UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

FILTRADO ESPACIAL DE RUIDO ADITIVO PARA INTERACCIÓN HUMANO ROBOT

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

RODRIGO MANUEL MAHU SINCLAIR

PROFESOR GUÍA:
NESTOR BECERRA YOMA

MIEMBROS DE LA COMISIÓN:
SANDRA CÉSPEDES UMAÑA
JOHN ATKINSON ABUTRIDY

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
POR: RODRIGO MANUEL MAHU SINCLAIR
FECHA: 2022
PROF. GUÍA: NESTOR BECERRA YOMA

FILTRADO ESPACIAL DE RUIDO ADITIVO PARA INTERACCIÓN HUMANO ROBOT

En el contexto de la interacción humano robot la comunicación mediante voz es de gran importancia. Por esto se requiere mejorar el desarrollo técnicas que permitan un reconocimiento robusto de voz en ambientes ruidosos. Estas técnicas pueden utilizar las capacidades típicas de un robot como son los múltiples sensores audiovisuales.

En este trabajo se aborda el uso de múltiples beamforming apuntando a las fuentes de voz y ruido. En conjunto con técnicas de separación de canal para obtener un mejor reconocimiento. Esto es posible utilizando robots con arreglos de micrófonos e información visual para detectar la localización de las fuentes.

Las señales de beamforming para fuentes de voz y ruido se consideran como conocidas y se aborda el uso de técnicas de separación de canal y deep learning para mejorar el desempeño del reconocedor de voz.

Se presenta un modelo para abordar la separación de la mezcla de ruido y voz. Esta se compara con otras técnicas de separación de señales y técnicas de deep learning para eliminar ruido.

Se genera una base datos que represente el problema abordado y se muestran resultados utilizando redes neuronales implementadas en TensorFlow. Las señales resultantes son evaluadas utilizando un sistema de reconocimiento de voz en el estado del arte.

El sistema propuesto presenta un desempeño similar a las técnicas en el estado del arte, requiriendo una una menor ventana de análisis.

A mis padres

Agradecimientos

Quisiera agradecer a mis padres Carolina y Francisco por su cariño y sus enseñanzas a lo largo de mi vida y por su apoyo y paciencia estos últimos años sin los cuales esta tarea hubiese sido imposible.

A mi hermano Javier por su compañía, cariño y apoyo.

A mis amigos de Club de Kung Chung Iy Tang por mantenerme cuerdo.

A mi profesor Néstor Becerra Yoma le agradezco la oportunidad de trabajar en el LPTV y por su guía constante.

Y a mis compañeros del LPTV por toda la ayuda y conocimiento que me han brindado.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

La colaboración entre humanos y robots será un componente estratégico en las aplicaciones civiles y de defensa en los próximos 10 o 20 años. Hay varias aplicaciones en defensa, entornos hostiles, minería, industria, silvicultura, educación y desastres naturales donde se requerirá la integración y colaboración entre humanos y robots [1]-[3]. Esto se debe a que el paradigma del robot totalmente autónomo pierde fuerza y la comunicación efectiva entre humanos y robots gana relevancia. La comunicación similar a la humana entre una persona y un robot es esencial para la interacción exitosa humano-robot (HRI) y la simbiosis colaborativa entre humano y robot [4]-[8]. Para este fin, el perfil del usuario y la adaptación del robot en los dominios físicos, cognitivos y sociales son cruciales [9]. Además, el habla es la forma más directa y sencilla en que los humanos se comunican [10] y debería ser la forma más natural de hacer posible una simbiosis colaborativa entre humano y robot.

La interacción social es un desafío muy complejo en el contexto de la robótica, en parte porque exige reconocer o detectar efectivamente las direcciones de la mirada, las expresiones faciales, el contenido lingüístico y la prosodia de la voz. Dependiendo del contexto cultural, la diferencia entre los estados emocionales humanos puede ser tan sutil como "un simple guiño o una creciente inflexión en un solo fonema" [11].

La colaboración entre humano y robot requiere una interacción social adecuada entre humanos y robots, lo que a su vez es un gran desafío. A pesar de que se han realizado algunos avances, la interacción social entre humano y robot bajo diversas condiciones naturales y limitaciones del mundo real no es posible hoy en día. Para lograr este propósito, los sistemas tendrán que combinar la entrada multimodal de diversos sensores e integrar esta información perceptiva con modelos predicativos de intenciones sociales [11]. En la figuras 1.1 se muestra un ejemplo en el cual el robot recibe información multimodal de múltiples personas y una fuente de ruido auditivo en el aire acondicionado cuya posición espacial puede determinarse utilizando información visual. En la figura 1.2 se observa un ejemplo donde el robot recibe información multimodal de multiples personas en un ambiente al aire libre.

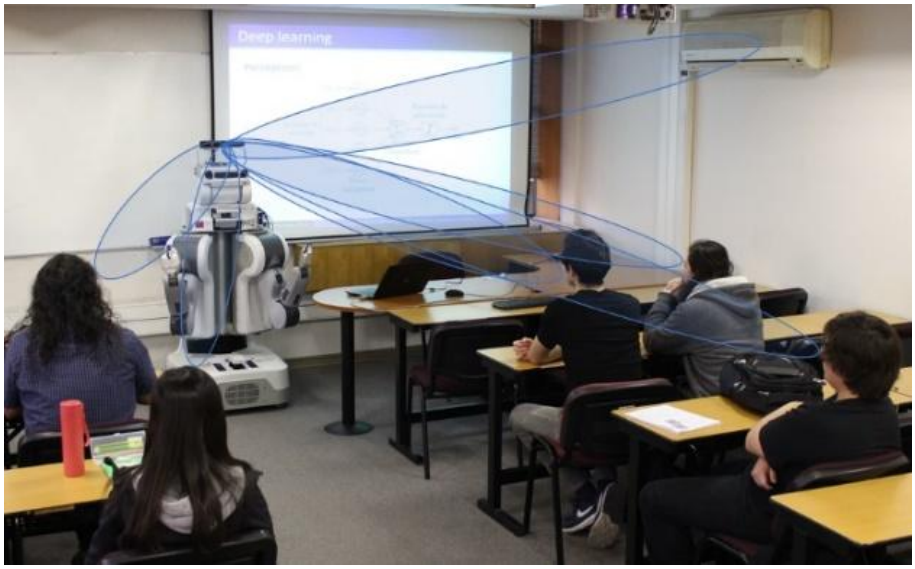

Figura 1.1: Escenario de HRI en interiores.

Figura 1.2: Escenario de HRI al aire libre.

Por entrada multimodal en este contexto, entendemos toda la información observada de usuarios humanos que pueden ser capturados o adquiridos por robots: habla, imagen y señales fisiológicas. Observe que estos modelos de interacción social dependen del contexto, la cultura y las normas morales. Los robots sociales necesitarán empatizar con los humanos, comprender la propiedad y cumplir la promesa de generar confianza y relaciones a largo plazo. También se requieren modelos de conocimiento, creencias, propósitos y emociones de las personas para que la comunicación entre humanos y robots sea más efectiva [9], [12], [13].

Para lograr una colaboración efectiva con las personas, los robots deben detectar y establecer un perfil para los usuarios con los que están interactuando, y modificar y adaptar su comportamiento de acuerdo con los modelos aprendidos que tienen sobre el usuario. HRI

requiere modelar y reconocer las acciones y capacidades humanas, revelar las intenciones y objetivos detrás de tales acciones, y determinar los parámetros que caracterizan la interacción social. Si los robots comprenden mejor a los usuarios, pueden adaptar su comportamiento con respecto a las características y preferencias de los usuarios, y mejorar la satisfacción del usuario y la aceptación del robot. En este punto, vale la pena resaltar que el perfil y la adaptación conductual pueden tener lugar y analizarse desde los puntos de vista de interacción física, cognitiva y social [9].

Perfil físico: este dominio comprende las características del usuario que están relacionadas con el cuerpo humano y los movimientos en el espacio. En consecuencia, la creación de perfiles de usuarios corresponde a la detección de las capacidades de movimiento relacionadas con el proceso de interacción y los movimientos previstos en el espacio. La capacidad de reconocer y modelar actividades humanas físicas define una tecnología habilitadora para lograr aplicaciones de HRI efectivas. La actividad física se puede clasificar en cuatro niveles diferentes, gestos, acciones, interacciones y actividades grupales [14]. Los gestos son movimientos básicos de la parte del cuerpo de una persona (por ejemplo, una mano que saluda) que podrían interpretarse también como comandos proporcionados al robot [15]. Las acciones son actividades realizadas por una persona que pueden estar compuestas de varios patrones elementales organizados secuencialmente como "caminar" [16]. Las interacciones corresponden a actividades que implican contacto físico entre dos o más personas u objetos [17]. Un ejemplo de interacción podría ser dos personas peleando. Finalmente, las actividades grupales son actividades físicas realizadas por grupos compuestos por varias personas que no requieren necesariamente contacto físico (por ejemplo, un grupo que tiene una reunión o un grupo que visita un museo) [18]. La investigación actual sobre la clasificación de la actividad humana generalmente sigue un enfoque de reconocimiento de patrones para los movimientos de los usuarios en el espacio. La idea es extraer información y características diferentes de los datos del sensor y utilizar el aprendizaje automático para detectar y clasificar patrones de actividad. Algunos estudios han empleado solo señales fisiológicas (ej., Frecuencia cardíaca, frecuencia respiratoria, temperatura de la piel, ECG) o una combinación de acelerómetros y datos fisiológicos (ej., Frecuencia cardíaca) para realizar el reconocimiento de actividad. Sin embargo, el uso de sensores portátiles puede ser incómodo y rechazado por el usuario [19].

Perfiles cognitivos: más allá de la clasificación de las actividades físicas humanas observables que resultan de la interacción con el mundo externo, existe la necesidad de predecir, detectar y reconocer las intenciones del agente observado [20]. La capacidad de inferir y reconocer las intenciones, los deseos, las creencias, los estados internos, la personalidad y las emociones de los individuos a menudo se conoce como Teoría de la Mente [21]. Los seres humanos muestran una habilidad natural para comunicarse e identificar su estado interno utilizando códigos no verbales como el lenguaje corporal, los gestos y las expresiones faciales. También pueden adaptar sus comportamientos en respuesta a su comprensión y predicción de las intenciones de los demás haciendo uso de acciones observables. Al interactuar con un humano, un robot debería ser capaz de generar una representación de sus acciones. Por ejemplo, una sola acción física y su significado pueden estar relacionados con otras acciones y con el posible objetivo del usuario detrás de la interacción misma. La identificación de intenciones, objetivos y planes son capacidades de asistencia para una colaboración proactiva entre humanos y máquinas basadas en comandos implícitos del usuario [22]. La mayoría de los estudios sobre reconocimiento de intenciones abordaron el problema de recuperar la intención detrás

de la comunicación verbal. Por ejemplo, en [23], [24], el problema del modelado del usuario (es decir, la comprensión de las creencias, objetivos y planes del usuario) se abordó en los sistemas de diálogo de inteligencia artificial, y la importancia de tal se resaltó el modelado de la interacción. En [25], la capacidad del robot para rastrear y actualizar las creencias de sus interlocutores en una interacción verbal se abordó proponiendo un esquema adaptativo de revisión y expresión de creencias. Enfoques probabilísticos para predecir eventos e intenciones futuros, principalmente mediante el uso de modelos ocultos de Markov (HMM) [26], máquinas de estado finito [27], [28], Bayesian Networks [29], [30] también se han propuesto. Cabe destacar que la personalidad también caracteriza el estado cognitivo de un usuario y es un aspecto clave de las interacciones sociales humanas. Los estudios de psicología han sugerido que existe una relación directa entre la personalidad y el comportamiento [31], [32]. Según [32], la personalidad se puede definir como "el patrón de carácter colectivo, rasgos de comportamiento, temperamental, emocional y mental de un individuo que tiene consistencia en el tiempo y las situaciones" [15]. Los estudios sobre psicolingüística se han centrado en los marcadores de personalidad en el lenguaje para clasificar los comportamientos de la personalidad a partir de observaciones indirectas. Por ejemplo, los extravertidos tienden a hablar más alto y más iterativamente con menos dudas y pausas que las personas introvertidas. En [33], las características prosódicas (tono vocal, energía y velocidad de conversación) se utilizan para clasificar automáticamente los rasgos de personalidad con más del 70 % de precisión.

Perfil social: para las interacciones sociales exitosas con las personas, los robots deben reconocer e interpretar las señales sociales que muestra un ser humano [34]. Las señales sociales pueden definirse como comportamientos observables que producen cambios de comportamiento durante la interacción [35]. Observe que en el perfil cognitivo el estado cognitivo del usuario puede ser representado por sus creencias y objetivos, o personalidad. Con respecto al perfil social, el enfoque se centra en el modelo y la identificación de señales sociales y preferencias relacionadas con el proceso de HRI. La elaboración de perfiles sociales también denota la posibilidad de reconocer fenómenos sociales (es decir, compromiso, conflicto, empatía, interés y emociones) que no pueden observarse directamente pero deben inferirse a través del análisis de señales indirectas. En una interacción cara a cara entre humanos, se pueden emplear varias modalidades para coordinar o suavizar la interacción: postura corporal, gestos, mirada, vocalización y expresiones faciales. En consecuencia, el reconocimiento de estas señales puede usarse para mejorar la HRI [36], [37]. En consecuencia, el procesamiento de señal social (SSP) gana relevancia en HRI. El objetivo principal del SSP es el análisis automático de señales verbales y no verbales para reconocer fenómenos sociales como la empatía, el conflicto, el interés, las actitudes, el dominio, el coqueteo, la atención, la cortesía o el acuerdo [35]. A diferencia de la personalidad que es una característica estable en el tiempo, las emociones dependen del tiempo y deben identificarse durante la interacción [9]. Las emociones del usuario pueden emplearse para caracterizar su estado cognitivo, pero también dependen de la interacción. La transferencia de las emociones humanas podría lograrse empleando la entrada visual (es decir, el reconocimiento de la expresión facial, el lenguaje corporal y el tacto), la entrada de audio (es decir, el análisis del tono) o las señales fisiológicas.

La adquisición de información relevante de los usuarios humanos es un problema difícil en aplicaciones reales que puede ser altamente dinámico y dependiente del tiempo. La creación de perfiles de usuario se puede realizar en los dominios físicos, cognitivos o sociales mediante el uso de entradas multimodales. Sin embargo, hay varias limitaciones prácticas. Por ejemplo,

como se mencionó anteriormente, el empleo de muchos sensores portátiles puede generar información muy relevante sobre la condición emocional o física de los humanos que interactúan con el robot, pero puede ser incómodo e invasivo desde el punto de vista de los usuarios. El reconocimiento de emociones basado en imágenes depende en gran medida de las condiciones de iluminación y del ángulo de la cara del usuario con respecto a las cámaras del robot. La voz transmite una gran cantidad de información lingüística y paralingüística (por ejemplo, prosodia). Además de los comandos de voz para los robots, el habla es una ventana a la condición psicológica, física y emocional de los humanos. Sin embargo, el análisis y el procesamiento del habla son muy sensibles a los entornos de ruido (incluido el efecto "cocktail party"), el canal acústico variable en el tiempo y la reverberación. En este trabajo se investiga el uso de técnicas de beamforming multimodales o audiovisuales (es decir, beamforming de audio utilizando con el seguimiento de imágenes) y los métodos basados en aprendizaje profundo para procesar el habla para extraer las características relevantes que pueden conducir a un reconocimiento de voz robusto y proporcionar información sobre las condiciones emocionales y psicológicas del usuario en escenarios dinámicos de HRI.

El beamforming multimodal o audiovisual denota la integración del procesamiento de vídeo o imagen con la tecnología de beamforming de audio, figuras 1.1 y 1.2 que es muy sensible al ruido aditivo y la reverberación, lo que a su vez causa un error en la estimación del ángulo de incidencia (AOI). El uso del seguimiento de objetivos basado en imágenes para mejorar la sensibilidad a la interferencia del beamforming de audio es un enfoque que no es nuevo. Sin embargo, el progreso de Deep Learning permite la implementación de tareas más sofisticadas con tecnología audiovisual como: detector de actividad de voz, VAD [38]; separación de altavoces [39]; diarización del orador [40]; y, reconocimiento automático de voz, ASR [41].

Uno de los enfoques de la inteligencia artificial (AI) más exitosos es el aprendizaje automático. Corresponde a la segunda ola de AI, el aprendizaje estadístico. El aprendizaje automático extrae patrones de datos no etiquetados (aprendizaje no supervisado) o categoriza eficientemente los datos de acuerdo con definiciones preexistentes incorporadas en un conjunto de datos etiquetados (aprendizaje supervisado). El aprendizaje automático permite que las computadoras aprendan sin ser programadas explícitamente. Los conjuntos de datos se pueden alimentar a estos sistemas y encontrarán relaciones ocultas en los datos de entrenamiento y mejorarán su rendimiento con el tiempo. Categorizar y etiquetar los datos a través del aprendizaje supervisado a menudo funciona mejor, especialmente con conjuntos de datos más pequeños. La tecnología de aprendizaje automático potencia muchos aspectos de la sociedad moderna: desde búsquedas en la web hasta filtrado de contenido en redes sociales y recomendaciones en sitios web de comercio electrónico, y está cada vez más presente en productos de consumo como cámaras y teléfonos inteligentes. Los sistemas de aprendizaje automático se utilizan para identificar objetos en imágenes, transcribir discursos en texto, relacionar elementos de noticias, publicaciones o productos con los intereses de los usuarios y seleccionar resultados relevantes de búsqueda. Las aplicaciones del aprendizaje automático han adquirido una gran relevancia gracias al aprendizaje profundo [42]. Este es un tipo de aprendizaje automático que utiliza capas de procesamiento jerárquicas adicionales (análogas a las estructuras de las neuronas humanas) y grandes conjuntos de datos para modelar abstracciones de alto nivel y reconocer patrones en datos complejos. Los sistemas de aprendizaje profundo son especialmente buenos para extraer patrones de la complejidad y están logrando

avances importantes en la resolución de problemas que han resistido los mejores intentos de la comunidad de inteligencia artificial durante muchos años. Ha producido resultados extremadamente prometedores para diversas tareas en la comprensión del lenguaje natural, en particular la clasificación de temas, el análisis de sentimientos, la respuesta a preguntas y la traducción de idiomas.

En este trabajo se investiga el uso de técnicas de beamforming, separación de canal y aprendizaje profundo para mejorar el reconocimiento automático de voz en el contexto de voz interacción humano-robot (Human-Robot Interaction en inglés, HRI). Si bien estos métodos son cada vez más robustos, hay aspectos que en los que aun se puede mejorar.

Los robot vienen equipados con una variedad de sensores para realizar diversas tareas. Entre ellos la presencia de múltiples micrófonos abre la posibilidad de utilizar técnicas de beamforming. A su vez la presencia de cámaras permite la detección de fuentes de audio utilizando información visual. Esto permite aplicar técnicas de beamforming a para obtener señales enfocadas en mas de una fuente.

En el presente trabajo se abordara la idea de utilizar múltiples beamforming y técnicas de separación de canal para mejorar el desempeño de un sistema de reconocimiento de voz.

1.2. Definición del problema

Se estudiara el caso de dos señales de beamforming generadas por de una fuente de voz y una fuente de ruido. La primera señal corresponde a un beamforming apuntando a la fuente de voz. En la segunda señal el beamforming apunta a la señal de ruido. Estas señales corresponden al tipo de datos que se pueden obtener con los sensores multimodales de un robot.

Estas señales serán utilizadas en un sistema de reconocimiento de voz. Se desea separar la señal de voz para obtener un reconocimiento mas exacto. Para esto se aplicaran técnicas de deep learning las cuales se compararan con técnicas de separación del fuentes en el estado del arte.

1.3. Hipótesis

El uso de técnicas de separación de canal en un sistema de reconocimiento de voz permite obtener resultados mas robustos. Estas técnicas requieren una gran cantidad de datos para determinar las estadísticas necesarias para descomponer las señales. En situaciones dinámicas como el caso de un robot en movimiento estas estadísticas varían por lo cual deben calcularse con menos datos. El uso de técnicas de aprendizaje profundo permite estimar las estadísticas de la señal utilizando ventanas de corto plazo. Con lo que es posible utilizarlas en ambientes dinámicos.

1.4. Objetivos

1.4.1. Objetivo general

Mejorar la robustez de un sistema de reconocimiento de voz en interacción humano robot en ambiente ruidoso con filtrado espacial i.e. beamforming.

1.4.2. Objetivos específicos

- Generar base de datos de entrenamiento y test, simulando dos beamforming con una fuente de voz y una fuente de ruido.
- Modelar la mezcla de señales de voz y ruido en un contexto de múltiples beamformings.
- Desarrollar e implementar un sistema para separar la señal del ruido.
- Evaluar la robustez de un sistema de reconocimiento de voz utilizando la técnica de separación de fuentes desarrollada en relación a otras técnicas del estado del arte.

1.5. Estructura de la memoria

El trabajo aquí contenido se divide en cinco partes siendo esta la primera, en la cual se introduce brevemente el tema a tratar y los objetivos.

En el capítulo 2 se describe las técnicas de deep learning, la interacción humano robot basada en voz, el reconocimiento automático de voz, técnicas de beamfoming y separación de fuentes.

En el capítulo 3 se propone un modelo matemático utilizado para describir el caso de estudio y como se relaciona otros modelos de la literatura. Se analizan posibles soluciones al problema utilizando el modelo propuesto. Finalmente se describen las soluciones que se implementaran y evaluaran en el capitulo siguiente.

En el capítulo 4 se describen los experimentos realizados y se discuten los resultados. Primero se describen las bases de datos utilizadas, y el sistema de reconocimiento de voz en el cual se evalúan los resultados. Luego se presentan los resultados de los experimentos realizados para evaluar las soluciones propuestas.

En el capítulo 5 se presentan las conclusiones del trabajo. Se discute si el método cumple las expectativas y posibles formas de mejorar en trabajos futuros.

Capítulo 2

Reconocimiento de voz en interacción humano robot y separación de fuente

2.1. Deep learning

2.1.1. Introducción

Los métodos de deep learning apuntan a aprender jerarquías de características. Las características de aprendizaje automático en múltiples niveles de abstracción permiten que un sistema aprenda funciones complejas mapeando la entrada a la salida directamente desde los datos, sin depender completamente de las características diseñadas específicamente para abordar el problema. Esto es especialmente importante para las abstracciones de alto nivel, donde los humanos tienen dificultades para especificar explícitamente en términos de los datos no procesados. La capacidad de aprender automáticamente funciones de alta complejidad aumentara su importancia a medida que la cantidad de datos y el rango de aplicaciones para los métodos de aprendizaje automático continúe creciendo [43].

2.1.2. Redes neuronales profundas

Inspirados por la arquitectura del cerebro, los investigadores de redes neuronales aspiraron a entrenar redes neuronales multicapa profundas DNN (Deep Neural Networks) durante décadas [44], [45]. Sin embargo los primeros intentos exitosos no fueron reportados hasta el año 2006. La literatura reportaba resultados experimentales positivos para redes de dos o tres niveles (es decir, una o dos capas ocultas). Por el contrario el entrenamiento de redes más profundas arrojaba de manera consistente resultados inferiores.

En 2006 Geoffrey Hinton y un grupo de científicos en la Universidad de Toronto introdujeron las Deep Belief Networks (DBN) [46], usando un método que realiza un entrenamiento no supervisado capa a capa mediante un algoritmo de aprendizaje llamado Restricted Boltzmann Machine (RBM) [47]. Poco después, se propusieron algoritmos relacionados basados en autocoders [48], [49], utilizando otro enfoque para explotar el mismo principio: guiar el entrenamiento de niveles intermedios de representación usando aprendizaje no supervisado, que puede realizarse localmente en cada nivel. Recientemente se han propuesto otros algo-

ritmos para arquitecturas profundas que no utilizan RBM ni autocoders y que explotan el mismo principio [50], [51].

Desde 2006, las redes profundas se han aplicado con éxito en tareas de clasificación [48], [52], [53], regresión [54], reducción de dimensionalidad [55], modelado de texturas [56], segmentación de objetos [57], recuperación de información [58], robótica [59], procesamiento de lenguaje natural [60] y filtrado colaborativo [61]. Aunque los autoencoders, las RBM y las DBN se pueden entrenar con datos no etiquetados, en muchas de las aplicaciones anteriores, se han utilizado con éxito para inicializar redes neuronales feedforward profundas supervisadas aplicadas a una tarea específica [62].

2.1.3. Redes neuronales recurrentes

Las redes neuronales recurrentes, o RNN [63], son una familia de redes neuronales usadas para procesar datos secuenciales y se pueden aplicar secuencias mucho más largas de lo que sería práctico para redes convencionales feedforward. La mayoría de las redes recurrentes también pueden procesar secuencias de longitud variable. La idea detrás de las RNN es hacer uso de la información secuencial, mientras en una red neuronal tradicional se suponen que todas las entradas (y salidas) son independientes entre sí, lo que para muchas tareas puede ser insuficiente o suboptimo. Las RNN se llaman recurrentes porque realizan la misma tarea para cada elemento de una secuencia, y la salida depende de los cálculos previos.

Las RNN, y en menor medida los MLPs, sufren del denominado problema de desvanecimiento del gradiente (vanishing gradient problem). En deep learning, el problema de desvanecimiento del gradiente es una dificultad que se encuentra en el entrenamiento de redes neuronales artificiales con métodos de aprendizaje basados en gradiente y backpropagation. En dichos métodos, cada uno de los pesos de la red neuronal se actualiza de manera proporcional a la derivada parcial de la función de costo con respecto al peso actual en cada iteración de entrenamiento. El problema es que, el gradiente puede ser muy pequeño, lo que evitará que el peso cambie su valor. En el peor de los casos, esto puede detener por completo el aprendizaje de una red neuronal.

Como un ejemplo de la causa del problema, las funciones de activación tradicionales, como la función de tangente hiperbólica, tienen gradientes en el rango (0, 1) y el algoritmo de backprogation calcula los gradientes según la regla de la cadena. Esto tiene el efecto de multiplicar n de estos pequeños números para calcular los gradientes de las capas "frontales" en una red de n capas, lo que significa que el gradiente (señal de error) disminuye exponencialmente con n, haciendo que las capas frontales se entrenen muy lentamente. Este problema, identificado por Hochreiter en 1991 [64], no solo afecta a las redes feedforward de muchas capas, sino también a las redes recurrentes. Estas últimas se entrenan desplegándolas en redes feedforward muy profundas, donde se crea una nueva capa para cada paso de una secuencia de entrada procesada por la red.

2.1.4. Long-Short Term Memory

Una forma de resolver el problemas del desvanecimiento del gradiente son las redes Long-Short Term Memory (LSTM). Estas son redes neuronales recurrentes que tiene bloques de

células LSTM en lugar de capas de redes neuronales estándar. Estas celdas se componen de una puerta de entrada, una puerta de olvido y una puerta de salida, las cuales se describen más adelante. En la figura 2.1 se muestra una representación gráfica de la celda LSTM.

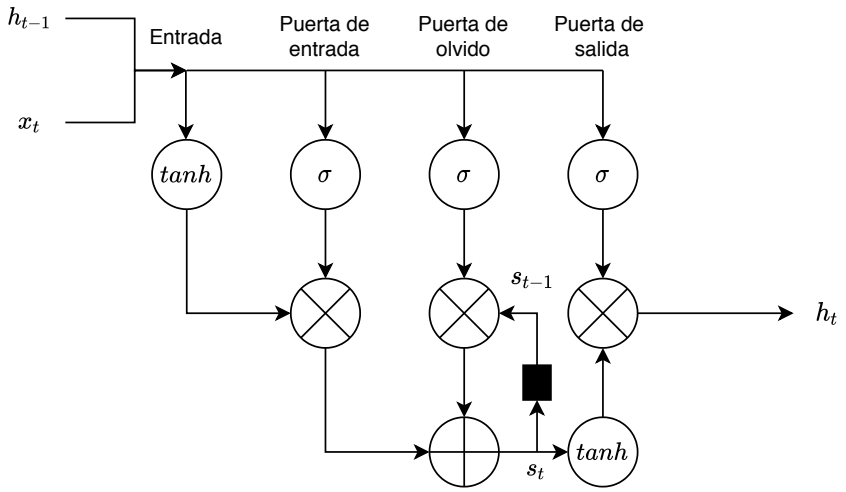

Figura 2.1: Esquema de una red LSTM

La entrada a la celda consiste en la secuencia x_t concatenada con el resultado anterior de la celda h_{t-1}. Esta entrada alimenta tanto a una capa *tanh* como a las puertas de entrada, olvido y salida. Una puerta de entrada es una capa de nodos con activación sigmoide cuya salida se multiplica por la salida de la capa tanh. Las activaciones sigmoides de la puerta de entrada actúan eliminando o conservando cualquier elemento del vector de entrada según sea necesario, entrenando los pesos que conectan la entrada a estos nodos para generar valores de salida cercanos a 0 o a 1 para distintos valores de entrada.

El siguiente paso en el flujo de datos a través de esta celda es el ciclo estado interno y puerta de olvido. Las células LSTM tienen una variable de estado interna s_t . Esta variable, se utiliza despues de un retardo es decir, s_{t-1} se agrega a los datos de entrada para crear una capa efectiva de recurrencia.

Esta operación de adición, en lugar de una operación de multiplicación, ayuda a reducir el riesgo de desvanecimiento de gradiente. Sin embargo, este ciclo de recurrencia está controlado por una puerta de olvido que funciona de manera similar a la compuerta de entrada, pero ayuda a la red a saber qué variables de estado deben "recordarse" u "olvidarse".

Finalmente, tenemos una función *tanh* como capa de salida, controlada por una puerta de salida. Esta puerta determina qué valores están realmente permitidos como salida de la celda h_t.

2.2. Interacción humano robot basada en a voz

Para que la robótica social se vuelva una realidad, es necesario desarrollar la integración entre humanos y robots. Así se podría mejorar significativamente la cooperación entre usuarios y máquinas. Aplicaciones como defensa, ambientes hostiles, minería, industria, silvicultura, educación y desastres naturales, son situaciones donde sera necesaria la integración y colaboración entre humanos y robots [65]. La interacción entre humanos y robots, conocido por sus siglas en ingles HRI (Human Robot Interaction) cobra especial relevancia en aquellas situaciones donde los robots no son totalmente autónomos. Ejemplo de esto es cuando requieren interacción con humanos para recibir instrucciones o información para la toma de decisiones [66]-[69].

En este contexto es esencial que la comunicación entre personas y robots emule la interacción entre humanos, para así establecer una exitosa colaboración humano-robot [4]-[8].

Además, el habla es la forma más sencilla y natural que los humanos emplean para comunicarse. Como consecuencia, el HRI basado en voz debe ser la forma más natural de facilitar una sinergia de colaboración entre humanos y robots[10], [70], [71]. Por lo tanto, la tecnología del habla, especialmente el reconocimiento de voz automático (ASR), debe jugar un papel importante en la robótica social.

Además, es bien sabido que la visión artificial es un tema de investigación importante en robótica. Los desafíos recientes como DARPA Robotics Challenge [72] y Robocup [73] han llevado a grandes mejoras en la visión por computacional. Por otro lado, también ha habido un progreso significativo en el ASR), pero este avance ha tenido lugar fuera del campo HRI. El uso de ASR ha ganado relevancia en robótica en los últimos años, pero su estado aún está lejos del que disfrutó la visión por computadora en la investigación robótica [69], [74]-[76]. Esto todavía es algo sorprendente, considerando que ambas tecnologías utilizan métodos de procesamiento de señales y de aprendizaje profundo similares, y pueden explicar en parte la menor penetración de ASR) en la comunidad robótica.

En [77] se genero una base de datos con un robot en movimiento para simular el diálogo entre un robot y humano mientras estos se encuentran en movimiento. Esta base de datos está inspirada en Aurora-4 [78], [79] que agrega ruido aditivo al Corpus de DARPA Wall Street Journal (WSJ0). La base de datos MChRSR (Multichannel Robot Speech Recognition Database) utiliza el PR2 (Personal Robot 2) como una plataforma móvil para grabar audio utilizando el arreglo de micrófonos de Microsoft Kinect. Las 330 señales del conjunto de datos de prueba se volvieron a grabar en 16 condiciones de movimiento, utilizando un altavoz fijo.

El primer uso de esta base de datos fue en [80] donde se entrenó un sistema DNN-HMM (redes neurnals concinadas con modelos ocultos de Markov)utilizando la plataforma Kaldi [81].

En [82] se utilizó la misma base de datos, con un entrenamiento basado en el ambiente (EbT). Este entrenamiento considera la respuesta al impulso de la sala como en [80] y el ruido del robot se agrega a los datos de entrenamiento con la misma metodología que en la base de datos Aurora-4 usando el software FaNT [83]. El ASR resultante supera a las API de ASR de propósito general proporcionadas por Google, Microsoft e IBM en la tarea dada.

2.3. Reconocimiento automático de voz

2.3.1. Formulación del problema

El proceso de reconocimiento automático de voz, conocido por sus siglas en ingles ASR (Automatic Speech Recognition), corresponde a obtener una transcripción a partir de una señal acústica que contenga voz. El problema se puede formular de la siguiente manera [84]:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\{p(W|X)\} \tag{2.1}$$

donde $\hat{W}$ es la secuencia óptima de etiquetas (palabras o fonemas), X es la secuencia de observaciones de entrada que representa un enunciado de voz dado; Sin embargo, dado que p (W | Y) es difícil de modelar directamente, el teorema de Bayes se usa para transformar 2.1 en el problema equivalente de encontrar

$$\hat{W} = \underset{W}{\operatorname{argmax}}\{p(X|W)p(W)\} \tag{2.2}$$

Dode $p(W)$ denota el modelo de lenguaje que describe las probabilidades de combinaciones de palabras y $p(X|W)$ indica el modelo acústico. En consecuencia, la tarea de un sistema ASR es encontrar la secuencia de etiqueta más probable $\hat{W}$ (mediante un proceso llamado decodificación, realizado con el algoritmo de Viterbi [85]).

El modelo de lenguaje puede ser representado con modelos estadísticos, gramática libre de contexto(en inglés Stochastic context-free grammar, SCFG), o modelos estocásticos de estado finito.[86]. En el caso de los modelos estadísticos, que son ampliamente usados en investigación, la probabilidad a priori de una secuencia de palabras $W = w_1, ..., w_n$ en la ecuaciónes 2.1, 2.2 puede ser aproximada con N-gramas (secuencia de N fonemas):

$$p(W) = \sum_{l=1}^{L} p(w_l|w_{l-1}, w_{l-2}, ..., w_{l-N+1}) \tag{2.3}$$

donde N es típicamente un entero entre 2 y 4. El modelo de lenguaje define la probabilidad de transición desde un N-grama a la siguiente palabra para guiar la búsqueda de una interpretación de la entrada acústica. Adicionalmente, el tamaño del vocabulario y la perplejidad son factores críticos para el desempeño del ASR. La perplejidad mide la incertidumbre de las palabras que pueden seguir a un N-grama dado. Un modelo de lenguaje con baja perplejidad definido por un contexto o una tarea dada limitará la decodificación y se desempeñará mejor que uno de alta perplejidad.

La modelación acústica define representaciones estadísticas para la secuencia de vectores de características acústicas X obtenidos de la forma de onda de la señal de voz. Las señales son divididas en ventanas, típicamente de 20 a 30 [ms] con traslape, por ejemplo, del 50 %.

Usualmente los vectores de caracteristicas se obtienen aplicando a cada ventana la transformada rápida de Fourier (fast Fourier transform en inglés, FFT) [87]-[89]. Los coeficientes de velocidad y aceleración (llamados coeficientes delta y delta-delta) también son típicamente usados, y el vector de características final se compone de las características estadísticas junto a los coeficientes recién mencionados. [90]. Los vectores de características pueden utilizar normalización como la de media y varianza en los coeficientes.

2.3.2. Modelos ocultos de Markov

Por muchos años la mayoría de las sistemas de reconocimiento de voz solían utilizar modelos ocultos de Markov (en inglés hidden Markov models, HMMs) para tratar la variabilidad temporal de la voz, y modelos gaussianos mixtos (en inglés Gaussian mixture models, GMMs) para representar $p(X|W)$. Dado un conjunto de vectores de caracteristicas de voz $X = \{x_t\}_{t=1}^{T}$, la función de densidad de probabilidad de observación del vector x_t en el estado s_i se expresa como [87]:

$$p(x_t|s_\mathrm{i}) = \sum_{m=1}^{M} c_{\mathrm{i},m} N(x_t, \mu_{\mathrm{i},m}, \Sigma_{\mathrm{i},m}) \tag{2.4}$$

donde $c_{\mathrm{i},m}$, $\mu_{\mathrm{i},m}$ y $\Sigma_{\mathrm{i},m}$ corresponden a los pesos de mezcla, vectores de medias, y matrices de covarianzas respectivamente, para M componentes de mezcla Gaussianos.

2.3.3. Redes neuronales profundas

En los últimos años, las redes neuronales artificiales (como por ejemplo las Deep Neural Networks, DNNs) han demostrado un desempeño significativamente mejor que los modelos basados en GMMs.

En un sistema GMM-HMM, la DNN entrega una pseudo-verosimilitud logarítmica definida como:

$$log[p(x_t|s_\mathrm{i})] = log[p(s_\mathrm{i}|x_t)] - log[p(s_\mathrm{i})] \tag{2.5}$$

donde s_j denota uno de los estados; y las probabilidades a priori $log[p(s_\mathrm{i})]$ pueden ser entrenadas usando los alineamientos de estado obtenidos con la base de datos de entrenamiento.

$$\hat{W} = \underset{W}{\mathrm{argmax}}\{log[p(X|W)] - \lambda log[p(W)]\} \tag{2.6}$$

donde probabilidad del modelo acústico $p(X|W)$ depende de la función de pseudo verosimilitud logarítmica $log[p(s)]$ entregada por la DNN. La variable λ permite controlar la importancia del modelo acústico y el modelo de lenguaje [62].

Ejemplo de uso de ASR con DNN y HMM se presenta en [91] donde el sistema permite una reducción de la tasa de error de palabras (Word Error Rate en inglés, WER) del 32 % relativo, al compararlo con el sistema GMM-HMM común en el Switchboard task [92].

La tarea de entrenar una DNN puede resultar difícil. La función objetivo puede ser altamente no-convexa y el algoritmo de entrenamiento puede converger fácilmente a un mínimo local sub-óptimo. Además, las redes neuronales artificiales necesitan una mayor cantidad de datos de entrenamiento que los sistemas GMM-HMM [93].

Es importante considerar que los sistemas de ASR basados en redes neuronales que han sido publicados emplean al menos miles de horas de datos de voz para ser entrenados [61], [94], [95].

Otras arquitecturas de redes neuronales artificiales aplicadas al problema del ASR son: LSTM [96], CNN [97] y RNN [98]. Los resultados obtenidos usando sistemas DNN-HMM son competitivos al compararlos con los resultados de otras arquitecturas [98]-[103]

En algunos casos es posible generar sistemas que superan el desempeño de las DNNs, LSTMs, o CNNs empleando combinaciones arquitecturas como very deep CNN [104] o fCNN [105]. Sin embargo, al aumentar el número de parámetros de una red neuronal artificial, se requiere una cantidad mayor de datos para su entrenamiento.

El uso de condiciones representativas de los datos de prueba en el entrenamiento, permiten que el sistema ASR presente una mejora en su desempeño. Por el contrario, los modelos presentarán dificultades al reconocer datos de prueba cuando estos difieren de los datos de entrenamiento. Por esta razón, la robustez al ruido de un sistema con redes neuronales artificiales se puede lograr usando un entrenamiento con condiciones múltiples. Por ejemplo, una DNN entrenada con distintos tipos de ruido y niveles de SNR puede llevar a mejoras de alta precisión en aplicaciones reales [106].

2.3.4. Word error rate

Una métrica usada en el reconocimiento de voz es el WER (Word Error Rate), el cual es la tasa de errores respecto al numero de palabras presentes en la señal. Los errores pueden ser sustituciones cuando se detecta una palabra distinta, omisiones cuando una palabra no se detecta y o inserciones cuando se detecta una palabra extra. Con esto el WER se calcula según la ecuación 2.7.

$$WER = 100\frac{Sustituciones + Omisiones + Inserciones}{Palabras} \tag{2.7}$$

Es importante notar que en WER se calcula sobre el alineamiento que minimice el wer. Es decir se busca asociar las palabras de la transcripción de referencia con las de la transcripción a evaluar de manera que exista el menor número de errores posibles antes de contar el número de errores.

En la tabla 2.1 se muestran múltiples posibles transcripciones para una misma frase y cómo se calcula el WER en cada caso. Además se incluye un ejemplo en el cual un mal

alineamiento de la transcripción 3 lleva a un calculo errado del WER.

Tabla 2.1: Ejemplos calculo de WER.

	Alineamiento							S	O	I	WER %
Referencia		La	casa	de	color	rojo					
Transcripción 1	Óptimo		casa	de	color	rojo			1		20 %
Transcripción 2	Óptimo	La	**caja**	de	color	rojo	**oscuro**	1		1	40 %
Transcripción 3	Óptimo	La	casa		color	**rosa**		1	1		40 %
Transcripción 3	Erróneo		**La**	**casa**	color	**rosa**		2	1	1	80 %

2.4. Técnicas de beamforming e información visual

2.4.1. Introducción

El proceso de beamforming se refiere a múltiples técnicas de filtrado espacial. La idea básica es utilizar un arreglo de sensores con diferentes ubicaciones en el espacio para separar una señal deseada del ruido y las señales interferentes, en función de la ubicación de su origen. Mientras que un filtro FIR es una combinación lineal de muestras a lo largo del tiempo, un beamforming es una combinación lineal de muestras en el espacio.

Estas técnicas se pueden aplicar en múltiples áreas utilizando diferentes tipos de sensores. En [107] se describen varias técnicas para abordar este problema utilizando arreglos de micrófonos para el reconocimiento de voz.

En [108] se realiza una revisión del estado del arte, y se describen múltiples aplicaciones que incluyen sonar, radar, comunicaciones, imágenes, astrofísica, geofísica y biomedicina. En el caso de las comunicaciones, las antenas se utilizan como sensor.

Una medida de la mejora obtenida por el uso de beamforming es la relación entre el SNR (ec. 2.8) de un sensor y la del arreglo. Esta razón se conoce como la ganancia G y se define por la ecuación 2.9.

$$SNR = \frac{P_S}{P_R} \tag{2.8}$$

Donde P_S es la potencia de la señal y P_R la potencia del ruido.

$$G = \frac{SNR_{Areglo}}{SNR_{Sensor}} \tag{2.9}$$

2.4.2. Retardo y suma

Retardo y suma [109] es la técnica de beamforming más simple. Esto supone retrasos conocidos entre los diferentes sensores N, que los alinea y luego los suma. Esto se define por la ecuación 2.10 en la que x_i corresponde a la señal recibida por el sensor i, τ_i corresponde al retraso de la señal para el sensor i.

$$y(t) = \sum_{\mathrm{i}=1}^{N} x_{\mathrm{i}}(t - \tau_{\mathrm{i}}) \tag{2.10}$$

Retardo y suma resultan ser un beamforming óptimo para el caso del ruido blanco gaussiano con una ganancia de $G = N$

2.4.3. Minimum variance distortionless response

MVDR (Minimum variance distortionless response) es una técnica de beamforming más avanzada. Esto minimiza la varianza de la señal resultante sujeta a tener una ganancia unitaria en la dirección de llegada de la señal.

Su función objetivo está definida por las ecuaciones 2.11 y 2.12 sujetas a la restricción de la ecuación 2.13.

$$R_x = E\{X^H X\} \tag{2.11}$$

$$min\{W^H R_x W\} \tag{2.12}$$

$$W^H \mathrm{d} = 1 \tag{2.13}$$

La solución para la matriz de peso está dada por la ecuación 2.14

$$W = \frac{R_x^{-1}\mathrm{d}}{\mathrm{d}^H R_x^{-1}\mathrm{d}} \tag{2.14}$$

Esta beamforming resulta óptima para un ruido no correlacionado y una señal de banda estrecha. Es importante tener en cuenta que la voz siempre es de banda ancha [110].

2.4.4. Estimación de retardos

Una forma de obtener los retardos es calcular la correlación entre la señal de cada sensor con un sensor de referencia. Con esto, el máximo de la correlación entre ambas señales corresponderá al número de muestras de retardo.

Otra forma de estimar los retrasos entre las señales es conocer el la dirección de llegada DOA. Con esto es así y la geometría de la disposición se puede calcular retrasos. La figura 2.2 muestra un ejemplo de un frente de onda plana con una arreglo lineal de micrófonos.

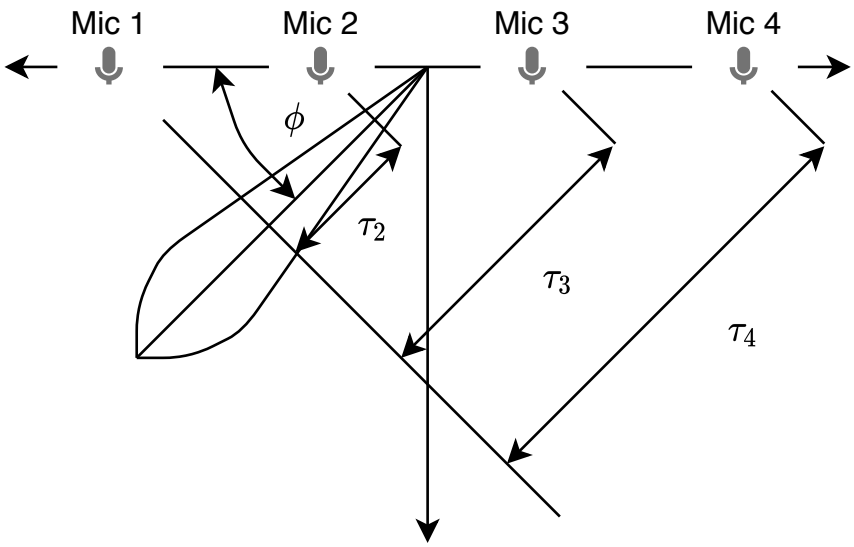

Figura 2.2: Arreglo de micrófono con frente de onda plana.

En este caso para el ángulo ϕ y conociendo la distancia l_i al punto de referencia, podemos calcular el τ_i como se muestra en la ecuación 2.15, donde V es la velocidad de sonido [111].

$$\tau_\mathrm{i} = \frac{l_\mathrm{i} \sin(\phi)}{V} \tag{2.15}$$

2.4.5. Beamforming con información visual

Uno de los puntos cruciales para utilizar las técnicas de beamforming es la estimación de retardos y en particular la estimación del ángulo de incidencia (AOI). En el caso de los robot se puede aprovechar la parecencia de cámaras para estimar esta dirección visualmente.

En [112], [113] se utilizó las cámaras de una Kinect para estimar la dirección de una fuente con el robot en movimiento y utilizarla para calcular un beamforming. Luego estas señales fueron utilizadas en un sistema de reconocimiento de voz.

Esta misma metodología puede ser utilizada para generar múltiples beamformings a múltiples fuentes. Ya que diferentes fuentes serán objetos distintos en la cámara y al conocer sus posiciones se podrá calcular retardos para cada uno de ellos.

2.5. Análisis de componentes independientes

El análisis de componentes independientes conocido como ICA por sus siglas en ingles (Independent Component Analysis) es una técnica de separación de fuentes [114].

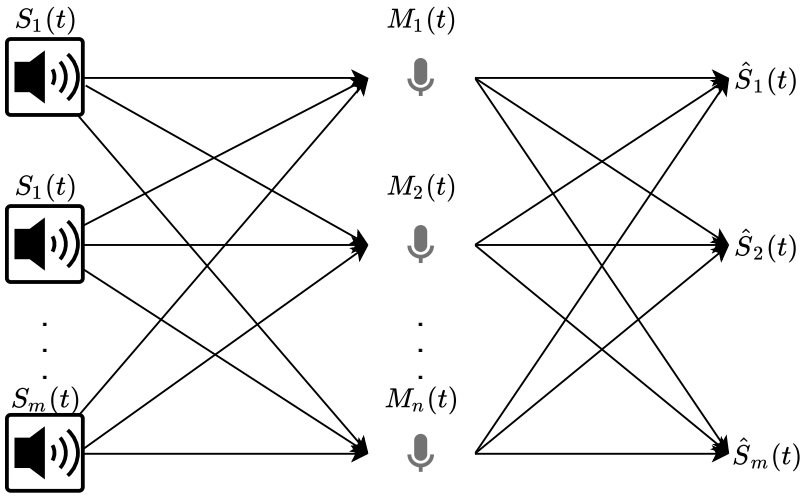

Figura 2.3: Esquema de modelo ICA en contexto de audio.

En un contexto de múltiples fuentes de audio y micrófonos como se muestra en la figura 2.3, se tiene que en cada sensor se mide una combinación de señales con diferentes ganancias para cada señal, como se muestra en:

$$\begin{bmatrix} M_1(t) \\ M_2(t) \\ \vdots \\ M_n(t) \end{bmatrix} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,m} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n,1} & g_{n,2} & \cdots & g_{n,m} \end{bmatrix} \begin{bmatrix} S_1(t) \\ S_2(t) \\ \vdots \\ S_m(t) \end{bmatrix} \tag{2.16}$$

Donde la señal para la fuente j se denota S_j la mezcla de señales medida en el micrófono i es M_{i} y la ganancia entre la fuente i y el sensor j se denota $g_{\mathrm{i},j}$. Esto se conoce como la matriz de mezcla.

La ecuación 2.16 puede resumir como:

$$M(t) = GS(t) \tag{2.17}$$

alternativamente se puede considerar el caso con ruido:

$$M(t) = GS(t) + n(t) \tag{2.18}$$

y el caso no lineal

$$M(t) = F(S(t)|K) + n(t) \tag{2.19}$$

El algoritmo ICA busca encontrar la matriz de mezcla G que permita recuperar las señales S al multiplicar su inversa por las observaciones M, obteniendo las señales $\hat{S}_{\text{i}}$ que estiman la señal S_{i}.

Eso se puede hacer de dos maneras:

- Minimizando la información mutua entre las señales S
- Maximizando la No-Gaussianidad las señales S

Para medir la información mutua o la no-gaussianidad existen muchas funciones. En el caso de la no-gaussianidad, la más común es la kurtosis, que se define como:

$$Kurt\big[X\big] = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big] = \frac{E[(X-\mu)^4]}{E[(X-\mu)^2]^2} = \frac{\mu^4}{\sigma^4} \tag{2.20}$$

Es importante notar que mientras más muestras se tomen para calcular la kurtosis, esta será más representativa de la señal, ya que las estadísticas calculadas serán más significativas. Por lo tanto, si se trabaja con ventanas de tiempo pequeños, este método se vuelve menos exacto.

En el caso de la información mutua, esta se obtiene a partir de las PDF (Función de Densidad de Probabilidades) de las variables, y se define para un conjunto de variables aleatorias como la diferencia entre la probabilidad conjunta de estas y el producto de las probabilidades marginales de cada una. Naturalmente, para que las funciones de densidad de probabilidad sean representativas, se requieren más datos, es decir, si se trabaja con ventanas de tiempo con pocas muestras, habrá en deterioro en la precisión de este método.

Se puede aplicar ICA directamente sobre las señales en el tiempo en el caso ideal (ICA sin ruido). Sin embargo, cuando hay presencia de ruido y canal, resulta conveniente aplicarlo en el dominio de la frecuencia. Para esto, se trabaja con la trayectoria de cada bin de frecuencia como observación separada, aunque esto conlleva a problemas ya que cada bin es estimado con diferentes precisiones y además se producen permutaciones entre los bines estimados. En [115] y [116] se presentan soluciones a este problema mediante la agrupación de bines. El mismo problema se produce en el tiempo al trabajar por ventana, ya que no se sabe a cuál de las fuentes corresponde cada estimación, y por lo tanto, de un frame a otro el orden de las variables estimadas puede cambiar, lo que dificulta enormemente la reconstrucción de la señal completa.

ICA se ha mostrado útil para el reconocimiento de voz de una fuente en ambientes ruidosos y reverberantes (problema denominado cocktail party) visto en [117], teniendo buenos resultados para SNR y WSS (Weighted Spectral Slope). También se usó para separación ciega de voz usando ICA con NLR (Nonparametric Likelihood Ratio) como función objetivo en [118], comparándolo con resultados obtenidos usando MMI (Minimum Mutual Information). Otro ejemplo es el uso de ICA en condiciones de múltiples interlocutor para el reconocimiento de voz usando en la salida de ICA un post-masking en el dominio del tiempo-frecuencia para el reconocimiento de características faltantes, así como también se aplica el modelo de

incertidumbre Gaussiano complejo para la estimación de la incertidumbre de características reconocidas [119]. Fuera de sus aplicaciones para el reconocimiento de voz, ICA ha probado ser útil en separación ciega de voz de datos en electroencefalografías (EEG) así como en EEG (Event-Related Potential) lo cual se puede ver en [120], como también analizando datos dados por magnetoencefalografías (MEG) en [121], siendo un método efectivo a la hora de separar fuentes de señales cerebrales. Relacionado a esto, ha probado ser un poderoso método para detectar activaciones relacionadas a tareas y activaciones no anticipadas en datos de imágenes por resonancia magnética funcional (fMRI) como se ve en [74]. También se ha usado ICA en el reconocimiento facial en [122], donde se usaron 2 diferentes arquitecturas y ambas dieron mejores resultados que los obtenidos con análisis de componentes principales (PCA) para reconocer caras a lo largo de los días y cambios de expresiones, donde un clasificador que combinaba ambas arquitecturas daba los mejores resultados.

2.6. Non-negative Matrix Factorization

De manera similar a ICA, NMF (Non-negative Matrix Factorization) se usa para aproximar la matriz M en la ecuación 2.17 considerando que las matrices G y S son no negativas [123]. En procesamiento de voz es posible usar este procedimiento ya que los espectrogramas de las señales de audio son justamente matrices no-negativas. El problema de NMF no tiene solución exacta, por lo que se suelen hacer métodos numéricos para aproximar las soluciones.

Los algoritmos de "expectation-maximization" (EM) o maximización del valor esperado son los más comunes. Estos consisten en encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables. En estos se alternan dos pasos:

- E: se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables.
- M: se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E.

Como se puede ver, los algoritmos EM incluyen el cálculo de la verosimilitud y el valor esperado. Por lo tanto, para que funcionen correctamente se debe trabajar con un mínimo de muestras, ya que de lo contrario los valores calculados se vuelven poco significativos y las estimaciones pierden validez. Es decir, los algoritmos EM presentan problemas si se trabaja con ventanas pequeñas. Una limitación que es común a todos los algoritmos de NMF, es que al trabajar con la amplitud del espectrogramas, la información de la fase se pierde, haciendo que la reconstrucción de la señal en el tiempo se distorsione.

En [124], los autores utilizaron NMF para abordar la tarea de mejora del habla en un entorno de ruido complejo en combinación con una matriz de micrófonos.

En [125], se presentó un enfoque combinado que utiliza la agrupación de NMF y K-means para la separación de la fuente, donde los autores probaron su método en múltiples conjuntos de datos. El método mostró un buen rendimiento para mezclas lineales de voz y música, pero un bajo rendimiento para mezclas convolutivas. En [126], se presentó un método basado en

la factorización de matriz no negativa (NMF) para la separación de la fuente. Los autores presentaron un marco probabilístico bayesiano para el modelado de múltiples fuentes. Este marco aborda una restricción significativa con los métodos clásicos de NMF, que es que se debe conocer el número de fuentes.

Capítulo 3

Separación de fuente usando deep learning

3.1. Modelo de mezcla propuesto

El procesamiento de voz a distancia es muy importante en la robótica social (por ejemplo, educación, servicios de salud, personas mayores y acompañantes de niños, servicio de turismo, etc.) y ofrece varios desafíos a la ciencia y tecnología del voz de última generación, particularmente ASR y sus problemas de ruido aditivo, canal acústico variable y reverberación en escenarios que varían en el tiempo. La separación de fuentes es una estrategia muy importante para abordar este problema. Para una matriz de micrófonos de múltiples canales, las señales capturadas se pueden representar como $Y = [y_0(t), y_1(t), ..., y_m(t), ..., y_{M-1}(t)]$, dónde $0 \leq m < M$, M es el número total de micrófonos y Yes la matriz de señal capturada. Considera que hay K fuentes $s_k(t)$, dónde $0 \leq k < K$. Cada micrófono recibe la suma de todas las K fuentes, lo que se expresa como:

$$y_m(t) = s_{0,m}(t) + s_{1,m}(t) + ... + s_{K-1,m}(t) = \sum_{k=0}^{K-1} s_{k,m}(t) \tag{3.1}$$

dónde $s_{k,m}(t)$ denota señal fuente $s_k(t)$ recibida en el micrófono m. Esta representación se puede comprimir como una función $y_m(t) = f[s_{k,m}(t)]$ que depende de la posición de cada fuente k. Si suponemos que la función es lineal, se puede obtener un modelo de mezcla: $Y = WS$. Como se definió anteriormente, Y es la matriz de señal capturada de tamaño $Mx1$ para cada instante de tiempo t, dónde $0 \leq t < T$ y T es la longitud en muestras del intervalo de señal que se analiza. Por consiguiente, Y puede interpretarse como una matriz de tamaño MxT, W es la matriz de mezcla de tamaño MxK y S es la matriz de fuentes KxT. Hablando genéricamente, los métodos que pueden explotar las propiedades estadísticas fundamentales de las señales del voz como que no son estacionarias, blancas ni gaussianas. Estas han sido muy populares para abordar tareas relacionadas con la voz [127]. ICA y NMF son ejemplos típicos de este tipo de enfoques. Estas técnicas estadísticas también se conocen como métodos de descomposición, y todas comparten el mismo principio que una

matriz dada V se puede factorizar en otras dos matrices W y H. Sin embargo, cada método tiene sus propias restricciones: ICA supone que las filas son independientes e idénticamente distribuidas [128]; y, NMF tiene la restricción de que todas las matrices tienen que ser no negativas [124].

Ambos métodos, ICA y NMF, se han utilizado para abordar múltiples tareas relacionadas con la voz como se encuentra en la literatura [124]-[126], [129], [130]. Como un inconveniente importante de la separación de fuentes basada en métodos estadísticos, podemos mencionar la dependencia de la cantidad de datos disponibles para optimizar las funciones objetivas, lo que a su vez impone restricciones a la aplicabilidad a entornos dinámicos que varían en el tiempo.

Como se explicó anteriormente, para tener éxito en aplicaciones reales, los robots sociales necesitan establecer perfiles para los usuarios de los dominios físicos, cognitivos y sociales.

Aquí, investigaremos la tecnología reconocimiento automático de voz (ASR), en entornos naturales. Por entornos naturales, entendemos escenarios esencialmente dinámicos en los que uno o más usuarios intentan interactuar con un robot en movimiento en presencia de ruido. Se abordarán los desafíos reconocimiento de voz distante y ruido aditivo. Para lograr estos objetivos, se investigará nuevos métodos de separación de fuentes y beamformings multimodales mediante la investigación y aplicación de nuevas arquitecturas avanzadas de aprendizaje profundo.

Ahora modelaremos el problema que se abordará para mostrar y contextualizar los desafíos de la tecnología de punta. Una matriz de micrófonos con un número arbitrario de micrófonos cuyas salidas pueden procesarse y combinarse para obtener un filtrado espacial generando una señal de beamforming. El uso de una matriz de micrófonos para realizar un beamforming puede reducir el efecto del ruido. En en el caso de retardo y suma, para apuntar el haz a la fuente p, las muestras $y_m(t)$ de cada micrófono m son retrasados por $\tau_{p,m}$ muestras y luego sumadas. Al hacerlo, la señal de salida con beamforming $b_p(t)$ en el dominio de tiempo discreto corresponde a:

$$b_p(t) = \sum_{m=0}^{M-1} y_m(t - \tau_{p,m}) \tag{3.2}$$

Se puede suponer un frente de onda plana si la distancia entre el conjunto de micrófonos y la fuente de sonido es mayor de 5-10 veces la longitud del arreglo de micrófonos [131] En consecuencia se puede utilizar el método de la ecuación 2.15 y el retardo para cada micrófono viene dado por:

$$\tau_{p,m} = \frac{l_m \sin(\phi_p)}{V} \tag{3.3}$$

dónde l_m es la distancia entre el micrófono m y al punto de referencia, ϕ_p es el ángulo de incidencia (AOI) correspondiente a la fuente S_p y V es la velocidad de propagación del sonido en el medio [111]. Remplazando $y_m(t - \tau_{p,m})$ por 3.1 en 3.2 y cambiando el orden de

suma, $b_p(t)$ se puede expresar como:

$$b_p(t) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} s_{k,m}(t - \tau_{p,m}) \tag{3.4}$$

Considere que, dada una fuente k, todos los $s_{k,m}$ tiene la misma energía para $0 \leq m < M$. Esta es una suposición razonable si la distancia entre el conjunto de micrófonos y la fuente de sonido es mayor que 5-10 veces la longitud del conjunto de micrófonos, como se mencionó anteriormente, y si los micrófonos son omnidireccionales. El esquema de beamforming de retardo y suma condicionará la energía de la señal k de la fuente recibida con una ganancia que depende de la dirección a la que apunta, que a su vez es una función de $\tau_{p,m}$:

$$b_p(t) = \sum_{k=0}^{K-1} g_{p,k} s_k(t) \tag{3.5}$$

dónde $g_{p,k}$ denota la ganancia de la fuente k cuando el beamforming apunta a la fuente p. El retraso de tiempo debido a la propagación desde la fuente hasta el punto de referencia se pueda omitir si se considera un escenario estático.

Considerando el caso de una fuente de audio deseada $s(t)$ y una fuente de ruido $n(t)$ es posible obtener una señal de beamforming para la señal de audio $b_0(t)$ y otra para el ruido $b_0(t)$, como se muestra en las figuras 3.1 y 3.1.

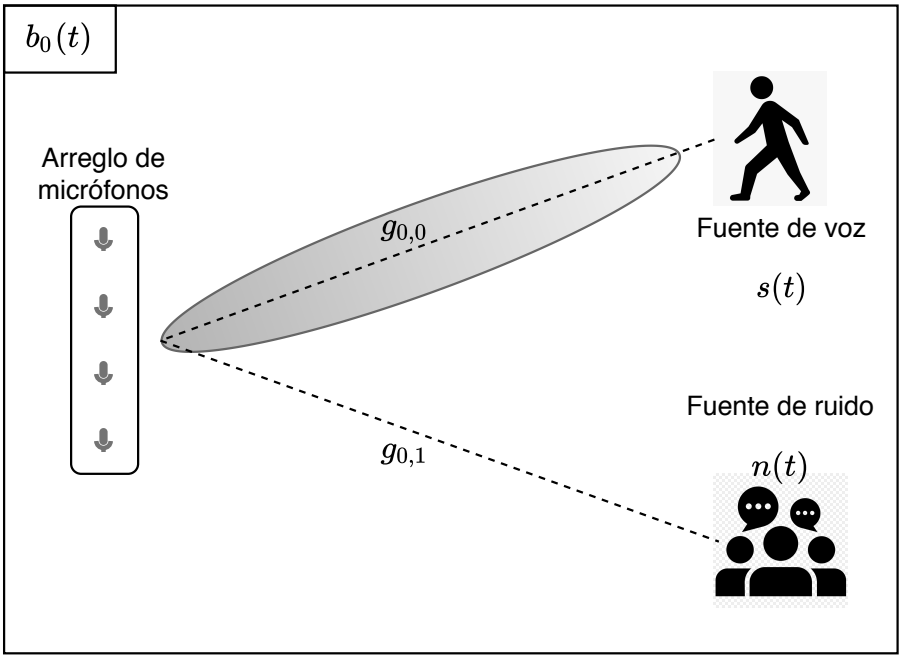

Figura 3.1: Esquema de beamforming apuntando a la fuente de voz.

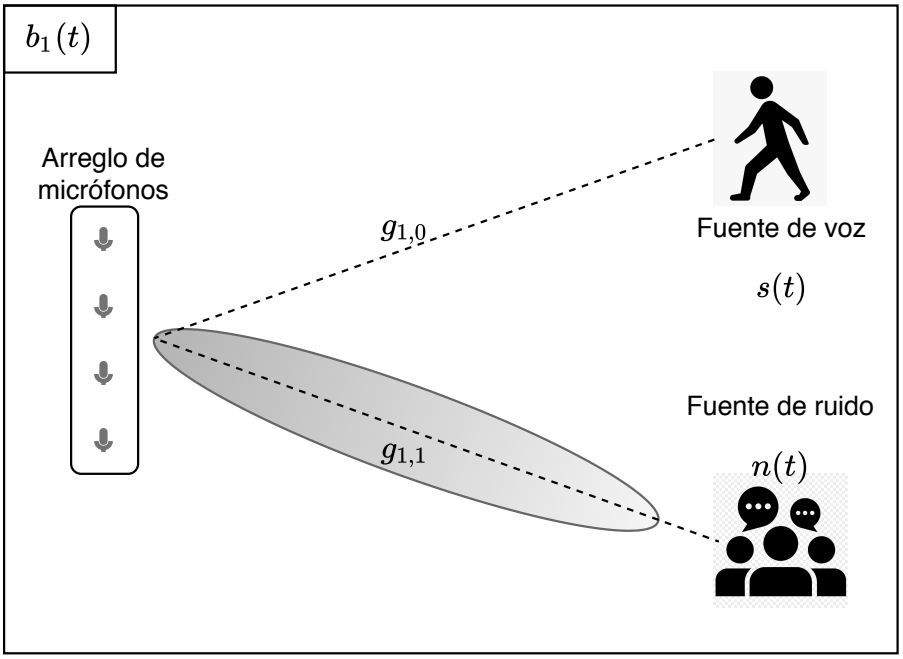

Figura 3.2: Esquema de beamforming apuntando a la fuente de ruido.

Aplicando la ecuacion 3.5 a este escenario, ambas contendrán una mezcla de ruido y señal a diferentes SNR como se muestra en la ecuación 3.6

$$\begin{aligned} b_0(t) &= g_{0,0}s(t) + g_{0,1}n(t) \\ b_1(t) &= g_{1,0}s(t) + g_{1,1}n(t) \end{aligned} \tag{3.6}$$

Este modelo es equivalente al modelo de mezcla utilizado por ICA y NMF. Sin embargo, en el caso dinámico las ganancias $g_{*,*}$ tendrán una dependencia temporal como se observa en la ecuación 3.7

$$\begin{aligned} b_0(t) &= g_{0,0}(t)s(t) + g_{0,1}(t)n(t) \\ b_1(t) &= g_{1,0}(t)s(t) + g_{1,1}(t)n(t) \end{aligned} \tag{3.7}$$

El mismo modelo (ec. 3.7) puede ser expresado en el dominio de la frecuencia (ec. 3.8), para la ventana k y la frecuencia ω.

$$\begin{aligned} B_0(k,\omega) &= g_{0,0}(k)S(k,\omega) + g_{0,1}(k)N(k,\omega) \\ B_1(k,\omega) &= g_{1,0}(k)S(k,\omega) + g_{1,1}(k)N(k,\omega) \end{aligned} \tag{3.8}$$

Alternativamente si se consideran señales normalizadas de manera que cada ventana k tenga potencia unitaria. Las señales normalizadas $\tilde{S}(k,\omega)$ y $\tilde{N}(k,\omega)$ quedan definidas por las ecuaciones 3.9 y 3.10.

$$\tilde{S}(k,\omega) = \frac{S(k,\omega)}{\sqrt{\sum_\omega |S(k,\omega)|^2}} \tag{3.9}$$

$$\tilde{N}(k,\omega) = \frac{N(k,\omega)}{\sqrt{\sum_\omega |N(k,\omega)|^2}} \tag{3.10}$$

Luego se puede definir un modelo de mezcla de señales normalizadas:

$$\begin{aligned} B_0(k,\omega) &= \tilde{g}_{0,0}(k)\tilde{S}(k,\omega) + \tilde{g}_{0,1}(k)\tilde{N}(k,\omega) \\ B_1(k,\omega) &= \tilde{g}_{1,0}(k)\tilde{S}(k,\omega) + \tilde{g}_{1,1}(k)\tilde{N}(k,\omega) \end{aligned} \tag{3.11}$$

Donde las ganancias $\tilde{g}_{*,*}(k)$ para cada ventana quedan definidas en función de las ganancias originales y la potencia de las señales originales según las ecuaciones en 3.12.

$$\begin{aligned} \tilde{g}_{0,0}(k) &= g_{0,0}(k)\sqrt{\textstyle\sum_\omega |S(k,\omega)|^2} \\ \tilde{g}_{0,1}(k) &= g_{0,1}(k)\sqrt{\textstyle\sum_\omega |N(k,\omega)|^2} \\ \tilde{g}_{1,0}(k) &= g_{1,0}(k)\sqrt{\textstyle\sum_\omega |S(k,\omega)|^2} \\ \tilde{g}_{1,1}(k) &= g_{1,1}(k)\sqrt{\textstyle\sum_\omega |N(k,\omega)|^2} \end{aligned} \tag{3.12}$$

La ventaja de este modelo es que separa la forma de la señal (ventana normalizada) de la potencia con la que se recibe, independiente de la razón por la que esta varia. Es decir no es necesario determinar si la variación de potencia es debido a una cambio en la ganancia o a la potencia con la que la señal fue emitida.

Esto permite analizar las señales por ventana independiente de si la señal tiene potencia variable o una ganancia variable. En el caso de utilizar aprendizaje automático, las señales con diferente ganancia y las señales con distinta potencia son matemáticamente equivalentes.

3.2. Solución del sistema de mezcla

3.2.1. Supresión de ruido

Una forma de obtener las señal de voz $S(t)$ a partir de la señal de beamforming $B_0(t)$ y $B_1(t)$ es utilizar un sistema de supresión de ruido. Esto se puede lograr entrenando una red neuronal que mapee la señal de beamforming a la señal de voz.

En este caso el problema de estimar $S(t)$ utilizando $B_0(t)$ o $B_1(t)$ son matemáticamente equivalente, siendo la única diferencia cuán severo es el SNR. Así es posible utilizar la misma red para ambos casos.

3.2.2. Separación de señales

3.2.2.1. Estimación de ganancia

Otra forma de obtener una estimación de $S(t)$ es estimar las ganancias $\tilde{g}$ y aplicar el modelo de ecuaciones en 3.11.

Es importante notar que la diferencia entre estimar $\tilde{g}_{*,0}$ y $\tilde{g}_{*,1}$ depende de las características de las señal de voz y la señal de ruido. Mientras que la diferencia entre estimar $\tilde{g}_{0,*}$ y $\tilde{g}_{1,*}$ depende solo del rango en el cual se mueven las ganancias y los SNR de ambas mezclas.

Así es posible entrenar una red que la recibir $B_0(t)$ estima $\tilde{g}_{0,0}$ y al recibir $B_1(t)$ estima $\tilde{g}_{1,0}$ y del mismo modo una segunda red puede estimar $\tilde{g}_{0,1}$ y $\tilde{g}_{1,1}$ con $B_0(t)$ y $B_1(t)$ respectivamente.

3.2.2.2. Solución del sistema de mezcla

Tomando el sistema de ecuaciones en 3.11 y resolviendo para $\tilde{S}$

$$\tilde{S}(k,\omega) = \frac{\frac{B_0(k,\omega)}{\tilde{g}_{0,1}(k)} - \frac{B_1(k,\omega)}{\tilde{g}_{1,1}(k)}}{\frac{\tilde{g}_{0,0}(k)}{\tilde{g}_{0,1}(k)} - \frac{\tilde{g}_{1,0}(k)}{\tilde{g}_{1,1}(k)}} \tag{3.13}$$

La ecuación 3.13 entrega un estimador de la señal normalizada, sin embargo si se desea utilizar la señal en un ASR es necesario que la señal estimada contenga información de la variación temporal de la potencia. Ya que al no conocer $g_{0,0}$ no es posible estimar S la alternativa es estimar $g_{0,0}S$

$$g_{0,0}(k)S(k,\omega) = \tilde{g}_{0,0}(k)\tilde{S}(k,\omega) \tag{3.14}$$

$$g_{0,0}(k)S(k,\omega) = \tilde{g}_{0,0}(k)\frac{\frac{B_0(k,\omega)}{\tilde{g}_{0,1}(k)} - \frac{B_1(k,\omega)}{\tilde{g}_{1,1}(k)}}{\frac{\tilde{g}_{0,0}(k)}{\tilde{g}_{0,1}(k)} - \frac{\tilde{g}_{1,0}(k)}{\tilde{g}_{1,1}(k)}} \tag{3.15}$$

3.2.2.3. Aproximación con ganancias de ruido

En la ecuación 3.15 se logra estimar la señal limpia salvo una constante multiplicativa a partir de la estimación de 4 parámetros y utilizando 2 redes neuronales. Dado esto es deseable obtener un estimador de similar calidad que requiera estimar menos parámetros. Para eso se considera el estimador:

$$\hat{S}(k,\omega) = B_0(k,\omega) - B_1(k,\omega)\frac{\tilde{g}_{0,1}(k)}{\tilde{g}_{1,1}(k)} \tag{3.16}$$

Remplazando 3.11 en 3.16

$$\hat{S}(k,\omega) = \tilde{S}(k,\omega)\Big(\tilde{g}_{0,0}(k) - \frac{\tilde{g}_{1,0}(k)\tilde{g}_{0,1}(k)}{\tilde{g}_{1,1}(k)}\Big) \tag{3.17}$$

Luego se remplaza 3.12 en 3.17

$$\hat{S}(k,\omega) = \tilde{S}(k,\omega)\sqrt{\sum_{\omega}|S(k,\omega)|^2}\Big(g_{0,0}(k) - \frac{g_{1,0}(k)g_{0,1}(k)}{g_{1,1}(k)}\Big) \tag{3.18}$$

Finalmente se remplaza 3.9 en 3.18

$$\hat{S}(k,\omega) = S(k,\omega)\Big(g_{0,0}(k) - \frac{g_{1,0}(k)g_{0,1}(k)}{g_{1,1}(k)}\Big) \tag{3.19}$$

Con esto el estimador es S multiplicado por una ganancia al igual que en la ecuación 3.15, sin embargo solo es necesario entrenar una red y estimar dos parámetros.

3.3. Red neuronal propuesta

3.3.1. Supresión de ruido

Para evaluar la técnica propuesta de supresión de ruido se propone entrenar una red que reciba B_* como entrada y entregue como salida una estimación de la señal limpia S, como se muestra en la figura 3.3.

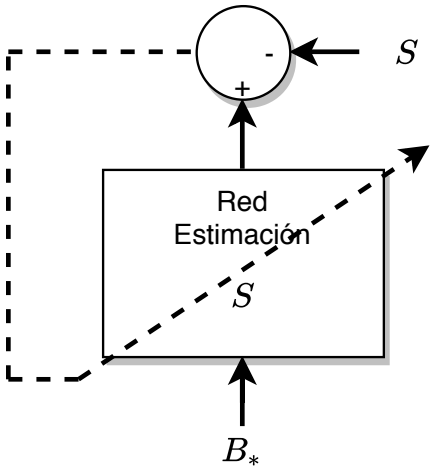

Figura 3.3: Entrenamiento DNN de supresión de ruido.

Una vez entrenada esta red se puede utilizar para eliminar el ruido de las señales de beamforming como se muestra en la figura 3.4

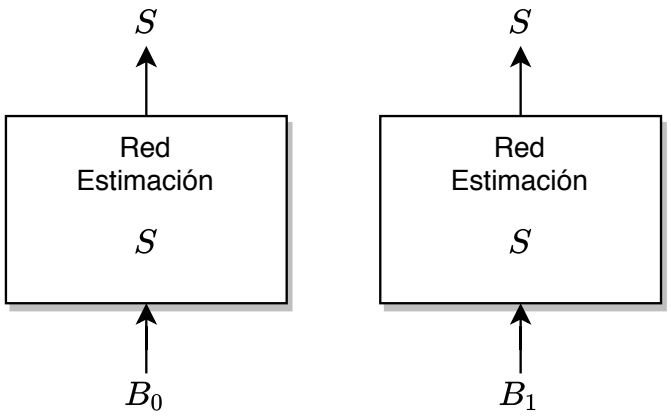

Figura 3.4: Aplicación de red de supresión de ruido sobre ambos beamformings para obtener S.

3.3.2. Estimación de ganancias

Para estimar las ganancias se entrena una red distinta para cada par de ganancias $\tilde{g}_{*,0}$ y $\tilde{g}_{*,1}$ como se muestra en la figura 3.5.

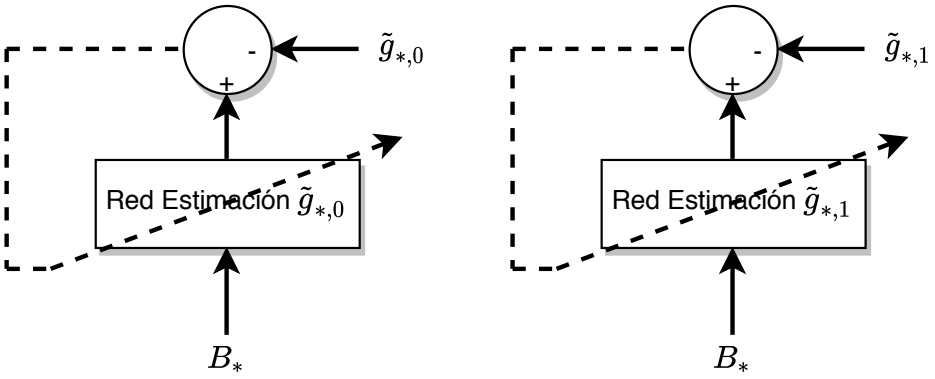

Figura 3.5: Entrenamiento DNNs de estimación de ganancias.

Luego al alimentar cada red con una señal de beamforming distinto se obtiene una de las ganancias como se muestra en la figura 3.6.

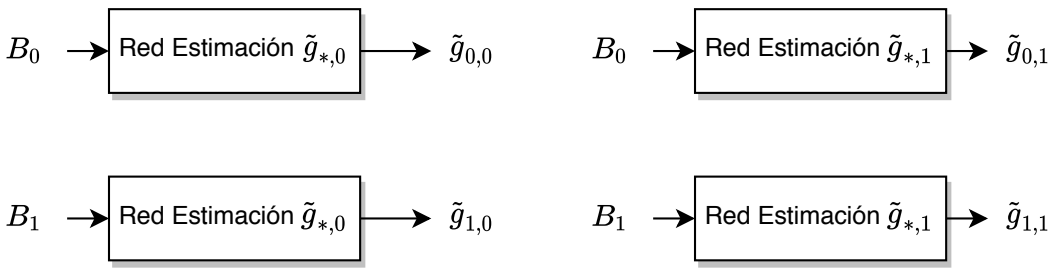

Figura 3.6: Aplicación de redes de de estimación de ganancias sobre múltiples beamformings para obtener las distintas ganancias.

Una vez obtenidas las ganancias se pueden utilizar para calcular una estimación de la señal limpia.

En la figura 3.7 se muestra el esquema completo para obtener la señal limpia a partir de ambas señales de beamforming. En azul se muestra como se procesa la señal de beamforming a la señal de voz B_0 y en rojo la señal de beamforming al ruido B_1. En líneas punteadas se muestran los elementos que se pueden omitir al utilizar la aproximación utilizando solo ganancias del ruido.

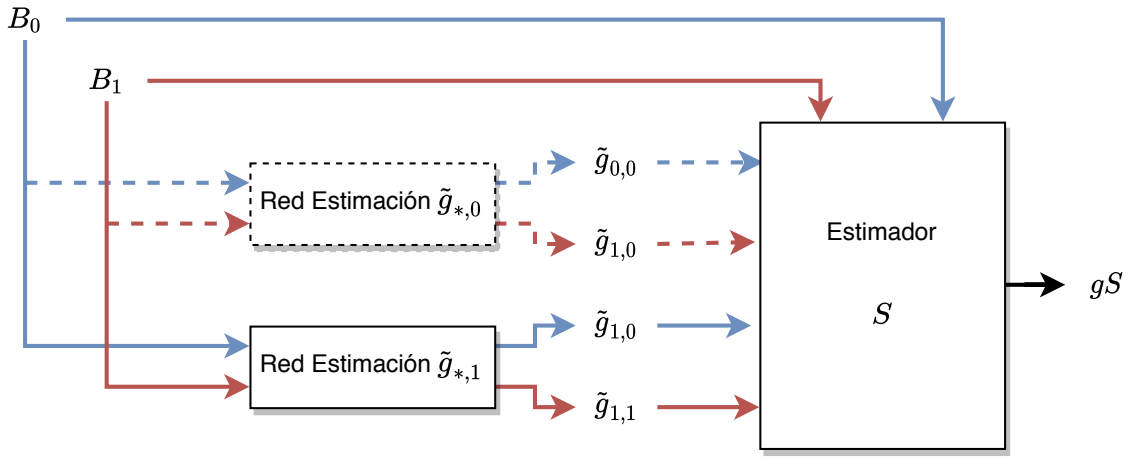

Figura 3.7: Esquema completo para estimar la señal limpia a partir de ambos beamformings utilizando redes de estimación de ganancias.

3.3.3. Arquitectura

Las DNN implementadas para evaluar las técnicas propuestas constan de tres capas ocultas y una capa lineal de salida. Las capas contienen 1024 unidades cada una. Los datos de entrenamiento ingresan a la red en batches (o "lotes") de 512 ventanas elegidos aleatoriamente de todo el conjunto de entrenamiento. Se utilizó el optimizador Adam [132] y la función de costo empleada fue el error cuadrático medio (MSE).

En la figura 3.8 se ilustra al arquitectura para las redes neuronales implementadas.

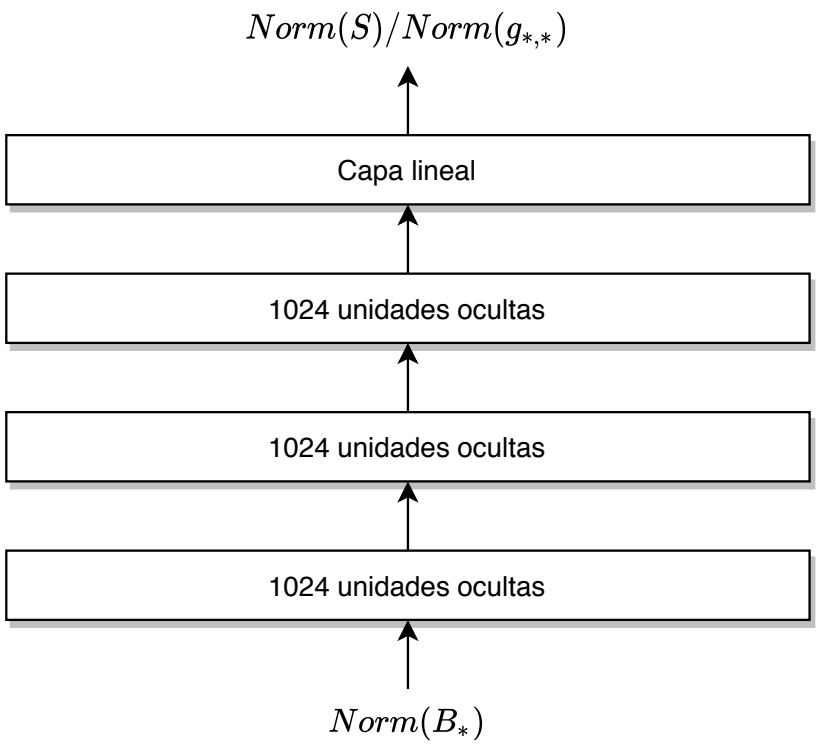

Figura 3.8: Arquitectura implementada en TensorFlow.

Para el pre-entrenamiento de las tres capas ocultas se utilizaron RBMs. Para la primera capa se utilizo un RBM Gaussian-Bernoulli (GBRBM) y una RBM Bernoulli-Bernoulli (BBRBM) para la segunda y tercera. La capa lineal se pre-entrenó de manera supervisada con el método de backpropagation implementado en Tensorflow.

Las señales fueron normalizadas para cada bin de frecuencia utilizando la potencia en el espacio lineal o normalización por media y varianza (MVN) en el espacio logarítmico.

Capítulo 4

Experimentos

4.1. Base de datos

4.1.1. Señales limpias

La bases de datos utilizadas en este trabajo están basadas en la base de datos Aurora-4 [78], [79]. Esta en su componente limpia contiene un conjunto de test compuesto de 7138 señales de audio de 83 locutores distintos, un conjunto de evaluación de 330 señales de 10 locutores y un conjunto de pruebas de 330 señales de otros 8 locutores. Las señales corresponden a frases en ingles leídas del Wall Street Journal y grabadas usando un micrófono Sennheiser HMD 414.

Aurora-4 generó apartir de la tarea de vocabulario de ciclo cerrado de 5000 palabras basada en el Corpus DARPA Wall Street Journal (WSJ0). El conjunto de entrenamiento limpio Aurora-4 corresponde a la base de datos WSJ0 SI-84 [133] de la CSR ARPA de noviembre de 1992 [134].

4.1.2. Ruido aditivo

Para simular el uso de dos beamformings en una situación con señales de voz y una fuente de ruido se utilizan la señales de limpias Aurora-4 y segmentos del ruido restaurant utilizado en Aurora-4. Este es un ruido que contiene múltiples eventos como murmullos de múltiples conversaciones, musica, ruidos de corta duración como choque de cubiertos contra platos, entre otros.

Para los datos de test se utilizaron SNR aleatorios con una distribución uniforme entre 5 dB y 15 dB para generar b_0, b_1 se genera con 3 dB menos que b_0, como se muestra en la figura 4.1.

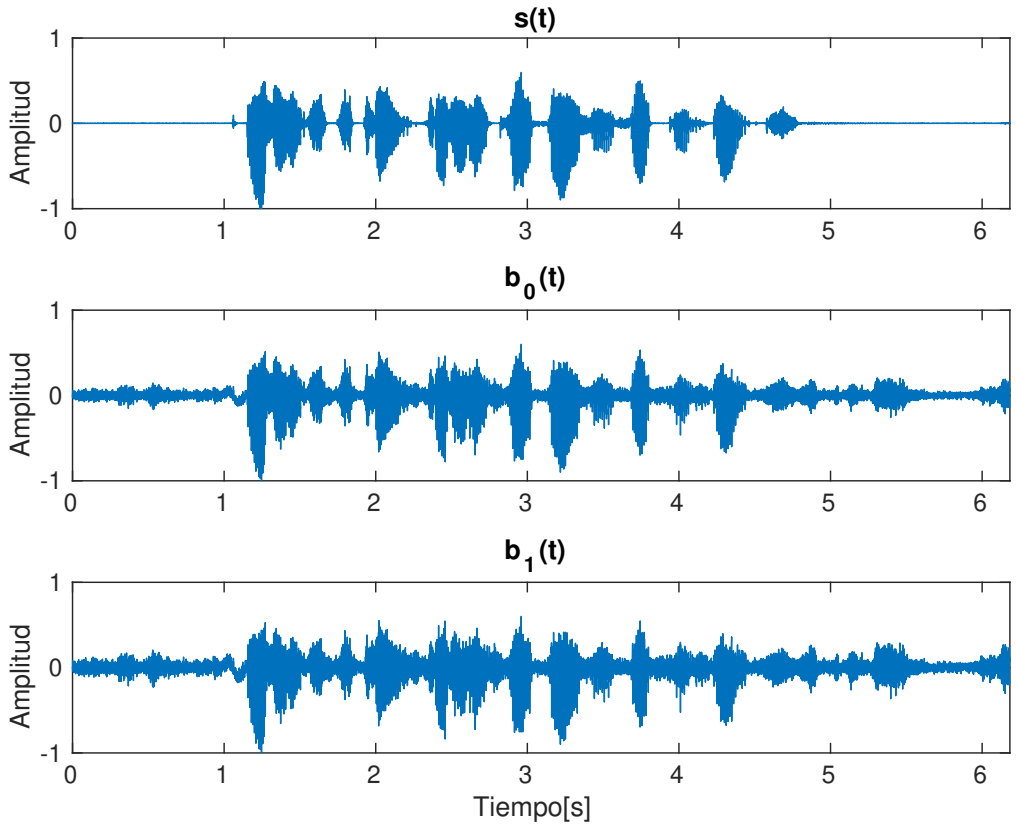

Figura 4.1: Ejemplo de una señal de test en el tiempo.

En el caso de datos de entrenamiento y evaluación se utilizaron SNR aleatorios con una distribución uniforme entre 0 dB y 15 dB para generar una única señal B_*. Este es un rango de SNR incluye los rangos de b_0 y b_1 en test y incluye casos con un ruido mas agresivo que en entrenamiento.

4.1.3. Datos en el espacio de frecuencia

Para cada señal $b_*(t)$ generada en la sección anterior se tiene el conjunto de señales en el tiempo $\{b_*(t), s_*(t), n_*(t)\}$ que cumplen la relación de la ecuación 4.1.

$$b_*(t) = s_*(t) + n_*(t) \tag{4.1}$$

Estas señales son separadas en ventanas de 25 ms con un traslape de 15 ms, que coincide con el tamaño de ventana utilizado en el ASR. Ya que las señales están muestreadas a 16000 Hz corresponde a ventanas de 400 muestras con 240 muestras de traslape. A cada una de estas ventanas se le aplica una transformada de fourier de 512 muestras, como se muestra en la figura 4.2. Luego se calculan las ganancias $\tilde{g}_{**}$ según las ecuaciones 4.2 y 4.3 como se muestra en la figura 4.3.

$$\tilde{g}_{*,0}(k) = \sqrt{\sum_{\omega} |S_*(k,\omega)|^2} \tag{4.2}$$

$$\tilde{g}_{*,1}(k) = \sqrt{\sum_{\omega} |N_*(k,\omega)|^2} \tag{4.3}$$

Con esto se genera el conjunto de datos Θ_* definido por la ecuación 4.4.

$$\Theta_*(k) = \left\{ B_*(k), S_*(k), \tilde{g}_{*,0}(k), \tilde{g}_{*,1}(k) \right\} \tag{4.4}$$

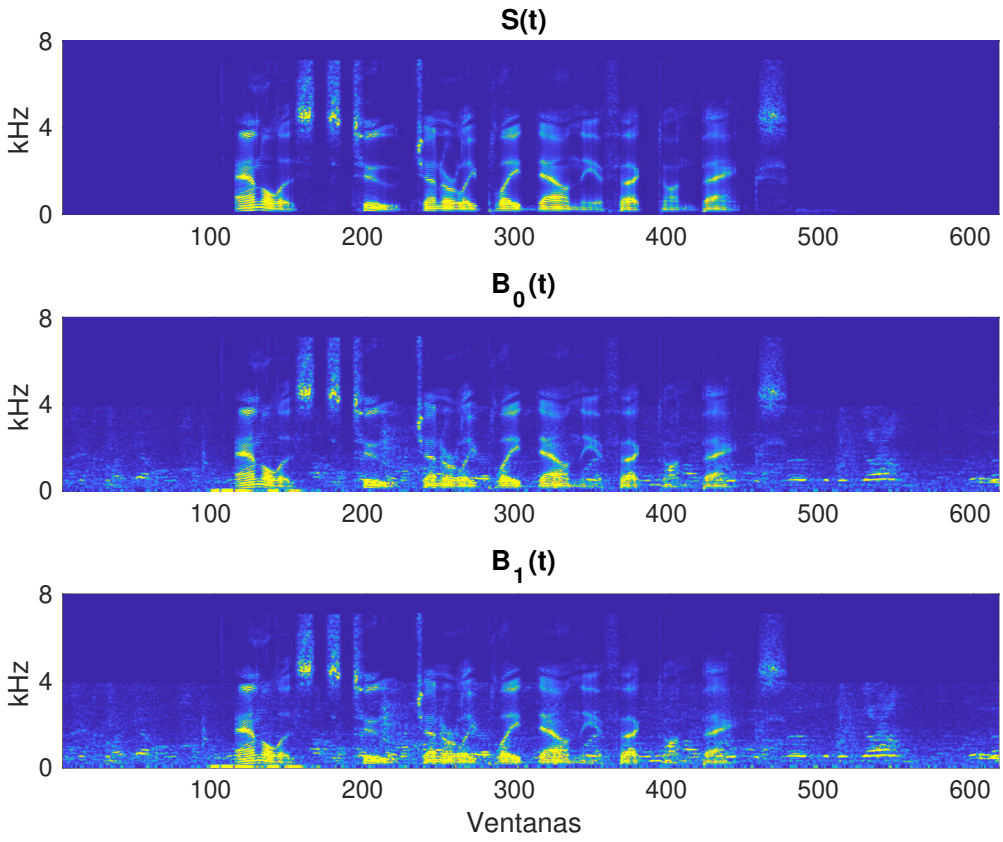

Figura 4.2: Ejemplo de una señal de test en el espacio de frecuencia.

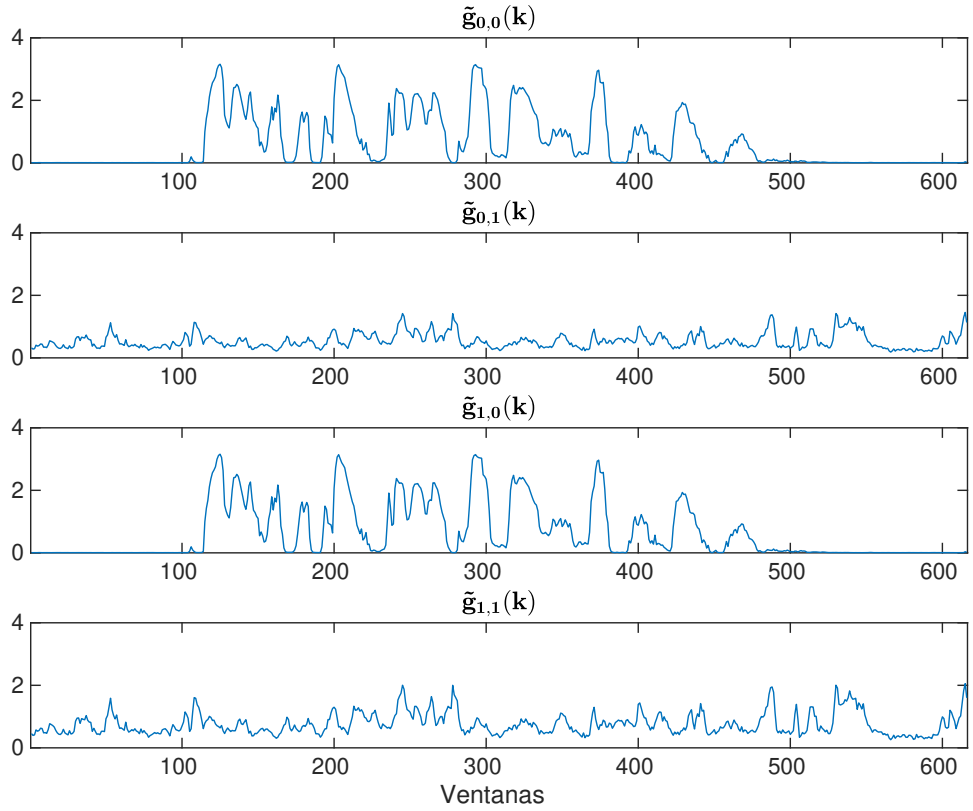

Figura 4.3: Ejemplo de ganancias para una señal de test.

4.2. Sistema de reconocimiento de voz utilizado

Los experimentos de reconocimiento de voz se realizaron con un ASR DNN-HMM utilizando el Kit de herramientas de reconocimiento de voz Kaldi [81]. Primero, se entrena un GMM-HMM de acuerdo con la receta tri2b Kaldi Aurora-4 con los datos de entrenamiento que se describen a continuación utilizando las características de MFCC, el análisis discriminante lineal (LDA) y las transformaciones lineales de máxima probabilidad (MLLT). Inicialmente se entrenó un sistema monófono; luego, las alineaciones de ese sistema se emplearon para generar un sistema de triphonemas inicial; finalmente, las alineaciones triphonema fueron empleadas para entrenar un sistema triphonema. El GMM en el sistema GMM-HMM entrenado se reemplazó con un DNN compuesto por siete capas ocultas y 2048 unidades por capa cada una, y la entrada considera una ventana de contexto de 11 frames. El número de unidades de la capa DNN de salida es igual al número de gaussianas en el sistema GMM-HMM correspondiente. El DNN fue entrenado usando las características del banco de filtros Mel (MelFB). El DNN se entrenó en primer lugar utilizando el criterio de entropía cruzada. Luego, el sistema final se obtiene al volver a entrenar el DNN con el entrenamiento discriminativo sMBR [135]. Para la decodificación, se utilizó el modelo de lenguaje estándar con el léxico de 5K y trigrama de la base de datos DARPA Wall Street Journal (WSJ) [136].

4.3. Resultados

4.3.1. Resultados preliminares

Como referencia se realizo una decodificación de las señales de test de ambos beamforming sin ninguna técnica de mejora y la señal limpia. La tabla 4.1 muestra como la adición de ruido genera un deterioro considerable del WER.

Tabla 4.1: Valores de referencia.

Señal	**WER**
Señal Limpia	2.19 %
B_0	15.9 %
B_1	27.89 %

4.3.2. Red supresión de ruido

Para evaluar la supresión de ruido utilizando redes neuronales se entrenaron dos redes utilizando normalización por potencia y por MVN en espacio logarítmico. La figura 4.4 muestra los resultados de aplicar estas redes a las señales B_0, comparados contra la decodificación de señales limpias y B_0. La tabla 4.2 muestra los valores de WER y las mejoras relativas al comparar con los resultados de B_0

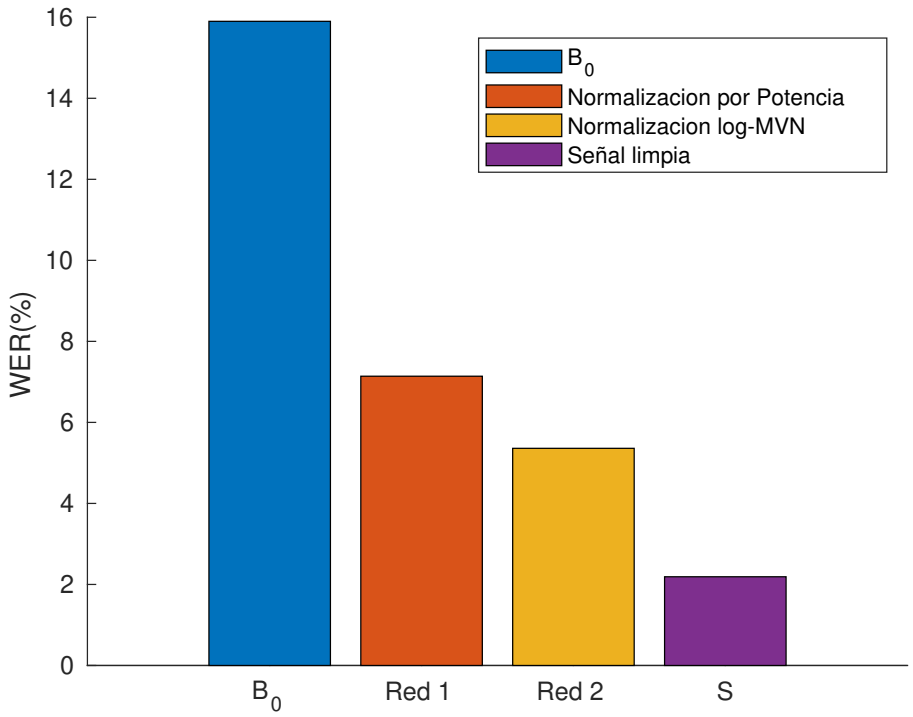

Figura 4.4: Resultados red de supresión de ruido.

El uso de estas redes permiten obtener mejoras de hasta 66 % al utilizar la normalización

Tabla 4.2: WER y mejoras red de supresión de ruido.

Señal	WER	Mejora
Normalizacion por Potencia	7.14 %	55.1 %
Normalizacion log-MVN	5.36 %	66.3 %

MVN en espacio logarítmico. Esta es la misma normalización que se utiliza en el ASR.

4.3.3. Análisis de componentes independientes y non-negative matrix factorization

Para fines de comparación, los experimentos de separación de fuentes se llevaron a cabo con ICA y NMF utilizando las herramientas ICAmatlab [137] y FASST [138] respectivamente.

En la figura 4.5 se muestran espectrogramas para las señales se ICA y NMF comparadas con las señales S y B_0. Se observa como ambas técnicas muestran una reducción considerable en los niveles de ruido.

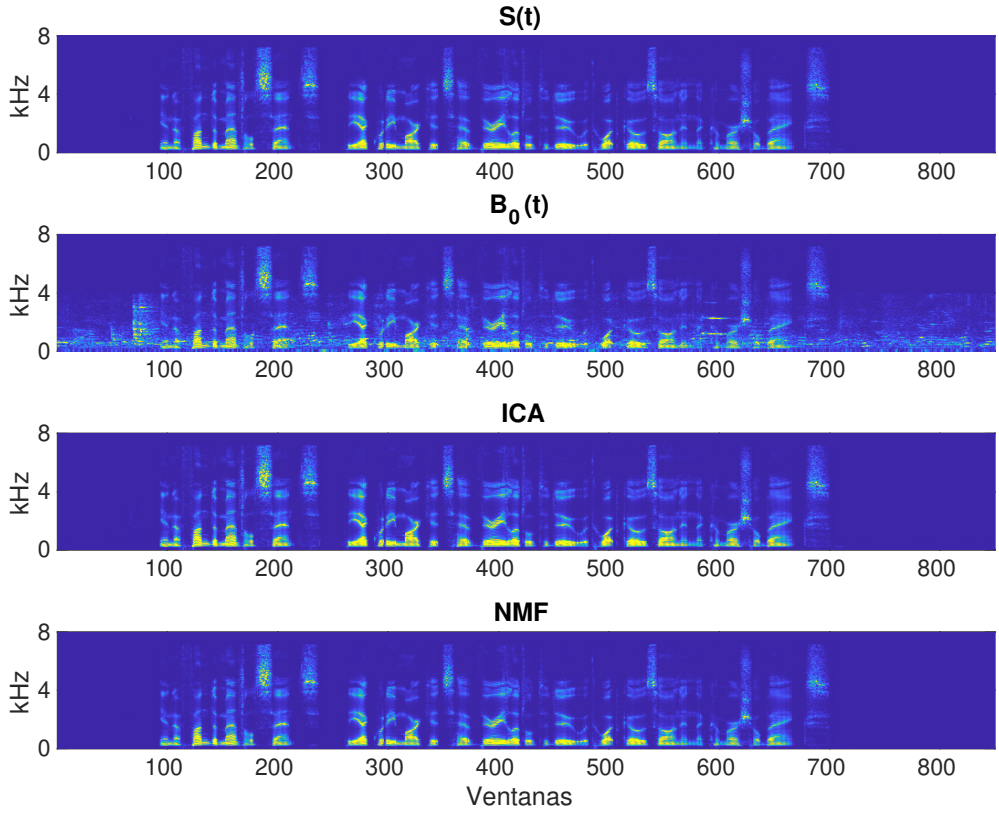

Figura 4.5: Espectrograma para la señal limpia, B_0, ICA y NMF

En la figura 4.6 se muestra los resultados de WER de ICA y NMF comparados con B_0 y la señal limpia. En la tabla 4.3 se muestran los resultados de WER y las mejoras relativas

respecto a B_0.

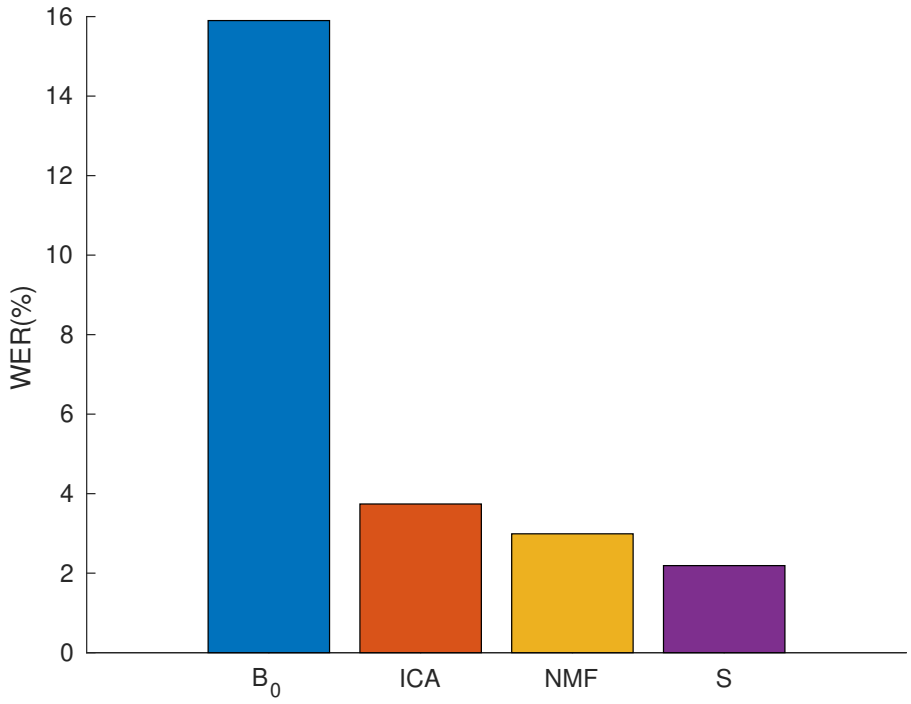

Figura 4.6: Resultados técnicas de separación de fuente.

Tabla 4.3: WER y mejoras técnicas de separación de fuente.

Señal	**WER**	**Mejora**
ICA	3.74 %	76.5 %
NMF	2.99 %	81.2 %

Se observa que las técnicas de separación de fuente entregan mejoras de hasta 81 %. Además se observa que estas técnicas entregan mejor resultado que el obtenido con las redes de supresión de ruido. En este caso se observan mejoras de 30 % para ICA y 44 % para NMF al compararlo con la mejor red de supresión de ruido.

4.3.4. Estimación de ganancias

Para evaluar el modelo de estimación de ganancia se entrenó la red para estimar $\tilde{g}_{*,0}$ y $\tilde{g}_{*,1}$ con normalización por potencia. En la figura 4.7 se muestran la estimación de $\tilde{g}_{*,0}$, y en la figura 4.8 para $\tilde{g}_{*,1}$. En ambos casos se observa cualitativamente como ambas redes son capaces de generar una estimación de los parámetros.

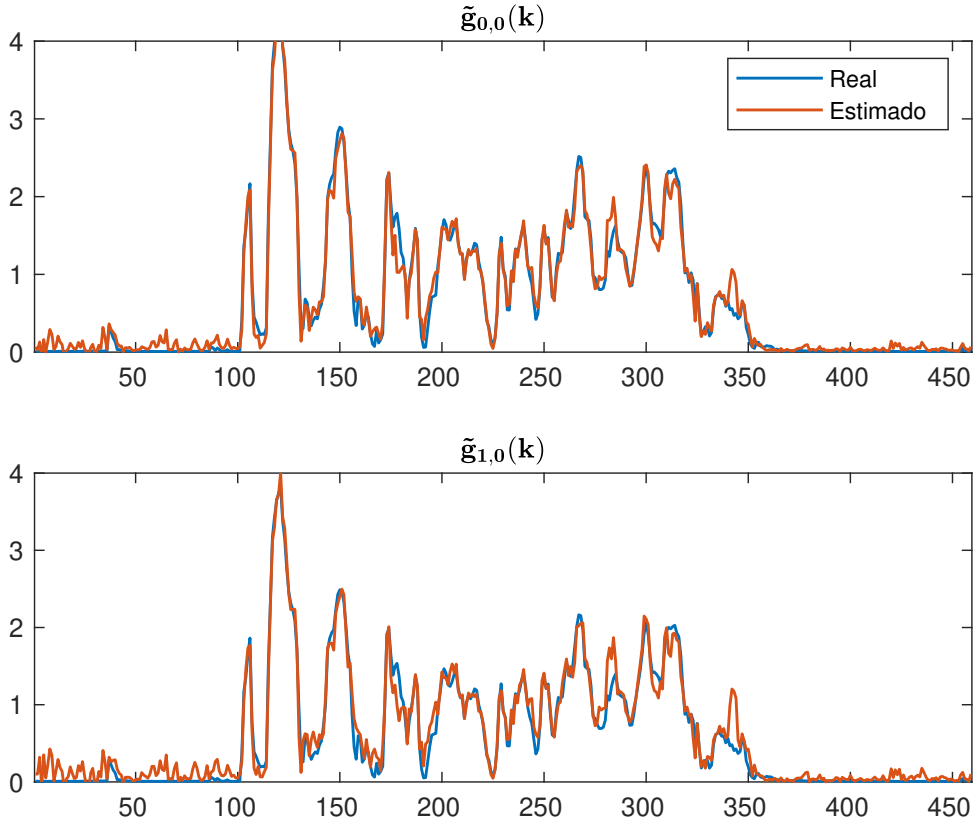

Figura 4.7: Resultados estimación de ganancias de la red de estimación $\tilde{g}_{*,0}$.

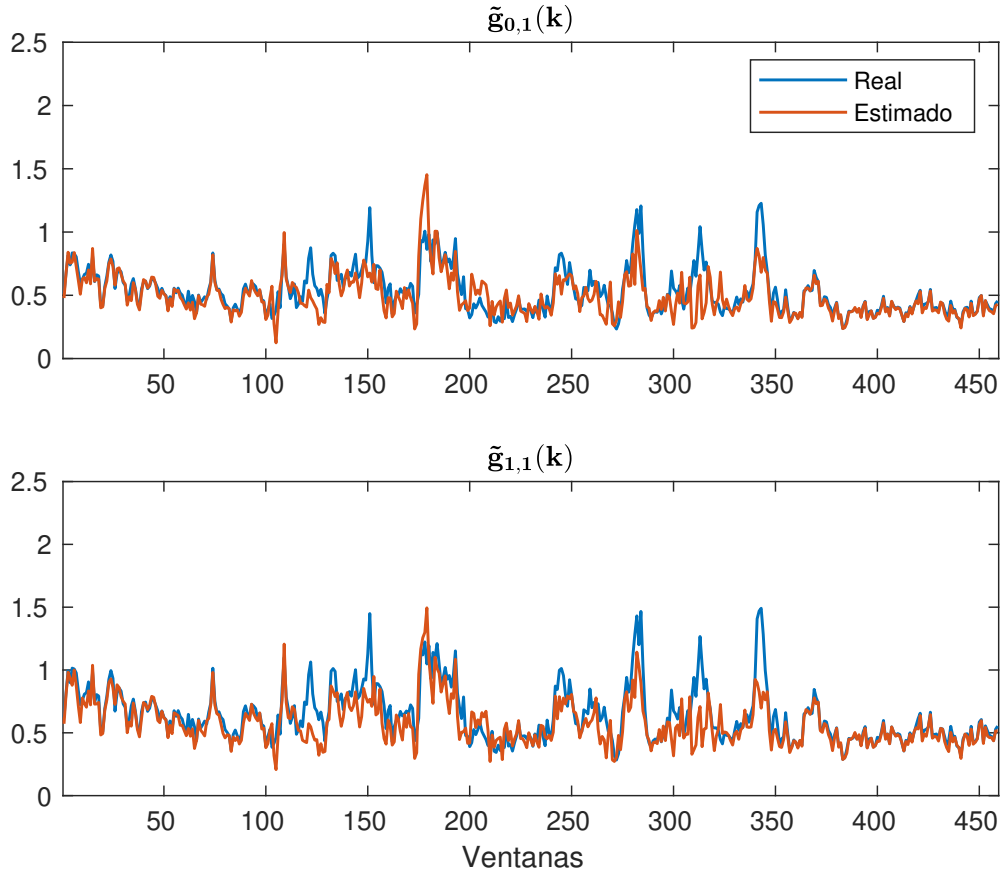

Figura 4.8: Resultados estimación de ganancias de la red de estimación $\tilde{g}_{*,1}$.

Con la red de estimacion para $\tilde{g}_{*,1}$ se obtuvieron los parámetros para aplicar el estimador de la ecuación 3.16. Considerando que el estimador esta modelado para aplicarlo sobre la FFT compleja y la red utiliza el modulo de la FFT como entrada, se evaluará también el efecto de aplicar el estimador sobre el modulo de la FFT.

En la figura 4.9 se muestran los espectrogramas para las señales obtenidas con este modelo. En ambos casos se muestra una reducción considerable en el nivel de ruido.

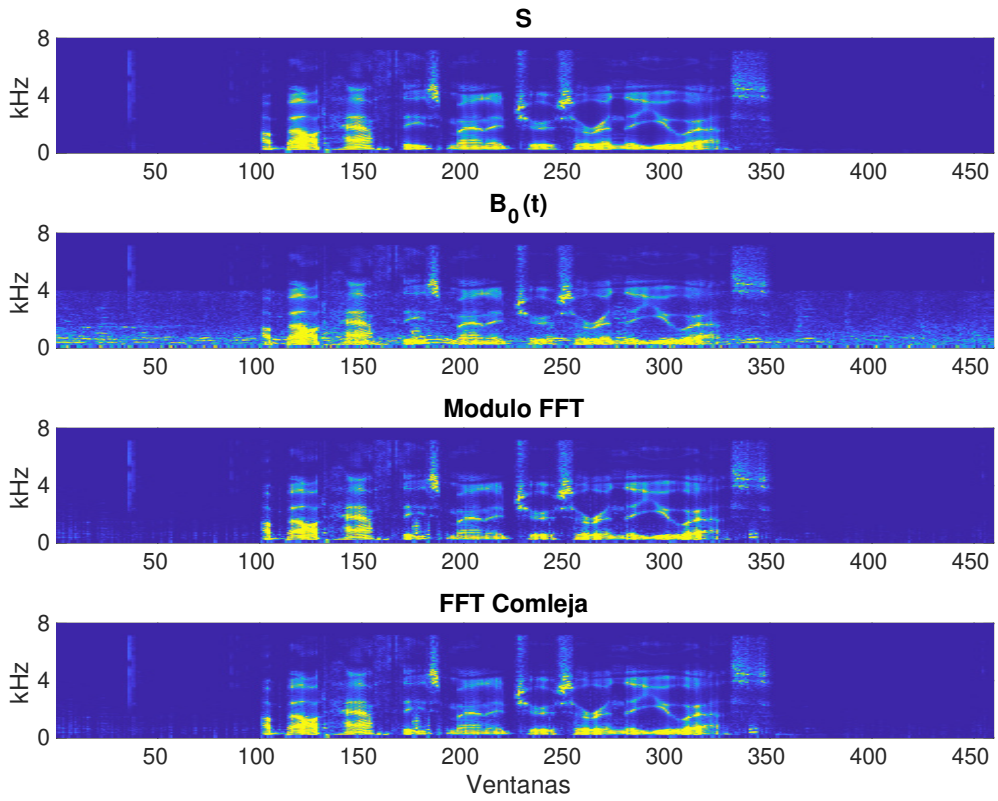

Figura 4.9: Espectrograma para la señal limpia, B_0, estimación con modulo FFT y FFT compleja

En la figura 4.10 se muestran los resultados de aplicar el estimador del modelo propuesto sobre la FFT y el modulo de la FFT con las ganancias estimadas. Estos resultados se comparan con la aplicación del modelo con las ganancias reales para obtener un resultado oráculo. La tabla 4.4 se muestran los valores de WER y las mejoras relativas.

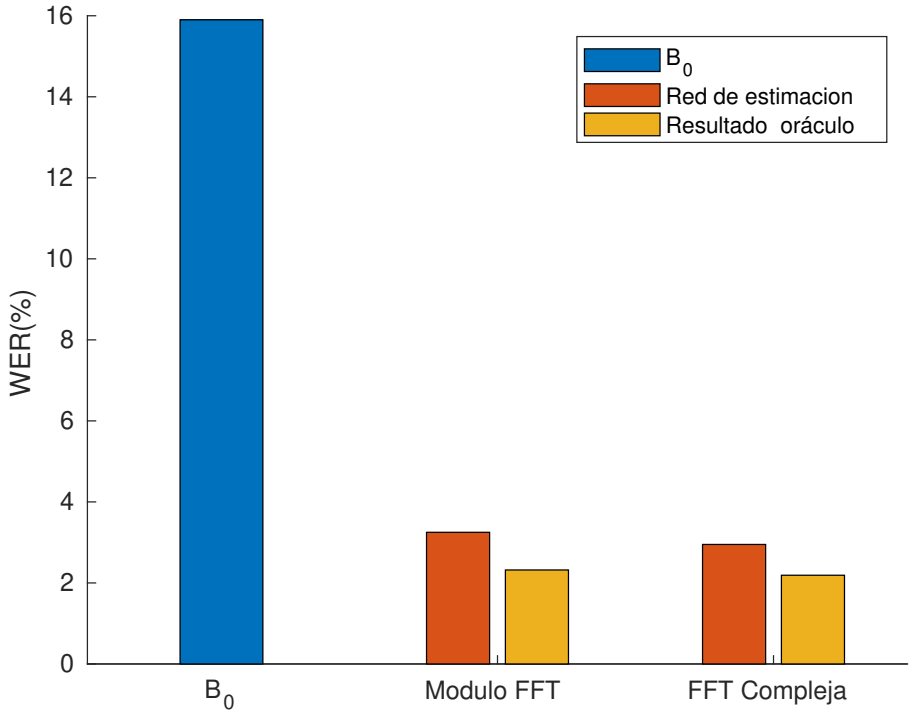

Figura 4.10: Resultados modelo de estimación de ganancias.

Tabla 4.4: WER y mejoras modelo de estimación de ganancias.

Señal	**WER estimación**	**WER oráculo**	**Mejora**
Modulo FFT	3.25 %	2.32 %	79.6 %
FFT Compleja	2.95 %	2.19 %	81.5 %

Utilizando el modelo propuesto se obtuvieron mejoras de hasta 81 %. Se observa que al generar un resultado oráculo para la FFT compleja se obtiene el mismo resultado que la señal limpia, lo que valida el uso de este estimador. El usos de el modulo de la FFT en el modelo degrada el resultado en un 6 % para el caso oráculo y en un 10 % con las ganancias estimadas. Esto comprueba la importancia de la fase en el modelo propuesto.

En la figura 4.11 se muestra un resumen de los resultados mas importantes. En esta se muestra la red de supresión de ruido con normalización log-MVN y el modelo de ganancias aplicado a la FFT compleja. Los resultados se muestran ordenados de mayor a menor WER.

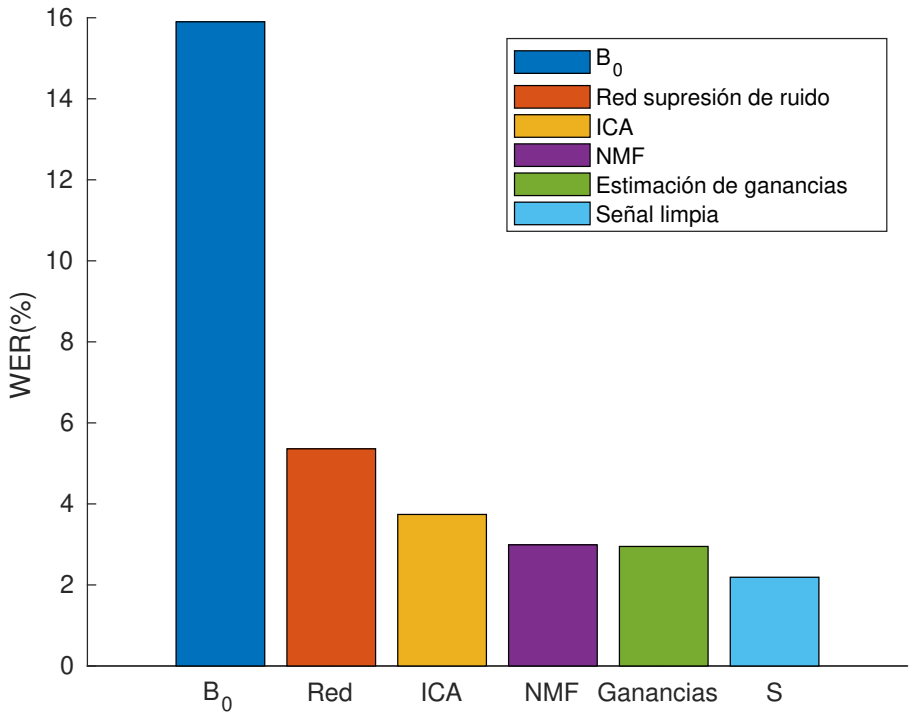

Figura 4.11: Resumen de los mejores resultados para cada técnica.

Se observa que la técnica propuesta entrega mejoras de 45 % respecto a la red de supresión de ruido, 21 % respecto a ICA y 1 % respecto a NMF. Este ultimo valor es bajo sin embargo es importante considerar que NMF se aplica por señales y la técnica propuesta se aplica por ventanas lo que permitiría usarla en situaciones donde las ganancias varían entre ventanas. Estas situaciones incluyen los casos donde las fuentes se mueven por ejemplo una persona en movimiento, casos donde los sensores se mueven por ejemplo un robot en movimiento o el ambiente cambia por ejemplo la persona y el robot entrando o saliendo de un edificio.

Capítulo 5

Conclusión

5.1. Conclusiones

En el presente trabajo se estudio el uso de múltiples beamforming y técnicas de separación de canal, para generar un ASR mas robusto. Esta metodología tiene utilidad en el contexto de la interacción humano robot, donde la comunicación verbal es de gran importancia.

Se analizó el escenario donde se tiene una fuente de voz y una fuente de ruido ambas con posición conocida. Esto considera el caso donde el robot utiliza sus cámaras para encontrar la posición de ambas fuentes. Así es posible establecer dos beamformings uno a la fuente de voz y otro a la fuente de ruido.

Considerando el escenario anterior se genero bases de datos para entrenamiento validación y test que simulen la situación descrita. Estas bases de datos fueron utilizadas para evaluar diversas metodologías para mejorar el reconocimiento de voz. Entre las técnicas de la literatura se encuentran el uso de redes neuronales para suprimir el ruido. Esta presentaron mejoras de hasta 66 %.

Luego se evaluaron las técnicas de reaparición de canal, como ICA y NMF, las cuales presentaron mejoras de hasta 81 %.

Finalmente se propuso un modelo de mezcla que considera ganancias variables en el tiempo y se utilizó este modelo para obtener mejoras de hasta un 81 %. Estas mejoras son similares a las de NMF, sin embargo la técnica propuesta tiene las ventajas de ser aplicado en ventanas de corto plazo mientras NMF requiere estadísticas de una señal mas larga. Esto causa que en condiciones variables en el tiempo como robots y personas en movimiento las estadísticas utilizadas por NMF varían mientras en el corto plazo utilizado por la técnica propuesta es posible aplicarlas.

5.2. Resumen

- Se logró generar una base de datos de test que simula la diferencia de SNR obtenidas al utilizar múltiples beamformings con un mismo arreglo de micrófonos.
- Se logró generar una base de datos de entrenamiento para redes neuronales que cubre las condiciones de los múltiples beamformings de la base de datos de test.
- Se elaboró un modelo de mezcla que describa el problema propuesto y se presentó una posible solución. Esa logró obtener mejores resultados que otras técnicas de supresión de ruido y separación de canal.
- La técnica desarrollada genero resultados de ASR mas robustos con menos datos de los requeridos por otras técnicas de separación de canal.

5.3. Trabajo futuro

En los resultados presentados se observa que aun existe espacio para mejorar el desempeño del sistema. En particular una estimación mas robusta de las ganancias permitiría acercarse a los resultados oráculo. Para esto es posible utilizar otras herramientas de deep learning como son las LSTM o CNN.

En este trabajo se presento un modelo para una fuente de voz y una de ruido. Se puede extender el modelo para abordar el caso de múltiples fuentes de voz o ruido.

Finalmente se puede abordar el caso de señales con reverberación, las cuales presentan una tarea desafiante para los sistemas de separación de señales.

Bibliografía

[1] B. Hayes y B. Scassellati, «Autonomously constructing hierarchical task networks for planning and human-robot collaboration,» en *2016 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2016, págs. 5469-5476.

[2] A. Bauer, D. Wollherr y M. Buss, «Human–robot collaboration: a survey,» *International Journal of Humanoid Robotics*, vol. 5, n.º 01, págs. 47-66, 2008.

[3] M. A. Goodrich y A. C. Schultz, *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[4] Y. Kondo, K. Takemura, J. Takamatsu y T. Ogasawara, «A gesture-centric android system for multi-party human-robot interaction,» *Journal of Human-Robot Interaction*, vol. 2, n.º 1, págs. 133-151, 2013.

[5] S. Yamamoto, K. Nakadai, M. Nakano et al., «Real-time robot audition system that recognizes simultaneous speech in the real world,» en *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2006, págs. 5333-5338.

[6] D. Wang, H. Leung, A. P. Kurian, H.-J. Kim y H. Yoon, «A deconvolutive neural network for speech classification with applications to home service robot,» *IEEE Transactions on Instrumentation and Measurement*, vol. 59, n.º 12, págs. 3237-3243, 2010.

[7] Y. Zhan, H. Leung, K.-C. Kwak y H. Yoon, «Automated speaker recognition for home service robots using genetic algorithm and Dempster–Shafer fusion technique,» *IEEE Transactions on Instrumentation and measurement*, vol. 58, n.º 9, págs. 3058-3068, 2009.

[8] H. G. Okuno, T. Ogata, K. Komatani y K. Nakadai, «Computational auditory scene analysis and its application to robot audition,» en *International Conference on Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004.*, IEEE, 2004, págs. 73-80.

[9] S. Rossi, F. Ferland y A. Tapus, «User profiling and behavioral adaptation for HRI: a survey,» *Pattern Recognition Letters*, vol. 99, págs. 3-12, 2017.

[10] S. Han, J. Hong, S. Jeong y M. Hahn, «Robust GSC-based speech enhancement for human machine interface,» *IEEE Transactions on Consumer Electronics*, vol. 56, n.º 2, págs. 965-970, 2010.

[11] G.-Z. Yang, J. Bellingham, P. E. Dupont et al., «The grand challenges of Science Robotics,» *Science robotics*, vol. 3, n.º 14, eaar7650, 2018.

[12] G. Sukthankar, C. Geib, H. H. Bui, D. Pynadath y R. P. Goldman, *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014.

[13] C. A. Corneanu, M. O. Simón, J. F. Cohn y S. E. Guerrero, «Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends,

and affect-related applications,» *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, n.º 8, págs. 1548-1568, 2016.

[14] J. K. Aggarwal y M. S. Ryoo, «Human activity analysis: A review,» *ACM Computing Surveys (CSUR)*, vol. 43, n.º 3, págs. 1-43, 2011.

[15] S. Rossi, E. Leone, M. Fiore, A. Finzi y F. Cutugno, «An extensible architecture for robust multimodal human-robot communication,» en *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, págs. 2208-2213.

[16] B. Bruno, F. Mastrogiovanni y A. Sgorbissa, «Wearable inertial sensors: Applications, challenges, and public test benches,» *IEEE Robotics & Automation Magazine*, vol. 22, n.º 3, págs. 116-124, 2015.

[17] C. Coppola, D. R. Faria, U. Nunes y N. Bellotto, «Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data,» en *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2016, págs. 5055-5061.

[18] M. Vázquez, A. Steinfeld y S. E. Hudson, «Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation,» en *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2015, págs. 3010-3017.

[19] O. D. Lara, A. J. Pérez, M. A. Labrador y J. D. Posada, «Centinela: A human activity recognition system based on acceleration and vital sign data,» *Pervasive and mobile computing*, vol. 8, n.º 5, págs. 717-729, 2012.

[20] Y. Demiris, «Prediction of intent in robotics and multi-agent systems,» *Cognitive processing*, vol. 8, n.º 3, págs. 151-158, 2007.

[21] B. Scassellati, «Theory of mind for a humanoid robot,» *Autonomous Robots*, vol. 12, n.º 1, págs. 13-24, 2002.

[22] A. G. Hofmann y B. C. Williams, «Intent Recognition for Human-Robot Interaction.,» en *Interaction Challenges for Intelligent Assistants*, 2007, págs. 60-61.

[23] W. Wahlster y A. Kobsa, «User models in dialog systems,» en *User models in dialog systems*, Springer, 1989, págs. 4-34.

[24] I. Zukerman y D. Litman, «Natural language processing and user modeling: Synergies and limitations,» *User modeling and user-adapted interaction*, vol. 11, n.º 1-2, págs. 129-158, 2001.

[25] G. Briggs y M. Scheutz, «Facilitating mental modeling in collaborative human-robot interaction through adverbial cues,» en *Proceedings of the SIGDIAL 2011 Conference*, 2011, págs. 239-247.

[26] C. L. McGhan, A. Nasir y E. M. Atkins, «Human intent prediction using markov decision processes,» *Journal of Aerospace Information Systems*, vol. 12, n.º 5, págs. 393-397, 2015.

[27] M. Awais y D. Henrich, «Proactive premature intention estimation for intuitive human-robot collaboration,» en *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, págs. 4098-4103.

[28] T. Liu, J. Wang y M. Q.-H. Meng, «Evolving hidden Markov model based human intention learning and inference,» en *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2015, págs. 206-211.

[29] V. Magnanimo, M. Saveriano, S. Rossi y D. Lee, «A bayesian approach for task recognition and future human activity prediction,» en *The 23rd IEEE international symposium on robot and human interactive communication*, IEEE, 2014, págs. 726-731.

[30] W. Y. Kwon e I. H. Suh, «A temporal bayesian network with application to design of a proactive robotic assistant,» en *2012 IEEE International Conference on Robotics and Automation*, IEEE, 2012, págs. 3685-3690.
[31] R. B. Ewen et al., *An introduction to theories of personality*. Psychology Press, 2014.
[32] L. W. Morris, *Extraversion and introversion: An interactional perspective*. Halsted Press, 1979.
[33] G. Mohammadi y A. Vinciarelli, «Automatic personality perception: Prediction of trait attribution based on prosodic features,» *IEEE Transactions on Affective Computing*, vol. 3, n.º 3, págs. 273-284, 2012.
[34] D. McColl, A. Hong, N. Hatakeyama, G. Nejat y B. Benhabib, «A survey of autonomous human affect detection methods for social robots engaged in natural HRI,» *Journal of Intelligent & Robotic Systems*, vol. 82, n.º 1, págs. 101-133, 2016.
[35] A. Vinciarelli y A. S. Pentland, «New social signals in a new interaction world: the next frontier for social signal processing,» *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, n.º 2, págs. 10-17, 2015.
[36] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter y A. Knoll, «Social behavior recognition using body posture and head pose for human-robot interaction,» en *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, págs. 2128-2133.
[37] D. McColl y G. Nejat, «Affect detection from body language during social HRI,» en *2012 IEEE RO-MAN: the 21st IEEE international symposium on robot and human interactive communication*, IEEE, 2012, págs. 1013-1018.
[38] I. Ariav e I. Cohen, «An end-to-end multimodal voice activity detection using wavenet encoder and residual networks,» *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, n.º 2, págs. 265-274, 2019.
[39] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa y T. Nakatani, «Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues.,» en *INTERSPEECH*, 2019, págs. 2718-2722.
[40] J. S. Chung, B.-J. Lee e I. Han, «Who said that?: Audio-visual speaker diarisation of real-world meetings,» *arXiv preprint arXiv:1906.10042*, 2019.
[41] L. Qu, C. Weber y S. Wermter, «LipSound: Neural Mel-Spectrogram Reconstruction for Lip Reading.,» en *INTERSPEECH*, 2019, págs. 2768-2772.
[42] Y. LeCun, Y. Bengio y G. Hinton, «Deep learning,» *nature*, vol. 521, n.º 7553, págs. 436-444, 2015.
[43] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.
[44] Y. Bengio, Y. LeCun et al., «Scaling learning algorithms towards AI,» *Large-scale kernel machines*, vol. 34, n.º 5, págs. 1-41, 2007.
[45] P. E. Utgoff y D. J. Stracuzzi, «Many-layered learning,» *Neural Computation*, vol. 14, n.º 10, págs. 2497-2529, 2002.
[46] G. E. Hinton, S. Osindero e Y.-W. Teh, «A fast learning algorithm for deep belief nets,» *Neural computation*, vol. 18, n.º 7, págs. 1527-1554, 2006.
[47] Y. Freund y D. Haussler, «Unsupervised learning of distributions on binary vectors using two layer networks,» en *Advances in neural information processing systems*, 1992, págs. 912-919.
[48] Y. Liu, G. Cao, Q. Sun y M. Siegel, «Hyperspectral classification via deep networks and superpixel segmentation,» *International Journal of Remote Sensing*, vol. 36, n.º 13, págs. 3459-3482, 2015.

[49] M. Ranzato, C. Poultney, S. Chopra e Y. L. Cun, «Efficient learning of sparse representations with an energy-based model,» en *Advances in neural information processing systems*, 2007, págs. 1137-1144.

[50] H. Mobahi, R. Collobert y J. Weston, «Deep learning from temporal coherence in video,» en *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, págs. 737-744.

[51] J. Weston, F. Ratle, H. Mobahi y R. Collobert, «Deep learning via semi-supervised embedding,» en *Neural networks: Tricks of the trade*, Springer, 2012, págs. 639-655.

[52] A. Ahmed, K. Yu, W. Xu, Y. Gong y E. Xing, «Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks,» en *European Conference on Computer Vision*, Springer, 2008, págs. 69-82.

[53] H. Larochelle, D. Erhan, A. Courville, J. Bergstra e Y. Bengio, «An empirical evaluation of deep architectures on problems with many factors of variation,» en *Proceedings of the 24th international conference on Machine learning*, 2007, págs. 473-480.

[54] G. E. Hinton y R. R. Salakhutdinov, «Using deep belief nets to learn covariance kernels for Gaussian processes,» en *Advances in neural information processing systems*, 2008, págs. 1249-1256.

[55] G. E. Hinton y R. R. Salakhutdinov, «Reducing the dimensionality of data with neural networks,» *science*, vol. 313, n.º 5786, págs. 504-507, 2006.

[56] S. Osindero y G. E. Hinton, «Modeling image patches with a directed hierarchy of Markov random fields,» en *Advances in neural information processing systems*, 2008, págs. 1121-1128.

[57] I. Levner, «Data driven object segmentation,» 2009.

[58] M. Ranzato y M. Szummer, «Semi-supervised learning of compact document representations with deep networks,» en *Proceedings of the 25th international conference on Machine learning*, 2008, págs. 792-799.

[59] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller e Y. LeCun, «Deep belief net learning in a long-range vision system for autonomous off-road driving,» en *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2008, págs. 628-633.

[60] R. Collobert y J. Weston, «A unified architecture for natural language processing: Deep neural networks with multitask learning,» en *Proceedings of the 25th international conference on Machine learning*, 2008, págs. 160-167.

[61] R. Salakhutdinov, A. Mnih y G. Hinton, «Restricted Boltzmann machines for collaborative filtering,» en *Proceedings of the 24th international conference on Machine learning*, 2007, págs. 791-798.

[62] L. Bahl, «Language-model/acoustic channel balance mechanism,» *IBM Technical Disclosure Bulletin*, vol. 23, n.º 7, págs. 3464-3465, 1980.

[63] D. E. Rumelhart, G. E. Hinton y R. J. Williams, «Learning representations by back-propagating errors,» *nature*, vol. 323, n.º 6088, págs. 533-536, 1986.

[64] S. Hochreiter, «Untersuchungen zu dynamischen neuronalen Netzen,» *Diploma, Technische Universität München*, vol. 91, n.º 1, 1991.

[65] M. A. Goodrich, A. C. Schultz et al., «Foundations and Trends in Human–Computer Interaction,» *Foundations and Trends in Human–Computer Interaction*, vol. 1, n.º 3, págs. 203-275, 2008.

[66] L. S. Lopes y A. Teixeira, «Human-robot interaction through spoken language dialogue,» en *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, IEEE, vol. 1, 2000, págs. 528-534.

[67] G. Hoffman y K. Vanunu, «Effects of robotic companionship on music enjoyment and agent perception,» en *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, págs. 317-324.

[68] C.-Y. Lin, K.-T. Song, Y.-W. Chen et al., «User identification design by fusion of face recognition and speaker recognition,» en *2012 12th International Conference on Control, Automation and Systems*, IEEE, 2012, págs. 1480-1485.

[69] K. Zheng, D. F. Glas, T. Kanda, H. Ishiguro y N. Hagita, «Designing and implementing a human–robot team for social interactions,» *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, n.º 4, págs. 843-859, 2013.

[70] E. L. Meszaros, M. Chandarana, A. Trujillo y B. D. Allen, «Compensating for limitations in speech-based natural language processing with multimodal interfaces in uav operation,» en *International Conference on Applied Human Factors and Ergonomics*, Springer, 2017, págs. 183-194.

[71] M. Staudte y M. W. Crocker, «Investigating joint attention mechanisms through spoken human–robot interaction,» *Cognition*, vol. 120, n.º 2, págs. 268-291, 2011.

[72] H. A. Polido, «DARPA Robotics Challenge,» 2014.

[73] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa y H. Matsubara, «RoboCup: A challenge problem for AI,» *AI magazine*, vol. 18, n.º 1, págs. 73-73, 1997.

[74] T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee y T. J. Sejnowski, «Imaging brain dynamics using independent component analysis,» *Proceedings of the IEEE*, vol. 89, n.º 7, págs. 1107-1122, 2001.

[75] W. Burger y M. J. Burge, *Digital image processing: an algorithmic introduction using Java*. Springer Science & Business Media, 2009.

[76] J. Nakamura, *Image sensors and signal processing for digital still cameras*. CRC press, 2017.

[77] J. Novoa, J. P. Escudero, J. Fredes, J. Wuth, R. Mahu y N. B. Yoma, «Multichannel Robot Speech Recognition Database: MChRSR,» *arXiv preprint arXiv:1801.00061*, 2017.

[78] G. Hirsch, «Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task,» *ETSI STQ Aurora DSR Working Group, Dec. 2002*, 2002.

[79] D. Pearce y J. Picone, «Aurora working group: DSR front end LVCSR evaluation AU/384/02,» *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[80] J. Novoa, J. Wuth, J. P. Escudero et al., «Robustness Over Time-Varying Channels in DNN-HMM ASR Based Human-Robot Interaction.,» en *Proceedings of Interspeech*, 2017.

[81] D. Povey, A. Ghoshal, G. Boulianne et al., «The Kaldi speech recognition toolkit,» en *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[82] J. Novoa, J. Wuth, J. P. Escudero, J. Fredes, R. Mahu y N. B. Yoma, «DNN-HMM based automatic speech recognition for HRI scenarios,» en *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2018, págs. 150-159.

[83] H. G. Hirsch, «Fant-filtering and noise adding tool,» *Niederrhein University of Applied Sciences, http://dnt.-kr. hsnr. de/download. html*, 2005.

[84] S. Young, «HMMs and related speech recognition technologies,» en *Springer Handbook of Speech Processing*, Springer, 2008, págs. 539-558.

[85] X. D. Huang, Y. Ariki y M. A. Jack, *Hidden Markov models for speech recognition*. Edinburgh University Press, 1990, vol. 2004.

[86] R. Justo y M. I. Torres, «Integration of complex language models in ASR and LU systems,» *Pattern Analysis and Applications*, vol. 18, n.º 3, págs. 493-505, 2015.

[87] B. H. Juang y L. R. Rabiner, «Hidden Markov models for speech recognition,» *Technometrics*, vol. 33, n.º 3, págs. 251-272, 1991.

[88] M. Chetouani, B. Gas, J. Zarader y C. Chavy, «Discriminative training for neural predictive coding applied to speech features extraction,» en *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 1, 2002, págs. 852-857.

[89] N. Dave, «Feature extraction methods LPC, PLP and MFCC in speech recognition,» *International journal for advance research in engineering and technology*, vol. 1, n.º 6, págs. 1-4, 2013.

[90] S. Furui, «Speaker-independent isolated word recognition using dynamic features of speech spectrum,» *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, n.º 1, págs. 52-59, 1986.

[91] G. Hinton, L. Deng, D. Yu et al., «Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,» *IEEE Signal processing magazine*, vol. 29, n.º 6, págs. 82-97, 2012.

[92] J. J. Godfrey y E. Holliman, «Switchboard-1 Release 2,» *Linguistic Data Consortium, Philadelphia*, vol. 926, pág. 927, 1997.

[93] J. Schröder, J. Anemüller y S. Goetze, «Performance comparison of GMM, HMM and DNN based approaches for acoustic event detection within task 3 of the DCASE 2016 challenge,» en *Proc. Workshop Detect. Classification Acoust. Scenes Events*, 2016, págs. 80-84.

[94] B. Li, T. N. Sainath, A. Narayanan et al., «Acoustic Modeling for Google Home.,» en *Interspeech*, 2017, págs. 399-403.

[95] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang y A. Stolcke, «The Microsoft 2017 conversational speech recognition system,» en *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, págs. 5934-5938.

[96] S. Hochreiter y J. Schmidhuber, «Long short-term memory,» *Neural computation*, vol. 9, n.º 8, págs. 1735-1780, 1997.

[97] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn y D. Yu, «Convolutional neural networks for speech recognition,» *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, n.º 10, págs. 1533-1545, 2014.

[98] A. Graves, A.-r. Mohamed y G. Hinton, «Speech recognition with deep recurrent neural networks,» en *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, págs. 6645-6649.

[99] A. Tsilfidis, I. Mporas, J. Mourjopoulos y N. Fakotakis, «Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing,» *Computer Speech & Language*, vol. 27, n.º 1, págs. 380-395, 2013.

[100] J. Li, A. Mohamed, G. Zweig e Y. Gong, «LSTM time and frequency recurrence for automatic speech recognition,» en *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, IEEE, 2015, págs. 187-191.

[101] T. N. Sainath y B. Li, «Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks,» 2016.

[102] Y. Liu y K. Kirchhoff, «Novel Front-End Features Based on Neural Graph Embeddings for DNN-HMM and LSTM-CTC Acoustic Modeling.,» en *INTERSPEECH*, 2016, págs. 793-797.

[103] Y. Zhao, D. Wang, B. Xu y T. Zhang, «Late reverberation suppression using recurrent neural networks with long short-term memory,» en *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, págs. 5434-5438.

[104] Y. Qian, M. Bi, T. Tan y K. Yu, «Very deep convolutional neural networks for noise robust speech recognition,» *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, n.º 12, págs. 2263-2276, 2016.

[105] V. Mitra y H. Franco, «Coping with Unseen Data Conditions: Investigating Neural Net Architectures, Robust Features, and Information Fusion for Robust Speech Recognition.,» en *INTERSPEECH*, 2016, págs. 3783-3787.

[106] C. Weng, D. Yu, M. L. Seltzer y J. Droppo, «Single-channel mixed speech recognition using deep neural networks,» en *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, págs. 5632-5636.

[107] M. Brandstein y D. Ward, «Microphone arrays: signal processing techniques and applications,» en Springer Science & Business Media, 2013, cap. 2, págs. 19-38.

[108] B. D. Van Veen y K. M. Buckley, «Beamforming: A versatile approach to spatial filtering,» *IEEE assp magazine*, vol. 5, n.º 2, págs. 4-24, 1988.

[109] M. Omologo, M. Matassoni y P. Svaizer, «Speech recognition with microphone arrays,» en *Microphone arrays*, Springer, 2001, cap. 15, págs. 331-353.

[110] M. Brandstein y D. Ward, «Microphone arrays: signal processing techniques and applications,» en Springer Science & Business Media, 2001, cap. 2, págs. 19-38.

[111] M. Crocco y A. Trucco, «Stochastic and analytic optimization of sparse aperiodic arrays and broadband beamformers with robust superdirective patterns,» *IEEE Transactions on audio, speech, and language processing*, vol. 20, n.º 9, págs. 2433-2447, 2012.

[112] J. Novoa, R. Mahu, A. Díaz, J. Wuth, R. Stern y N. B. Yoma, «Weighted delay-and-sum beamforming guided by visual tracking for human-robot interaction,» *arXiv preprint arXiv:1906.07298*, 2019.

[113] A. Díaz, R. Mahu, J. Novoa, J. Wuth, J. Datta y N. B. Yoma, «Assessing the effect of visual servoing on the performance of linear microphone arrays in moving human-robot interaction scenarios,» *Computer Speech & Language*, vol. 65, pág. 101 136, 2020.

[114] P. Comon, «Independent component analysis, a new concept?» *Signal processing*, vol. 36, n.º 3, págs. 287-314, 1994.

[115] L. E. Di Persia y D. H. Milone, «Using multiple frequency bins for stabilization of FD-ICA algorithms,» *Signal Processing*, vol. 119, págs. 162-168, 2016.

[116] D. D. Sai, K. Kishor y K. S. R. Murty, «Speech Source Separation using ICA in Constant Q Transform Domain,»

[117] A. K. Barros, T. Rutkowski, F. Itakura y N. Ohnishi, «Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets,» *IEEE Transactions on Neural Networks*, vol. 13, n.º 4, págs. 888-893, 2002.

[118] J.-T. Chien y B.-C. Chen, «A new independent component analysis for speech recognition and separation,» *IEEE transactions on audio, speech, and language processing*, vol. 14, n.º 4, págs. 1245-1254, 2006.

[119] D. Kolossa, R. Fernandez Astudillo, E. Hoffmann y R. Orglmeister, «Independent component analysis and time-frequency masking for speech recognition in multitalker conditions,» *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, págs. 1-13, 2010.

[120] S. Makeig, A. J. Bell, T.-P. Jung y T. J. Sejnowski, «Independent component analysis of electroencephalographic data,» en *Advances in neural information processing systems*, 1996, págs. 145-151.

[121] S. Ikeda y K. Toyama, «Independent component analysis for noisy data—MEG data analysis,» *Neural Networks*, vol. 13, n.º 10, págs. 1063-1074, 2000.

[122] M. S. Bartlett, J. R. Movellan y T. J. Sejnowski, «Face recognition by independent component analysis,» *IEEE Transactions on neural networks*, vol. 13, n.º 6, págs. 1450-1464, 2002.

[123] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca y R. J. Plemmons, «Algorithms and applications for approximate nonnegative matrix factorization,» *Computational statistics & data analysis*, vol. 52, n.º 1, págs. 155-173, 2007.

[124] D.-x. Wang, M.-s. Jiang, F.-l. Niu, Y.-d. Cao y C.-x. Zhou, «Speech enhancement control design algorithm for dual-microphone systems using β-NMF in a complex environment,» *Complexity*, vol. 2018, 2018.

[125] A. Ozerov y C. Févotte, «Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,» *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n.º 3, págs. 550-563, 2009.

[126] C. Narisetty, «A Unified Bayesian Source Modelling for Determined Blind Source Separation.,» en *INTERSPEECH*, 2019, págs. 1343-1347.

[127] L. Wang, H. Ding y F. Yin, «Speech separation and extraction by combining superdirective beamforming and blind source separation,» en *Blind Source Separation*, Springer, 2014, págs. 323-348.

[128] H. Barfuss, K. Reindl y W. Kellermann, «Informed Spatial Filtering Based on Constrained Independent Component Analysis,» en *Audio Source Separation*, Springer, 2018, págs. 237-278.

[129] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee y K. Shikano, «Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,» *IEEE Transactions on Audio, speech, and language processing*, vol. 14, n.º 2, págs. 666-678, 2006.

[130] T. Nishikawa, H. Saruwatari, K. Shikano, S. Araki y S. Makino, «Multistage ICA for blind source separation of real acoustic convolutive mixture,» 2003.

[131] I. J. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons, 2009.

[132] D. P. Kingma y J. Ba, «Adam: A method for stochastic optimization,» *arXiv preprint arXiv:1412.6980*, 2014.

[133] D. B. Paul y J. Baker, «The design for the Wall Street Journal-based CSR corpus,» en *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[134] P. C. Woodland, J. J. Odell, V. Valtchev y S. J. Young, «Large vocabulary continuous speech recognition using HTK,» en *Proceedings of ICASSP'94. IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing, IEEE, vol. 2, 1994, págs. II-125.

[135] K. Veselý, A. Ghoshal, L. Burget y D. Povey, «Sequence-discriminative training of deep neural networks.,» en *Interspeech*, vol. 2013, 2013, págs. 2345-2349.

[136] J.-L. Gauvain, L. Lamel y M. Adda-Decker, «Developments in continuous speech dictation using the ARPA WSJ task,» en *1995 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, 1995, págs. 65-68.

[137] B. Moore, *PCA and ICA Package*, MATLAB Central File Exchange, 2020. dirección: `https://www.mathworks.com/matlabcentral/fileexchange/38300-pca-and-ica-package`.

[138] A. Ozerov y E. Vincent, «Using the FASST source separation toolbox for noise robust speech recognition,» en *Machine Listening in Multisource Environments*, 2011.

[139] J. P. Escudero, V. Poblete, J. Novoa et al., «Highly-Reverberant Real Environment database: HRRE,» *arXiv preprint arXiv:1801.09651*, 2018.

[140] J. P. Escudero, J. Novoa, R. Mahu et al., «An improved DNN-based spectral feature mapping that removes noise and reverberation for robust automatic speech recognition,» *arXiv preprint arXiv:1803.09016*, 2018.

[141] J. Novoa, R. Mahu, J. Wuth, J. P. Escudero, J. Fredes y N. B. Yoma, «Automatic Speech Recognition for Indoor HRI Scenarios,» *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, n.º 2, págs. 1-30, 2021.

[142] A. Díaz, D. Pincheira, R. Mahu y N. B. Yoma, «Short-time deep-learning based source separation for speech enhancement in reverberant environments with beamforming,» *arXiv preprint arXiv:2011.01965*, 2020.

Anexos

Anexo A

Publicaciones del autor

A.1. Trabajo previo en ASR y HRI

A.1.1. Publicaciones de conferencia

2017 J. Novoa, J. Wuth, J. P. Escudero et al., «Robustness Over Time-Varying Channels in DNN-HMM ASR Based Human-Robot Interaction.,» en *Proceedings of Interspeech*, 2017

2018 J. Novoa, J. Wuth, J. P. Escudero et al., «DNN-HMM based automatic speech recognition for HRI scenarios,» en *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2018, págs. 150-159

A.1.2. Publicaciones preprint

2017 J. Novoa, J. P. Escudero, J. Fredes et al., «Multichannel Robot Speech Recognition Database: MChRSR,» *arXiv preprint arXiv:1801.00061*, 2017

2018 J. P. Escudero, V. Poblete, J. Novoa et al., «Highly-Reverberant Real Environment database: HRRE,» *arXiv preprint arXiv:1801.09651*, 2018

2018 J. P. Escudero, J. Novoa, R. Mahu et al., «An improved DNN-based spectral feature mapping that removes noise and reverberation for robust automatic speech recognition,» *arXiv preprint arXiv:1803.09016*, 2018

A.2. Trabajo durante el Magister

A.2.1. Publicaciones en revistas del Institute for Scientific Information (ISI)

2020 A. Díaz, R. Mahu, J. Novoa et al., «Assessing the effect of visual servoing on the performance of linear microphone arrays in moving human-robot interaction

scenarios,» *Computer Speech & Language*, vol. 65, pág. 101 136, 2020

A.2.2. Publicaciones en revistas

2021 J. Novoa, R. Mahu, J. Wuth et al., «Automatic Speech Recognition for Indoor HRI Scenarios,» *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, n.º 2, págs. 1-30, 2021

A.2.3. Publicaciones preprint

2019 J. Novoa, R. Mahu, A. Díaz et al., «Weighted delay-and-sum beamforming guided by visual tracking for human-robot interaction,» *arXiv preprint arXiv:1906.07298*, 2019

2020 A. Díaz, D. Pincheira, R. Mahu et al., «Short-time deep-learning based source separation for speech enhancement in reverberant environments with beamforming,» *arXiv preprint arXiv:2011.01965*, 2020

Anexo B

Acrónimos

AI Inteligencia Artificial.

AOI Ángulo de Incidencia.

API Application Programming Interface.

ARPA Advanced Research Projects Agency.

ASR Reconocimiento Automático de Voz.

BBRBM Bernoulli-Bernoulli Restricted Boltzmann Machine.

CNN Convolutional Neural Networke.

CSR Continuous Speech Recognition.

DARPA Defense Advanced Research Projects Agency.

DBN Deep Belief Networks.

DNN Deep Neural Networks.

DOA Direction Of Arrival.

EbT Entrenamiento Basado en el Ambiente.

ECG Electrocardiograma.

EEG Electroencefalografías.

EM Expectation-Maximization.

FaNT Fant-Filtering and Noise Adding Tool.

fCNN Fused Convolutional Neural Networke.

FFT Fast Fourier Transform.

FIR Respuesta Finita al Impulso.

fMRI Magnetoencefalografías.

GBRBM Gaussian-Bernoulli Restricted Boltzmann Machine.

GMM Gaussian Mixture Models.

HMM Modelos Ocultos de Markov.

HRI Interaccion Humano Robot.

ICA Análisis de Componentes Independientes.

LDA Análisis Discriminante Lineal.

LSTM Long-Short Term Memory.

MChRSR Multichannel Robot Speech Recognition Database.

MEG Magnetoencefalografías.

MelFB Banco de Filtros Mel.

MFCC Mel-Frequency Cepstral Coefficients.

MLLT Transformaciones Lineales de Máxima Probabilidad.

MLP Multi Layer Perceptron.

MMI Minimum Mutual Information.

MSE Error Cuadrático Medio.

MVDR Minimum Variance Distortionless Response.

MVN Normalización por Media y Varianza.

NLR Nonparametric Likelihood Ratio.

NMF Non-negative Matrix Factorization.

PCA Análisis de Componentes Principales.

PDF Función de Densidad de Probabilidades.

PR2 Personal Robot 2.

RBM Restricted Boltzmann Machine.

RNN Redes Neuronales Recurrentes.

SCFG Stochastic Context-Free Grammar.

sMBR State-Level Minimum Bayes Risk.

SNR Signal to Noise Ratio.

SSP Procesamiento de Señal Social.

VAD Detector de Actividad de Voz.

WER Word Error Rate.

WSJ Wall Street Journal.

WSS Weighted Spectral Slope.