Investigación
doi: https://doi.org/10.16925/in.v23i13.2006

Técnicas de detección de la frecuencia fundamental de la voz en entornos reales

María Manuela Silva Zambrano[1], Harold Armando Romo Romero[2], Jesús Mauricio Ramírez Viáfara[3], Diana María Galvis Zambrano[4]

[1] Ingeniera Electrónica y de Telecomunicaciones. Docente de la Facultad de Ingeniería Electrónica y Telecomunicaciones. Universidad del Cauca. Popayán, Colombia
Correo electrónico: mariasilva@unicauca.edu.co
[2] Magíster en Electrónica y de Telecomunicaciones. Docente Facultad de Ingeniería Electrónica y Telecomunicaciones. Universidad del Cauca. Popayán, Colombia
[3] Magíster en Electrónica y de Telecomunicaciones. Docente Facultad de Ingeniería Electrónica y Telecomunicaciones. Universidad del Cauca. Popayán, Colombia
[4] Médico General. Pontificia Universidad Javeriana. Bogotá, Colombia

Fecha de recibido: 28 de abril del 2017 **Fecha de aprobado:** 25 de agosto del 2017

Cómo citar este artículo: M. M. Silva-Zambrano, H. A. Romo-Romero, J. M. Ramírez -Viáfara y D.M. Galvis-Zambrano, "Técnicas de detección de la frecuencia fundamental de la voz en entornos reales", *Ingeniería Solidaria*, vol. 13, n.º 23, pp. 122-137, Sept. 2017. doi: https://doi.org/10.16925/in.v23i13.2006

Resumen: *Introducción*: este artículo de revisión se desarrolló como parte de un trabajo de maestría de la Universidad del Cauca en el 2017. Buscó encontrar los métodos de detección de la frecuencia fundamental de la voz más propicios para implementar en entornos reales como parte de una solución para mejorar la comunicación de las personas con discapacidad auditiva e incluirlas en la sociedad, ya que la mayoría de las propuestas solo plantean mejorar el canal de comunicación en el que la persona con discapacidad auditiva es el transmisor. *Metodología*: se realizó una revisión actualizada de la literatura, por lo que se basó principalmente en artículos científicos publicados en los últimos cinco años. Para la inclusión de artículos se hizo un mapeo sistemático sobre los diferentes métodos de detección de la frecuencia fundamental de la voz. *Resultados*: los fenómenos contemplados por los diferentes algoritmos para definir el entorno van desde el ruido y la interferencia hasta la reverberación; el desempeño del algoritmo depende de la calidad del audio grabado, lo que se ve en las variaciones obtenidas que dependen de la base de datos utilizada; se pueden detectar hasta dos frecuencias fundamentales diferentes. *Conclusiones*: se han implementado novedosos métodos para hacer más eficiente la detección de la frecuencia fundamental de la voz; sin embargo, aún queda mucho trabajo por hacer en esta área.

Palabras clave: detección, entornos reales, frecuencia, fundamental, voz.

ISSN 2357-6014 (en línea)

doi: https://doi.org/10.16925/in.v23i13.2006

Techniques for detecting voice fundamental frequency in real environments

Abstract. *Introduction*: This review article was prepared as part of a graduate thesis at Universidad del Cauca in 2017. It sought to find the most appropriate methods for detecting voice fundamental frequency to be implemented in real environments. This is part of a solution to improve the communication of people with hearing disabilities and include them in society, since most of the proposals only aim to improve the communication channel in which the hearing-impaired individual is the transmitter. *Methodology*: An updated review of the literature was carried out, based mainly on scientific articles published in the last five years. For the inclusion of articles, a systematic mapping was performed on the different methods for detecting voice fundamental frequency. *Results*: the phenomena considered by the various algorithms to define the environment range from noise and interference to reverberation; the performance of the algorithm depends on the quality of the recorded audio, which is observed in the variations obtained which depend on the database used; up to two different fundamental frequencies can be detected. *Conclusions*: Novel methods have been implemented to make the detection of voice fundamental frequency more efficient; however, there is still much work to be done in this area.

Keywords: detection, real environments, frequency, fundamental, voice.

Técnicas de detecção da frequência fundamental da voz em entornos reais

Resumo. *Introdução*: o artigo de revisão foi desenvolvido como parte de um trabalho de mestrado da Universidade de Cauca em 2017. Buscou-se encontrar os métodos de detecção da frequência fundamental da voz mais propícios para implementar em entornos reais como parte de uma solução para melhorar a comunicação das pessoas com deficiência auditiva e incluí-las na sociedade, já que a maioria das propostas enfocam apenas em melhorar o canal de comunicação no qual a pessoa com deficiência auditiva é o transmissor. *Metodologia*: é realizada uma revisão atualizada da literatura, baseada principalmente em artigos científicos publicados nos últimos cinco anos. Para a inclusão de artigos, foi feito um mapeamento sistemático sobre os diferentes métodos de detecção da frequência fundamental da voz. *Resultados*: mostram que os fenômenos contemplados pelos diferentes algoritmos para definir o entorno vão desde o ruído e a interferência até a reverberação; o desempenho do algoritmo depende da qualidade do áudio gravado, o que se vê refletido nas variações obtidas que dependem da base de dados utilizada; é possível detectar até duas frequências fundamentais diferentes. *Conclusões*: métodos inovadores foram implementados para tornar a detecção da frequência fundamental da voz mais eficiente; no entanto, ainda há muito trabalho a ser feito nessa área.

Palavras-chave: detecção, entornos reais, frequência fundamental, voz.

1. Introducción

La comunicación juega un papel primordial en la vida diaria de los seres humanos. Las personas con discapacidad auditiva tienen dificultad a la hora de conseguir empleo [1] y acceder a servicios básicos, por ejemplo la educación y la salud [2], [3]. Es imperioso desarrollar e implementar soluciones que ayuden a mejorar la comunicación de estas personas; sin embargo, muchas de las soluciones encontradas en la literatura no se pueden llevar a una implementación real porque carecen de usabilidad. Las soluciones encontradas para mejorar la comunicación en personas con discapacidad auditiva son:

- *Pendiente*. En [4] proponen un pendiente que, al colgarse del cuello de la persona, detecta las señas que se están realizando frente a él; el problema radica en que este pendiente limita el área donde se pueden realizar las señas y algunas de estas requieren de movimientos más amplios.
- *Anillos*. En [5] y [6] proponen anillos que se encargan de detectar el movimiento de las manos y se comunican con un dispositivo encargado de realizar el procesamiento y la traducción. Esta forma es mucho más portable y cómoda para la persona con discapacidad auditiva; no obstante, se requieren algoritmos sin mucho peso computacional para que el procesamiento se pueda realizar en un dispositivo portátil y la traducción sea rápida.
- *Procesamiento de imágenes*. En [7] proponen un sistema en dos direcciones para mejorar la comunicación entre personas sordas y oyentes. El sistema captura la imagen de la seña, la reconoce y la convierte a texto; además es capaz de detectar audio y convertirlo en texto o en la imagen de la seña. El lenguaje de señas se basa en movimientos, los movimientos para describir las letras del abecedario son, en la mayoría de los casos, posiciones estáticas de los dedos de la mano, lo que facilita su detección y procesamiento, así como el despliegue de la letra detectada en una imagen para el caso del sentido opuesto de la comunicación; sin embargo, los movimientos que describen las palabras que comúnmente utilizamos —hola, casa, perro— requieren no solo de posiciones específicas de los dedos de las manos, sino también del movimiento de los brazos, de tal forma que el procesamiento de estos datos mediante el reconocimiento de imágenes y la conversión de audio a una secuencia de imágenes que describa el movimiento implica un alto peso computacional. Lo anteriormente expuesto, junto con la necesidad de una cámara que en todo momento esté capturando los movimientos de la persona con discapacidad auditiva, hacen de esta una solución poco práctica para implementar en la vida real.
- *Sistemas de control gestual*. Actualmente, se están utilizando tecnologías como el Kinect® [8] o Leap Motion® [9] para implementar soluciones que permitan detectar el lenguaje de señas. Las pruebas realizadas hasta el momento muestran buenos resultados en el desempeño de estos detectores, siempre y cuando la persona que está realizando las señas se encuentre en el rango de visión del dispositivo. El principal factor desfavorable de estas propuestas consiste en la falta de portabilidad.

Las soluciones anteriormente expuestas muestran que existe un interés muy grande por encontrar soluciones que permitan mejorar la comunicación con discapacidad auditiva; empero, se considera que no se han tenido en cuenta ni la usabilidad de la solución, ni el caso en el que la persona con discapacidad auditiva es la receptora.

El hecho de que no existan muchas soluciones para ayudar a las personas con discapacidad auditiva a recibir la información que se brinda de forma oral puede basarse en que muchas de ellas son capaces de leer los labios y así detectar lo que se les está diciendo; incluso, algunas personas son capaces de hablar al imitar los movimientos de los labios. Ahora bien, se debe tener en cuenta varios factores: el primero de ellos es que no todas las personas vocalizan bien, por lo que no es posible detectar el movimiento de los labios de todas las personas; el segundo es que la persona con discapacidad auditiva necesitaría poder ver todo el tiempo la boca de la persona con la que está hablando, y el tercero es que no es posible detectar el movimiento de los labios de más de una persona, mientras que nuestro cerebro es capaz de detectar, separar y comprender lo que dicen varias personas al tiempo. La solución que busca superar las limitaciones mencionadas y hacerlo de tal forma que se centre en la usabilidad —con el fin de que pueda ser implementada en la vida real— requiere un algoritmo capaz de detectar diferentes señales de voz, que se pueden traslapar en el tiempo, separarse y traducirse a texto individual, de tal forma que la persona

con discapacidad auditiva vea en una pantalla la conversación que está ocurriendo a su alrededor.

El propósito de esta revisión de la literatura es caracterizar el proceso de generación de la voz para entender así las dificultades al momento de detectar su frecuencia fundamental y, finalmente, determinar la viabilidad de implementar los métodos de detección de esta frecuencia en entornos reales, donde se tiene presencia de ruido e interferencias.

2. Materiales y métodos

Para esta revisión de la literatura se consultaron bases de datos especializadas como Scopus, IEEE y ScienceDirect, y además se realizaron búsquedas en Google Scholar. Todas las búsquedas se limitaron al periodo comprendido entre el 2013 y el 2017. Las búsquedas en bases de datos especializadas se realizaron mediante cadenas de búsqueda que abarcaran los principales temas que deberían abordar los artículos.

Con el fin de encontrar qué tan aptos son los métodos de detección de la frecuencia fundamental de la voz, propuestos hasta el momento, para ser implementados en entornos reales, se plantea realizar un mapeo sistemático (MS) siguiendo las directrices expuestas en [10]. El MS se realiza en cinco pasos, mostrados en la figura 1: en el primer paso, se definen las preguntas que se quieren responder con la investigación; el segundo paso consiste en realizar la búsqueda y en el tercer paso, los resultados son analizados para determinar si se incluyen o no dentro del MS los artículos incluidos se clasifican en el cuarto paso, y, en el quinto paso, se mapean.

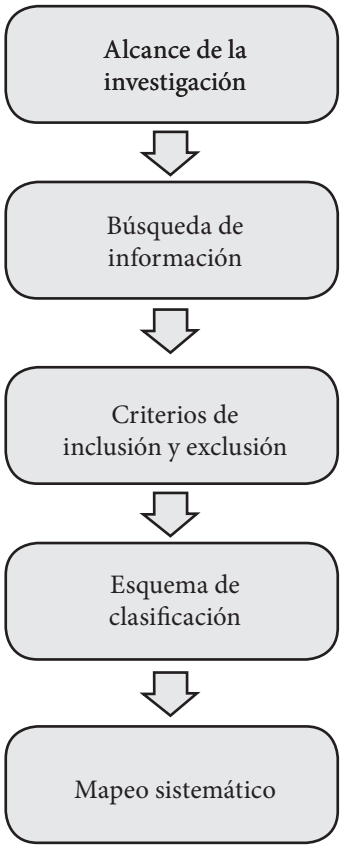

Figura 1. Proceso de mapeo sistemático
Fuente: elaboración propia

2.1 Definición de las preguntas de investigación

Se busca conocer los principales métodos para el reconocimiento de la frecuencia fundamental de la voz, averiguar si existen métodos que contemplen la presencia de ruido y, si es posible, realizar el reconocimiento simultáneo de múltiples frecuencias fundamentales.

De acuerdo con el propósito, se definen tres preguntas de investigación:

PI1: ¿cuáles son las principales formas con las que se realiza la detección de la frecuencia fundamental de la voz?

PI2: ¿es posible realizar la detección de la frecuencia fundamental de la voz en presencia de ruido?

PI3: ¿es posible realizar la detección simultanea de múltiples frecuencias fundamentales?

2.2 Búsqueda de información

El segundo paso del MS es la búsqueda de información. Para esto es necesario definir cadenas de búsqueda (CB) y las bases de datos especializadas en temas de ingeniería donde se van a realizar estas búsquedas.

Se definieron tres CB:

CB1: "pitch detection" AND "voice"
CB2: "pitch detection" AND "voice" AND "noise"
CB3: "pitch detection" AND "voice" AND "multi"

Las búsquedas realizadas en las tres bases de datos especializadas arrojaron los resultados que se muestran en la tabla 1.

Tabla 1. Resultados de la búsqueda inicial

Cadena	IEEE	ScienceDirect	Scopus
CB1	62	47	23
CB2	28	35	6
CB3	4	1	1
Total	94	83	30

Fuente: elaboración propia

2.3 Inclusión o exclusión de artículos

Los artículos incluidos deben ser aquellos que aporten a la resolución de las preguntas de investigación planteadas. Para esto con cada artículo se analizan su título, sus palabras clave, su resumen, y, en caso de que después de esto no exista claridad sobre la posible inclusión o exclusión del artículo, se leen la introducción y las conclusiones. Los artículos incluidos se muestran en la tabla 2.

Tabla 2. Artículos incluidos

Cadena	IEEE	SciencDirect	Scopus
ss1	2	0	1
ss2	1	3	1
ss3	1	2	1
Total	4	5	3

Fuente: elaboración propia

2.4 Esquema de clasificación

Los artículos seleccionados se clasifican en tres grandes grupos, de acuerdo a su aporte en la resolución de alguna de las tres preguntas de investigación.

Una segunda clasificación se realiza entre los artículos para determinar cuáles son los principales métodos empleados en la detección de la frecuencia fundamental de la voz; las categorías definidas en esta clasificación se basan en la planteada en [11]. La lectura de los resúmenes de los artículos permite dilucidar en cada caso el método base para dicha detección.

Las categorías definidas se muestran en la figura 2.

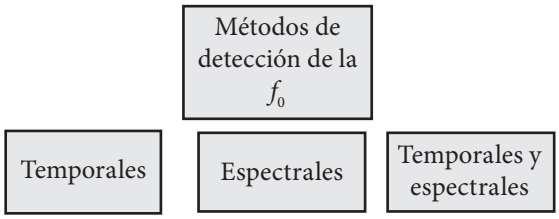

Figura 2. Esquema de clasificación
Fuente: elaboración propia

i. Temporales

La detección de la frecuencia fundamental en el dominio del tiempo se basa en la cuasiperiodicidad temporal de la señal de voz. De esta forma, se busca detectar la envolvente temporal correspondiente a la frecuencia fundamental. Algunos métodos temporales calculan la correlación entre diferentes funciones y las muestras de la señal.

ii. Espectrales

En el dominio de la frecuencia, la detección de la frecuencia fundamental trata de encontrar la presencia de frecuencias con mayor intensidad para encontrar así los armónicos más cercanos a la frecuencia fundamental.

iii. Temporales y espectrales

En el análisis realizado, tanto en el dominio temporal como en el espectral, se descompone la señal en múltiples subbandas y se aplican las técnicas en el dominio del tiempo a cada una de ellas.

3. Resultados

El procesamiento de las señales de voz tiene varias etapas: la primera consiste en identificar los momentos en que existe o no presencia de una señal de voz. En [12] realizan un estudio comparativo de las diferentes técnicas para la detección de la voz. La fase que corresponde al reconocimiento de la señal de voz ha sido ampliamente estudiada. En [13] estudian los mecanismos más robustos para el reconocimiento automático de las señales de voz.

Con el fin de que la persona sorda detecte la fuente de información, es necesario realizar la detección

y reconocimiento de voz, así como identificar a la persona o personas que están hablando.

3.1 Detección de la frecuencia fundamental de la voz

El sonido es producto de la vibración de las partículas del aire, que se genera por los cambios de presión que produce un cuerpo al vibrar. La propagación del sonido se da en forma de onda sinusoidal que se caracteriza, entre otras cosas, por su frecuencia. La voz en los seres humanos se produce por el aparato fonador cuando el aire que sale de los pulmones pasa por la tráquea y provoca una oscilación en las cuerdas vocales [14]. Estas no son en realidad cuerdas, sino más bien pliegues, como se observa en la figura 3, cuyo periodo de oscilación es el inverso de la frecuencia fundamental de la voz y le da su sonido característico [15]. Los órganos que componen el aparato fonador no tienen como función principal la generación de la voz —son parte de los sistemas respiratorio y digestivo—; esto lleva a pensar que la voz surgió como una necesidad evolutiva de los seres humanos [16].

La detección de la frecuencia fundamental de la voz tiene múltiples aplicaciones en las áreas de la salud, la seguridad, la computación, el entretenimiento y el control, entre otras.

3.2 Salud

En el área de la salud, existen diferentes propuestas para realizar la detección de enfermedades mediante el análisis de la frecuencia fundamental de la voz [18].

La detección de la enfermedad de Parkinson mediante las características de la voz del paciente se propone en [19]. Para esto, utilizaron la relación entre los armónicos y el ruido, y las variaciones que ciclo a ciclo podría tener la frecuencia fundamental. Los resultados muestran que, al comparar los resultados de las personas que padecen la enfermedad de Parkinson y quienes no, no hay una diferencia significativa, con lo cual se concluyó que este método no es efectivo para detectar la enfermedad. Sin embargo, es posible detectar diferentes patologías de la voz, ya que estas se deben a problemas en las cuerdas vocales.

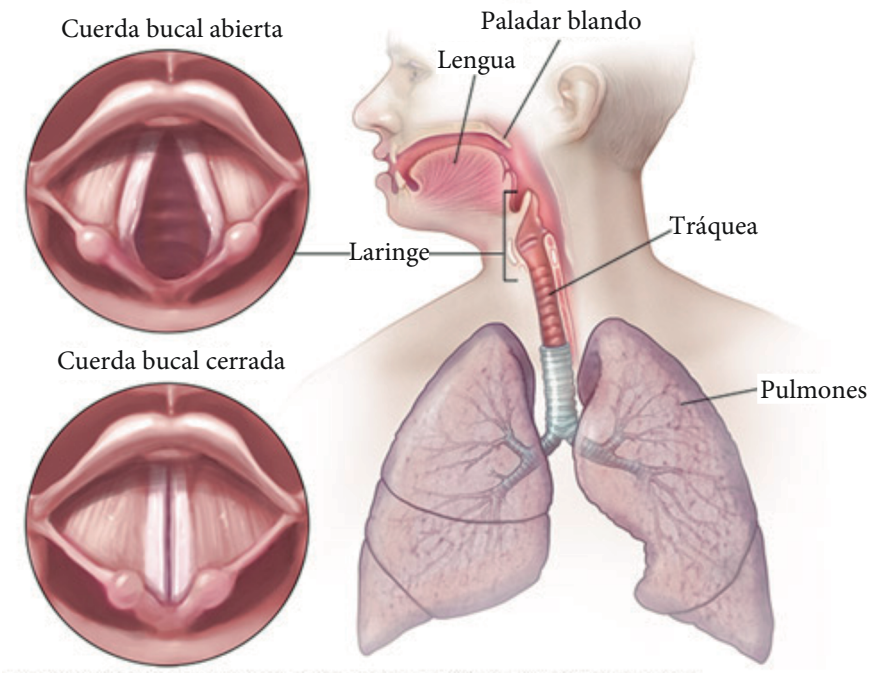

Figura 3. Aparato fonador
Fuente: [17]

En [20] se analizan los parámetros más relevantes al momento de realizar la detección de diferentes patologías de la voz; para esto se basan en los datos almacenados en tres bases de datos diferentes, lo que permite obtener así una precisión de hasta el 99,68%. Se concluye que la detección de la frecuencia fundamental de la voz es una herramienta que busca ayudar a los médicos a diagnosticar enfermedades en las primeras etapas. Se aclara que el diagnóstico no debe basarse únicamente en estos resultados, sino que además debe realizarse un examen médico al paciente.

Los ciclos de apertura y cierre de las cuerdas vocales se estudian para detectar patologías de la voz en [21], donde mediante un filtraje pasa bajas sucesivo encuentran irregularidades en dichos ciclos. Al evaluar sus resultados en dos bases de datos diferentes obtuvieron una precisión, en ambos casos, superior al 94%; con esto se concluye que las irregularidades en los ciclos de las cuerdas vocales son útiles para detectar patologías de la voz. En [22] también se propone la detección de las diferentes patologías de la voz, pero en este caso se realiza un análisis temporal y espectral de la señal que excita la cavidad glótica —caja cartilaginosa ubicada al final de la tráquea, donde se encuentran las cuerdas vocales— los resultados alcanzados mediante simulación muestran una alta precisión, alrededor del 90%; se concluye que aumentar el orden de derivación aumentaría la precisión. Finalmente, en [23], se utilizan las máquinas de vectores de soporte (SVM)[1] para clasificar las diferentes patologías de la voz. Se concluye a partir de los resultados obtenidos que este estudio es un paso inicial para alcanzar la modulación automática de tono para la electro laringe y las interfaces de habla silenciosa.

La detección de la diafonía en pacientes se propone en [24]. El algoritmo propuesto está basado en el método de estimación espectográfico de la relación entre los armónicos y el ruido, y los resultados obtenidos permiten concluir que con este método es posible detectar la diafonía. Por otro lado, un método no invasivo, con capacidad de aprender y adaptar los límites de decisión, para detectar enfermedades en la laringe es propuesto en [25]; los resultados de simulación muestran una especificidad del 94%, con lo cual se concluye que este método es útil tanto para detectar enfermedades en la laringe como para monitorear el progreso de los tratamientos.

3.3 Seguridad

En seguridad, es deseable identificar a una persona de acuerdo a su voz. Para lograr esto, en [26] parten de la creación de una base de datos tomando personas de diferente sexo. La base de datos se crea con muestras de señales de voz diciendo un texto predefinido; luego, se realiza el reconocimiento de patrones utilizando los modelos ocultos de Markov (HMM). Los resultados muestran que se obtiene una mayor precisión cuando se modelan por separado las señales de hombres y mujeres. El método propuesto es eficiente al momento de identificar a una persona según su voz; sin embargo, las pruebas realizadas fueron ideales, por lo que se deben realizar pruebas y ajustes para lograr un método eficiente en presencia de ruido.

En [27] se propone un sistema de reconocimiento de voz basado en redes neuronales artificiales, con el fin de que sea capaz de imitar el proceso de aprendizaje del cerebro y logre, así, diferenciar muchas voces —dada la gran cantidad de voces que un sistema de seguridad podría llegar a percibir—. Los resultados muestran que, en efecto, las redes neuronales permiten realizar el reconocimiento imitando el proceso de aprendizaje del cerebro; no obstante, se concluye que existe una falla en lo referente a seguridad, ya que este sistema solo puede identificar las voces que ya ha aprendido con anterioridad.

Otra aplicación relacionada con el área de seguridad se muestra en [28], en el cual se propone un esquema de autenticación mediante el reconocimiento de la voz. Este esquema se utiliza para identificar a las partes que se desean comunicar antes de establecer la comunicación. Los resultados obtenidos muestran que el método propuesto permite realizar con precisión el reconocimiento de la voz, pero además se concluye que el protocolo de conexión debe continuar mejorándose, con el fin de evitar ataques de seguridad.

En el marco de la investigación forense, se estudia la viabilidad de diferenciar las voces [29] incluso si se comparan las de gemelos. Los resultados muestran que, cuando se comparan la voz de una misma persona y la voz de diferentes personas existe una diferencia apreciable, por lo que se concluye que es posible realizar la identificación haciendo uso de la distancia euclidiana para calcular la similitud entre las voces.

3.4 Computación

El desarrollo de la tecnología, y en particular del *big data*, ha impulsado nuevos servicios en los dispositivos electrónicos, como lo son la traducción simultánea de voz a texto y las búsquedas por voz en la Web [30].

El avance hacia las nuevas interfaces entre humanos y computadores busca que su interacción sea más intuitiva, para lo cual se quiere que las máquinas sean capaces de interpretar las señales sociales que se encuentran inmersas en la voz y los gestos, entre otros [31]. Con el fin de que el computador pueda garantizar la satisfacción del usuario, se quiere que este sea capaz de detectar sus emociones. Dotarlo con propiedades afectivas permitirá que el computador tenga la capacidad de realizar acciones con el fin de aumentar la satisfacción del usuario. En [32] se plantea un algoritmo de detección de emociones en la voz que emplea SVM en el proceso de clasificación; los resultados muestran que este es un método eficiente y sencillo para realizarla. Además, se concluye que las diferentes funciones de Kernel influyen en la precisión de los resultados, siendo las funciones lineales y cuadráticas las que mayor eficiencia alcanzar para detectar emociones como miedo, felicidad y tristeza; la función polinómica es la de peor desempeño.

Detectar las emociones en la voz es el objetivo que se quiere alcanzar en [33], dado que se ha demostrado que los algoritmos para detectar emociones que hacen distinción en el género de la persona que habla son más precisos que aquellos que no toman este factor en consideración. Los autores proponen una etapa inicial en la cual se detecte el género de la persona para, de acuerdo a esto, realizar el procesamiento que busca detectar las emociones en la señal de voz. Esto les permitió alcanzar una precisión de alrededor del 90% en la detección de emociones entre dos estados, por lo que se concluye que la diferenciación de género es crucial para la detección de emociones, pero es necesario definir más parámetros para diferenciar las distintas emociones. En [34], los autores proponen un algoritmo para detectar emociones en la voz cuando se tienen conversaciones normales y no relacionadas con la actuación —que son el tipo de datos que analizan la mayoría de estos estudios—. Los datos alcanzados permiten concluir que sí es posible detectar emociones en conversaciones normales. El principal aporte de este estudio fue mostrar que, con el fin de obtener mayor información de las señales de voz, se debe implementar un sistema con memoria para determinar las emociones en la conversación.

3.5 Entretenimiento

En el entretenimiento, la habilidad de identificar quién está hablando o cantando es muy útil, como se muestra en [35], trabajo en el que se propone un esquema que, combinando las señales de audio y video, es capaz de rastrear a las personas que están hablando en el escenario, aunque la cámara se encuentre alejada y las personas estén dándole la espalda. Se concluye que los métodos propuestos permiten mejorar el desempeño, y se plantea como paso a seguir el estudio de la correlación entre los gestos y el habla para mejorar así el sistema planteado. Modificar el tono de voz de una persona o mejorar las características de diálogos susurrados son retos que se abordan en la transformación y conversión de la voz. En [36] se realiza una revisión de los sistemas de conversión de voz; el rango de aplicación para los algoritmos de transformación y conversión de la voz es muy amplio y no se limita únicamente al entretenimiento: abarca desde la reconstrucción de diálogos para mejorar el desempeño de sistemas de telecomunicaciones, hasta la implementación de ayudas para personas con problemas de audición.

3.6 Control

El espectro de una palabra dicha por diferentes personas puede llegar a variar de tal manera que la palabra no llegue a ser reconocida si se compara con una muestra espectral de dicha palabra. En los sistemas de control se busca distinguir comandos de voz para controlar los sistemas; por esto, en [37] se plantea un sistema de reconocimiento de comandos de voz que, a pesar de las posibles variaciones de la señal de entrada (niños, adultos, voces enfermas), sea capaz de reconocer los comandos. Para esto, se utiliza un método para detectar individualmente los fonemas que conforman cada comando. Los resultados obtenidos muestran que se consigue una mayor precisión que otros métodos de reconocimiento. Se concluye que es posible realizar el reconocimiento de los comandos incluso en ruso, idioma conocido por su complejidad fonética.

Un sistema de traducción de voz a lenguaje de señas se introduce en [38]; su implementación se realiza mediante el control de una mano robótica de bajo costo, ya que se puede imprimir en 3D. Esta mano es capaz de reproducir el alfabeto mediante dactilografía, las etapas de control y procesamiento se realizan por separado. En este caso se utiliza una tarjeta *Arduino* para el control y una *Raspberry Pi* para el procesamiento. Los resultados de este trabajo corroboran la conclusión de que la implementación de manos robóticas para la traducción de habla a lenguaje de señas es razonable y realizable.

Hasta el momento se han mostrado algunas de las múltiples aplicaciones que existen con respecto a la detección de la frecuencia fundamental de la voz, pero el procesamiento digital de la misma es una disciplina con más de dos décadas de antigüedad que aún tiene mucho por perfeccionar. En [39] se explica por qué es un gran reto el procesamiento de la voz; las señales de este tipo son muy complejas y de una alta variabilidad, tanto en su composición espectral como en su intensidad en el tiempo. Existen muchos factores que afectan el reconocimiento de la voz: el ruido del ambiente, la estructura del lugar y las características anatómicas y fisiológicas de la persona que habla, entre otros. Sin embargo, las señales de voz se pueden asumir cuasiperiódicas para facilitar el procesamiento.

En [40] se analizan las dificultades de determinar la frecuencia fundamental debidas a las características de las señales de voz. La articulación de palabras trasmitidas mediante la voz, que se conoce como habla, tiene las siguientes características:

- El habla no es un proceso estacionario, ya que las características del aparato fonador pueden cambiar abruptamente en el tiempo.
- Es posible que el tiempo total cuando hay presencia de habla solo dure unos pocos periodos fundamentales.
- Las combinaciones posibles entre la vocalización generada en el tracto vocal y las voces hacen que las estructuras temporales lleguen a ser casi infinitas.
- El rango espectral en el que puede estar la frecuencia fundamental es muy grande (50-800 Hz).
- La excitación que genera la voz no siempre es uniforme, aun en condiciones normales (la persona no padece patologías en los órganos del aparato fonador).

3.7 Métodos de detección de la frecuencia fundamental de la voz

Los artículos seleccionados de acuerdo al MS realizado son los siguientes:

1. Improving pitch estimation by enhancing harmonics [41]

Se propone un método de detección de la frecuencia fundamental de la voz en el dominio de la frecuencia. La señal de voz pasa por una etapa de preprocesamiento cuando se mejora la diferenciación de los armónicos, lo que da como resultado un mejor desempeño en presencia del ruido.

2. Multi-band summary correlogram-based pitch detection for noisy speech [42]

En este artículo se propone un método para la detección tanto en el dominio del tiempo como en el de la frecuencia. La señal se divide por subbandas y un proceso de mejoramiento de armónicos. Posteriormente, se encuentra la correlación.

La correlación de resumen de multibanda (MBSC) presenta un mejor desempeño que los métodos tradicionales del dominio del tiempo; además, tiene un buen comportamiento en presencia de ruido.

3. Two-pitch tracking in co-channel speech using modified group delay functions [43]

Las funciones de retardo de grupo están comenzando a cobrar importancia tanto en la detección como en el reconocimiento del habla, dado que con estas se obtiene una mejor resolución que con el espectro de magnitud calculado mediante la transformada de Fourier. El trabajo citado rastrea la frecuencia predominante, la cual una vez detectada se elimina junto con sus armónicos mediante un filtro de barrido para detectar la segunda frecuencia predominante. El rastreo de las frecuencias se realiza mediante el análisis de las funciones de retardo de grupo.

4. Multiple comb filters and autocorrelation of the multi-scale product for multi-pitch estimation [44]

En el estudio citado, se procesa la señal de voz en diferentes ambientes ruidosos. Con el fin de detectar las frecuencias fundamentales, se hace uso de la autocorrelación temporal. Luego, se detecta la frecuencia fundamental predominante; y una vez conocida, se elimina junto con sus componentes armónicas mediante un filtro de barrido. La señal resultante es procesada nuevamente para detectar la segunda frecuencia fundamental.

5. Tracking pitch period using particle filters [45]

En el estudio nombrado se utiliza un filtro de partículas para realizar la detección continua de la frecuencia fundamental de la voz en presencia de ruido; el método de detección se realiza en el dominio del tiempo. El proceso consiste en calcular las variaciones de la frecuencia fundamental cuando se tiene habla muy ruidosa. Los resultados muestran que con este método es posible determinar continuamente la frecuencia fundamental de la voz en casos no ideales.

6. Multipitch tracking with continuous correlation feature and hybrid DBNS/HMM model [46]

En este trabajo se propone un método para estimar continuamente la frecuencia fundamental de la voz en el dominio del tiempo, cuando otras señales del mismo tipo interfieren en la señal de voz. Utilizan las redes de creencias profundas (DBN) para encontrar las frecuencias más probables y después emplean HMM para determinar la frecuencia, las continuas iteraciones generan un contorno del habla del cual se puede determinar su periodo. Este método muestra ser más eficiente frente a otros que no realizan el análisis de continuidad.

7. A novel method for pitch detection via instantaneous frequency estimation using polynomial chirplet transform [47]

La frecuencia instantánea de una señal de voz se puede determinar mediante la transformada *wavelet*, la transformada de Fourier de corto tiempo o la distribución Wigner Ville; sin embargo, es posible obtener un mayor orden de generalización al representar la frecuencia como una función de mayor orden polinómica del tiempo. Esto se logra utilizando la transformación polinómica de Chirplet (PCT). Con la PCT no es necesario asumir el habla como un proceso estacionario, por lo que su frecuencia puede variar en el tiempo, los resultados obtenidos muestran precisión para determinar las frecuencias fundamentales de diferentes hombres y mujeres.

8. Automatic speaker recognition using a unique personal feature vector and gaussian mixture models [48]

El reconocimiento de la frecuencia fundamental de la voz se realiza utilizando la herramienta de simulación Matlab. Para esto se define el vector de la huella de la voz y se utilizan los modelos mezclados gaussianos para el proceso de clasificación. La huella de la voz corresponde a las características de esta. Con el fin de obtener dichas características, se realiza el análisis cepstral. Una vez conformado el vector, se utilizará para realizar el reconocimiento. Se observa, sin embargo, que la información aportada por el análisis en el dominio del tiempo puede llegar a ser redundante.

9. A method of speech periodicity enhancement using transform-domain signal decomposition [49]

Partiendo de la característica de la cuasiperiodicidad de las señales de voz, se plantea la mejora de la periodicidad aplicada al residuo de la predicción lineal del habla con el fin de mejorar el desempeño en situaciones con habla ruidosa. El algoritmo propuesto consiste en un proceso adaptativo para definir los pesos que determinan la porción del habla periódica y la aperiódica. La mejora en la periodicidad del habla depende de la lengua, siendo más eficiente en lenguas tonales que no tonales.

10. A multipitch tracking algorithm for noisy and reverberant speech [50]

La detección de la frecuencia fundamental se realiza tanto en el dominio del tiempo como en el

de la frecuencia, y se consideran efectos adversos sobre la señal de habla, como el ruido y la reverberación producida por el multitrayecto. Utilizando descriptores temporales y espectrales se calculan las probabilidades de las posibles frecuencias mediante HMM. Los resultados obtenidos muestran mayor eficiencia que otros métodos de detección, sobre todo en habla con reverberancia.

11. Pitch contour extraction of singing voice in polyphonic recordings of Indian classical music [51]

Las grabaciones de la música clásica india se clasificaron y se seleccionaron aquellas en las cuales predomina el canto y son armónicas. La estimación se realiza mediante el análisis de Fourier con la transformada de Fourier ejecutando la interpolación parabólica entre los picos espectrales. El método propuesto no cae en el error de estimación de la frecuencia por octava y tiene una gran precisión considerando la calidad de las grabaciones.

12. Pitch detection method based on morphological filtering and HHT [52]

Se propone un método de detección en el dominio de la frecuencia para habla ruidosa. Esta señal pasa por un filtro morfológico para remover el ruido. Posteriormente la transformada Hilbert-Huang (HHT) es aplicada y se calcula la energía instantánea. Esta energía permite detectar los momentos de cierre y apertura de las cuerdas vocales. Los ciclos de las cuerdas vocales permiten inferir el valor de la frecuencia fundamental de la voz.

El número de artículos que aporta a la resolución de las tres preguntas de investigación se muestra en la figura 4. Dado que en todos los textos se realiza la detección de la frecuencia fundamental de la voz, todos brindan información sobre los métodos actualmente más utilizados para esto; mientras que cinco artículos (1, 2, 5, 9 y 12) contemplan la presencia de ruido y cuatro la detección de múltiples frecuencias fundamentales (3, 4, 6 y 10).

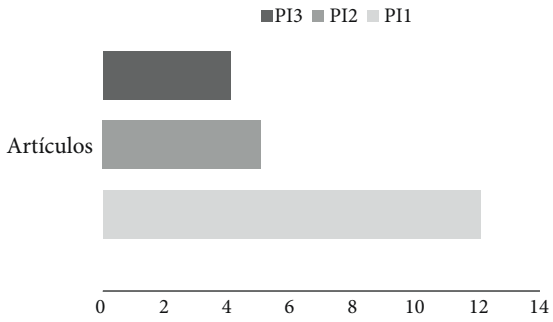

Figura 4. Artículos por pregunta de investigación
Fuente: elaboración propia

Para responder a la primera pregunta de investigación, que busca identificar los principales métodos para la detección de la frecuencia fundamental de la voz, se agruparon los diferentes artículos en las categorías definidas en el esquema de clasificación, los resultados obtenidos se sintetizaron en la figura 5.

Los métodos más explorados son los espectrales, en los cuales se realizan transformaciones basadas en la transformada de Fourier para mejorar la detección de la frecuencia fundamental, tales como las funciones de retardo de grupo, la transformada Cepstrum, la transformada Chirplet polinómica y la transformada Hilbert-Huang. El dominio del tiempo emplea modelos probabilísticos para encontrar la autocorrelación de la señal. En estos casos, se busca mejorar la periodicidad de la señal y se emplean redes neuronales. El método que considera análisis en los dominios del tiempo y la frecuencia es el que toma en cuenta múltiples frecuencias fundamentales, ruido y además reverberación en la señal de voz recibida. Con esto se busca que la información adicional brindada al realizar el análisis en los dos dominios mejore la precisión en la detección.

Es importante determinar cuáles son los métodos utilizados cuando se considera la presencia de ruido en la señal de voz. Los resultados encontrados se muestran en la figura 6, en la que se observa que el análisis espectral es el más estudiado. Los artículos 1 y 9 buscan mejorar la periodicidad de la señal para garantizar la correcta detección, según [41] y [49], mientras que otras propuestas utilizan métodos que implican un mayor procesamiento, como

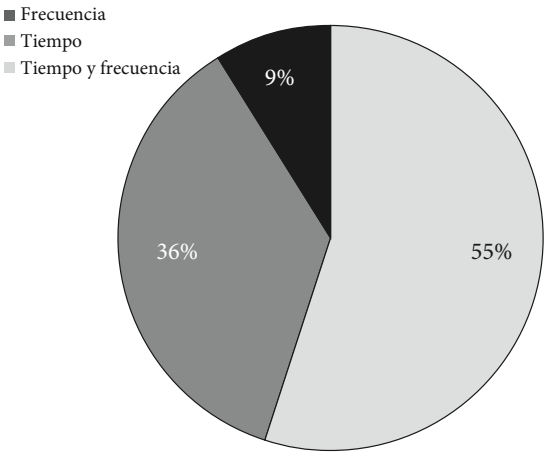

Figura 5. Agrupación de acuerdo al método de detección de la frecuencia fundamental
Fuente: elaboración propia

el artículo 5, según [45], en el que se utiliza un filtro de partículas.

El uso de filtros como una etapa de preprocesamiento para mitigar la presencia de ruido en la señal es útil cuando se tienen componentes de ruido que no se solapan con el ancho de banda de la señal.

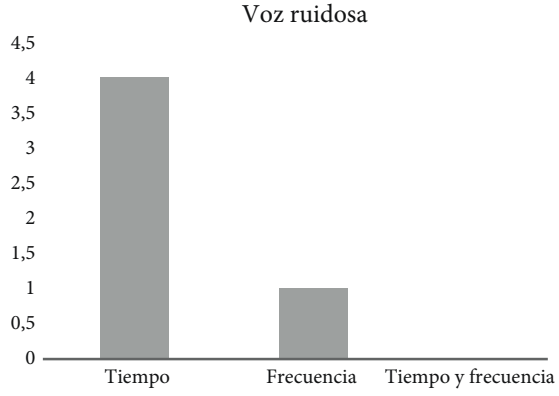

Figura 6. Métodos de detección de la frecuencia fundamental con ruido
Fuente: elaboración propia

En los casos en los que se consideran múltiples frecuencias fundamentales, los métodos más utilizados son los del dominio del tiempo, aunque la diferencia no es tan marcada como en el caso del ruido. Los resultados obtenidos se muestran en la figura 7.

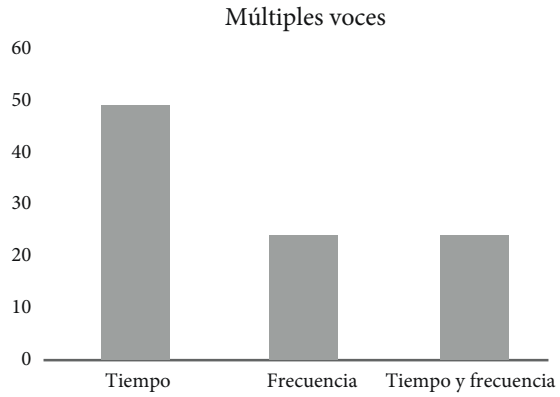

Figura 7. Métodos de detección de múltiples frecuencias fundamentales
Fuente: elaboración propia

De los cuatro artículos que consideran múltiples voces combinadas en la señal de audio, solo [43] y [44] consideran la detección simultánea de dos frecuencias fundamentales diferentes. Para esto proponen un procesamiento por fases, en las cuales, una vez detectada la primera frecuencia, se elimina junto con sus armónicos y la señal resultante se analiza nuevamente para detectar la segunda frecuencia. Los dos artículos realizan su análisis en dominios diferentes (tiempo y frecuencia).

4. Discusión y conclusiones

Los resultados obtenidos por las diferentes aplicaciones que realizan detección de la frecuencia fundamental de la voz revelan la importancia del proceso de adquisición de datos, dado que es posible obtener diferentes niveles de precisión de acuerdo a la base de datos consultada. Además, factores como la frecuencia de muestreo, el entorno y las posibles anomalías en la voz de las personas analizadas pueden influir notablemente en los resultados.

La eficacia de los métodos para la detección de la frecuencia fundamental de la voz varía de acuerdo a la aplicación en la que se utilicen; por ejemplo, en el área de la salud, para detectar enfermedades en la

voz se solicita al paciente que diga de forma sostenida una vocal. Esto es de gran ayuda cuando se asume la periodicidad de las señales de voz, pero no se lograrían los mismos resultados con un proceso de habla aleatorio.

Los resultados arrojados por el presente mapeo sistemático muestran que los métodos en el dominio del tiempo se utilizan cuando se requiere una mayor precisión, pero conllevan un mayor tiempo de procesamiento, por lo que habría que evaluar su viabilidad para aplicaciones en tiempo real.

La revisión de la literatura arrojó que existen pocos estudios que comparen los métodos más novedosos para la detección de la frecuencia fundamental de la voz, como las diferentes variaciones de la transformada de Fourier para la detección en el dominio de la frecuencia o el uso de diferentes filtros.

El procesamiento por fases permite identificar dos frecuencias fundamentales diferentes, pero no se encontraron trabajos que planteen la detección de un número mayor. La dificultad radica en que se asume que la señal tendrá una frecuencia fundamental dominante que se identifica en la primera fase del procesamiento. Esta, junto con sus armónicos, es eliminada mediante un filtro y la señal resultante se analiza nuevamente para encontrar la segunda frecuencia fundamental; sin embargo, otras frecuencias fundamentales podrían no diferenciarse lo suficiente del resto de componentes espectrales para ser identificadas.

Se debe tener en cuenta, además, que los humanos cuentan con dos oídos. El tener un oído en cada lado de la cabeza permite identificar la dirección del sonido y ubicar la fuente, por lo que se debe estudiar la posibilidad de diseñar arreglos de micrófonos para mejorar la detección de múltiples voces.

En dos artículos [44] y [45] se proponen métodos para detectar la frecuencia fundamental de la voz de forma continua, pero no se contempla la detección simultánea de diferentes frecuencias fundamentales, por lo que no existen trabajos que aborden todos los factores necesarios para realizar el reconocimiento de múltiples voces en entornos reales y efectuar el procesamiento en tiempo real.

Aunque el campo del procesamiento digital de la voz ha estado presente por varias décadas, todavía existen retos por superar en esta área. Esto se evidencia en las dificultades que presentan aún los sistemas que intentan entender la voz.

Los métodos actuales para la detección de la frecuencia fundamental de la voz utilizan diferentes transformas, como la HHT o la PCT, para obtener una mayor resolución en el espectro; además, se hace uso de filtros de alta discriminación como los filtros de barrido. Con el fin de aumentar la precisión se emplean algoritmos basados en redes neuronales o filtros de partículas, pero los métodos que emplean estos algoritmos tienen un mayor tiempo de procesamiento. Finalmente, en el dominio del tiempo son muy utilizados los modelos ocultos de Markov.

El procesamiento que se le da a la señal de voz antes de ser procesada para detectar su frecuencia fundamental es crucial, dado que factores como el muestreo de la señal y la ecualización previa al procesamiento, entre otros, son cruciales a la hora de determinar el desempeño del método de detección; ello se muestra en las variaciones obtenidas de acuerdo a la base de datos seleccionada en las aplicaciones de la detección de la frecuencia fundamental de la voz.

Referencias

[1] D. R. Terry, L. Quynh y B. Hoang, "Moving forward with dignity: Exploring health awareness in an isolated Deaf community of Australia", *Disabil. Health J.*, vol. 9, n.º 2, pp. 281-288, 2016. [Online]. doi: http://dx.doi.org/10.1016/j.dhjo.2015.11.002.

[2] A. Iglesias, J. Jiménez, P. Revuelta y L. Moreno, "Avoiding communication barriers in the classroom: the APEINTA project", *Interact. Learn. Environ.*, vol. 4820, n.º September, pp. 1-15, 2014. [Online]. doi: https://doi.org/10.1080/10494820.2014.924533

[3] R. Perkins, T. Battle, J. Edgerton y J. Mcneill, "A Survey of Barriers to Employment for Individuals Who Are Deaf", *J. Am. Deaf. Rehabil. Assoc.*, vol. 49, n.º 1, pp. 66-85, 2015. [Online]. Disponible en http://repository.wcsu.edu/jadara/vol49/iss2/3/.

[4] T. Starner, J. Auxier, D. Ashbrook y M. Gandy, "The Gesture Pendant: a self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring", *Iswc*, pp. 87–94, 2000. [Online]. doi: https://doi.org/10.1109/ISWC.2000.888469

[5] J. Wang, "Magic Ring: A Self-contained Gesture Input Device on Finger", *Proc. 12th Int. Conf. Mob.*

Ubiquitous Multimed., vol. 13, pp. 3-6, 2013. [Online]. doi: https://doi.org/10.1145/2541831.2541875

[6] M. Wilhelm, D. Krakowczyk, F. Trollmann y S. Albayrak, "eRing: multiple finger gesture recognition with one ring using an electric field", *Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction - WOAR '15*, 2015, pp. 1-6. [Online]. doi: https://doi.org/10.1145/2790044.2790047

[7] S. R. Ghorpade and S. K. Waghamare, "Full Duplex Communication System for Deaf & Dumb People", vol. 5, n.º 5, pp. 224-227, 2015. [Online]. Disponible en http://www.ijetae.com/files/Volume5Issue5/IJETAE_0515_38.pdf.

[8] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang y X. Chen, "Sign Language Recognition and Translation with Kinect", en *The 10th IEEE International Conference on Automatic Face and Gesture Recognition*, 2013, pp. 22-26. [Online]. Disponible en http://iip.ict.ac.cn/sites/default/files/publication/2013_FG_xjchai_Sign%20Language%20Recognition%20and%20Translation%20with%20Kinect.pdf.

[9] L. E. Potter, J. Araullo y L. Carter, "The Leap Motion controller", *Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration - OzCHI '13*, 2013, pp. 175-178. [Online]. Disponible en http://dl.acm.org/citation.cfm?doid=2541016.2541072.

[10] K. Petersen, R. Feldt, S. Mujtaba y M. Mattsson, "Systematic mapping studies in software engineering", *12th International Conference on Evaluation and Assessment in Software Engineering*, pp. 68-77, 2008. [Online]. Disponible en http://dl.acm.org/citation.cfm?id=2227123.

[11] L. Sukhostat y Y. Imamverdiyev, "A Comparative Analysis of Pitch Detection Methods Under the Influence of Different Noise Conditions", *J. Voice*, vol. 29, n.º 4, pp. 410-417, 2014. [Online]. doi: http://dx.doi.org/10.1016/j.jvoice.2014.09.016.

[12] M. Maky H. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations", *Comput. Speech Lang.*, vol. 28, n.º 1, pp. 295-313, 2014. [Online]. doi: http://dx.doi.org/10.1016/j.csl.2013.07.003.

[13] J. Li, L. Deng, Y. Gong y R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition", *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 22, n.º 4, pp. 745-777, Abr. 2014. [Online]. doi: https://doi.org/10.1109/TASLP.2014.2304637

[14] B. Torres, *Anatomía funcional de la voz*, España: Paidotribo, 2008. [Online]. Disponible en http://www.medicinadelcant.com/cast/1.pdf.

[15] D. Talkin, W. B. Kleijn y K. K. Paliwal, "A Robust Algorithm for Pitch Tracking (RAPT)", *Speech Coding and Synthesis*, Netherlands: Elsevier, 1995. [Online]. Disponible en https://www.ee.columbia.edu/~dpwe/papers/Talkin95-rapt.pdf.

[16] R. Dosal González, "Producción de la voz y el habla. La fonación", pp. 1-27, 2014. [Online]. Disponible en http://repositorio.unican.es/xmlui/bitstream/handle/10902/5583/DosalGonzalezR.pdf?sequence=1

[17] Clínica de Mayo, "Cuerdas vocales abiertas y cerradas - Mayo Clinic" [Online]. Disponible en http://www.mayoclinic.org/es-es/diseases-conditions/vocal-cord-paralysis/multimedia/vocal-cords-open-and-closed/img-20008069.

[18] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Noth y A. Revelo-García, "Detection of Different Voice Diseases Based on the Nonlinear Characterization of Speech Signals", *Expert Syst. Appl.*, vol. 82, pp. 184-195, 2017. [Online]. doi: https://doi.org/10.1016/j.eswa.2017.04.012

[19] A. Kacha, C. Mertens, F. Grenez, S. Skodda y J. Schoentgen, "On the harmonic-to-noise ratio as an acoustic cue of vocal timbre of Parkinson speakers", *Biomed. Signal Process. Control*, p. 7, 2016. [Online]. doi: http://dx.doi.org/10.1016/j.bspc.2016.09.004.

[20] A. Al-Nasheri *et al.*, "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification", *J. Voice*, vol. 31, n.º 1, pp. 113-118, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.jvoice.2016.03.019

[21] G. Muhammad, G. Altuwaijri y M. Alsulaiman, "Automatic Voice Pathology Detection and Classification Using Vocal Tract Area Irregularity", *EMS '13 Proceedings of the 2013 European Modelling Symposium*, 2016, vol. 6, pp. 164-168. [Online]. doi: https://doi.org/10.1016/j.bbe.2016.01.004

[22] G. Muhammad *et al.*, "Pathology Detection Using Interlaced Derivative Pattern on Glottal Source Excitation", *Biomed. Signal Process. Control*, vol. 31, pp. 156-164, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.bspc.2016.08.002.

[23] W. De Armas, K. A. Mamun y T. Chau, "Vocal Frequency Estimation and Voicing State Prediction with Surface EMG Pattern Recognition", *SPEECH Commun.*, vol. 63-64, pp. 15-26, 2014. [Online]. doi: http://dx.doi.org/10.1016/j.specom.2014.04.004.

[24] N. Vieira and P. H. Sansa, "Measurement of Signal-to-Noise Ratio in Dysphonic Voices by Image Processing of Spectrograms", vol. 62, pp. 17-32, 2014. [Online]. doi: https://doi.org/10.1016/j.specom.2014.04.001

[25] H. Ghasemzadeh, M. Tajik y M. Khalil, "Detection of Vocal Disorders Based on Phase Space Parameters and Lyapunov Spectrum", *Biomed. Signal Pro-*

cess. Control, vol. 22, pp. 135-145, 2015. [Online]. doi: http://dx.doi.org/10.1016/j.bspc.2015.07.002.

[26] K. Selvan, A. Joseph y K. K. Anish Babu, "Speaker Recognition System for Security Applications", *IEEE Recent Advances in Intelligent Computational Systems Speaker*, 2013, pp. 1-5. [Online]. Disponible en http://ieeexplore.ieee.org/document/6745441/.

[27] R. Achkar, M. El-halabi, E. Bassil, R. Fakhro y M. Khalil, "Voice Identity Finder Using the Back Propagation Algorithm of an Artificial Neural Network", *Procedia - Procedia Comput. Sci.*, vol. 95, pp. 245-252, 2016. [Online]. doi: http://dx.doi.org/10.1016/j.procs.2016.09.322.

[28] S. Adibi, "Telematics and Informatics A low overhead scaled equalized harmonic-based voice authentication system", *Telemat. Informatics*, vol. 31, n.º 1, pp. 137-152, 2014. [Online]. doi: http://dx.doi.org/10.1016/j.tele.2013.02.004.

[29] E. San Segundo, A. Tsanas y P. Gómez-vilda, "Euclidean Distances as measures of speaker similarity including identical twin pairs : A forensic investigation using source and fi lter voice characteristics", *Forensic Sci. Int.*, vol. 270, pp. 25-38, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.forsciint.2016.11.020.

[30] J. H. Ahnn, "Scalable Big Data Computing for the Personalization of Machine Learned Models and its Application to Automatic Speech Recognition Service", in *IEEE International Conference on Big Data (Big Data)*, 2014, pp. 1-8. [Online]. Disponible en http://ieeexplore.ieee.org/abstract/document/7004349/.

[31] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler y E. André, "The Social Signal Interpretation (SSI) Framework Multimodal Signal Processing and Recognition in Real-Time", *Proceedings of the 21st ACM International Conference on Multimedia*. pp. 21-25, 2013. [Online]. Disponible en http://dl.acm.org/citation.cfm?id=2502223.

[32] P. P. Dahake, K. Shaw y P. Malathi, "Speaker Dependent Speech Emotion Recognition using MFCC and Support Vector Machine", *International Conference on Automatic Control and Dynamic Optimization Techniques*, 2016, pp. 1080-1084. [Online]. Disponible en http://ieeexplore.ieee.org/document/7877753/.

[33] E. André and T. Vogt, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", *Language Resources and Evaluation Conference*, 2006. pp. 1-6. [Online]. Disponible en https://www.informatik.uni-augsburg.de/lehrstuehle/hcm/publications/2006-LREC/.

[34] R. Chakraborty, M. Pandharipande y S. Kopparapu, "Event Based Emotion Recognition for Realistic Non-Acted Speech", *TENCON 2015 - 2015 IEEE Reg. 10 Conf.*, 2015. pp. 1-5. [Online]. Disponible en http://ieeexplore.ieee.org/document/7372953/?reload=true&arnumber=7372953.

[35] E. D'Arca, N. M. Robertson y J. Hopgood, "Using the Voice Spectrum for Improved Tracking of People in a Joint Audio-Video Scheme", en *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3622-3626. [Online]. Disponible en http://ieeexplore.ieee.org/document/6638333/.

[36] S. H. Mohammadi and A. Kain, "An Overview of Voice Conversion Systems", *Speech Commun.*, vol. 88, pp. 65-82, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.specom.2017.01.008.

[37] A. V Savchenko and L. V Savchenko, "Towards the Creation of Reliable Voice Control System Based on a Fuzzy Approach", *Pattern Recognit. Lett.*, vol. 65, pp. 145-151, 2015. [Online]. doi: http://dx.doi.org/10.1016/j.patrec.2015.07.013.

[38] J. Gatti, C. Fonda, L. Tenze y E. Canessa, "Voice-Controlled Artificial Handspeak System", International Journal of Artificial Intelligence & Applications, vol. 5, n.º 1, pp. 107-112, 2014. [Online]. doi: https://doi.org/10.5121/ijaia.2014.5108

[39] C. G. Le Prell and O. H. Clavier, "Effects of noise on speech recognition : Challenges for communication by service members Hearing in Noise Test Speech Perception in Noise", *Hear. Res.*, vol. 349, pp. 76-89, 2016. [Online]. doi: http://dx.doi.org/10.1016/j.heares.2016.10.004.

[40] W. J. Hess, "Pitch and Voicing Determination of Speech with an Extension Toward Music Signals", *Springer Handbook of Speech Processing*, Springer, 2008, pp. 181-212. [Online]. Disponible en http://link.springer.com/10.1007/978-3-540-49127-9_10.

[41] K. Wu, D. Zhang y G. Lu, "iPEEH : Improving pitch estimation by enhancing harmonics", *Expert Syst. Appl.*, vol. 64, pp. 317-329, 2016. [Online]. doi: http://dx.doi.org/10.1016/j.eswa.2016.08.018

[42] L. N. Tan y A. Alwan, "Multi-Band Summary Correlogram-Based Pitch Detection for Noisy Speech", *Speech Commun.*, vol. 55, n.º 7-8, pp. 841-856, 2013. [Online]. doi: http://dx.doi.org/10.1016/j.specom.2013.03.001.

[43] R. Rajan y H. A. Murthy, "Two-Pitch Tracking in Co-Channel Speech Using Modified Group Delay Functions", *Speech Commun.*, vol. 89, pp. 37-46, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.specom.2017.02.004.

[44] J. Zeremdini, M. Anouar, B. Messaoud y A. Bouzid, "Multiple Comb Filters and Autocorrelation of the Multi-Scale Product for Multi-Pitch Estimation", *Appl. Acoust.*, vol. 120, pp. 45-53, 2017. [Online]. doi: http://dx.doi.org/10.1016/j.apacoust.2017.01.013

[45] G. Zhang y S. Godsill, "Tracking Pitch Period Using Particle Filters", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp.

1-4. [Online]. Disponible en http://ieeexplore.ieee.org/document/6701846/.

[46] J. ie Lin, G. Zhang, B. Fu y Y. Hao, "Multipitch Tracking With Continuous Correlation Feature and Hybrid DBNS / HMM Model", *11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing(ICCWAMTIP)*, 2014, pp. 218-221. [Online]. Disponible en http://ieeexplore.ieee.org/document/7073394/.

[47] G. Naganjaneyulu, M. V. Ramana y A. Narasimhadhan, "A Novel Method for Pitch Detection via Instantaneous Frequency Estimation using Polynomial Chirplet transform", *IEEE Region 10 Conference (TENCON)*, 2016, n.º 2, pp. 1250-1253. [Online]. Disponible en http://ieeexplore.ieee.org/document/7848211/.

[48] K. Kaminski, E. Majda y A. Dobrowolski, "Automatic speaker recognition using a unique personal feature vector and Gaussian Mixture Models", in *Signal Processing: Algorithms, Architectures, Arrangements y Applications (SPA)*, 2013, pp. 220-225. [Online]. Disponible en http://ieeexplore.ieee.org/document/6710629/.

[49] F. Huang, T. Lee, W. B. Kleijn y Y. Kong, "A Method of Speech Periodicity Enhancement Using Transform-Domain Signal Decomposition", *Speech Commun.*, vol. 67, pp. 102-112, 2015. [Online]. doi: http://dx.doi.org/10.1016/j.specom.2014.12.001.

[50] Z. Wang and J. DeLiang, "A Multipitch Tracking Algorithm for Noisy and Reverberant Speech", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4218-4221. [Online]. Disponible en http://ieeexplore.ieee.org/document/5495702/.

[51] K. Akant and S. Limaye, "Pitch contour extraction of singing voice in polyphonic recordings of Indian classical music", *International Conference on Electronic Systems, Signal Processing and Computing Technologies*, 2014, pp. 123-128. [Online]. Disponible en http://ieeexplore.ieee.org/document/6745358/.

[52] W. Yao-qi, W. Xiao-peng, L. Tao y L. Wei-wei, "Pitch Detection Method Based on Morphological Filtering and HHT", *J. China Railw. Soc.*, 2014. pp. 56-61. [Online]. Disponible en http://en.cnki.com.cn/Article_en/CJFDTOTAL-TDXB201407010.htm.