Reconocimiento de comandos de voz para el manejo de una silla de ruedas

Alejandro Calgaro, Agustín Chen, Esteban Menin
Estudiante de Ingeniería en Informática, alejandrocalgaro@gmail.com

***Resumen*—En este proyecto final para la asignatura Procesamiento Digital de Señales, se desarrolló un sistema para el reconocimiento de comandos de voz para el manejo de una silla de ruedas. Para personas que necesitan utilizar una silla de ruedas, el poder controlar la misma a través de comandos de voz, ya sea por comodidad o por una incapacidad de utilizar las manos para arrastrarla, puede ser una herramienta interesante para ayudar a mejorar su calidad de vida. Al utilizar éste sistema, la persona podría controlar el movimiento de la silla de ruedas a partir de cinco posibles comandos: adelante, atrás, derecha, izquierda y detener.**

Se propuso resolver el problema capturando la señal de voz generada por el usuario al pronunciar el comando que desea ejecutar, y analizar la misma para extraer como características relevantes los Coeficientes Cepstrales en escala Mel (MFCC), coeficientes delta, Coeficientes de Predicción Lineal (LPC) y energía, para luego realizar la comparación con las características previamente obtenidas para los cinco comandos disponibles, buscando con cuál de ellos existe mayor similitud y determinar cuál es el comando que el usuario desea ejecutar.

Se comprobó el funcionamiento del sistema creando un dataset con 60 audios de prueba y verificando la cantidad de aciertos y desaciertos al detectar el comando correcto, para lo cual se obtuvo un 91.6% de aciertos. Por otra parte, se analizó la robustez del sistema al añadir ruido a la señal recibida. Se comprobó que a medida que aumenta la potencia del ruido, aumenta el error en el reconocimiento del comando correcto, observando una mayor cantidad de errores en los comandos "derecha" y "atrás".

***Palabras clave*—comandos de voz, silla de ruedas, DTW, MFCC, LPC, coeficientes delta, energía.**

I. INTRODUCCIÓN

El objetivo es implementar un método que reciba la señal de voz del usuario que utiliza el sistema y permita reconocer la acción que desea llevar a cabo, permitiendo a la persona controlar su silla de ruedas mediante la voz, ya sea por una necesidad física o por comodidad en el uso de la misma. Para facilitar el aprendizaje del uso del sistema y la implementación, se contará con un grupo reducido de posibles comandos a reconocer [1], los cuales serán: "Adelante", "Atrás", "Derecha", "Izquierda" y "Detener". Además, como los patrones en el habla de una persona pueden ser muy diferentes a los de otra, para acotar el alcance en este trabajo, las señales de entrada pronunciando los distintos comandos posibles serán generadas por una sola persona, representando al usuario que utilizaría la silla de ruedas.

Para desarrollar este sistema, se ha visto en [2] que se puede procesar y analizar la señal de voz recibida para extraer ciertas características que resulten relevantes, algunas de las cuales pueden ser los Coeficientes Cepstrales en escala Mel (MFCC), Coeficientes de Predicción Lineal (LPC), coeficientes delta y energía, a partir de los cuales se puede realizar una comparación entre señales y determinar el comando o acción que desea ejecutar el usuario.

Luego de extraer características relevantes de la señal, un tema importante a tener en cuenta al momento de comparar la similitud con las señales de referencia, es que un usuario al pronunciar dos veces la misma palabra puede realizarlo con distinta velocidad, con lo cual, la medición de distancias entre dos señales en el reconocimiento de voz puede ser un problema. Para solucionar esto, como se explica en [3], existe la técnica Dynamic Time Warping (DTW), que es una alineación temporal de las señales, ya que consiste en alinear dos secuencias de vectores hasta encontrar la coincidencia óptima entre las dos, obteniendo un ajuste entre ellas incluso si existe un desfase en la velocidad o en el tiempo.

Una vez realizado todo el proceso de reconocimiento del comando de voz, el sistema indica cuál es la acción a ejecutar con la silla de ruedas.

II. RESOLUCIÓN

La resolución del problema se dividió en dos partes, por un lado, un algoritmo que reciba la señal de audio, le realice un acondicionamiento, la filtre, la analice para extraer características relevantes, realice la comparación contra los comandos de referencia disponibles en el sistema e indique cuál fue el comando que el usuario desea ejecutar y, por otro lado, la realización de las pruebas y validaciones correspondientes.

En esta sección nos centraremos en los detalles de la resolución de la primera parte, desglosando las etapas del diagrama que se presenta en la Fig. 1 donde se resume la resolución para el reconocimiento del comando.

A. Obtención de la señal

Para comenzar, el usuario emite una señal de voz con la acción que desea ejecutar en su silla de ruedas, la cual podrá ser uno de los cinco comandos disponibles, y sería capturada por un micrófono y recibida en el sistema. Por simplificación en este trabajo se utilizan señales previamente grabadas a través de un dispositivo móvil y almacenadas en archivos .wav, con una frecuencia de muestreo de 8 kHz.

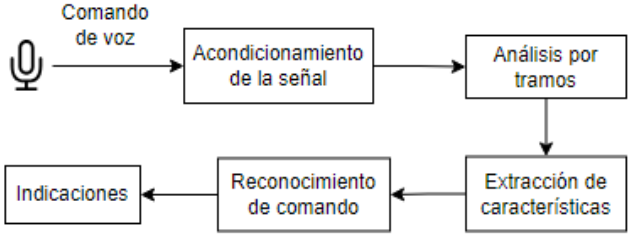

Fig. 1: Diagrama de solución para la identificación del comando

B. *Acondicionamiento de la señal*

Una vez que la señal es recibida en el sistema, se comienza aplicando un acondicionamiento a la misma, con la finalidad de preparar la señal y mejorar el desempeño de los métodos que se utilizarán luego. En este caso, el acondicionamiento de la señal consiste en tres pasos. En primer lugar, una remoción de la media, lo cual permite tener una señal con media 0 y eliminar así la componente continua, que es una desviación general de la señal, positiva o negativa, debida principalmente a micrófonos de mala calidad y que afectaría el análisis posterior. En segundo lugar, un filtrado para la eliminación de ruido, utilizando un filtro pasa-banda que deja pasar las componentes desde 75 Hz que es la frecuencia mínima a la que se encuentra la frecuencia fundamental para la voz masculina, hasta 3000 Hz, ya que se considera un rango de frecuencia de la voz aceptable donde poder encontrar información relevante, eliminando así componentes de baja frecuencia que puedan provenir de distintas fuentes. Y, en tercer lugar, un filtrado pre-énfasis donde se realza la señal en altas frecuencias.

C. *Enventanado*

Luego de acondicionar la señal, se realiza un proceso de enventanado de la misma, ya que la señal de voz es no estacionaria, pero se considera que en intervalos cortos de tiempo la señal se estabiliza y se aproxima a una señal estacionaria, entonces para su posterior análisis dividimos la señal en segmentos pequeños que llamaremos tramas. La duración de cada trama es un parámetro a definir, el cual dependerá de la velocidad de articulación, ya que cuanto más rápido se pronuncie, se necesitará un tamaño de ventana menor, pero en general, se utilizan ventanas de entre 10 y 30 milisegundos. Otro parámetro importante es la superposición entre las tramas, ya que es recomendable que dos tramas consecutivas se superpongan cierta cantidad, la cual también dependería de la velocidad de articulación, a mayor velocidad, mayor debería ser la superposición. En éste caso se han realizado diferentes pruebas y se ha elegido utilizar un tamaño de ventana de 25 milisegundos y una superposición de 10 milisegundos. Llevado a muestras, debido a la frecuencia de muestreo de 8 kHz utilizada, cada trama tendrá 200 muestras y una superposición de 80 muestras.

Y para la elección de la ventana a utilizar, se debe recordar que multiplicar la señal por la ventana en el dominio temporal, equivale a convolucionar ambos espectros en el dominio frecuencial, lo cual producirá distintos resultados de acuerdo al ancho del lóbulo central y la altura de los lóbulos laterales de la ventana, ya que, cuanto más estrecho sea el lóbulo central, más abrupta podrá ser la pendiente de transición, y mientras más bajos sean los lóbulos laterales menos se afectará al espectro con el rizado que se genera. La ventana más utilizada y elegida en este caso es la ventana de Hamming.

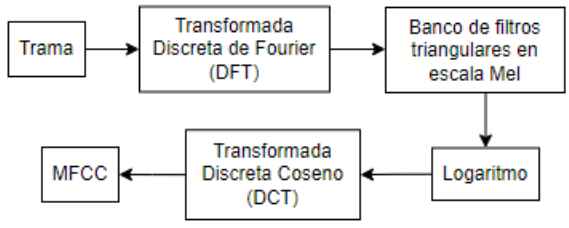

Fig. 2: Proceso de extracción de MFCC

D. *Coeficientes Cepstrales en escala Mel*

Una vez realizadas las operaciones anteriores, se tiene la señal sin ruido en bajas frecuencias, con las altas frecuencias enfatizadas y separada en tramas por el enventanado. A continuación, se explican los pasos que se ilustran en la Fig. 2 para extraer los MFCC.

El primer paso es calcular la Transformada Discreta de Fourier (DFT) a cada trama para obtener su contenido espectral, y nos quedamos solo con la magnitud que es la información que nos interesa.

Como segundo paso, se aplica un banco de filtros en escala Mel, ya que el comportamiento del oído humano no es lineal con la frecuencia, entonces es conveniente muestrear el espectro de la señal siguiendo una escala que se aproxime a dicho comportamiento, y una posible escala para lograrlo es la escala Mel, la cual tiene un comportamiento casi lineal hasta los 1000 Hz y luego logarítmico. El banco consiste en un conjunto de filtros triangulares de área unidad, equiespaciados según la escala Mel. Como principales parámetros para generar el banco de filtros se utilizan las frecuencias de corte máxima y mínima y el número de filtros, que generalmente dicho número de filtros se encuentra entre 20 y 40. En nuestro caso, como la señal de entrada tiene una frecuencia de muestreo de 8000 Hz, se ha tomado como frecuencia mínima 0 Hz y frecuencia máxima 4000 Hz, ya que nos interesa solo la parte positiva de la DFT, y se ha implementado un banco de 26 filtros distribuidos en la escala Mel, como se ilustra en la Fig. 3, donde en el eje horizontal se observan las frecuencias y en el eje vertical la magnitud del filtro. Estos filtros se aplican sobre cada trama de la señal.

Respecto al tercer paso, como en el dominio temporal la señal de excitación está convolucionada con la respuesta del tracto vocal, al aplicar la Transformada de Fourier dicha convolución se transforma en un producto que no nos permite separar ambas señales, entonces, para poder separarlas se requiere la utilización de un operador no lineal que convierta esa multiplicación en una suma. La opción elegida es utilizar el logaritmo como operador, entonces simplemente se aplica el logaritmo al resultado que se tiene actualmente.

Como último paso, se debe aplicar la Transformada de Fourier Inversa, pero por cuestiones de economía de cómputos se decide utilizar la Transformada Discreta Coseno (DCT) para pasar al dominio cepstral y obtener finalmente los coeficientes que se buscaban.

El número de coeficientes está dado por el número de filtros Mel aplicados, y en éste caso, nos hemos quedado con los primeros 13 coeficientes para cada trama, donde se tiene la información más importante respecto a la transferencia del tracto vocal.

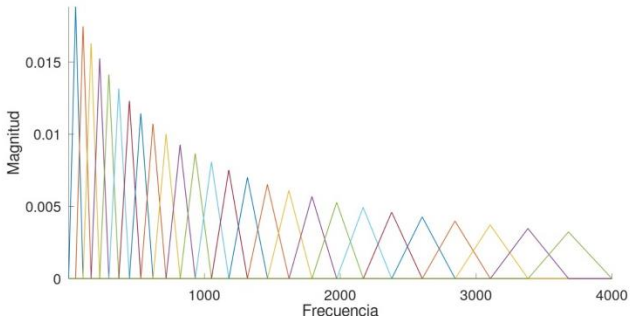

Fig. 3: Banco de filtros en escala Mel

E. Coeficientes de Predicción Lineal

La obtención de estos coeficientes se realiza mediante la resolución de un sistema de ecuaciones formado por la autocorrelación de la señal de entrada, el cual se puede resolver mediante el algoritmo de Levinson-Durbin y se extraen nuevamente 13 coeficientes en cada trama, que agregan más información para la reconocimiento del comando.

F. Coeficientes delta

Hasta ahora con todo el análisis que se hizo sobre las ventanas solo tenemos información estática, ya que una ventana no guarda información de lo que pasó en una ventana anterior, sin embargo, sabemos que existe una correlación porque el aparato fonador tiene una velocidad de cambio lenta, por lo tanto, entre una ventana y la siguiente el espectro debería ser similar, salvo si se produce un cambio abrupto como un cambio de fonema. Es por eso que, además de los MFCC y LPC, existen otros coeficientes derivados de éstos que permiten tener en cuenta la velocidad y variabilidad entre las pronunciaciones del usuario, por eso se denominan coeficientes dinámicos o delta, ya que incluyen información de la dinámica temporal y permiten mejorar el desempeño del sistema.

En este caso, se utilizan coeficientes derivados de los MFCC, que para calcularlos simplemente se agrega información de la derivada, midiendo la variación de los MFCC en cada instante de tiempo y generando otro vector de coeficientes para mejorar la información disponible. De esta manera, se obtienen 13 coeficientes delta para añadir al vector de características de cada trama.

Lo mismo se podría hacer con la segunda derivada si se quisiera, para obtener y agregar los llamados coeficientes de aceleración, viendo dos deltas sucesivos y calculando la diferencia. En nuestro caso no fueron utilizados.

G. Energía

El último coeficiente que se añadió tiene que ver con la energía, para lo cual, simplemente se calcula la energía en cada trama y se añade dicha información al final de su vector de características.

H. Dynamic Time Warping

Luego de obtener un vector para cada trama con las características relevantes de la señal de entrada, es momento de evaluar la similitud con las características de las señales de referencia. Como se mencionó en la introducción, al realizar esta comparación se debe tener en cuenta la variabilidad de la voz, debido a que puede haber un cambio de velocidad en la pronunciación del usuario entre una señal y otra, aunque pertenezcan al mismo comando, para lo cual, se utiliza la técnica DTW para obtener una alineación entre las señales y así poder compararlas de forma adecuada. En la Fig. 4 se observa una ilustración de dicha técnica, donde en el lado izquierdo se muestra el alineamiento original de dos secuencias y en el lado derecho el alineamiento utilizando DTW.

Al realizar el proceso de alineación de las señales, en nuestro caso se cuenta con un vector de características para cada trama de la señal de entrada y lo mismo para las señales de referencia, y para el cálculo de las diferencias entre las características obtenidas para la señal de entrada y las características de las cinco señales de referencia, se decidió utilizar la norma euclídea entre estos vectores,

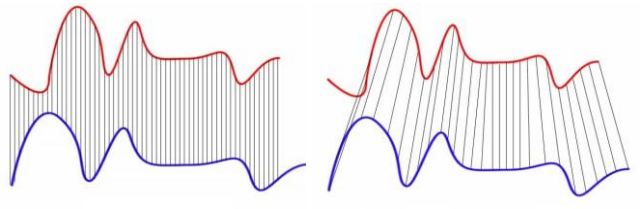

Fig. 4: Ilustración DTW

guardando dichas diferencias y, una vez obtenidas todas, se busca cuál fue la señal de referencia con la que se obtuvo la mínima diferencia, es decir, la señal de mayor similitud, decidiendo así cuál es el comando que el usuario desea ejecutar.

I. Comandos de referencia

Cabe señalar que, para los cinco comandos de referencia disponibles en el sistema, previamente se han analizado señales de audio correspondientes a los mismos y extraído sus características relevantes de la forma que se explicó anteriormente, creando vectores con dichas características para cada comando, que fueron almacenados en archivos .txt, para que, una vez que ingrese una nueva señal de voz al sistema, ya se encuentren disponibles los archivos de referencia para cargarlos y realizar la comparación para determinar la similitud con cada uno y elegir así el comando a ejecutar.

J. Tiempo real

Por otra parte, se añadió una función extra para simular el funcionamiento del sistema en tiempo real, donde el usuario puede presionar una tecla y pronunciar cuál es la acción que desea ejecutar, esa señal de voz es recibida a través del micrófono y el sistema guarda la información en un archivo de audio .wav que se analiza de la misma manera que los audios grabados anteriormente para determinar cuál es el comando que el usuario desea ejecutar.

III. DATOS Y PRUEBAS

Para la parte correspondiente a la evaluación del rendimiento del sistema, se reunieron 60 audios en total, correspondientes a los cinco comandos disponibles, los cuales fueron grabados con el micrófono de un dispositivo móvil y almacenados en el dataset de pruebas en formato .wav y con una frecuencia de muestreo de 8 kHz. Las señales de audio se probaron en primer lugar tal cual fueron grabadas, sin ruido adicional, y luego agregando distinto nivel de ruido sobre la señal a analizar, para comprobar la respuesta del sistema utilizando distinta relación señal ruido (SNR). Para cada uno de los cinco comandos disponibles se evaluó la cantidad de aciertos y desaciertos al analizar los grupos de señales de audio correspondientes a los mismos, comprobando si el comando se reconoció de forma correcta o no.

IV. RESULTADOS

En la Tabla I se muestran los resultados obtenidos al analizar las señales de audio originales, sin ruido adicional, para las cuales se obtuvo un 91.6% de aciertos. En las columnas de dicha tabla se puede observar el comando, la cantidad de aciertos y cantidad de desaciertos del sistema al reconocer el comando elegido.

TABLA I
RESULTADOS OBTENIDOS CON LAS SEÑALES ORIGINALES

Comando	Aciertos	Desaciertos
Adelante	12	0
Atrás	11	1
Derecha	8	4
Izquierda	12	0
Detener	12	0

TABLA II
RESULTADOS OBTENIDOS AL AÑADIR DISTINTOS NIVELES DE RUIDO A LAS SEÑALES ORIGINALES

SNR [dB]	Aciertos	Desaciertos
0	13	47
10	23	37
20	32	28
30	36	24
40	42	18
50	54	8

En la Tabla II se observan los resultados que se obtuvieron al agregar distintos niveles de ruido a la señal, utilizando una SNR desde 0 dB hasta 50 dB. En las columnas de dicha tabla se puede observar la SNR utilizada en cada caso, la cantidad de aciertos y cantidad de desaciertos del sistema al identificar el comando elegido.

Como es de esperar, al aumentar la SNR, la cantidad de aciertos aumenta, pero se pudo observar que el comando “derecha” fue el que presentó mayor dificultad para el reconocimiento al agregar ruido en la señal, el cual en los casos de error arroja mayor cercanía al comando “detener”, y el segundo comando con mayor cantidad de errores en presencia de ruido fue “atrás”. Mientras que el comando “detener” obtuvo 100% de efectividad incluso utilizando una SNR igual a 0 dB, el comando “adelante” tuvo 100% de efectividad luego de SNR igual a 10 dB y el comando “izquierda” obtuvo un 66.6% de efectividad con SNR igual a 20 dB, aumentando luego a 91.6% de efectividad.

V. CONCLUSIONES

El sistema desarrollado utilizando las técnicas mencionadas en la resolución para la extracción de características relevantes de las señales de voz, ha dado un resultado muy eficiente a la hora del reconocimiento de los comandos cuando se utilizaron las señales originales como fueron obtenidas a partir del micrófono, sin presencia de ruido adicional, mientras que, al realizar pruebas generando y sumando ruido a la señal, se ha visto cómo a medida que aumentaba el ruido presente aumentaba también la cantidad de desaciertos al detectar los comandos, con los resultados antes mencionados.

Para analizar otras variante del sistema, se ha probado extrayendo como características relevantes de las señales solo los MFCC, reduciendo la cantidad de cálculos y tiempo de procesado, y se ha comprobado que utilizando las señales de audio originales sin el agregado adicional de ruido, los resultados fueron muy similares a los que se indicaron en la Tabla I de la sección anterior, mientras que al agregar ruido a la señal, el sistema actual incluyendo como características los MFCC, deltas, LPC y energía tiene un mejor desempeño.

Otra variación que se probó fue cambiar el tamaño de las ventanas y la superposición entre las mismas, observando que con ventanas de menor duración algunos comandos mejoraron en su reconocimiento, pero otros empeoraron, además de aumentar el tiempo de procesamiento debido al menor tamaño de las ventanas, por lo cual, se decidió dejar el tamaño de ventana en 25 milisegundos y la superposición en 10 milisegundos como fue explicado.

Dos posibles trabajos futuros serían, por un lado, determinar un umbral de validez para cada comando, con el cual se pueda agregar seguridad al sistema al permitir que, cuando un comando no se reconoce adecuadamente o existe mucha diferencia con los comandos de referencia, no se deba ejecutar el de mayor similitud, sino que el sistema pueda indicar al usuario que el comando no fue válido y se deba repetir la operación de pronunciar la acción o comando a realizar. Y, por otro lado, se podría generar una segmentación de tramas sonoras para analizar solo ese conjunto y descartar las tramas sordas, lo cual podría reducir el tiempo de procesado y respuesta del sistema.

VI. REFERENCIAS

[1] R. C. Simpson and S. P. Levine, "Voice control of a powered wheelchair," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 2, pp. 122-125, June 2002, doi: 10.1109/TNSRE.2002.1031981.

[2] R. Gupte, S. Hawa and R. Sonkusare, "Speech Recognition Using Cross Correlation and Feature Analysis Using Mel-Frequency Cepstral Coefficients and Pitch," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bangluru, India, 2020, pp. 1-5, doi: 10.1109/INOCON50539.2020.9298320.

[3] Yurika Permanasari, “Speech recognition using Dynamic Time Warping (DTW)” *et al 2019 J. Phys.: Conf. Ser. 1366 012091*.