UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
VALPARAÍSO – CHILE

"DETECCIÓN DE LA ACTIVIDAD DE LA VOZ EN SEÑAL DE ACELERÓMETRO"

FELIPE ANDRÉS ACEVEDO LA RIVERA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERIA CIVIL ELECTRÓNICA MENCIÓN COMPUTADORES

PROFESOR GUÍA:	DR. AGUSTÍN GONÁLEZ VALENZUELA.
PROFESOR CORREFERENTE:	DR. MATÍAS ZAÑARTU SALAS.

ABRIL - 2016

AGRADECIMIENTOS

El siguiente trabajo fue posible gracias a Dr. Agustín González Valenzuela, Dr. Matías Zañartu Salas y el equipo detrás del proyecto de monitoreo móvil de las cuerdas vocales compuesto de integrantes de la Universidad Técnica Federico Santa María en conjunto con Massachusetts Institute of Technology, los que dieron la oportunidad de desarrollar una solución para optimizar sus procesos, además de ser guías y dar a disposición la base de datos para poder ejecutar las pruebas.

Muchas gracias

DETECCIÓN DE LA ACTIVIDAD DE LA VOZ EN SEÑAL DE ACELERÓMETRO

Memoria para optar al título de Ingeniero Civil Electrónico, mención Computadores

Felipe Andrés Acevedo La Rivera

Profesor Guía: Agustín González Valenzuela

Abril 2016

RESUMEN

El siguiente trabajo se encuentra enmarcado en el proyecto de investigadores de la Universidad Técnica Federico Santa María, en conjunto con el Massachusetts General Hospital, para monitorear continuamente la actividad de las cuerdas vocales de pacientes que presentan trastornos en las cuerdas vocales. Esto se logra utilizando un acelerómetro en la superficie de la piel del cuello del paciente, donde se ubican las cuerdas vocales y un teléfono inteligente.

Al estar continuamente monitoreando las cuerdas vocales, es necesario descartar de la señal obtenida, los períodos en los cuales las cuerdas vocales no se encuentran en uso, para así utilizar los recursos del teléfono de mejor manera y prevenir que segmentos de ruido entren al análisis posterior de la señal.

A partir de la problemática descrita, se propone un algoritmo de detección de la actividad de las cuerdas vocales en tiempo real, basado en técnicas tradicionales de la detección de la actividad de la voz utilizada en la actualidad, teniendo que ser adaptadas para funcionar con la señal de vibración en el acelerómetro en vez de una señal de presión acústica captada por el micrófono.

Durante el trabajo se estudian las heurísticas actualmente utilizadas en el post-procesamiento del proyecto, para analizar cuáles son más efectivas para aplicarse en el algoritmo en tiempo real con la señal del acelerómetro.

En conjunto con las heurísticas seleccionadas, se decide implementar el algoritmo estadístico de “Sequential Gaussian Mixture Model” estudiado para el análisis en base a la energía de la señal.

Como resultado se obtiene un algoritmo capaz de distinguir la actividad de las cuerdas vocales en una señal de acelerómetro utilizando las heurísticas más relevantes y un análisis estadístico de la energía de la señal.

Palabras Claves: Detección de la Actividad de las Cuerdas Vocales

ABSTRACT

The following work is part of a project of researchers of Universidad Técnica Federico Santa María together with people of Massachusetts General Hospital, to continuously monitor the activity of the vocal folds of patients with vocal cords disorders. This is achieved by using an accelerometer in the surface of the neck's skill of the patient, where the vocal folds are located.

To be continuously monitoring the vocal folds, it's necessary to discard of the captured signal, the periods of time in which the vocal folds are not active, to achieve a better utilization of the resources and to prevent noise fragments to enter the next stages of processing.

Because of the problem just described, an algorithm is proposed to detect activity of the vocal folds in real time, based on traditional techniques of voice activity detection that are currently implemented, which needs to be adapted to be able to work with the accelerometer's signal instead of a signal of a microphone.

During this work, the heuristics that are used in the post-processing of the project are studied, to analyze which one are more effective to be applied on the real time algorithm with the accelerometer signal.

Together with the selected heuristics, it's decided to implement a statistical algorithm based on "Sequential Gaussian Mixture Model" for the analysis of the signal's energy.

The result of this works is an algorithm capable of distinguish the activity of the vocal folds in an accelerometer's signal using the most relevant heuristics and a statistical analysis of the signal's energy.

GLOSARIO

DAV	Detector de la Actividad de la Voz
DACV	Detector de la Actividad de las Cuerdas Vocales
AMR	"Adaptive Multi Rate", algoritmo propuesto por ETSI para la detección de la actividad de la voz
ETSI	European Telecommunications Standards Institute
Ground Truth	Base de datos de señales previamente marcadas por un experto para utilizar como referencia al momento de comparar los resultados de la decisión de los algoritmos a implementar
SGMM	"Sequential Gaussian Mixture Model", es un modelo probabilístico Gaussiano, el que es aplicado al reconocimiento de la actividad de las cuerdas vocales.
RMS	"Root Mean Square", es la raíz del promedio de los cuadrados de los valores de una señal en un período determinado. El RMS es utilizado como medida de la energía de una señal oscilante. Su cálculo se describe en el Capítulo 3-1

ÍNDICE

Contenido

INTRODUCCIÓN

Aproximadamente 6.6% de la población activa en Estados Unidos padece de desórdenes de la voz, cifras similares se esperan del resto del mundo [1]. La mayoría de estos desordenes requieren procedimientos quirúrgicos, frecuentes terapias a la voz, o ambos, con un resultado que depende en gran parte de la habilidad de detectar y cambiar comportamientos específicos de la voz. Los desórdenes de la voz más comunes son crónicos, o con condiciones recurrentes, que son propensos a resultar en patrones inapropiados en el comportamiento de la voz, conocidos como hiperfunción de la voz [1].

Una evaluación clínica del funcionamiento de la voz es esencial para la evaluación de la presencia de hiperfunción de la voz, antes y después de una intervención vocal. Comúnmente, la forma de recolectar información acerca de la hiperfunción vocal es a través de pruebas dentro de laboratorios. Estas pruebas tienen la desventaja de ser muy costosas en tiempo, y no permiten monitorear el uso diario de la voz y modificar sus malos usos.

Es por este problema que, un grupo de académicos de la Universidad Técnica Federico Santa María, en conjunto con el Massachusetts General Hospital (MGH), buscan implementar un dispositivo no invasivo para el monitoreo continuo del uso de la voz, en base de un acelerómetro colocado en el cuello sobre las cuerdas vocales y un teléfono inteligente con Android, para poder recolectar información sobre problemas en la voz y poder dar informar al paciente lo más rápido posible sobre la utilización de su voz y así poder tomar medidas correctivas cuando corresponda.

Durante el monitoreo diario del paciente, aproximadamente un 6% del tiempo se encuentra utilizando las cuerdas vocales [1], lo que motiva el trabajo de desarrollar un algoritmo que permita discriminar cuándo efectivamente hay actividad en las cuerdas vocales, para así utilizar de forma óptima los recursos de procesamiento,

almacenamiento, energía y transmisión, además de proteger al sistema de análisis de la entrada de señales de ruido.

Problema a resolver

El principal problema que este trabajo busca resolver es optimizar el uso de recursos del sistema de monitoreo continuo de la actividad de las cuerdas vocales. Esto se consigue realizando un primer filtro en los datos capturados, que permite discriminar cuándo efectivamente hay actividad de las cuerdas vocales y descartar los otros períodos en que solo hay ruido.

Objetivo

El proyecto tiene como objetivo desarrollar un algoritmo para la detección en tiempo real de la actividad de las cuerdas vocales como primer filtro en una aplicación de monitoreo y diagnóstico de desórdenes de la voz, y cuyo propósito es concentrar el esfuerzo de procesamiento solo en las señales de interés y así ahorrar en recursos valiosos, como el consumo energético y el uso de canales de comunicación.

Estrategia

La estrategia para lograr el objetivo fue la siguiente:

1. Realizar un estudio del tipo de señal que entrega el acelerómetro en presencia de la actividad de las cuerdas vocales.
2. Analizar el algoritmo actual y las heurísticas que utiliza, observando la efectividad de cada una para discriminar la actividad de las cuerdas vocales en un segmento de señal. Además, se ajustan los rangos para los cuales se espera tener actividad de las cuerdas vocales en un segmento de señal, de forma de minimizar el error resultante al comparar con el ground truth.
3. Implementar el algoritmo de Sequiential Gaussian Mixture Models (SGMM) [4] como mejora al análisis de RMS del algoritmo actual. De igual forma que en análisis de características, se buscan los parámetros a utilizar que minimizan el error al comparar con el ground truth.
4. Analizar el resultado del algoritmo resultante.

CAPÍTULO 1. ALGORITMOS DE DETECCIÓN DE LA ACTIVIDAD DE LA VOZ ACTUALES

Este trabajo se basó en los algoritmos de Detección de la Actividad de la Voz (DAV) encontrados en la literatura, adaptándose para ser utilizados en la señal de la actividad de las cuerdas vocales obtenida con el acelerómetro.

Los algoritmos DAV son utilizados en varias áreas del análisis y transmisión para mejorar y optimizar su rendimiento [5]. En la transmisión de audio, DAV es utilizado para enviar solo las señales cuando hay voz presente, disminuyendo así el tráfico necesario para realizar la comunicación. En el reconocimiento de voz, se utiliza principalmente para analizar solo los instantes en donde existe voz para no deteriorar el algoritmo con entradas de otro tipo (ruido). Además, en esta área es de importancia predecir el ruido para mejorar el análisis posterior (i.e. sustraer una predicción del ruido a la señal de voz para mejorar su calidad), lo que es posible detectar con el resultado del DAV, el que permite distinguir una señal solo de ruido con una señal de ruido junto con voz.

1.1.Estructura de un Detector de la Actividad de la Voz

Los algoritmos DAV se pueden separar en tres módulos principales:

1. **Extracción de Características:** El objetivo de este módulo es extraer mediante cálculos, un vector de características discriminativas de la voz. Para realizar esta tarea, los algoritmos utilizan distintas técnicas, como Coeficientes Cepstrales en las Frecuencias de Mel (MFCC) [3] [4] [5], tono y timbre en sub-bandas (ETSI "Adaptive Multi Rate" AMR-1), energía y potencia espectral en sub-bandas (ETSI AMR-2), una combinación de las anteriores [6].
2. **Clasificación:** Este módulo busca determinar, con el vector de características, si en el segmento a analizar hay presencia de voz o no. Para este objetivo, los algoritmos utilizan técnicas basadas en la heurística ([3], ETSI AMR), modelos estadísticos [4] [7] [8], y aprendizaje de máquina [6] [9].
3. **Suavizado:** Una vez clasificado un segmento, es necesario realizar un suavizado de la decisión tomada. Esto es debido a problemas como detección tardía de un segmento de voz, retraso en detección de ausencia de voz, y el no reconocimiento de la voz en silencios dentro de una frase [5]. Es por esto que los algoritmos generalmente utilizan técnicas para suavizar la salida de la clasificación, ya sea introduciendo retrasos en el cambio de clasificación o utilizando "Hidden Markov Models" (HMM) [8]

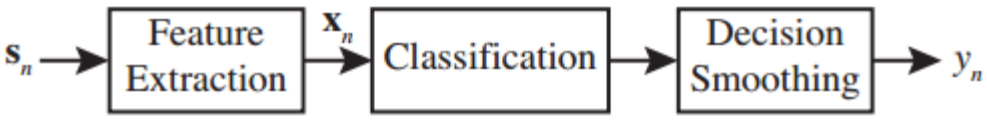

Figura 1: Diagrama de bloques de un DAV [5].

1.2. Tipos de Detectores de la Actividad de la Voz

Dentro de los DAV estudiados, estos pueden ser clasificados según sus técnicas de clasificación en los siguientes tipos:

- **Heurísticos** ([3], ETSI AMR): Estos DAV utilizan directamente el vector de características extraídas, con lo que son capaces de estimar el ruido y, junto con la comparación con parámetros obtenidos empíricamente, obtener una clasificación. Estos algoritmos continúan siendo los más populares en la industria debido a su larga historia y su bajo costo computacional.
- **Estadísticos** [4] [7] [8]: Utilizando modelos estadísticos, como "Gaussian Mixture Model" (GMM) [7] [8], "Sequential Gaussian Mixture Model" (SGMM) [3], entre otros, este tipo de DAV busca dar respuesta a la clasificación basándose, además, en aprendizaje supervisado [7] como no supervisado [4] [7]. Estos DAV presentan un mejor desempeño en entornos en donde el ruido llega a ser comparable con la señal (SNR < 0).
- **Aprendizaje de Máquina** [6] [9]: En los últimos años se ha tratado de incluir técnicas de aprendizaje de máquina para resolver el problema de detección de la voz. Estos algoritmos utilizan "Deep Belief Networks" [6] y "Deep Neural Networks" [9], entre otros. Estos algoritmos presentan una mejora en robustez con respecto a los otros DAV, pero su carga computacional también es considerablemente superior, por lo que algunos autores los caracterizan como no implementables en tiempo real [9], mientras otros mencionan que sí son implementables, pero el margen de tiempo para futuras operaciones se ve considerablemente disminuido [6].

1.3.Detectores de la Actividad de la Voz Estudiados

A continuación, se describen los algoritmos DAV estudiados, cada uno corresponde a cada tipo descrito anteriormente:

- **Heurísticos – ETSI Adaptive Multi Rate Opción 2 (AMR-2):** parte del estándar del Instituto de estándares de Telecomunicaciones Europea (ETSI) que define un DAV para la transmisión discontinua en GSM. Este algoritmo se basa en características espectrales y estimaciones de energía y ratios entre energía y ruido (SNR) para realizar la discriminación.

  En la Figura 1-1, se muestra el diagrama de bloques propuesto por el Instituto para detectar la actividad de la voz:

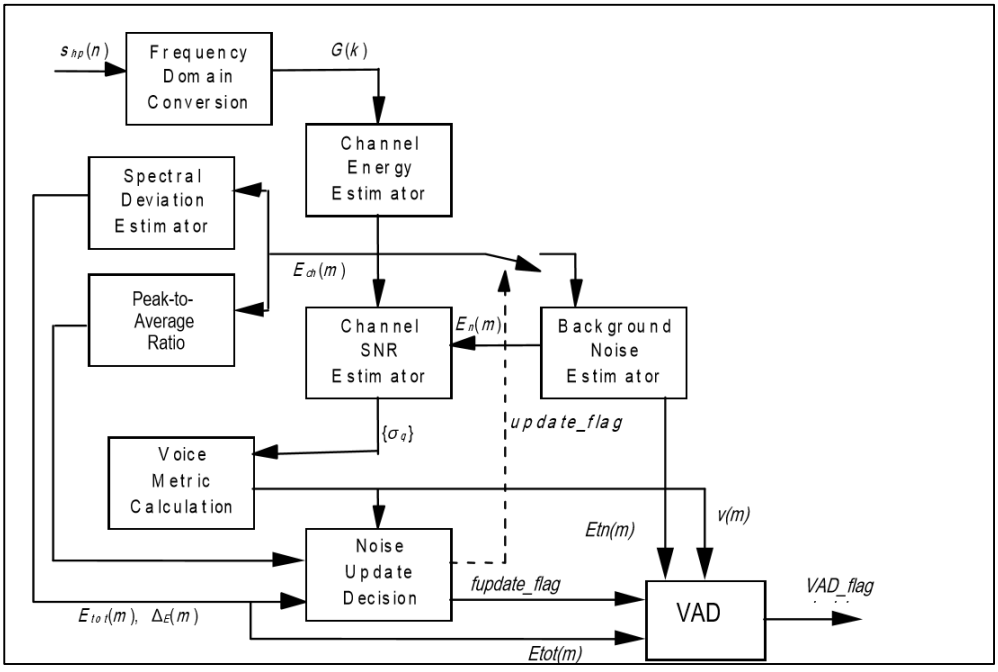

Figura 1-1:Diagrama de Bloques del algoritmo ETSI AMR-2

- **Estadísticos – Sequential Gaussian Mixture Models (SGMM):** La idea base de este algoritmo es agrupar la energía en el espacio de la frecuencia en dos modelos Gaussianos, uno correspondiente a la presencia de voz, y otro a la energía del ruido asociado, para luego determinar si un segmento, por medio de su modelo Gaussiano, pertenece al modelo del ruido o al de la voz. Con estos modelos, se puede calcular un límite en la intersección, igualando ambos modelos, con el que se puede decidir si hay presencia de la voz dependiendo si la energía es mayor o menor a éste.

  La principal ventaja es que se adapta a los distintos niveles que puede tomar la señal de forma no supervisada, lo que ahorra calibraciones iniciales y es más flexible a los distintos contextos en los que se puede encontrar el paciente, en los cuales puede hablar muy despacio o muy fuerte.

  La Figura 1-2 muestra las distribuciones gaussianas de la presencia de la vox y del silencio, donde el eje **x** es el logaritmo de la energía, y el eje **y** son las ocurrencias de esa energía.

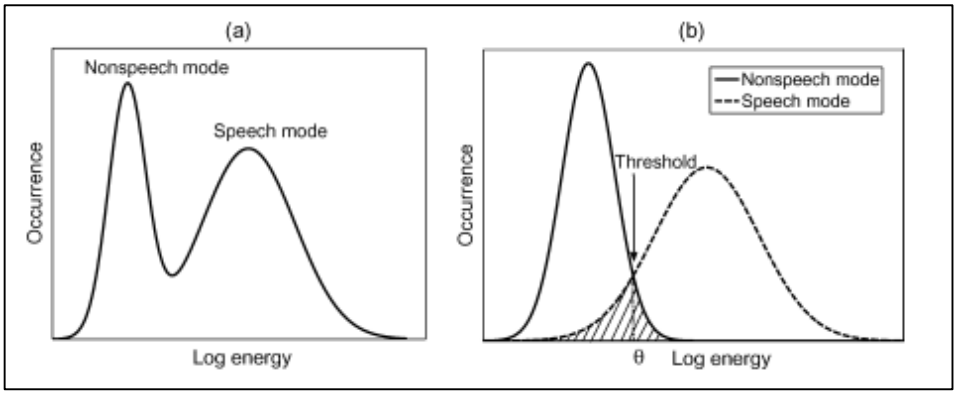

Figura 1-2:Distribuciones Gaussianas del Ruido y la Actividad de la Voz [4]

- **Aprendizaje de Máquina – Deep Belief Networks:** La estrategia del algoritmo es lograr una fusión de características de la señal de audio, detalladas en la Tabla 1-1, para generar nuevas características más relevantes para realizar una clasificación.

  Para lograr esto se utiliza Deep Belief Networks, un tipo de Red Neuronal que realiza un aprendizaje no supervisado para generar esta fusión de características, para luego utilizarla en la red de clasificación entrenada de forma supervisada.

  La principal desventaja de este algoritmo es su demanda computacional, haciéndolo inviable para un sistema de tiempo real, sobre todo en un dispositivo móvil.

Tabla 1-1: Características utilizadas en Deep Belief Networks [6]

Característica	**Dimensiones**
Tono	1
Coeficientes de la Transformada de Fourier	48
Coeficientes Cepstrales en las Frecuencias de Mel	60
Codificación predictiva lineal	12
RASTA Predicción perceptual lineal	17
AMS	135
Total	273

1.4. Comparación de Detectores de la Actividad de la Voz

Los algoritmos DAV pueden ser comparados de distintas formas, dentro de las más utilizadas están:

- **Speech Hit Rate y Non-speech Hit Rate** [2]: Tasas de acierto en la clasificación de un segmento con presencia de voz y un segmento con ausencia de voz respectivamente. Entre mayor sea el valor es mejor.
- **Receiver operating characteristics curves (curvas ROC)** [2] [3] [6]: Estas curvas están compuestas por la tasa de acierto de silencio con la tasa de error de falsos positivos (no detección de segmentos con voz). Éstas buscan describir completamente la tasa de error de un DAV.

  La curva es generada variando la flexibilidad del algoritmo, para lograr un mínimo de falsos positivos (detectar actividad de la voz en todos los segmentos) hasta lograr una máxima tasa de acierto de silencio (detectar como silencio todos los segmentos). El punto óptimo es aquel que se acerca más al punto (0, 100%).
- **Complejidad**: complejidad del algoritmo, es lo que influye en el cumplimiento de ejecución en tiempo real y en la carga que somete al sistema.

Dado estos parámetros y los resultados presentados en la literatura [2, 4, 5, 6, 7, 8, 9], se opta por la implementación del algoritmo estadístico SGMM, principalmente por su simpleza y adaptabilidad, además de obtener buenos resultados en las pruebas que se muestran en el trabajo del autor.

1.5.Error de Discretización Temporal de Segmentos

Para analizar una señal en tiempo real, es necesario capturar fragmentos de ésta y analizarlos a medida que se vayan obteniendo. Durante el estudio de los algoritmos, la comunidad científica [1, 11] utiliza por convención un análisis en segmentos de 50 [ms], por ende, en este trabajo también se utiliza dicha medida.

La utilización de segmentos de 50 [ms] puede generar errores de decisión al analizar un segmento compuesto por una mitad de silencio y la otra de actividad, como se muestra en la Figura 1-3, en donde, por utilizar segmentos de análisis, se clasifica erróneamente el silencio previo a la actividad como si fuese parte de ésta también, éste error se le llama *error de discretización temporal de segmentos*.

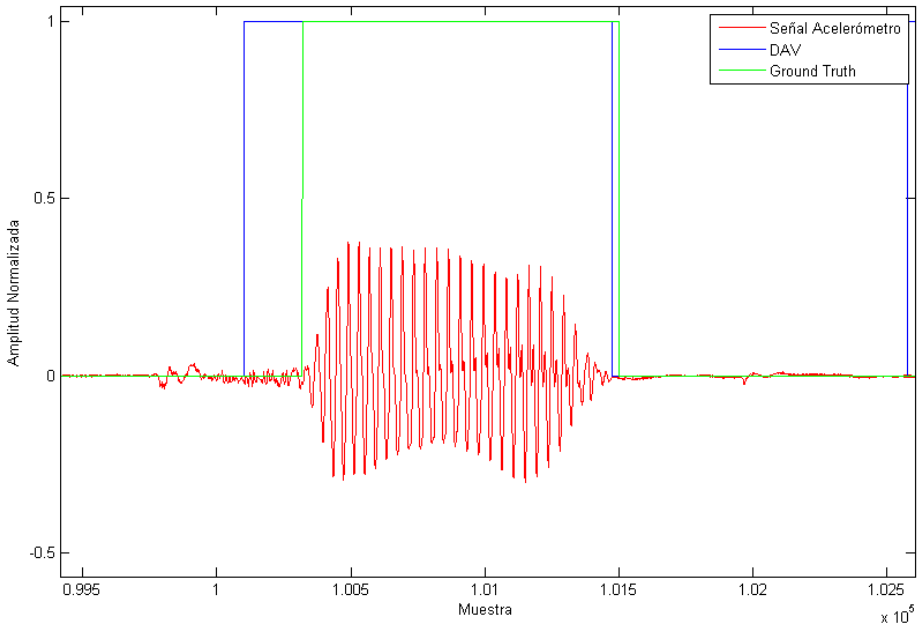

Figura 1-3: Gráfico que muestra Error de Discretización Temporal

Se puede calcular el máximo error de discretización temporal tomando como promedio que el segmento contiene 50% silencio y 50% de actividad, por lo que el error considerando el peor caso, es del 50% del tamaño del segmento en cada cambio, lo que se traduce a la Ecuación 1-1, en donde $\boldsymbol{ls}$ es el largo del segmento a analizar, $\boldsymbol{gt}(\boldsymbol{x})$ es el ground truth en la muestra **x**, representando la presencia de actividad como 1 y el silencio como 0, y **n** es el largo de la señal:

$$Error_d = \frac{\sum_{x=1}^{n-1} |gt(x-1) - gt(x)| * \frac{ls}{2}}{n} \quad \textbf{(1-1)}$$

CAPÍTULO 2. ANÁLISIS DE LA SEÑAL DE ACELERÓMETRO

En este capítulo se estudia la señal de la actividad de las cuerdas vocales capturadas por el acelerómetro, además de sus componentes espectrales y comparación con la señal capturada por un micrófono. El objetivo de este estudio es obtener un conocimiento base de la señal que permita diseñar y ajustar un algoritmo que discrimine la presencia de la actividad de las cuerdas vocales del silencio o ruido, por medio del análisis de sus características.

El sistema de medición se basa en un acelerómetro colocado en la superficie de la piel a la altura de las cuerdas vocales, como se muestra en la Figura 2-1. Este sensor se conecta por medio de un cable a la entrada de micrófono a un celular con sistema operativo Android.

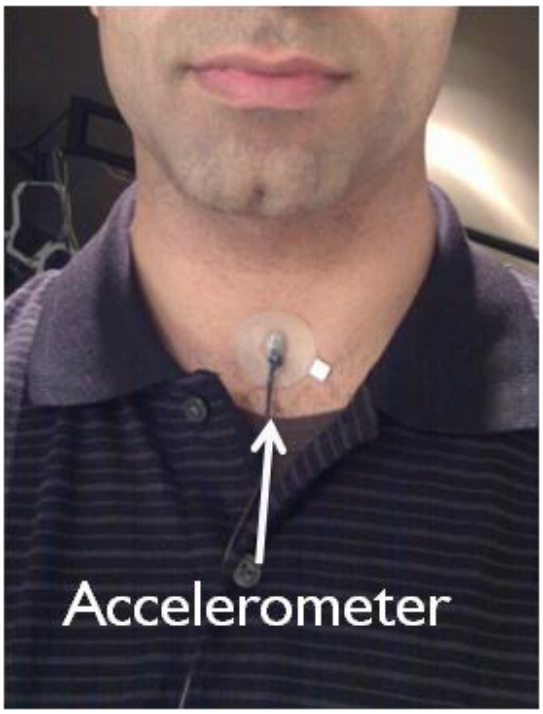

Figura 2-1: Imagen del sensor instalado en el cuello del paciente

2.1.Base de Datos de Pruebas

Para analizar la señal de la actividad de las cuerdas vocales, se tuvo acceso a una base de datos de señales de micrófono y acelerómetro de pruebas en pacientes sanos y con patologías, los que recitan un texto específico para este tipo de muestras llamado "rainbow passage". Este texto se distingue por contener la mayoría de los fonemas del idioma inglés. Estas muestras fueron previamente marcadas, clasificando cada segmento como sonoro ("voiced") cuando el sonido se basa en la actividad de las cuerdas vocales, sordos ("unvoiced") cuando el sonido se basa en otro tipo de mecanismos, como las fricativas (sonido de la letra "s").

En este trabajo se consideran los segmentos "unvoiced" como silencio debido a que la fuente de su sonido no depende de las cuerdas vocales.

La tasa de muestreo de los datos es de 11.025 Hz.

Las señales utilizadas para los análisis son las siguientes:

- **N1f, N3f, N4f:** Pacientes femeninas sin patologías
- **N5m, N6m:** Pacientes masculinos sin patologías
- **P1m_MI, P4m_SE, P5m_MO:** Pacientes masculinos con trastorno de la voz
- **P2f_SE, P3f_MI, P6_MO:** Pacientes femeninos con trastorno de la voz

2.2. Análisis del Espectro de la Señal

En Figura 2-2 se muestra un segmento de la señal del acelerómetro en el momento en que hay actividad en las cuerdas vocales.

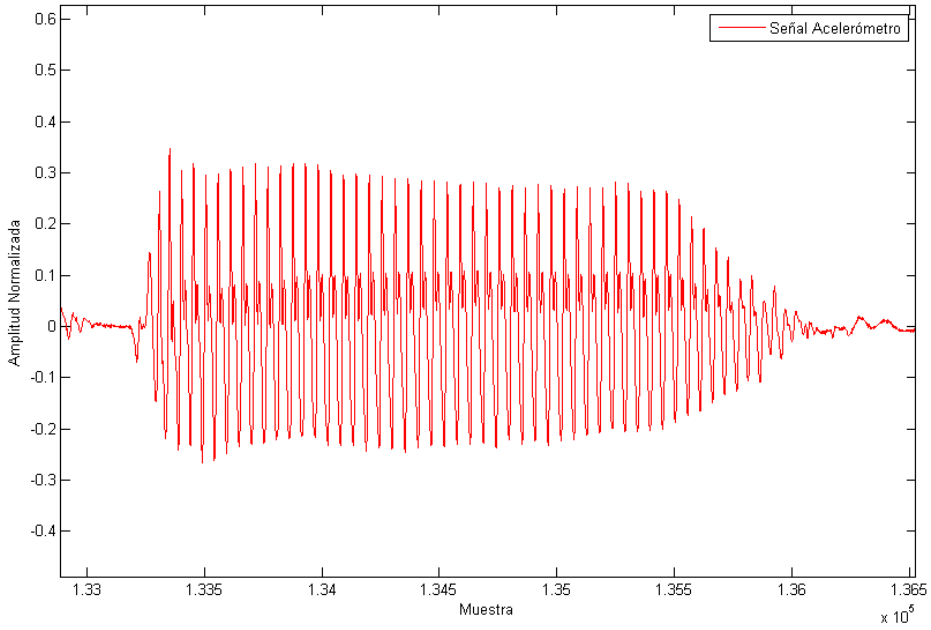

Figura 2-2:Muestra de la actividad de las cuerdas vocales en acelerómetro

Como se puede apreciar en la Figura 2-2, la señal sobresale claramente del ruido, por lo que da un buen indicio para poder ser discriminada por su energía.

El contenido espectral de un segmento con presencia de la actividad de las cuerdas vocales se puede apreciar en la Figura 2-3, en donde se puede ver que la señal se conforma de una frecuencia fundamental de frecuencia cerca de los 200 [Hz] seguida de dos armónicos en 400 [Hz] y 600 [Hz].

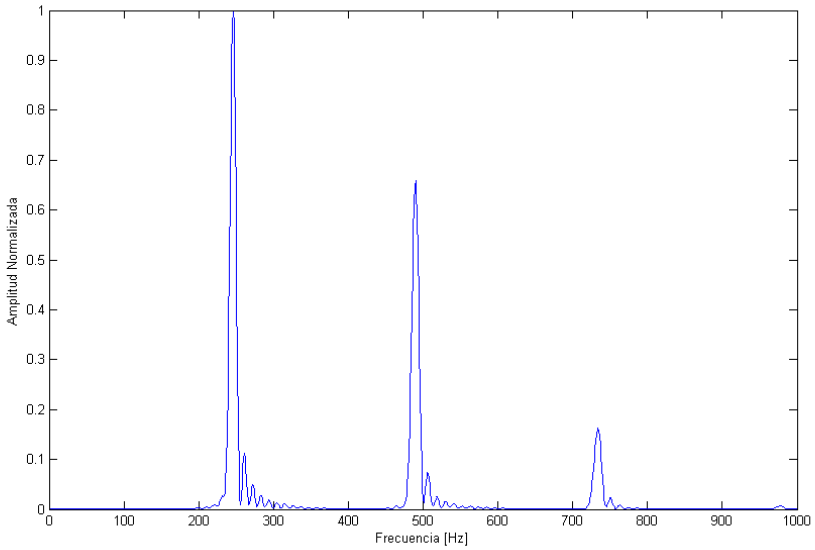

Figura 2-3: Diagrama espectral de la actividad de las cuerdas vocales en acelerómetro

La presencia de los armónicos no se da en todos los casos, como se muestra en la Figura 2-4, donde solo la frecuencia fundamental está presente en el segmento de señal de un hombre sano.

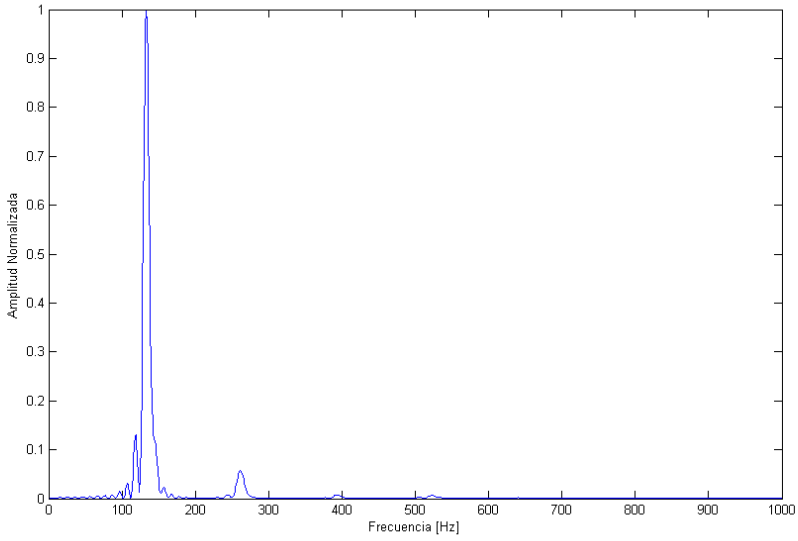

Figura 2-4: Diagrama espectral de la actividad de las cuerdas vocales sin armónicos

Como se puede ver en la Figura 2-5, la cantidad de armónicos presentes en la señal de acelerómetro (gráfico superior) es mucho menor a la presente en la señal de micrófono (gráfico inferior).

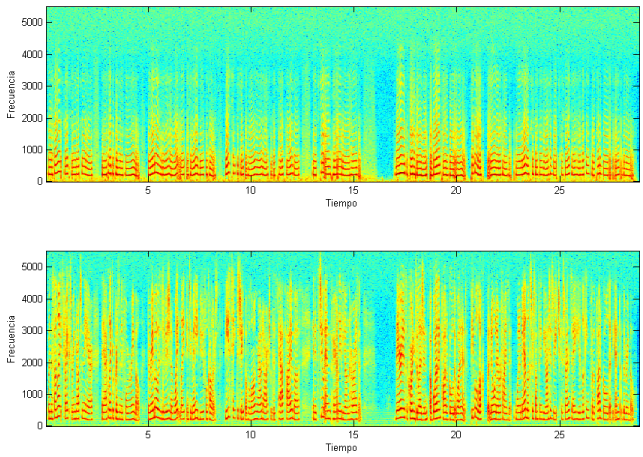

Figura 2-5: Espectrograma de la señal en acelerómetro (superior) y micrófono (inferior)

Si bien el acelerómetro, por la naturaleza de señales que mide, es poco susceptible al ruido acústico, si se ve afectado por otra especie de ruidos, como lo son vibraciones al caminar u otro tipo de movimientos. Queda fuera del alcance de este trabajo realizar un estudio más detallado de la influencia de estas perturbaciones.

2.3. Error de Discretización Temporal de Segmentos en los Datos Estudiados

Para el análisis del resultado de los algoritmos a implementar, es necesario tener presente el error de discretización temporal que se puede esperar en cada una de las señales de la base de datos.

A continuación, se muestran los errores por señal calculado según la Ecuación 1-1:

Tabla 2-1: Resultados de Error de Discretización Temporal

Muestra	Error de Discretización Temporal (%)
N1f	10.37 %
N3f	8.48 %
N4f	9.98 %
N5m	7.96 %
N6m	11.43 %
P1m_MI	9.82 %
P2f_SE	9.87 %
P3f_MI	11.11 %
P4m_SE	5.16 %
P5m_MO	8.05 %
P6f_MO	11.14 %
PROM	**9.40 %**

Los resultados a obtener por un algoritmo que utilice ventanas de análisis sin traslape se pueden ver afectados por un error base superior al 10% en algunas señales, debido a la discretización temporal.

CAPÍTULO 3. ANÁLISIS DE LAS CARACTERÍSTICAS UTILIZADAS EN EL ALGORITMO ACTUALMENTE UTILIZADO

Además de los algoritmos estudiados en el estado del arte, se cuenta con el algoritmo actual utilizado para el pre-procesamiento del proyecto aplicado a señales previamente grabadas.

3.1. Algoritmo Actual

En este algoritmo se evalúan las siguientes características para discriminar si en un segmento de señal hay actividad de las cuerdas vocales:

- **Valor cuadrático medio (RMS):** representación de la amplitud promedio de la señal. Su cálculo está dado por la Ecuación 3-1, en donde **y(x)** es la amplitud de la señal en la muestra **x**, y **n** es el largo del período a calcular:

$$RMS = \sqrt{\frac{1}{n}\sum_{x=1}^{n} y(x)^2} \quad (3\text{-}1)$$

- **Frecuencia fundamental (F0)**: Frecuencia base generada por las cuerdas vocales, las que puede generar más armónicos. Para el caso de las cuerdas vocales, la frecuencia fundamental normalmente toma valores entre los 50 [Hz] y los 500 [Hz] [1].
- **Periodicidad**: Es la característica de una señal compuesta por ciclos repetitivos. Se puede calcular basándose en la auto-correlación de la señal y calculando la relación entre los picos formados en la frecuencia fundamental y el más próximo.
- **Pico Normalizado**: similar a la periodicidad, pero su cálculo no depende de la frecuencia fundamental, sino de los dos primeros picos en la auto-correlación.

- **Pico de Amplitud Cepstral (“Cepstral Prominence Peak” CPP)**: Pico en la amplitud de la representación cepstral (frecuencias de Mel) de la señal.
- **Ratio Energía de Frecuencias Bajas/Altas (“Low-High Ratio” LHR)**: Relación entre la energía de las frecuencias bajas y las frecuencias altas. Se divide el espectro de Fourier en una frecuencia central, para estas pruebas ésta es de 2.000 [Hz].
- **Inclinación Espectral**: Pendiente formada al buscar una recta que aproxime el decaimiento de energía entre la frecuencia fundamental y sus armónicos.
- **Ratio de Cruces por Cero (“Zero-crossing Rate” ZCR)**: Proporción de cantidad de cruces por cero que hay en el segmento de la señal comparado con el tiempo del segmento analizado.

Una vez calculada una característica, se revisa que su valor esté en un rango definido previo a la ejecución. Para poder identificar que en un segmento de señal hay presencia de actividad de las cuerdas vocales, este algoritmo analiza que todas las características definidas previamente tienen que estar dentro del rango esperado, de lo contrario el segmento es descartado como silencio.

3.2.Estrategia de Implementación

A continuación, se procede a evaluar la efectividad de cada característica y buscar qué valores límites son los óptimos para clasificar correctamente las señales según las marcas de la base de datos. Para esto se realiza lo siguiente:

1. Sincronizar las señales del micrófono y del acelerómetro por posible desfase.
2. Por cada característica, discriminar si hay presencia de la actividad de las cuerdas vocales evaluando el resultado numérico de dicha característica con respecto a un parámetro.
3. Iterar un número N=20 de veces, variando el parámetro, para encontrar el que minimiza el error.
4. Repetir por cada señal de la base de datos.
5. Promediar los parámetros óptimos de cada característica para así obtener un valor único.
6. Volver a evaluar en cada señal con el parámetro obtenido.

3.3.Resultados

Al aplicar el método en las señales de micrófono y acelerómetro, los resultados obtenidos se muestran en la Figura 3-1:

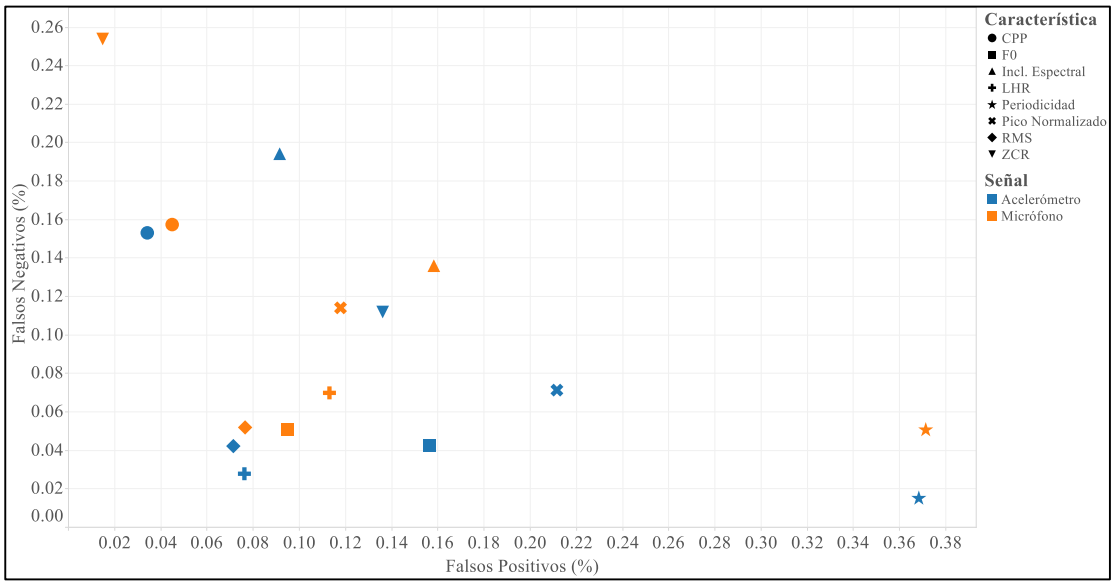

Figura 3-1: Diagrama de dispersión de Errores de clasificación por característica y tipo de señal

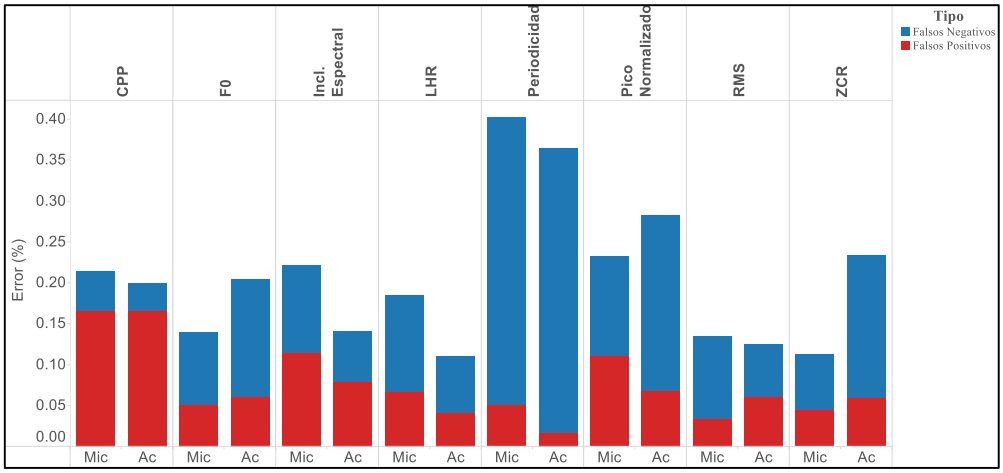

Figura 3-2: Diagrama acumulado de Errores de clasificación por característica y tipo de señal

A partir de estos gráficos se puede ver que distintas características tienen distintos resultados para señales de micrófono y acelerómetro, por lo que se puede concluir que no todo algoritmo de Detección de la Voz funcionaría de forma óptima en una señal de acelerómetro detectando la actividad de las cuerdas vocales.

Las características que más sobresalen por el bajo error total obtenido son el RMS, Ratio Energía de Frecuencias Bajas/Altas y la Inclinación Espectral, por lo que son candidatos para ser utilizados en el algoritmo final.

Además, se incluye la Frecuencia Fundamental por su bajo nivel de falsos positivos, por lo que ayuda a obtener un resultado más estricto al momento de clasificar un segmento con presencia de la actividad de las cuerdas vocales.

CAPÍTULO 4. IMPLEMENTACIÓN DE ALGORITMO ADAPTIVO PARA ANALISIS DE ENERGÍA

El algoritmo adaptivo para el análisis de energía, propuesto por Dr. Ying [4], utiliza una estrategia con modelos estadísticos basado en modelos Gaussianos, llamado "Sequential Gaussian Mixture Model" (SGMM).

La idea base de este algoritmo es agrupar la energía del espacio de la frecuencia de Mel en dos modelos Gaussianos, uno correspondiente a la presencia de voz, y otro a la energía del ruido asociado, para luego determinar si un segmento de señal, por medio de su energía, pertenece al modelo del ruido o al de la voz. Estos modelos van siendo actualizados de acuerdo se van analizando nuevos segmentos.

Como se muestra en la Figura 4-1, este modelamiento se aplica a las señales obtenidas por el acelerómetro al medir la actividad de las cuerdas vocales.

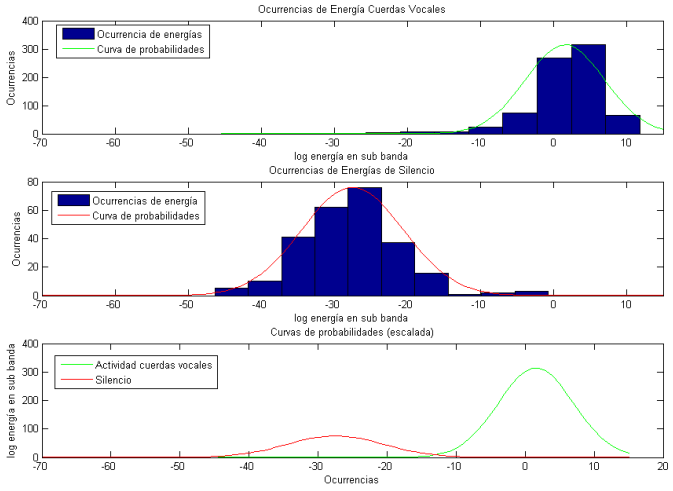

Figura 4-1: Gráficos que ilustran la formación de curvas Gaussianas

4.1. Algoritmo basado en SGMM

El algoritmo se describe con el siguiente seudocódigo:

Para los primeros N segmentos
Calcular FFT y separar en sub-bandas
Por cada sub-banda
Extraer el logaritmo de la energía
Establecer los parámetros de los dos modelos Gaussianos usando k-means clustering
Determinar el límite entre ambos modelos (Ecuación 4-3)
Si logaritmo de la energía > límite
Clasificar como actividad de las cuerdas vocales
Si no
Clasificar como silencio
Fin
Fin
Realizar la votación de cada sub-banda por segmento
Fin
Por cada nuevo segmento recibido ***k***
Calcular FFT y separar en sub-bandas
Por cada sub-banda
Extraer el logaritmo de la energía ***x***
Calcular la probabilidad de ***x*** *en cada modelo Gaussiano (Ecuación 4-1)*
Actualizar los parámetros de los modelos Gaussianos (Ecuaciones 4-4, 4-5, 4-6)
Determinar el límite entre ambos modelos (Ecuación 3)
Si logaritmo de la energía > límite
Clasificar como actividad de las cuerdas vocales
Si no
Clasificar como silencio
Fin
Fin
Clasificar el segmento según la votación por sub-bandas
Fin

Las siguientes ecuaciones fueron extraídas de la publicación del autor [4].

La Ecuación 1 muestra el cálculo de la probabilidad en cada modelo Gaussiano **z** dado el logaritmo de la energía **x** en el segmento **k**, en una sub-banda. Siendo $p(x_k|z,\lambda_k)$ la probabilidad de $\boldsymbol{x_k}$ en el modelo **z**, $\boldsymbol{\lambda}$ los parámetros que definen los modelos, y $\boldsymbol{w_{k,z}}$ el peso asociado a un modelo en el segmento **k**, el cual lleva el promedio de las probabilidades $p(z|k_k,\lambda)$ según la Ecuación 4-4.

$$p(z|x_k,\lambda_k) = \frac{w_{k,z}p(x_k|z,\lambda_k)}{\sum_z w_{k,z}p(x_k|z,\lambda_k)} \textbf{ (4-1)}$$

La ecuación 2 muestra el cálculo de la probabilidad de $\boldsymbol{x_k}$ en el modelo **z**. Siendo $\boldsymbol{\kappa_z}$y $\boldsymbol{\mu_z}$ la varianza y la media del modelo **z**.

$$p(x_k|z,\lambda) = \frac{1}{\sqrt{2\pi\kappa_z}} e^{\frac{-(x_k-\mu_k)^2}{2\kappa_z}} \textbf{(4-2)}$$

La Ecuación 3 muestra la expresión que para buscar $\boldsymbol{\theta}$, que representa el valor del logaritmo de la engería, cuya probabilidad es igual en ambos modelos Gaussianos. Siendo **z=1** representa la actividad de las cuerdas vocales, **z=0** representa el silencio

$$p(\theta|z=1,\lambda)p(z=1) = p(\theta|z=0,\lambda)p(z=0) \textbf{ (4-3)}$$

La Ecuación 5 muestra el cálculo del peso $\boldsymbol{w_{k+1}}$ en un segmento **k+1**, a partir del peso del segmento anterior $\boldsymbol{w_k}$, la probabilidad $p(x_k|z,\lambda_k)$ descrita en la Ecuación 1, y un factor de retención $\boldsymbol{\alpha}$, el que determina cuánta más influencia tiene el valor anterior con respecto al nuevo valor calculado. El punto de partida se define arbitrariamente en 0.5.

$$w_{k+1} = \alpha w_k + (1-\alpha)\, p(x_k|z,\lambda_k) \textbf{ (4-4)}$$

La Ecuación 6 y 7 muestran cómo se actualizan la media $\mu_{k+1,z}$ y la varianza $\kappa_{k+1,z}$ del modelo Gaussiano **z**

$$\mu_{k+1,z} = \alpha w_{k,z}\mu_{k,z} + \frac{(1-\alpha)\, x_k\, p\left(x_k \middle| z, \lambda\right)}{w_{k+1,z}} \textbf{ (4-5)}$$

$$\kappa_{k+1,z} = \alpha w_{k,z}\kappa_{k,z} + \frac{(1-\alpha) p\left(x_k \middle| z, \lambda\right)\left(x_k - \mu_{k+1,k}\right)^2}{w_{k+1,z}} \textbf{ (4-6)}$$

4.2.Estrategia de Implementación

Al igual que en las pruebas del algoritmo actual, se realizan una serie de iteraciones para buscar los parámetros que disminuyen el error al clasificar las señales de la base de datos, para luego obtener un promedio de los mejores valores.

Los parámetros a evaluar son los siguientes:

- **Factor de retención:** Factor que define qué tan relevante es la información pasada, afectando qué tan rápido el sistema varía ante variaciones de amplitud de la señal.
- **Número de sub-bandas:** Número de sub-bandas a analizar por separado (Figura 4-2). Es importante que en las sub-bandas se encuentren las frecuencias de la actividad de las cuerdas vocales, ya sea la frecuencia fundamental o alguno de sus armónicos, y no abarque solo frecuencias entre armónicos o demasiado altas.
- **Frecuencia máxima:** Frecuencia máxima a analizar. La señal generada por la actividad de las cuerdas vocales en el acelerómetro no cubre todo el espectro de frecuencias con el que se está muestreando la señal, por lo que acotar a una frecuencia máxima es necesario para no tener sub-bandas sin información de la señal a analizar.

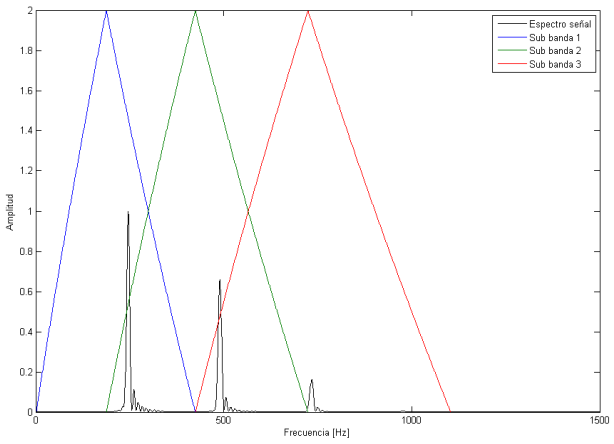

Figura 4-2: Gráfico que muestra la distribución del banco de filtros en escala Mel con 3 sub-bandas

4.3.Resultados

El resultado de la prueba en busca de los parámetros óptimos presentados en la Tabla 4-1, muestra una mejora promedio de aproximadamente 1.7% en el error total. De los parámetros encontrados, se puede ver una tendencia a mantener un factor de retención alto, lo que hace al algoritmo menos dispuesto a variar ante variaciones cortas, como momentos de silencio entre frases.

Sobre la frecuencia máxima de análisis, se puede ver una tendencia a sólo analizar la frecuencia fundamental en la mayoría de los casos, lo que elimina la información adicional que pueden entregar los armónicos generados, esto debido principalmente a la falta de ruido, lo que hace que el análisis de la fundamental sea suficiente. Además, la presencia de armónicos no se da en todos los casos, como fue mostrado en el Capítulo 2, por lo que su análisis podría llevar a falsos negativos.

Tabla 4-1: Resultados parámetros óptimos algoritmo SGMM

Muestra	Factor de Retención	Frecuencia Máxima [Hz]	Sub Bandas	Error Total (%)	Error RMS (%)
N1f	0.84	1212.75	3	5.01 %	7.25 %
N3f	0.78	882	3	4.16 %	6.13 %
N4f	0.69	882	4	5.36 %	8.40 %
N5m	0.66	882	8	5.99 %	7.83 %
N6m	0.63	882	8	9.92 %	10.97 %
P1m_MI	0.84	882	5	5.92 %	7.62 %
P2f_SE	0.78	1212.75	5	11.39 %	13.61 %
P3f_MI	0.84	882	4	5.56 %	9.18 %
P4m_SE	0.87	882	3	4.66 %	4.50 %
P5m_MO	0.66	1543.5	3	6.73 %	7.98 %
P6f_MO	0.6	2205	6	5.04 %	6.48 %
PROM	**0.74**	**1122.55**	**5**	**6.34 %**	**8.18 %**

En la Tabla 4-2 se pueden ver los resultados con la implementación con los promedios de los parámetros aplicados a todas las señales.

Tabla 4-2 : Resultados parámetros óptimos promediados

Muestra	Falsos Positivos (%)	Falsos Negativos (%)	Error Total (%)
N1f	5.08	1.40	6.48
N3f	2.40	2.27	4.67
N4f	4.08	2.58	6.66
N5m	7.23	1.35	8.58
N6m	0.87	23.80	24.67
P1m_MI	3.46	7.09	10.55
P2f_SE	10.96	3.80	14.76
P3f_MI	0.3	41.17	41.47
P4m_SE	1.38	6.19	7.57
P5m_MO	1.95	12.67	14.62
P6f_MO	32.39	0.19	32.58
PROM	**6.37**	**9.32**	**15.69**

Los resultados obtenidos muestran resultados variados, teniendo muestras en donde el error llega a ser menor al 5%, mientras que en otras llega a ser superior al 30%.

De las señales con peor desempeño, dos de ellas tienen una gran cantidad de falsos positivos, los que pueden ser rectificados al usar el análisis por característica descrito en el Capítulo 3.

CAPÍTULO 5. IMPLEMENTACIÓN FINAL. PRUEBAS Y RESULTADOS

Luego de haber realizado las pruebas de cada una de las características utilizadas en el algoritmo actual, junto con la implementación del algoritmo en base a GMM, se procede a idear un algoritmo que contemple las características con mejor resultado junto con GMM para así lograr mejores resultados.

5.1. Algoritmo Final

El algoritmo cuenta principalmente de dos partes, como se muestra en la Figura 5-1, la primera es el análisis de energía a través de SGMM, el que funciona como primer filtro capaz de identificar de forma temprana los segmentos de silencio. Si esta primera parte entrega un resultado positivo, se prosigue a verificar si es una señal de la actividad de las cuerdas vocales, clasificando por las características de mejor rendimiento durante las pruebas anteriores, frecuencia fundamental, inclinación espectral, y ratio de energía de frecuencias Bajas/Altas, de las cuales, basta con que dos de ellas entregue positivo para marcar el segmento como producto de la actividad de las cuerdas vocales.

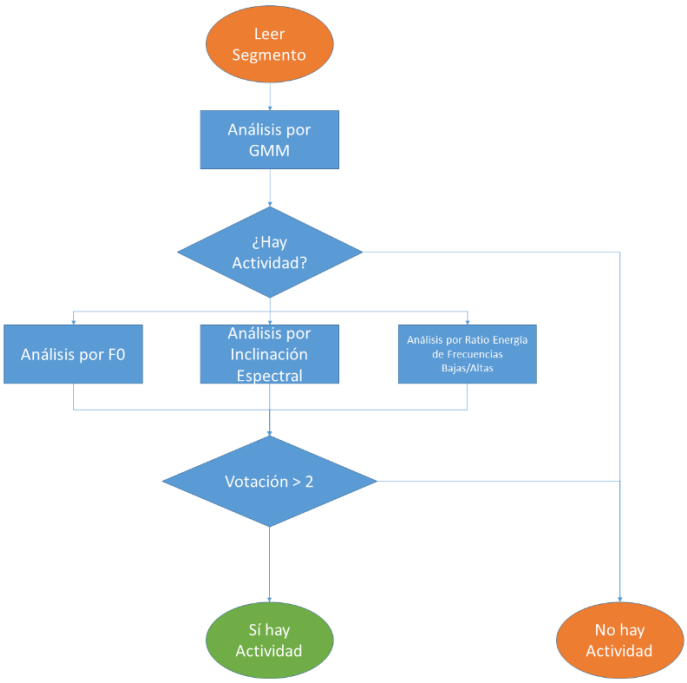

Figura 5-1: Diagrama de flujo del algoritmo final

5.2. Estrategia de Implementación

Para la implementación, se utiliza el promedio de los mejores valores obtenidos en las pruebas anteriores.

Además, se evalúa el tiempo de ejecución, para dar un primer acercamiento sobre si se alcanza a ser ejecutado en tiempo real. Para este análisis se calcula el costo en tiempo del algoritmo comparado con el tiempo de la señal (Ecuación 5-1)

$$Costo_t = \frac{tiempo_{algoritmo}}{tiempo_{señal}} < 1 \textbf{ (5-1)}$$

También se hace una proyección con el error obtenido y los resultados obtenidos en estudios pasados para calcular cuánto tiempo de señal se está ahorrando procesar, almacenar o transmitir. Para el cálculo del tiempo ahorrado, se utiliza la Ecuación 5-2 y se aproxima el tiempo promedio diario a 10 horas y utilización promedio de 6% según los datos presentados en el trabajo anterior [1], siendo fn el porcentaje de falsos negativos y fp el porcentaje de falsos positivos

$$Tiempo_{ahorro} = 10 * ((1 - (0.06 * (1 - fn))) * (1 - fp)) \, [hrs] \textbf{ (5-2)}$$

5.3. Resultados

Utilizando el algoritmo de detección de la actividad de las cuerdas vocales se obtienen los resultados mostrados en la Tabla 5-1, los que indican un error promedio del 11.37%, logrando un resultado con 4% menos de error que el algoritmo de SGMM y logrando un balance entre el promedio de falsos positivos y falsos negativos. Cabe destacar que los falsos positivos, el tipo de error más perjudicial ya que deja pasar ruido al análisis posterior, se mantiene menor al 8% en todas las muestras.

Tabla 5-1: Resultados del Algoritmo Final

Muestra	Falsos Positivos	Falsos Negativos	Error Total
N1f	0.0628	0.0279	0.0907
N3f	0.0311	0.0412	0.0723
N4f	0.0764	0.0512	0.1276
N5m	0.0748	0.0284	0.1032
N6m	0.0759	0.1185	0.1944
P1m_MI	0.0514	0.0334	0.0848
P2f_SE	0.0724	0.1737	0.2461
P3f_MI	0.0493	0.0299	0.0792
P4m_SE	0.0425	0.0202	0.0627
P5m_MO	0.0520	0.0559	0.1079
P6f_MO	0.0639	0.0179	0.0818
PROMEDIO	**0.0593**	**0.0544**	**0.1137**

El costo de tiempo del algoritmo final, por cada señal se muestra en la Tabla 5-2, en la que se puede ver que se requiere menos del 7% del tiempo de cada segmento para realizar el análisis de detección de actividad de las cuerdas vocales, cumpliendo con la ejecución en tiempo real y dejando un 93% del tiempo para procesos siguientes.

Las pruebas para el análisis del costo de tiempo fueron realizadas con el algoritmo implementado en Matlab 2014, en un computador portátil ASUS G501J con Windows 10, por lo que se esperan diferencias al momento de ser implementado en un teléfono inteligente, siendo los resultados obtenidos un primer acercamiento para validar la posibilidad de ejecución en tiempo real.

Tabla 5-2: Carga del algoritmo

Muestra	Señal [s]	Procesamiento [s]	Costo (%)
N1f	28.9252	1.919	6.63%
N3f	35.3911	1.921	5.42%
N4f	27.0512	1.607	5.94%
N5m	40.2049	2.439	6.07%
N6m	28.8724	1.659	5.75%
P1m_MI	27.9955	1.553	5.55%
P2f_SE	50.6492	2.342	4.62%
P3f_MI	31.0575	1.550	4.99%
P4m_SE	59.1209	2.718	4.60%
P5m_MO	30.4236	1.721	5.66%
P6f_MO	35.4435	1.636	4.61%
PROMEDIO	**35.9214**	**1.915**	**5.44%**

Si se implementase este algoritmo en el dispositivo de monitoreo continuo, se estaría el ahorrando 88.73% de los datos diarios, equivalentes a 8 horas y 52 minutos de señal que no sería procesada, almacenada o transmitida. Solo el 11.27% de la señal sería procesada, lo que equivale a solo a 1 hora y 8 minutos diarios.

CONCLUSIONES

A partir del trabajo desarrollado se pueden realizar las siguientes conclusiones:

- Al utilizar un acelerómetro para monitorear la actividad de las cuerdas vocales se puede observar claramente la señal de éstas en los datos capturados, lo que lo hace posible distinguir y estudiar esta señal por este medio. Otra ventaja es que, al utilizar vibraciones en la superficie de la piel, es inmune al ruido, pero no así a los movimientos que naturalmente se realizan en la vida cotidiana, como es el caminar.
- Utilizar algoritmos de detección de la voz tradicionales para detectar la actividad de las cuerdas vocales en la señal de un acelerómetro es un buen punto de partida, pero no es óptimo en la mayoría de los casos. Se requiere un ajuste en los parámetros a utilizar en cada caso y las características se pueden discriminar de mejor o peor manera dependiendo si la señal es de micrófono o acelerómetro.
- Al realizar un análisis sobre ventanas de tiempo movibles sin traslape, no se obtuvo error cercano a cero, esto debido a la presencia del error de discretización temporal. Por lo tanto, no se puede asegurar que a los siguientes procesos del proyecto se va a tener una señal completamente formada por la actividad de las cuerdas vocales.
- Aun considerando los errores del algoritmo propuesto, este es capaz de ahorrar gran parte de la señal diaria, consumiendo pocos recursos, lo que generaría un gran impacto en el uso de recursos de un sistema móvil, ya sea en energía consumida, almacenamiento utilizado y en tiempos de transferencia.

Trabajo futuro

Como trabajo futuro se propone:

- Realizar un análisis detallado de los tipos de ruidos vibratorios a los que están sometidos los pacientes (movimientos propios, transporte público, conciertos).
- Con este análisis, realizar los ajustes necesarios al algoritmo para mantener o mejorar su eficacia en dichos casos.
- Realizar un estudio para disminuir el error de discretización temporal utilizando, por ejemplo, ventanas de tiempo traslapadas.
- Implementar el algoritmo final en un celular con Android y confirmar la ejecución en tiempo real en el dispositivo.

REFERENCIAS

[1] Daryush D. Mehta, Matías Zañartu, Shengran W. Feng, Harold A. Cheyne II, and Robert E, Hillman, et al. "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform." Biomedical Engineering, IEEE Transactions on 59.11 (2012): 3090-3096.

[2] J. Ramirez, J. M. Górriz, and J. C. Segura. “Voice activity detection. fundamentals and speech recognition system robustness.” na, 2007.

[3] François G. Germain, Dennis L. Sun, and Gautham J. Mysore. "Speaker and noise independent voice activity detection." INTERSPEECH. 2013.

[4] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework." Audio, Speech, and Language Processing, IEEE Transactions on 19.8 (2011): 2624-2633.

[5] Phillip De Leon, and Salvador Sanchez. "Voice activity detection using a sliding-window, maximum margin clustering approach." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[6] Xiao-Lei Zhang, and Ji Wu. "Deep belief networks based voice activity detection." Audio, Speech, and Language Processing, IEEE Transactions on21.4 (2013): 697-710.

[7] T. Kinnunen, and P. Rajan. "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data." ICASSP. 2013.

[8] J. Sohn, N. Kim, and W. Sung. "A statistical model-based voice activity detection." Signal Processing Letters, IEEE 6.1 (1999): 1-3.

[9] Xiao-Lei Zhang. "Unsupervised domain adaptation for deep neural network based voice activity detection." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.

[10] Md Jahangir Alam, Patrick Kenny, Pierre Ouellet, Themos Stafylakis, Pierre Dumouchel, et al. "Supervised/Unsupervised Voice Activity Detectors for Text-dependent Speaker Recognition on the RSR2015 Corpus."

[11] Lauwereins, S., Meert, W., Gemmeke, J. F., & Verhelst, M. (2014, September). Ultra-low-power voice-activity-detector through context-and resource-cost-aware feature selection in decision trees. In Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on (pp. 1-6). IEEE.

ANEXOS

Los siguientes anexos se encuentran en el CD correspondiente a esta memoria:

- ANEXO 1: Código fuente en Matlab con las pruebas realizadas
- ANEXO 2: Código fuente en Matlab con el algoritmo final
- ANEXO 3: Base de Datos de señales de pruebas
- ANEXO 4: Bibliografía utilizada