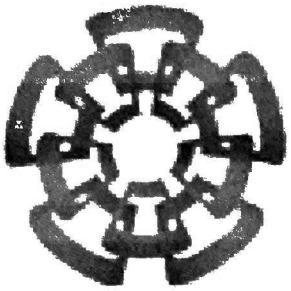
xx C10 1585.1)



CINVESTAV

Centro de Investigación y de Estudios Avanzados del IPN
Unidad Guadalajara

DESARROLLO DE UN CALCULADOR DE SONORIDAD

TESIS QUE PRESENTA
NOEL TRUJILLO MORALES

PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS

EN LA ESPECIALIDAD DE
INGENIERÍA ELÉCTRICA

Guadalajara, Jal. Noviembre de 2001

DESARROLLO DE UN CALCULADOR DE SONORIDAD

Tesis de Maestría en Ciencias
Ingeniería Eléctrica

por:

Noel Trujillo Morales

Ingeniero en Electrónica
Universidad Autónoma de Guadalajara, 1995-1999

Becario del CONACYT, expediente no. 143910

Directores de Tesis:

Dr. Deni Librado Torres Roman
Profesor. Michel Naranjo

CINVESTAV del IPN Unidad Guadalajara, Noviembre de 2001

RECONOCIMIENTOS

A Frederick y Renaud de la Universidad Blaise Pascal de Clermont-Ferrant, Francia, por su colaboración en el desarrollo de la herramienta calculador de sonoridad.

A María Andueza y Agurtzane Echegaray de la Universidad de Bilbao-España, por su colaboración en el tema de morfología matemática.

Al Dr. Eduardo Bayro-Corrochano y Juan Francisco Alvarado Casas por su colaboración en el análisis cuaterniónico de las imágenes de la sonoridad.

AGRADECIMIENTOS

A Dios por darme el tiempo y la salud para poder realizar éste trabajo. A mis padres por darme el apoyo, aliento y sus sabios consejos para continuar con mis estudios. A mis cuatro pilares: Claudia, Olivia, Melina y Mayra por estar ahí siempre, las quiero mucho y éste trabajo va dedicado para ustedes. Al Profesor Michel Naranjo por su paciencia y dedicación en la realización de éste trabajo. A mis grandes e inseparables amigos, que ya sea de cerca o a cientos de kilómetros de distancia estuvieron conmigo para apoyarme y escucharme en los momentos buenos y malos. También quisiera agradecer a aquellas personas que, en su tiempo, estuvieron conmigo y me dieron ánimos, consejos y fuerzas para terminar éste trabajo. Gracias.

CONTENIDO

1 INTRODUCCIÓN

En la actualidad el modelo del oído esta siendo muy utilizado en diversas aplicaciones como: codificación de audio (mp3), evaluación de la calidad de los sistemas de transmisión así también como en la industria de la televisión, cine, conciertos musicales, etc. Esto es debido a la gran importancia que tiene el trabajar con las señales de audio basadas en parámetros perceptuales como lo es la sonoridad ya que, a fin de cuentas, nos da la información más cercana de lo que en realidad escucha el ser humano. Así pues, y debido a la importancia que esto tiene, surge la motivación de trabajar con las señales de voz basadas en parámetros perceptuales.

A lo largo de éste trabajo se estudiará desde el funcionamiento del oído humano, para posteriormente construir un calculador de sonoridad, hasta introducir una nueva forma de estudio para el análisis de la voz, pasando por algunas aplicaciones como lo es el reconocimiento del habla.

En el capítulo 2 se da una introducción a los conceptos teóricos básicos del sonido, la descripción del funcionamiento del oído humano y el efecto que tiene éste sobre las ondas del sonido. Después se introduce el concepto de "sonoridad", se revisa el modelo matemático del oído humano para finalizar con la construcción de la herramienta "Calculador de Sonoridad" Ya teniendo una representación de las señales del sonido en parámetros perceptuales, se trata de explotar las ventajas que esto nos pueda ofrecer.

La primera aplicación que se le da a las señales basadas en parámetros perceptuales es la de realizar el reconocimiento del habla, tratar de evaluar el desempeño del sistema y ver si mejora en comparación con los sistemas actuales de reconocimiento. Esta primera aplicación se estudia en el capítulo 3.

En el capítulo 4, se introduce un concepto nuevo que es el de "imágenes de la sonoridad" En la actualidad se ha trabajado con las señales de voz representadas en un espectograma o en sonogramas. El espectograma es una representación tiempo-frecuencia de la señal de voz, en los sonogramas se estudia la voz pero con diferente resolución en cada banda frecuencial, lo que da una mejor descripción de la señal de voz debido a las características que ésta presenta. En un principio pareciera que las imágenes de la sonoridad no son mas que una representación tiempo-frecuencia con una resolución diferente en cada banda frecuencial, al igual que los sonogramas. La gran diferencia que hay es que las imágenes de la sonoridad están basadas en parámetros perceptuales, y se aprovecha las

ventajas que tiene el trabajar con éste tipo de parámetros. La ventaja más importante, que se describirá en el capítulo 4, es el efecto que tiene el ruido que perturba a una señal de voz, en las imágenes de la sonoridad. Ahora, ¿Por qué llamarles imágenes de la sonoridad y no una simple representación tiempo-frecuencia de la señal de voz basada en parámetros perceptuales?. La respuesta es muy simple, debido a las características que dicha representación presenta, surge la motivación de analizar la voz a través de la representación tiempo-frecuencia-sonoridad pero con herramientas de tratamiento de imágenes, es por eso que se le da el nombre de "imagen de la sonoridad". Con éste concepto se abre un nuevo campo de estudio para el análisis de la voz y en el capítulo 4 se muestran las ventajas que tiene el analizar la voz mediante el estudio de las imágenes de la sonoridad.

En el capítulo 5 se realiza un estudio sobre los métodos existentes para la evaluación de la calidad en los sistemas de transmisión y se muestran las ventajas y desventajas que éstos presentan, esto con la idea de aplicar, de alguna forma, los resultados obtenidos en éste trabajo y ver si pueden ayudar a mejorar el desempeño de los sistemas de evaluación de la calidad del habla.

Definición de objetivos:

Programar el modelo psicoacústico del oído

Crear una base de sonidos en formato WAV

Programar la herramienta Calculador de sonoridad

Utilizar el modelo psicoacústico del oído para realizar el reconocimiento de voz basado en parámetros perceptuales y analizar los resultados

Realizar la construcción de las imágenes de la sonoridad y analizarlas de acuerdo a las características que éstas presentan

Ver las ventajas que presenta el trabajar con señales de voz basadas en parámetros perceptuales y más aún, con las imágenes de la sonoridad

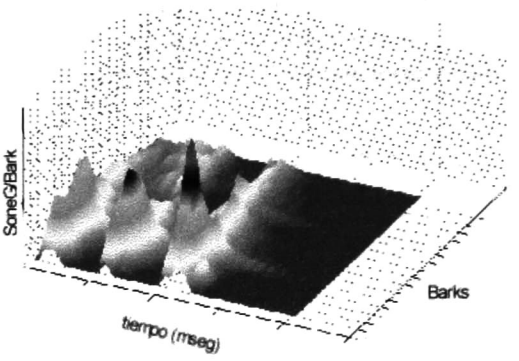
Así pues, comencemos nuestro estudio de la representación de las señales de voz basadas en parámetros perceptuales.

CAPÍTULO 2

2 MODELO PSICOACÚSTICO DEL OÍDO

2.1 Introducción

En éste capítulo se presentan los conceptos más importantes sobre la percepción del sonido. Básicamente se encuentra dividido en tres partes las cuales se describen a continuación: En la primera parte se pretende dar un repaso de conceptos teóricos básicos del sonido como lo son la naturaleza y características del sonido, *nivel de presión sonora* (SPL, por sus siglas en inglés), la *intensidad* y *potencia sonora*. En la segunda parte se da una breve explicación de la estructura del oído humano y se presenta cómo es procesado el sonido a través del oído. Por último, se introducen los conceptos básicos de psicoacústica, se presentan los elementos necesarios para el modelado psicoacústico del oído, el diagrama a bloques para la construcción de un medidor de sonoridad y finalmente la construcción de la herramienta "calculador de sonoridad".



2.2 Generalidades y Medición del Sonido

El sonido se produce mediante un tipo de ondas longitudinales que viajan a través de un medio sólido, líquido o gaseoso. La transmisión de la energía de las ondas sonoras se lleva a cabo por la vibración de las moléculas del medio de transmisión, dichas moléculas vibran en dirección de propagación de la onda y ésta vibración se refleja en una variación de la presión atmosférica, esto es a lo que se llama *presión sonora p* y tiene como unidad el *pascal (Pa)*. El sonido puede ser fácilmente descrito mediante la variación en tiempo de la presión sonora *p(t)*. Las variaciones de presión sonora son mucho menores en relación con las variaciones de la presión atmosférica. Las variaciones entre 10^{-5}Pa para el umbral absoluto y 10^{2}Pa para el umbral de dolor son relevantes.

2.2.1 Nivel de Presión Sonora

El *nivel de presión sonora (SPL)* es una medida que relaciona el valor RMS de la presión sonora con el mínimo audible promedio y está dado por la siguiente ecuación:

$$SPL = 20 \cdot \log \frac{P_{rms}}{p_0} dB, \quad donde$$
$$p_0 = 2 * 10^{-5} Pascal \tag{2.1}$$

A lo largo de éste trabajo nos estaremos refiriendo al *nivel de presión sonora (SPL)* solamente como L.

2.2.2 Nivel de Intensidad Sonora

La *intensidad (o flujo de energía) sonora I* es la potencia (en energía/seg.) transmitida por la onda sonora que cruza una unidad de área ($1m^2$), perpendicular a la dirección de propagación de dicha onda y sus unidades están dadas en W/m^2.

En ondas planas progresivas, el *nivel de presión sonora L* está relacionado con el *nivel de intensidad sonora I* mediante la siguiente ecuación:

$$L = 20 \cdot \log \frac{I}{I_0} dB, \quad donde$$
$$I_0 = 10^{-12} W / m^2 \quad que\ corresponde\ al\ sonido\ más\ débil \tag{2.2}$$
$$que\ el\ humano\ puede\ escuchar.$$

2.3 Procesamiento de la Información en el Sistema Auditivo

Nos detengamos por un momento a escuchar lo que nos rodea. Podemos escuchar ruidos desde el movimiento del aire hasta el tronar de los motores en las avenidas; el canto de un grillo, el agitar de las hojas de los árboles. Por medio de nuestro sentido del oído podemos percibir el exterior y tener un elemento más para identificar el ambiente en el que estamos. La percepción del sonido da lugar a una de las funciones más importantes para el ser humano: el lenguaje. Esto nos permite tener uno de los tres elementos necesarios para establecer una comunicación con nuestro propio ambiente, el elemento receptor.

Hemos dicho con anterioridad que el sonido es una variación de la presión sonora en el tiempo pero, ¿Cómo y en qué forma llegan esas variaciones de presión sonora a nuestro cerebro?. A continuación se presentan los elementos que conforman el sistema auditivo, se describe su funcionamiento y con esto veremos cómo es convertido el sonido a impulsos eléctricos para posteriormente ser procesados en el cerebro.

2.3.1 Región Auditiva

El oído no responde a todas variaciones de presión sonora ni tampoco a todas las frecuencias de la onda sonora. Es ésta la razón por la cual se definió una *región auditiva*. La región auditiva corresponde al área en la cual el sonido es audible y que no causa daño alguno para el sistema auditivo, o dicho de otra forma, es el área que cae dentro del *umbral de audición* y el *umbral de dolor*. A continuación se presenta una gráfica en donde se muestra la región auditiva.

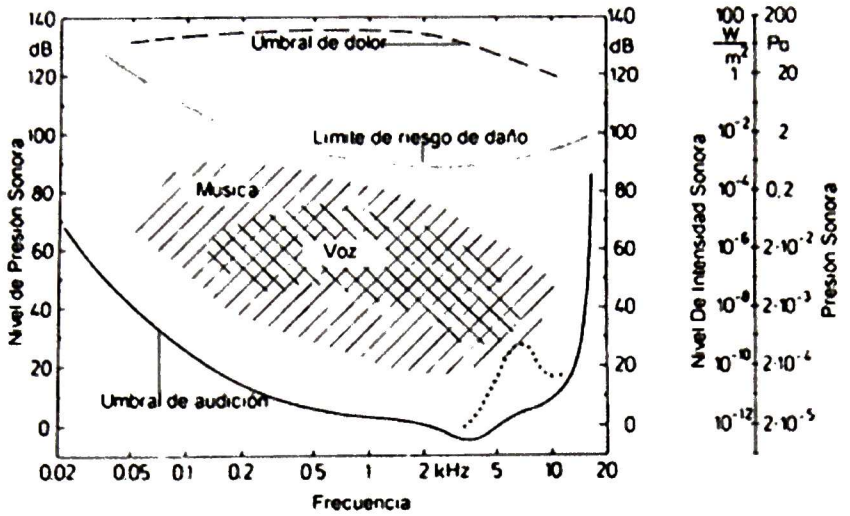


Fig. 2.1. Región auditiva.

En la gráfica anterior las escalas son logarítmicas:
frecuencia en la abscisa
presión sonora (dB) como la ordenada.

A la derecha tenemos una escala logarítmica de la intensidad y presión sonora para tener una relación con el nivel de presión sonora. Como se había mencionado anteriormente, el mínimo sonido audible está en el rango de presión sonora de $2x10^{-5}$ Pa y el rango de frecuencias audibles se encuentra entre 20Hz y 20kHz.

Como ***Umbral de audición*** se le conoce al límite de nivel de presión sonora sobre el cuál un sonido es audible. Todos los sonidos con nivel de presión sonora que estén por debajo de este umbral el oído no es capaz de detectarlos.

El ***umbral de dolor*** es en el que el sonido causa daño al oído y corresponde aproximadamente a los 140dB de nivel de presión sonora L. También se usa representar un límite de riesgo de daño al oído.

El *umbral auditivo* es calculado poniendo un tono de prueba variando su nivel de presión sonora hasta que el sujeto no escuche nada, e ir variando la frecuencia del tono desde frecuencia 0 hasta arriba de 20kHz. Este umbral cambia para las frecuencias arriba de los 1.5kHz dependiendo de la edad de las personas.

Un *estímulo* se genera cuando las ondas sonoras alcanzan el órgano sensorial. Cuando el sonido que provoca dicho estímulo corresponde a un sonido dentro del rango audible, el estímulo da lugar a una *sensación*.

2.3.2 Pre-procesamiento del Sonido en el Sistema Periférico

Consideremos como pre-procesamiento del sonido a todas las modificaciones y conversión del sonido en el sistema mecánico. Como procesamiento de la información le llamaremos al procesamiento a nivel neural. Dependiendo de la dirección proveniente del sonido se experimentan dos tipos de campos sonoros: *campo libre* y *campo difuso*. En el campo sonoro libre el sonido proviene de una sola dirección mientras que en el campo sonoro

difuso el sonido proviene de múltiples direcciones. Básicamente son tres aspectos los que nos interesan sobre el sonido: la ubicación espacial de la fuente sonora, contenido frecuencial e intensidad. Cuando tenemos un sonido proveniente de un campo libre/difuso, la cabeza y los hombros sirven como paredes para rebotar el sonido hacia el oído y así concentrar la mayor parte posible de energía.

El oído periférico básicamente se divide en tres partes: Oído externo, oído medio y oído interno de acuerdo a su ubicación en el cráneo, como se observa en la fig. 2.2. Nuestro objetivo es conocer el funcionamiento del oído como sistema, la forma como la información es procesada dentro del cerebro no es de interés para nosotros por el momento.

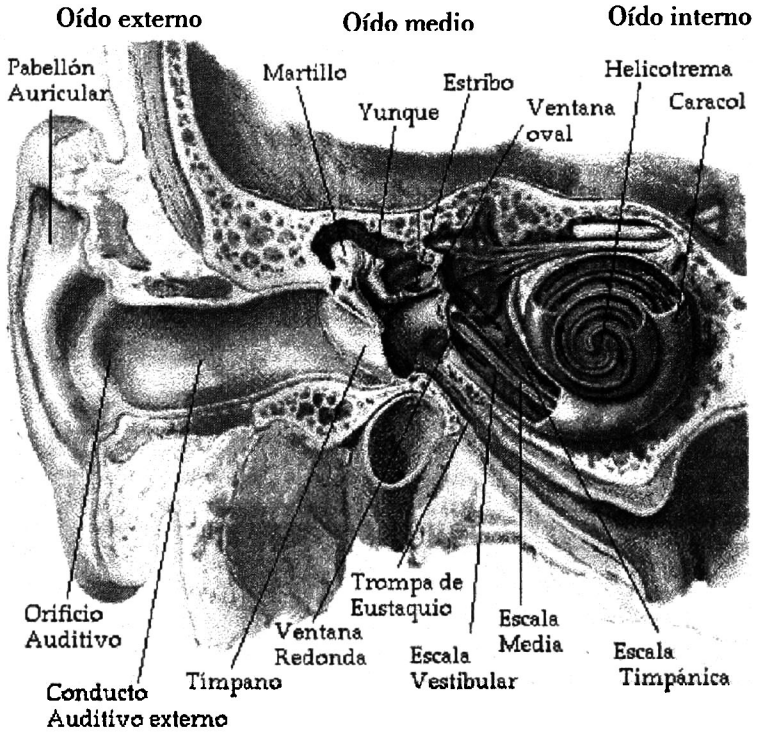


Fig. 2.2. Anatomía del oído.

2.3.2.1 Oído Externo

Se había mencionado anteriormente que la cabeza y los hombros ayudan a dirigir las ondas acústicas hacia el sistema auditivo. El ***pabellón***, que constituye una parte del oído externo, tiene una funcionalidad parecida. Este sirve como una especie de director de ondas para dirigir las ondas sonoras hacia el conducto auditivo externo y así lograr concentrar la mayor energía posible del sonido. Otra función del pabellón es que nos permite hacer una localización espacial de la fuente sonora.

El ***canal auditivo externo***, cuya longitud es aproximadamente 2cm, sirve para conducir las ondas sonoras hacia la ***membrana timpánica*** o ***tímpano*** que se encuentra en el extremo

final del conducto auditivo. La membrana timpánica es la que recibe las ondas sonoras y transfiere la vibración al oído medio. El canal auditivo no responde de la misma forma para todas las frecuencias, en particular, debido a que la longitud del canal auditivo externo es de 0.02m y tomando en cuenta que el canal auditivo es un tubo cerrado en el extremo final por la membrana timpánica, la longitud corresponde a ¼ de la longitud de onda de la onda sonora; tomando que el sonido viaja a una velocidad de 334m/seg en el aire, la frecuencia de resonancia corresponde a 4175Hz, aprox. 4kHz como se puede observar en la figura 2.3. El canal auditivo tiene dos funciones primordiales: Proteger la delicada cadena de huesos que se encuentra en el oído medio, y disminuir la longitud de los nervios provocando una velocidad mayor en la transmisión del los impulsos eléctricos hacia el cerebro.

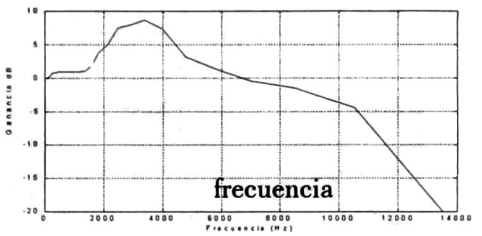


Fig. 2.3. Respuesta en frecuencia del canal auditivo externo.
La frecuencia resonante está en aprox. 4kHz.

2.3.2.2 Oído Medio

El oído medio corresponde a un sistema de acoplamiento de impedancias para lograr la máxima transferencia de la energía. Debido a que las ondas acústicas por lo general viajan en el medio de transmisión aire, y la energía de éstas ondas sonoras tiene que ser transmitida a un medio líquido (oído interno, como se verá mas adelante); Por la diferencia de medios de transmisión no se logra transmitir la máxima energía de las ondas acústicas. Es por esto que se necesita de un mecanismo que realice un acoplamiento de impedancias entre los dos medios para poder hacer la máxima transferencia de la energía. Este mecanismo lo forma una cadena de huesos articulados entre sí que son: el *martillo* (pegado a la membrana timpánica), *yunque* y *estribo* que se encuentra pegado a la *ventana oval* que vendría siendo la puerta al oído interno. El oído medio también funciona como un sistema protector del oído ya que cuando el tímpano recibe sonidos muy fuertes, el músculo del martillo se contrae provocando una tensión en la membrana timpánica y así dificultar la transmisión de las ondas hacia el medio líquido (oído interno).

2.3.2.3 Oído Interno

En el oído interno es en donde el sonido es convertido de ondas acústicas a impulsos eléctricos y su componente principal es la cóclea o caracol. La cóclea es un conducto en forma de espiral que se encuentra dividido internamente por tres escalas: *vestibular* y *media* (divididas por la membrana de Reissner) y escala *timpánica* (dividida por la membrana basilar). La escala vestibular y la escala timpánica están llenas de un líquido salino llamado perilinfa y la escala media esta llena de un líquido llamado endolinfa.

El estribo, del oído medio, esta conectado a la ventana oval que corresponde a la entrada de la escala vestibular. Así pues, las ondas sonoras son recolectadas por el pabellón y

dirigidas hacia el conducto auditivo externo que a su vez conduce las vibraciones de presión hacia la membrana timpánica. El tímpano vibra con las variaciones de presión y conduce dicha vibración hacia el oído interno a través de la cadena de huesecillos, y así es como las ondas acústicas son convertidas a vibraciones en un medio líquido.

La *membrana basilar*, de 32mm de longitud, funciona como un analizador de espectros ya que resuena en diferentes posiciones de acuerdo a la frecuencia. A esto se le llama "código de lugar".

En la figura 2.4 se muestra el movimiento de la membrana basilar para tonos de diferentes frecuencias.

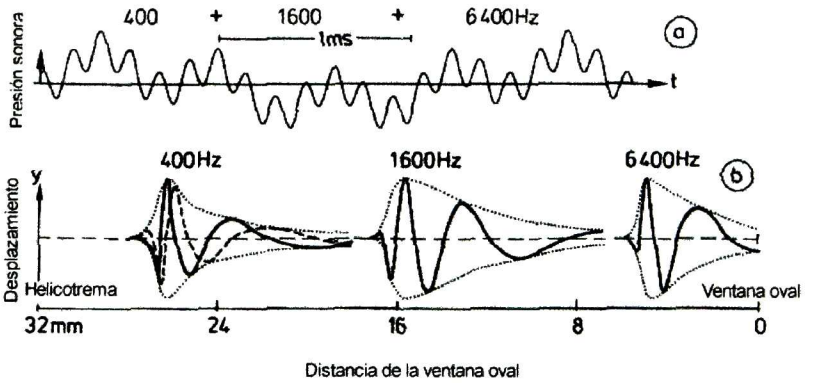


Fig 2.4. a) Señal de sonido en tiempo. b) Movimiento de la membrana basilar para las frecuencias de 400, 1600 y 6400Hz. La línea punteada para la frecuencia de 400Hz corresponde a la señal presentada ¼ del período después de la original. Se puede observar que no ocurren nodos y anti-nodos.

La dirección de propagación de las ondas va de la ventana oval hacia el helicotrema (parte final de la cóclea) en forma creciente hasta llegar a una amplitud máxima para luego llegar a su estado inicial de una forma abrupta. Cerca de la ventana oval, la membrana basilar resuena para frecuencias altas (alrededor de 20kHz). Si vamos recorriendo la membrana basilar en dirección hacia el helicotrema, las frecuencias de resonancia van disminuyendo en forma logarítmica hasta llegar a las frecuencias bajas (en el rango de 20Hz) que corresponde a la parte final de la membrana basilar.

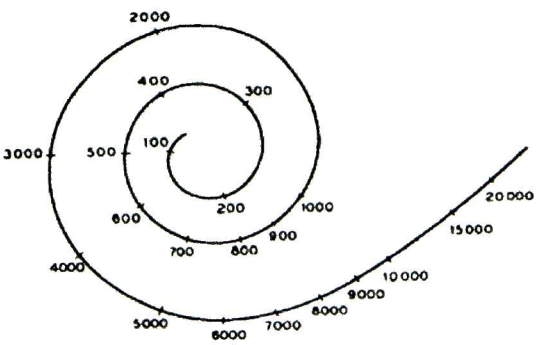


Fig 2.5. Posición de las frecuencias resonantes a lo largo de la membrana basilar.

Por encima y a lo largo de la membrana basilar se encuentra el órgano de Corti el cuál es el encargado de detectar, por medio de las células ciliares, la posición de vibración de la membrana basilar. Las células ciliares internas (aprox. 3600) generan los impulsos eléctricos y son enviados al cerebro por medio del nervio auditivo, las células ciliares externas (eferentes) reciben impulsos del cerebro y sirven como tensores de la membrana basilar para adaptar la sensibilidad a los diferentes niveles de presión sonora.

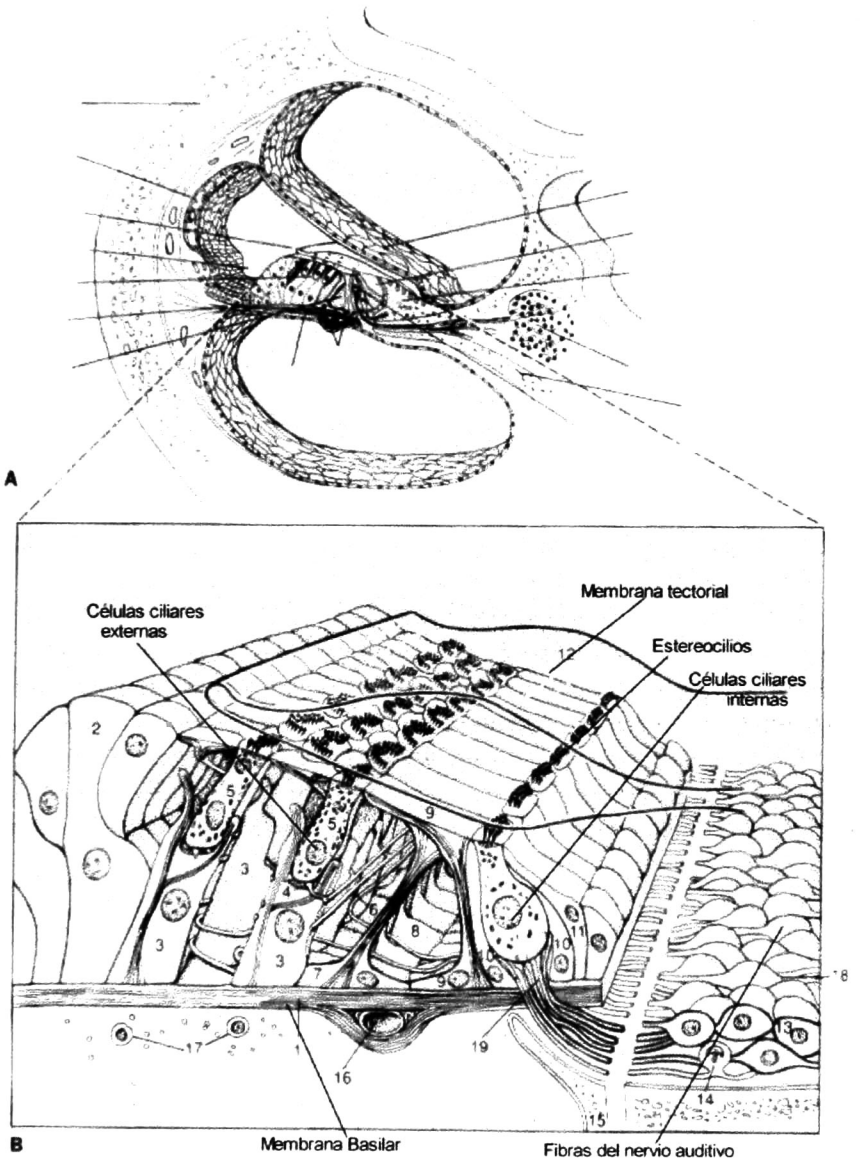


Fig. 2.6. Organo de Corti.

Las células ciliares internas tienen unos bellos llamados estereocilios que están en contacto con la *membrana tectorial*. Cuándo la membrana basilar vibra, los estereocilios se

deflectan (tal como se ve en la figura 2.7 y es cuando se genera la diferencia de potencial que será transmitida al cerebro.

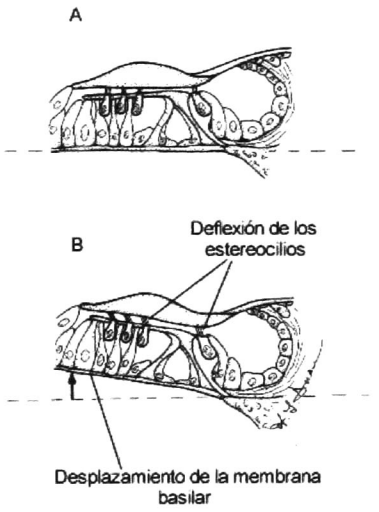


Fig. 2.7. Descripción del movimiento de las células ciliares internas. En la figura A, se presenta la membrana basilar sin movimiento. En la figura B, se presenta el movimiento de la membrana basilar y la deflexión de los estereocilios.

2.3.3 Atributos Subjetivos del Sonido

Cuatro atributos son usados frecuentemente para describir el sonido: sonoridad, tono, timbre y duración. Cada una de éstas cualidades subjetivas depende de uno o mas parámetros físicos. Por ejemplo, la sonoridad depende principalmente de la presión sonora pero también depende del contenido frecuencial del sonido en cuestión. El tono depende principalmente de la frecuencia aunque muestra una pequeña dependencia con la presión sonora y la envolvente. A continuación se muestra la dependencia de las cualidades subjetivas sobre los parámetros físicos.

Tabla 2.1 Dependencia de las cualidades subjetivas del sonido sobre los parámetros físicos.

ATRIBUTOS SUBJETIVOS DEL SONIDO

Parámetros físicos	Sonoridad	Tono	Timbre	Duración
Presión sonora	+++	+	+	+
Frecuencia	+	+++	++	+
Espectro	+	+	+++	+
Duración	+	+	+	+++
Envolvente	+	+	++	+

2.4 Dimensión Subjetiva de la Intensidad Sonora: Sonoridad o Intensidad Subjetiva

Hemos estudiado, hasta este momento, el funcionamiento del sistema auditivo, y sabemos que las ondas sonoras que llegan hasta el oído interno no son exactamente las mismas que generó la fuente sonora, es decir, la percepción que tenemos sobre el sonido es diferente a la que realmente sale de la fuente sonora. Ahora es de interés para nosotros estudiar una de las cuatro características subjetivas del sonido: *la sonoridad*.

De acuerdo a mediciones objetivas que se han hecho de la calidad de la voz en base a cualidades subjetivas, muchos investigadores están de acuerdo en que la medición objetiva en base a cualidades subjetivas, como lo es la sonoridad, es altamente correlacionable con la medición subjetiva como lo es el ***MOS*** por sus siglas en inglés *Mean Opinión Score [105]*. Esto se basa asumiendo que la calidad de la voz está directamente relacionada con la sonoridad de la voz.

La *sonoridad* es la impresión subjetiva de la intensidad sonora percibida; es qué tan fuerte o qué tan débil escuchamos un sonido pero, como mencionamos en la sección anterior, la sonoridad depende del contenido espectral del sonido. Esta dependencia es debida a las características físicas del oído como veremos más adelante.

Tan solo en el oído externo, el canal auditivo es no lineal y debido a esto no todas las frecuencias son escuchadas con la misma intensidad aunque tengan el mismo nivel de L. Mas adelante veremos que otros factores son de importancia para la percepción de la intensidad sonora.

Así pues, comenzamos nuestro estudio para calcular la intensidad sonora percibida por el oído: la sonoridad.

2.4.1 Sensación y Excitación

Como se había mencionado en la sección 2.3.2.3, la membrana basilar resuena en diferentes posiciones dependiendo de el contenido frecuencial del sonido. Si hacemos una prueba con un tono simple, por ejemplo un tono de 1kHz., el patrón de excitación en la membrana basilar debido al estímulo es como se muestra en la figura 2.8.

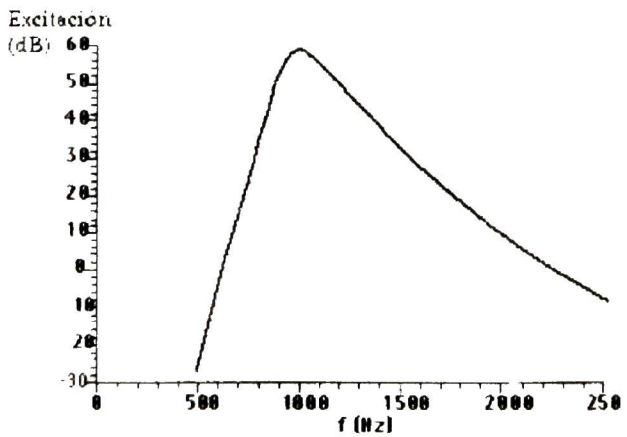


Fig. 2.8. Patrón de excitación de la membrana basilar debido a un tono de 1kHz.

El nivel de excitación máximo en la curva corresponde al nivel de presión sonora del tono [109].

Puesto que lo que nos interesa son los sonidos que dan lugar a las sensaciones, el nivel de excitación máximo del estímulo debe de estar por encima del umbral auditivo.

En muchos casos en lugar de excitación, el nivel de excitación definido como L_E y dado por la ecuación (2.3), es utilizado [109].

$$L_E = 10 \cdot \log \frac{E}{E_0} dB \qquad (2.3)$$

donde:

$-E$ *es la excitacion,*

$-E_0$ *es la excitacion correspondiente a la Intensidad de referencia* $I_0 = 10^{-12} W/m^2$

Hay que tomar en cuenta que lo que nos interesa es el nivel de presión sonora a nivel de tímpano. Esto da lugar a que un *factor de transmisión* a_0 del conducto auditivo externo sea considerado.

2.4.2 Enmascaramiento

El enmascaramiento sonoro puede definirse como el proceso en el cuál el umbral auditivo correspondiente a un sonido se eleva, debido a la presencia de otro sonido [96]. El enmascaramiento sonoro tiene lugar debido a que los receptores auditivos situados en el órgano de Corti tienen un tiempo de respuesta determinado. Por otra parte se cree que los receptores que se encuentran estimulados por una señal A deben de recibir un nuevo nivel de estimulación o excitación debido a otra señal B, tal que la diferencia entre la excitación debida a A y B juntas supere a la debida a A en una determinada magnitud: si eso ocurre, el sonido B será percibido; en caso contrario, B será inaudible [17].

El enmascaramiento sonoro depende de el contenido frecuencial del sonido (separación entre las frecuencias) y en el tiempo, así también como el nivel de potencia sonora del sonido enmascarante y enmascarado.

Para describir los efectos de enmascaramiento se hace la definición de un *umbral de enmascaramiento* que vendría siendo el "nuevo" umbral de audición que corresponde a escuchar o dejar de escuchar un tono de prueba (tono enmascarado) que se va recorriendo desde frecuencias bajas hacia frecuencias altas variando su nivel de presión sonora de manera que sea apenas audible en presencia de un tono enmascarador (de frecuencia y nivel de presión sonora fija). Si no existe ningún tono enmascarador, obviamente el umbral de enmascaramiento corresponde al umbral auditivo.

Existen tres tipos de enmascaramiento: *pre-enmascaramiento*, *enmascaramiento simultáneo* y *post-enmascaramiento*. El pre-enmascaramiento se refiere a que, de acuerdo a la ubicación temporal del tono de prueba con respecto al tono enmascarador, el tono de prueba se presenta antes del tono enmascarador. El enmascaramiento simultáneo es cuando el tono enmascarador y el tono de prueba (enmascarado) se presentan al mismo tiempo. Y el post-enmascaramiento se presenta cuando el tono de prueba (tono enmascarado) se presenta después del tono enmascarador.

El enmascaramiento sonoro tiene dos tipos de efectos: el *efecto temporal* y el *efecto frecuencial* o *espectral*. En la siguiente sección se describen cada uno de éstos efectos.

2.4.2.1 Efecto Frecuencial del Enmascaramiento Sonoro

En la Fig. 2.9. se muestra el umbral de enmascaramiento para el ruido pasa-banda como señal enmascaradora, cuya frecuencia central está a 1kHz para diferentes niveles de presión sonora. Como se puede apreciar, el tono de prueba es enmascarado cuando su frecuencia es cercana a la frecuencia central del ruido enmascarante. La banda de frecuencias va aumentando de acuerdo al nivel de presión sonora del ruido, pero el aumento no es significativo cuando el tono de prueba se acerca por la izquierda a la frecuencia central del ruido. Por el contrario, cuando el tono se aleja por la derecha de la frecuencia central, el tono de prueba es enmascarado en mayor medida y el enmascaramiento está presente aún cuando la frecuencia del tono de prueba es mucho mayor a la frecuencia central del ruido pasa-banda. Mas adelante veremos que dicha banda corresponde al ancho de banda crítica.

Este efecto espectral da lugar a que tonos con frecuencias cercanas a la frecuencia central del ruido, que estén por encima del umbral auditivo, no sean detectadas a menos que su nivel de presión sonora sea mayor que el indicado por el umbral de enmascaramiento.

El efecto frecuencial del enmascaramiento sonoro parece ser razonable si tomamos en cuenta el patrón de excitación que sigue la membrana basilar, ya que la excitación decae mas lento hacia frecuencias altas.

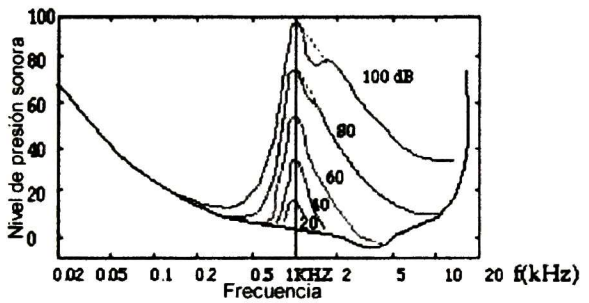


Fig. 2.9. Dependencia del nivel de presión sonora del tono enmascarador en función de la la frecuencia del tono de prueba.

El efecto espectral del enmascaramiento sonoro se refleja pues en una discriminación de las componentes frecuenciales de la señal de sonido. Cabe mencionar que, la banda de frecuencias para las cuales el tono de prueba es enmascarado, depende de la posición de la frecuencia central del ruido.

2.4.2.2 Efecto Temporal del Enmascaramiento Sonoro

Debido a que las células ciliares tienen un tiempo de polarización-depolarización, su tiempo de recuperación es finito. Éste tiempo de recuperación depende de la duración del sonido. Si las células son excitadas durante ese tiempo de recuperación, el sonido no es detectado en su totalidad. En la figura 2.10 se muestra la amplitud de la señal neuronal debido a un tono burst de diferente tiempo de duración.

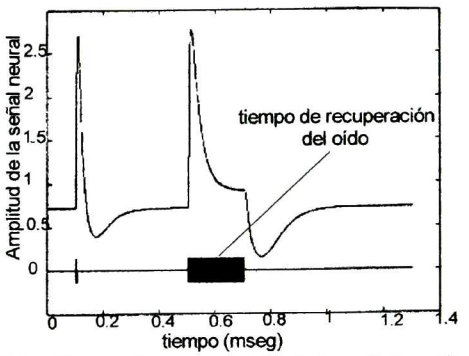


Fig. 2.10. Tiempo de recuperación de las células ciliares dependiente de la duración del sonido.

El efecto temporal tiene lugar cuando la duración del sonido en cuestión es menor a los 200mseg. Como se aprecia en la Fig. 2.11, para sonidos de duración menor a 200mseg, el umbral auditivo aumenta 10dB por cada década de tiempo, después de los 200mseg. el umbral auditivo permanece constante. Ésta dependencia sugiere que, para duraciones del sonido menores a 200mseg. el oído funciona como un detector de energía [109].

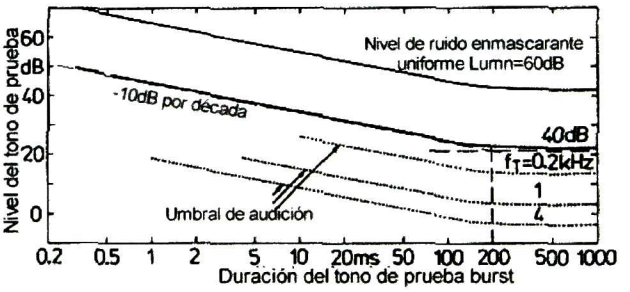


Fig. 2.11. Umbral auditivo (líneas punteadas) y de enmascaramiento (líneas sólidas) en función de la duración del tono de prueba. Tono de prueba burst de 0.2, 1 y 4kHz y ruido enmascarante uniforme con L=60dB.

2.4.3 Bandas Críticas

Una *banda crítica* corresponde a la banda frecuencial en la que las componentes frecuenciales del sonido que caigan dentro de dicha banda, no provocan un incremento en la sonoridad del sonido. En la figura 2.12 se muestra la sonoridad de ruido pasa-banda centrado a 2kHz como una función de su ancho de banda. Se puede apreciar que dentro del ancho de banda crítico, la sonoridad se mantiene constante.

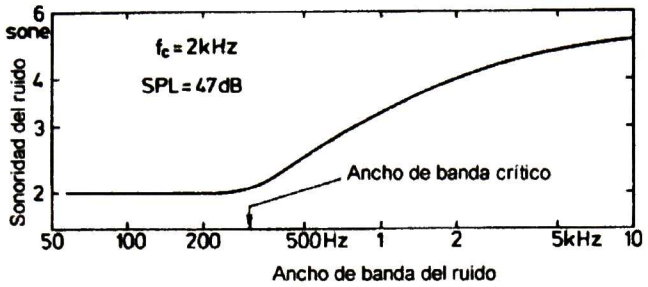


Fig. 2.12. Sonoridad de ruido pasa-banda centrado a 2kHz con una potencia *SPL* total de 47dB como una función de su ancho de banda.

Debido al gran uso de éstas bandas críticas, fue creada una *tasa de bandas críticas* (*z*) la cuál es una escala que describe el rango audible en una forma lineal. Tiene como unidad el *Bark* en honor al científico Barkhausen quién introdujo el *phon*, un valor para describir el nivel de sonoridad. El rango audible fue dividido en 24 bandas críticas y como consecuencia la tasa de bandas críticas crece desde 0 hasta 23 Barks.

El *ancho de banda crítico* depende de la frecuencia central de dicha banda. En la tabla 2.2. se encuentran sumarizadas las 24 bandas críticas.

Tabla 2.2. Tasa de bandas críticas *z*, frecuencia baja f_l, frecuencia alta f_u, límite frecuencial del ancho de banda crítico, Δf_G, centrada a f_c.

z	f_l	f_u	f_c	z	Δf_G
Bark	Hz	Hz	Hz	Bark	Hz
0	0	100	50	0.5	100
1	100	200	150	1.5	100
2	200	300	250	2.5	100
3	300	400	350	3.5	100
4	400	510	450	4.5	110
5	510	630	570	5.5	120
6	630	770	700	6.5	140
7	770	920	840	7.5	150
8	920	1080	1000	8.5	160
9	1080	1270	1170	9.5	190
10	1270	1480	1370	10.5	210
11	1480	1720	1600	11.5	240
12	1720	2000	1850	12.5	280
13	2000	2320	2150	13.5	320
14	2320	2700	2500	14.5	380
15	2700	3150	2900	15.5	450
16	3150	3700	3400	16.5	550
17	3700	4400	4000	17.5	700
18	4400	5300	4800	18.5	900
19	5300	6400	5800	19.5	1100
20	6400	7700	7000	20.5	1300
21	7700	9500	8500	21.5	1800
22	9500	12000	10500	22.5	2500
23	12000	15500	13500	23.5	3500

La *intensidad por banda crítica* puede verse como la suma de intensidades que caen dentro de una ventana frecuencial con el ancho de banda crítico [109], y puede ser calculado por la siguiente ecuación.

$$I_G(z) = \int_{z-0.5Bark}^{z+0.5Bark} \frac{dI}{dz} dz \tag{2.4}$$

El *nivel de banda crítica* L_G está dado por la relación

$$L_G = 10 \cdot \log \frac{I_G}{I_0} dB \quad donde\ I_0\ corresponde\ al\ valor\ de\ referencia\ I_0 = 10^{-12} W/m^2 \tag{2.5}$$

2.4.4 Sonoridad

Al principio de ésta sección se definió la sonoridad y se dijo que es una impresión subjetiva de la intensidad sonora; así también se mencionó la importancia de dicha cualidad del sonido. A continuación se describen lo que es el *nivel de sonoridad* así también como su función. Posteriormente se establece una relación de la sonoridad con la excitación.

2.4.4.1 Nivel de Sonoridad

Para definir de una manera cuantitativa a la sonoridad tenemos lo que es el *nivel de sonoridad*, esto vendría siendo como cuantificar a las sensaciones producidas por el sonido.

El nivel de sonoridad está definido como el nivel de presión sonora de un tono de 1kHz necesario para que suene igual de fuerte que un tono en cuestión y es expresado en phons. Las curvas iso-sonoras, representan el nivel de presión sonora de cada una de las frecuencias para que tengan la misma sonoridad así como se puede observar en la fig. 2.13. Las curvas iso-sonoras tienen un comportamiento parecido al umbral auditivo, como se observó en la figura 2.1.

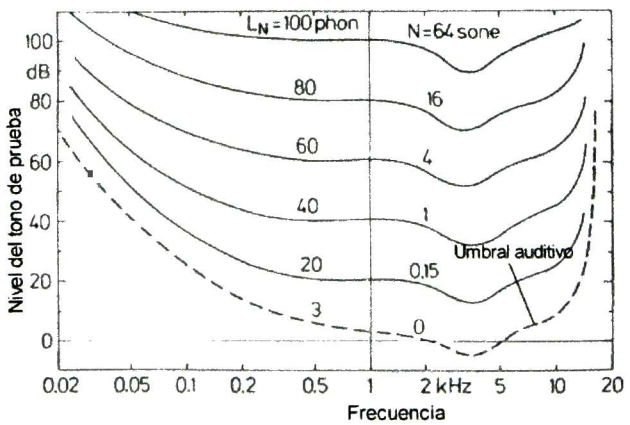


Fig. 2.13. Curvas iso-sonoras. El nivel de sonoridad para un tono 1kHz se considera como la referencia y corresponde directamente al nivel de presión sonora.

Un *sone* es una unidad utilizada para expresar la sonoridad; doblando el número de sones debería de describir un sonido dos veces más fuerte y corresponde al tono de referencia de 1kHz a 40dB de presión sonora.

Hasta ahora hemos definido la sonoridad pero no la hemos relacionado con el nivel de excitación.

Para el patrón de excitación de la membrana basilar, la *sonoridad o sonoridad relativa N*, corresponde al área bajo la curva de la excitación, como se muestra en la figura 2.14. La *sonoridad relativa N* es la suma de todas las *sonoridades específicas*, llamadas *N'*, sobre la tasa de bandas críticas *z*. Matemáticamente la *sonoridad relativa N* se puede expresar como:

$$N = \int_0^{24\,Bark} N'\,dz \tag{2.6}$$

En base a cálculos realizados y experimentaciones, E. Zwicker y colaboradores han propuesto una ecuación para calcular la sonoridad relativa y está dada por:

$$N' = 0.068\left(\frac{E_{TQ}}{s \cdot E_0}\right)^{0.25}\left[\left(1 - s + s \cdot \frac{E}{E_{TQ}}\right)^{0.25} - 1\right]\frac{sone}{Bark} \tag{2.7}$$

donde E_{TQ} corresponde a la excitación en el umbral auditivo, E_0 es la excitación correspondiente al nivel de intensidad de referencia $I = 10^{-12}W/m^2$, E es la excitación y s corresponde a el índice de enmascaramiento.

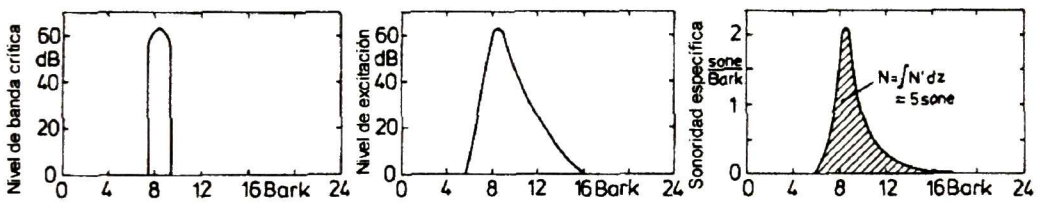


Fig. 2.14. Relación entre el nivel de excitación *E* y la sonoridad específica. A la izquierda se muestra la intensidad de banda crítica. El área sombreada en la figura de la derecha corresponde a la sonoridad relativa

2.5 CALCULADOR DE SONORIDAD

2.5.1 Diagrama a Bloques de un Medidor de Sonoridad

A continuación se hace una descripción del diagrama a bloques mostrado en la figura 2.15. La señal sonora es tomada mediante un micrófono y es amplificada. Posteriormente se alimenta a un filtro (libre/dif) para dar los efectos del campo sonoro libre/difuso mediante el factor de transmisión a_0. En la tabla 2.3 se presentan los valores para el factor de transmisión a_0 en función de la tasa de bandas críticas *z*.

Tabla 2.3. Factor de transmisión a_0.

z	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
a_0	0	-0.2	-0.8	-0.8	-0.9	-0.92	-0.93	-0.94	-0.95	-0.97	-1	-1.8	-3.8	-5	-7.5	-8	-8.7	-7.3	-3.1	-1.4	-0.5	-1.5	-4.4	20

En seguida la señal es pasada por el banco de filtros de banda crítica. Para el objetivo de este trabajo, únicamente se utilizan 16 bandas debido a que las señales de interés para nosotros son las correspondientes al rango de voz.

En la etapa siguiente, junto con el filtro pasa-bajos (LP) para dar la envolvente temporal de la salida de los filtros, se realiza el cálculo de la energía de la señal, ya que lo que nos interesa de las bandas críticas es la energía que se encuentra en dicha banda. En la etapa N', la sonoridad específica es calculada y el efecto frecuencial del enmascaramiento sonoro es realizado. La etapa NL es un filtro de primer orden con una constante de tiempo τ de aproximadamente τ=200mseg., el cuál se simula el efecto temporal del enmascaramiento sonoro. Por último se realiza el cálculo de la sonoridad relativa que es la sumatoria de todas las sonoridades específicas como se había mencionado con anterioridad.

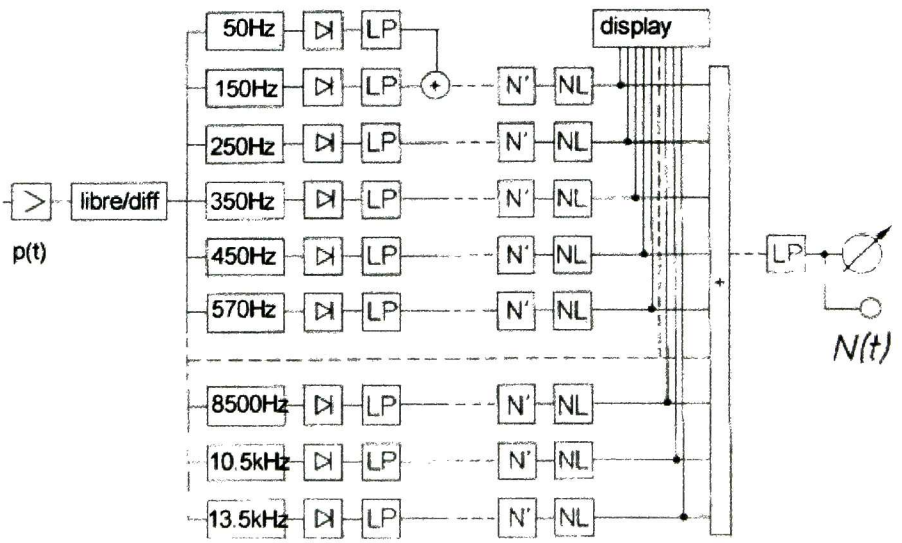


Fig. 2.15. Diagrama a bloques para un medidor de sonoridad propuesto por E. Zwicker en Psychoacoustics: Facts and models. Springer Series & information series. 1999

2.5.2 PROGRAMACIÓN EN MATLAB

En la actualidad existen varias herramientas que facilitan el modelado de sistemas físicos, como lo son: MATLAB y LabView. Estas dos herramientas son una opción excelente para nuestras necesidades. La ventaja significante que nos ofrece LabView sobre Matlab es que el modelo puede ser ejecutado en tiempo real. Sin embargo, la herramienta "calculador de sonoridad" fue programada en MATLAB debido a que era el lenguaje con el que se contaba y cumplía con las necesidades requeridas.

El objetivo de construir esta herramienta es facilitar el análisis de la voz basada en parámetros perceptuales, como lo es la sonoridad.

La herramienta tiene las siguientes características:

Interfase gráfica con el usuario

Trabaja con diferentes formatos de sonido: SAM y WAV.

Graba sonidos en formato WAV

Permite escuchar el sonido

Permite visualizar la señal de sonido

Mezcla la señal original con diferentes tipos de ruido como son: blanco, ambiental, interior de un carro, tanque militar y ruido de fábrica. La SNR es definida por el usuario y está dada en dB.

Realiza el cálculo de la sonoridad así también como su visualización.

Permite grabar la señal bidimensional de la sonoridad: tiempo-bandas críticas, como imagen en formato TIFF.

Cuenta con un sistema de aprendizaje y reconocimiento de voz basado en redes neuronales.

Técnicas para el análisis de las imágenes del sonido: morfología matemática y transformada discreta de Fourier cuaterniónica (DQFT).

Servicio de impresión.

La interfase gráfica con el usuario es mostrada en la fig. 2.16.

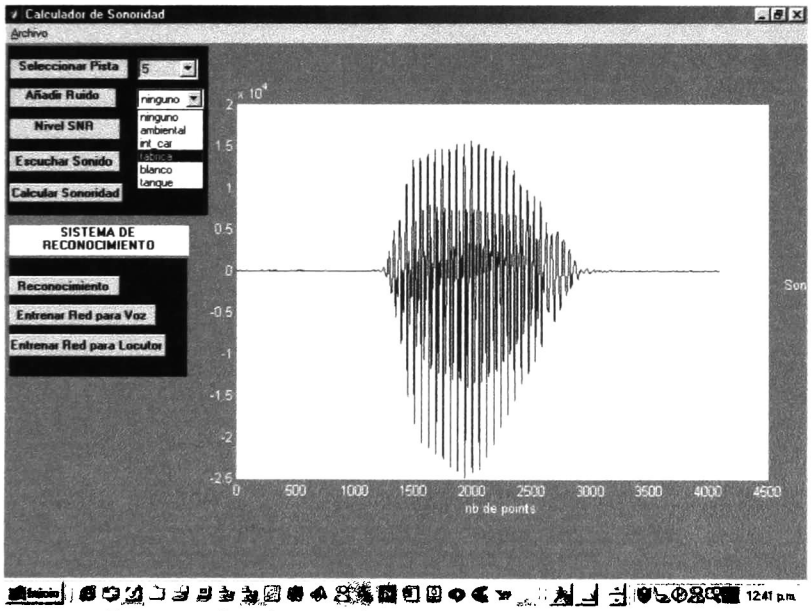


Fig. 2.16. Interfase con el usuario del calculador de sonoridad.

2.5.3 Cálculo y Visualización de la Sonoridad

Para hacer el calculo de la sonoridad, como se dijo en la sección anterior, se utilizó el modelo propuesto por E. Zwicker.

Puesto que lo que nos interesa es hacer el análisis para señales de voz, nos limitaremos a trabajar en el rango de frecuencias de 20Hz a 4kHz, esto significa que solo serán tomadas 16 de las 24 bandas críticas.

Puesto que los filtros que necesitamos son filtros adyacentes, es decir, la frecuencia de corte de un filtro es la frecuencia baja del siguiente filtro pasa-banda, se diseñaron filtros elípticos. En la figura 2.17 se presentan las respuestas de los 16 filtros pasa-banda.

Como se había explicado anteriormente, para el cálculo de la sonoridad se necesita el patrón de excitación de la membrana basilar. También se mencionó que el nivel de banda crítica, que corresponde a la suma de las intensidades que caen dentro de una banda crítica, equivale al valor máximo del patrón de excitación de la membrana basilar. A la sonoridad calculada para el máximo valor del patrón de excitación se le llama sonoridad principal. Por el momento nos limitaremos a calcular ésta sonoridad principal, teniendo este valor, podemos aproximar la "sonoridad de flanco" que es la sonoridad de la excitación de la membrana basilar hacia frecuencias altas.

El cálculo de la sonoridad principal se llevará a cabo utilizando la ecuación 2.7. Así pues, los datos necesarios para resolver dicha ecuación son:

E: Nivel de excitación máximo

S: Índice de enmascaramiento

Etq: Umbral auditivo

El nivel máximo de excitación, como mencionamos anteriormente, corresponde a la energía contenida dentro de cada banda crítica. El índice de enmascaramiento se obtuvo mediante tabulación y se da en la tabla 2.4.

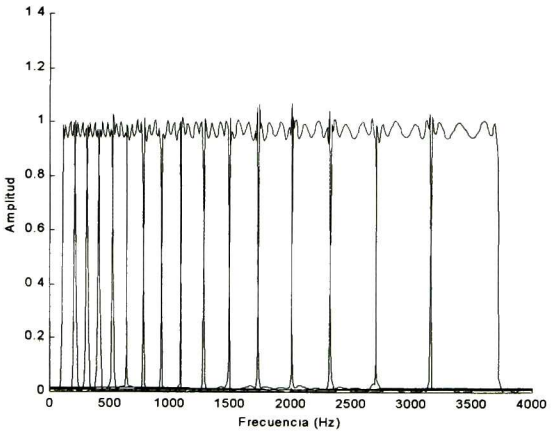


Fig. 2.17. Respuesta de los 16 filtros elípticos pasa-banda.

Tabla 2.4. Índice de enmascaramiento a_v

Barks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a_v(dB)	-2	-2	-2	-2	-2	-3	-3	-3	-3	-3	-3	-4	-4	-4	-4	-4

El umbral auditivo también se obtuvo mediante tabulación y se presenta en la siguiente figura.

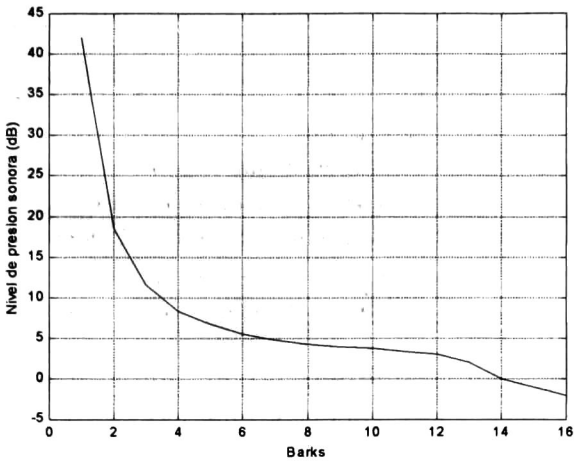


Fig. 2.18. Umbral auditivo en función de la tasa de bandas críticas.

El factor de transmisión a_0 no es tomado en cuenta puesto que se mantiene invariante para el rango de frecuencias de 0-4kHz.

A continuación se presenta una prueba que se hizo con el programa para un tono de 1kHz a 60dB.

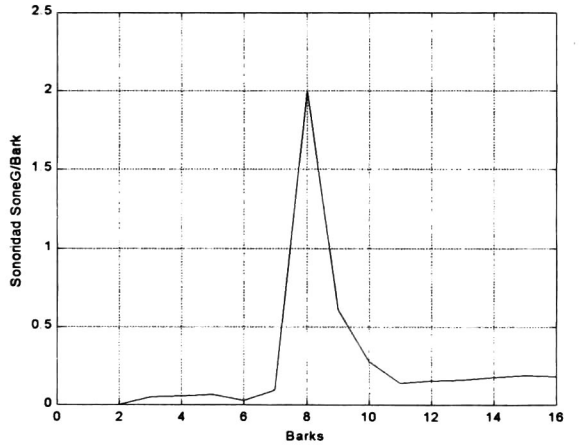


Fig. 2.19. Sonoridad específica de un tono de 1kHz a 60dB.

Se puede hacer la comparación de la fig. 2.19 con la fig. 2.14 y vemos que sí corresponde el valor de la sonoridad específica, que es 2 sonesG/Bark.

Para visualizar la sonoridad, se construyó una gráfica tri-dimensional tiempo-Barks-Sonoridad y a continuación se da un ejemplo, que corresponde a la palabra "inicia".

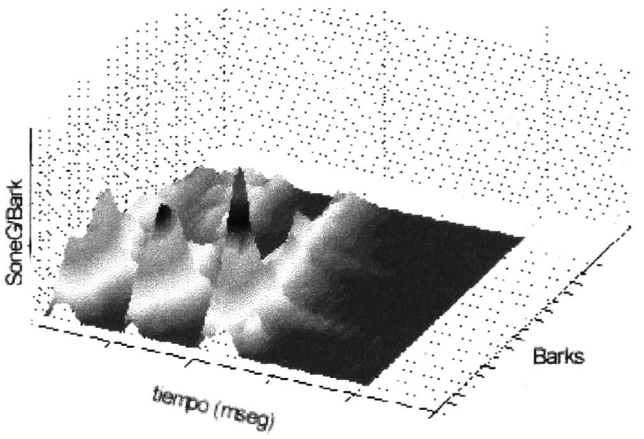


Fig. 2.20. Visualización de la evolución de la sonoridad calculada para la palabra "inicia" en función del tiempo y bandas críticas.

2.6 Conclusión.

Con los resultados obtenidos, comparados con la bibliografía, ésta herramienta presenta una buena aproximación para el cálculo de la sonoridad. La sonoridad principal es calculada satisfactoriamente puesto que solo depende de la energía de cada banda crítica. Por el contrario, la sonoridad específica varía un poco con respecto a la bibliografía. Esto es debido a que se está utilizando una aproximación del patrón de excitación de la membrana basilar. Sin embargo, se han hecho pruebas con diferentes valores del patrón de excitación y las gráficas construidas no tienen cambios significativos, la diferencia es significativa únicamente al hacer la sumatoria de todas las sonoridades específicas, es decir para el cálculo de la sonoridad relativa. Para el ejemplo del tono de 1kHz a 60dB, la bibliografía presenta un cálculo de la sonoridad relativa de aproximadamente 5 sones. Con la herramienta que se programó, la sonoridad relativa tiene un valor de aproximadamente 4.4 sones.

Teniendo ya una representación de la señal de voz en base a parámetros perceptuales, como lo es la sonoridad, procederemos a aplicar dicha representación para el reconocimiento de locutor así también como reconocimiento del habla.

La red neuronal y las herramientas para el análisis de las imágenes del sonido serán presentadas en capítulos posteriores.

CAPÍTULO 3

3 RECONOCIMIENTO DEL HABLA

3.1 Introducción

En la actualidad ya existen muchos sistemas de reconocimiento de voz bastante efectivos. Sin embargo, cuando estos sistemas trabajan en ambientes ruidosos el desempeño disminuye significativamente. El ser humano es capaz de reconocer, en un ambiente ruidoso y con baja relación señal a ruido, lo que otra persona le esta diciendo. Uno de los objetivos que se pretende lograr con esta aplicación es poder hacer el reconocimiento de voz utilizando señales de voz basadas en parámetros perceptuales y ver las ventajas y desventajas que presenta. Una de las ventajas que puede presentar es la robustez al ruido, y esto es de suponerse debido a los efectos frecuenciales del enmascaramiento sonoro en el oído humano. A lo largo del capítulo se describe el desarrollo del sistema de reconocimiento y al final se hace un análisis de los resultados obtenidos.

3.2 Concepto Común en Todos los Sistemas de Reconocimiento

Como reconocimiento de voz se le llama al proceso de convertir una señal acústica, capturada por un micrófono, a un conjunto de palabras. Las palabras reconocidas pueden ser utilizadas para diferentes aplicaciones como lo son: aplicaciones de control mediante comandos, captura de datos, etc. También pueden ser utilizadas para el procesamiento lingüístico para entender el habla.

Los parámetros más importantes que caracterizan un sistema de reconocimiento son los siguientes:

Parámetros	Rango
Modo del habla	Palabras aisladas o habla continua
Estilo del habla	Leída o espontánea
Dependencia	Dependiente del locutor o independiente del locutor
Vocabulario	Pequeño (<20 palabras) o grande (>20,000 palabras)
Modelo del lenguaje	Estado finito o sensitivo al contexto
Perplejidad	Pequeño (<10) o grande (>100)
SNR	Alta (>30dB) o baja (<10dB)
Transductor	Micrófono cancelador de eco o teléfono

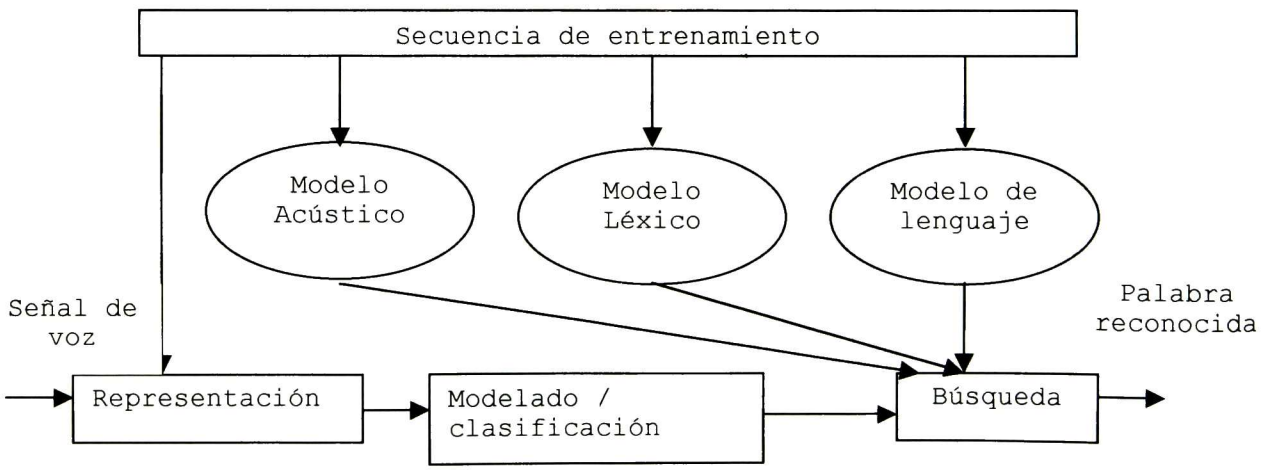
El reconocimiento puede ser de palabras aisladas o palabras continuas. En el caso de palabras aisladas, el locutor tiene que hacer una pausa entre palabra y palabra, cosa que para el reconocimiento de palabras continuas, no es necesario.

El habla espontánea es mucho más difícil de reconocer que cuando se está leyendo un texto. Así también, si el vocabulario es muy grande, el reconocimiento se vuelve mas complejo ya que existen un mayor número de palabras con sonidos parecidos. En éste caso, se utilizan reglas gramaticales para restringir determinado número de palabras. Dependiendo de la aplicación, el reconocimiento puede ser dependiente del locutor o independiente del locutor.

El modelo de lenguaje mas simple puede ser construido por una red de estado finito, en la cual las palabras permisibles que siguen a otra palabra se dan explícitamente. Un modelo mas general puede ser construido utilizando reglas gramaticales. La perplejidad es una medida muy popular para determinar que tan difícil puede ser una tarea, y está definida como la media geométrica del número de palabras que pueden seguir a otra después de haber aplicado el modelo del lenguaje. Finalmente, los factores que afectan al reconocimiento es el ruido ambiental y la posición del micrófono.

El reconocimiento del habla es una tarea muy difícil debido a que depende de muchos factores. Como ejemplo tenemos la dependencia de las realizaciones acústicas de los fonemas sobre el contexto en el cuál aparecen. También el reconocimiento se torna más difícil debido a la variabilidad acústica por los cambios en el ambiente, así también como la posición del transductor. Otro factor importante es la variabilidad del locutor. Esto se refiere a que la forma como habla una persona depende de el estado físico y emocional o la calidad de la voz. Finalmente se tienen las diferencias en el fondo sociolingüísticos, el dialecto, etc.

A continuación se presentan los componentes típicos de un sistema de reconocimiento.



3.3 Redes Neuronales

Las redes neuronales artificiales han tomado un papel muy importante en aplicaciones como: reconocimiento de patrones, sistemas de identificación, clasificación, voz, visión y sistemas de control, esto es debido a la gran habilidad de aprendizaje que poseen, el alto nivel de no linealidad, su alto nivel de paralelismo y la posibilidad de generalizar.

En la actualidad existe un sin número de tareas específicas en las cuáles las redes neuronales artificiales pueden ser aplicadas. Dependiendo de la aplicación, será la arquitectura de la red que se utilizará. Nuestro estudio se limitará a las redes neuronales con aplicación al reconocimiento de patrones.

3.3.1 Conceptos Teóricos Básicos

Las redes neuronales están compuestas por unidades simples de procesamiento llamadas neuronas. La forma de actuar de estas unidades fue inspirado por la forma como funciona el sistema nervioso biológico, y tiene un cierto parecido al funcionamiento del cerebro en los siguientes aspectos: el conocimiento del ambiente adquirido por la red es a través de un proceso de aprendizaje y, las conexiones interneurales son utilizadas para almacenar el conocimiento adquirido [16].

El modelo de la neurona artificial (a veces llamada perceptrón simple) es como se presenta en la figura 3.1. Como podemos ver, la neurona artificial consta de señales de entrada las cuales son las descriptoras del ambiente, pesos sinápticos, un sumador para generar el potencial de acción v debido a las entradas, un ajustador o Bias b_k, una función de activación (puede ser no lineal) para limitar la amplitud de la salida de la neurona, y la salida y_k.

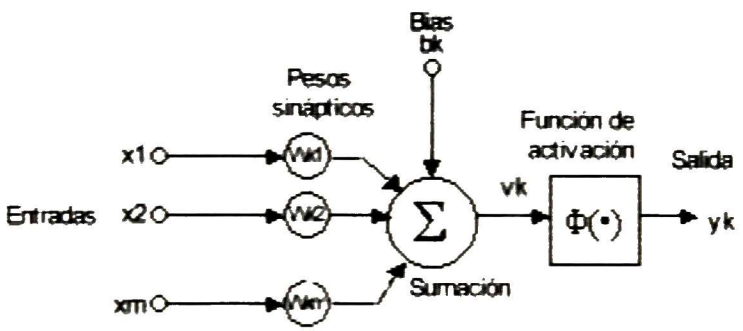


Fig. 3.1. Modelo no lineal de una neurona

La salida y_k está definida por la siguiente ecuación:

$$y_k = \varphi(u_k + b_k) \tag{3.1}$$

donde:

$$u_k = \sum_{j=1}^{m} w_{kj} x_{kj} \tag{3.2}$$

$$v_k = u_k + b_k \tag{3.3}$$

y_k es la salida
v_k es el potencial de acción
w_k son los pesos sinápticos
x_k son las entradas
b_k corresponde al bias.

Dependiendo de la función de activación será la salida de la neurona. Por ejemplo, si tuviéramos al escalón unitario como función de activación, la salida sería:

$$Si \quad v_k \geq 0 \Rightarrow y_k = 1,$$

$$Si \quad v_k < 0 \Rightarrow y_k = 0$$

Aquí la neurona permite de realizar una separación de los puntos de entrada en dos clases, denotadas por 0 y 1.

El perceptrón de una sola capa es unicamente varias neuronas en paralelo completamente conectadas (full-connected), el número de neuronas depende del número de salidas que se desee. El perceptrón multicapa (MLP por sus siglas en inglés multi-layer perceptron) como su nombre lo dice, no solamente tiene una capa sino varias capas: una capa de entrada, una de salida y capas intermedias llamadas capas ocultas. Cabe mencionar que las capas ocultas ayudan a extraer características propias de las secuencias de entrada.

Un perceptrón comprende tres elementos principales:

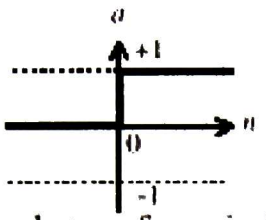
- Una retina: Consiste de células en la cuál es registrado el estímulo.

Una capa de células de asociación: Cada una de éstas células pueden ser conectadas a otras células de la retina, células de asociación, y células de decisión. Estas hacen la suma de los impulsos los cuáles vienen a ésta de otras células a la cuál esta conectada. La célula responde de acuerdo a una ley definida después de haber comparado la suma obtenida previamente con un umbral. La dirección de las conexiones es hecha de la retina hacia las células de asociación.

Una capa de células de decisión: Estas células funcionan como las células de asociación. Reciben sus entradas de otras células de asociación u otras células de decisión. Estas células representan la salida del perceptrón. La dirección de las conexiones entre las células de asociación y las células de decisión es bidireccional, las cuales permiten un retorno de la salida sobre la red.

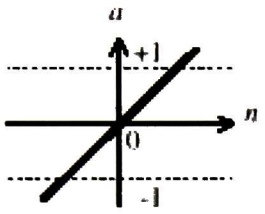
3.3.1.1 Funciones de Activación

Las funciones de activación junto con la forma de ponderación son las que determinarán las características de la neurona artificial. La función de activación puede ser lineal o no lineal y las más comunes son las que se presentan a continuación:



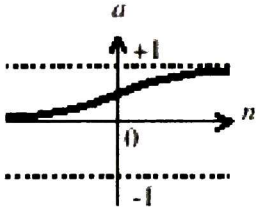
Función de transferencia de límite duro

$$\varphi(n) = \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases}$$



Función de transferencia lineal

$$\varphi(n) = n$$



Función de transferencia sigmoide

$$\varphi(n) = \frac{1}{1+\exp(-n)}$$

3.3.2 Algoritmos de Aprendizaje

Para que la red pueda aprender, se necesita hacer un entrenamiento. Existen varios tipos de entrenamiento pero los más comunes son los siguientes: entrenamiento supervisado y entrenamiento no supervisado. En el supervisado, un "supervisor" presenta a la red estímulos de entrada y la salida deseada. En un entrenamiento no supervisado se le permite a la red que se organice por sí misma y construya su propia representación de los datos de entrada. Con el entrenamiento, los pesos sinápticos de la red son modificados hasta que la red cumpla con la tarea para la cuál se le está entrenando. La manera como estos pesos sinápticos son modificados depende del algoritmo de aprendizaje que se utilice. El algoritmo más común y sencillo para el entrenamiento de las redes multi-capa es el método de retro-propagación de gradiente. Podemos decir que la red está bien entrenada cuando la red no memoriza, y en cambio hace una generalización. También se necesitan escoger buenos ejemplos de entrenamiento, es decir los que se encuentren más dispersos para cubrir el espacio de variación de los parámetros de entrada. Y por último, escoger la arquitectura de red que sea la más óptima para nuestras necesidades.

3.3.2.1 Algoritmo de Retropropagación de Gradiente (Back-Propagation)

El algoritmo de retropropagación es un algoritmo de entrenamiento supervisado, con corrección de errores el cuál consiste en calcular los pesos sinápticos hasta que la diferencia entre la salida de la red y la salida deseada sea mínima. Eventualmente en la última capa es en donde se estima el gradiente del error, puesto que son las únicas neuronas visibles para la cual las señales de error pueden ser calculadas directamente. Es común utilizar el error cuadrático medio (mse por sus siglas en inglés mean squared error) como medida a ésta diferencia.

A continuación se hace una descripción del algoritmo de retropropagación.

La señal de error a la salida de la neurona j en la iteración n, esta dada por la siguiente ecuación:

$$e_j(n) = d_j(n) - y_j(n) \tag{3.4}$$

donde
d_j es la salida deseada, y
y_j es la salida de la neurona j.

Calculando la energía del error para cada neurona en esa capa y sumando las energías de todas las neuronas en esa capa, tenemos

$$\varepsilon(n) = \frac{1}{2}\sum_{j \in C} e_j^{\,2}(n) \tag{3.5}$$

donde C incluye todas las neuronas en la capa de salida de la red.

El error promedio habiendo mostrado todos los patrones (ejemplos) está dado por

$$\varepsilon_{av}(n) = \frac{1}{N}\sum_{1}^{N} \varepsilon(n) \tag{3.6}$$

donde
N es el número de ejemplos.

El objetivo es cambiar los pesos sinápticos de la red hasta minimizar la función de error ε_{av}. Al igual que el algoritmo LMS, calculando el gradiente de la función de error ε_{av} nos dará el factor de corrección Δw_{ji} para modificar los pesos sinápticos.

El gradiente de la función de error ε se indica en la siguiente ecuación

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}} = \frac{\partial \varepsilon(n)}{\partial e_j(n)} \cdot \frac{\partial e_j(n)}{\partial y_j(n)} \cdot \frac{\partial y_j(n)}{\partial v_j(n)} \cdot \frac{\partial v_j}{\partial w_{ji}} \tag{3.7}$$

Simplificando, tenemos que el factor de corrección está dado por la siguiente ecuación

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}} = \Delta w_{ji}(n) = \eta \delta_j(n) y_i(n) \tag{3.8}$$

donde

δ_j (n) es el gradiente local y está definido como

$$\delta_j(n) = e_j(n)\varphi_j'(v_j(n)) \tag{3.9}$$

y η es la tasa de aprendizaje.

Aquí se actualizan los pesos sinápticos w_{ji} de la neurona j y se propaga el error hacia la capa anterior. Los nuevos pesos sinápticos se actualizan como se indica en la siguiente ecuación

$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji}(n) \tag{3.10}$$

Lo que en realidad se propaga a la capa anterior es el gradiente local, y está definido como el producto de la función derivativa asociada $\varphi_j'(v_j(n))$ y la sumatoria de todos los gradientes locales δ_s calculados para las neuronas de la capa siguiente.

3.4 Sistema de Reconocimiento de Voz

La idea de este trabajo, como se mencionó anteriormente, es realizar el reconocimiento de voz utilizando señales de voz basadas en parámetros perceptuales como lo es la sonoridad. Nuestro objetivo no es hacer reconocimiento del habla continua puesto que el análisis es mucho más complejo, en cambio nos limitaremos a realizar reconocimiento de palabras aisladas. Una aplicación de esto es una interfase hombre-máquina. Como ejemplo podemos tener la industria, los supervisores de línea en lugar de darle instrucciones a las máquinas por medio de un teclado, le podrían dar instrucciones a distancia por voz mediante un micrófono inalámbrico. Para éste tipo de aplicación el vocabulario es limitado y son únicamente comandos. La instrucción que se le de a la máquina posiblemente se requiera que sea reconocida por cualquier operario o por un operario en particular. Si éste fuera el caso, también es necesario hacer reconocimiento de locutor.

Uno de los factores importantes que hay que tomar en cuenta es el ambiente en el que se requiere realizar el reconocimiento de voz. En la actualidad hay muchos sistemas de reconocimiento de voz que son bastante eficientes en ambientes sin ruido, pero uno de los más grandes problemas con los que se han enfrentado es hacer el reconocimiento de voz en ambientes ruidosos.

3.4.1 Descripción de la Base de Datos

Para el desarrollo de éste trabajo se utilizó una base de datos en francés. Esto fue debido a que eran las herramientas con las que se contaba. La base de datos fue enviada por el laboratoro LASMEA de la universidad Blaise Pascal de Clermont-Ferrant, Francia.

Las ventajas que ofrece utilizar una base de datos ya hecha es que las grabaciones son de alta calidad, los locutores son de ambos sexos y de edades diferentes, con un número suficiente de locutores para las necesidades requeridas.

En un principio se pensaba en construir una base de datos en español pero por cuestiones de tiempo y falta de equipo adecuado no fue posible.

La base de datos "base de dones des son du français" tiene las locuciones en formato SAM conforme a las recomendaciones CEE-ESPRIT "SAM" No. 2589. A continuación se describe las características de la base de datos utilizada:

32 locutores: 16 hombres y 16 mujeres
Números aislados (del 0 al 99)
Letras
Habla continua
Frases (Fonéticamente balanceadas, silábicas, sonidos nasales)

Para éste trabajo se escogió trabajar con los números en francés como palabras aisladas para el reconocimiento de la palabra y se utilizaron los números del 0 al 9. Para el reconocimiento de locutor se utilizaron números del 0 al 99.

Para efectos de simular el ruido ambiental, se adquirió una base de datos de diferentes tipos de ruido. Los diferentes tipos de ruido fueron adquiridos de la internet. El formato de los archivos de ruido son WAV grabados a 8kHz mono.

Los tipos de ruido son:

Conversación: Conversación de un grupo de personas en un ambiente de restaurante.

Interior de un coche: Ruido grabado en el interior de un coche Volvo en condiciones de lluvia.

Interior de una fábrica: Grabado en el interior de una planta de manufactura.

Blanco

Interior de un tanque militar: Ruido grabado en el interior de un tanque militar en movimiento.

Con éstos tipos de ruido se intenta simular diferentes ambientes y ver si el sistema de reconocimiento de voz es capaz de reconocer en ambientes ruidosos y a diferentes niveles de SNR.

3.4.2 Red Neuronal en Matlab

Para facilitar el trabajo, la red neuronal fue diseñada utilizando el toolbox de redes neuronales de Matlab®.

En la figura 3.2 se muestra el diagrama a bloques del sistema de reconocimiento programado en Matlab. Como señal de entrada tenemos las señales de voz obtenidas de la

base de datos ya mencionada anteriormente. A la señal de voz se le calcula su sonoridad con la herramienta descrita en el capitulo I y el resultado es arrojado en una matriz M_{16x80}. que corresponde a las 16 bandas críticas y a 80 cuadros de la sonoridad calculada. Cada cuadro equivale a la sonoridad calculada en el tiempo en el que la voz se comporta estacionaria, para nuestro caso la trama de voz se dividió en intervalos de 10mseg. Cabe mencionar que si la trama de voz es menor a 800 mseg, la trama es rellenada con ceros.

La matriz M es convertida a un vector L de 1280x1 para adecuarla a la entrada de la red neuronal. Así pues, la red necesita tener 1280 entradas en la capa de entrada.

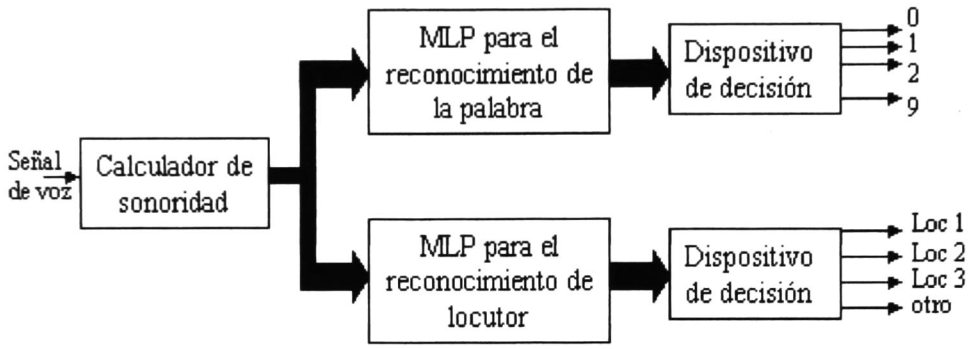


Fig. 3.2 Diagrama a bloques del sistema de reconocimiento de voz

3.4.2.1 Diseño del MLP con Aplicación al Reconocimiento de la Palabra

En primera instancia, se hará el diseño de un MLP con aplicación para el reconocimiento de la palabra. Las palabras a reconocer serán los números en francés del 0 al 9 dicho por 32 locutores diferentes (hombres y mujeres) y perturbadas con diferentes ruidos a diferentes niveles de SNR. De acuerdo a los criterios de diseño de un MLP, el perceptrón multi-capa que se construyó consta de una capa de 1280 entradas, dos capas ocultas de 30 y 15 neuronas respectivamente, puesto que con eso es suficiente para lograr el reconocimiento con un desempeño aceptable. En la tabla 3.1 se encuentran sumarizados los resultados que se hicieron para estructuras con 10, 30 y 50 neuronas en la capa oculta. Cómo podemos ver, la estructura que tiene mejor desempeño es la que tiene 2 capas ocultas, se probó también una estructura de una capa oculta con 100 neuronas pero el tiempo de entrenamiento era demasiado grande y el desempeño no mejoró significativamente. Para la capa de salida, tenemos 10 neuronas que corresponden a los números del 0 al 9. En la figura 3.3 se muestra la estructura del perceptrón que se utilizó.

No. De neuronas en las capas ocultas	No. De secuencias de entrenamiento (por clase)	No. De iteraciones	mse
10 (una capa)	30	5000	0.00232027
30 (una capa)	20	5000	0.00128264
30,15 (dos capas ocultas)	10	5000	0.000795734

Tabla 3.1. Evaluación del desempeño del perceptrón con diferente número de neuronas en la capas ocultas.

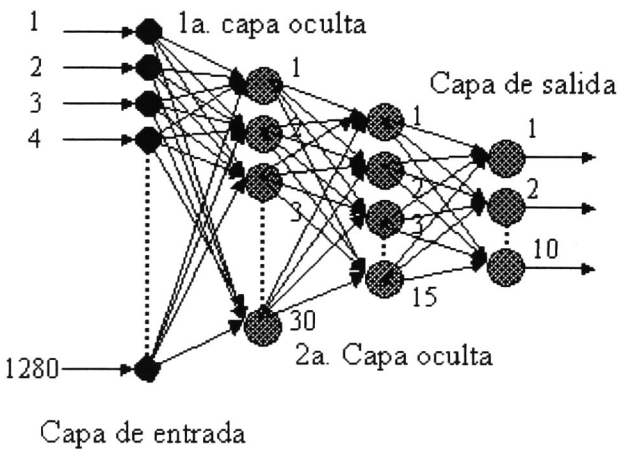


Fig. 3.3. Estructura del perceptrón utilizado para el sistema de reconocimiento de la palabra.

3.4.2.2 Entrenamiento

La red fue entrenada con el algoritmo de retropropagación de gradiente descendente con momento y tasa de aprendizaje variable. Como secuencia de entrenamiento se utilizaron 10 muestras de cada una de las clases, 5 de las muestras fueron presentadas sin ruido y 5 perturbadas por ruido con una SNR de 0dB. Las 10 muestras fueron de 10 de los 32 locutores de la base de datos (6 hombres y cuatro mujeres). Los ruidos utilizados para perturbar las secuencias de entrenamiento fueron: blanco, ambiental, interior de un auto, interior de un tanque militar y ruido de fábrica.

En la figura 3.4 se muestra la función de error durante el entrenamiento. Cabe mencionar que la red también fue entrenada con secuencias de entrenamiento sin ruido. Se hizo una comparación con la red entrenada con 10 secuencias sin ruido y con la red entrenada con secuencias 5 sin ruido y 5 perturbadas con ruido y el desempeño de la red fue casi el mismo, por un 1% tiene mejor desempeño la red entrenada con secuencias perturbadas por ruido así que ésta red fue la que se utilizó.

En la figura 3.4. se muestra la función del error con respecto del número de iteraciones durante el entrenamiento.

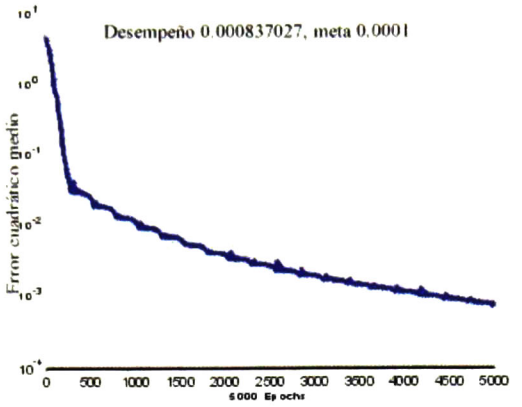


Fig. 3.4. Función del error cuadrático medio durante el entrenamiento.

3.4.2.3 Diseño del MLP con Aplicación al Reconocimiento del Locutor

La estructura interna del perceptrón básicamente fue la misma, es decir, también se utilizaron 2 capas ocultas de 30 y 15 neuronas respectivamente. El sistema debe de ser capaz de identificar a 3 locutores dentro de un grupo de 29. Así pues el número de neuronas de salida será de 4, 3 que corresponden a cada uno de los 3 locutores que queremos identificar y una cuarta que corresponderá a cualquier otra persona.

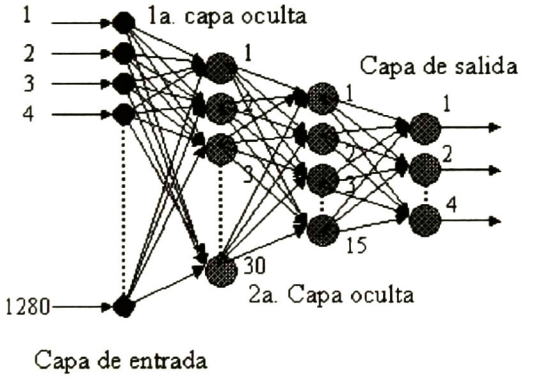


Fig. 3.5. Estructura del perceptrón utilizado para el sistema de reconocimiento de locutor.

3.4.2.4 Entrenamiento

El algoritmo de entrenamiento utilizado fue también el de retropropagación de gradiente descendente con momento y tasa de aprendizaje variable. Como secuencia de entrenamiento se utilizó lo siguiente:

Locutor 1	20 números aleatorios entre el 0 y el 99
Locutor 2	20 números aleatorios entre el 0 y el 99
Locutor 3	20 números aleatorios entre el 0 y el 99
8 locutores que no se desea reconocer	5 números aleatorios entre el 0 y el 99 dichos por cada locutor.

Tabla 3.2. Ejemplos utilizados para entrenar al perceptrón para reconocimiento de locutor.

Cabe mencionar que las secuencias de entrenamiento no fueron perturbadas por ningún tipo de ruido.

En la figura 3.6 se muestra la función de error con respecto del número de iteraciones durante el entrenamiento.

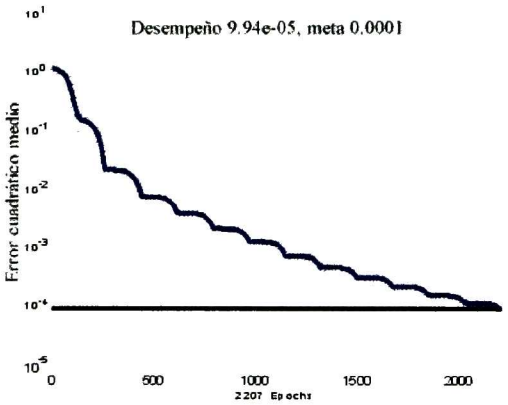


Fig. 3.6. Función del error cuadrático medio durante el entrenamiento del perceptrón dedicado al reconocimiento de locutor.

3.5 Resultados

En ésta sección se hace un análisis de los resultados obtenidos. Para probar el desempeño de la red dedicada al reconocimiento de la palabra se hizo la siguiente prueba:

Números del 0 al 9 sin ruido dicho por 29 locutores (ambos sexos)

Números del 0 al 9 con ruido ambiental a diferentes niveles de SNR, dicho por 29 locutores (ambos sexos).

Números del 0 al 9 con ruido del interior de un auto a diferentes niveles de SNR, dicho por 29 locutores (ambos sexos).

Números del 0 al 9 con ruido de fábrica a diferentes niveles de SNR, dicho por 29 locutores (ambos sexos).

Números del 0 al 9 con ruido del interior de un tanque a diferentes niveles de SNR, dicho por 29 locutores (ambos sexos).

En la tabla 3.2. están sumarizados los resultados obtenidos después de la prueba.

R\RUIDO	CONVERSACIÓN (% de reconocimiento)	INT. DE AUTO (% de reconocimiento)	FABRICA (% de reconocimiento)	INT. TANQUE MILITAR (% de reconocimiento)	SIN RUIDO (% de reconocimiento)
10dB	64%	83%	23%	71%	84%
5dB	56%	83%	17%	64%	X
0dB	45%	82%	14%	57%	X
-5dB	34%	80%	11%	48%	X

3.2. Desempeño de la red, en porcentaje de reconocimiento, para diferentes ambientes.

Para probar el desempeño de la red dedicada al reconocimiento del locutor, se realizó la siguiente prueba:

números del 0 al 99 dichos por el locutor 1
números del 0 al 99 dichos por el locutor 2
números del 0 al 99 dichos por el locutor 3
números del 0 al 99 dichos, cada uno, por 9 locutores diferentes al 1, 2 y 3.
En total se probó con 1200 números y la red hizo un reconocimiento correcto en un 75%.

3.6 Conclusión

Haciendo el análisis de resultados, vemos que no es del todo satisfactorio el desempeño del sistema de reconocimiento. Hay que tomar en cuenta que no se hizo un análisis exhausto a la hora de escoger los ejemplos para entrenar a la red. Otro factor que hay que tomar en cuenta es que, por ejemplo para el reconocimiento de la palabra, las secuencias de entrenamiento fueron de 10 palabras dichas por 10 locutores diferentes (ambos sexos), no se tenía en la base de datos la posibilidad de que un mismo locutor repitiera la misma palabra varias veces. Si escogíamos un número mayor de secuencias de entrenamiento no se iba a poder probar la red debido a que ya no se contaba con mas ejemplos. Es decir, a la hora de hacer las pruebas de desempeño se habría tenido que probar con los ejemplos que se le dio a la red para que aprendiera y no hubiésemos visto si la red estaba generalizando. Con respecto a la robustez al ruido, no resultó lo que se esperaba, al menos no para ambientes de ruido como el de conversación de varias personas. Quizá se le tenga que hacer una modificación a la estructura de la red o al entrenamiento porque visualmente se puede apreciar que las señales bidimensionales barks-tiempo-sonoridad son bastante robustas al ruido.

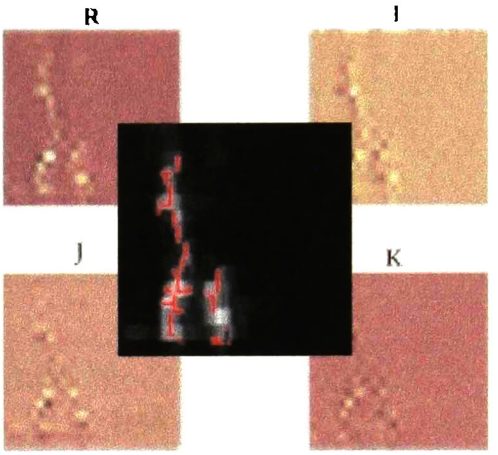
Las redes neuronales únicamente nos dan la oportunidad de reconocer o no reconocer la voz. En otros estudios que han realizado, se ha querido obtener valores intermedios, es decir, que la red pueda indicar con qué tanta exactitud se ha hecho el reconocimiento, pero esto no es posible y debido a esto nuestro estudio no puede llegar más allá que un simple reconocimiento.

CAPÍTULO 4

4 IMÁGENES DE LA SONORIDAD Y SUS TRATAMIENTOS

4.1 Introducción

En el capítulo anterior se mostró una aplicación de las señales de voz basadas en parámetros perceptuales. Ahora queremos llegar más allá explotando las ventajas que nos ofrece trabajar con éste tipo de parámetros. Así pues, surgió la idea de crear una "imagen de la sonoridad", que no es ni mas ni menos que la representación en 2-D de la señal de voz basada en parámetros perceptuales. El objetivo es hacer el análisis de la voz a través de sus imágenes y la idea de tratar a las señales de voz como imágenes surgió debido a que las "imágenes de la sonoridad" presentan una característica muy particular que es su forma, es decir, una misma palabra dicha por diferentes personas presentan una forma muy similar. Otra característica es el efecto que tiene el ruido sobre las imágenes. Las ventajas que esto puede presentar son mostradas a lo largo del capítulo.



4.2 Imágenes de la Sonoridad

La "imagen de la sonoridad" es construida haciendo una representación en 2-D de la sonoridad. Como eje x tenemos el tiempo, como eje y tenemos la tasa de bandas críticas (Barks) y la intensidad corresponde al nivel de sonoridad. En la figura 4.1 se muestra la imagen de la sonoridad de la palabra "zero" en francés.

Haciendo una comparación entre las imágenes de la sonoridad de la palabra "zero", en francés, dicho por 29 personas diferentes, vemos que las imágenes tienen una forma muy particular, que prácticamente podríamos decir que es la forma representativa del número "zero", en comparación con la forma que presentan los demás números. En el apéndice 2 se muestran las imágenes para que se pueda apreciar lo dicho anteriormente.

Así podemos hacer un análisis para cada uno de los números del 0 al 9 y vemos que cada número tiene una forma particular, ya sea por la forma en sí o por las zonas en donde se concentra la mayor parte de la energía.

Debido a éstas observaciones se sugiere, como primera opción, utilizar técnicas morfológicas para el análisis de imágenes de la sonoridad ya que la herramienta de morfología matemática es especialista en el análisis de forma de las imágenes. El objetivo es extraer las características propias de cada número, ya sea la forma o las regiones en donde se concentra la

mayor cantidad de energía. Así pues, procederemos al estudio de las técnicas morfológicas para el tratamiento de imágenes.

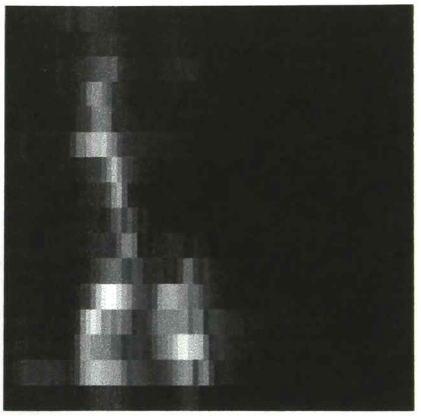


Figura 4.1. Imagen de la sonoridad de la palabra "cero" en francés.

4.3 Análisis Morfológico de las Imágenes de la Sonoridad

La morfología matemática es una herramienta cuyo lenguaje esta basado en la teoría de conjuntos y, como su nombre lo dice, estudia las formas de las imágenes. Se parte de que cualquier imagen es representada mediante un conjunto de pixeles y que se pueden hacer operaciones entre esos conjuntos. Ésta herramienta es muy utilizada para la extracción de componentes de las imágenes que son útiles para la representación y descripción de su forma.

A continuación se describirán algunos de los conceptos teóricos básicos para la comprensión de la teoría de morfología matemática.

4.3.1 Conceptos Teóricos

Como se mencionó anteriormente, la morfología matemática esta basada en la teoría de conjuntos y, para comenzar, partiremos con algunas definiciones básicas para posteriormente pasar a los operadores primitivos de la morfología matemática.

4.3.1.1 Definiciones Básicas

Sean A y B conjuntos en Z^2, con componentes $a=(a_1,a_2)$ y $b=(b_1,b_2)$, respectivamente.

La traslación de A por $x=(x_1,x_2)$, denotada por $(A)_x$, es definida como

$$(A)_x = \{c \mid c = a + x, \quad para \ \ toda \ \ a \in A\} \tag{4.1}$$

La reflexión de B, denotada por B^, es definido como

$$B^{\wedge} = \{x \mid x = -b, \quad para \ \ b \in B\} \tag{4.2}$$

El complemento del conjunto A es

$$A^{c} = \{x \mid x \notin A\} \tag{4.3}$$

La diferencia de dos conjuntos A y B, denotada por A-B, es definida como

$$A - B = \{x \mid x \in A, x \notin B\} = A \cap B^{c} \tag{4.4}$$

4.3.1.2 Operadores Primitivos: Dilación y Erosión

La erosión y dilación son las bases de las cuales se parte para formar muchos de los operadores de la morfología matemática.

Dilación

La dilación está definida como

$$A \oplus B = \{x \mid (B^{\wedge})_{x} \cap A \neq \varnothing\} \tag{4.5}$$

Es decir, los valores de x para los cuáles la intersección de A y la reflexión de B desplazada por x, es un subconjunto de A. A "B" se le conoce como elemento estructurante.

Erosión

La erosión está definida como

$$A \ominus B = \{x \mid (B)_{x} \subseteq A\} \tag{4.6}$$

Es decir, los valores de x para los cuáles el elemento estructurante B, desplazado por x, es un subconjunto de A.

4.3.1.3 Apertura y Cerradura

La apertura y cerradura corresponden a dos de las transformaciones morfológicas más importantes. La apertura es el proceso de erosionar la imagen X con el elemento estructurante B y después realizar una dilación con dicho elemento estructurante. La apertura se define matemáticamente como:

$$X \circ B = (X \ominus B) \oplus B \tag{4.7}$$

La transformación de apertura es muy útil para suavizar contornos y eliminar áreas pequeñas de la imagen X. Ideal para el estudio de la distribución del tamaño del objeto y para resaltar determinadas características de la imagen.

La transformación de cerradura es la transformación dual a la apertura y corresponde al proceso de realizar una dilación a la imagen X por el elemento estructurante B, seguido de una erosión con dicho elemento estructurante. La cerradura se define matemáticamente como:

$$X \bullet B = (X \oplus B)\Theta B \qquad (4.8)$$

Esta transformación es muy útil para cerrar canales pequeños en la imagen X.

Otra transformación de gran importancia es la esqueletización que equivale a obtener una representación regenerativa de determinado objeto.

4.3.2 Análisis de las imágenes

Habiendo observado las imágenes, vemos que el tratamiento que se le dará a cada imagen depende de la aplicación que le demos a las imágenes resultantes. Por el momento se presentan tres opciones de aplicación que pueden ser:

Reconocimiento de la palabra (por diferentes locutores) en condiciones ambientales optimas, es decir que no hay presencia de ruido.

Reconocimiento de la palabra (por diferentes locutores) en condiciones ambientales de ruido.

Evaluación de la degradación de la información en la imagen cuando ésta se encuentra perturbada por ruido.

A continuación se presentan las observaciones hechas a las imágenes:

Las imágenes de la sonoridad de una misma palabra, dicha por diferentes locutores, presentan una forma similar.

Lo primero que se pierde en la imagen, cuando hay presencia de ruido, son las consonantes. Esto es debido a que tienen un contenido bajo de energía y el ruido fácilmente las enmascara.

Las invariantes de las imágenes se presentan en un nivel intermedio en la escala de grises.

Las zonas de energía mas altas de la forma que presentan los números pueden ser ligeramente visibles hasta bajas relaciones de señal a ruido. La SNR depende del ruido con el que se este simulando pero, para un ruido de conversación, tenemos que aún se puede ver la forma del número cuando éste es perturbado a una SNR de –15dB. Y para el ruido del interior de un coche, la forma se mantiene aún cuando la SNR es de –45dB. Estas zonas de energía corresponden a las vocales.

Para la aplicación de reconocimiento de la palabra independiente del locutor, en condiciones ambientales óptimas, trataremos de resaltar algunas características únicas de cada palabra. En primera instancia observamos que, lo que más caracteriza a una misma palabra, es su forma. Así pues el primer tratamiento que se dará a las imágenes es tratar de extraer la forma típica de cada palabra. Para hacer esto se tomaran en cuenta los niveles bajos de sonoridad, que corresponden a las consonantes.

Primero se convierte la imagen a escala de grises y se umbraliza la imagen para quedarnos con los niveles que describen la forma de cada palabra, y esto es a un nivel de gris

de 15. Posteriormente, se realiza la operación de cerradura para resaltar la forma y eliminar todos los huecos que pueda tener en el centro. Como elemento estructurante se utilizó una imagen de disco de radio 2. Cabe mencionar que, todas las vocales a este nivel presentan la misma forma así que para extraer las características de cada una de las vocales se realizó otra umbralización de la imagen original pero ahora a un nivel de gris de 120, y también se le hizo una cerradura. Por último se le aplica una operación de negación a la imagen umbralizada a un nivel de 15, y se hace una unión con la imagen umbralizada al nivel de 120. Para resaltar las características de las vocales se realiza una apertura a la imagen resultante. Como resultado se obtiene una imagen que describe la forma de cada palabra. En la figura 4.2 se muestra la imagen original y la imagen procesada descriptiva de la forma. Para ver las imágenes de los diferentes números dichos por diferentes locutores referirse al apéndice 2.

a) b)

Fig. 4.2. a) Imagen de la sonoridad. b) Forma descriptiva de la imagen.

Para realizar el reconocimiento se puede utilizar una red neuronal como la que se describió en el capítulo II y los resultados debieran de ser mucho mas satisfactorios que los que se obtuvieron utilizando las imágenes de la sonoridad sin ningún tipo de tratamiento. Esto suena lógico ya que con las imágenes tratadas se han resaltado las características que describen más a cada palabra.

Como segunda aplicación tenemos el reconocimiento de la palabra independiente del locutor pero en condiciones de ambientes ruidosos. Para ésta aplicación, es obvio que no podemos tratar a las imágenes de la sonoridad de la misma manera como se hizo anteriormente, puesto que el ruido se hace presente aún cuando la SNR es relativamente alta, como por ejemplo 40dB.

Para poder tener una forma que describa a las palabras aún cuando la SNR sea pequeña, nos tenemos que basar en las zonas en donde la energía sea suficientemente grande o que se vea invariante aún cuando se tiene una SNR muy baja, pero que estas zonas sigan describiendo a la palabra.

Para realizar esto se hizo una umbralización a la imagen original en escala de grises a un nivel de 100, así también como las operaciones realizadas en el análisis anterior. Con esto las invariantes permanecen hasta un nivel de SNR de 0dB. En la figura 4.3 se muestran los resultados obtenidos.

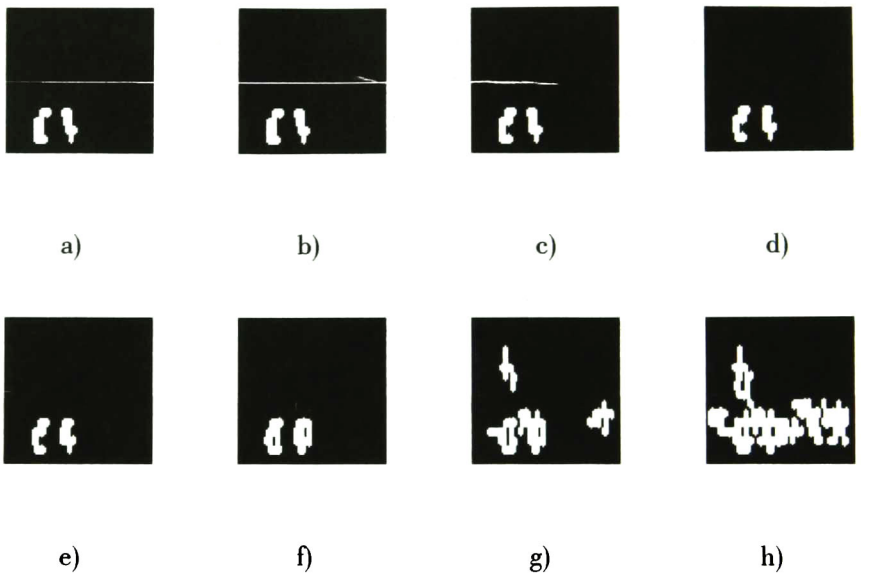


Fig. 4.3. Extracción de las invariantes de la palabra en francés "zero". a) Sin ruido. b) SNR=40dB. c) SNR=30dB. d)SNR=20dB. e) SNR=10dB. f) SNR=0dB. g) SNR=-5dB. h) SNR=-10dB.

Cabe mencionar que para éste análisis se pierden completamente las consonantes. Para el reconocimiento de los números, debido a que cada uno presenta diferentes invariantes, el reconocimiento puede ser correcto hasta bajos niveles de SNR, como lo es hasta –5dB. Pero si utilizamos palabras en las que solo varíen las consonantes como podrían ser: "pez", "res", seguramente el reconocimiento será erróneo puesto que las invariantes corresponderían únicamente a la vocal "e".

Se podría decir que hay un nivel de SNR a partir de la cual las consonantes ya se pierden debido a que son enmascaradas por el ruido. Ahora, el ser humano es capaz de reconocer la palabra debido a que hace una discriminación de vocabulario, es decir, si se le dice a una persona que la palabra que va a escuchar es un número y está entre el cero y el nueve, aunque la persona escuche solo "e o", sabrá que era el "cero". Esto significa que éste tratamiento puede ser de mucha ayuda para realizar el reconocimiento de la palabra en ambientes bastante ruidosos, y que el vocabulario sea limitado y lo suficientemente separado para que no hayan confusiones debido a la degradación de las consonantes. Para ésta prueba se utilizó un umbral muy alto para poder detectar las invariantes hasta bajos niveles de SNR, pero en un ambiente real, si una persona se comunica con una máquina a través de un micrófono seguramente la SNR no sería tan baja como la que se presentó en ésta prueba.

Otro punto importante es que cada una de las palabras debe de tener su propio tratamiento morfológico. Esto es debido a que, por ejemplo la palabra en francés "zero", presenta áreas conectadas verticalmente y para detectar estas áreas se utiliza una línea a 90° como elemento estructurante. Si utilizamos el mismo elemento estructurante para la palabra

en francés "due", se perderá mucha información debido a que ésta presenta áreas conectadas en forma horizontal.

Un tercer análisis que se puede realizar es el de ver que pasa con el esqueleto de una determinada palabra cuando ésta es perturbada por ruido. Una herramienta muy utilizada para la descripción de la forma de una imagen es la de esqueletización. Con ésta herramienta y aplicando un tratamiento adecuado se pretende ver que tanto se degrada o deforma el esqueleto de una imagen perturbada por ruido con respecto a la imagen original.

Fig. 4.4. a) Imagen de la sonoridad de la palabra en francés "zero". b) Esqueleto de la imagen a.

Para obtener el esqueleto de una imagen, en primera instancia, se realizó una umbralización a la imagen a un nivel de 75 en escala de grises. Esto fue para obtener los niveles de energía que describen, de una manera mas clara, la forma de la imagen. Posteriormente se realizó una apertura de área para eliminar las regiones de área menor a 5 pixeles, esto con la intención de dejar solo las componentes conectadas y eliminar el ruido. En seguida se hizo una apertura para resaltar la forma de la imagen y eliminar los huecos que esta pudiera tener o las variaciones pequeñas. Aquí como elemento estructurante se utilizó una línea de longitud 5 girada 90 grados, esto con la intención de no perder las componentes conectadas verticalmente. Para finalizar, el esqueleto se obtuvo con la función "mmthin" del toolbox de morfología matemática de Matlab® que corresponde a una operación de adelgazamiento. Los resultados son mostrados en la figura 4.4.

Una prueba interesante es ver que pasa con el esqueleto de una imagen cuando ésta es perturbada por ruido. Para saber que tanta diferencia hay entre dos esqueletos, uno que corresponde a una imagen que no se encuentra perturbada por ruido y otro que corresponde a la imagen perturbada por ruido, se realiza una operación de intersección entre los dos esqueletos. Los puntos que queden en la imagen resultante serán los puntos que se encuentran en los dos esqueletos. Con esto podríamos ver, en cierta forma, la degradación del esqueleto en función del ruido. Los resultados a ésta prueba se muestran a continuación.

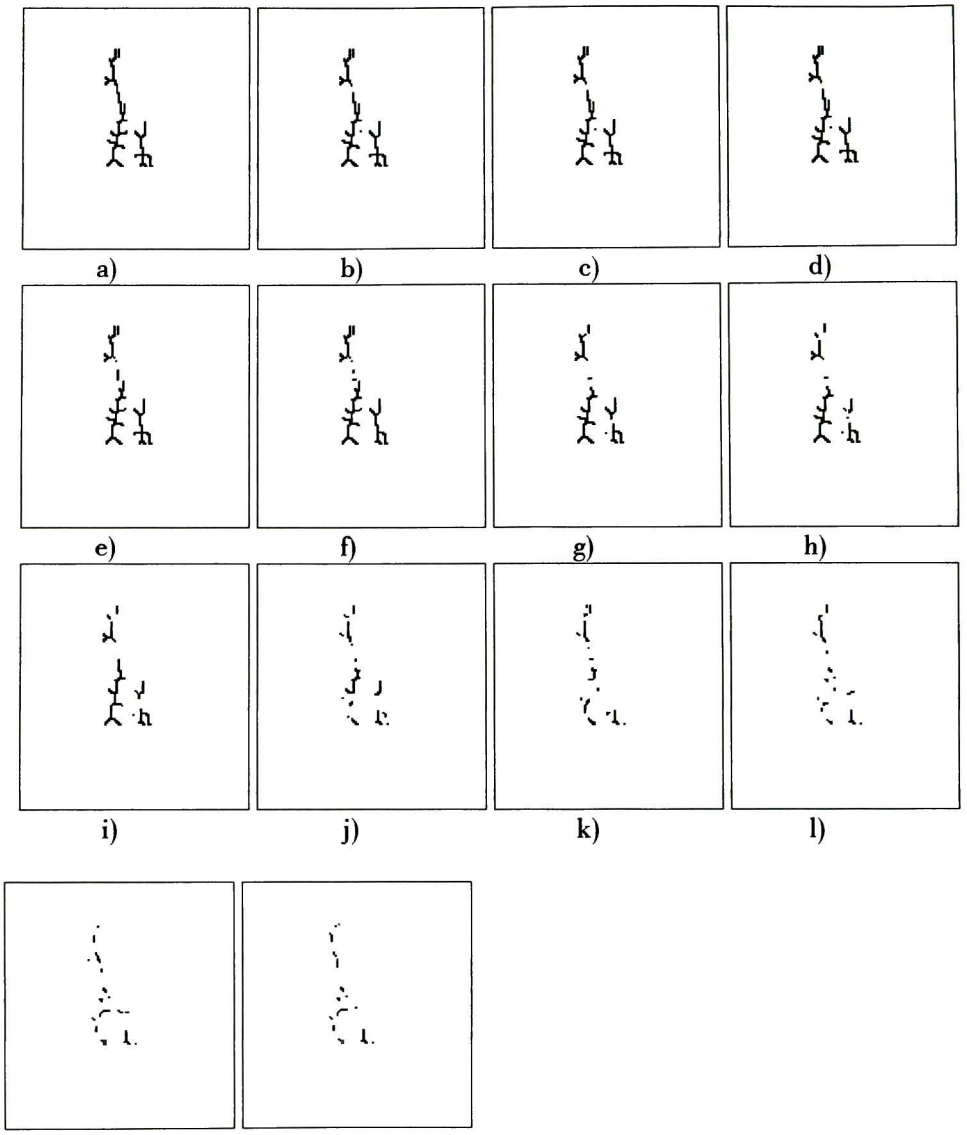


Fig. 4.5. Degradación del esqueleto de la imagen de la palabra en francés "zero" en función del ruido. a) Esqueleto de la imagen sin ruido. Las imágenes de la b) a la n) corresponden al resultado de la intersección entre el esqueleto original y el esqueleto obtenido cuando la imagen es perturbada por ruido a diferentes niveles de SNR. b) SNR=40dB, c) SNR=35dB, d) SNR=30dB, e) SNR=25dB, f) SNR=20dB, g) SNR=15dB, h) SNR=10dB, i) SNR=5dB, j) SNR=0dB, k) SNR=-5dB, l) SNR=-10dB, m) SNR=-15dB, n) SNR=-20dB.

Como podemos apreciar, el esqueleto se va degradando conforme aumenta el nivel de ruido. Algo importante que hay que señalar es que el esqueleto se degrada pero no de forma progresiva, es decir, algunas partes del esqueleto pueden estar mas completas cuando el nivel de SNR es de –10dB que cuando el nivel de SNR es de –5dB.

4.4 Análisis de las Imágenes de la Sonoridad con la Transformada de Fourier Cuaterniónica Discreta (DQFT)

Hasta el momento hemos realizado un análisis de las imágenes de la sonoridad con técnicas morfológicas para el tratamiento de imágenes. Ahora procederemos a estudiar a las imágenes desde otro punto de vista haciendo un análisis de tipo multi-resolución y analizaremos los resultados. Así pues, procederemos al estudio de las imágenes de la sonoridad con la transformada cuaterniónica de Fourier.

Los filtros de Gabor han sido muy utilizados en el análisis de imágenes debido a la ventaja que tienen de ser localizados óptimamente y simultáneamente en el dominio del tiempo y la frecuencia [7]. Algunas de las aplicaciones están en reconocimiento de patrones, clasificación, análisis de textura y estimación de frecuencia y fase local.

En 2-D, las componentes de fase son las que portan la mayor información sobre la imagen [7]. Debido a esto, surge la motivación de utilizar un filtro de Gabor para la estimación de la fase local, pero más aún, para tener mayor información sobre la fase, se utilizará un filtro de Gabor cuaterniónico.[1]

4.4.1 Conceptos Teóricos

A continuación se presentan los conceptos teóricos básicos como lo son la definición de los cuaterniones, sus propiedades, etc., para después continuar con la definición de la Transformada de Fourier Cuaterniónica Discreta y el filtro de Gabor cuaterniónico.

4.4.1.1 Cuaterniones

Los cuaterniones surgen por la idea de extender los números complejos C y fueron desarrollados por el físico William Rowan Hamilton. El sistema numérico de los cuaterniones ha sido aplicado en gráficos e investigación robótica debido a que pueden ser utilizados en el control de las rotaciones en el espacio tri-dimensional.

El cuaternión usualmente es escrito como,

$$q = a + bi + cj + dk \tag{4.9}$$

donde a, b, y c son valores escalares, e i, j, y k son cantidades imaginarias definidas como,

$$i^2 = j^2 = k^2 = -1 \tag{4.10}$$

[1] Para mayor información sobre el filtro de Gabor cuaterniónico favor referirse a [6] y [7].

y tienen la propiedad de que

$$ij = k, \quad jk = i, \quad ki = j, \quad ji = -k, \quad kj = -i, \quad ik = -j \tag{4.11}$$

La suma, multiplicación y división de los cuaterniones se definen enseguida.

Sean $q_1 = a_1 + b_1 i + c_1 j + d_1 k$, y $q_2 = a_2 + b_2 i + c_2 j + d_2 k$ cuaterniones,

$$q_1 + q_2 = (a_1 + a_2) + (b_1 + b_2)i + (c_1 + c_2)j + (d_1 + d_2)k \tag{4.12}$$

y

$$q_1 q_2 = (a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2) + (a_1b_2 + a_2b_1 + c_1c_2 - d_1c_2)i + (a_1c_2 + a_2c_1 + d_1b_2 - b_1d_2)j + (a_1d_2 + a_2d_1 + b_1c_2 - c_1b_2)k \tag{4.13}$$

Cabe mencionar que la multiplicación de los cuaterniones no es conmutativa.

$$\frac{q_1}{q_2} = q_1 q_2^{-1} \quad donde \tag{4.14}$$

$$q_2^{-1} = \left(\frac{a}{|q_2|}, -\frac{b}{|q_2|}, -\frac{c}{|q_2|}, -\frac{d}{|q_2|}\right), \quad y \tag{4.15}$$

$$|q_2| = \sqrt{a^2 + b^2 + c^2 + d^2} \tag{4.16}$$

El conjugado de un cuaternión

$q = a + bi + cj + dk$

está definido como,

$q^* = a - bi - cj - dk.$ (4.17)

4.4.1.2 Transformada de Fourier Cuaternónica Discreta (DQFT)

La transformada de Fourier cuaterniónica discreta está definida como

$$F_{quv} = \sum_{m=0}^{M-1} \sum_{n=}^{N-1} e^{\left(\frac{-i2\pi um}{M}\right)} fmne^{\left(\frac{-j2\pi vn}{N}\right)} \tag{4.18}$$

y la transformada de Fourier cuaterniónica inversa esta dada por,

$$f_{mn} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=}^{N-1} e^{\left(\frac{i2\pi um}{M}\right)} Fquv^{\left(\frac{j2\pi vn}{N}\right)} \tag{4.19}$$

4.4.1.3 Filtro de Gabor Cuaterniónico

Los filtros de Gabor están definidos como filtros lineales invariantes al desplazamiento con las funciones de base de ventana Gaussiana de la transformada de Fourier como sus funciones base.

El filtro de Gabor complejo (1-D) tiene una respuesta al impulso definida como

$$h(x;N,u_0,\sigma) = g(x;N;\sigma)e^{(i2\pi u_0 x)} \qquad \text{donde} \tag{4.20}$$

$g(x;N;\sigma)$ es la función Gaussiana $\qquad g(x;N;\sigma) = Ne^{-\frac{x^2}{2\sigma^2}}$ (4.21)

Los filtros de Gabor tienen como parámetros la constante de normalización N, la frecuencia central u_0 y la varianza σ de la función Gaussiana.

El filtro de Gabor cuaterniónico tiene respuesta al impulso

$$h^q(x;u_0;\sigma;\varepsilon) = g(x;\sigma;\varepsilon)e^{(i2\pi u_0 x)}e^{(j2\pi v_0 y)} \tag{4.22}$$

Y la transformada de Fourier cuaterniónica de la respuesta al impulso del filtro cuaterniónico de Gabor está dada por

$$h^q(x;u_0,\sigma,\varepsilon) \xrightarrow{H} H^q(u;u_0,\sigma,\varepsilon) = e^{\left(\frac{-2\pi^2\sigma^2\left[u-u_0\right]^2}{\varepsilon^2}\right)} \tag{4.23}$$

4.4.2 Análisis de las Imágenes

Como primer análisis, nos interesa saber que es lo que pasa con determinada imagen de la sonoridad que es perturbada por ruido a diferentes niveles de SNR. Así pues, se tomó la imagen de la sonoridad de la palabra "zero" en francés como imagen de prueba.

El ruido con el que se perturbo la señal del sonido fue con el ruido de conversación y los niveles de SNR van desde los 40dB hasta –20dB.

Realizando la convolución de una señal 2-D real con un la respuesta al impulso de un filtro 2-D cuaterniónico de Gabor, nos da como resultado una señal 2-D analítica cuaterniónica. Cabe mencionar que el análisis cuaterniónico de las imágenes de la sonoridad fue realizado con la herramienta del Dr. Eduardo Bayro-Corrochano. De las imágenes resultantes, vemos que las cuatro componentes *real*, *i*, *j* y *k* obtenidas después de filtrar la señal con un filtro de Gabor, nos da información de mucho interés para nosotros.

Haciendo una observación entre todas las pruebas realizadas, vemos que al descomponer la imagen 2-D en una señal 2-D cuaterniónica, las imágenes resultantes nos permiten ver ciertas características propias de la imagen y lo mejor aún es que, dichas características son los elementos invariantes de la imagen de la sonoridad cuando ésta es perturbada por ruido. Cabe mencionar que cada componente extrae ciertas características de la imagen de la sonoridad.

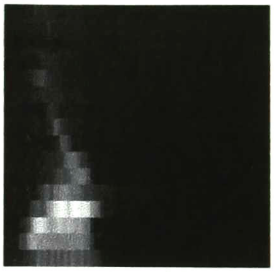


Figura 4.6. Imagen de la sonoridad de la palabra "zero" en f rancés dicha por el locutor 12.

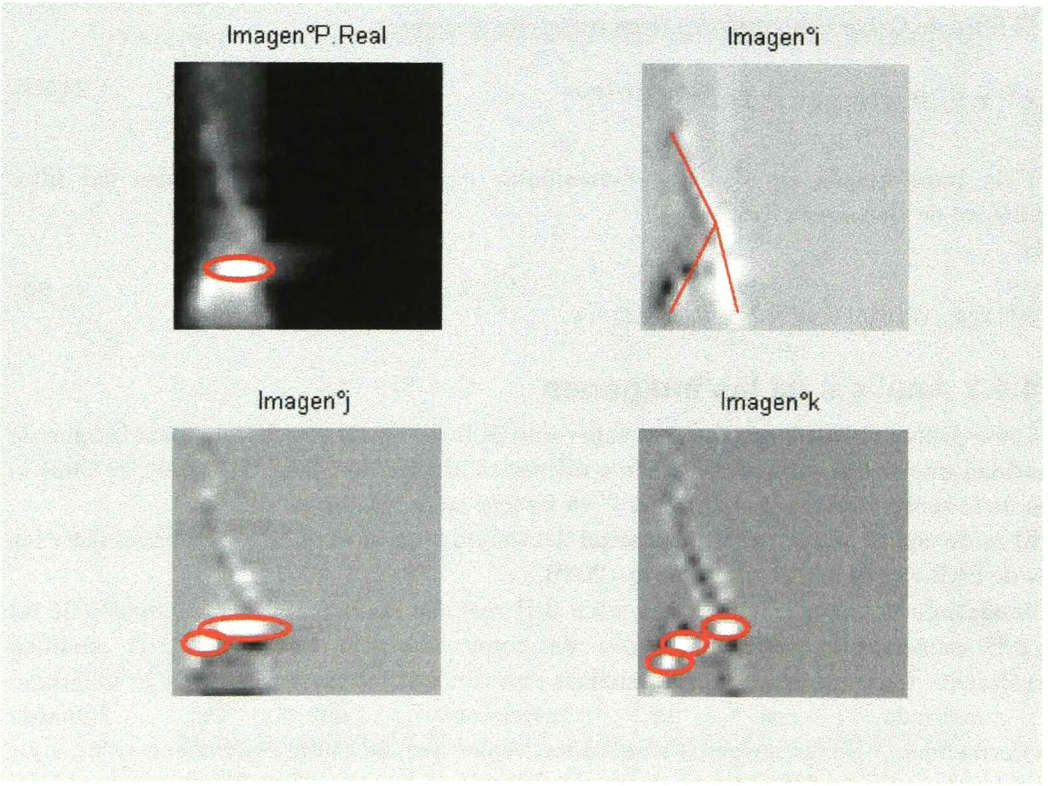


Figura 4.7. Componentes real, i, j y k de la imagen de la sonoridad mostrada en la figura 4.6. Los círculos muestran las componentes invariantes de la imagen.

En la figura 4.6 se muestra la imagen de la sonoridad de la palabra "zero" en francés dicha por el locutor 12 de la base de datos, y en la figura 4.7 se muestra la descomposición de la señal 2-D en sus cuatro componentes, *real*, *i*, *j*, y *k*. Los círculos indicados en las imágenes muestran los elementos invariantes de las imágenes aún cuando éstas presentan perturbación

por ruido. Para ver las imágenes de la sonoridad y la descomposición cuaterniónica de las imágenes probadas a diferentes niveles de SNR, favor de referirse al apéndice 2.

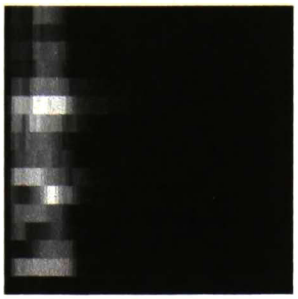


Figura 4.8. Imagen de la sonoridad de la palabra "un" en francés dicha por el locutor 1.

En la figura 4.9 se muestra la descomposición cuaterniónica de la palabra "un" en francés. De la misma manera, los círculos muestran las características invariantes de las imágenes cuando éstas presentan perturbación por ruido.

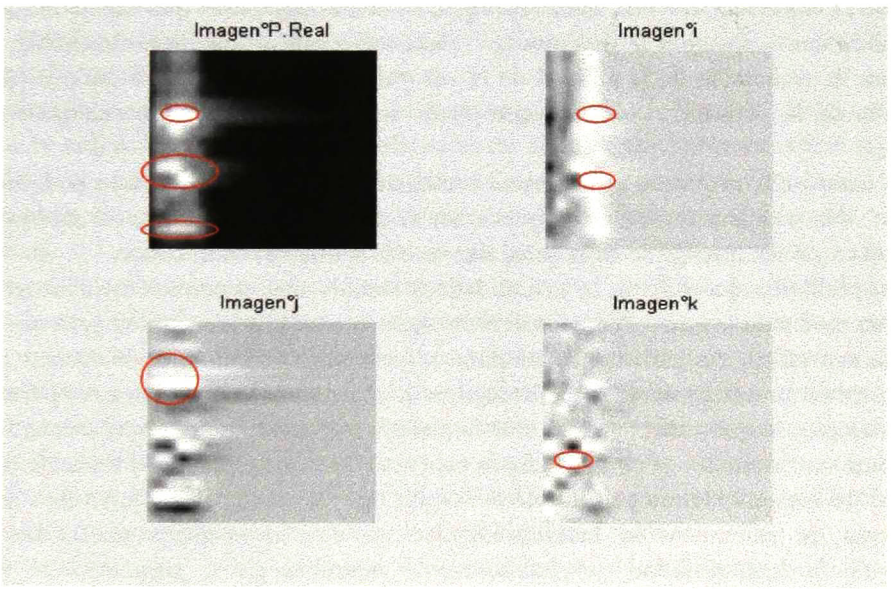


Figura 4.9. Componentes real, i, j y k de la imagen de la sonoridad mostrada en la figura 4.8. Los círculos muestran las componentes invariantes de la imagen.

4.5 Estudio Sobre La Medición Objetiva De La Calidad En Los Sistemas de Transmisión

En éste sub-tema se hace un estudio sobre los métodos de evaluación de la calidad en los sistemas de transmisión. Se da la diferencia que hay entre los conceptos de "calidad del habla" e "inteligibilidad del habla" El objetivo es hacer un estudio sobre los métodos

existentes, tanto subjetivos como objetivos, para conocer el estado actual de los sistemas de evaluación y, en trabajo futuro, tratar de aplicar los resultados obtenidos en la sección anterior para aportar un factor más para la evaluación de la calidad del habla y obtener un índice de correlación más alto a los actuales.

4.5.1 Necesidad de la Evaluación de la Calidad de los Sistemas de Transmisión

Cuando se desea realizar una llamada telefónica, uno espera poder no solo entender el mensaje que se esta recibiendo, sino también poder reconocer a la persona con la que estamos hablando. Cada vez los usuarios somos mas exigentes y queremos servicios de mejor calidad, sin embargo, debido a las necesidades de velocidades de transmisión cada vez mas altas para poder mandar por un solo cable: voz, datos, imagen, etc. el ancho de banda requerido es muy grande. Para poder cumplir con las necesidades del usuario y utilizar las tecnologías actuales de transmisión de datos, es necesario ocupar el menor ancho de banda posible. Para lograr esto se utilizan técnicas de codificación de voz, datos y video, sin embargo al utilizar estas técnicas, puesto que hay cierta perdida de información, la calidad de las señales transmitidas disminuye.

En un codificador de voz, como ejemplo, es necesario saber que tan parecida es la señal de salida con respecto a la de entrada. Para realizar este tipo de evaluaciones existen métodos para la evaluación de la calidad de la voz y éstos son de carácter subjetivo (depende de la opinión de la persona) u objetivo que puede ser realizado mediante cualquier sistema automático.

Las compañías necesitan realizar una evaluación de sus productos para poderlos sacar al mercado. Un codificador de voz no puede ser sacado al mercado sin antes saber con que tanta calidad o que tan inteligible es la señal de voz a la salida del codificador.

Otro problema que hay en la actualidad es cuando nos queremos comunicar con las máquinas por medio de la voz. Podemos decirle algún comando a la máquina para que realice alguna tarea específica, sin embargo la máquina no responde a la instrucción dada. Aquí es en donde entra un problema de diferencias lógicas que hay entre la máquina y el operario. El operario quizá piense que está haciendo todo lo posible por pronunciar bien el comando y aún así la máquina no responde. Por otro lado, la máquina quizá no entiende el mensaje debido a que la señal de voz esta siendo perturbada por algún tipo de ruido. Una forma de solucionar éste problema de comunicación hombre-máquina es que la máquina pueda hacer una evaluación de la inteligibilidad del habla, así la máquina puede responder al operario indicando, según el porcentaje de inteligibilidad, si hay demasiado ruido en el ambiente y necesita que le digan mas fuerte el comando o la pronunciación de la palabra no es la adecuada. También en los sistemas de reconocimiento de voz a veces es conveniente tener un índice de seguridad, por así decirlo, de que tan segura está la máquina de haber reconocido el comando que se le ha dado, por medio de éste porcentaje la máquina puede pedir al operario que repita el comando y así evitar que la máquina cometa errores por no haber entendido un comando correctamente.

Para el caso anterior forzosamente necesitaríamos que el sistema haga la evaluación de la inteligibilidad del habla, es decir, la evaluación sería de carácter objetivo. Más delante

estudiaremos los métodos subjetivos y objetivos más utilizados actualmente y veremos las ventajas y desventajas que presentan cada uno.

Hasta ahora hemos hablado de codificadores de voz, interfase hombre-máquina en las que quizá si no es entendido completamente el mensaje, simplemente o es una palabra no comprendida o la máquina realiza la tarea equivocada. Pero, ¿qué pasa por ejemplo en un combate de guerra?, en donde las condiciones ambientales son extremadamente ruidosas. Al piloto de un avión no le interesa si el mensaje que está recibiendo es de buena calidad o no, lo que más interesa es poder entender el mensaje aunque no se reconozca de que persona viene.

Para esto se necesita que los cascos de los pilotos sean evaluados, de una manera exhaustiva, en condiciones extremadamente ruidosas y así poder decidir si son o no aptos para la comunicación en combates de guerra. En este tipo de aplicaciones, si no se tiene una evaluación de la calidad de los equipos de transmisión, los efectos pueden ser catastróficos.

Así como estas aplicaciones en donde existe la necesidad de la evaluación de la calidad del habla, existen muchas más como pueden ser evaluar la inteligibilidad en una sala de conferencias, teatro, sala de conciertos, etc.

4.5.2 Métodos para la Evaluación de la Calidad del Habla

En la actualidad existen muchos métodos para la evaluación de la calidad de la voz. Como ya se mencionó anteriormente existen métodos de carácter subjetivo y métodos de carácter objetivo. En los métodos subjetivos el juicio lo dan las personas y en los métodos objetivos el juicio es dado por una máquina.

Hasta el momento hemos hablado de la evaluación de la calidad del habla pero, ¿Cómo se define la calidad del habla?. La definición de calidad del habla es difícil darla en una forma simple. La palabra "calidad" implica una evaluación estética de la voz. Muchas mediciones de la calidad de voz buscan determinar las preferencias del locutor, usualmente preguntando al locutor que extienda un juicio que va desde, el más preferido, hasta el menos preferido. En aplicaciones para la evaluación de la voz artificial, la calidad del habla se refiere a la naturalidad de la voz y el rango va desde lo más natural hasta lo menos natural. En otras aplicaciones como la evaluación de la calidad de los sistemas de transmisión como lo son los teléfonos, vocoders, etc., el juicio de la calidad de la voz se basa en factores como el reconocer al hablante, si se entiende correctamente el mensaje, presencia de ruido en los intervalos de silencio, etc. En general podemos decir que para la evaluación de la calidad del habla intervienen tres factores que son: naturalidad, comprensibilidad y preferencia. Estos factores son tomados en cuenta dependiendo de el sistema a evaluar.

Otro término que se encuentra en gran medida, en el contexto de la evaluación de la calidad de voz, es el de "inteligibilidad del habla" A menudo cuando se habla de calidad de voz, también se habla de la inteligibilidad y la diferencia entre una y otra no es muy clara.

La inteligibilidad del habla es un índice que nos indica el número de palabras que se entendieron correctamente dentro de un determinado número de palabras que se presentan a un locutor, está directamente relacionada con la articulación de las palabras y se dice que la inteligibilidad está dada por las consonantes[53]. También definen la inteligibilidad como el grado o nivel de entendimiento de una palabra.

La relación que existe entre calidad del habla e inteligibilidad del habla es un poco compleja. Algunos escritores como Nakatani y Dukes en 1972[76], han descrito la

inteligibilidad como uno de muchos atributos los cuales afectan la calidad del habla. Otros autores presentan que la calidad del habla está directamente relacionada con su inteligibilidad.

Las pruebas de inteligibilidad han sido comúnmente usadas en la evaluación de la calidad de la voz y asumen que la calidad y la inteligibilidad son lo mismo, pero esto no va de acuerdo con un grupo de pruebas que Nakatani y Dukes hicieron las cuales llamaron "métodos de escalamiento subjetivo" en los cuales se les hicieron preguntas a los escuchas sobre la naturalidad, comprensibilidad y preferencia. Usando estas pruebas, los sistemas que daban resultados de inteligibilidad iguales a menudo daban resultados de calidad diferentes.

En 1980, Voiers examinó las interdependencias entre calidad del habla e inteligibilidad medido por su método de "Prueba de diagnóstico de aceptabilidad", (DAM, por sus siglas en inglés) y la "prueba de diagnóstico por rima", y encontró que la aceptabilidad o calidad total depende en gran medida, pero no totalmente a la inteligibilidad.

Así pues, podemos decir que el hacer una evaluación de la calidad del habla o una evaluación de la inteligibilidad del habla depende del sistema que se necesita evaluar.

A continuación se presentan algunos de los métodos subjetivos y objetivos utilizados en la actualidad, tanto para medir la calidad del habla como su inteligibilidad.

4.5.2.1 Métodos subjetivos

En éste apartado estudiaremos los métodos para la evaluación de la inteligibilidad del habla así también como los métodos para la evaluación de la calidad del habla.

4.5.2.2 Métodos para la Evaluación de la Inteligibilidad del Habla

Los métodos subjetivos o estadísticos, como también se les conoce, se plantearon en 1910 (para la medición de la inteligibilidad del habla) y se perfeccionaron con la introducción del teléfono y el advenimiento de los sistemas de comunicación en la segunda guerra mundial. Este tipo de métodos son los más exactos y confiables que hay en la actualidad, en cuanto a evaluación de la inteligibilidad se refiere.

Para llevar a cabo el proceso del método estadístico para la evaluación de la inteligibilidad, se tienen que tomar en cuenta las siguientes consideraciones:

Los oradores deben de ser oradores entrenados

Deben de tener un vocabulario fluido y que pronuncien listas estandarizadas de palabras a través de sistemas de comunicación

Los escuchas deben de tener una capacidad auditiva fluida de acuerdo a un estándar

Las personas que dirigen las pruebas deben de ser altamente especializadas en el área

Mínimo deben de ser cinco oradores y cinco escuchas

Todos deben de hablar el mismo idioma

Deben de pertenecer a la misma región

Algunas de las listas de palabras especiales para la evaluación de la inteligibilidad del habla se muestran a continuación:

La prueba de diagnóstico por rima modificada (MRT, por sus siglas en inglés): La lista consiste de 6 conjuntos de 50 palabras monosilábicas que difieren, la primera mitad, en la consonante inicial y la segunda mitad en la consonante final.

La prueba de diagnóstico por rima (DRT. por sus siglas en inglés): Consiste de 96 pares de palabras monosilábicas que difieren únicamente de la consonante inicial.

Lista de palabras fonéticamente balanceadas (PB, por sus siglas en inglés): Los materiales de prueba consisten en 20 listas de 50 palabras fonéticamente balanceadas.

4.5.2.3 Métodos para la Evaluación de la Calidad del Habla

4.5.2.3.1.1 Medida de diagnóstico de aceptabilidad (DAM, por sus siglas en inglés)

El DAM fue desarrollado en Dynastat como un método para la medición subjetiva de la calidad o aceptabilidad de los sistemas de comunicación de voz. Éste método combina un acercamiento directo e indirecto a la evaluación de la calidad de voz. De esto, el escucha tiene la oportunidad de indicar, en forma directa, no solamente que tan aceptable es una muestra de voz, sino también indicar que factores de calidad son presentes en la muestra sin olvidar de cómo puede afectar su evaluación de aceptabilidad.

4.5.2.3.1.2 Calificación de la Opinión Media (MOS)

El MOS es probablemente el método más usado para la evaluación de la calidad de la voz. Consiste de una prueba para medir la aceptabilidad o calidad de la voz sobre un sistema de comunicaciones. El MOS pide al escucha que evalúe la calidad (en todos los aspectos) de un sistema de comunicaciones o realizar la evaluación sobre una escala de 5 categorías (excelente, bueno, aceptable, pobre y malo), conocidas como calificación de categoría absoluta (ACR, por sus siglas en inglés), para propósitos de comunicación telefónica. También es utilizada una escala DMOS, que corresponde a lo opuesto del MOS. En la tabla 5.1 se muestran los puntos a evaluar en una prueba MOS y DMOS.

Escala	MOS (ACR)	DMOS (DCR)
5	Excelente	Inaudible
4	Bueno	Audible pero no molesta
3	Aceptable	Poco molesto
2	Pobre	Molesto
1	Malo	Muy molesto

Tabla 4.1. Escalas usadas en el MOS y DMOS.

4.5.3 Métodos Objetivos

Los métodos objetivos son realizados por máquinas y a continuación se describen algunos de los métodos utilizados en la actualidad.

4.5.3.1 Métodos para la Evaluación de la Inteligibilidad del Habla

4.5.3.1.1 %ALcons

El %ALcons es un porcentaje de pérdida de Articulación de Consonantes. Ésta medición de inteligibilidad está estrechamente asociada con el analizador TEF (Tiempo-

Energía-Frecuencia). Se calcula a partir de mediciones de la Relación de sonido directo a reverberante y del tiempo de caída temprano.

4.5.3.1.2 Relación de sonido directo a reverberante

Es la relación entre las intensidades del sonido directo y de la reverberación. Hay varias mediciones para esta cantidad. La norma C50, presentada como una de las más populares, expresa la claridad del habla como la proporción entre la energía de los primeros 50mseg. De sonido directo y la energía total de la reverberación.

4.5.3.1.3 Relación de sonido útil a destructivo

Es el logaritmo entre la energía de los sonidos que benefician la inteligibilidad y aquellos que son perjudiciales para ella y es expresado en dB.

4.5.3.1.4 Relación de energía sonora temprana a tardía

Similar al C50 pero es aplicada al habla e incorpora mediciones en más de una banda de frecuencias.

También existen otros métodos para la inteligibilidad del habla como lo son: el índice de articulación (AI), el índice de transmisión del habla (STI), el índice de rapidez de transmisión del habla (RASTI) y el índice de inteligibilidad del habla (SII). De éstos el más utilizado es el SII y fue propuesto como proyecto preliminar para el estándar ANSI S3.5-1997. El SII tiene la ventaja de utilizar bandas frecuenciales como lo son: bandas críticas(21 bandas), bandas de un tercio de octava (18 bandas), bandas críticas de igual contribución (17 bandas) y bandas de una octava (6 bandas). Este método presenta una buena correlación entre los métodos estadísticos.[2]

4.5.3.2 Métodos para la Evaluación de la Calidad del Habla

Los métodos para la evaluación objetiva de la calidad del habla, han variado desde una simple medición de relación señal a ruido, medición de distancia espectral, medición de distancia paramétrica hasta la utilización de modelos perceptuales. Los sistemas que utilizan modelos perceptuales son los que tienen más alta correlación con los métodos subjetivos, como lo es el MOS. A continuación se presentan dos de los equipos más utilizados en la actualidad para la medición objetiva de la calidad del habla.

4.5.3.2.1 La Evaluación perceptual de la Calidad de la Voz (PESQ)

La evaluación Perceptual de la Calidad de la voz, PESQ, por sus siglas en inglés, fue oficialmente aprobada como la nueva recomendación ITU-T P.862 en febrero de 2001. El PESQ fue desarrollado por KPN Research, las telecomunicaciones holandesas y británicas,

[2] Para obtener mayor información sobre el AI, STI y RASTI, favor de referirse a [53].

mediante la combinación de dos medidas de calidad de voz, PSQM (Perceived speech quality measure) y PAMS.

4.5.3.2.2 La Medición perceptual de la Calidad de la voz (PSQM)

El sistema de Medición perceptual de la calidad de la voz PSQM por sus siglas en inglés, actualmente es el sistema con índice de correlación mas alto con el método subjetivo MOS. Éste sistema utiliza el modelo perceptual (modelo del oído) para realizar una representación interna de la señal en base a parámetros psicoacústicos, como lo es la sonoridad. A la salida del modelo se calcula la diferencia que hay entre la representación interna de la señal de referencia y la representación interna de la salida del sistema a evaluar. Dicha diferencia es la entrada a un modelo cognoscitivo para que a la salida se obtenga la evaluación de la calidad. En [25] se muestra que tomando en cuenta aspectos cognoscitivos como la asimetría y los intervalos de silencio, el índice de correlación es mayor que 0.97, lo cuál significa un alto acercamiento con el método subjetivo MOS.

4.5.4 Métodos Subjetivos Vs Métodos Objetivos

Los métodos subjetivos, como ya se describió anteriormente, necesitan de mucha preparación. Tiene que ser llevado a cabo por gente calificada, el reclutamiento de la gente que cumpla con las necesidades se vuelve cada vez más difícil después de determinado tiempo, puesto que las personas que ya realizaron una evaluación no pueden realizar otra sino hasta 3 meses después y todo el proceso es muy costoso. Además el tiempo promedio para obtener una evaluación subjetiva con el MOS es de aproximadamente un año, lo cuál implica mucho tiempo para que un producto esté sin salir al mercado.

Por otro lado, los métodos objetivos presentan una mejor opción debido a que es mucho más barato y el tiempo de evaluación se reduce significativamente, pero tienen la desventaja que no dan en un 100% el mismo resultado que se obtendría con la calificación de la opinión media. Algunas compañías prefieren aún los métodos subjetivos para la evaluación de sus sistemas de transmisión en lugar de un método objetivo que no sea totalmente confiable.

4.6 Conclusiones

Hemos visto que el análisis morfológico de las imágenes para las imágenes sonoras, presenta una muy buena opción para la extracción de características propias de las imágenes. Como se mencionó anteriormente, tiene el inconveniente que el análisis es diferente para cada palabra y no se puede hacer una generalización, pero para realizar un reconocimiento con vocabulario limitado, el resultado podría ser bastante satisfactorio. Con la esqueletización vemos que el esqueleto se va degradando pero, sería interesante poder establecer alguna métrica, dependiendo de la degradación del esqueleto, de que tan degradada se encuentra la palabra.

Por otro lado, con el análisis cuaterniónico, los resultados son sorprendentemente buenos. Cada componente de la imagen cuaterniónica nos da determinadas características y hemos visto que las características se mantienen hasta muy bajos niveles de SNR. Lo interesante es que no importa que palabra sea, ni que locutor diga la palabra y las invariantes siempre están a un nivel alto en la escala de grises.

Se podrían obtener resultados excelentes haciendo una combinación entre los dos métodos de análisis estudiados en éste capítulo. Si primero descomponemos la imagen en sus cuatro componentes, *real*, i, j y k y luego aplicamos las técnicas morfológicas a cada una de sus componentes, podemos extraer esas invariantes y obtener una imagen binaria en la que se encuentren presentes únicamente las invariantes de las imágenes. El ruido presente de fondo se puede eliminar realizando una umbralización y posteriormente hacer una apertura de área para eliminar las componentes menores a determinada área. Algo que hay que hacer notar es que, aplicando las técnicas morfológicas a las componentes de la imagen cuaterniónica, el análisis morfológico siempre puede ser el mismo, con esto queremos decir que determinado análisis lo podemos utilizar para cualquier palabra y para cualquier locutor, puesto que la umbralización no tiene que moverse, a diferencia de lo que se mencionó en la sección 4.3.2.

Estos resultados pueden tener una aplicación inmediata en el reconocimiento del habla en ambientes ruidosos. Sería interesante hacer un estudio de las imágenes de la sonoridad de los fonemas para observar si hay diferencias significativas entre un "fa" y un "pa", como ejemplo. Si se logra extraer características invariantes a éstos niveles, y lo mejor aún, si presentan una forma determinada, el reconocimiento de voz mediante imágenes de la sonoridad puede presentar una opción excelente para los sistemas de reconocimiento del habla.

5 CONCLUSIÓN

Hemos programado el modelo psicoacústico del oído obteniendo como resultado una buena aproximación para el cálculo de la sonoridad. Las pequeñas diferencias que existen en el cálculo de la sonoridad, no son de gran relevancia para la aplicación que le hemos dado a los resultados. Los efectos frecuenciales de enmascaramiento es la principal ventaja que ofrece el trabajar con señales de voz basadas en parámetros perceptuales en ambientes ruidosos, y esto lo podemos ver por el efecto que tiene el ruido sobre las imágenes de la sonoridad.

El desempeño del sistema de reconocimiento de voz no fue del todo satisfactorio para el reconocimiento en ambientes ruidosos por las razones que se dan al final del capítulo 3. Sin embargo, utilizando los resultados del capítulo 4, el reconocimiento en ambientes ruidosos puede mejorar significativamente.

El analizar la voz, mediante análisis de las imágenes de la sonoridad, presentó muchos más beneficios y ventajas de las que se esperaba. Con las herramientas morfológicas para el tratamiento de las imágenes se pudieron obtener resultados satisfactorios y hemos visto que es una herramienta bastante adecuada para éste tipo de imágenes debido a las características que éstas presentan.

Los resultados obtenidos con la transformada cuaterniónica de Fourier motivan a continuar el estudio de la voz utilizando imágenes de la sonoridad. En éste trabajo únicamente se utilizaron dos herramientas para el tratamiento de las imágenes. Por el momento, las dos herramientas fueron muy adecuadas para el análisis de las imágenes de la sonoridad pero aún falta probar más herramientas y tratar de explotar mas éste campo de estudio. Una herramienta interesante sería la transformada wavelet y la transformada cuaterniónica wavelet.

Como trabajo a futuro al parecer queda mucho. En primera instancia, realizar el tratamiento de las imágenes de la sonoridad pero utilizando una mezcla de las dos herramientas para el análisis de imágenes: descomposición cuaterniónica de las imágenes y morfología matemática. Esto con la intención de extraer las características invariantes de cada palabra y utilizar esas invariantes para realizar el reconocimiento de la palabra.

Otra cosa que queda por hacer es el analizar las imágenes de la sonoridad de los fonemas, estudiar si se pueden extraer características descriptivas de cada fonema así también como ver que pasa con la forma. Todo esto con intención de saltar al campo de reconocimiento del habla continua.

En este trabajo únicamente se analizó que pasa con cada palabra, es decir, ver si habían invariantes cuando determinada palabra se encontraba perturbada por ruido a determinado nivel de SNR. Algo interesante sería ver si la descomposición de la imagen de la sonoridad, en su representación analítica cuaterniónica, es capaz de extraer las invariantes de cada locutor, es decir, las características propias de un locutor. El trabajo sería hacer un análisis exhaustivo de cada componente y hacer la comparación entre las imágenes de las componentes, *real*, *i*, *j*, y *k*, de diferentes palabras dichas por el mismo locutor y tratar de encontrar elementos que aparezcan en todas las imágenes.

Con esto concluimos la realización de éste trabajo. Por el momento los objetivos fueron cumplidos pero reiteramos que aún falta mucho trabajo que hacer.

APÉNDICE 1: CÓDIGO DE PROGRAMACIÓN DE LA HERRAMIENTA "CALCULADOR DE SONORIDAD"

```
function fig = principal()
% This is the machine-generated representation of a Handle Graphics
object
% and its children.  Note that handle values may change when these
objects
% are re-created. This may cause problems with any callbacks written to
% depend on the value of the handle at the time the object was saved.
% This problem is solved by saving the output as a FIG-file.
%
% To reopen this object, just type the name of the M-file at the MATLAB
% prompt. The M-file and its associated MAT-file must be on your path.
%
% NOTE: certain newer features in MATLAB may not have been saved in this
% M-file due to limitations of this format, which has been superseded by
% FIG-files.  Figures which have been annotated using the plot editor
tools
% are incompatible with the M-file/MAT-file format, and should be saved
as
% FIG-files.

load principal

h0 = figure('Color',[0.250980392156863 0.501960784313725
0.501960784313725], ...
      'Colormap',mat0, ...
      'FileName','D:\Back_up\Backup_matlab\principal.m', ...
      'MenuBar','none', ...
      'Name','Calculador de Sonoridad', ...
      'NumberTitle','off', ...
   'PaperPosition',[18 180 576 432], ...
      'PaperUnits','points', ..
      'Position',[5 25 800 534], ...
      'Tag','principal', ...
      'ToolBar','none');
h1 = uimenu('Parent',h0, ...
      'Callback',mat1, ...
      'Label','&Archivo', ...
      'Tag','uimenu1');
h2 = uimenu('Parent',h1, ...
      'Callback','openbds', ...
      'Label','A&brir base de sonido *.?fo', ...
      'Tag','openbds');
h2 = uimenu('Parent',h1, ...
      'Callback','openwav', ...
      'Label','Abrir archivo  .&wav', ...
      'Tag','openwav');
h2 = uimenu('Parent',h1, ...
      'Callback','recwav', ...
```

```
      'Label','&Grabar sonido en formato wav'. ...
      'Separator'.'on'. ...
      'Tag'.'recwav');
h2 = uimenu('Parent',h1, ...
      'Callback','close(gcf) . ...
      'Label','&Cerrar...'. ..
      'Separator','on', ..
      'Tag'.'&Archivo&Cerrar...1');
h2 = uimenu('Parent',h1, ...
      'Callback',mat2, ...
      'Label','Pre&ferencias'. ...
      'Separator','on', ...
      'Tag'.'&Archivo&Cerrar...uimenu1');
h2 = uimenu('Parent',h1, ...
      'Callback',mat3, ...
      'Label ,'Configuracion de pa&gina'. ..
      'Separator','on'. ...
      'Tag'.'&ArchivoPre&ferenciasuimenu1');
h2 = uimenu('Parent',h1, ...
      'Callback',mat4,  ..
      'Label'.'Configuracion de Impre&sion'  ...
      'Tag',                    ');
h2 = uimenu('Parent',h1, ...
      'Callback',mat5, ...
      'Label'.'&Imprimir'. ...
      'Tag'.                ');
h1 = uicontrol('Parent',h0, ...
      'Units'.'points', ...
      'BackgroundColor'.[0.250980392156863 0 0.250980392156863],  ..
      'ListboxTop',0,  ..
      'Position',[3.75 150.75 132.75 88.5], ...
      'Style','frame'.  ..
      'Tag'.'Frame3');
h1 = uicontrol('Parent',h0, ...
      'Units'.'points',  ..
      'BackgroundColor'.[0.250980392156863 0 0.250980392156863], ...
      'ListboxTop',0, ...
      'Position',[2.25 271.5 150 126.75], ...
      'Style','frame'. ...
      'Tag','Frame2');
h1 = uicontrol('Parent',h0, ..
      'Units'.'points', ...
      'BackgroundColor'.[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'Callback'.'xtrak'. ...

      'ListboxTop',0, ...
      'Position'.[99.75 369.75 45 19.5], ...
      'Style','popupmenu'. ...
      'Tag','PopupMenu1'. ...
   'Value',18);
Noises=['ninguno' '|  'ambiental'  |' 'int_car'  |' 'fabrica'  |'
'blanco' '|  'tanque'];
h1 = uicontrol('Parent',h0, ...
      'Units'.'points', ...
      'BackgroundColor'.[1 1 1],  ..
```

```
      'Position',[99 351.5625 50 15], ...
   'String',Noises, ...
      'Style','popupmenu',  ..
   'Tag','ruido', ...
   'Callback','', ...
      'Value',1);
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588],  ..
      'Callback','playwav',  ..
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 303.75 83.25 15.75], ...
      'String','Escuchar Sonido', ...
      'Tag','Pushbutton1');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'FontWeight','bold', ..
      'ListboxTop',0, ..
      'Position',[3.75 352.5 79.5 14.25], ...
      'String','Añadir Ruido', ..
      'Style','text', ..
      'Tag','StaticText1');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588],
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 328.875 82.5 15], ...
      'String','Nivel SNR',  ..
      'Style','text', ..
      'Tag','StaticText2');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[1 1 1], ...
      'ListboxTop',0, ..
      'Position',[99 328.875 45 15],  ..
   'Style','edit', ...
   'String',0, ..
   'Tag','SNR',  ..
   'callback','plusnoise', ...
      'UserData', [ ]');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'Callback','filtering', ...
      'FontWeight','bold'   ..
      'ListboxTop',0,  ..
      'Position',[3.75 280.5 85.5 15], ...
      'String','Calcular Sonoridad', ...
      'Tag','Pushbutton2');
```

```
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[1 1 0.501960784313725], ...
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 241.5 134.25 22.5], ...
      'String','SISTEMA DE RECONOCIMIENTO', ...
      'Style','text', ...
      'Tag','StaticText3');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position' [3.75 210.75 83.25 15], ...
   'String','Reconocimiento', ...
   'callback','recognize', ...
      'Tag','Pushbutton3');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ..
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 188.25 110.25 16.5], ...
      'String','Entrenar Red para Voz', ...
      'Tag','Pushbutton4');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 165 117.75 17.25], ...
      'String' 'Entrenar Red para Locutor', ...
      'Tag','Pushbutton5');
h1 = axes('Parent',h0, ...
      'Box','on', ...
      'CameraUpVector',[0 1 0], ...
      'CameraUpVectorMode','manual', ...
      'Color',[1 1 1], ...
      'ColorOrder',mat6, ...
   'Position',[0.29 0.1853932584269663 0.6625 0.700374531835206], ...
   'Visible','off', ...
      'Tag','Axes1'   ..
      'XColor',[1 1 1], ..
      'YColor',[1 1 1], ...
      'ZColor',[1 1 1]);
h2 = text('Parent',h1, ...
      'Color',[0 0 0], ...
      'HandleVisibility','off', ...
      'HorizontalAlignment','center', ...
      'Position',[0.4990548204158791 -0.06434316353887404
9.160254037844386], ...
      'Tag','Axes1Text4', ...
```

```
      'VerticalAlignment','cap');
set(get(h2,'Parent'),'XLabel',h2);
h2 = text('Parent',h1, ...

      'Color',[0 0 0], ...
      'HandleVisibility','off', ...
      'HorizontalAlignment','center', ...
      'Position',[-0.05482041587901704 0.4959785522788204
9.160254037844386], ...
      'Rotation',90, ...
      'Tag','Axes1Text3', ...
      'VerticalAlignment','baseline');
set(get(h2,'Parent'),'YLabel',h2);
h2 = text('Parent',h1, ...
      'Color',[0 0 0], ...
      'HandleVisibility','off', ...
      'HorizontalAlignment','right', ...
      'Position',[-0.4404536862003781 1.160857908847185
9.160254037844386], ...
      'Tag','Axes1Text2', ...
      'Visible','off');
set(get(h2,'Parent'),'ZLabel',h2);
h2 = text('Parent',h1, ...
      'Color',[0 0 0], ...
      'HandleVisibility','off', ...
      'HorizontalAlignment' 'center', ...
      'Position',mat7, ...
      'Tag','Axes1Text1', ...
      'VerticalAlignment','bottom');
set(get(h2,'Parent'),'Title',h2);
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[0.752941176470588 0.752941176470588
0.752941176470588], ...
      'FontWeight','bold', ...
      'ListboxTop',0, ...
      'Position',[3.75 374.625 88.5 14.25], ...
      'String','Seleccionar Pista', ...
      'Style','text', ...
      'Tag','StaticText4');
h1 = uicontrol('Parent',h0, ...
      'Units','points', ...
      'BackgroundColor',[1 1 0], ...
      'FontWeight','bold', ...
      'ForegroundColor',[1 0 0], ...
      'ListboxTop',0, ...
      'Position',[14.25 14.25 209.25 15], ...
      'Style','text', ...
      'Tag','display', ...
      'Visible','off');
if nargout > 0, fig = h0; end

pause(1);
init;
```

```
function init

%*********************************************************************
%               Definicion de variables globales
%*********************************************************************


% fe -> Frecuencia de muestreo
% coeff -> Curvas de equal-loudness
% son -> senial de sonido
% Bl, Al, Bm, Am, Bh, Ah -> coeficientes de los filtros de banda critica
% index ->
% chem -> archivo de la base de sonido en formato SAM
% fich -> ruta del archivo de la base de sonido
global fe ;
global coeff son ;
global Bl Al Bm Am Bh Ah ;
global index chem fich nbech ;
global MMf ;
global Aff ;
global index;
global it P nbapp ;
%*********************************************************************
% Inicializacion de variables globales.
      nbapp=30 ;
      P=zeros(16*80,nbapp) ;
      it=0 ;
      Aff=0;
      fe=8000 ;
%*********************************************************************
%   parametros del banco de filtros de banda critica                *
%*********************************************************************
if (exist('n') & exist('Bh'))== 0
   set(findobj('tag','display'),'visible'.'on'. ...
      'string','Critical-Band filter calculating...');
% ** orden de los filtros
      n(1)=4 ;
      n(2)=6 ;
      n(3)=8 ;
% ** Rp : rizo en pasa banda
% ** Rs : atenuacion en la banda de corte
      Rp(1)=5 ;
      Rs(1)=40 ;
      Rp(2)=0.5 ;
      Rs(2)=40 ;
      Rp(3)=0.5 ;
   Rs(3)=36 ;
% ** Ancho de banda critico (Hz)
      w(1,:)=[100 198] ;
      w(2,:)=[200 298] ;
      w(3,:)=[300 398] ;
      w(4,:)=[400 508] ;
      w(5,:)=[510 627] ;
      w(6,:)=[630 765] ;
      w(7,:)=[770 915] ;
      w(8,:)=[920 1073] ;
```

```
      w(9,:)=[1080 1265] ;
      w(10,:)=[1270 1474] ;
      w(11,:)=[1480 1716] ;
      w(12,:)=[1720 1998] ;
      w(13,:)=[2000 2317] ;
      w(14,:)=[2320 2697] ;
      w(15,:)=[2700 3148] ;
      w(16,:)=[3150 3706] ;
      w=w./(fe/2) ;
      % coeficientes de equal-loudness para un tono de 1kHz a 40dB. de
acuerdo
   % con Zwicker.
      coeff=[1.0833;1.04164;1.02;1.01;1;1;1;1;1;1;1;0.966;0.958;0.958;0.9
58;0.958];
% ** modelo del patron de excitacion
c=([1.5 1.5 0.75 0.5 0.25 0.125 0 0 0 0 0 0 0 0 0 0 ])  ;
c=c./max(abs(c));
      for i=1:16
        MMf(i:i+16-1,i)=c ;
      end
      MMf=MMf(1:16,1:16) ;
      clear c ;
% ** funciones de transferencia de los filtros
      [Bl(1,:),Al(1,:)]=ellip(n(1),Rp(1),Rs(1),w(1,:));
      [Bl(2,:),Al(2,:)]=ellip(n(1),Rp(1),Rs(1),w(2,:));
      [Bl(3,:),Al(3,:)]=ellip(n(1),Rp(1),Rs(1),w(3,:));
      [Bl(4,:),Al(4,:)]=ellip(n(1),Rp(1),Rs(1),w(4,:));
      [Bm(1,:),Am(1,:)]=ellip(n(2),Rp(2),Rs(2),w(5,:));
      [Bm(2,:),Am(2,:)]=ellip(n(2),Rp(2),Rs(2),w(6,:));
      [Bm(3,:),Am(3,:)]=ellip(n(2),Rp(2),Rs(2),w(7,:));
      [Bm(4,:),Am(4,:)]=ellip(n(2),Rp(2),Rs(2),w(8,:));
      [Bm(5,:),Am(5,:)]=ellip(n(2),Rp(2),Rs(2),w(9,:));
      [Bm(6,:),Am(6,:)]=ellip(n(2),Rp(2),Rs(2),w(10,:));

      [Bh(1,:),Ah(1,:)]=ellip(n(3),Rp(3),Rs(3),w(11,:));
      [Bh(2,:),Ah(2,:)]=ellip(n(3),Rp(3),Rs(3),w(12,:));
      [Bh(3,:),Ah(3,:)]=ellip(n(3),Rp(3),Rs(3),w(13,:));
      [Bh(4,:),Ah(4,:)]=ellip(n(3),Rp(3),Rs(3),w(14,:));
      [Bh(5,:),Ah(5,:)]=ellip(n(3),Rp(3),Rs(3),w(15,:));
   [Bh(6,:),Ah(6,:)]=ellip(n(3),Rp(3),Rs(3),w(16,:));
   set(findobj('tag'.'display'),'visible'.'off');
end


% Programa que realiza el calculo de la sonoridad
% el resultado lo deja en un archivo MAT llamado "sonie"
% la matriz de sonoridad se encuentra en la variable E.
function filtering
global fe ;
global Bl Al Bm Am Bh Ah ;
global coeff son backup_son;
global MMf ;
global flag t;
global beg ;
global test_noise ;
plusnoise ;
```

```
flag = 0;
lg=size(backup_son,2) ;
% ** Ancho de cada cuadro de energia (10 ms)
tsd=10e-3 ;
% ** K : muestras/cuadro
K=fe*tsd ;
% ** numero de cuadros
t=floor(lg/K) ;
% ** Longitud usual del archivo de sonido : 800 ms
% ** 80 sq
t2=(800e-3)/tsd ;
% ** Matriz de energias (nrj)
E=zeros(16,t2) ;
% ** Desviacion estandar del sonido, solo deja un cuadro de longitud K
for i=1:t
  Ec(i)=std((backup_son(((i-1)*K+1):(i*K)))') ;
end
% ** Inicio del sonido util
   seuil=700 ;
   beg=find(Ec>seuil) ;
   beg=beg(1) ;
% ** Matriz de energias utiles
if (t-beg)>=t2-1
   son=son((beg*K)-(K-1):(beg*K)-(K-1)+K*t2-1) ;
  disp('Sound > 80 ms');
else
  son=[son((beg*K)-(K-1):lg) zeros(1,(t2*K-(lg-((beg*K)-(K-1))))-1)] ;
  disp('Sound < 80 ms => completed with 0') ;
end
size(son);
son = (son./max(abs(son)))*1414; %normalizacion a 60dB
% ****************
% Filtrado de la senial
% ****************
sf=zeros(16,t2*K) ;
sf(1,:)=filter(Bl(1,:),Al(1,:),son) ;
sf(2,:)=filter(Bl(2,:),Al(2,:),son) ;
sf(3,:)=filter(Bl(3,:),Al(3,:),son) ;
sf(4,:)=filter(Bl(4,:),Al(4,:),son) ;
sf(5,:)=filter(Bm(1,:),Am(1,:),son) ;
sf(6,:)=filter(Bm(2,:),Am(2,:),son) ;
sf(7,:)=filter(Bm(3,:),Am(3,:),son) ;
sf(8,:)=filter(Bm(4,:),Am(4,:),son) ;
sf(9,:)=filter(Bm(5,:),Am(5,:),son) ;
sf(10,:)=filter(Bm(6,:),Am(6,:),son) ;
sf(11,:)=filter(Bh(1,:),Ah(1,:),son) ;
sf(12,:)=filter(Bh(2,:),Ah(2,:),son) ;
sf(13,:)=filter(Bh(3,:),Ah(3,:),son) ;
sf(14,:)=filter(Bh(4,:),Ah(4,:),son) ;
sf(15,:)=filter(Bh(5,:),Ah(5,:),son) ;
sf(16,:)=filter(Bh(6,:),Ah(6,:),son) ;
% ****************
% Calculo de la energia
% ****************
ffor i=1:t
E(:,i)=(10.*log10(sum((sf(:,((i-1)*K+1):(i*K)) .^2))'./80)) ;
```

```
   ind=find(isinf(E(:,i)));
   for r=1:length(ind)
      E(ind(r),i)=0;
   end
end
%indice de enmascaramiento av
av = [0.63 0.63 0.6 0.58 0.57 0.55 0.55 0.525 0.5 0.49 0.48 0.436 0.428
0.422 0.416 0.39]' ;
Etq = [42 18.5  11.5 8.3 6.7 5.5 4.8 4.3 4 3.8 3.4 3 2 0 -1 -2]';

for i=1:t2
   N(:,i) = (0.068.*av.^(-
0.25)) *10.^(0.025.*Etq). (((ones(1,16)'+av.*10.^(0.1*(E(:,i)-
Etq))).^0.25)-ones(1,16)');
end
% ** Freq. masking
%sfm=zeros(16,16) ;
%for i=1:t2
%  for j=1:16
%    sfm(:,j)=MMf(:,j).*N(j,i) ;
%  end
%  N(:,i)=(max(sfm')') ;
%end
   clear sfm
   E=N;
[a b c]=find(N) ;
N=N/max(max(c)) ;
clear a b c ;
% *************
% Visualizacion del sonido
% *************
global a1 ;
h = figure;
set(gcf,'numbertitle','off');
set(gcf,'name','Sonoridad');
set(gcf,'color' [0 0 0]);
set(gcf,'menubar','none');
axes ;
cla
pcolor([N zeros(size(N,1),1) ; zeros(1,size(N,2)+1)]) ;
shading('interp') ;

%colormap(hsv);
%colormap(jet);
colormap(gray(255));
caxis('auto') ;
disp('Grafica de N') ;
size(N);
title('Sonie' 'color',[0 0 0],'units', 'normalized', 'position', [1.05
0.5])
set(gca,'xticklabels',(t*10e-3*10*get(gca,'xtick'))') ;
set(gca,'ytick',[0:16]) ;
set(gca, 'xcolor', [1 1 1]) ;
set(gca, 'ycolor', [1 1 1]) ;
set(gca, 'tickdir', 'out') ;
xlabel('ms', 'color', [1 1 1]) ;
```

```
ylabel('Barks','color',[1 1 1]) ;

sonie_3d=uicontrol (gcf,'style', 'pushbutton','string','3D', ...
   'units','normalized','position',[0.40 0.94 0.16 0.04], ...
   'tag','3dgraph', ...
   'callback','sonie3d') ;
writeimage=uicontrol (gcf,'style', 'pushbutton','string','Guardar
Imagen', ...
   'units','normalized','position',[0.60 0.94 0.16 0.04], ...
   'tag','', ...
   'callback','saveimage2') ;
snr = get(findobj('tag','SNR'),'string');
snr = ['SNR=   snr 'dB'];
disp_snr=uicontrol (gcf,'style', 'text','string',snr, ...
   'units','normalized','position',[0.25 0.94 0.16 0.04], ...
   'BackgroundColor',[0 0 0], ..
   'foregroundcolor',[1 1 1], ..
   'tag','disp_snr', ...
   'callback','') ;
handle=findobj('tag','ruido');
indice=get(handle,'value');
if indice == 1
   tipo_ruido='ninguno';
elseif indice == 2
   tipo_ruido='ambiental  ;
elseif indice == 3
   tipo_ruido='int_car' ;
elseif indice == 4
   tipo_ruido='fabrica' ;
elseif indice == 5
   tipo_ruido='blanco' ;
elseif indice == 6
   tipo_ruido='tanque' ;
end
tipo_ruido=['Ruido:   tipo_ruido];
disp_ruido=uicontrol (gcf,'style', 'text','string',tipo_ruido, ...
   'units','normalized','position',[0.05 0.94 0.16 0.04], ...
   'BackgroundColor',[0 0 0], ...
   'foregroundcolor',[1 1 1], ...
   'tag','disp_ruido'  ..
   'callback','') ;
h1 = uicontrol(gcf, ...
      'Units','points', ...
      'BackgroundColor',[1 1 0], ...
      'FontWeight','bold', ...
      'ForegroundColor' [1 0 0], ...
      'ListboxTop',0, ...
      'Position',[14.25 14.25 209.25 15], ...
      'Style','text', ...
      'Tag','display1', ...
      'Visible','off');
save sonie E ;
clear E ;
```

```
%Programa para generar los numeros en formato WAV.
%Este programa lee la base de datos en formato SAM y convierte las pistas
en
%formato WAV.
global index E chem fich nbech son pistas ind;
global backup_son ;
global test_noise track file_name;
openbds ;
track=[2 17 22 23 41 52 53 65 70 85] ;
xtrak2write ;



function autowav()
% ************************************************
% Generacion automatica de una matriz para el
% aprendizaje para la red neuronal
% con archivos WAV de 16 bits/8kHz.
% ************************************************
global P ;
global nsons ;
global it ;
global nbapp ;
global son ;
nbapp=20;
nsons = 3 ;
nsons_t = 10;
for i=1:nsons_t,
   if i==1 chem='D:\Back_up\Backup_matlab\numbers\cero\' ;
   elseif i==2 chem='D:\Back_up\Backup_matlab\numbers\uno\' ;
   elseif i==3 chem='D:\Back_up\Backup_matlab\numbers\dos\' ;
   elseif i==4 chem='D:\Back_up\Backup_matlab\numbers\tres\' ;
   elseif i==5 chem='D:\Back_up\Backup_matlab\numbers\cuatro\' ;
   elseif i==6 chem='D:\Back_up\Backup_matlab\numbers\cinco\' ;
   elseif i==7 chem='D:\Back_up\Backup_matlab\numbers\seis\' ;
   elseif i==8 chem='D:\Back_up\Backup_matlab\numbers\siete\' ;
   elseif i==9 chem='D:\Back_up\Backup_matlab\numbers\ocho\' ;
   elseif i==10 chem='D:\Back_up\Backup_matlab\numbers\nueve\' ;
   end
   for j=1:nbapp,
      fich=['son' num2str(j)  .wav'];
      A=wavread([chem fich]) ;
      son=A' ;
      son=(2^16)/max(max(son)).*son ;
      fich;
      filtering;
      load sonie ;
      it=it+1 ;
      [L,C]=size(E) ;
      for ic=1:C
         for il=1:L
            Pt(il+(ic-1)*L,1)=[E(il,ic)];
         end
      end
      P(:,it)=Pt ;
      clear Pt ;
```

```
      disp('Matriz creada');
   end
end
save bdswavnbr P ;
size(P);
it;

learnsp


%funcion para abrir los archivos en formato SAM
function openbds
clear global pistas ;
global index E chem fich nbech son pistas;
nbech=0 ;
MAX=10 ;
[info, chem]=uigetfile('* ?fo', 'abrir base de sonido') ;
if chem~=0
      fich=[info(1:11) 'S']
      fid=fopen([chem info]) ;
      if fid==-1, disp ('Error de lectura'), break, end ;
      e1=zeros(1,MAX) ;
      e2=e1 ;
      j=1 ;
      while 1
            ligne = fgetl(fid) ;
            if ~isstr(ligne), break, end % ** EOF
            if ligne(1:3)=='ELF', break,
            elseif ligne(1:3)=='LBR'
                  nbech=nbech+1 ;
                  ligne = ligne(6:size(ligne,2)) ;
                  while 1
                        i=sscanf(ligne,'%c',1) ;
                        ligne=ligne(2:size(ligne,2)) ;
                        if i=='.' ;
                              break ;
                        else
               e1=[e1 i];
                              j=j+1 ;
                        end
                  end
                  j=1 ;
                  while 1
                        i=sscanf(ligne,'%c',1) ;
                        ligne=ligne(2:size(ligne,2)) ;
                        if i=='.' ;
                              break ;
                        else
               e2=[e2 i];
                              j=j+1 ;
                        end
                  end
                  j=1 ;
                  e1=str2num(e1) ;
         e2=str2num(e2) ;
```

```
        pistas(nbech,:)=[e1, e2] ;
        e1=0;
        e2=0;
           end
     end
     fclose(fid) ;
index=findobj('tag','PopupMenu1');
p = get(index,'position');
set (index,'string', [1:nbech],'position',p);
end
```

```
% Funcion para abrir un archivo de sonido en formato WAV
function openwav
global son backup_son fich;
[fich,chem]=uigetfile('D:\Backup\Backup_Matlab\numbers\*.wav', 'Ouvrir un
son WAV 8kHz/16bits/mono') ;
if chem~=0
   [A,FS]=wavread([chem fich]) ;
      son=A' ;
   son=(2^16)/max(max(son)).*son ;
   backup_son = son ;
   plusnoise ;
   handle=get(findobj('tag','principal'),'currentaxes');
   axes(handle);
   plot(son,'r') ;
   title('Son','color',[1 1 1],'units', 'normalized', 'position', [1.05
0.5])
   set(gca, 'xcolor', [1 1 1]) ;
   set(gca, 'ycolor', [1 1 1]) ;
   xlabel('nb de points', 'color', [1 1 1]) ;
end
```

```
% funcion para tocar un archivo en formato wav
function playwav
global son fe backup_son;
if exist('son')
   plusnoise ;
   son_=son./max(abs(son));
      sound (son_, fe) ;
end
```

```
% funcion para aniadir ruido a los sonidos
function plusnoise
global son backup_son noise_d ;
if (isempty(son)) == 0
   handle=findobj('tag','ruido');
   index=get(handle,'value');
   if index == 1
```

```
      son = backup_son ;
      break ;

   elseif index == 2
      file2open='ambiental.wav' ;
   elseif index == 3
      file2open='int_car.wav' ;
   elseif index == 4
      file2open='fabrica.wav' ;
   elseif index == 5
      file2open='blanco' ;
   elseif index == 6
      file2open='tanque' ;
   end
   son = backup_son ;
   path='D:\Backup\Backup_Matlab\noisedatabase\ ;
   [noise,Fsn] = wavread([path file2open]) ;
   Fsn ;
   noise = noise(1:length(son)) ;
   Enoise = sum(noise.^2) ;
   Eson = sum(son.^2)
   snr = get(findobj('tag','SNR'),'string') ;
      snr = str2num(snr) ;
      Enoise_d = Eson./(10^(snr./20)) ;
      k = Enoise_d./Enoise ;
      noise_d = sqrt(k).*noise ;
      Enoise_k = sum(noise_d.^2)
      signalplusnoise = son + noise_d' ;
      backup_son = son ;
      son = signalplusnoise ;
      handle=get(findobj('tag','principal'),'currentaxes');
   axes(handle);
   plot(son,'r') ;
   title('Son','color',[1 1 1],'units','normalized','position',[1.05
0.5])
   set(gca, 'xcolor', [1 1 1]) ;
   set(gca, 'ycolor', [1 1 1]) ;
   xlabel('nb de points', 'color', [1 1 1]) ;
end


% Funcion para grabar el sonido en formato wav
function recwav
global son fe chemin nom;
if exist('son')
      A=son ;
   if chemin~=0
      A=A./max(abs(A));
            wavwrite (A, fe, [chemin nom]) ;
      end
end


% Funcion para graficar la imagen de la sonoridad en 3-D
function sonie3d
```

```
global E;
global flag t;
load sonie;
flag= flag+1;
if flag == 1
   surf([E zeros(size(E,1),1) ; zeros(1,size(E,2)+1)]);
   title('Sonie','color',[0 0 0],'units', 'normalized', 'position', [1.05
0.5])
   set(gca,'xticklabels',(t*10e-3*10*get(gca,'xtick'))') ;
   set(gca,'ytick',[0:16]) ;
   set(gca, 'xcolor', [1 1 1]) ;
   set(gca, 'ycolor', [1 1 1]) ;
   set(gca, 'tickdir', 'out') ;
   xlabel('ms', 'color', [1 1 1]) ;
   ylabel('Barks','color',[1 1 1]) ;
   zlabel('Nivel de Sonoridad','color',[1 1 1]) ;
   rotate3d on ;
   set(findobj('tag','3dgraph'),'string','2D') ;
   flag = -1 ;
else
   pcolor([E zeros(size(E,1),1) ; zeros(1,size(E,2)+1)]) ;
   title('Sonie','color',[0 0 0],'units', 'normalized', 'position'  [1.05
0.5])
   set(gca,'xticklabels',(t*10e-3*10*get(gca,'xtick'))') ;
   set(gca,'ytick',[0:16]) ;
   set(gca, 'xcolor', [1 1 1]) ;
   set(gca, 'ycolor', [1 1 1]) ;
   set(gca, 'tickdir', 'out') ;
   xlabel('ms', 'color', [1 1 1]) ;
   ylabel('Barks','color',[1 1 1]) ;
   rotate3d off;
   set(findobj('tag','3dgraph'),'string','3D');
   flag = 0;
end
shading('interp');
set(gca,'color',[0 0 0]);



% Funcion para seleccionar la pista de la base de sonido
% en formato SAM
function xtrak
global index E chem fich nbech son pistas ind;
global backup_son ;
global test_noise ;
test_noise = 1;
set(findobj('tag','ruido'),'value',1);
ind=get(index, 'value') ;
did=fopen([chem fich]  ) ;
if did==-1, disp ('Erreur de lecture'), break, end ;
OK=fseek(did,pistas(ind,1)*2,'bof') ;
if OK==-1, disp ('Erreur d''accès'), break, end ;
[A nb]=fread(did, pistas(ind,2)-pistas(ind,1),'short') ; % reading of 16
bits
% ** conversion de  16kHz -> 8kHz
```

```
S=zeros(1,(size(A,1)/2)-1) ;
for k=2:2:size(A,1)-1
      S(1,k/2)=A(k) ;
end
S=round(S) ;
son=S ;
backup_son = son ;
plot(son,'r') ;
title('Son','color',[1 1 1],'units', 'normalized', 'position', [1.05
0.5])
set(gca, 'xcolor', [1 1 1]) ;
set(gca, 'ycolor', [1 1 1]) ;
xlabel('nb de points', 'color', [1 1 1]) ;
fclose(did) ;
plusnoise ;



% Seleccion de las pistas para la generacion automatica
% de los archivos wav
function xtrak2write
global index E chem fich nbech son pistas ind;
global backup_son ;
global test_noise ;
global track nom chemin;
test_noise = 1;
set(findobj('tag','ruido'),'value',1);
locutor='loc29\';
chemin ='D:\Backup\Backup_Matlab\numbers\';
for i=1:10
   ind=track(i);
   if i==1
      nom=[locutor 'siete'];
   elseif i==2
      nom=[locutor 'nueve'];
   elseif i==3
      nom=[locutor 'tres'];
   elseif i==4
      nom=[locutor 'cuatro'];
   elseif i==5
      nom=[locutor 'cinco'];
   elseif i==6
      nom=[locutor 'uno'];
   elseif i==7
      nom=[locutor 'seis'];
   elseif i==8
      nom=[locutor 'ocho'];
   elseif i==9
      nom=[locutor 'cero'];
   elseif i==10
      nom=[locutor 'dos'];
   end
   did=fopen([chem fich]  ) ;
   if did==-1, disp ('Erreur de lecture'), break, end ;
      OK=fseek(did,pistas(ind,1)*2,'bof') ;
```

```
   if OK==-1, disp ('Erreur d''accès'), break, end ;
      [A nb]=fread(did, pistas(ind,2)-pistas(ind,1),'short') ; % reading
of 16 bits
   % ** Passage 16kHz -> 8kHz
   S=zeros(1,(size(A,1)/2)-1) ;
   for k=2:2:size(A,1)-1
      S(1,k/2)=A(k) ;
   end
      S=round(S) ;
   son=S ;
   backup_son = son ;
      plot(son,'r') ;
      title('Son','color',[1 1 1],'units', 'normalized', 'position',
[1.05 0.5])
   set(gca, 'xcolor', [1 1 1]) ;
   set(gca, 'ycolor', [1 1 1]) ;
   xlabel('nb de points', 'color'  [1 1 1]) ;


   fclose(did) ;
      plusnoise ;
   recwav ;
end



%Grabar Imagen del sonido en formato *.tiff.
function saveimage
global fich ind loc rl;
fich1=fich;%(1:8);
set(findobj('tag','display1'),'visible','on', ...
      'string','Saving image.. ');
handle=findobj('tag','ruido');
ruido=get(handle,'value');
snr = get(findobj('tag','SNR'),'string') ;
filename=['D:\Backup\Backup_Matlab\imagenes\loc' num2str(rl) '\  fich1 '-
'num2str(ind) '-' num2str(ruido) '-' snr '.tiff'];
load sonie;
E=E./max(max(E));
colormap(gray(255));
map=colormap;
%construccion de una imagen del sonido, cada cuadro de energia es de
10x16 pixeles.
imagen=[];
nrjsquare=[];
for i=1:16
   for j=1:80
      for r=1:5
         for l=1:1
            nrjsquare(r,l)=E(i,j);
         end
      end
      imageni=[imageni nrjsquare];
   end
   imagen=[imagen; imageni];
```

```
   imageni = [];
end
[I,J]=size(imagen);
im2write=flipdim(imagen,1);
imwrite((im2write.*length(map)),map,filename,'tiff');
set(findobj('tag','display1'),'visible','off' ...
      'string','Saving image...');


%Grabar Imagen del sonido en formato *.tiff.
function saveimage
global fich ind loc rl;
fich1=fich;%(1:8);
set(findobj('tag','display1'),'visible','on' ...
      'string','Saving image...');
handle=findobj('tag','ruido');
ruido=get(handle,'value');
snr = get(findobj('tag','SNR'),'string')
filename=['D:\Backup\Backup_Matlab\imagenes\qft\' fich1 '-'num2str(ind)
'-' num2str(ruido) '-' snr '.tiff'];
load sonie;
E=E./max(max(E));
Egray = E.*255;
name = [fich1 '-'num2str(ind) '-' num2str(ruido) '-' snr];
colormap(gray(255));
map=colormap;
%construccion de una imagen del sonido, cada cuadro de energia es de
10x16 pixeles.
imagen=[];
nrjsquare=[];
for i=1:16
   for j=1:80
      for r=1:5
         for l=1:1
            nrjsquare(r,l)=E(i,j);
         end
      end
      imageni=[imageni nrjsquare];
   end
   imagen=[imagen; imageni];
   imageni=[];
end
[I,J]=size(imagen)

imagegray=imagen.*255;
save  name imagegray
im2write=flipdim(imagen,1);
imwrite((im2write.*length(map)),map,filename,'tiff');
set(findobj('tag','display1'),'visible','off' ...
      'string','Saving image...');

%programa para guardar imagenes automaticamente.
%Este programa crea imagenes de la sonoridad con niveles de SNR
%que van desde low hasta upp de las palabras de los numeros en frances
%dicho por 29 locutores diferentes.
```

```
%Nombre del programa: autosaveimage.m
global loc rl;
global son backup_son fich;
noisetype=get(findobj('tag'.'ruido'),'value');
upp=40;
low=-20;
for ruido=upp:-1:low
   if noisetype==1
      ruido=0;
      upp=0;
      low=0;
   end
set(findobj('tag'.'SNR'),'string',ruido);
for rl=1:29
   loc = ['loc' num2str(rl) '\'] ;
   for i=1:10
      if i==1
         fich='cero';
      elseif i==2
         fich='uno';
      elseif i==3
         fich='dos';
      elseif i==4
         fich='tres';
      elseif i==5
         fich='cuatro';
      elseif i==6
         fich='cinco';
      elseif i==7
         fich='seis';
      elseif i==8
         fich='siete';
      elseif i==9
         fich='ocho';
      elseif i==10
         fich='nueve';
      end
      chem= 'D:\Backup\Backup_Matlab\numbers\';
      [A,FS]=wavread([chem loc fich '.wav']) ;
      son=A' ;
      son=(2^16)/max(max(son)) *son ;
      backup_son = son ;
      %plusnoise ;
      filtering ;
      close gcf ;
      saveimage ;
   end
end
end
```

```
%Programa para generar los numeros en formato WAV.
%Este programa lee la base de datos en formato SAM y convierte las pistas
en
%formato WAV.
global index E chem fich nbech son pistas ind;
global backup_son ;
global test_noise track file_name;
openbds ;
track=[2 17 22 23 41 52 53 65 70 85] ;
xtrak2write ;


% Programa para el análisis de las imágenes con morfología maetemática
function rmnoise(file,noise,loc)
global Tenergia;
Tenergia=0;
posicion=[
      3   384   122   132; ...
      132   382   122   132;  ..
      262   381   122   132;  ..
      392   381   122   132;  ..
      522   380   122   132;  ..
      652   380   122   132;  ..
      3   201   122   132; ...
      133   201   122   132;  ..
      263   201   122   132; ...
      393   201   122   132; ...
      523   202   122   132;
      653   202   122   132; ..
      3    20   122   132; ...
      133    20   122   132; ...
      262    19   122   132; ...
      392    19   122   132; ...
      522    19   122   132; ..
      652    19   122   132];

sonie2analyze = ['D:\Backup\Backup_Matlab\imagenes\loc' loc '\' file '--'
noise '-'] ;
s=1;
[xin_orig map] = imread(['D:\Backup\Backup_Matlab\imagenes\loc' loc '\
file '--1-0.tiff']);
b=mmsebox; %structurant element.
umbral=75;        %en 35 salen cosas interesantes.
l = mmthreshad( xin_orig, umbral );          %en 65 estaba bien.
y = mmareaopen(l,5,b);
z= mmopen(y,mmseline(5,90));
esqorig = mmthin( z, mmhomothin, -1, 45, 'clockwise' ) ;
figure(s);
   handle=gcf;
   set(handle,'position',posicion(s,:));
mmshow(mmneg(esqorig));
Eorig = sum(sum(esqorig));
for i=40:-5:-20
s=s+1;
[xin map] = imread([sonie2analyze num2str(i)  .tiff']);
%figure(1);
```

```
%mmshow(xin);            %visualizacion de la imagen sin ruido
l = mmthreshad( xin, umbral );
%figure(2) ;
%mmshow(l) ;
y = mmareaopen(l,5,b);
%figure(3);
%mmshow(y);
z= mmopen(y,mmseline(5,90));
%figure(4) ;
%mmshow(z) ;
r = mmthin( z, mmhomothin, -1, 45, 'clockwise' ) ;
%pause;
%figure(s) ;
%r=mmneg(r);
%mmshow(r) ;
p = mmintersec(r,esqorig) ;
%p = mmareaopen(p,5);
figure(s);
   handle=gcf;
   set(handle,'position',posicion(s,:));
p=mmneg(p);
%mmshow(mmneg(p),mmneg(esqorig)) ;
mmshow(p);
[M,N]=size(p);
p = mmneg(p);
p = double(p);
Tenergia=[Tenergia sum(sum(p))];
end
(100.*ones(1,length(Tenergia))-(Tenergia./Eorig).*100)
figure(s+1)
handle=gcf;
   set(handle,'position',posicion(s+1,:));
mmshow(xin_orig);
%[xin map] = imread('D:\Backup\Backup_Matlab\imagenes\loc1\ambiental.wav-
-1-0.tiff');
%l = mmthreshad( xin, umbral );
%y = mmareaopen(l,5,b);
%z= mmopen(y,mmsebox);%mmseline(5,90));
%r = mmthin( z, mmhomothin, -1, 45, 'clockwise' ) ;
%r=mmneg(r);
%q=mmneg(q);
%p = mmintersec(r,q) ;
%figure(s+1);
%p=mmneg(p);
%mmshow(p) ;
```

Apéndice 2: Imágenes de la Sonoridad

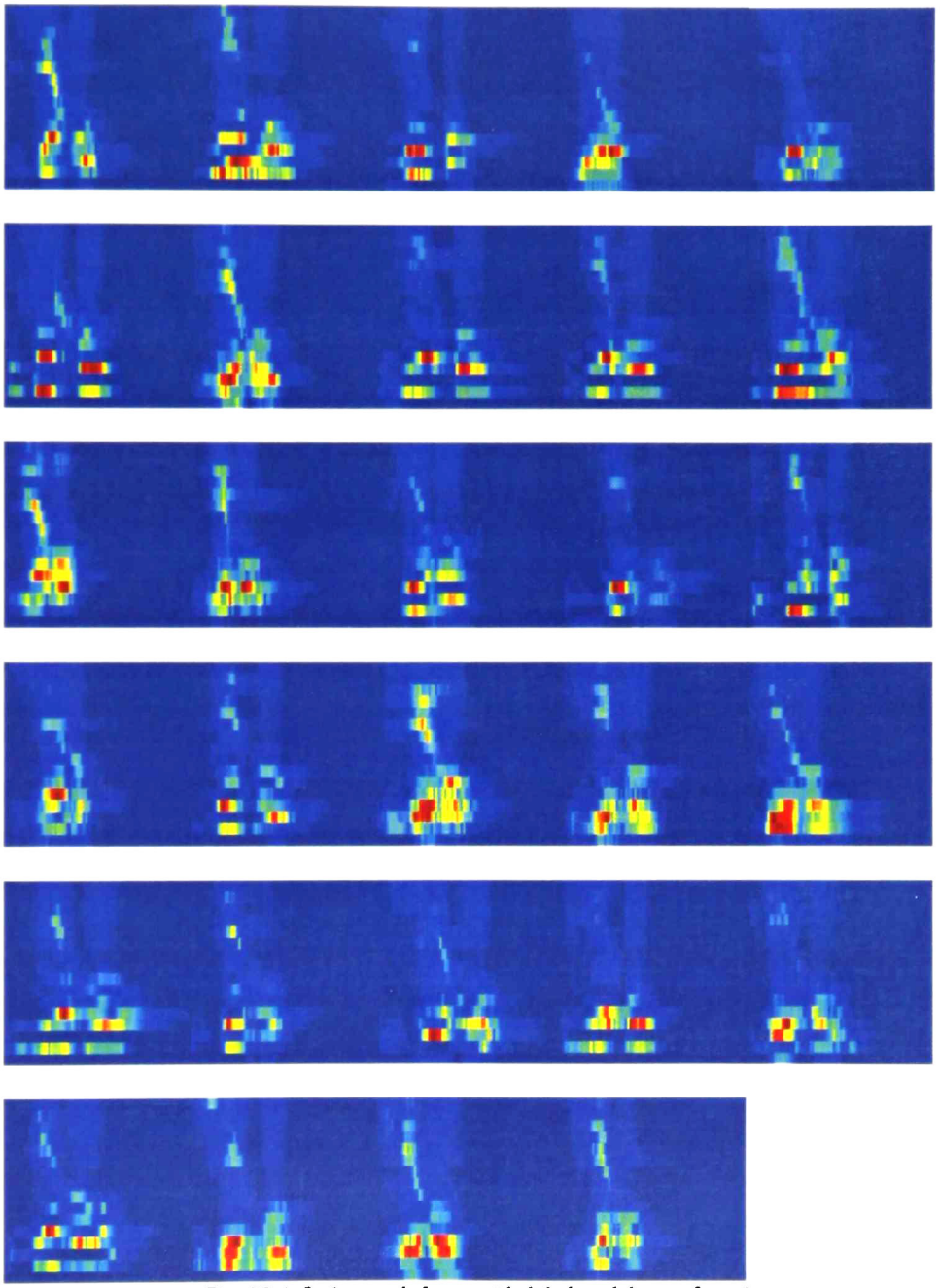


Fig. A2-1. Imágenes de la sonoridad de la palabra en francés "zero" dicho por 29 locutores diferentes.

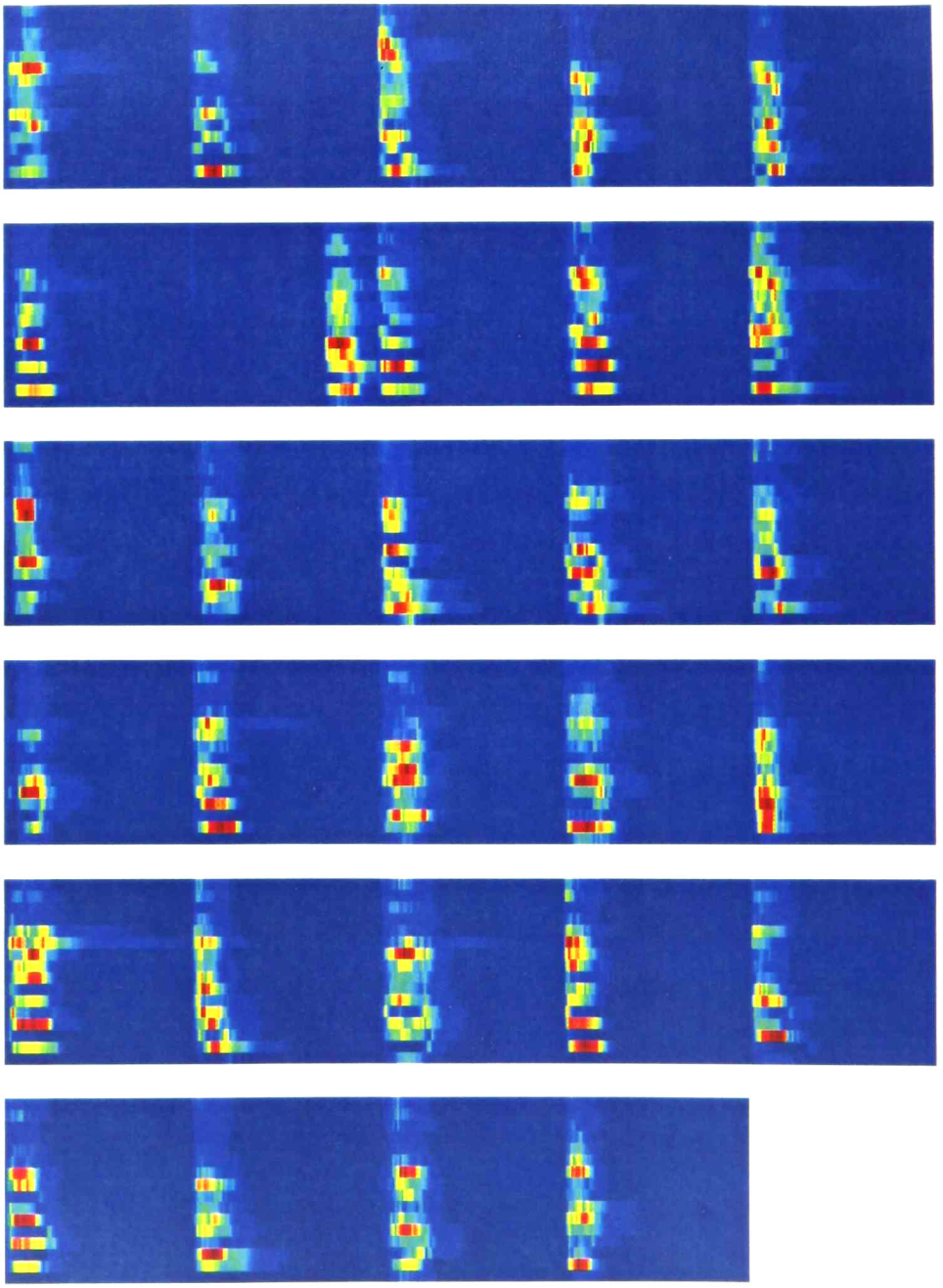


Fig. A2-2. Imágenes de la sonoridad de la palabra en francés "un" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

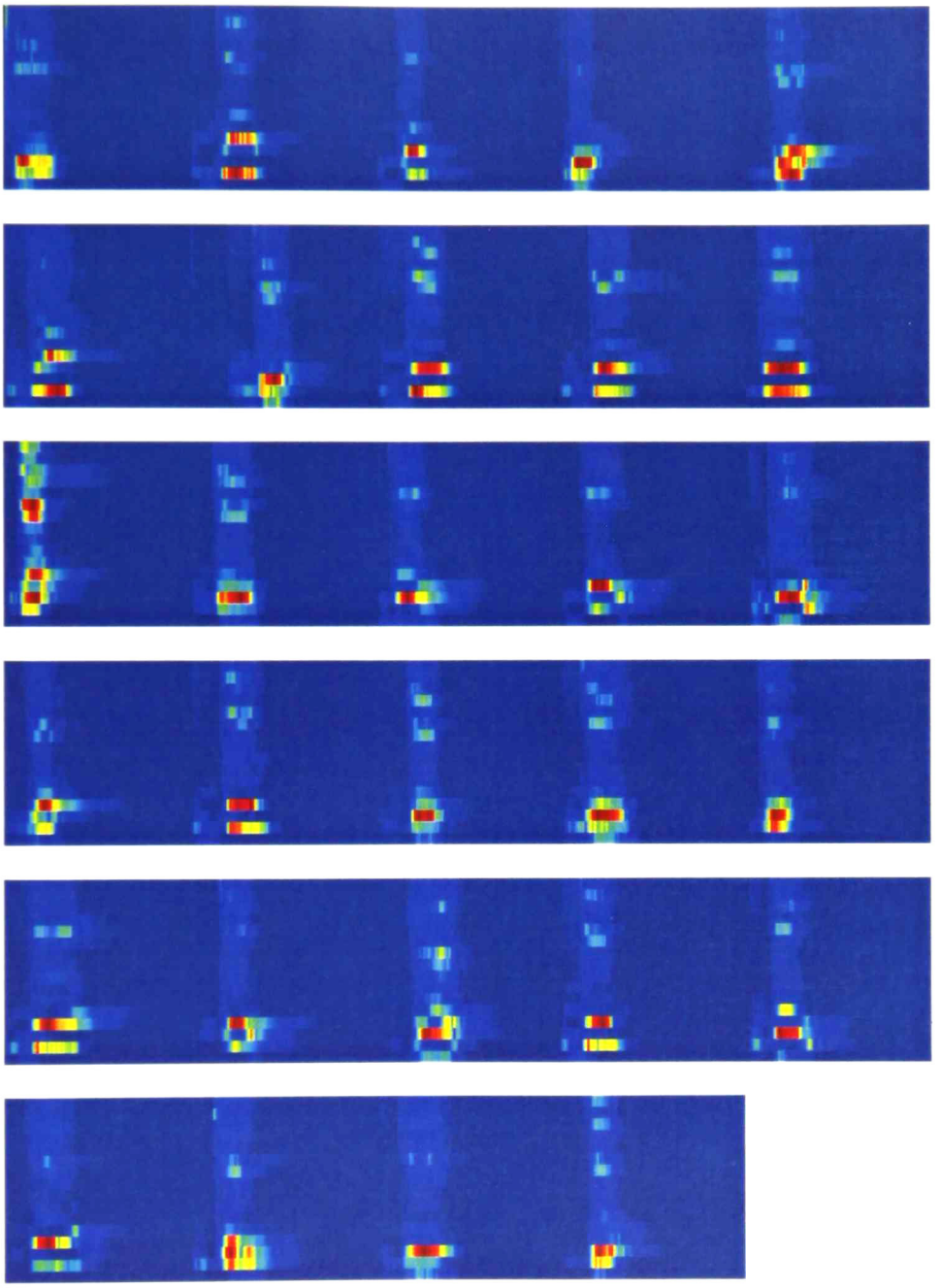


Fig. A2-3. Imágenes de la sonoridad de la palabra en francés "due" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

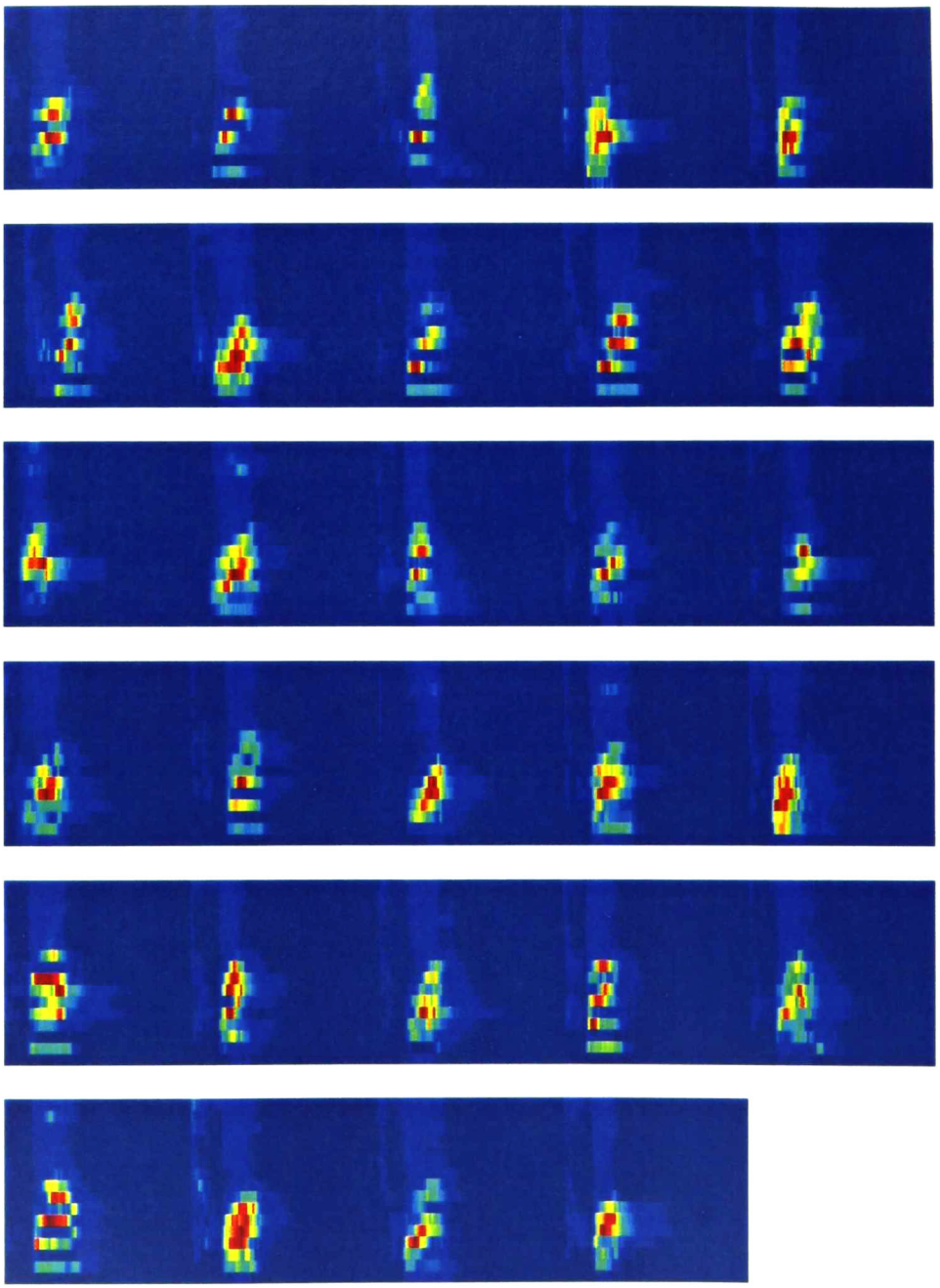


Fig. A2-4. Imágenes de la sonoridad de la palabra en francés "trois" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

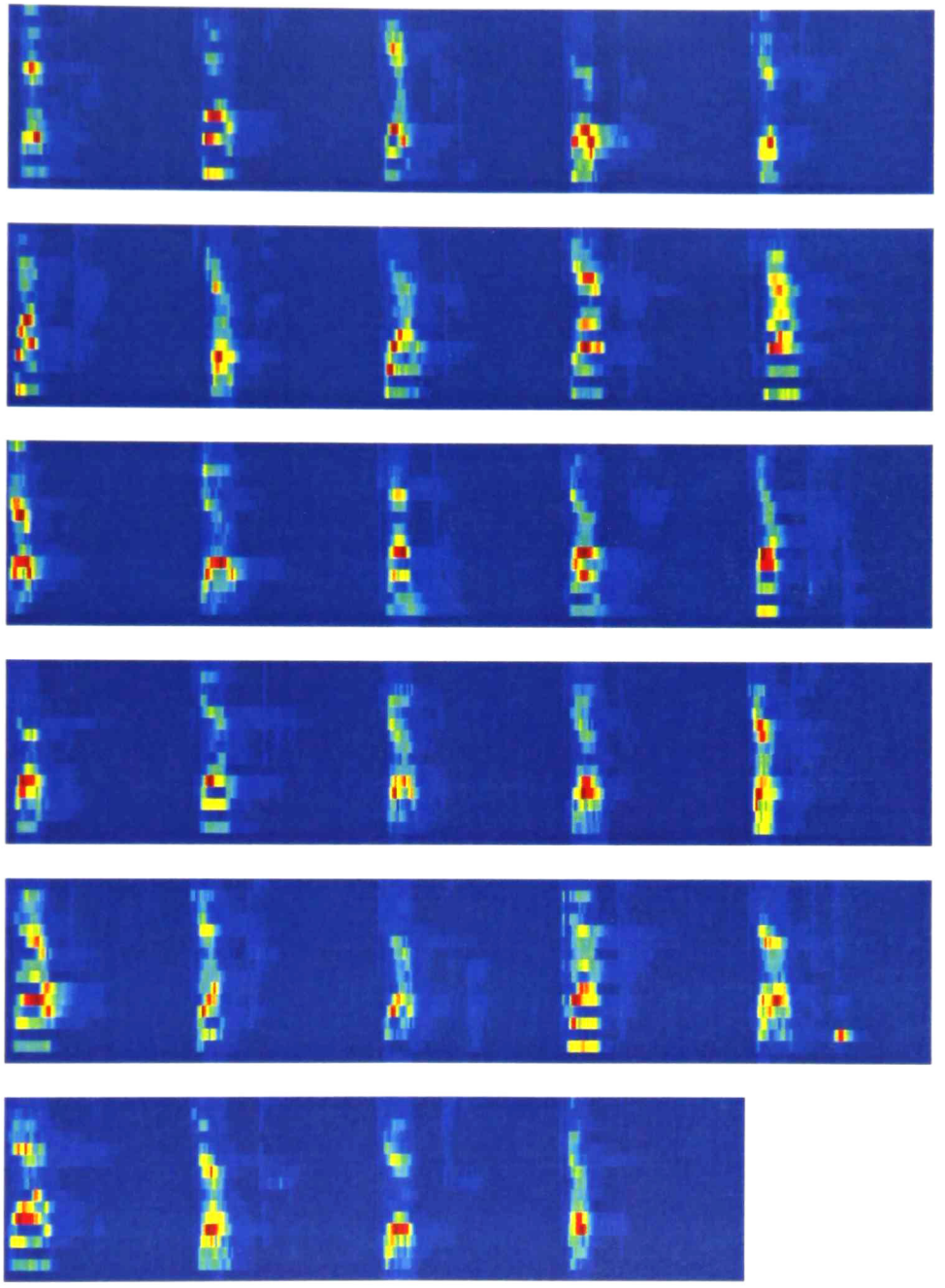


Fig. A2-5. Imágenes de la sonoridad de la palabra en francés "quatre" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

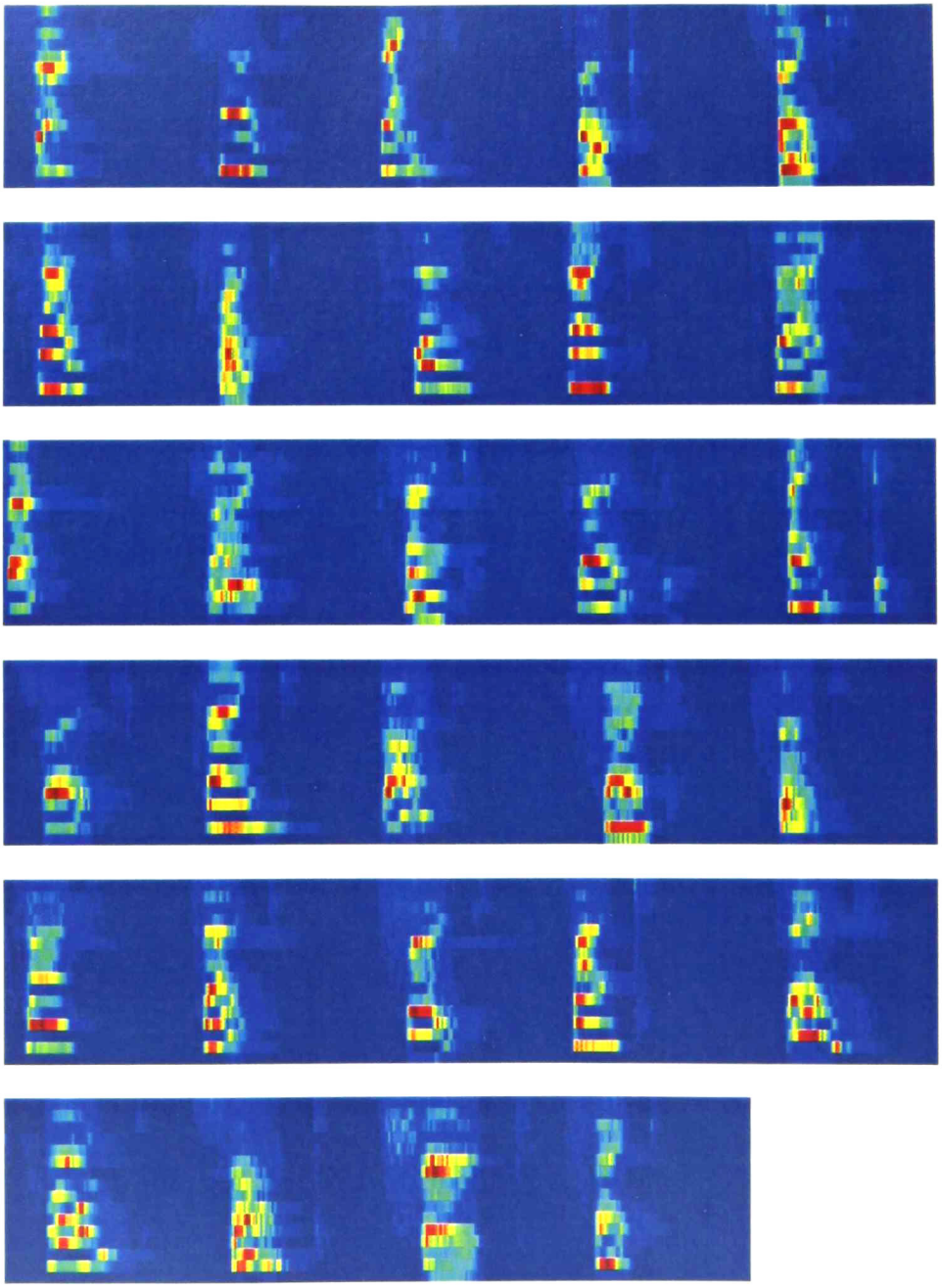


Fig. A2-6. Imágenes de la sonoridad de la palabra en francés "cinq" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

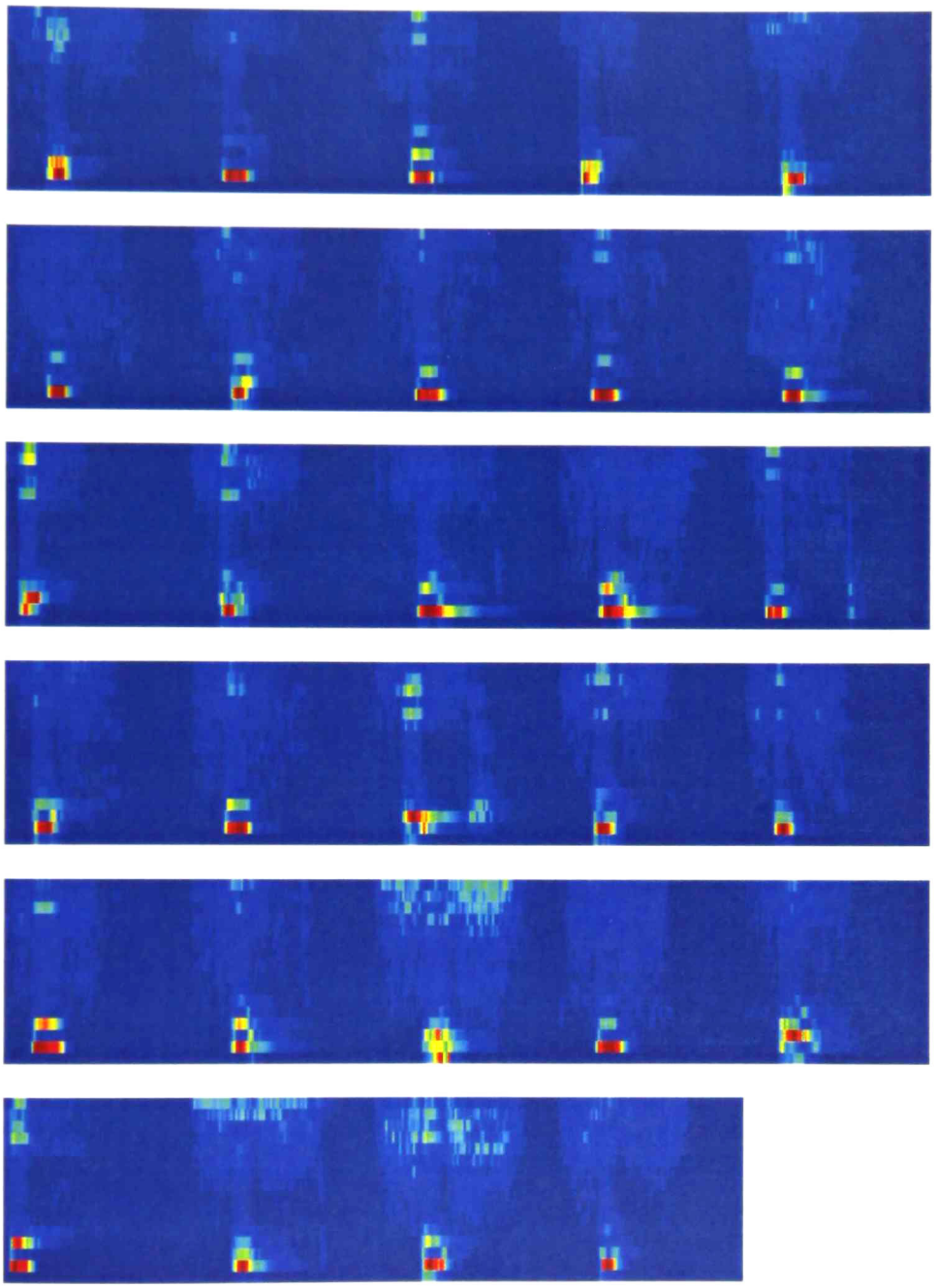


Fig. A2-7. Imágenes de la sonoridad de la palabra en francés "six" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

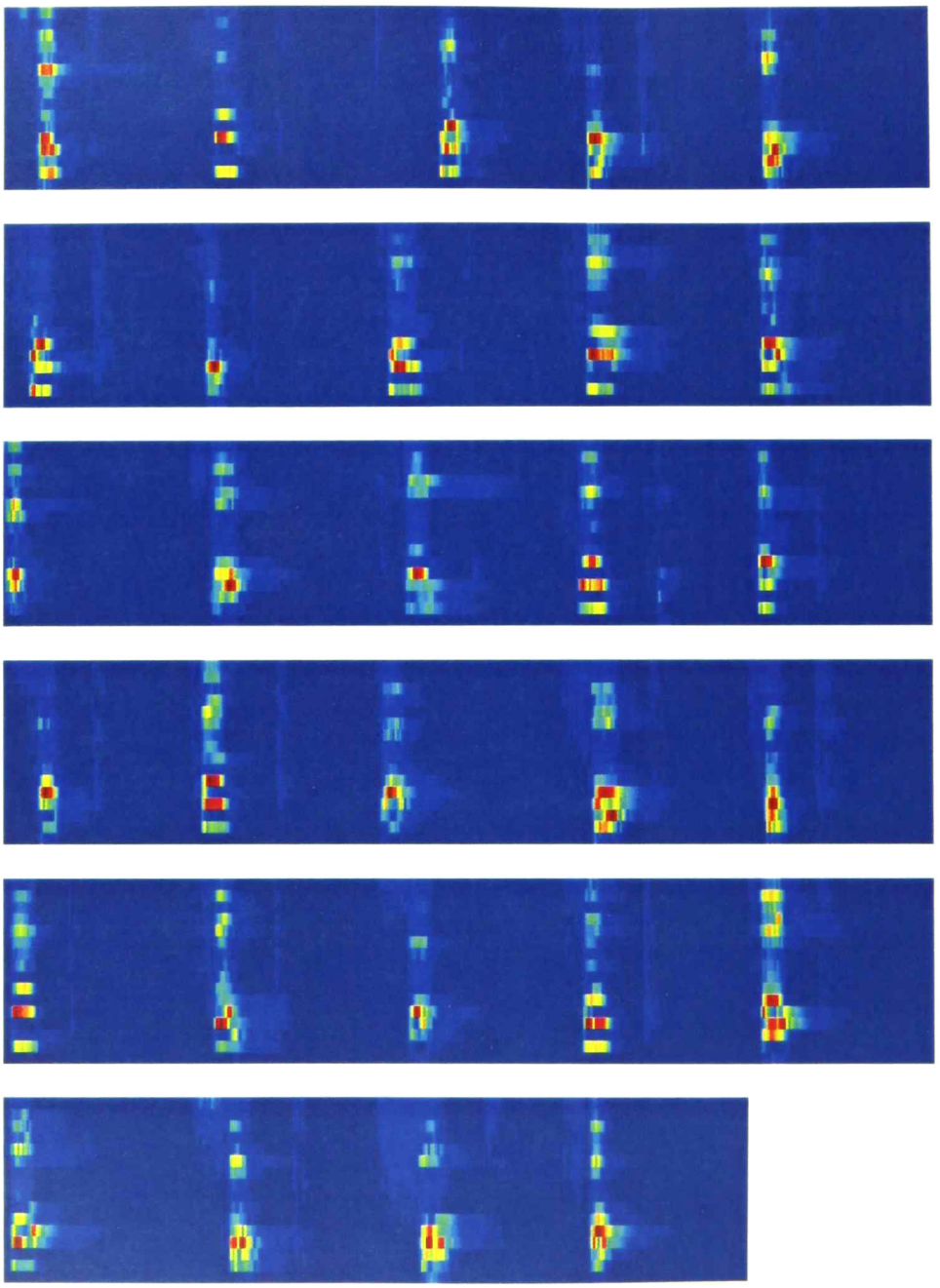


Fig. A2-8. Imágenes de la sonoridad de la palabra en francés "sept" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

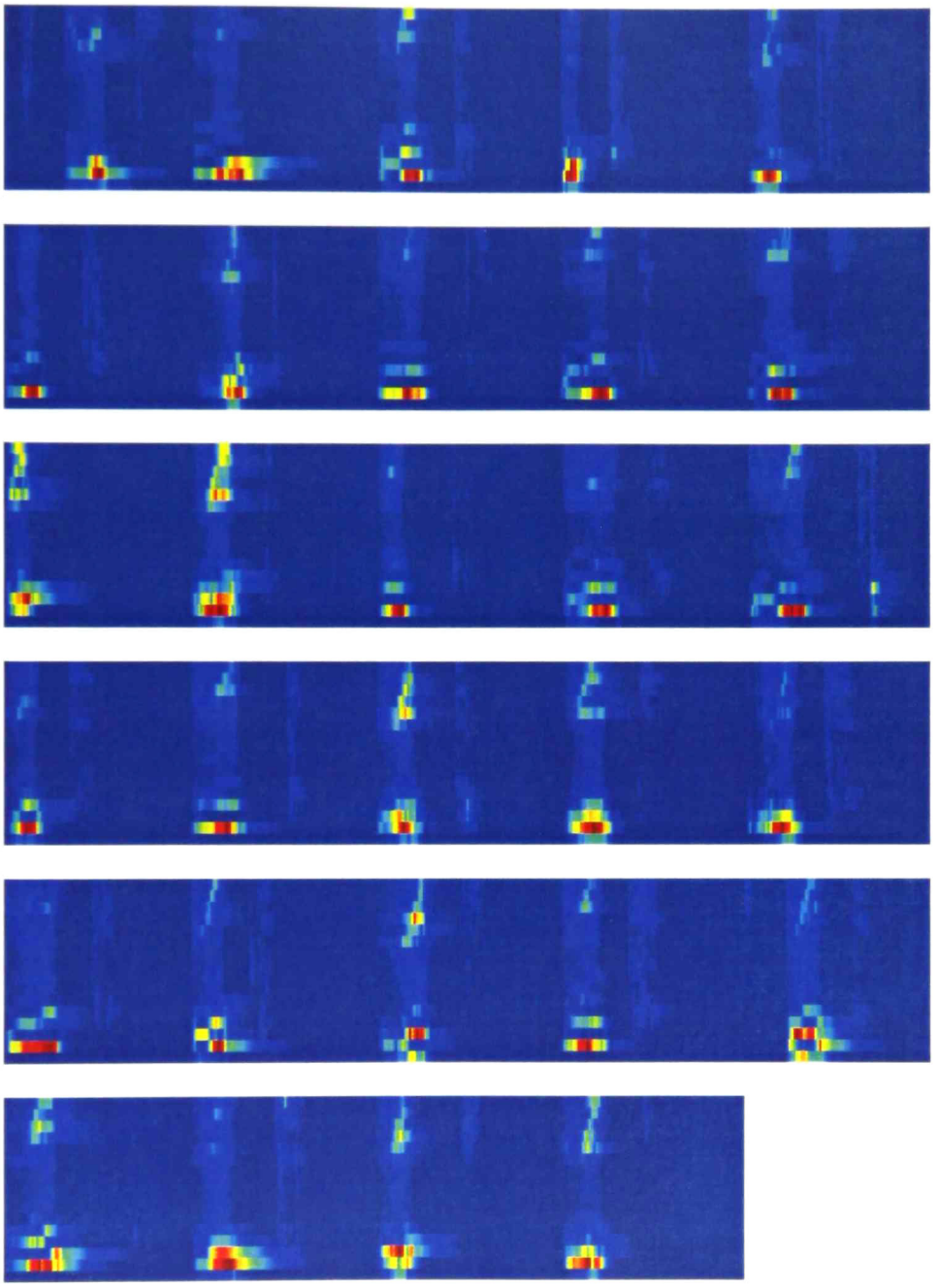


Fig. A2-9. Imágenes de la sonoridad de la palabra en francés "huit" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

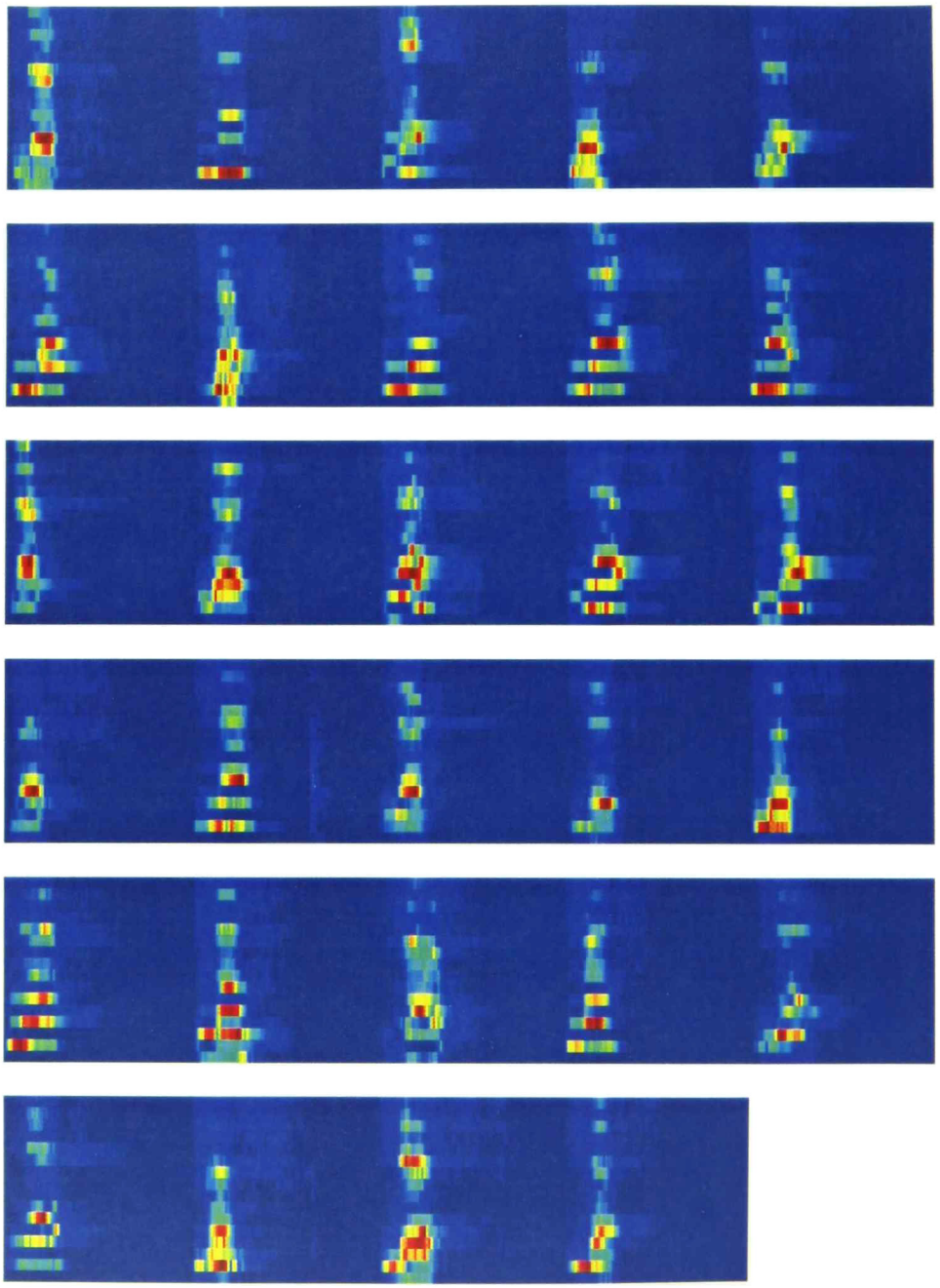


Fig. A2-10. Imágenes de la sonoridad de la palabra en francés "neuf" dicho por 29 locutores diferentes. Las señales de voz no han sido perturbadas por ningún tipo de ruido.

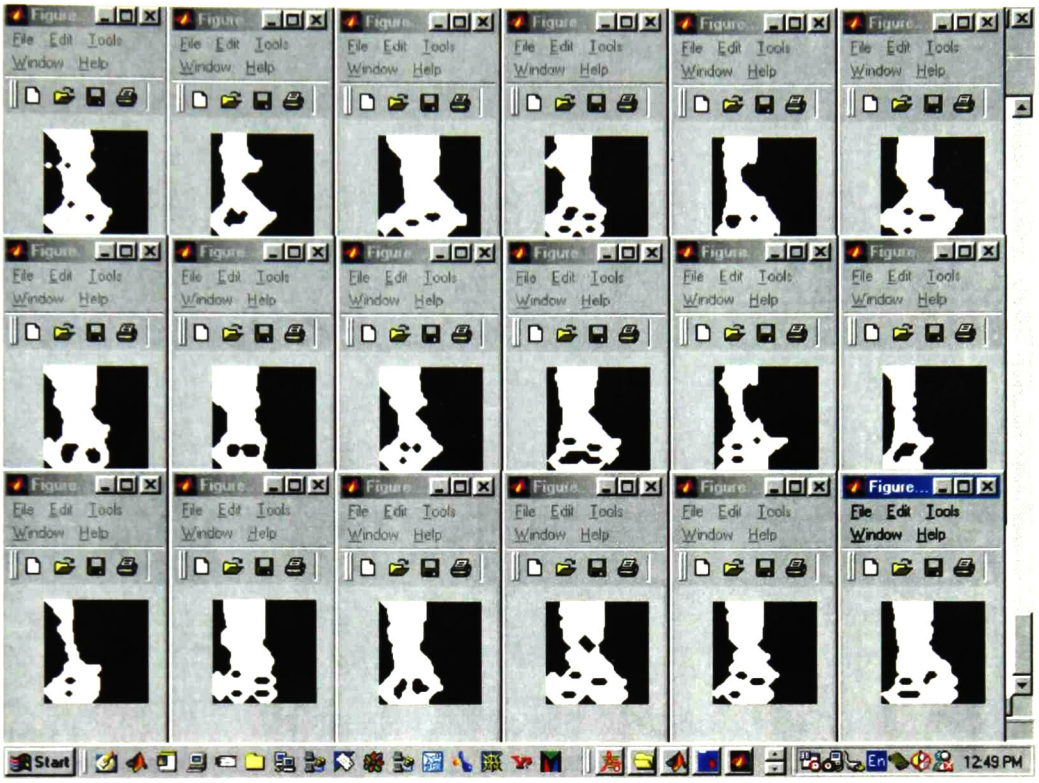


Fig. A2-11. Imágenes de la palabra "zero" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

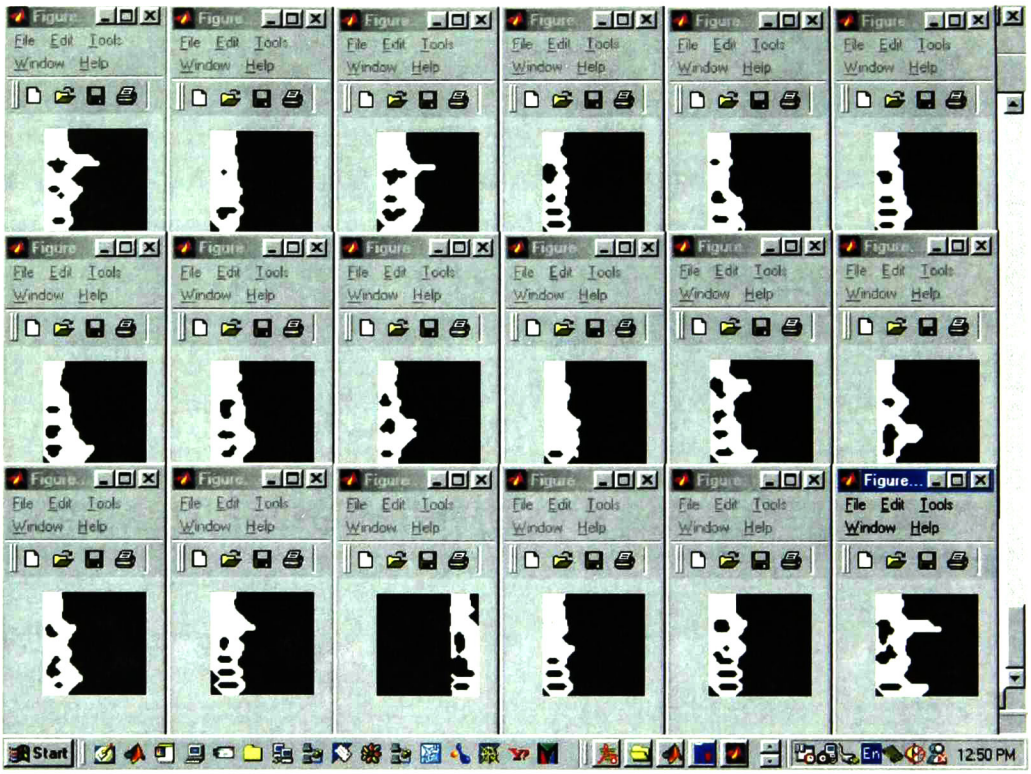


Fig. A2-12. Imágenes de la palabra "un" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

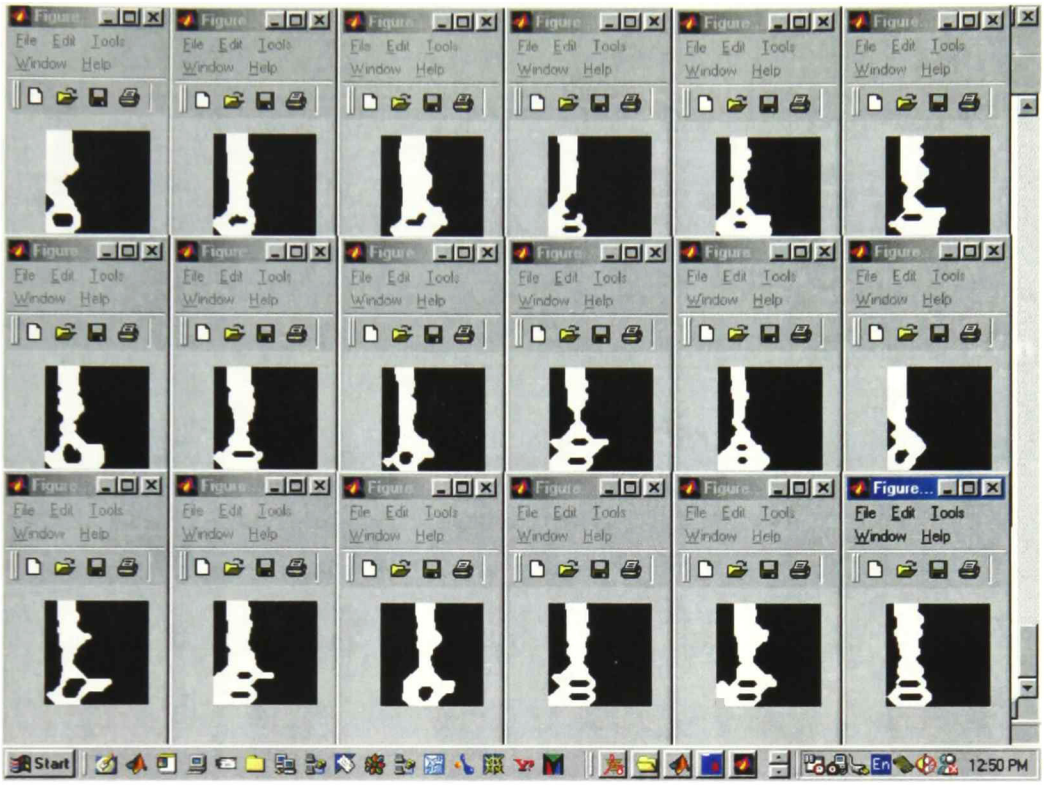


Fig. A2-13. Imágenes de la palabra "due" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

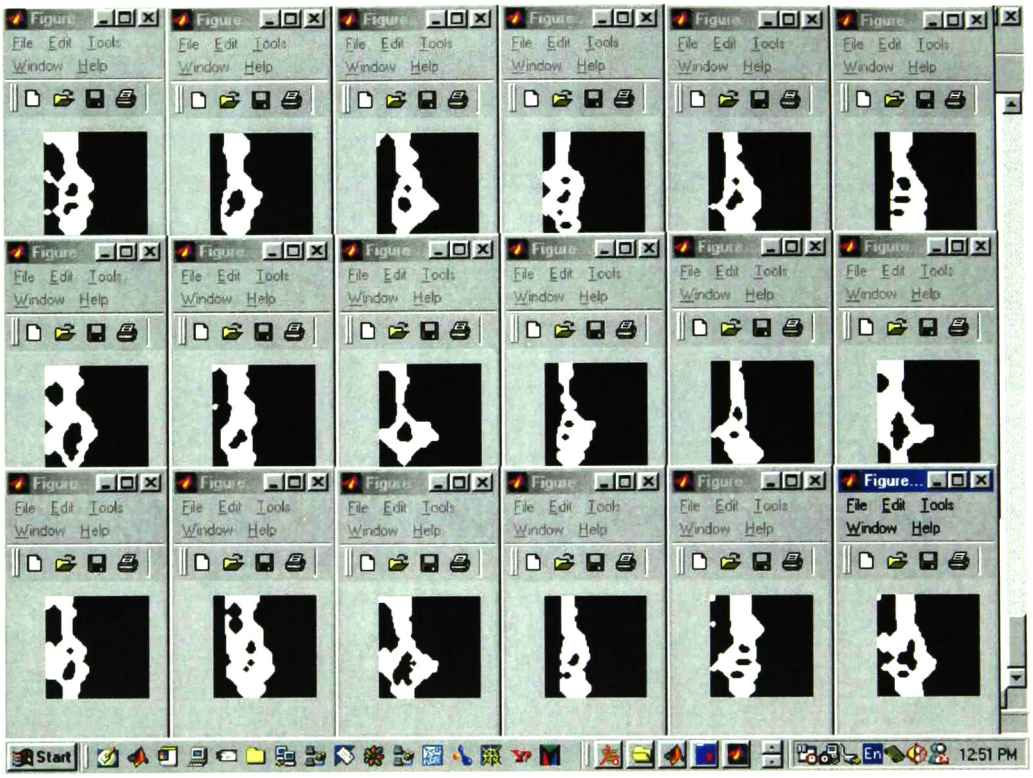


Fig. A2-14. Imágenes de la palabra "trois" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

Fig. A2-15. Imágenes de la palabra "quatre" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

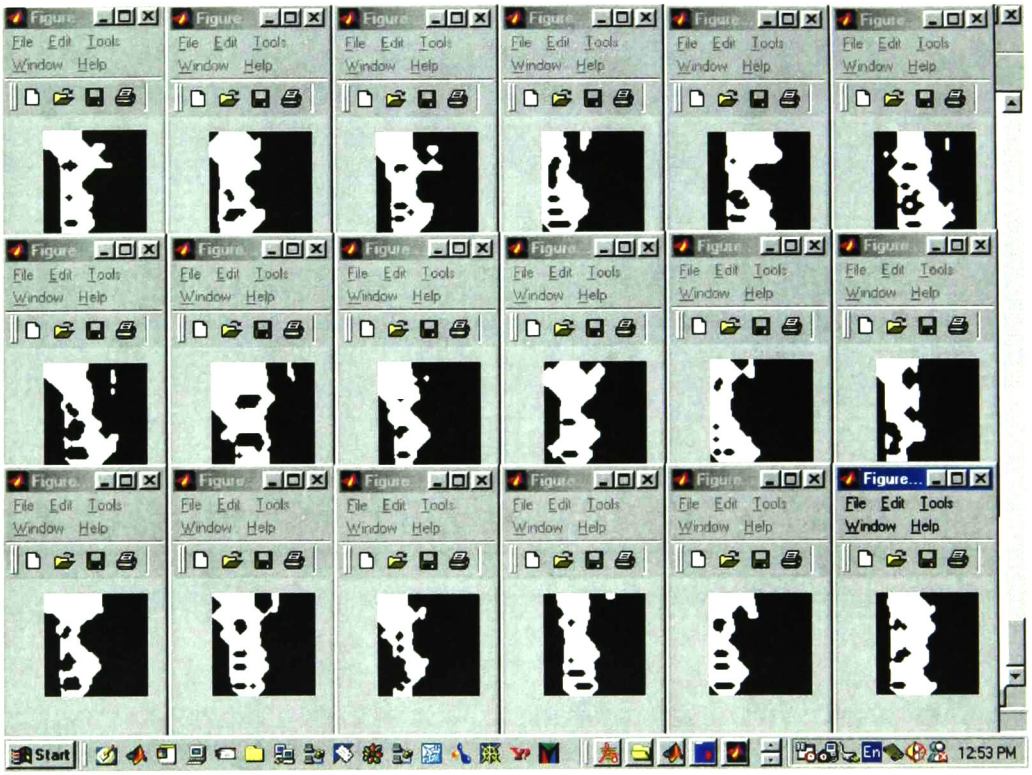


Fig. A2-16. Imágenes de la palabra "cinq" en francés, dicha por 29 locutores diferentes, tratadas con morfología matemática para extraer su forma.

Descomposición cuaterniónica de las imágenes de la Sonoridad.

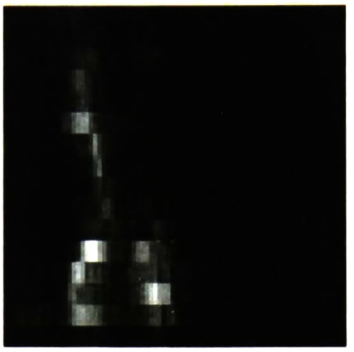


Figura A2-17. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 40dB.

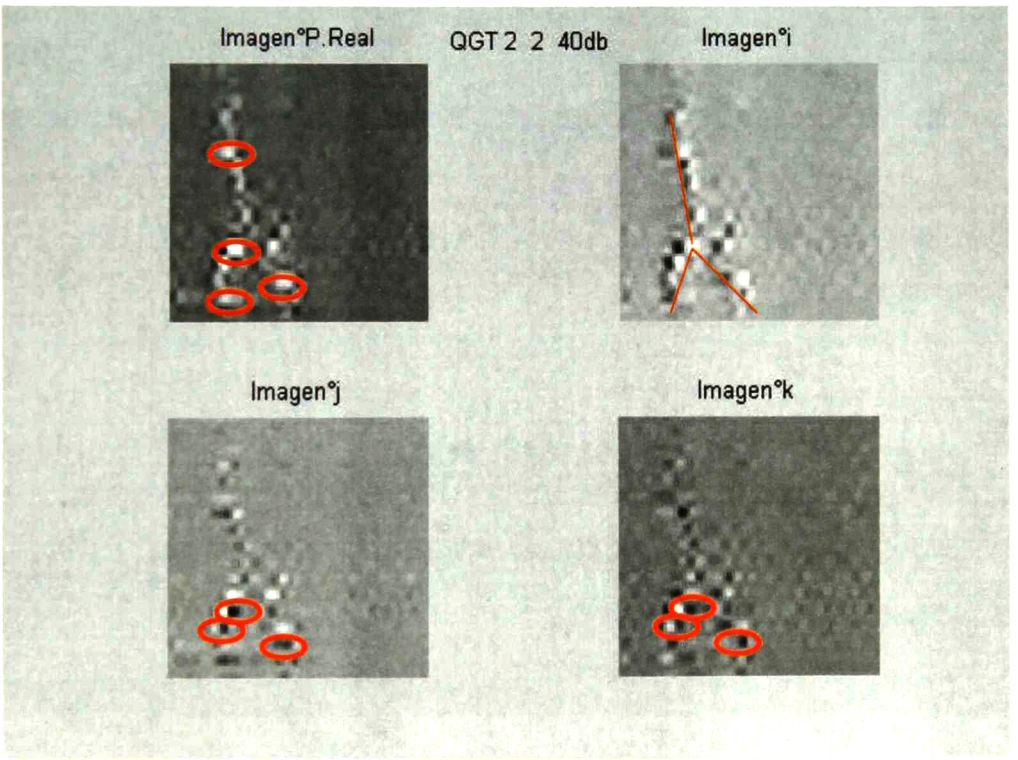


Figura A2-18. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

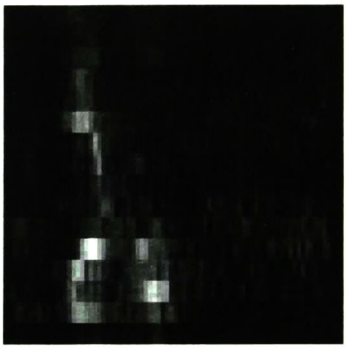


Figura A2-19. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 20dB.

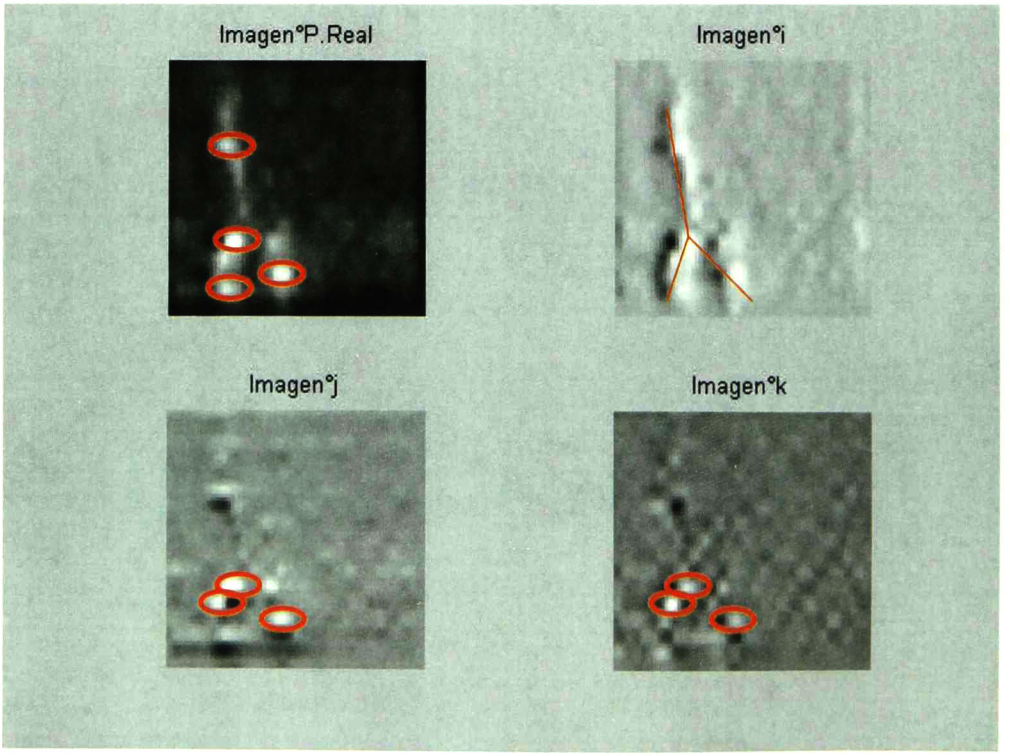


Figura A2-20. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

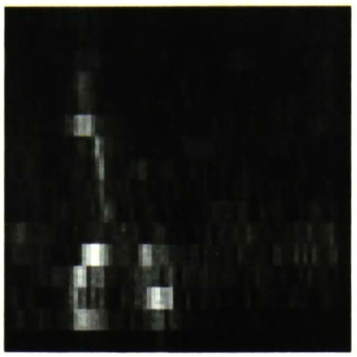


Figura A2-21. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 10dB.

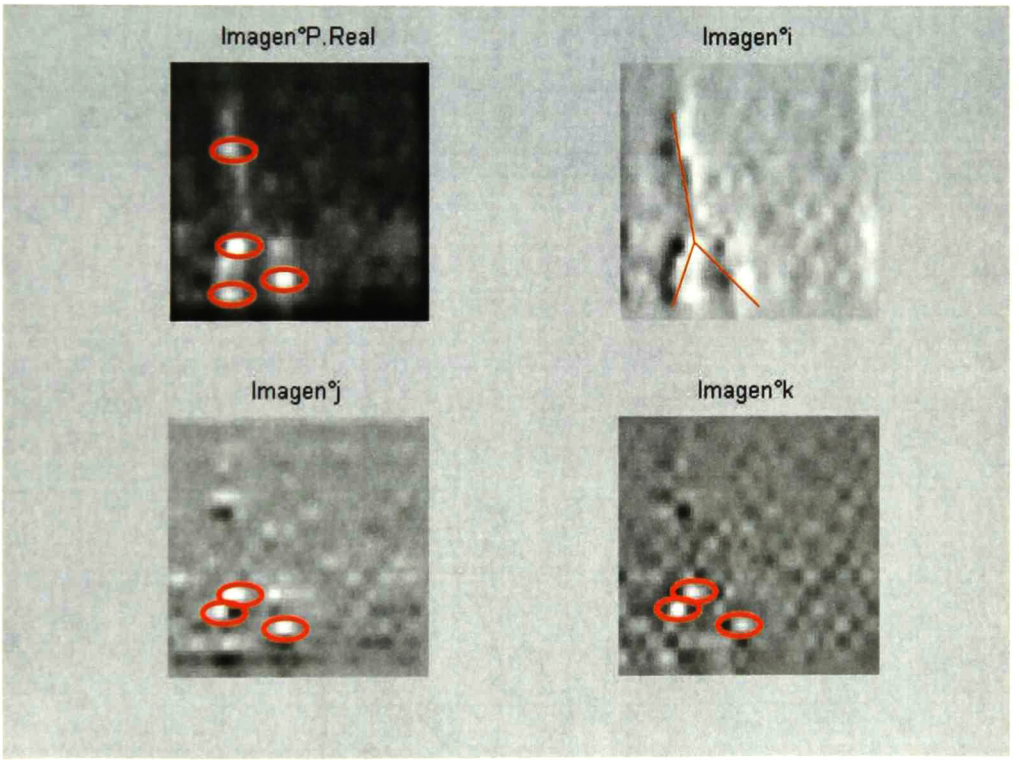


Figura A2-22. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

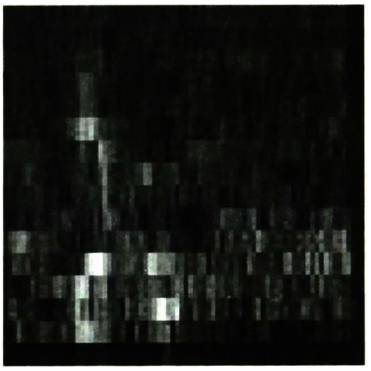


Figura A2-23. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 0dB.

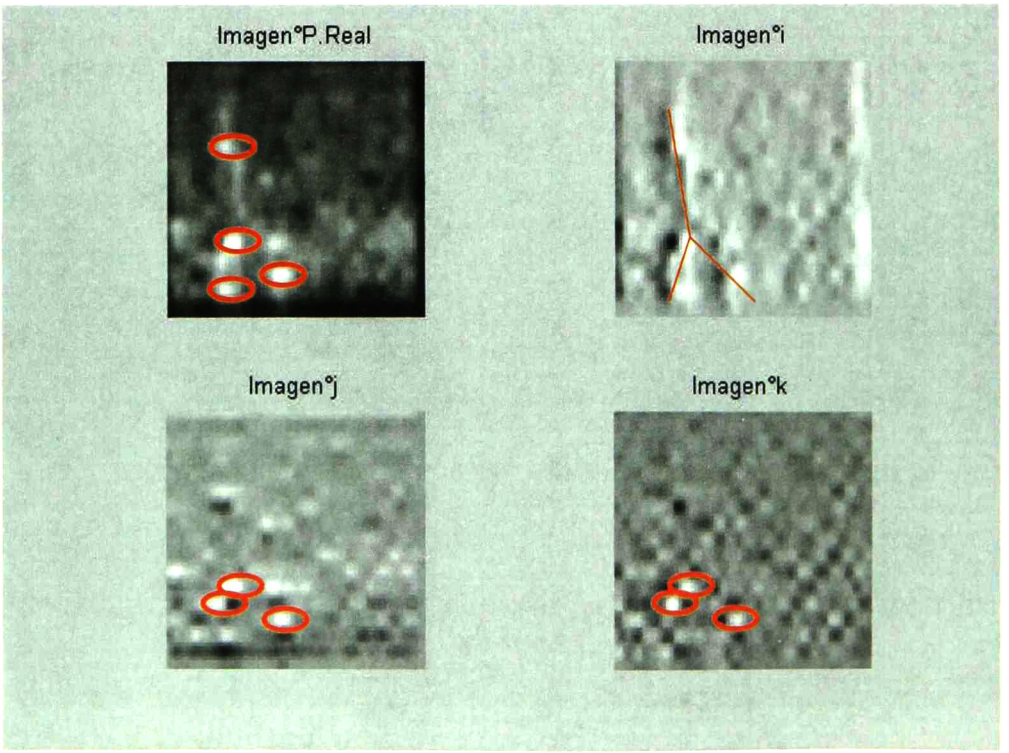


Figura A2-24. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

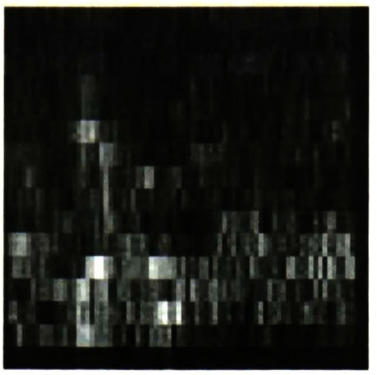


Figura A2-25. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -5dB.

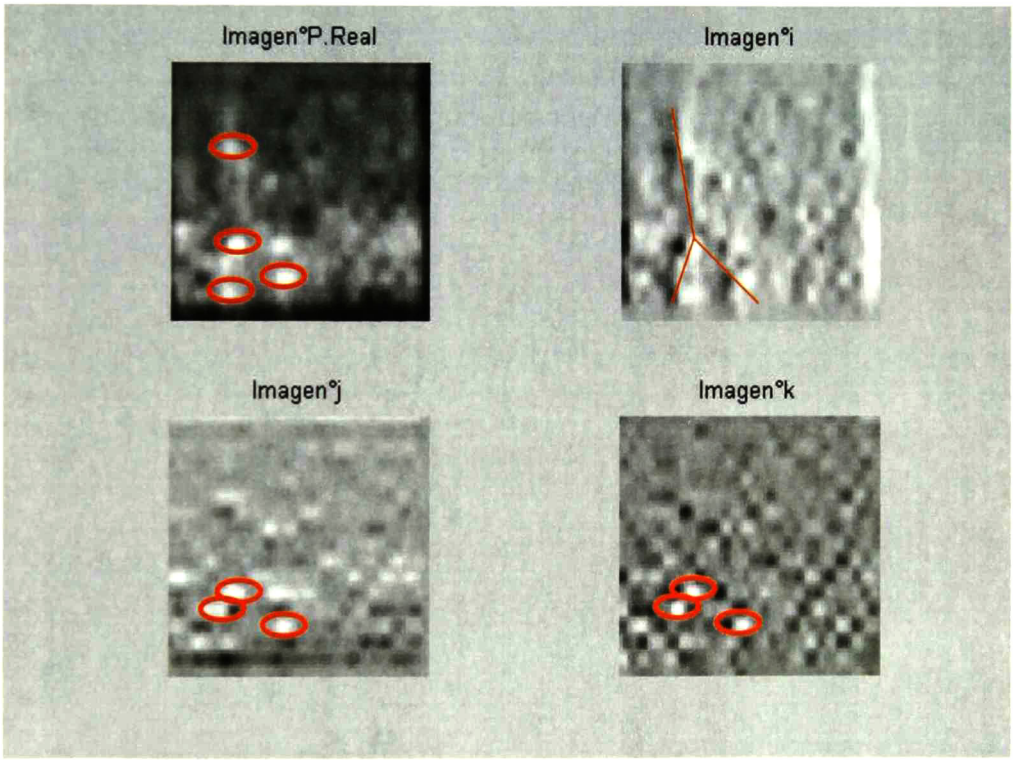


Figura A2-26. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

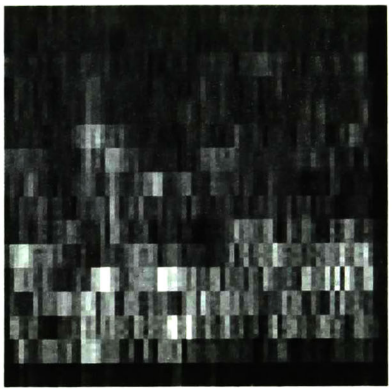


Figura A2-27. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -10dB.

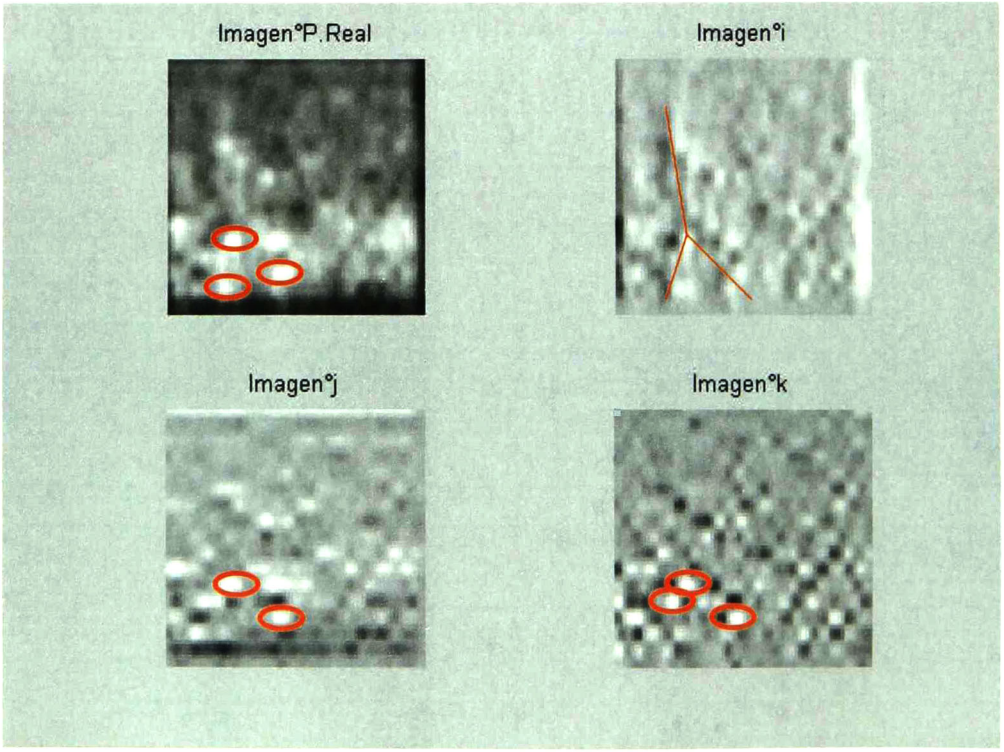


Figura A2-28. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

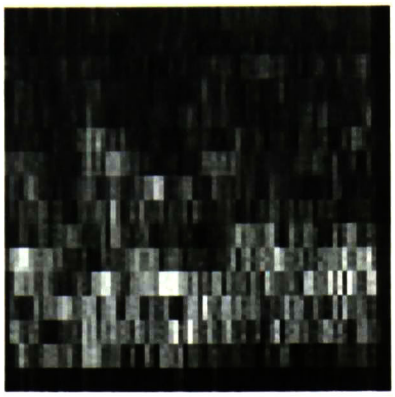


Figura A2-29. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -15dB.

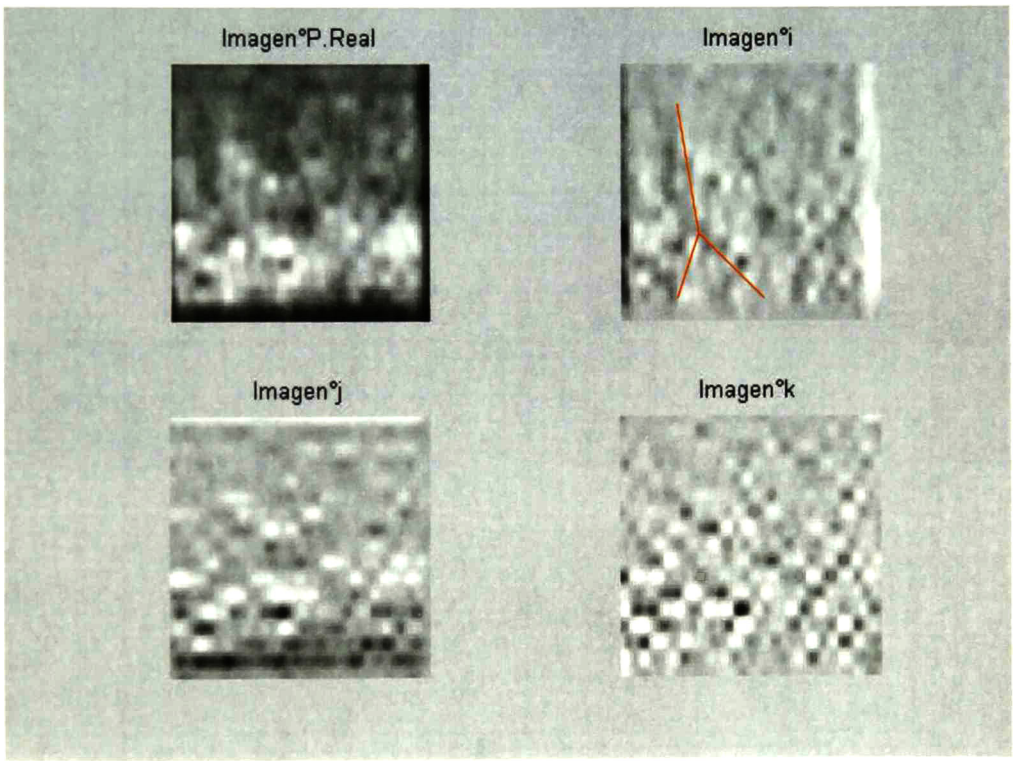


Figura A2-30. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

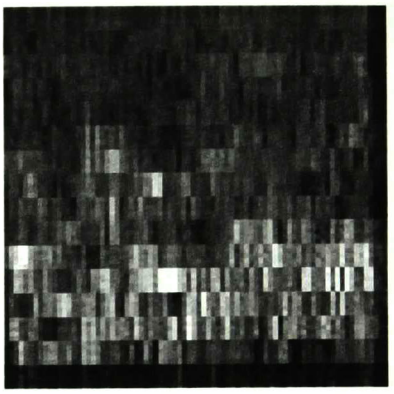


Figura A2-31. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -20dB.

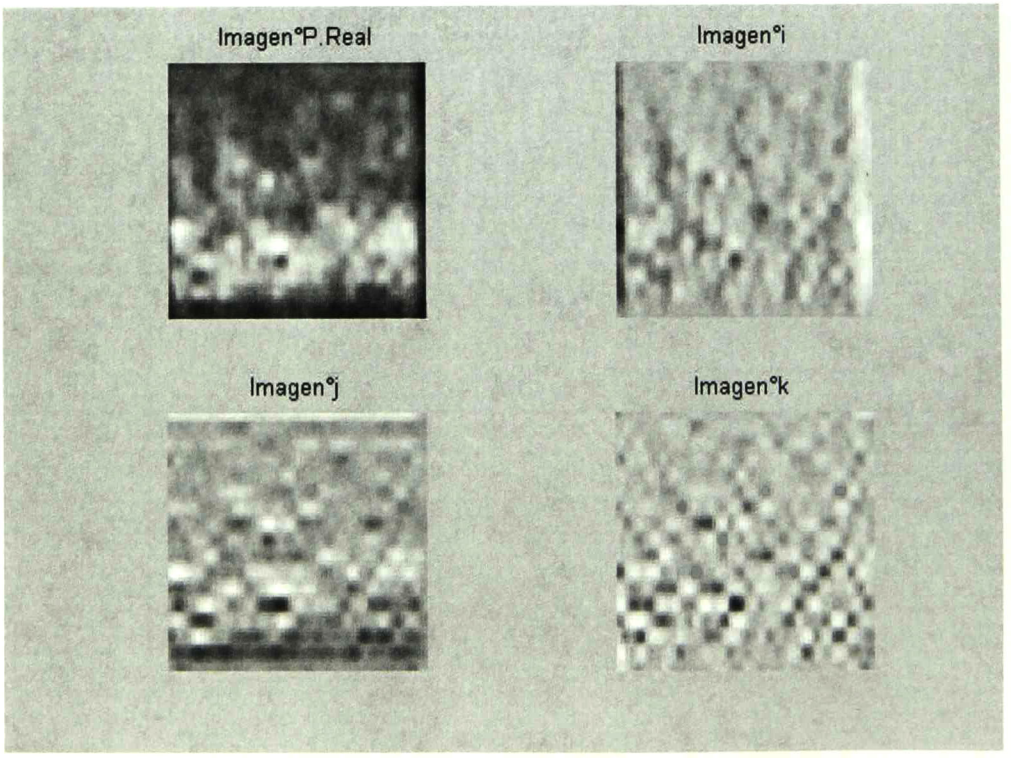


Figura A2-32. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

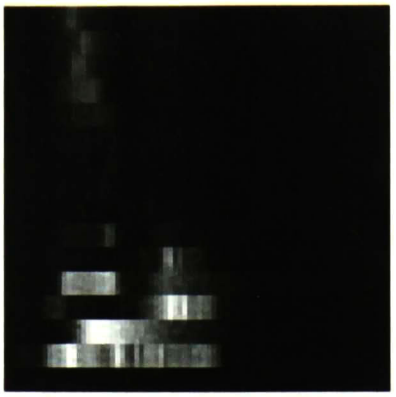


Figura A2-33. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Sin ruido. SNR= 0dB.

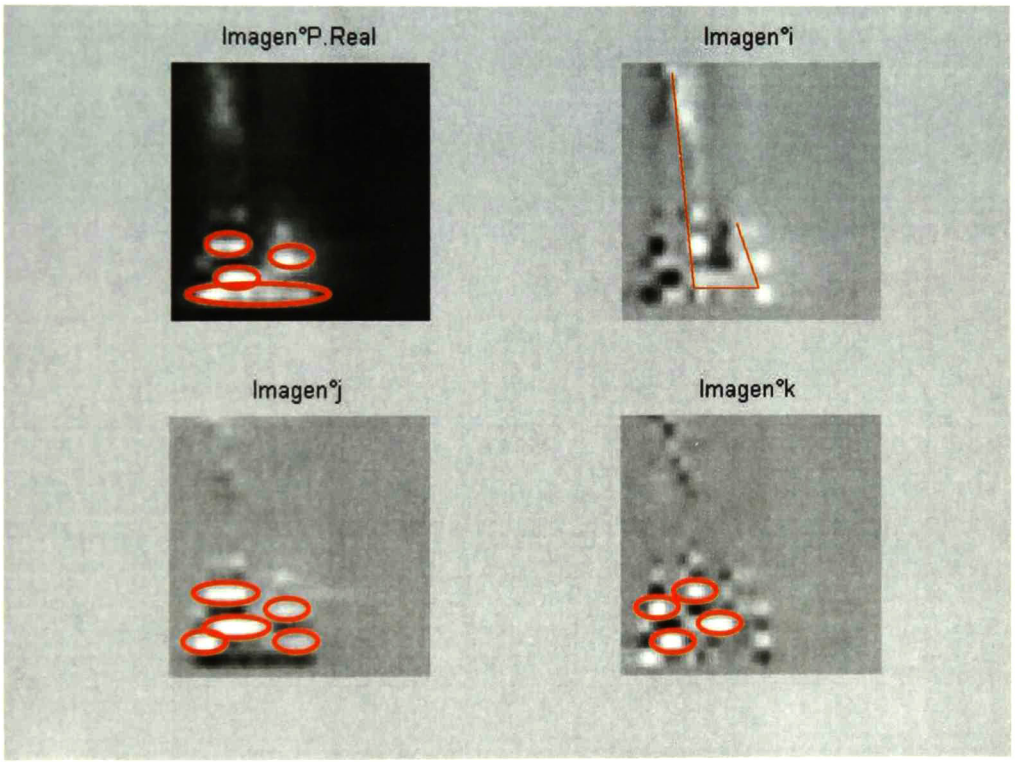


Figura A2-34. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

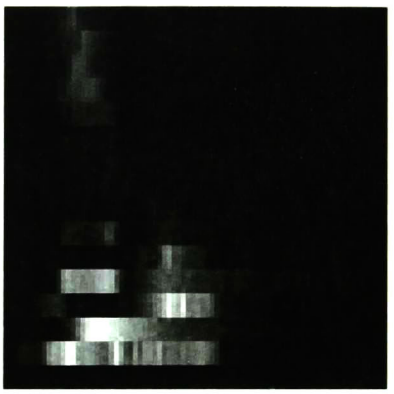


Figura A2-35. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= 40dB.

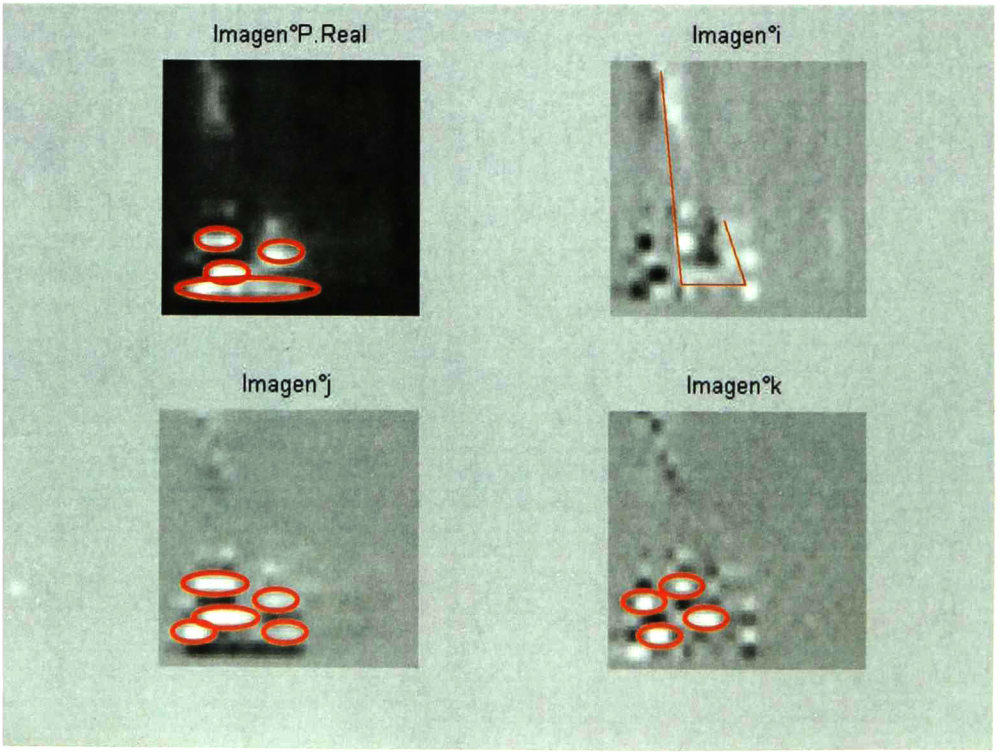


Figura A2-36. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

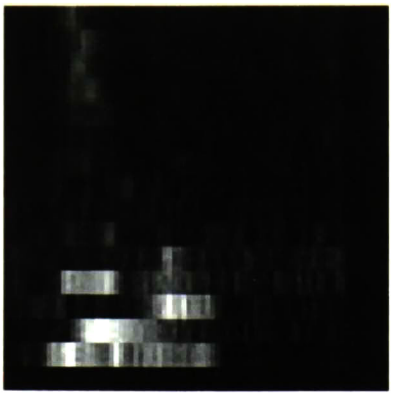


Figura A2-37. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= 20dB.

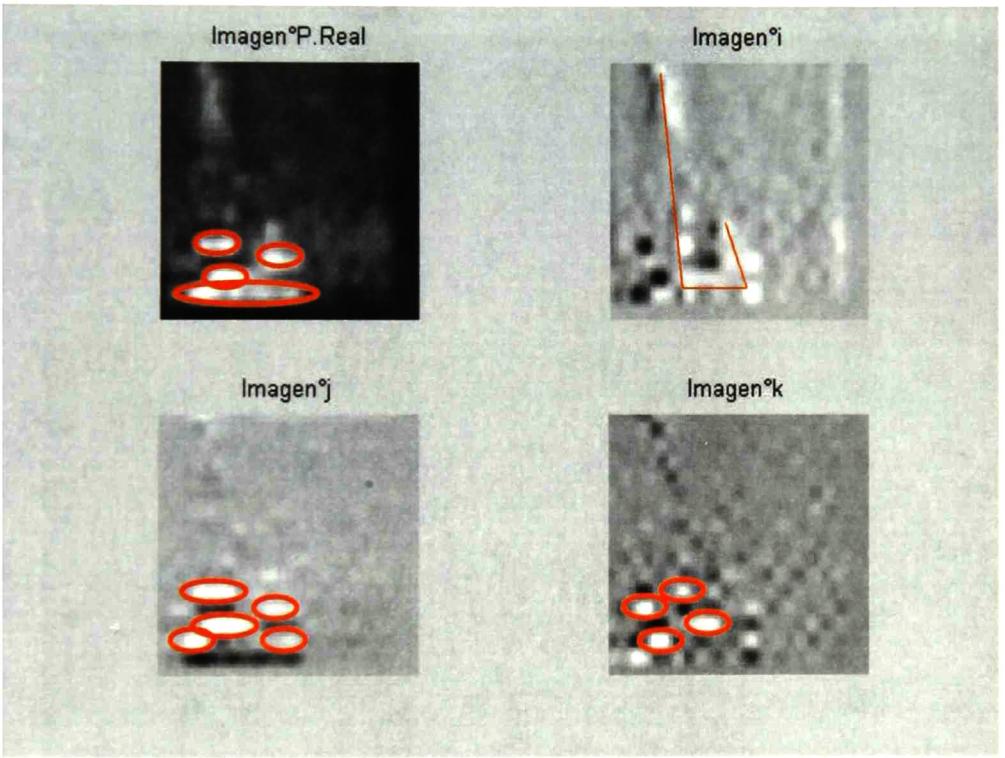


Figura A2-38. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

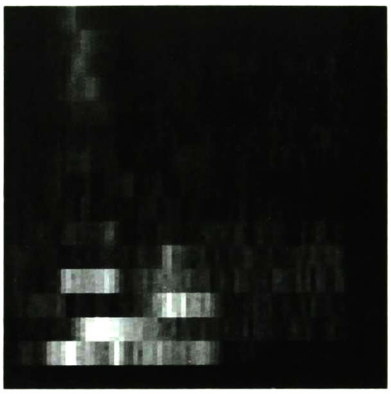


Figura A2-39. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= 10dB.

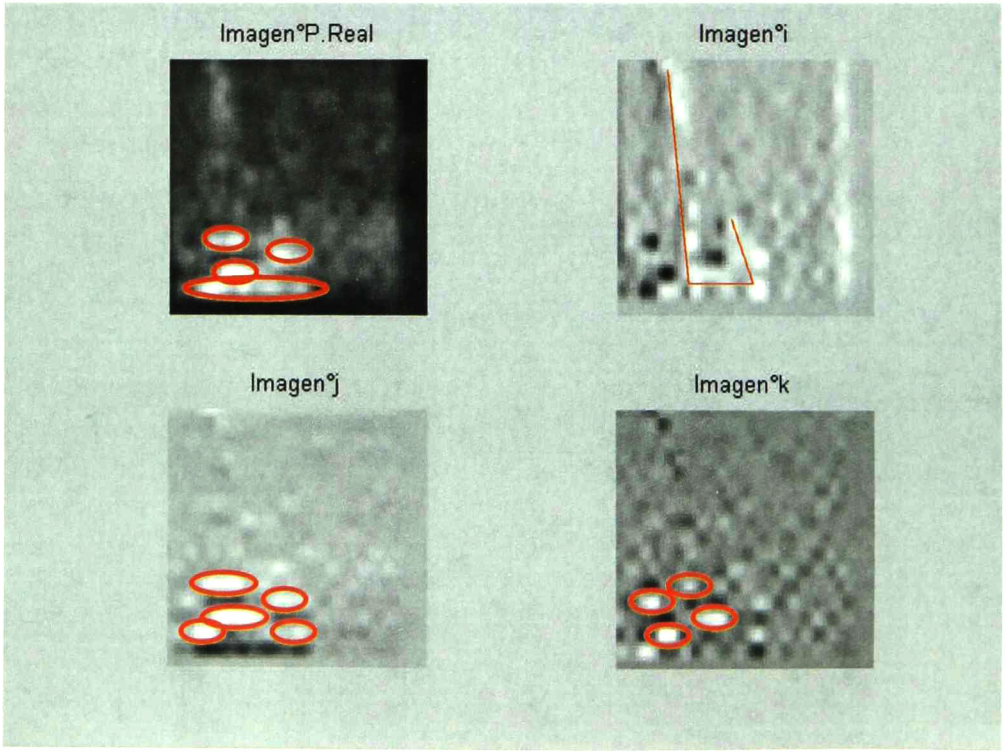


Figura A2-40. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

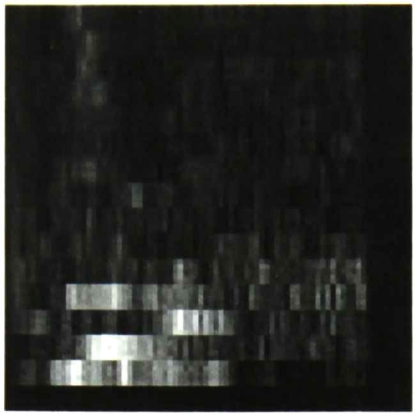


Figura A2-41. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= 5dB.

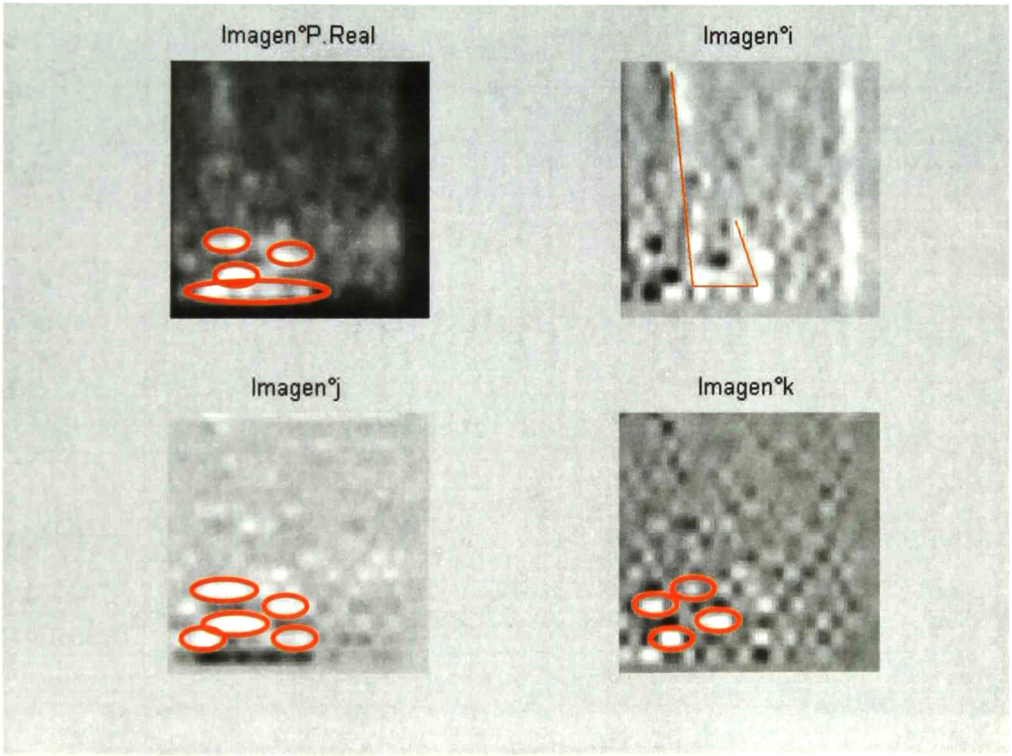


Figura A2-42. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

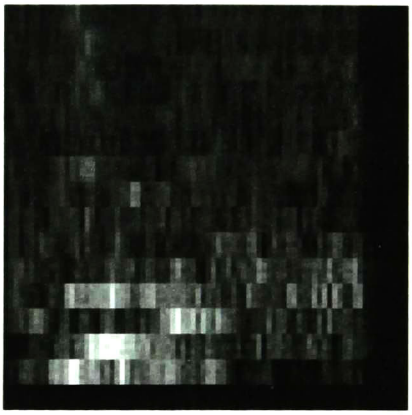


Figura A2-43. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= 0dB.

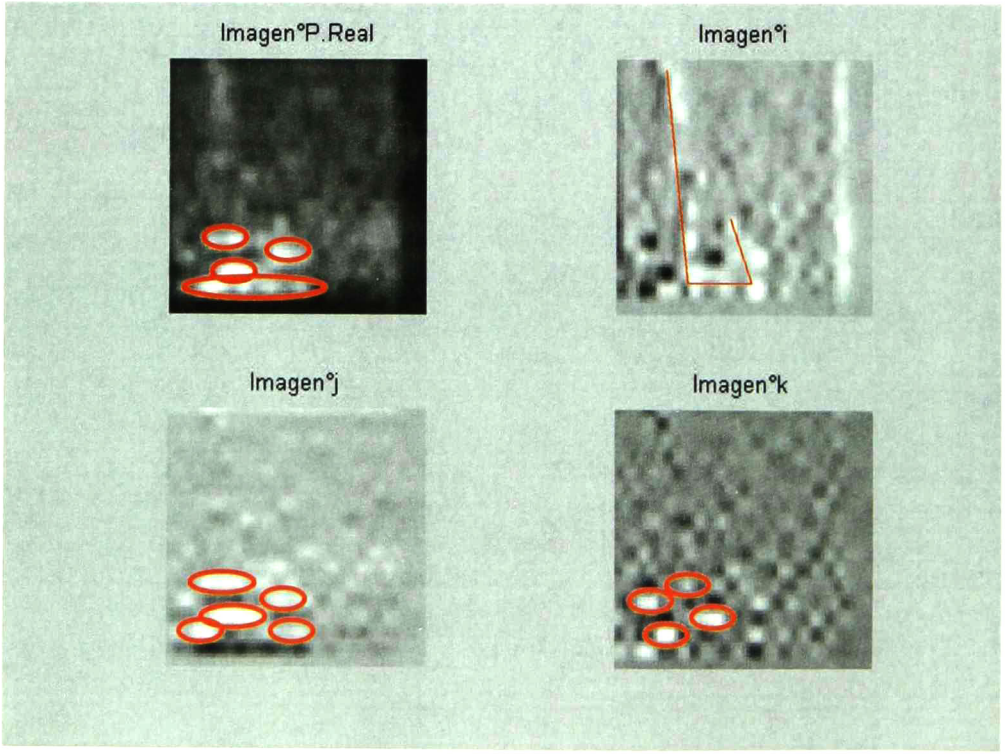


Figura A2-44. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

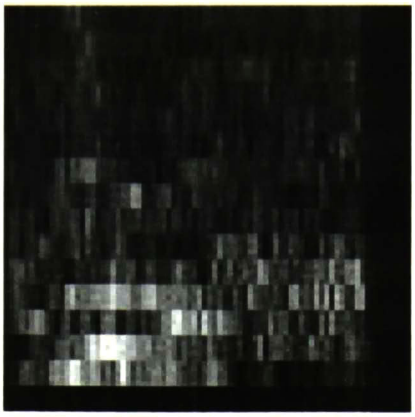


Figura A2-45. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= -5dB.

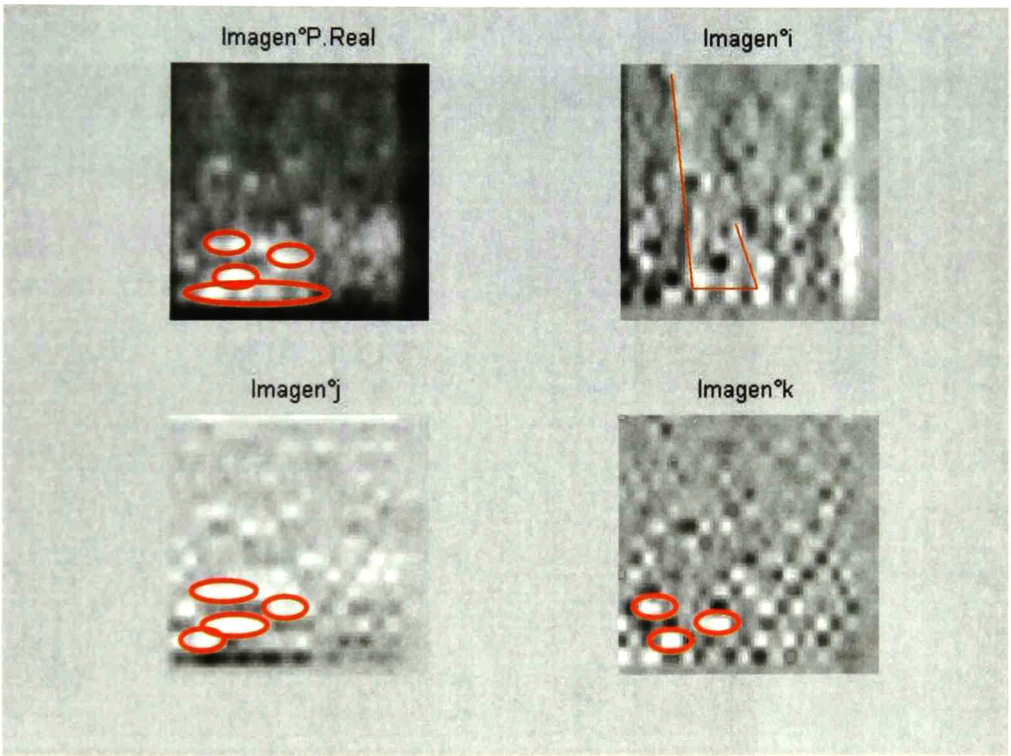


Figura A2-46. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

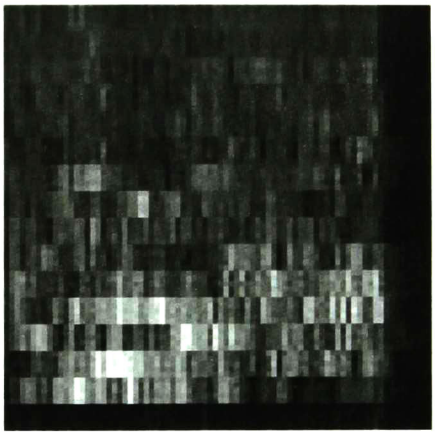


Figura A2-47. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 10. Tipo de ruido: Conversación. SNR= -10dB.

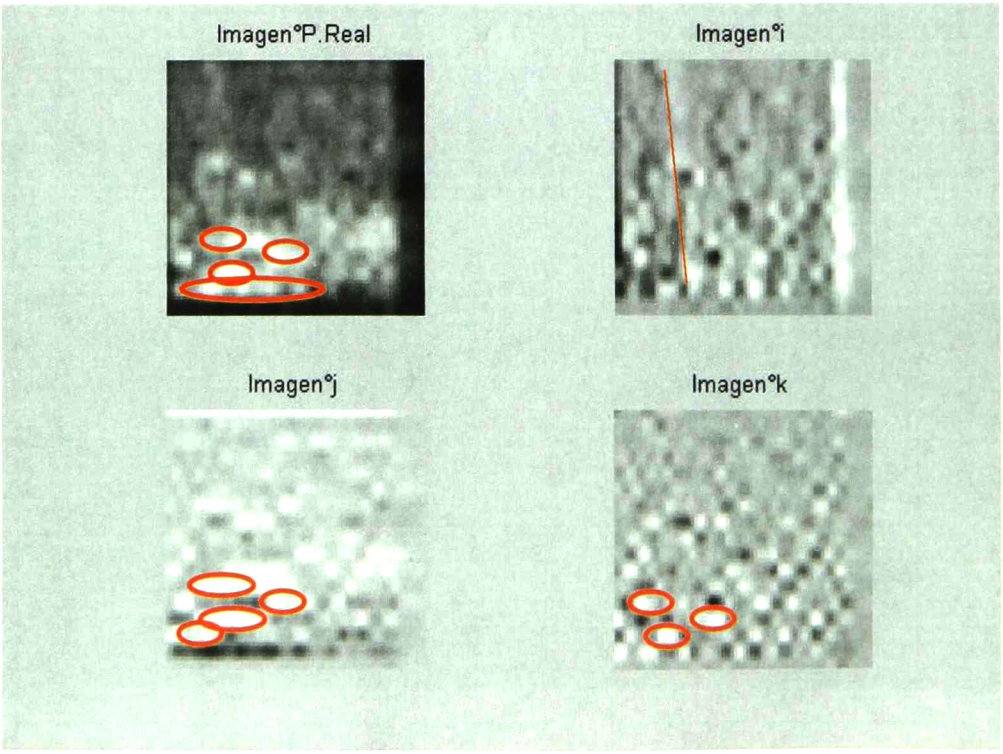


Figura A2-46. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

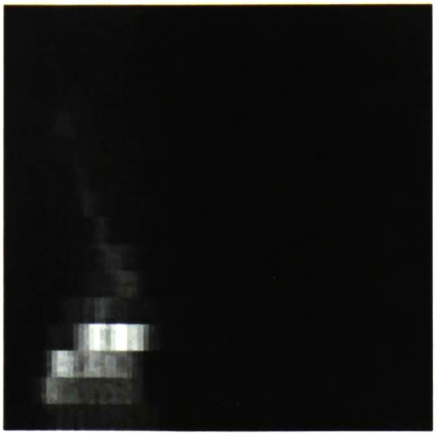


Figura A2-47. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Sin ruido. SNR= 0dB.

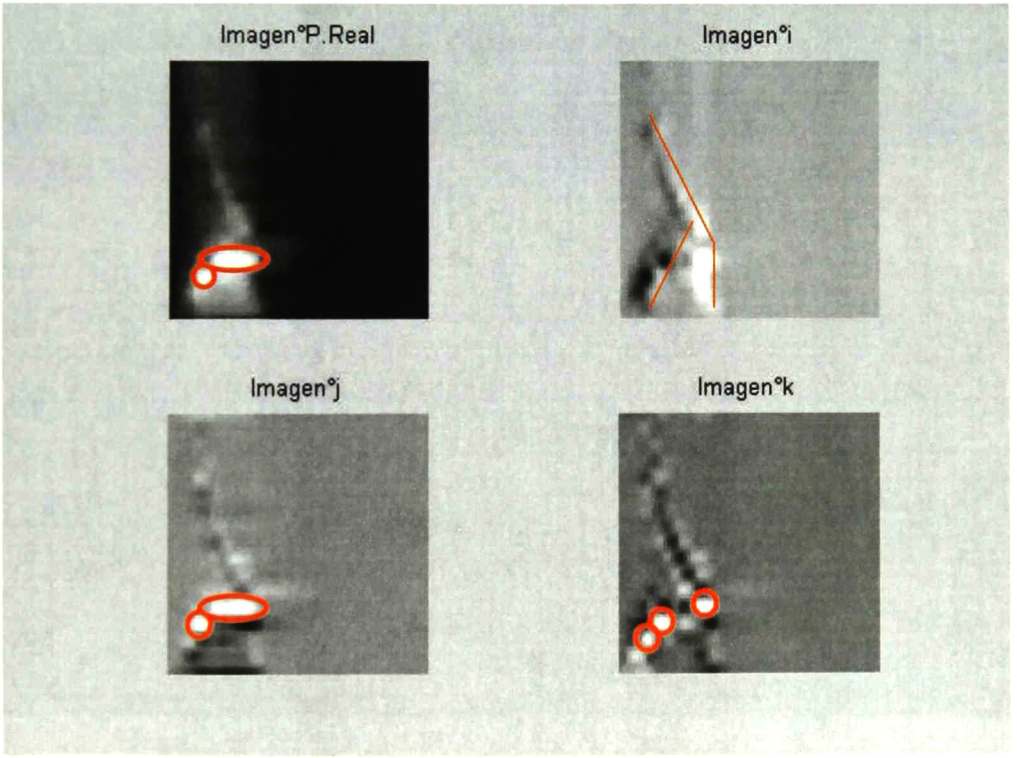


Figura A2-48. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

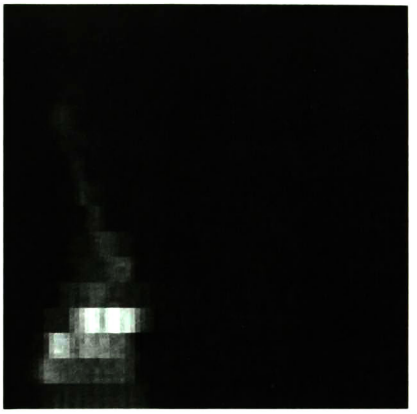


Figura A2-49. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= 40dB.

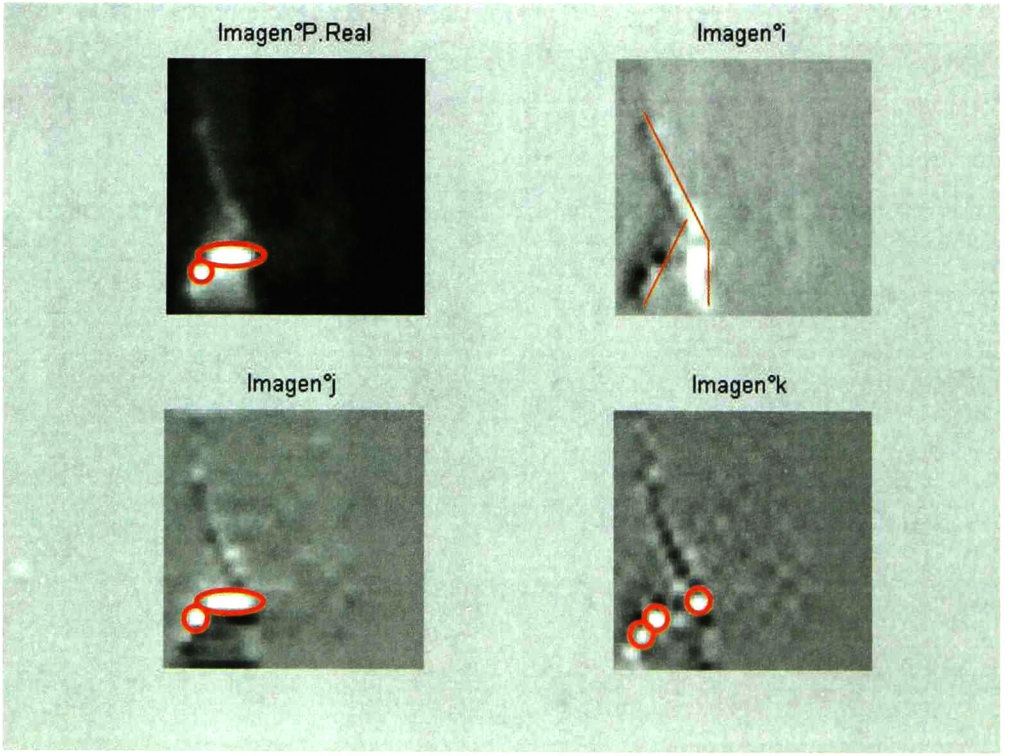


Figura A2-50. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

Figura A2-51. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= 20dB.

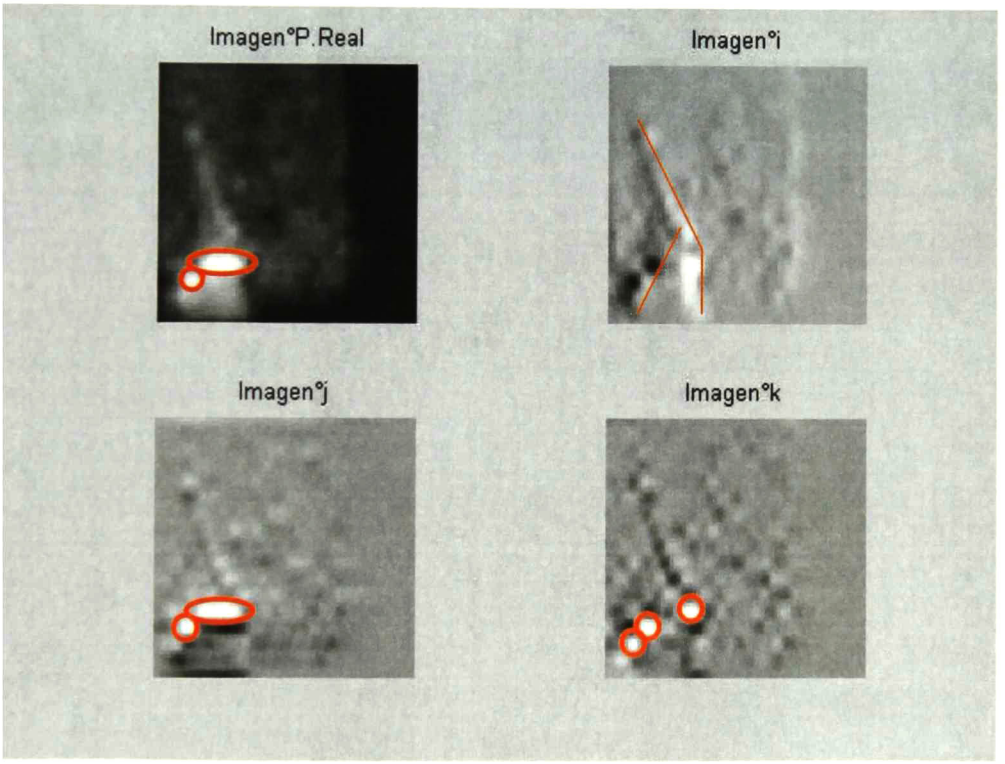


Figura A2-52. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

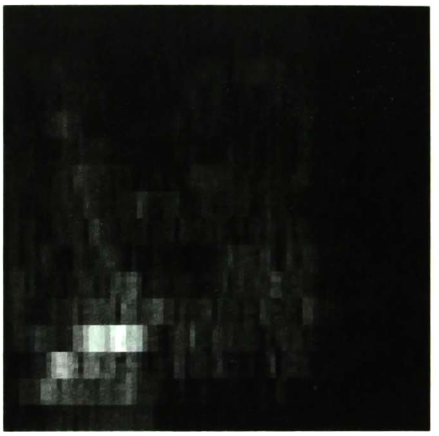


Figura A2-53. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= 10dB.

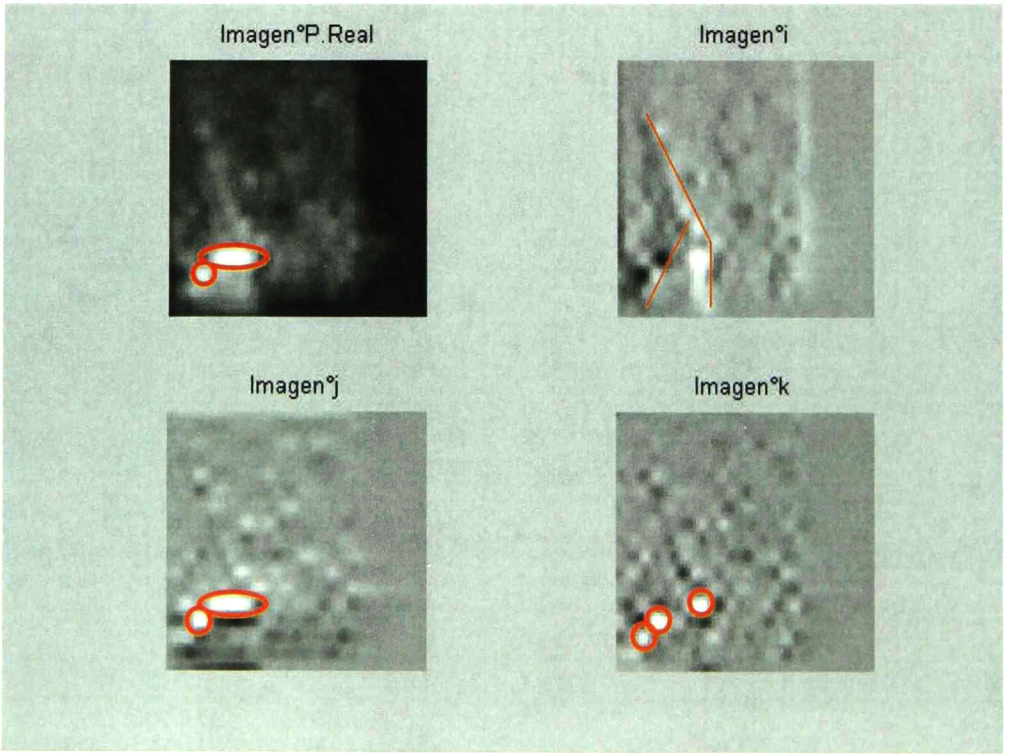


Figura A2-54. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

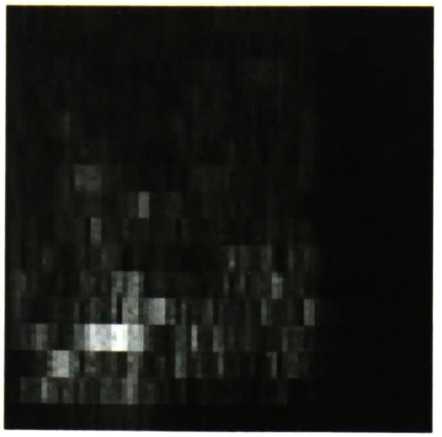


Figura A2-55. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= 0dB.

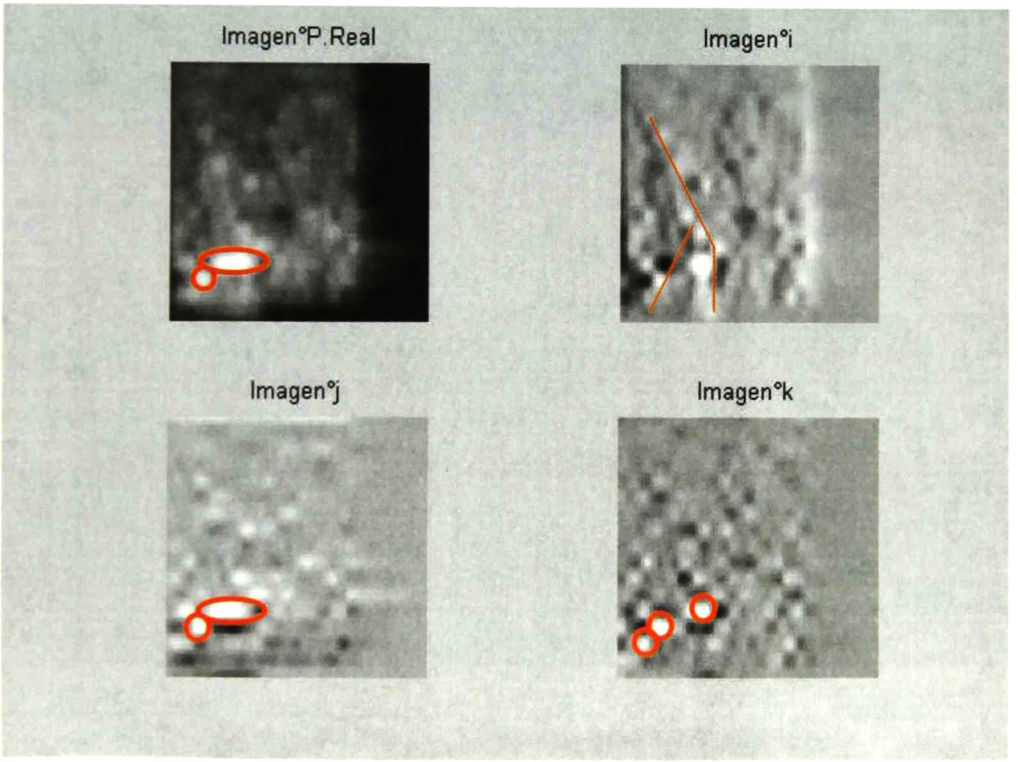


Figura A2-56. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

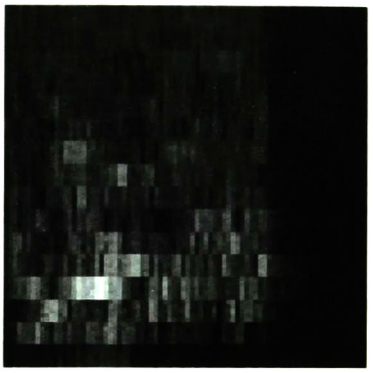


Figura A2-57. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= -5dB.

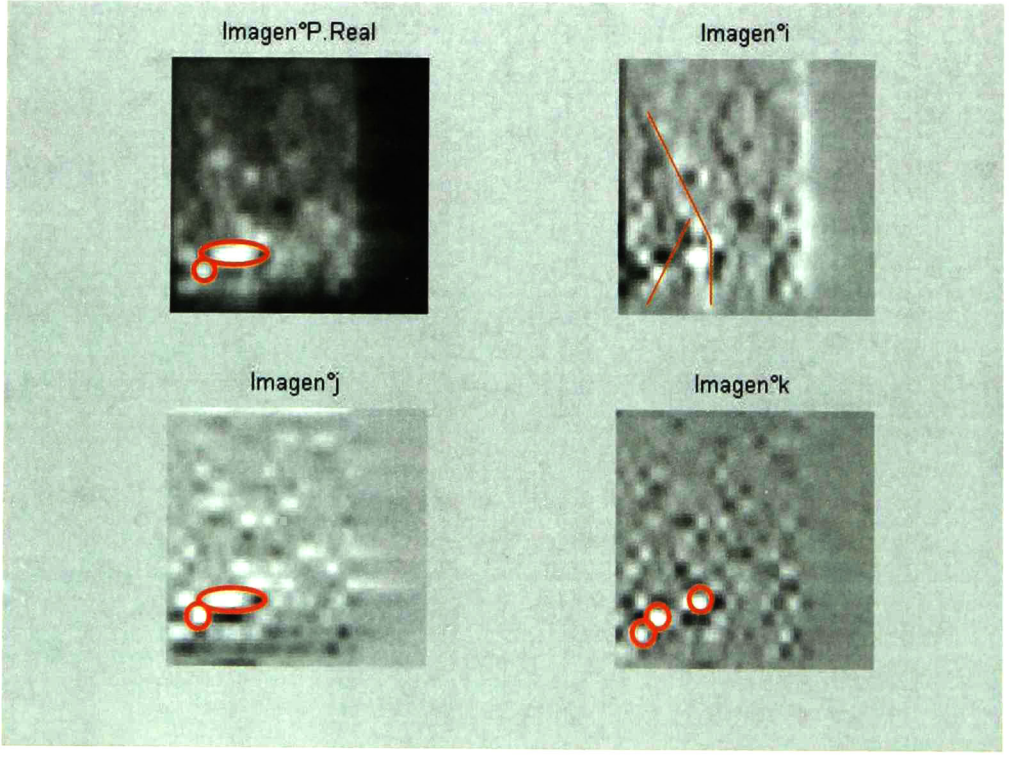


Figura A2-58. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

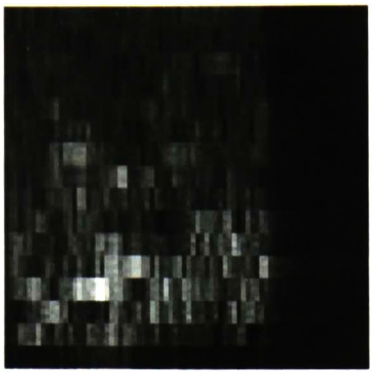


Figura A2-59. Imagen de la sonoridad de la palabra "zero" en francés, dicha por el locutor 12. Tipo de ruido: Conversación. SNR= -10dB.

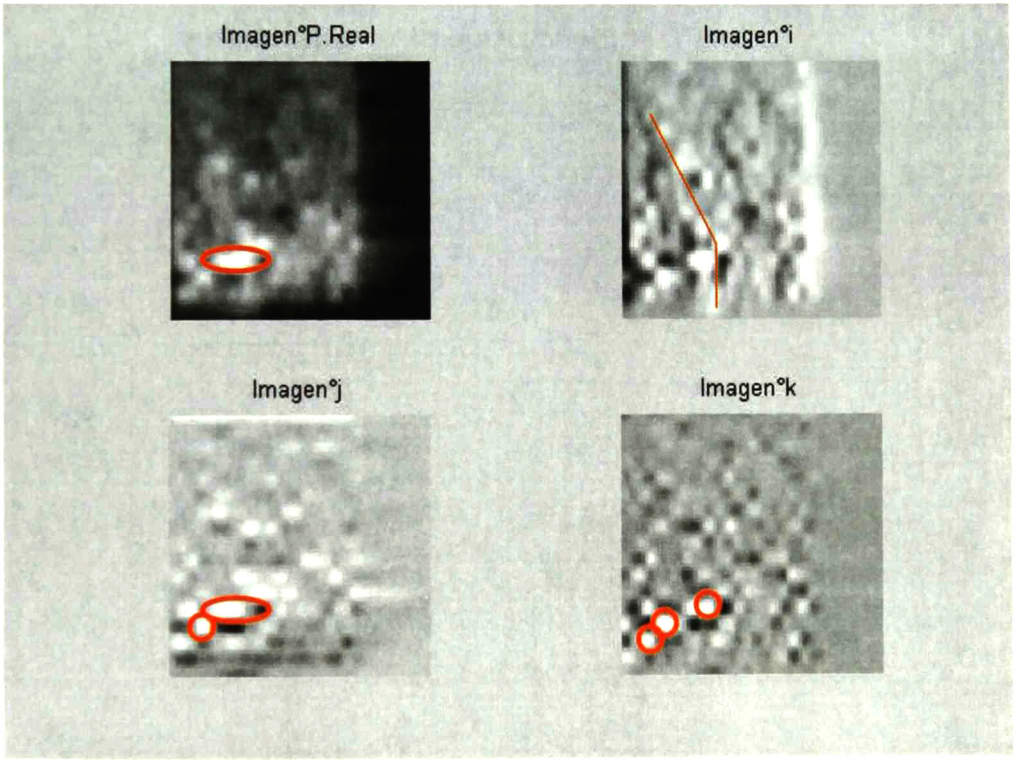


Figura A2-60. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

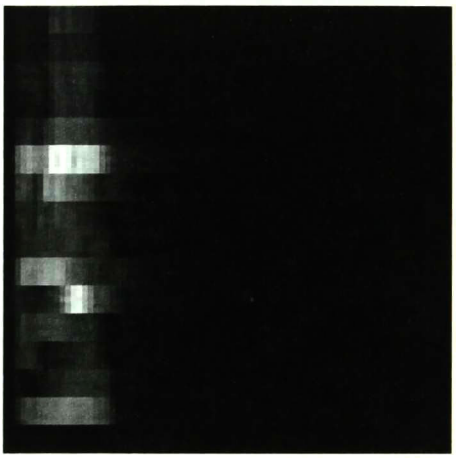


Figura A2-61. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Sin ruido.

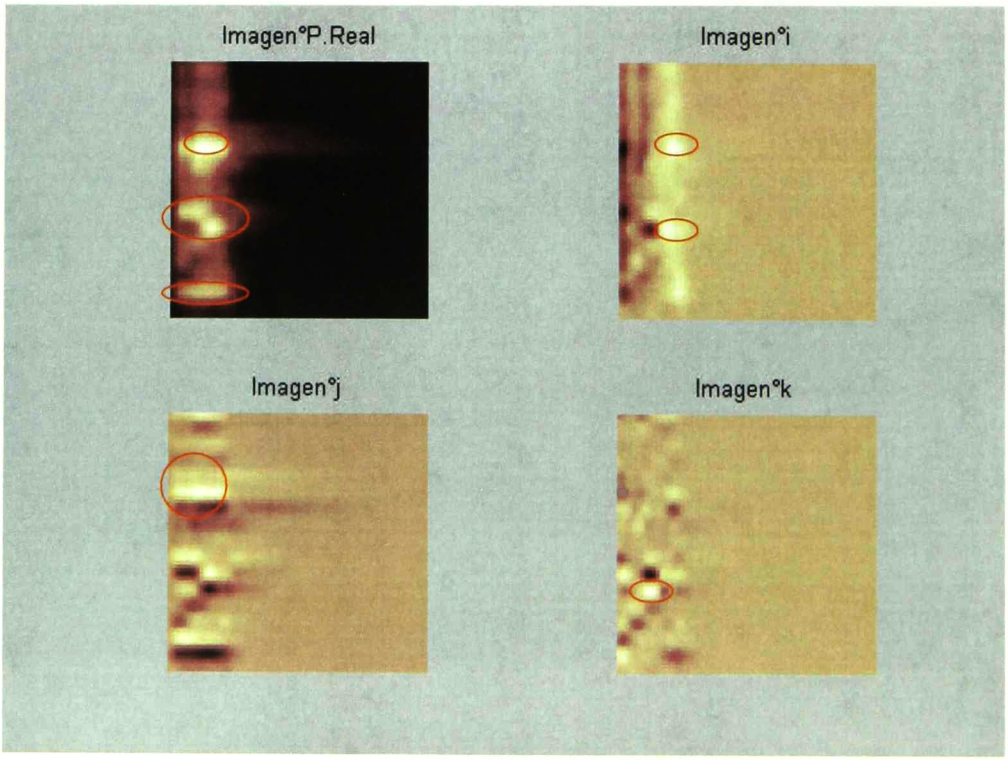


Figura A2-62. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

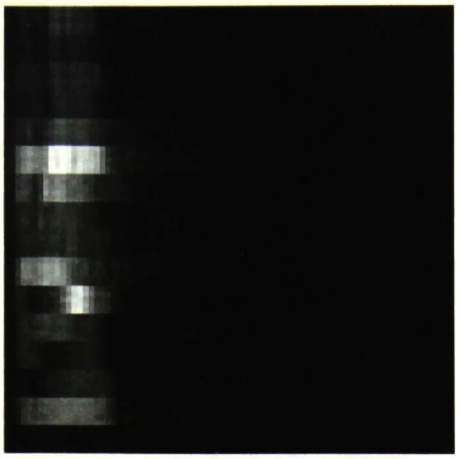


Figura A2-63. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 40dB.

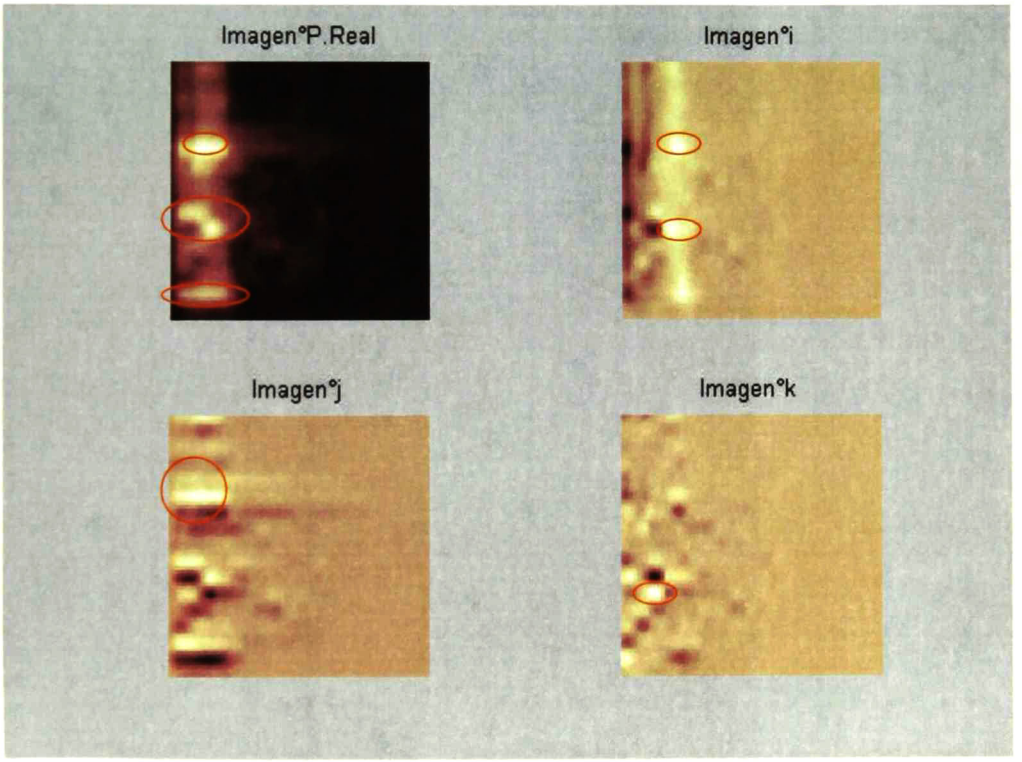


Figura A2-64. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

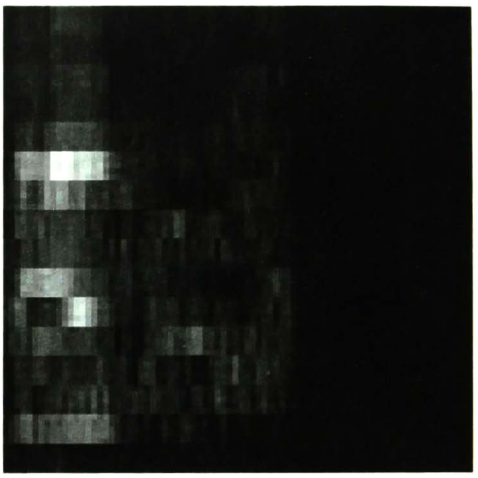


Figura A2-65. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 10dB.

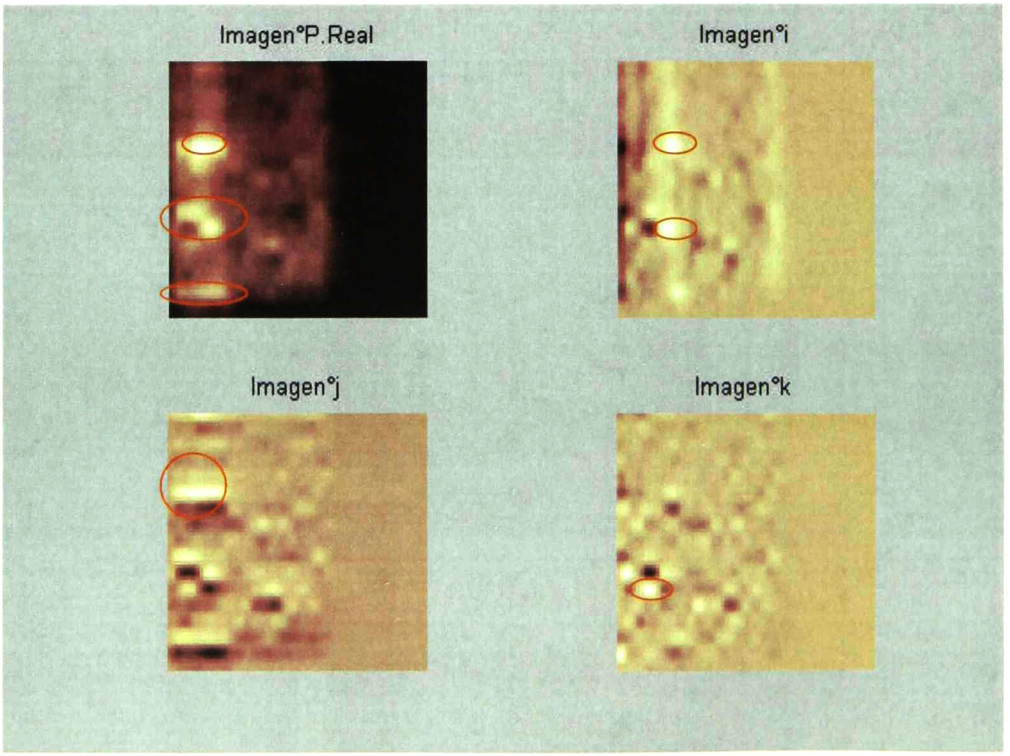


Figura A2-66. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

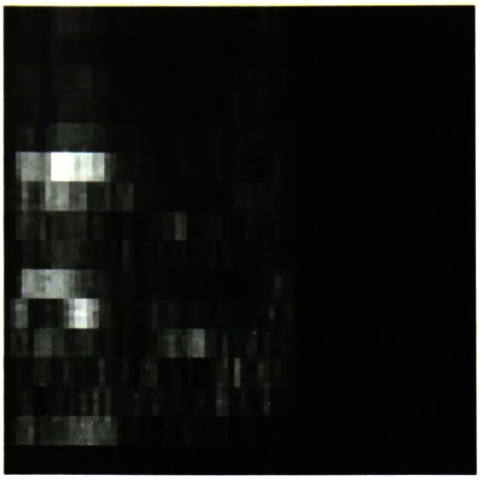


Figura A2-67. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 5dB.

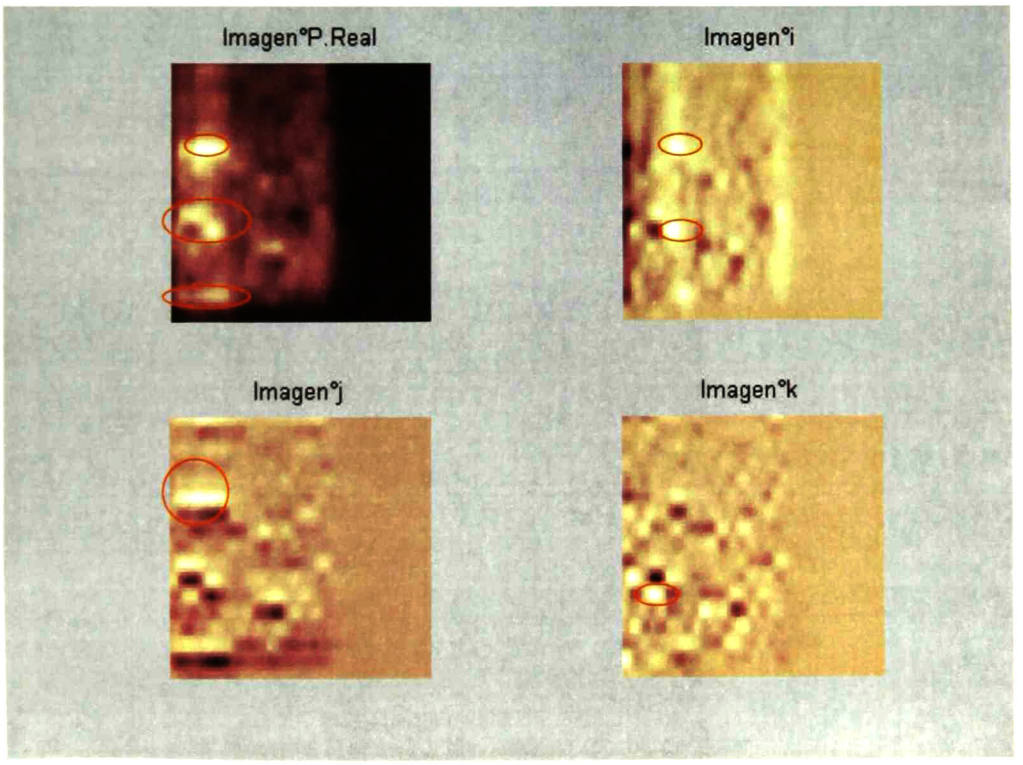


Figura A2-68. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

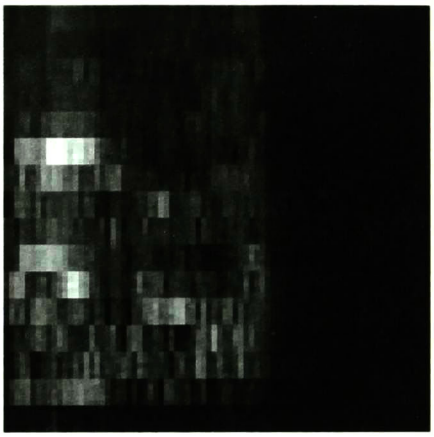


Figura A2-69. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= 0dB.

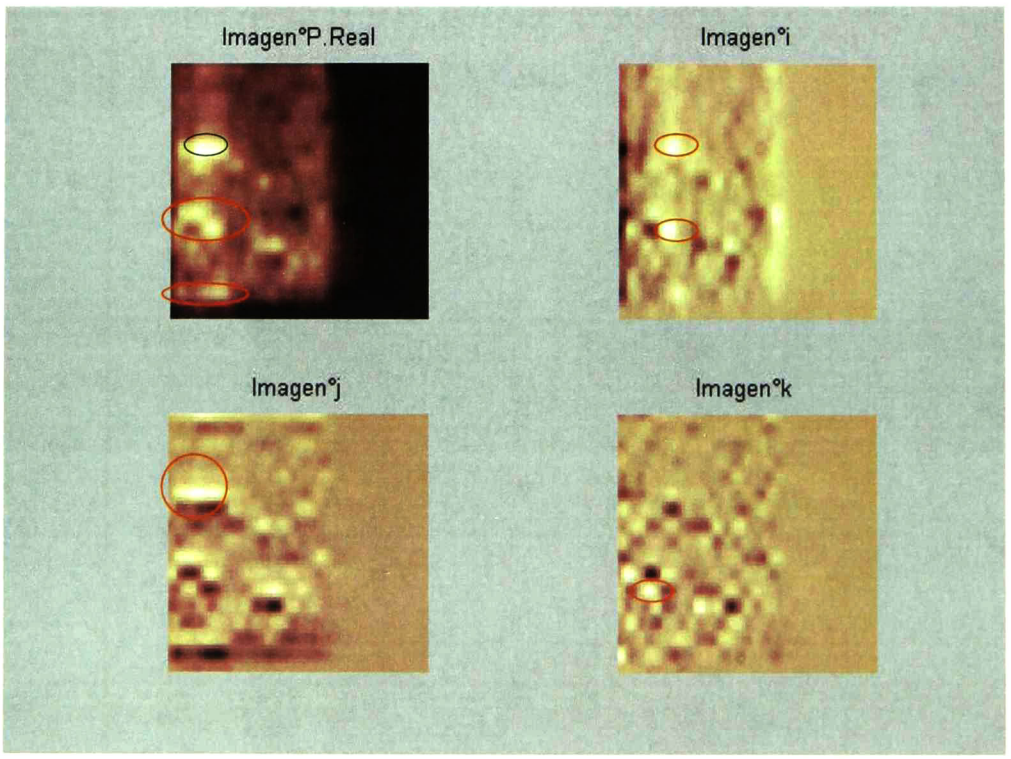


Figura A2-70. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

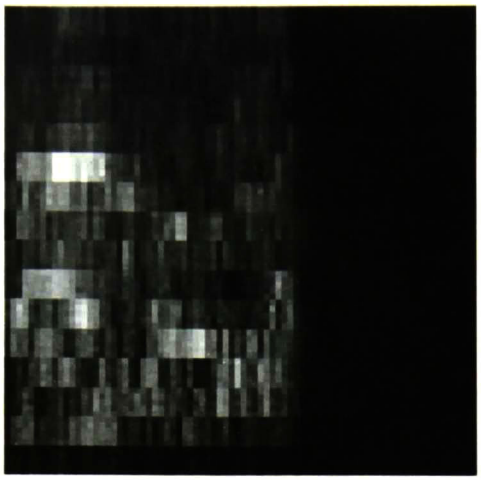


Figura A2-71. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -5dB.

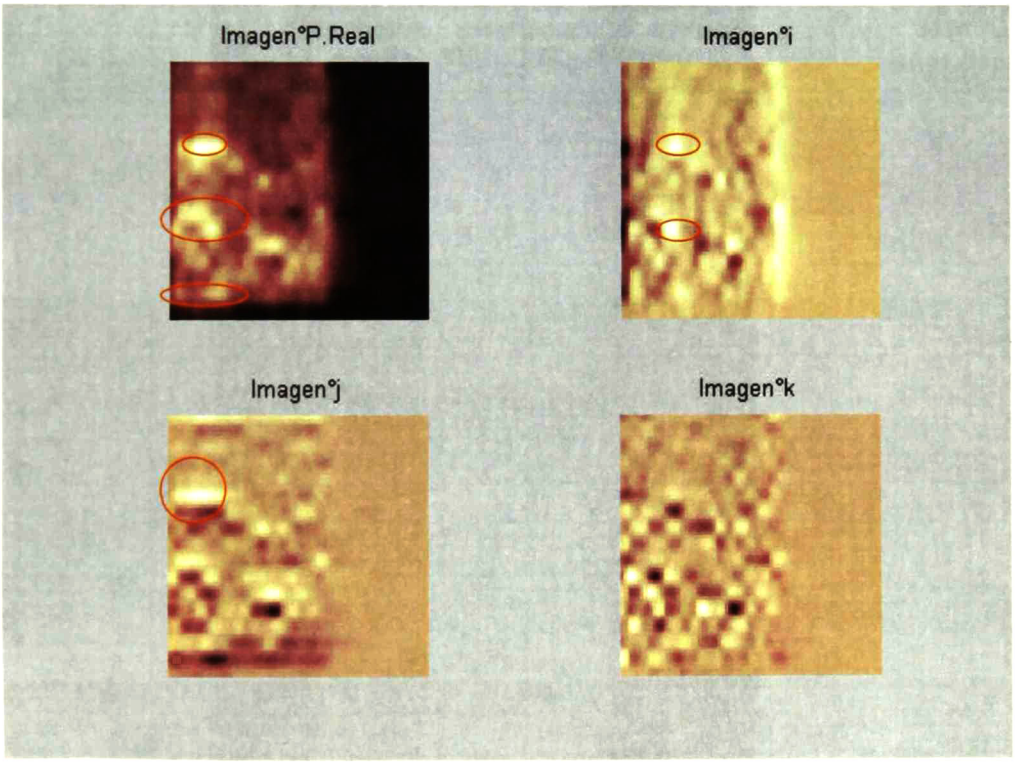


Figura A2-72. Componentes ***real***, ***i***, ***j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

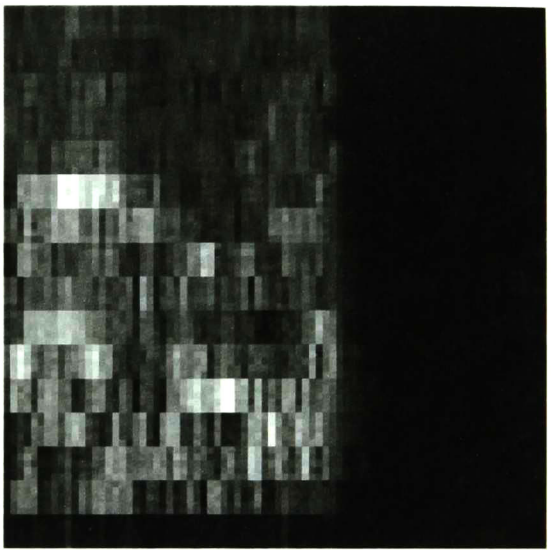


Figura A2-73. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -10dB.

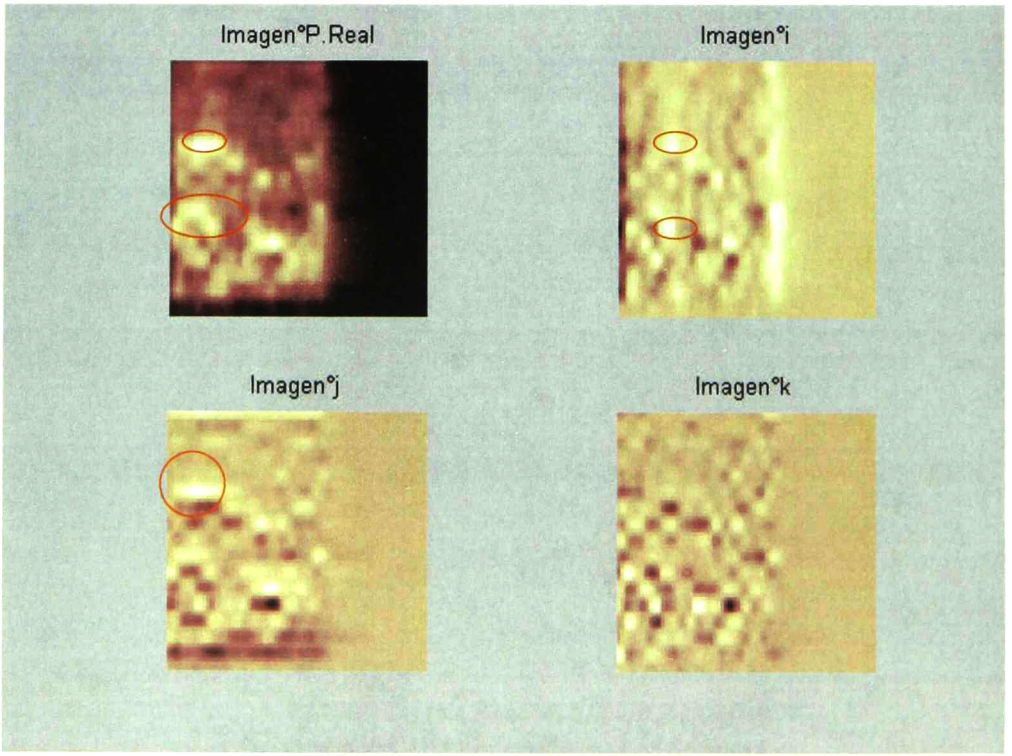


Figura A2-74. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

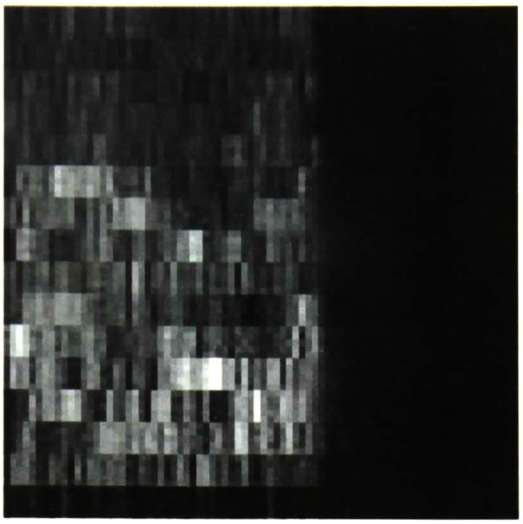


Figura A2-75. Imagen de la sonoridad de la palabra "un" en francés, dicha por el locutor 1. Tipo de ruido: Conversación. SNR= -20dB.

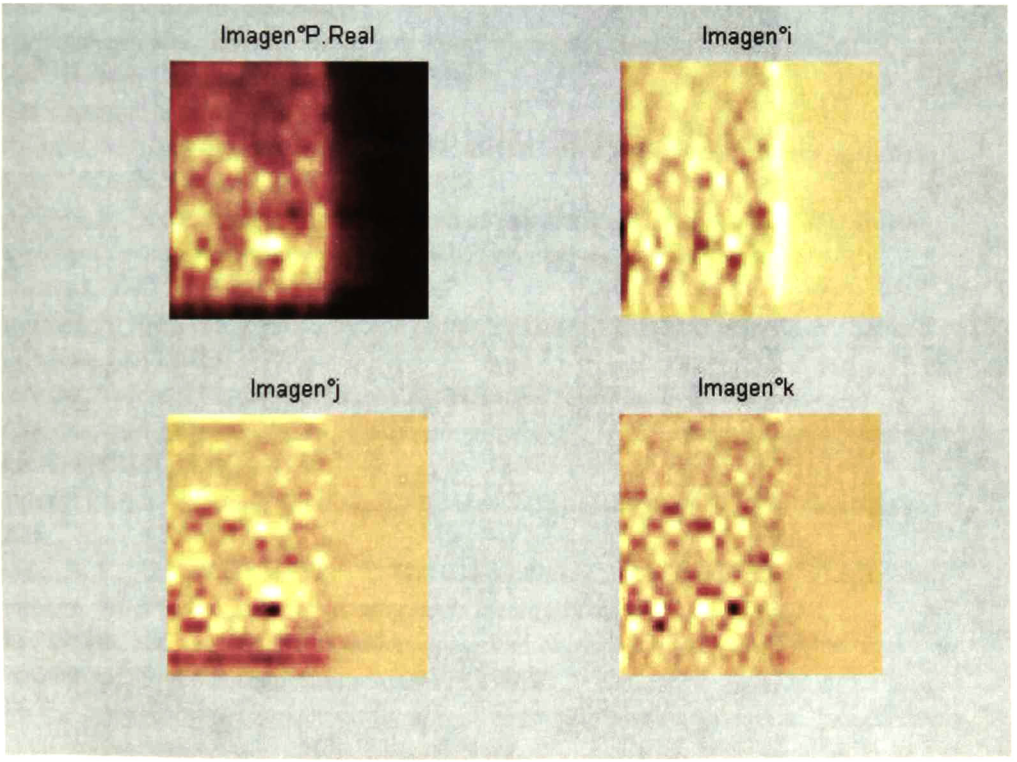


Figura A2-76. Componentes ***real, i, j*** y ***k*** de la imagen de la sonoridad de la figura anterior.

BLIOGRAFÍA

[1] Bayro-Corrochano, Eduardo. "Geometric Computing for Perception actino systems". Springer Verlag. May 2001, Boston.
[2] Beerends, John G. "Modeling Cognitive Effects that play a role in the perceptron of speech quality". Workshop "Speech quality Assessment", November 10-11, 1994 at RuhrUniversitat Bochum, Germany
[3] Berger, Jens. Merkel, Andrea. "An experimental System for objective speech quality measurements". Workshop "Speech quality Assessment", November 10-11, 1994 at RuhrUniversitat Bochum, Germany.
[4] Bradley, J.S. "Predictors of speech intelligibility in rooms". JASA vol. 80, no. 3. (1986)
[5] Bulow, Thomas. Felsberg, Michael and Sommer, Gerald. "Non-commutative Hypercomplex Fourier transforms of multidimensional signals".
[6] Bulow, Thomas. Sommer, Gerald. "Local hypercomplex signal representations and applications"
[7] Bulow, Thomas. Sommer, Gerald. "Quaternionic Gabor filters for local structure classification".
[8] Castleman, Kenneth R. "Digital Image Processing". Prentice Hall. 1996.
[9] Chan, C.P. Wong, Y.W. "Two-dimensional multi-resolution analysis of speech signals and its applications to speech recognition".
[10] Chen, Guangyi. Bui, Tien D. "Invariant Fourier-wavelet descriptor for pattern recognition". Pattern Recognition. Vol. 32 (1999).
[11] Digital Image Fundamentals
[12] Dimolitsas, S. "Characterization of low-rate digital voice coder performance with non-voice signals". Speech Communication 12 (1993).
[13] Dimolitsas, S. "Experimental quantification of voice transmission quality of mobile-satellite personal communications systems". IEEE Journal on selected areas in communications. Vol. 13. No.2. February 1995
[14] Dimolitsas, S. "Non-voice performance of the 16 kbit/s LD-CELP algorithm". Speech Communications 12 (1993).
[15] Dimolitsas, Spiros. "Transmission quality of North American cellular, personal communications, and public switched telephone networks". IEEE transactions on vehicular technology. Vol. 43. No.2. May 1994.
[16] Fulchiero, Ralph. Spanias, Andreas S. "Speech Enhancement susing the bispectrum". IEEE, 1993.
[17] Ganong, W.F. "Fisiología médica, El Manual Moderno", México, 1988, 11ª edición.
[18] Greenberg, Steven. On the origins of speech intelligibility in the real world
[19] Haika, Ulrich. Heute, Ulrich. "A new approach to objective quality-measures based on attribute-matching". Speech communications 11 (1992)
[20] Halka U. "Speech-Model processes for objective quality measurements of speech-coding systems"
[21] Halka U. & AL. "A new approach to objective quality-measures based on attribute-matching"

[22] Halka U. & AL. "A new objective quality measure for speech-coding systems based on the estimation of their nonlinear properties"
[23] Halka, Ulrich. "A new objective quality measure for speech-coding systems based on the estimation of their nonlinear properties". IEEE 1991.
[24] Halka, Ulrich. "Speech-model processes for objective quality measurements of speech-coding systems".
[25] Hamilton, William Rowan. "On a new species of imaginary quantities connected with a theory of quaternions". Edited by David R. Wilkins, 1999.
[26] Harris, Richard W. Brey, Robert H. "The effects of digital quantization error on speech intelligibility and perceived speech quality". Journal of Speech and Hearing Research. Vol. 34. February 1991.
[27] Hauenstein M. "Comparative study of psychoacoustics-based objective speech quality measures using Markov-SIRPS"
[28] Hauenstein, Markus. "Comparative study of psychoacoustics-Based objective Speech-quality measures Using Markov-SIRPS.Workshop "Speech quality Asessment", November 10-11, 1994 at RuhrUniversitat Bochum, Germany.
[29] Haykin, Simon. "Neural networks: A comprehensive foundation". 2nd. Edition. Prentice Hall. 1999.
[30] Hollier, M.P. "Characterization of communications systems using a speechlike test stimulus". J. Audio Eng. Soc., vol 41, No. 12, December 1993
[31] Hollier, M.P. % AL. "Algorithms for assessing the subjectivity of perceptually weighted audible errors"
[32] Hollier, M.P. & AL. "Characterization of communications systems using a speechlike test stimulus".
[33] Hollier, M.P. & AL. "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain"
[34] Hollier, M.P. & AL. "Objective perceptual analysis: Comparing the audible performance of data reduction schemes"
[35] Houtgast, T. & AL. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in audition"
[36] Howes, Davis. "On the relation between the intelligibility and frequency of occurrence of english words". The journal of the acoustical society of America. Volume 29, No. 2. February, 1957
[37] http//www3.labc.usb.ve/
[38] http://carini.physics.indiana.edu/
[39] http://cs'www.uchigado.edu/
[40] http://hyperphysics.phy'astr.gsu.edu/
[41] http://online.anu.edu.au/
[42] http://www.acoustics.hut.fi/
[43] http://www.cenamec.org.ve/
[44] http://www.cpl.umn.edu/
[45] http://www.cslu.cse.ogi.edu/
[46] http://www.dynastat.com/
[47] http://www.hearnet.com/
[48] http://www.home.att.net/

[49] http://www.ient.rwth'aachen.de/
[50] http://www.itcr.ac.cr/
[51] http://www.mcsquared.com/
[52] http://www.measure.demon.co.uk/
[53] http://www.meyersound.com/
[54] http://www.mlssa.com/
[55] http://www.opticom.de/
[56] http://www.pesq.org/
[57] http://www.psqm.com/
[58] http://www.science.wayne.edu/
[59] http://www.sfu.ca/
[60] http://www.utm.edu/
[61] http://wwwmmorph.com/
[62] Image Analysis and Computer Vision.
[63] Ingle, Vinay K. Proakis, John G. "Digital Signal Processing Using Matlab V.4.". PWS Publishing company. 1997
[64] Irii H. & AL. "Objective assessment of 16Kbits/s LD-CELP speech quality"
[65] Irii H. & AL. "Objective measurement method for estimating speech quality of low-bit-rate speech coding"
[66] Iru H. & A-L. "Promote-A system for estimating speech transmission quality in telephone network"
[67] Ishizuka, Toshio. Noguchi, Akira. Hanada, Eisuke. "Evaluation of speech quality for a packetized speech communication system using multipulse coding".Nec Res. & Develop.. Vol. 32. No.2. April 1991.
[68] Itoh K. & AL. "A new artificial speech signal for objective quality evaluation of speech coding systems"
[69] Joy, Kenneth I. "On-Line computer graphics notes"
[70] Keidser, Gitte. "Computerized measurement of speech intelligibility" Scan Audiol, 1991.
[71] Kirchhoff, Katrin. "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments".
[72] Kirschning, Ingrid. Aoe, Jun-Ichi. "Redes neuronales recurrentes de correlación en cascada para reconocimiento de voz" Congreso iberoamericano de inteligencia artificial, IBERAMIA'96.
[73] Kitawaki, N. "Research of objective speech quality assessment".
[74] Kubichek, R.F. "Standards and technology issues in objective voice quality assessment"
[75] Liu, Derong. "A new synthesis approach for feedback neural networks based on the perceptron training algorithm". IEEE transactions on neural networks, vol. 8. No. 6. November 1997.
[76] Mannell, Robert H. "Natural and Synthetic speech Intelligibility and Quality Testing".
[77] Mc Daniel, D. Michael. Cox, Robyn M. "Evaluation of the speech intelligibility rating (SIR) test for hearing aid comparisons". Journal of Speech and Hearing Research, Volume 35. Jun 1992.
[78] Metz D.E. & AL. The use of artificial neural networks to estimate speech intelligibility from acoustic variables: A preliminary analysis".

[79] Metz, Dale E. "The use of artificial neural networks to estimate speech intelligibility from acoustic variables: A preliminary analysis". J. Commun. Disord. Vol 25. (1992).
[80] Mush, Hannes. "Fletcher and Galt{s method for calculating the articulation index.
[81] Nagabuchi, Hiromi. Kitawaki, Nobuhiko. "Artificial conversational Speech signal for evaluating speech device performance".
[82] Nielsen-Schmidt, A. "Comments on the use of physical measures to assess speech intelligibility"
[83] Nomura, H. & AL. "Speech intelligibility and modulation transfer function in no-exponential decay fields"
[84] Peterson, Patrick. Jeanrenaud, Philippe. "Improving Intelligibility of a 300 b/s segment vocoder".
[85] Pollack, Irwin and Pickett, J. M. "Stereophonic listening and speech intelligibility against voice babble". The Journal of The Acoustical Society of America. Volume 30, No. 2. February 1958.
[86] Proakis, John G. "Digital Communications". 3rd. Edition. McGraw'Hill International editions.1995
[87] Proakis, John G. Manolakis, Dimitris G. "Digital Signal Processing: Principles, algorithms, and applications". 3rd. Edition. Prentice Hall. 1996
[88] Projet A23: Analyse D'images auditives par morphologie mathematique. Project report.
[89] PW Ellis, Daniel. "A computer implementation of Psychoacoustic Grouping Rules" 12th International Conference on Pattern Recognition, Jerusalem, October 1994
[90] Robinson, David J.M. "The Human Auditory System".
[91] Robinson, David J.M. & Hawksford, Malcolm J. "Time-domain auditory model for the assessment of high-quality coded audio"
[92] Rossing, D. Thomas. "The science of sound". 2nd. Edition. Addison-Wesley Publishing company. 1989.
[93] Silipo, Rosaria. "Spectro-Temporal Constraints on Speech Intelligibility".
[94] Steeneken, H.J. "The evaluation of speech transmission quality"
[95] Steward, Oswald. "Functional Neuroscience". Springer Verlag. 2000.
[96] Stuart, J.R.: "Predicting the audibility, detectability and loudness of errors in Audio Systems", Audio Engineering Society Preprint, presentado en la 91a convención de la AES, Nueva York, 1991.
[97] Tardelli, John D. "A systematic investigation of the Mean Opinion Score (MOS) and the Diagnostic Acceptability Measure (DAM) for use in the selection of digital speech compression algorithms". Technical report. ARCON Corporation.
[98] Veste, Stéphane. Tuffeli, Denis. Naranjo, Michel. "Mesure objective de l'intelligibilite de la parole a travers les systemes de transmission (Rapport R8).
[99] Voran, S. "Objective estimation of perceived speech quality using measuring normalizing blocks". NTIA report 98-347.April 1998.
[100] Voran, S. "Perception-based objective estimators of speech quality"
[101] Watta, Paul B. Wang, Kaining. "Recurrent Neural Nets. As Dynamical Boolean Systems with application tro associative memory". IEEE transactions on neural networks. Vol. 8. No. 6, November 1997.
[102] Yamazaki T. "A telecommunication speech-quality assessment method using the likehood for degraded speech pattern as an input in a neural network".

[103] Yamazaki, Tetsuro. "A telecommunication speech-quality assessment method using the likelihood for degraded speech pattern as an input in a neural network". Workshop "Speech quality Assessment", November 10-11, 1994 at RuhrUniversitat Bochum, Germany.
[104] Yang, Wonho. Benbouchta, Majid. "Performance of the modified Bark spectral distortion as an objective speech quality measure"
[105] Yang, Wonho. Dixon, Myron. "A modified Bark spectral distortion measure which uses noise masking threshold".
[106] Yang, Wonho. Yantorno, Robert. "Comparison of two objective speech quality measures: MBSD and ITU-T recommendation P.861".
[107] Yang, Wonho. Yantorno, Robert. "Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS"
[108] Zwicker, Eberhard. Feldtkeller, Richard. "Psychoacoustique: L'oreille, récepteur de'information". Msson. 1981.
[109] Zwicker, Eberhard. "Psychoacoustics: Facts and models".2nd. Edition. Springer-Verlag. 1999.

Centro de Investigación y de Estudios Avanzados del IPN

Unidad Guadalajara

El Jurado designado por la Unidad Guadalajara del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, aprobó la tesis: DESARROLLO DE UN CALCULADOR DE SONORIDAD del(a) C. Noel TRUJILLO MORALES el día 16 de Noviembre de 2001

Dr. Valeri Korjik
Investigador Cinvestav 2A
CINVESTAV GDL
--

Dr. Eduardo Jose Bayro Corrochano
Investigador Cinvestav --
CINVESTAV GDL
--

Dr. Michel Naranjo
Profesor
UNIVERSIDAD BLAISE PASCAL
Francia