Sistema de identificación de cantidad de hablantes

Martín L. Córdoba, Emiliano E. Kalafatic, Leandro G. Panessidi
Ingeniería en informática – FICH – UNL

***Resumen*—En este articulo se presenta un enfoque alternativo para el desarrollo de un sistema de identificación de hablantes en una habitación. Para lograr dicho enfoque, se ha investigado sobre los diferentes métodos que existen para lograr dicha finalidad; desde el análisis de fonemas a través de la identificación del pulso glótico del hablante, como así también del uso de clusters a través de técnicas de Inteligencia Artificial. Cada uno de estos métodos presentan diferentes limitaciones, tales como la necesidad de que cada hablante pronuncie al menos alguna fonante, disponibilidad de una base de datos, cantidad de datos a utilizar y costos computacionales elevados. Se graba una conversación sin interrupciones ni solapamiento de voces en una habitación con poco ruido y a partir de allí se obtiene la señal de audio digital. Luego se identifican los tramos de la señal en donde se encuentre que una persona esté hablando mediante el análisis y medición de la amplitud a través de un umbral. Se analizan cada uno de estos tramos entre sí para determinar si corresponden a personas diferentes a través de la identificación y comparación de ciertos parámetros característicos. A partir de allí, se hace una discriminación por sexo y luego un análisis más fino para determinar la cantidad de hablantes por sexo y en total. Con las medidas de desempeño calculadas se determinó que el nivel de aciertos del sistema fue aproximadamente del 77% para el método de Kullback-Leibler, mientras que para el método del producto punto, fue del 55%**

***Palabras clave*— hablantes, amplitud, parámetros, género.**

I. Introducción

El reconocimiento del hablante es la identificación de personas a través del análisis de características propias de la voz. También suele ser denominado como reconocimiento de voz. Existe una diferencia entre *reconocimiento de voz* y *reconocimiento del habla* (donde se identifica *qué* se dice). Estos dos términos son usualmente confundidos y el término de *reconocimiento de voz* suele ser utilizado para ambos.

La técnica de reconocimiento de voz tiene una historia de más de 40 años y su principal herramienta de análisis es la identificación de propiedades acústicas del habla, en las cuales se ha descubierto que difieren persona a persona. Estas propiedades reflejan patrones de la anatomía humana (tamaño y forma del tracto vocal, boca, etc) como de los pulsos glóticos generados por las cuerdas vocales, donde a su vez estas proveen información tales como el pitch, y características del hablante en general.

En este proyecto, se pretende abarcar la problemática de identificar la cantidad de hablantes en una conversación grabada en una señal de audio.

Se han encontrado diferentes métodos para resolver problemáticas similares. En [1], [4] se estudia la relación entre el pulso glótico y la estructura del tracto vocal para determinar el sexo de los hablantes, y en [3] se hace énfasis en el pulso glótico como herramienta para la identificación de hablantes en general. En [2] se utiliza la información de las frecuencias formantes y a través de un clasificador y de métodos estadísticos se determina la cantidad de hablantes.

El enfoque a implementar descripto en este texto apunta a realizar un sistema que, a partir de la señal de audio de una conversación entre personas, determine la cantidad de hablantes en la misma a través de la identificación, análisis y comparación de parámetros característicos en los tramos de la señal donde se encuentre que una persona esté hablando. Para determinar cuando una persona habla, se analiza la amplitud de la señal a lo largo del tiempo, y a través de un umbral, se determina si se trata de ruido de fondo o de una persona hablando (ver Fig. 1). Una vez obtenidos cada uno de los tramos descriptos, se procede a calcular diferentes parámetros característicos por cada tramo.

Los parámetros característicos a analizar son los siguientes:

- Frecuencia Fundamental F0
- Densidad de frecuencia por bandas
- Coeficientes LPC

Con la F0 ya calculada, y utilizando la misma, se procede a separar a los hablantes según su sexo. Luego, se determina la cantidad de hablantes por sexo utilizando diferentes métodos, para así obtener la cantidad de hablantes en total.

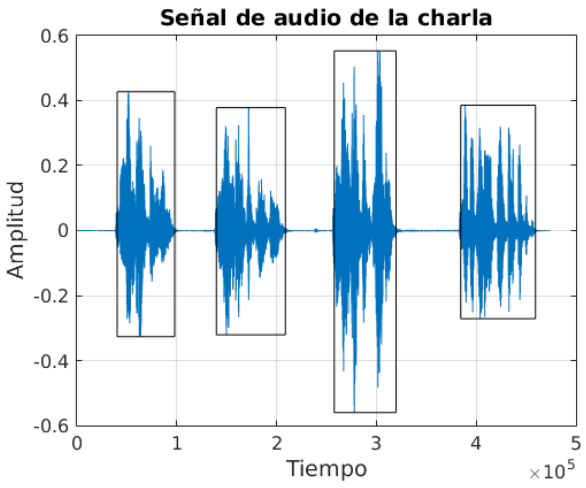

Fig. 1: Detección de charlas.

II. Algoritmo propuesto

Se procede a ventanear la señal utilizando la ventana cuadrada con una longitud de 0.03 segundos [5]. Se calcula

el pico de amplitud de todas la ventanas correspondientes a los primeros dos segundos, luego se lo multiplica por una tolerancia para definir así un umbral. A partir de esto, se evalúan los picos de las ventanas siguientes, y aquellos que superen el umbral se las consideraran como ventanas pertenecientes a una frase (o charla). Cada una de las frases obtenidas se guardaran en diferentes vectores (ver Fig. 2).

Por cada frase obtenida se identificarán y analizarán un conjunto de parámetros que determinan características propias de un hablante. Estos son:

- Frecuencia fundamental (F0)
- Densidad por banda de frecuencia
- Coeficientes LPC

cada uno de estos parámetros conformarán un vector característico definido en R^{16} por cada frase encontrada y tendrá la estructura siguiente:

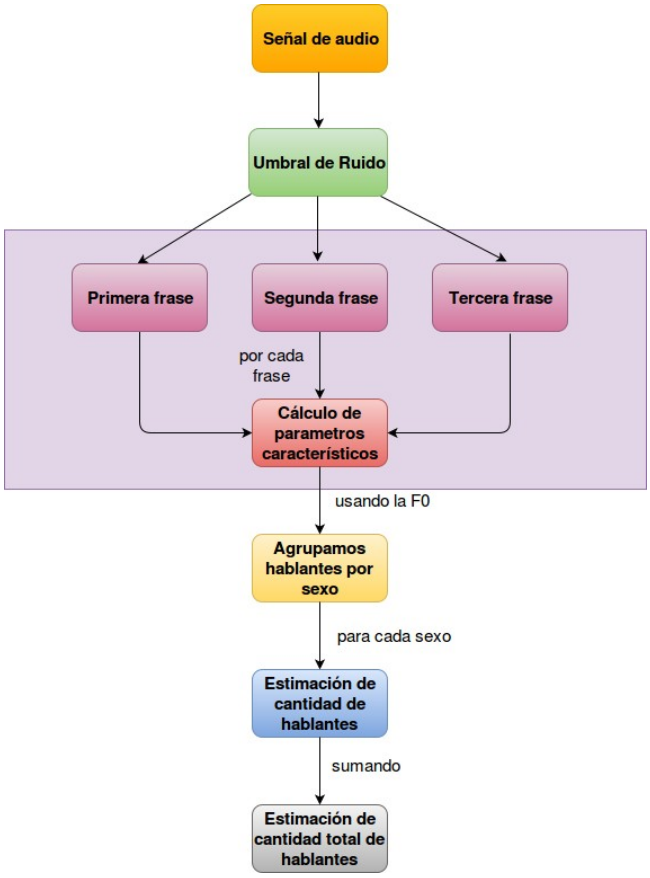

Fig. 2: Estructura de la implementación para una señal con tres frases.

F0	Bandas de frecuencia	Coeficientes LPC

donde F0 conformará un parámetro de entrada, las bandas de frecuencias conformarán siete y se definen en los rangos (Hz) definidos en la Tabla I:

TABLA I

BANDAS DE FRECUENCIA

min	100	250	500	750	1000	1500	2000
max	250	500	750	1000	1500	2000	3000

luego se tienen los coeficientes LPC que conformarán los últimos ocho parámetros. En la Fig. 3 se obtiene una representación gráfica de los vectores característicos para una conversación determinada.

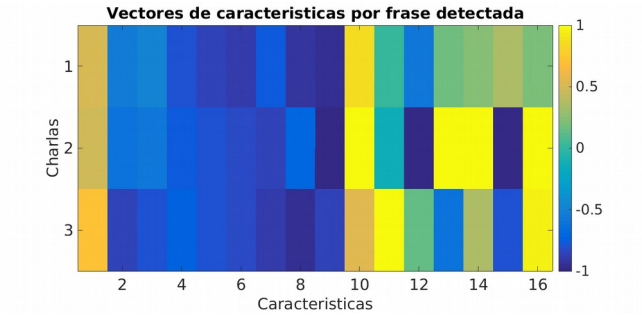

Fig. 3: Representación de vectores característicos.

A continuación se clasifican las frases de acuerdo al sexo del hablante, para esto se utiliza la F0. De acuerdo a [1] y a datos experimentales, se asume que frecuencias fundamentales mayores a 200Hz corresponden a mujeres, mientras que si la frecuencia es menor a este valor, la frase corresponde a un hombre. Con las frases ya clasificadas por sexo, se procede a realizar la identificación de la cantidad de hablantes en cada grupo.

Para determinar la cantidad de hablantes, utilizamos los vectores característicos y aplicamos el método de Divergencia de Kullback-Leibler [6], el cual se define como en (1).

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \qquad (1)$$

donde P está dado por los posibles hablantes y Q por el conjunto de hablantes ya identificados. Para la primer iteración se asigna el primer vector característico automáticamente como el primer elemento del conjunto Q. Luego se realiza una comparación entre P y Q de manera tal de que si ambos son parecidos bajo una tolerancia, P y Q se promedian, y se actualizan los datos de Q. De no ser así, P se agrega al conjunto Q. Como método alternativo se propone un enfoque similar, utilizando los vectores P y Q donde la comparación entre ellos se realiza a través del producto punto.

Finalmente, con la cantidad de hablantes por sexo ya identificada se procede a sumar ambas cantidades para obtener así la cantidad de hablantes en total (ver Fig. 4).

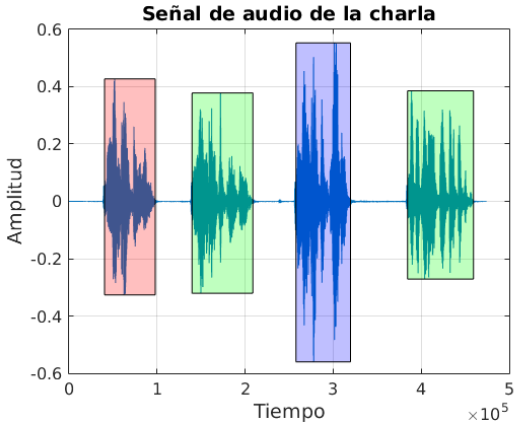

Fig. 4: Identificación de hablantes por color.

III. Generación de los datos

Para darle solución al problema planteado, se dispone de un archivo correspondiente a la señal digital de audio de una conversación en una habitación. Las características de la conversación son las siguientes:

- Ruido de fondo despreciable
- No hay interrupciones
- Volumen constante
- Se empieza a hablar a partir de los dos segundos

La señal fue muestreada con una frecuencia de muestreo de 44100 Hz.

IV. Resultados

A. Porcentaje de acierto

Se utilizaron nueve audios diferentes para evaluar el porcentaje de acierto de los dos métodos propuestos. Los audios varían en cantidad de hablantes como así también en la cantidad de sexos. Los resultados obtenidos se muestran en la Tabla II:

TABLA II

Porcentaje de aciertos

Producto Punto	**55%**
Kullback-Leibler	**77%**

se puede apreciar que el método mas efectivo es el originalmente propuesto debido a que trabaja con distribuciones de probabilidades.

B. Cantidad de hablantes vs. Resultados obtenidos

En la Fig. 5 se comparan los resultados obtenidos por el método de Kullback-Leibler de cada audio con la cantidad de hablantes reales de cada uno.

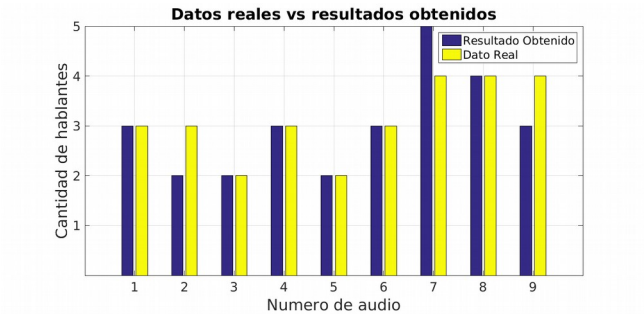

Fig. 5: Cantidad de hablantes y resultados.

como se puede ver, se corroboran los datos de la Tabla II. Se pudo apreciar que en gran cantidad de los audios donde el programa arrojó un resultado erróneo, fue debido a que los audios no cumplían con las características previamente definidas.

C. Medición del error en función del ruido de la señal

Para analizar la robustez del programa, se procede a utilizar señales de entrada artificialmente contaminadas con ruido blanco y de esta manera evaluar su desempeño. Se generan señales cuya componente de ruido se incrementa gradualmente, y para cada caso se calcula el error absoluto del resultado obtenido respecto del dato original. Este proceso se repite para diferentes audios con las que el programa funciona correctamente, realizando un promedio truncado de los errores calculados por cada nivel de ruido. La señal se contamina de forma proporcional a la relación señal-ruido (SNR dB) de la misma (ver Fig. 6). A

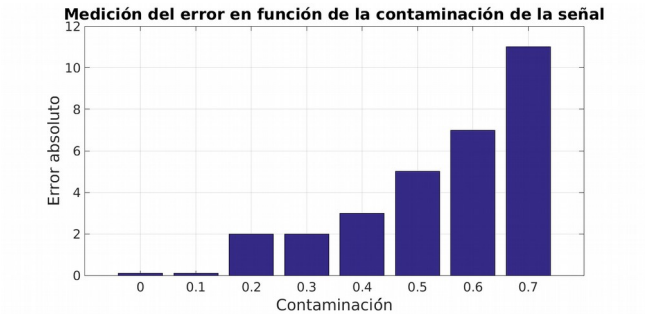

Fig. 6: Medición de error de señales contaminadas por ruido.

partir de valores de contaminación mayores a 0.7 no es posible la identificación de hablantes ya la excesiva contaminación no permite la identificación de frases.

V. Conclusiones

Se desarrolló un sistema para detectar la cantidad de hablantes en una conversación a través del análisis de la señal de audio. Para esto, se detectaron los tramos de la señal correspondiente a las frases, y por cada una de estas, se evaluaron parámetros característicos del habla para determinar la cantidad de hablantes por sexo, y finalmente obtener la cantidad total. Con las medidas de desempeño calculadas se determinó que el nivel de aciertos del sistema fue aproximadamente del 77% para el método de Kullback-Leibler, mientras que para el método del producto punto, fue del 55%. Para mejorar este resultado, se podría aplicar un filtro de manera tal de reducir considerablemente el ruido, además se podría aumentar la cantidad de parámetros característicos, como así también encontrar una relación entre la tolerancia y el ruido. Adicionalmente se podría mejorar el equipamiento utilizado para realizar las grabaciones.

Agradecimientos

Matías F. Gerard, por su ayuda y seguimiento durante el desarrollo del trabajo.

Leandro E. Di Persia, por su ayuda y sugerencias.

Referencias

[1] *David R. R. Smith, Thomas C. Walters, and Roy D. Patterson,"Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled", 2007.*

[2] *Oliver Baumann Æ Pascal Belin,"Perceptual scaling of voice identity: common dimensions for different vowels and speakers", 2008.*

[3] *Elizabeth Vera de Payer,"El pulso glótico como fuente de I nformación para identificación del hablante", 2001.*

[4] *Jessica Junger, Katharina Pauly, Sabine Bröhr,"Sex matters: Neural correlates of voice gender perception", 2013.*

[5] *D. H. Milone, H. L. Rufiner, R. C. Acevedo, L.E. Di Persia, H.M. Torres, "Introducción a las señales y los sistemas digitales", Aprobado por Consejo Editorial UNER.*

[6] "Kullback–Leibler divergence", en Wikipedia, 12 de junio de 2018, de *https://en.wikipedia.org/wiki/Kullback-Leibler_divergence*