesta al impulso del recinto y usualmente se denota como $h(t)$. Este será diferente para cualquier punto en el espacio dentro del recinto. Haciendo un análisis temporal de una respuesta al impulso, podemos identificar 3 partes: en primer lugar el nivel de sonido directo (producido por la onda que viaja a través de camino directo), las reflexiones tempranas (cuyo limite temporal vendrá definido por las características propias de cada recinto) y por último la cola reverberante. Esto se ve representado en la Figura 2. Se puede distinguir la parte de reflexiones tempranas y la cola reverberante partiendo de la suposición de que las reflexiones tempranas ocurren en un proceso determinístico, siendo altamente sensibles a pequeños cambios en la geometría del recinto, mientras que la cola reverberante es mas bien un proceso estocástico, y al depender de un mayor número de reflexiones no varia drásticamente frente a pequeños cambios de geometría.

Idealmente, el micrófono captura una señal que corresponde a la convolución entre la respuesta al impulso del recinto y la señal fuente, como se ve en la ecuación 5. Esto equivale a una multiplicación en el dominio de la frecuencia de acuerdo con la transformada de Fourier, como se ve en la ecuación 6.

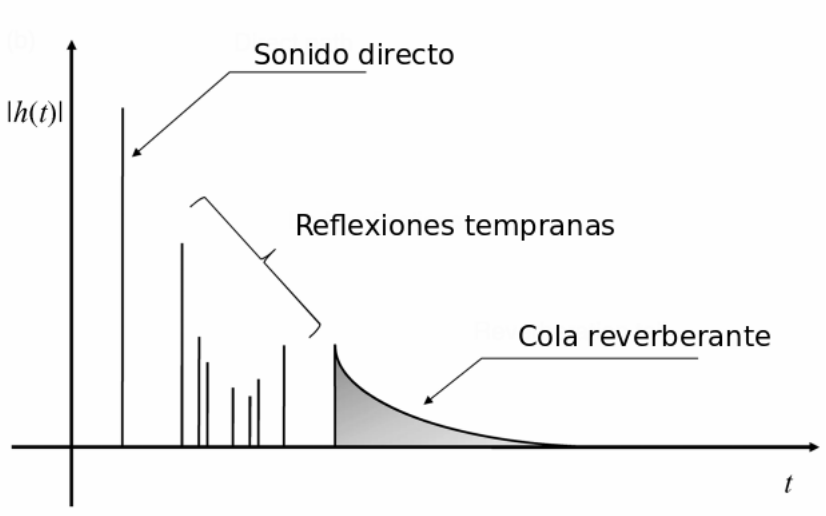

Figura 2: Secciones temporales de una respuesta al impuulso

$$x(t) = h(t) * s(t) \tag{5}$$

$$X(f) = H(f)S(f) \tag{6}$$

De esta manera se puede ver que la respuesta al impulso conserva toda la información sobre la influencia de la reverberación del recinto sobre la señal captada por el micrófono.

3.3. Inteligibilidad y parámetros de calidad de percepción

Para caracterizar la señal del habla propagándose en condiciones reverberantes se utilizan métricas objetivas derivadas de la respuesta al impulso del recinto en cuestión, como por ejemplo el tiempo de reverberación o la relación energética entre la señal directa y el campo reverberado. En cambio, al considerar el proceso de dereverberación de estas señales, las respuestas al impulso requieren ser estimadas, lo que usualmente conduce a una caracterización de baja calidad. Además, los algoritmos de dereverberación pueden introducir artefactos audibles a la señal voz, los cuales no son contemplados por las respuestas al impulso estimadas. Es por esto que es preciso utilizar métodos de medida de calidad basados en la señal dereverberada. Las pruebas subjetivas son el método mas confiable para evaluar la calidad percibida de una señal de habla dereverberada. Sin embargo, este método es costoso y requiere mucho tiempo, por lo cual se vuelve inviable su aplicación para procesamientos en tiempo real. Para aplicaciones prácticas se definieron entonces métodos objetivos de medición de calidad basados en la señal dereverberada como reemplazo de las pruebas subjetivas. Estos métodos consisten en algoritmos que de manera objetiva y repetible buscan estimar la calidad percibida de la señal, por lo cual, un método resulta efectivo cuando logra obtener una alta correlación con las respuestas subjetivas. Estos métodos se clasifican en intrusivos o no intrusivos, dependiendo de si requieren o no una señal de referencia para realizar la estimación. Poder contar con una señal de referencia para realizar estas estimaciones es usualmente una dificultad, por lo cual se presta mayor interés en aquellos métodos no intrusivos.

3.3.1. Relación energía de modulación de voz a reverberación

Este parámetro de medida de calidad para señales dereverberadas se basa en obtener características de la reverberación partiendo del espectro de modulación de la señal [31]. La formulación de este parámetro se basa en el hecho de que la cola reverberante de cualquier respuesta al impulso puede ser modelada como ruido blanco Gaussinano exponencialmente amortiguado. Esta característica puede ser explotada en el análisis del espectro de modulación de la señal bajo análisis para obtener descriptores del efecto de la reverberación.

3.3.2. Inteligibilidad objetiva de corto termino extendida

Este parámetro está basado en características extraídas a partir de la correlación de corto término entre la señal limpia y la señal procesada. Es aplicable para evaluar aquellos procesos que realizan transformaciones no lineales [32]. Su funcionamiento se basa en aplicar una ventana de análisis de $384\ ms$ en las envolventes de amplitud de las subbandas de la señal analizada. Estas ventanas temporales se aplican en pos de contemplar frecuencias de modulación que son relevantes para la inteligibilidad. En estos lapsos temporales se calculan coeficientes de correlación espectrales que son luego promediados. De esta manera, este parámetro puede ser interpretado en términos de una descomposición ortogonal de espectrogramas energéticamente normalizados que son luego ordenados de acuerdo a su contribución a la inteligibilidad estimada.

3.3.3. Relación señal a distorsión

Este descriptor fue ampliamente utilizado en tareas de separación de fuentes y refuerzo de señales de habla. Esta basado en el cómputo de la relación señal a interferencia (SIR), y en la relación señal a artefacto (SAR) [33]. En las tareas de dereverberación, estas medidas pueden ser interpretadas como proporcionales a la supresión de componentes reverberantes tardías e inversamente proporcionales a la distorsión en la señal del habla, respectivamente. Contemplando estos valores, el parámetro final contempla la calidad general de la señal dereverberada.

3.4. Redes neuronales y algoritmos de aprendizaje

La base de los algoritmos de aprendizaje profundo se encuentra en la neurona artificial. Esta consiste en un modelo que parte de los principios de funcionamiento de las neuronas biológicas [34]. El perceptrón [35] fue de las primeras arquitecturas formalmente implementadas y que se considera la unidad básica de estos algoritmos. Un esquema de una neurona artificial básica se puede ver en la Figura 3. Las componentes básicas de una neurona artificial son:

- **Entradas**: Recibe los datos que van a ser procesados en esta unidad.
- **Pesos sinápticos**: Parámetros de ponderación. Cada entrada se asocia a uno de estos parámetros. Es el valor que se va ajustando cuando el modelo se encuentra en la etapa de entrenamiento. En ellos se ve reflejado la propagación del error.
- **Suma ponderada**: Los pesos sinápticos se asocian a cada entrada a partir de una regla de propagación que consiste en una suma ponderada.
- **Función de activación**: Función que se aplica a la salida de la suma ponderada, cuya salida representa la salida final de la unidad. Esta se determina de manera de poder agregar complejidad al modelo. Algunos ejemplos de funciones de activación se pueden ver en la figura 4.
- **Salida**: Es el resultado de aplicar el proceso completo al conjunto de entradas. En una estructura, esta salida puede ser la entrada de una o varias unidades subsiguientes.

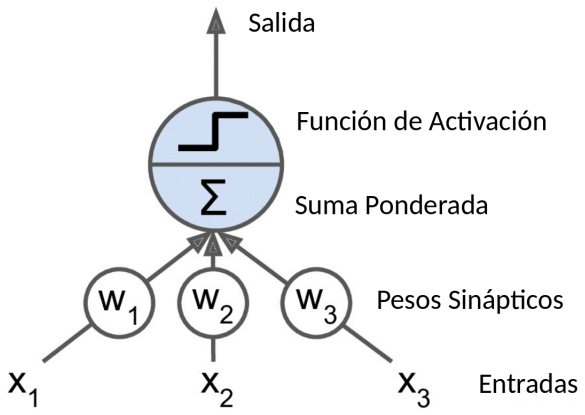

Figura 3: Esquema de neurona artificial

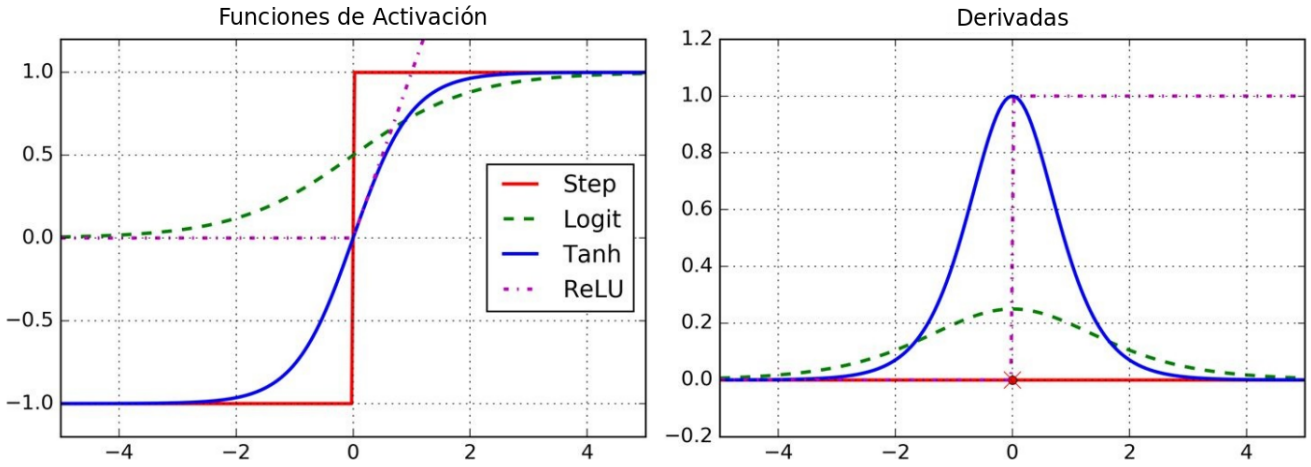

Figura 4: Funciones de activación y sus derivadas

3.4.1. Modelos basados en redes neuronales

Los modelos basados en redes neuronales son sistemas compuestos por capas que agrupan unidades computacionales (neuronas artificiales). En una capa, las entradas y salidas de las neuronas artificiales que la componen están agrupadas. Diferentes capas se relacionan formando sistemas de acuerdo al problema que se busque resolver. En general, un sistema se compone por una capa de entrada, una capa de salida, y un número finito de capas ocultas intermedias. De esta forma, el sistema recibe valores de entrada, los procesa a través de las distintas capas que componen la red, y otorga valores de salida. Un esquema de este funcionamiento se puede ver en la Figura 5.

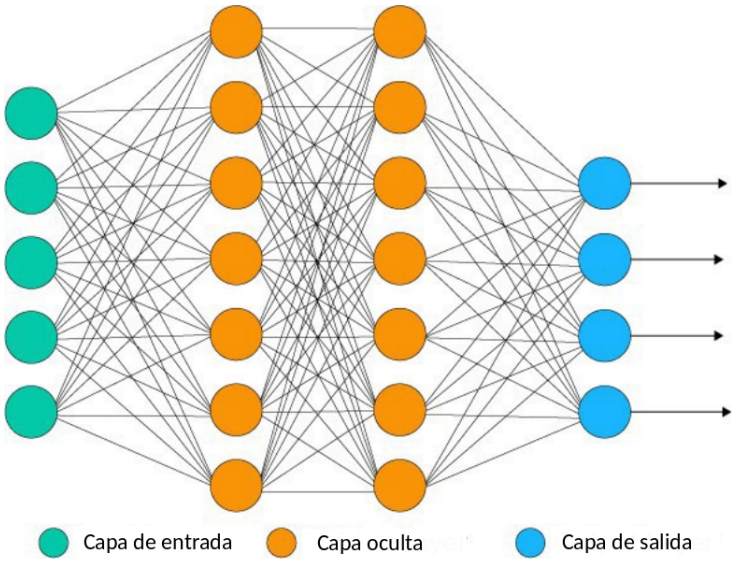

Figura 5: Esquema básico de red neuronal

Catalogar a estos sistemas como de aprendizaje 'profundo' hace referencia al hecho de tener sucesivas capas de representación [36]. Cuando un modelo se compone de un mayor numero

de capas, se lo considera mas 'profundo'.

Estas estructuras de redes neuronales aprenden automáticamente al ser expuestas a un conjunto de datos. El proceso de aprendizaje se puede pensar como la evaluación de un mapeo de valores de entrada a ciertos valores objetivos de salida. Esto es, se toman valores de entrada, se los transforman a lo largo de las capas que componen la red produciendo valores de salida, y se comparan estas salidas con los valores de salida objetivos. Entonces, la especificación del proceso que está siendo implementado por el sistema se encuentra reflejado en los pesos sinápticos de las neuronas que componen cada capa. El aprendizaje se obtiene a partir de poder modificar estos pesos sinápticos acorde a las diferencias que se obtengan entre las salidas producidas por la red y las salidas que se tienen como objetivo. Para lograr esto último, los algoritmos de redes neuronales utilizan determinadas funciones:

- **Función de costo**: También denominada función objetivo, o función de pérdida. Recibe las salidas de la red y las salidas esperadas y evalúa que tanto difieren entre sí. Estas diferencias las traduce a una medida de distancia a partir de una expresión matemática que se define en función del problema que se busca resolver. Entonces, para cada estimación de la red, esta función otorga un puntaje que explica cuan lejos está el valor estimado del valor pretendido.

- **Función de optimización**: Esta función aplica el algoritmo de propagación del error hacia atrás, que es una parte fundamental de un algoritmo de aprendizaje profundo. Este cálculo permite estimar el aporte que tiene cada peso sináptico en el error final de la estimación de la red, y por lo tanto permite ajustar los valores de estos pesos sinápticos estratégicamente para conseguir minimizar la distancia computada por la función de costo.

Entonces, la salida de la función de costo se utiliza como realimentación del sistema a través de la función de optimización. De este modo, el entrenamiento consiste en un bucle en el cual en cada iteración el sistema evalúa una instancia de los datos de entrenamiento (valores de entrada y de salida), y ajusta los pesos sinápticos en pos de reducir el error calculado. Un esquema que expone este funcionamiento se ve en la Figura 6. Repetir este ciclo un número suficiente de veces conduce a la convergencia del valor de error.

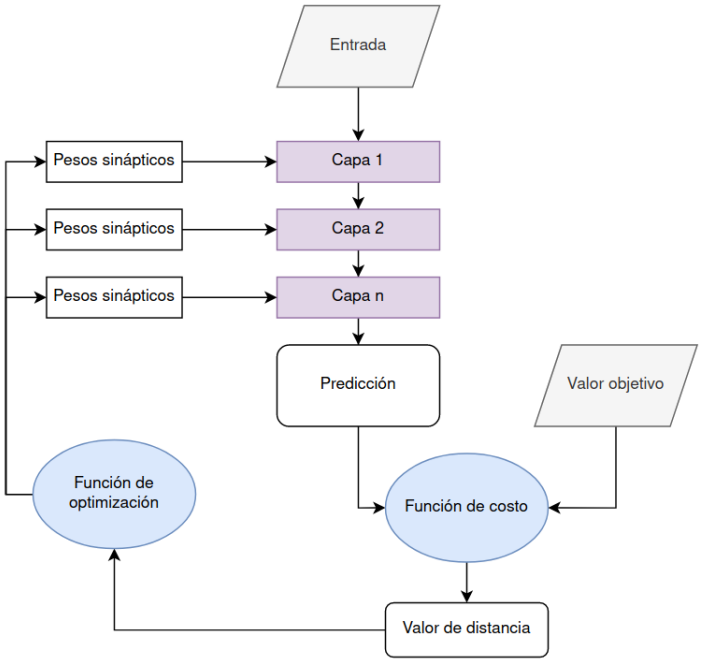

Figura 6: Diagrama de flujo del bucle de entrenamiento de una sistema de red neuronal

Por último, la manera en que el sistema de red neuronal recibe y procesa los datos también influye en el desempeño de la misma. El objetivo final del sistema es alcanzar un grado de generalización que le permita procesar adecuadamente instancias de datos que no hayan sido reveladas ante la red en la etapa de entrenamiento. Por esto, el conjunto total de los datos se divide en tres subgrupos:

- **Conjunto de entrenamiento**: Este conjunto de datos es el que se utiliza en la etapa de entrenamiento para optimizar los parámetros de la red. Aquí se concentra el mayor volumen de datos.
- **Conjunto de validación**: Sobre este conjunto se mide el desempeño del sistema a lo largo de su entrenamiento. Los resultados obtenidos del procesamiento de este conjunto sirven para ajustar variables que requieren ser especificadas de manera previa al entrenamiento. Estos parámetros se denominan hiper parámetros.
- **Conjunto de prueba**: Este conjunto es el que se utiliza para medir el rendimiento final del sistema. Como contiene instancias que no fueron utilizadas en las etapas de entrenamiento y ajuste de parámetros, el análisis del procesamiento de este conjunto sirve para medir el nivel de generalización que el sistema logró alcanzar.

El conjunto de datos de entrenamiento se segmenta en lotes. En cada iteración de entrenamiento la red neuronal recibe un lote, lo procesa, aplica la función de costo y ajusta los pesos sinápticos de cada capa. Cuando la red procesó todos los lotes que componen el conjunto de datos de entrenamiento se dice que transcurrió una época. El proceso de entrenamiento depende en cierta medida del tamaño de los lotes [36]. Si consideramos la curva de la función de

costo de un parámetro como la de la Figura 7, vemos que existen mínimos locales y mínimos globales a lo largo de la misma. En el proceso de entrenamiento se busca minimizar este valor de costo. Si se toman lotes muy pequeños, lo que se traduce en desplazamientos pequeños a lo largo de esta curva, se corre el riesgo de quedar confinado en un mínimo local. De igual manera, un conjunto demasiado grande produciría saltos demasiado grandes en comparación a las fluctuaciones de esta curva, haciendo que se obtengan valores de costo aleatorios.

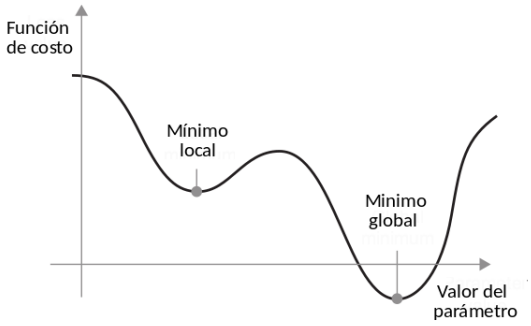

Figura 7: Curva de costo para un parámetro

3.4.2. Redes neuronales convolucionales

Las redes neuronales convolucionales emergen del estudio de la corteza visual del cerebro. En los últimos años, estas estructuras fueron utilizadas para resolver tareas visuales complejas (análisis de imágenes). El componente principal es la capa convolucional. Para el análisis de imágenes, estas capas se concatenan de manera que la primera capa no contempla cada píxel de la imagen, sino que solo se enfoca un número acotado de píxeles que caen dentro de su campo perceptivo. De igual manera, las capas subsiguientes se enfocan en las salidas de un conjunto acotado de neuronas de la capa precedente. Este funcionamiento se ilustra en la Figura 8.

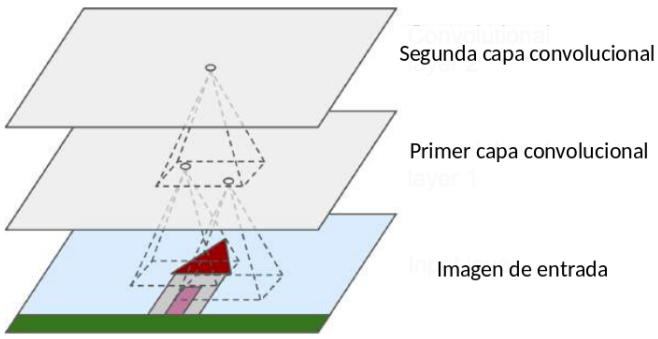

Figura 8: Capas convolucionales con campos receptivos locales rectangulares

Formar esta estructura le permite a la red aprender diferentes patrones estructurales locales de manera jerárquica [37]. Estas capas se distinguen por dos propiedades fundamentales:

- Los patrones que se aprenden son invariantes al desplazamiento. Esto quiere decir, que si se aprende de un patrón ubicado en un lugar específico de una imagen, este mismo patrón puede ser identificado en cualquier otra ubicación dentro de la imagen.

- Cuando estas capas se concatenan formando redes logran aprender jerarquías espaciales de patrones. Esto les permite aprender de manera eficiente conceptos visuales cada vez más complejos y abstractos.

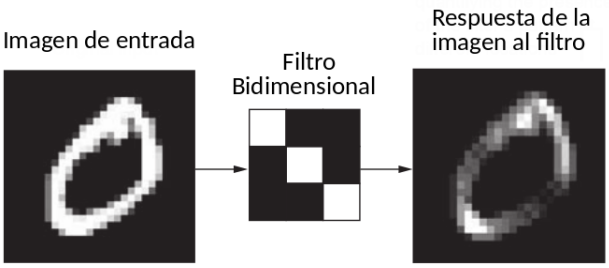

Figura 9: Representación de la aplicación de un filtro bidimensional sobre una imagen

El procesamiento que ocurre en una capa convolucional consiste en la aplicación de uno o varios filtros bidimensionales sobre la imagen de entrada, lo cual genera mapas de respuestas que representan la presencia del patrón del filtro a lo largo de la imagen, como se puede apreciar en la Figura 9. En este tipo de capas, el aprendizaje se traduce en determinar la forma de los filtros que se deben aplicar para conseguir los resultados esperados. Teniendo en cuenta este proceso, las variables que se deben definir en cada capa son:

- **Tamaño del filtro**: Define el tamaño del campo perceptivo de cada unidad de procesamiento de la capa. Valores comunes son $3x3$ o $5x5$. En la Figura 10 se ve un ejemplo de un filtro de tamaño $3x3$.

- **Tamaño del salto**: Determina la distancia horizontal y vertical entre campos perceptivos de dos unidades contiguas. Hacer que este valor sea mayor a uno permite reducir las dimensiones de la imagen de entrada al atravesar la capa convolucional. Esto se puede apreciar en la Figura 10 en donde se aplica un tamaño de salto igual a dos (tanto en sentido vertical como horizontal).

- **Relleno de ceros**: Cuando se pretende mantener invariables las dimensiones de entrada y salida de una capa convolucional, se suele aplicar un relleno con ceros en los contornos de la imagen. La cantidad de ceros agregados dependerá de las características del filtro a aplicar. Aplicar un relleno de ceros produce un fenómeno denominado efecto de borde [37].

- **Cantidad de filtros aplicados**: El número de filtros computados por la convolución.

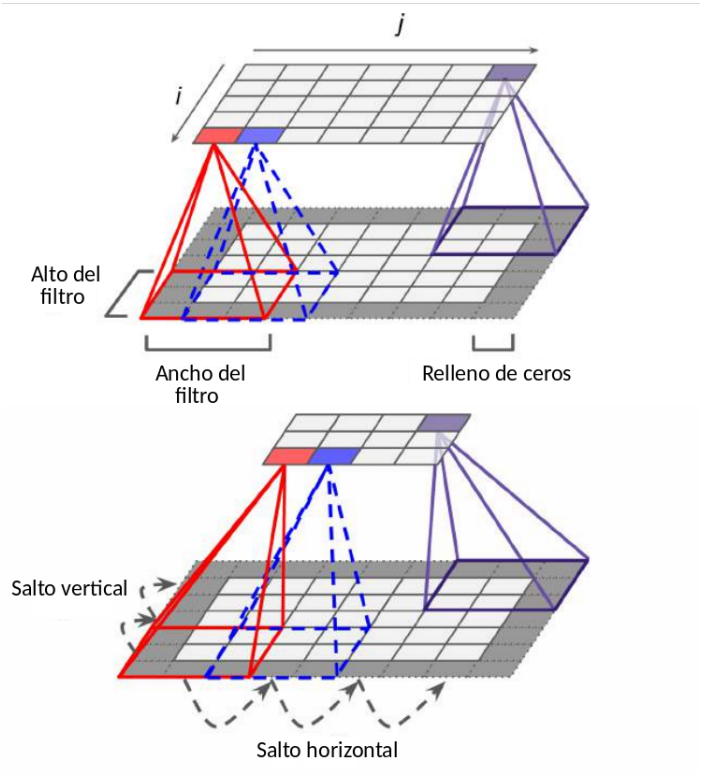

Figura 10: Parámetros de procesamiento en capas convolucionales

De esta manera, las capas convolucionales permiten construir modelos con menos cantidad de parámetros a entrenar, distribuidos adecuadamente para conseguir la generalización de conceptos visuales complejos.

4. Metodología

4.1. Análisis de datos

4.2. Sistema propuesto

5. Resultados

5.1. Influencia del armado de datos

5.1.1. Variables del sistema

6. Discusión de los resultados

7. Conclusiones

8. Lineas futuras de investigación

Bibliografía

[1] G. Hinton L. Deng y B. Kingsbury. “New types of deep neural network learning for speech recognition and related applications: An overview”. En: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process (ICASSP)* (2013).

[2] H. Chung y col. “Noise-adaptive deep neural network for single-channel speech enhancement”. En: *Proc. Int. Workshop on Machine Learning for Signal Process. (MLSP)* (2018).

[3] J. Pearson y col. “Robust distant-talking speech recognition”. En: *IEEE International Conference on. Vol. 1* (1996). Acoustics, Speech, and Signal Processing.

[4] Pierre J. Castellano, S. Sradharan y David Cole. “Speaker recognition in reverberant enclosures”. En: *IEEE International Conference on. Vol. 1* (1996). Acoustics, Speech, and Signal Processing.

[5] Joseph H. DiBiase, Harvey F. Silverman y Michael S. Brandstein. “Robust localization in reverberant rooms”. En: *Springer Berlin Heidelberg* (2001). Microphone Arrays.

[6] Masato Miyoshi y Yutaka Kaneda. “Inverse Filtering of Room Acoustics”. En: *IEEE Transactions on Acoustics, Speech and Signal Processing* (1988).

[7] Stephen T. Neely y Jont B. Allen. “Invertibility of a room impulse response”. En: *Acoustics Research Department, Bell Laboratories, Murray Hill, New Jerse* (1997).

[8] Laurence R. Rabiner y W. Schafer Ronald. *Digital Speech Processing*. Prentice- Hall, 1975.

[9] B. Yegnanarayana y P. Satyanarayana. “Enhancement of Reverberant Speech Using LP Residual Signal”. En: *IEEE Transactions on Speech and Audio Processing* (2000).

[10] S. Gannot y M. Moonen. “Subspace Methods for Multimicrophone Speech Dereverberation”. En: *EURASIP journal on advances in signal processing* (2003).

[11] N. Roman y D. L. Wang. “Pitch-based monaural segregation of reverberant speech”. En: *Journal of Acoustical Society of America* (2006).

[12] M. Avendano y H. Hermansky. “Study on the dereverberation of speech based on temporal envelope filtering”. En: *Proc. of ICSLP* (1996).

[13] J. M. Boucher K. Lebart. “A New Method Based on Spectral Subtraction for SpeechDereverberation”. En: *Acta Acustica united with Acustica* (2001).

[14] K. Lebart, J.-M. Boucher y P. Denbigh. “A new method based on spectral subtraction for speech dereverberation”. En: *Acta Acustica united with Acustica* (2001).

[15] M. Wu y D. L. Wang. “A two-stage algorithm for one-microphone reverberant speech enhancement”. En: *IEEE Transactions on Acoustics, Speech and Signal Processing* (2006).

[16] D. L. Wang y Eds. G. J. Brown. “Computational Auditory Scene Analysis: Principles, Algorithms, and Applications”. En: *Proc. of ICSLP* (1996).

[17] A. S. Bregman. “Auditory Scene Analysis”. En: *Cambridge, MA: MIT Press* (1990).

[18] D. Wang. *On ideal binary mask as the computational goal of auditory scene analysis*. Kluwer, 2005, págs. 181-197.

[19] John Woodruff Nicoleta Roman. “Intelligibility of reverberant noisy speech with ideal binarymasking”. En: *Journal of Acoustical Society of America* (2011).

[20] John Woodruff Nicoleta Roman. “Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold”. En: *Journal of Acoustical Society of America* (2013).

[21] O. Hazrati, J. Lee y P. C. Loizou. “Blind binary masking for reverberation suppression in cochlear implants”. En: *Journal of Acoustical Society of America* (2013).

[22] Z. Jin y D. L. Wang. “A supervised learning approach to monaural segregation of reverberant speech”. En: *IEEE Transactions on Acoustics, Speech and Signal Processing* (2009).

[23] DeLiang Wang Kun Han Yuxuan Wang. “Learning spectral mapping for speech dereverberation”. En: *IEEE International Conference on Acoustic, Speech and Signal Processing* (2014).

[24] F. Weninger y col. “Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition”. En: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2014).

[25] W. Shi D. S. Wang Y. X. Zou. “A Deep Convolutional Encoder-Decoder Model for Robust Speech Dereverberation”. En: *22nd International Conference on Digital Signal Processing* (2017).

[26] Ori Ernst y col. “Speech Dereverberation Using Fully Convolutional Networks”. En: *22nd International Conference on Digital Signal Processing* (2017).

[27] Chunlei Liu, Longbiao Wang y Jianwu Dang. “Deep Learning-Based Amplitude Fusion for Speech Dereverberation”. En: *Discrete Dynamics in Nature and Society* (2020).

[28] Joshua D. Reiss y Andrew P. McPherson. *Audio Effects: Theory, Implementation and Application*. CRC Press, 2014.

[29] Alan V. Oppenheim, Ronald W. Shafer y John R. Buck. *Discrete-Time Signal Processing*. Prentice- Hall, 1989.

[30] E. Oran Brigham. *The Fast Fourier Transform And Its Aplications*. Prentice- Hall, 1988.

[31] Tiago H. Falk y Wai-Yip Chan. “A Non-intrusive Quality Measure of Dereverberated Speech”. En: *Department of Electrical and Computer Engineering* (2008). Queen’s University, Kingston, Ontario, Canada.

[32] Tiago H. Falk y Wai-Yip Chan. “A Non-intrusive Quality Measure of Dereverberated Speech”. En: *Department of Electrical and Computer Engineering* (2008). Queen’s University, Kingston, Ontario, Canada.

[33] Emmanuel Vincent, Rémi Gribonval y Cédric Févotte. “Performance Measurement in Blind Audio Source Separation”. En: *IEEE Transactions on Audio, Speech and Language Processing* (2006).

[34] Chunlei Liu, Longbiao Wang y Jianwu Dang. “A Logical Calculus of Ideas Immanent in Nervous Activity”. En: *Bulletin of Mathematical Biophysics* (1943).

[35] Frank Rosenblatt. “The Perceptron: A Perceiving and Recognizing Automaton”. En: *Cornell Aeronautical Laboratory* (1957).

[36] Francois Chollet. *Deep Learning with Python*. Manning Publications, 2017.

[37] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'reilly media, 2019.