DETECCIÓN DE LOS PRIMEROS FORMANTES DE LA VOZ POR MEDIO DEL ANÁLISIS DE LA SEÑAL EN TIEMPO Y FRECUENCIA

Jairo Alejandro Rodríguez Martínez

Resumen

Se aplicarán las técnicas de representaciones en tiempo y frecuencia para observar algunas características de la voz humana. Compararemos las ventajas y desventajas de esas representaciones. Veremos cómo con estas técnicas se puede observar la rápida variación de las estructuras de los armónicos y de los formantes.

Palabras claves

Formantes, análisis tiempo -frecuencia

Abstract

The techniques to represent time and frequency will be applied in order to observe some characteristics about the human voice. We will compare the advantages and disadvantages of such representations. Also, it could be observed how applying these techniques one can find out the quick variations in the harmonic and formant structures.

Index Terms

Formant, analysis time-frequency

El habla es uno de los más útiles y complejos medios de comunicación. Podemos considerar que la señal de voz se trasmite a través de un canal por medio de ondas de presión. La estructura de dicho canal es intrincada; y como consecuencia de ello surgen elementos aleatorios en la señal de voz generada.

Los seres humanos poseemos "transductores" que producen y captan esta señal: el aparato fonador y auditivo respectivamente.

El aparato fonador lo constituyen tres elementos: un generador de energía, un sistema vibrante y una cavidad resonante[1].

La principal función vocal de los pulmones (*los generadores de energía*) es la de producir una diferencia de presión generando así un chorro de aire. El flujo de aire atraviesa la glotis, que es un espacio situado en la base de la laringe limitado por las cuerdas vocales (*el sistema vibrante*). La corriente de aire hace vibrar las cuerdas vocales con lo cual se producen en la laringe variaciones periódicas de la presión. La laringe, que es de forma tubular, se encuentra acoplada a una cavidad más amplia (*la faringe*) que une la boca con el esófago. El techo de la faringe lo forma el paladar blando cuya misión es franquear la entrada a la cavidad nasal; es decir, cuando se pronuncian las vocales se alza, cerrando así el paso de aire hacia la nariz[2].

La laringe, la faringe y la boca constituyen el conducto vocal (*la cavidad resonante*) que se comporta como una cámara de resonancia similar al tubo de un fagot o al cuerpo de una guitarra. La forma de la cavidad la determinan los articuladores: la laringe, la mandíbula, la lengua y los labios. Por ejemplo; al bajar la laringe crece la longitud del conducto, mientras que el movimiento de los otros articuladores contraen o dilatan el conducto en ciertos puntos.

Consideraremos ahora con más detalle cómo es que se genera la voz.

El chorro de aire impelido por los pulmones produce una diferencia de presión por debajo de la glotis, induciendo a que las cuerdas vocales se separen; el aire fluyendo a través de los ligamentos genera *el efecto Bernoulli,* el cual, en combinación con las propiedades mecánicas de las cuerdas, cierra casi inmediatamente la glotis. La diferencia de presión se incrementa obligando a que las cuerdas se separen nuevamente[3].

Figura. 1. La diferencia de presión mueve la cuña de tal manera que cierra el conducto derecho, y el aire fluirá con mayor rapidez por la abertura; sin embargo, ahora, debido a la diferencia de velocidades, en el conducto aumentará la presión y forzará a la cuña a moverse hacia la izquierda.

Se puede suponer la laringe como un tubo dividido en dos recámaras separadas por una cuña triangular que vendría a ser la glotis (figura 1). La abertura lateral constituye la separación entre las cuerdas vocales, mientras que la de la derecha es la región que desemboca en la faringe, la cual se halla protegida por la epiglotis durante la ingestión.

El ciclo de movimiento oscilatorio de la cuña genera un tren de pulsos de aire que pasan al conducto vocal. La frecuencia de la vibración está determinada por la presión del aire en los pulmones y por las propiedades mecánicas de las cuerdas vocales. Por lo general, cuanto mayor es la presión del pulmón y cuanto más delgados sean y más tensos estén los ligamentos vocales, mayor será la frecuencia a la que vibren estos y emitan

[1] F. Casacubierta, E. Vidal. *Reconocimiento automático del habla.* Editorial Marcombo.

[2] J. Sundberg. Investigación y ciencia. Enero de 1998.

[3] D. Jou, J. E. Llebot, C. P. García. *Física para las ciencias de la vida.* Mc Graw Hill.

pulsos de aire. El tren de pulsos produce una presión de aire con una oscilación muy rápida en el conducto vocal, es decir, produce un sonido. Este sonido generado por el chorro de aire entrecortado por la vibración de las cuerdas vocales se denomina fuente vocal (figura 2), y constituye la materia prima del habla. Es un sonido complejo con una frecuencia fundamental (*el pitch*) determinada por el ritmo vibratorio de los ligamentos vocales, y un gran número armónicos o sobretonos.

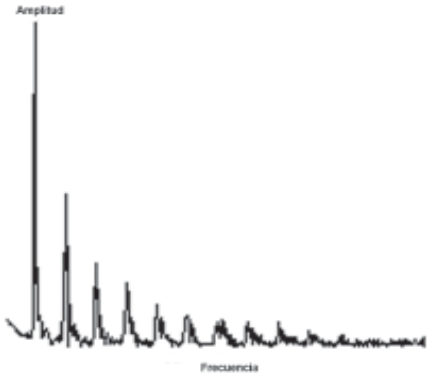


Figura 2. El chorro de aire procedente de los pulmones se interrumpe periódicamente por la vibración de los ligamentos vocales. El sonido resultante, la fuente de voz, tiene un espectro compuesto de una gran cantidad de armónicos cuya amplitud decrece uniformente con la frecuencia.

El conducto vocal es un resonador y la transmisión de un sonido a través de un resonador acústico es función de la frecuencia; así, los sonidos correspondientes a la frecuencia de resonancia propia de cada resonador son menos atenuados que otros sonidos, y, por tanto, son emitidos con una mayor amplitud relativa que otros sonidos. El conducto posee cuatro o cinco resonancias importantes llamadas formantes (figura 3), y todos los armónicos de la fuente vocal pasan por el conducto vocal con más o menos éxito, según su frecuencia; cuanto más cerca esté un armónico de la frecuencia de un formante, tanto más aumentará su amplitud en los labios. La presencia de los formantes interrumpe la envolvente de pendiente uniforme de la fuente sonora, creando picos en las frecuencias del formante. Esta perturbación de la envolvente de la fuente sonora es la responsable de que se produzcan sonidos inteligibles al hablar; es decir, ciertas frecuencias de los formantes se manifiestan en el espectro emitido como picos en la envolvente (figura 4), los cuales son característicos de ciertos sonidos.

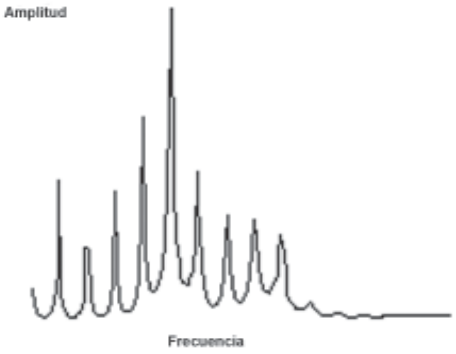


Figura 3. La columna de aire que discurre por el conducto vocal posee unos modos de vibración muy característicos llamados formantes.

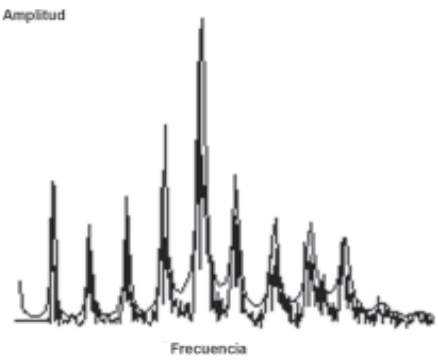


Figura. 4. Como la fuente vocal pasa a través del conducto vocal, cada armónico se atenúa en proporción a su distancia del formante más cercano a su frecuencia. Obsérvese como las frecuencias de la fuente vocal son moduladas por la curva de los formantes, lo cual constituye el espectro emitido.

Las frecuencias de los formantes dependen de la forma del conducto.

Vamos a considerar dicho conducto como un tubo acústico de longitud *L* cerrado por un extremo (la glotis) y abierto por el otro (los labios).

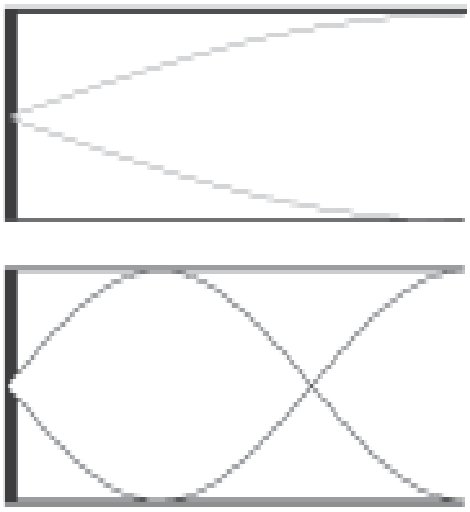
La ecuación que gobierna las vibraciones de una columna de aire dentro de un resonador en forma de cilindro esta dada por:

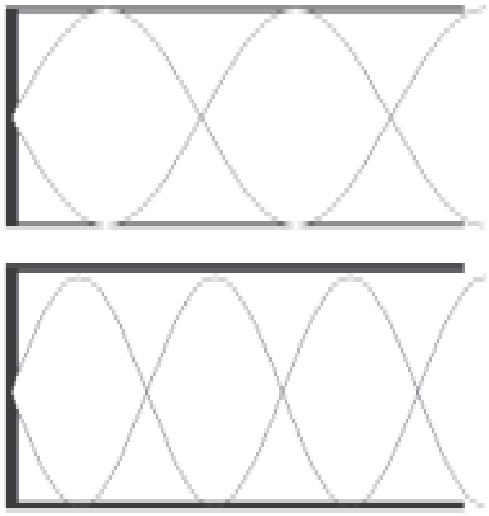
$$\frac{\partial^2 \xi}{\partial x} = \frac{1}{v^2}\frac{\partial^2 \xi}{\partial t^2} \qquad (1)$$

Donde es el desplazamiento a partir de la posición de equilibrio, *V* la velocidad del sonido en el medio. Un extremo abierto representa (aproximadamente, para cualquier frecuencia) una condición de variación de presión nula durante la oscilación y un lugar de máximo movimiento del aire[4]. La solución de la ecuación (1) conduce a las frecuencias naturales del tubo; es decir:

$$f_n = \frac{2n-1}{4L}V,\ n = 1,2,3... \qquad (2)$$

Suponiendo que la longitud del conducto es de 17.5 centímetros, lo cual resulta verosímil en el caso de un varón adulto, entonces los primeros cinco formantes tendrán frecuencias cercanas a 500, 1500, 2500, 3500 y 4500 hertzios. Si el conducto vocal fuera más largo, o más corto, estas frecuencias serían algo inferiores, o algo superiores.





Las figuras ilustran los primeros cuatro modos de vibración de un tubo sonoro; sin embargo, es pertinente advertir que tratándose el sonido de una onda longitudinal, esos diagramas representan vibraciones transversales las cuales son más sencillas de visualizar. Las formas de onda de esos modos se ajustan a las condiciones de frontera impuestas a la ecuación (1), tanto en la glotis como en los labios, lo que conduce a la solución:

$$\xi(x,t) = \xi_0 sen\left(\frac{2n-1}{2L}Vx\right)\cos(2\pi f_n t) \qquad (3)$$

Ahora bien; cada formante está asociado con una onda estacionaria o modo de vibración. El más grave, por ejemplo, corresponde a un cuarto de longitud de onda, lo que equivale a decir que un cuarto de longitud de onda cabe dentro del conducto vocal.

Cualquier variación en la sección transversal del conducto vocal modifica las distintas frecuencias de los formantes. Una contracción del conducto vocal es un punto en donde la onda estacionaria de un formante presenta oscilaciones de presión cuya amplitud mínima determina, generalmente, la reducción de la frecuencia del formante; por el contrario, una dilatación del conducto en estos mismos puntos incrementará la frecuencia.

[4] A. P. French. *Vibraciones y ondas.* MIT.

El conducto vocal se contrae y se dilata de muchas formas, y su contracción en un punto incide en la frecuencia de todos los formantes de diferentes maneras. Hay sin embargo, tres elementos muy importantes para cambiar la forma del conducto, de suerte que la frecuencia de un formante particular puede variar en un cierto ancho de banda. Estos elementos son la mandíbula, el cuerpo y la punta de la lengua. La apertura de la mandíbula puede reducir el conducto en la zona adyacente a la glotis y ampliarlo en la región de los labios; tal apertura resulta decisiva para la frecuencia del primer formante, que aumenta a medida que se va abriendo la mandíbula. Por otra parte, la frecuencia del segundo formante es muy sensible a la posición de la lengua.

Justamente en problemas tales como los que se han descrito, es que el análisis en tiempo y frecuencia posee enorme utilidad.

El objetivo básico de éste tipo de análisis es desarrollar una función que pueda describir simultáneamente en tiempo y frecuencia la densidad de energía que posee una señal. Una de estas funciones es la *transformada de Fourier de tiempo corto* o también llamada *transformada enventanada de Fourier* (short time Fourier transform), que constituye el método más ampliamente usado para el estudio de señales no estacionarias. El concepto que subyace es simple y poderoso. Supongamos que nos trasladamos con rapidez constante en un vehículo que posee ventanas de distinto tamaño y que observamos el paisaje a través de una de las ventanas. La observación por una ventana grande nos permitirá ver una imagen panorámica del paisaje, pero perderemos la misma al hacerlo por la ventana pequeña. Sin embargo, la apertura menor nos permitirá detectar detalles locales, mientras que la otra nos mostrará detalles globales. El paisaje cambiante formado por casas, árboles y montañas desfilará ante nuestros ojos cuando miramos por la ventana mayor, pero lo mismo no ocurrirá cuando lo hagamos por la otra, ya que quizá únicamente percibamos las casas pero no el entorno adyacente a las mismas. Seguramente nos parecerá monótono el ver discurrir las casas cuando elegimos la ventana pequeña, en contraste con el mosaico variopinto que nos permitiría ver la ventana grande. Se podría pensar que, salvo algunos detalles menores como el color, la forma y el tamaño, la imagen de las casas se puede considerar estacionaria y la imagen de las casas más su entorno como no estacionaria.

De manera semejante, para el estudio de las propiedades de una señal no estacionaria en un cierto tiempo *t*, uno enfatiza la señal en ese instante y la suprime para otros. Esto se logra multiplicando la señal por una función ventana *h(t),* centrada en *t*, para producir una versión ahora enventanada de la señal,

$$s_t(\tau) = s(\tau)h(\tau - t) \qquad \textbf{(4)}$$

La señal modificada de esta manera es función de dos tiempos; el tiempo fijo *t* de nuestro interés, y el tiempo *t* que transcurre. La función ventana es elegida para dejar más o menos inalterada la señal alrededor del tiempo *t*, pero para suprimir la señal en instantes lejanos del tiempo de interés. Esto es:

$$s_t(\tau) = \begin{cases} s(\tau) \ \ para\ \tau\ cercanos\ a\ t \\ 0 \quad\ para\ \tau\ lejanos\ de\ t \end{cases} \qquad \textbf{(5)}$$

Como la señal enventanada enfatiza la señal alrededor del tiempo *t*, la transformada de Fourier puede reflejar la distribución de frecuencias alrededor del tiempo *t*:

$$S_t(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} s_t(\tau) e^{-jw\tau} d\tau \qquad \textbf{(6)}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} s\ (\tau) h(\tau - t) e^{-jw\tau} d\tau \qquad \textbf{(7)}$$

la densidad espectral de energía en el instante *t* es por lo tanto:

$$P_{SD}(t,w) = \left|S_t(w)\right|^2$$

$$= \left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} s(\tau)h(\tau - t)e^{-jw\tau} d\tau \right|^2 \quad \textbf{(8)}$$

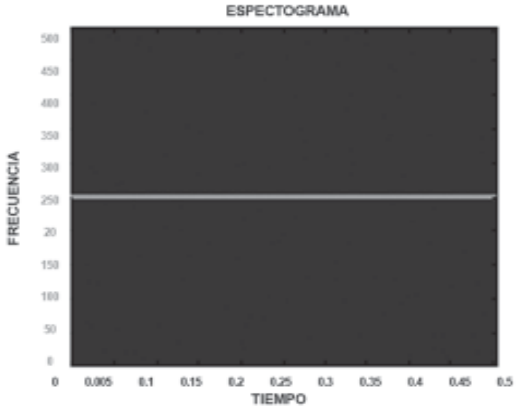


Figura 5. Espectrograma de una onda senoidal de 250Hz. Es un típico ejemplo de señal estacionaria.

para cada instante de tiempo *t* tendremos un espectro distinto y la totalidad de esos espectros es la distribución en tiempo-frecuencia, $P_{SD,}$ también conocida con el nombre de *espectrograma* [5].

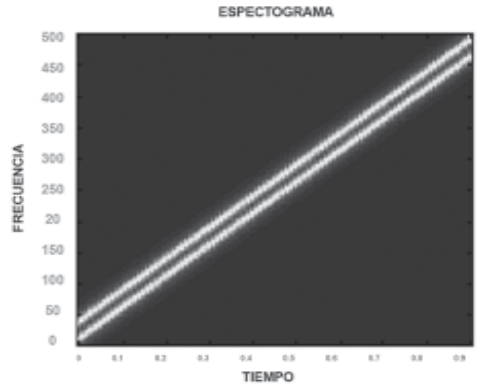


Figura 6. Espectrograma de una señal no estacionaria llamada "chirp" lineal. Nótese cómo la frecuencia varía uniformemente con el tiempo.

[5] L. Cohen. *Time frequency analisis.* Prentice Hall.

Un aspecto de interés central en el uso de la transformada de Fourier de tiempo corto *STFT* es la resolución de tiempo-frecuencia. Supongamos dos ondas seno con frecuencias semejantes y espaciadas Δw.

El valor más pequeño de Δw para el cual las dos señales son distinguibles se llama *resolución en frecuencia*. La duración correspondiente de la ventana *h(t)* se conoce como *resolución de tiempo* y se denota por Δt. La resolución en frecuencia y la resolución en tiempo se relacionan en forma inversa, como lo indica:

$$\Delta w \Delta t \geq 1/2 \qquad (9)$$

que refleja una propiedad de dualidad de la STFT, heredada de la transformada de Fourier.

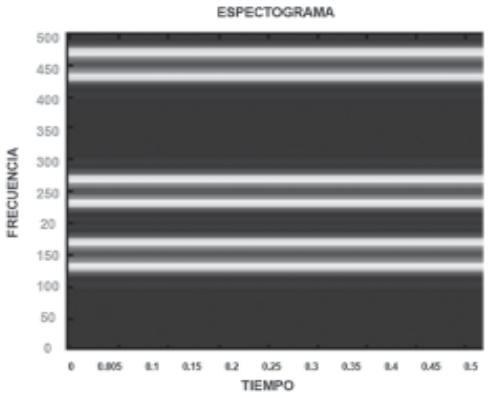


Figura 7. Tres sinusoides analizadas mediante una ventana estrecha. Se puede observar la poca resolución en frecuencia.

Así, si la ventana de análisis es pequeña (figura 7), se pierde resolución en frecuencia y se aumenta en tiempo; pero si la ventana de análisis es grande (figura 8), se gana resolución en frecuencia y se disminuye en tiempo.

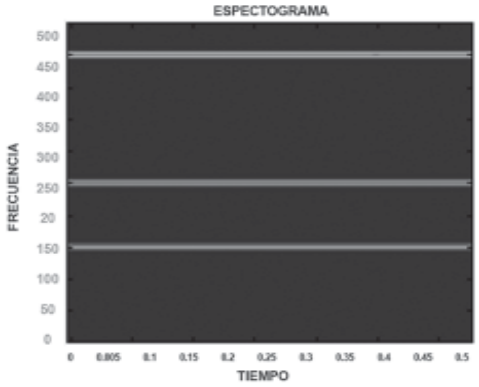


Figura 8. Las mismas sinusoides pero analizadas con una ventana más amplia. En contraste con la ventana pequeña (Figura 7), ahora son claramente distinguibles las frecuencias de las ondas.

Como ya se mencionó, la variación del tracto vocal trae consigo que la voz sea una señal no estacionaria; por tanto, las distribuciones tiempo-frecuencia son una excelente herramienta para estudiarla.

Consideraremos en primer lugar el uso de espectrogramas de banda estrecha y de banda ancha de la vocal "a" pronunciada por un varón de cuarenta años de edad.

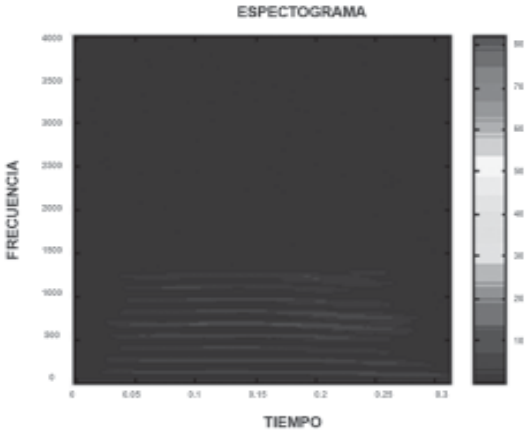


Figura 9. Se ha añadido una barra de color para poner de manifiesto la energía de los distintos armónicos. Compárese con lo exhibido en la gráfica 2. Se ha elegido una ventana Hamming de 1000 muestras para una señal de 3500 muestras.

En la figura de arriba se muestra el espectrograma de la vocal "a". Se nota la ausencia de frecuencias por arriba de 1000 hertzios debido al efecto de la glotis. Para suprimir ese efecto, la señal de voz se filtra por medio de un filtro FIR, denominado de preénfasis[6].

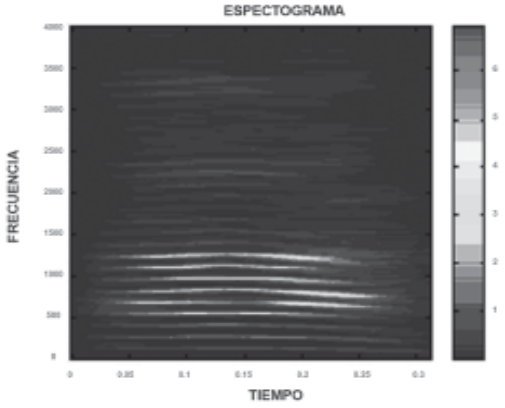


Figura 10. El efecto del filtrado tiene como consecuencia la aparición de frecuencias por encima de los 1000 hertzios. Ahora, claramente son distinguibles los formantes por la energía que contienen.

Luego del preénfasis se pueden apreciar nuevas bandas de frecuencias, así como la detección de los picos de energía para algunos formantes tales como el ubicado cerca de los 700 hertzios, y otro entre 1100 y 1150 hertzios.

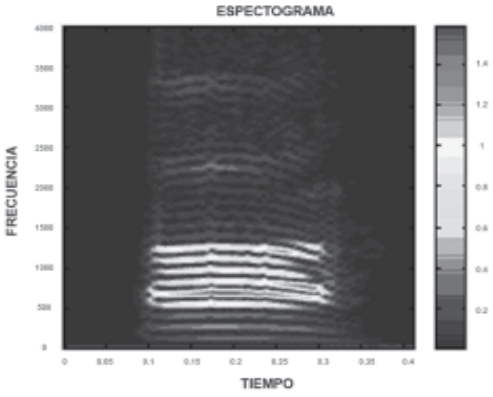


Figura 11. Hay una ostensible mejora en la resolución temporal, y se hacen discernibles dos formantes adicionales cercanos a los 2200 hertzios y 3300 hertzios respectivamente. La ventana de análisis se eligió con longitud de 200 muestras.

[6] C. Sidney Burrus, Alan V. Oppenheim. *Tratamiento de la señal*. Prentice Hall.

Las figuras 10 y 11 ponen de manifiesto la diferencia entre espectrogramas de banda ancha y de banda angosta. Por lo general los primeros sirven para medir las frecuencias de los distintos armónicos y los segundos para detectar los formantes.

Se ha recalcado a lo largo de este artículo el efecto que tienen las articulaciones en la ubicación frecuencial de los formantes.

El siguiente experimento pretende medir esos efectos. Para tal propósito, se le pide a un locutor que pronuncie la vocal "a" y que simultáneamente de apertura a la boca en forma gradual.

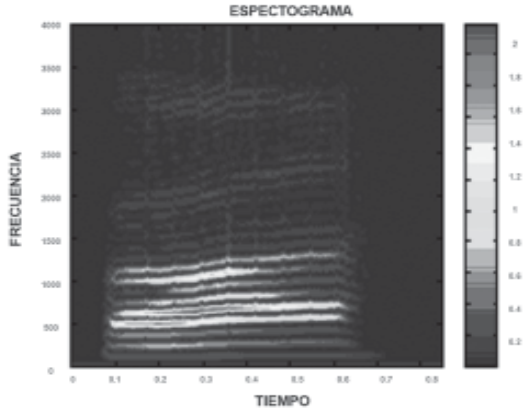


Figura 12. Se distingue con mucho detalle la evolución frecuencial del primer formante, el cual se inicia a los 130ms con una frecuencia 625 hertzios, deteniéndose en 725 hertzios al cabo de 600ms.

En la figura 12 se pone de manifiesto la proporcionalidad directa entre la frecuencia y el tiempo, tanto de los armónicos de la fuente de voz como de los formantes del tracto vocal, en consonancia con la apertura gradual de la boca. Esa linealidad se observa incluso hasta el tercer formante, que aumenta desde 1900 hasta 2350 hertzios.

Para concluir este análisis preliminar, se incluirá un experimento en el cual el locutor curva la lengua hacia arriba al pronunciar la misma vocal.

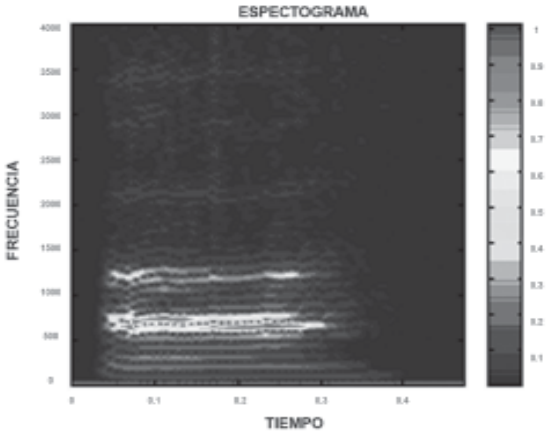


Figura 13. Si se compara este resultado con la gráfica anterior, se notará que el primer formante no cambia demasiado su estructura; pero en cambio el segundo formante se torna más débil. Ello pone de manifiesto la gran sensibilidad del segundo formante con respecto a la postura de la lengua.

La caracterización del habla ha sido posible gracias al desarrollo de los analizadores del habla y los sintetizadores de voz. El análisis en tiempo y frecuencia permite extraer las características acústicas contenidas en la señal y presentarlas en el plano tiempo frecuencia. De estas representaciones se pueden extraer las propiedades genéricas que permiten una caracterización de la señal: dimensiones físicas, frecuencia, energía y duración. Desde este punto de vista, las vocales aisladas se pueden caracterizar mediante la localización de los dos primeros formantes, ya que el tercero permanece relativamente estable para todas.

Otras representaciones tiempo-frecuencia son las denominadas *distribuciones bilineales*[7], las cuales se caracterizan por no ser afectadas por el principio de incertidumbre tiempo y frecuencia.

La distribución pseudo suavizada de Wiener-Ville pertenece a este grupo, que será la base de nuestros siguientes experimentos, destinados a localizar el primer formante de las vocales españolas.

Una distribución que permite estudiar simultáneamente energía, frecuencia y duración de una señal es cono-

[7] A. P. French. *Vibraciones y ondas.* MIT.

cida con el nombre de *distribución de Wiener-Ville,* la cual se define de la siguiente manera:

$$W_x(t,f)=\int_{-\infty}^{+\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j2\pi f\tau}d\tau \qquad \textbf{(10)}$$

Para ilustrar el uso de esta expresión consideraremos el caso en el cual un vehículo pasa frente a un observador. La señal del motor percibida por la persona cambia a medida que transcurre el tiempo; es decir, la frecuencia principal respecto del observador disminuye. Este fenómeno es conocido con el nombre de *efecto doppler*, el cual expresa la dependencia de la frecuencia del sonido escuchada por un oyente, debido al movimiento relativo entre la fuente y el observador[8].

La eficacia de las técnicas estudiadas en el presente artículo queda evidenciada al poder ver cómo evoluciona en el tiempo, y cual es la energía de la frecuencia instantánea escuchada por un oyente, que se encuentra inmerso en el campo acústico de una fuente que se desplaza respecto de ella.

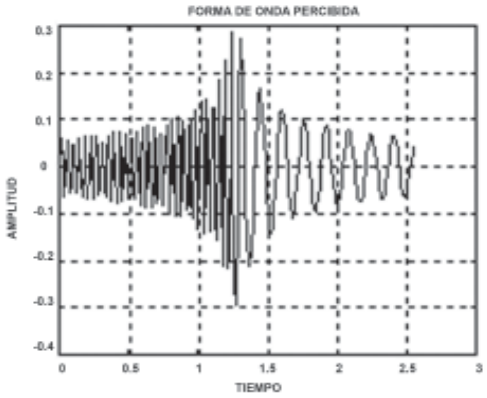


Figura 14. Señal escuchada por un receptor en reposo. La frecuencia de la fuente mide 15Hz y se desplaza con una rapidez de 200m/s.

Se puede observar en la figura la variación del periodo de la onda respecto del detector, así como también el cambio de amplitud. Los tres parámetros se pueden distinguir fácilmente en el plano tiempo-frecuencia.

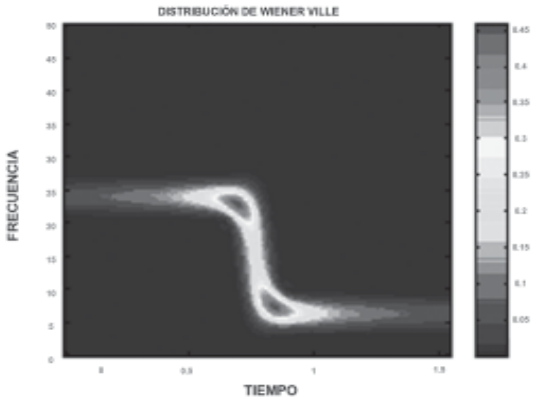


Figura 15. Se ha elegido una frecuencia de muestreo de 100Hz, y la fuente pasando a 10 metros del observador.

La frecuencia inicial percibida es de 23.82 hertzios. Posteriormente, la frecuencia cae drásticamente a 6.17 hertzios cuando la fuente pasa por el frente del observador. También se puede extraer información concerniente a la energía de la onda; por ejemplo, los instantes para los cuales ésta es mayor.

Regresando a nuestro problema acerca de la detección de los formantes, es el momento de mencionar la dificultad para distinguirlos de entre los armónicos de la fuente vocal (ver figuras 10, 11, 12 y 13).

Esas dificultades pueden ser resueltas a través de la distribución de Wiener-Ville.

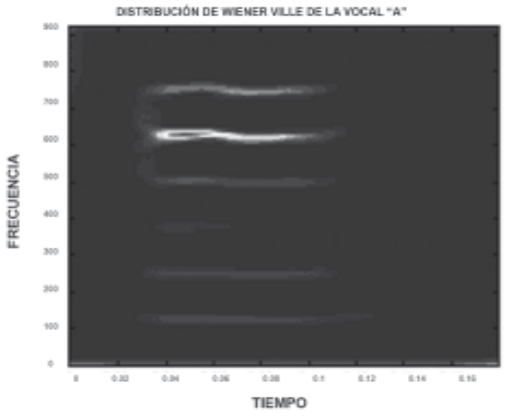


Figura 16. Distribución de Wiener-Ville de la vocal *a.* El formante es claramente distinguible alrededor de los 600Hz.

8 Matlab 5.3

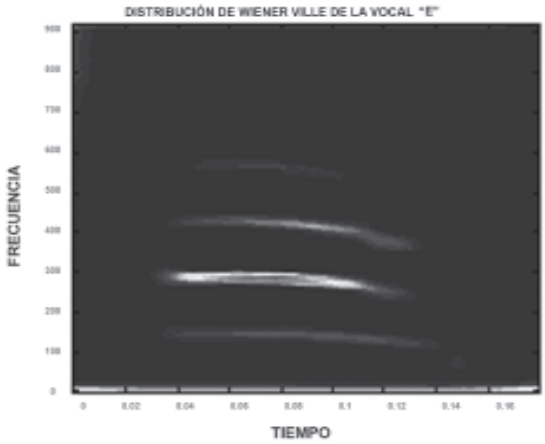


Figura 17. Distribución de Wiener-Ville de la vocal *e*. El formante es claramente distinguible alrededor de los 600Hz.

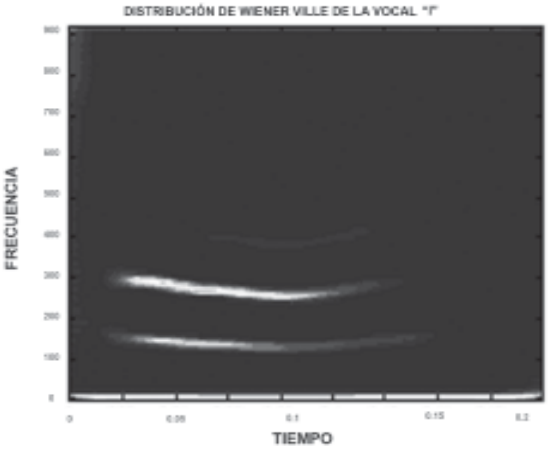


Figura 18. Distribución de Wiener-Ville de la vocal *i*. El formante es claramente distinguible un tanto por debajo de los 300Hz.

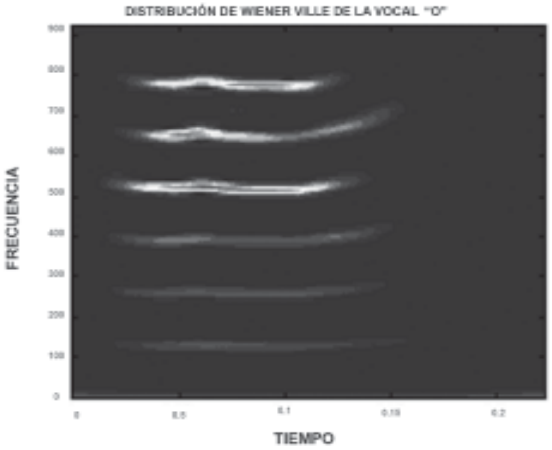


Figura 19. Distribución de Wiener-Ville de la vocal *o*. Ahora la energía de la señal codificada permite detectar el formante alrededor de 500Hz.

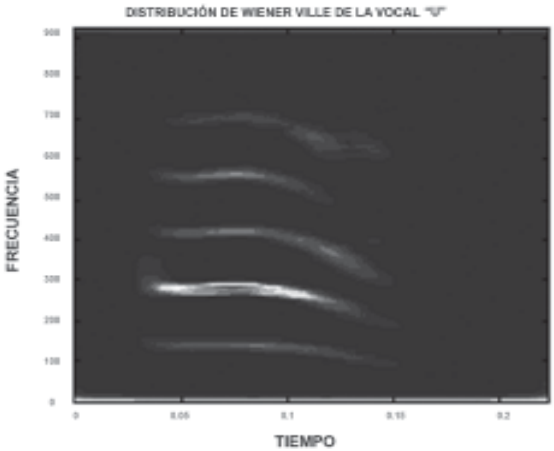


Figura 20. Distribución de Wiener-Ville de la vocal *u*. Se observa que por debajo de los 300Hz está situado el primer formante de esta vocal.

Las cinco gráficas precedentes muestran la utilidad de este tipo de análisis de una señal, pues permite la detección de los primeros formantes de las vocales de nuestro idioma. En algunos casos es la energía la que hace resaltar el formante de entre los armónicos (figura 18) y en otros es simplemente que se logra diferenciar con facilidad de los mismos.