UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

De-reverberación del habla a partir de algoritmos de aprendizaje profundo†

Autor: Meza, Martin Bernardo
Tutores:

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

Septiembre 2020

Plan de Tesis

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. 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 baja en el rendimiento de aquellas aplicaciones que dependen de la integridad de dicha señal, como ser:

- Reconocimiento del habla [1]
- Verificación del hablante[1] [2]
- Localización del hablante [3]
- Aumento de la 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, este trabajo 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] 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. Objetivos

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

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 [4]. 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 [5], 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 [6]. 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 [7]. También se realizaron análisis de dereverberación a partir del uso de varios transductores, aplicando técnicas de descomposición sobre el conjunto de señales captadas [8]. Otras características propias de la señal del habla fueron explotadas

en pos de eliminar los efectos de la reverberación, tales como la estructura armónica [9], y el espectro de modulación [10].

Posteriormente, se aplicó la idea de la sustracción espectral [11] [12] que básicamente consiste en la estimación del espectro de potencia generado por la reverberación a partir de modelos estadísticos. Este enfoque combinado con el de la estimación de filtros inversos logró presentar avances importantes en la efectividad de los algoritmos [13].

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 (CASA por sus siglas en inglés)[14] que está inspirado en la teoría perceptiva del Análisis de la Escena Auditiva (ASA por sus siglas en inglés)[15] 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 [16]. 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. Este concepto se aplicó entonces para tratar el problema de dereverberación [17], 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 especifico como parte de la señal deseada o de la interferencia [18]. Hazrati et al. [19] 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 and Wang [20] 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.[21] proponen el uso de redes neuronales profundas (DNN por sus siglas en inglés) para aprender el mapeo espectral de señales reverberantes hacia señales anecoicas. Esto quiere decir, en otras palabras, que se entrena una DNN 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 sera 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. [22] 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 [23]. Estas

ofrecen una mayor habilidad de modelado, permitiendo considerar patrones locales presentes en la representación temporal-frecuencial. Ademas, 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 dimension 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 [24], en las cuales se contempla el conjunto completo de puntos de tiempo-frecuencia. Este ultimo presento 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 [25].

3. Marco Teórico

3.1. Representación temporal y frecuencial de señales

3.2. Respuesta al impulso y reverberación

Si en un recinto tengo una fuente y un micrófono captando a una cierta distancia de la fuente, las ondas sonoras que emite la fuente se reflejaran en las paredes del recinto y alcanzaran el micrófono inmediatamente después que la onda sonora directa. IMAGEN. 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 (RIR por sus siglas en inglés) y usualmente se denota como $h(t)$. Este sera 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 ultimo la cola reverberante. 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.

$$x(t) = h(t) * s(t) \quad (1)$$

3.3. Inteligibilidad y parámetros de calidad de percepción

3.4. Redes neuronales y algoritmos de aprendizaje

4. Diseño de la Investigación

Esta investigación busca estudiar el rendimiento de algoritmos de aprendizaje profundo en la tarea de la segregación de las componentes reverberantes de una señal de voz. Para conseguirlo, se propone organizar la metodología en las siguientes etapas:

Etapa 1: Revisión del estado del arte y planteo de objetivos

Lo primero será estudiar detalladamente los documentos precedentes en este campo específico de aplicación. Se debe prestar especial atención a los trabajos mas recientes para poder determinar claramente cuales son los puntos débiles que presentan las metodologías actuales y de ese modo dilucidar de que manera se puede aportar al problema en cuestión y desde que enfoque conviene hacerlo. Con esto, se busca tener una mayor perspectiva sobre el problema para poder determinar con mayor eficiencia los objetivos pretendidos a corto y largo plazo.

Etapa 2: Determinar las herramientas a utilizar

Con los objetivos en mente, el siguiente paso es determinar con que herramientas desarrollar los algoritmos para poder resolver cada paso de la cadena de procesamiento requerida. Existiendo tanta variedad de opciones, se debe analizar detalladamente que elección resulta mas fructífera para la implementación de los procesos pretendidos y para el posterior análisis de dicho proceso. Esto incluye tanto las herramientas de software o frameworks, tanto como los datos o bases de datos a utilizar.

Etapa 3: Análisis y Pre-procesamiento de los datos

Habiendo elegido las herramientas y los datos a utilizar, el siguiente paso sera realizar un análisis exploratorio de dichos datos. Este análisis tiene como finalidad lograr comprender de una

mejor manera las características de los datos con los que se va a trabajar, y determinar cuales son los procesamientos previos que se deben aplicar para que estos datos sean útiles a la hora de alimentar el algoritmo de aprendizaje profundo. Probablemente esta etapa sea la de mayor importancia en cuanto a que es el campo en donde se puede realizar un mayor aporte desde el punto de vista de la Ingeniería de Sonido. En esta etapa se deberá determinar con que datos alimentar el algoritmo y de que manera (o que características) serán estos introducidos a la cadena de procesamiento. Los resultados de este análisis y las decisiones que se deriven de estos marcaran el rumbo de la investigación.

Etapa 4: Modelado y puesta a punto de la/las estructuras de aprendizaje profundo a utilizar, junto con su posterior entrenamiento
En esta etapa se realiza la implementación del algoritmo de aprendizaje profundo. Consiste tanto en armar la estructura de red elegida previamente, como también el resto de la cadena de procesamiento (previo y posterior a la red). Se deben definir las métricas a utilizar para cuantificar el rendimiento de los algoritmos, las funciones de optimización para la propagación del error, y otros parámetros relativos al algoritmo.

Etapa 5: Análisis de los resultados de cada modelo, validación y comparación con los modelos existentes.
En esta etapa de la investigación se busca hacer un contraste con el trabajo realizado y los antecedentes existentes, de manera de valorar los resultados obtenidos y ponerlos en contexto. Se deben elegir adecuadamente los parámetros a comparar y deben considerarse todas las decisiones tomadas a lo largo del procesamiento para poder abordar a conclusiones representativas sobre el aporte de la investigación al campo de estudio.

Etapa 6: Armado de interfaz gráfica
Por último, la etapa restante consiste en el armado de una herramienta que permita visualizar el funcionamiento del algoritmo propuesto. Esta etapa también contempla el armado de una función global que permita utilizar un modelo entrenado para ser insertado en una cadena de procesamiento posterior, para ser utilizado en otras aplicaciones como el de-noising.

4.1. Cronograma

En el siguiente diagrama de Grantt se propone el cronograma de actividades para la realización de esta investigación.

Actividad / Quincenas	1	2	3	4	5	6	7	8	9	10	11	12
Revision de estado del arte y planteo de objetivos	X	X	X									
Determinación de herramientas		X	X									
Análisis y pre-procesamiento de datos				X	X	X	X					
Modelado de las estructuras de aprendizaje profundo							X	X				
Análisis y validación de resultados									X	X	X	
Desarrollo de Interfaz de visualización de datos											X	X

Bibliografia

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

[2] 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.

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

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

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

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

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

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

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

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

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

[12] 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).

[13] 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).

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

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

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

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

[18] 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).

[19] 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).

[20] 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).

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

[22] 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).

[23] 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).

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

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