ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES Y DE TELECOMUNICACIÓN

Titulación :

INGENIERO TÉCNICO DE TELECOMUNICACIÓN, ESPECIALIDAD EN SONIDO E IMAGEN

Título del proyecto:

ESTUDIO SOBRE EL VOCODER HOMOMÖRFICO

Alumno: Ion Aritz Dallo

Tutor: Miroslav Zivanovic Jeremic

Pamplona, 19 de Junio del 2012

INDICE

INTRODUCCION

RESUMEN/PROBLEMÁTICA

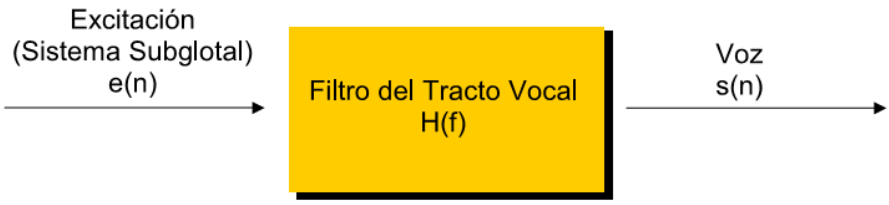
La voz humana consiste en sonidos generados por la apertura y cierre de la glotis (cuerdas vocales), lo que produce una onda periódica con muchos sonidos armónicos. Este sonido básico es entonces filtrado por la nariz y la garganta (un complicado sistema resonante conocido como el tracto vocal) de forma controlada, creando la amplia variedad de sonidos de habla. Hay otro conjunto de sonidos, conocidos como sordos, que no son generados por la vibración de las cuerdas vocales, pues estas permanecen abiertas cuando el aire pasa a través de ellas y sólo influye el tracto vocal.

Existen varias representaciones para la señal de voz. Se puede representar la forma de onda basándose en procesos de muestreo y cuantificación. Esto pretende mantener la forma de onda original. Otra forma es la representación paramétrica. En ella se considera la señal de voz como la salida de un sistema de producción de voz que puede ser representado como un conjunto de parámetros. También existen vocoders que son mezcla de los anteriores, se usan técnicas paramétricas y de forma de onda mezcladas, dando lugar a lo que se conoce como vocoder híbrido. El vocoder que se diseñará en este proyecto es de este tipo.

Un Vocoder (nombre derivado de voice-coder, «codificador de voz») es un analizador y sintetizador de voz. Fue desarrollado en la década de 1930 como un codificador de voz para telecomunicaciones. Su primer uso fue la seguridad en radiocomunicaciones, donde la voz tiene que ser digitalizada, cifrada y transmitida por un canal de ancho de banda estrecho.

El vocoder examina el habla encontrando los parámetros de interés, y midiendo cómo cambian las características espectrales con el tiempo grabando el habla. Esto da como resultado una serie de números representando esas frecuencias modificadas en un tiempo particular a medida que el usuario habla. Al hacer esto, el vocoder reduce en gran medida la cantidad de información necesaria para almacenar el habla. Para recrear el habla, el vocoder simplemente revierte el proceso. El resultado es habla inteligible, aunque algo mecánica en los vocoders de tipo paramétrico. Con los vocoders híbridos se consiguen mejores resultados a costa de sacrificar Ancho de Banda (BW).

El vocoder híbrido diseñado es un vocoder homomórfico, basado en un modelo de representación de la voz convolucional, en el que la señal de voz se considera como la salida de un filtro lineal al que se le aplica una señal de excitación. La señal de voz s(n) es la convolución entre la excitación e(n) y el filtro lineal H(f) que modela el tracto vocal.



En tiempo tenemos que:

$$s(n) = e(n) * h(n)$$

Y en espectro:

$$S(\Omega) = E(\Omega)\, H(\Omega)$$

Separar las dos señales no es un problema trivial, pues no están superpuestas aditivamente. Si las señales ocuparan intervalos temporales distintos o bandas frecuenciales separadas sería posible separarlas fácilmente en cualquiera de los dos dominios.

El problema que se pretende resolver es una deconvolución. El procesado homomórfico sirve de ayuda transformando la convolución en una suma. La deconvolución se define como la operación inversa de la convolución. Por lo tanto si se supone que una señal se ha formado por la convolución de dos señales, es posible separar dichas señales aplicando la operación inversa o deconvolución.

$$s(n) = e(n) * h(n) \rightarrow \hat{s}(n) = \hat{e}(n) + \hat{h}(n)$$

El procesado homomórfico o análisis cepstral, tiene como objetivo transformar la señal de voz, de manera que las dos componentes se superpongan aditivamente, y se puedan separar linealmente, en el dominio cepstral.

La transformación se hace mediante un operador D*:

$$\hat{s}[n] = D^*(\hat{s}[n]) = D^*(\hat{s}_1[n] * \hat{s}_2[n]) = D^*(\hat{s}_1[n]) + D^*(\hat{s}_2[n]) = \hat{s}_1(n) + \hat{s}_2(n)$$

Se logra pasar así de una convolución de difícil resolución a una suma de logaritmos, en definitiva a una suma de dos señales de muy distintas características en frecuencia, ya que la señal de excitación tiene variaciones temporales muy superiores a la respuesta de filtro. Si se realizara una nueva transformada de Fourier al resultado actual, se separarían claramente las señales en regiones apartadas dado que la respuesta del filtro ocupará las bajas frecuencias y la excitación las regiones de más alta frecuencia.

Una vez separados se puede deshacer la transformación mediante el operador inverso D^{*-1}, y si se quiere recuperar la señal original sólo hay que realizar la convolución entre la señal de excitación y el tracto vocal

$$D^{*-1}(\hat{s}_1[n]) = s_1[n]$$

$$s_1[n] * s_2[n] = s[n]$$

APLICACIONES

Las aplicaciones que se le pueden dar a los vocoders son variadas; codificación (almacenamiento y transmisión), síntesis (generación de voz humana), reconocimiento (sistemas de seguridad), artísticas (el vocoder como efecto)…

El objetivo de este proyecto es el análisis de dos tipos de vocoder para su uso en codificación-decodificación con el objeto de buscar una representación de la señal de voz que permita comprimirla al máximo manteniendo una calidad aceptable.

Se busca el mínimo número de bits para una calidad máxima, de esta forma se ahorra ancho de banda en transmisión y espacio de almacenamiento.

Hoy en día la reducción de la tasa de bits con la máxima calidad se hace necesaria por el continuo aumento del tráfico de datos en telefonía móvil, pues la voz ahora tiene que compartir canal con todo el tráfico multimedia.

- Reducción de la velocidad binaria en los teléfonos móviles y llamadas usando las redes 3G/4G sin pérdida de calidad.
- Seguridad: encriptación de la información.

El codificador LPC debido a la baja calidad pero poca demanda de ancho de banda resulta útil en situaciones donde la alta calidad no es imprescindible, pero si la capacidad de encriptación y el poder usar el sistema en lugares con redes de telecomunicación precarias como podría ser en campañas militares en países remotos. No obstante la calidad conseguida es muy baja e inaceptable en sistemas de uso civil o doméstico. La calidad conseguida con el vocoder homomórfico, es mucho mayor como podrá ver más adelante.

ESTADO DEL ARTE

Hace unos cincuenta años que empezó la investigación en el campo de la codificación de la voz. El pionero fue Homer Dudley, que trabajaba en los laboratorios de la Bell Telephone. La motivación para realizar esta investigación surgió por la necesidad de transmitir voz por los cables de telegrafía de pequeño ancho de banda. La idea del vocoder de Dudley era analizar la voz para extraer una serie de características y que el emisor enviase esas características, cuando éstas le llegasen al receptor reconstruiría la voz original. Este codificador recibió gran atención durante la Segunda Guerra Mundial, debido a su potencial en cuanto a eficiencia y posibilidad de encriptación se refiere.

Las primeras implementaciones del vocoder eran analógicas, sin embargo, con el nacimiento de los sistemas digitales y de las posibilidades que éstos ofrecen, pronto se pasó a las implementaciones digitales. Durante la década de los 40 hubo una gran actividad en la Codificación por Modulación de Impulsos (PCM). Este tipo de codificación no sigue la filosofía del vocoder de Dudley (y de los vocoders en general), sino que simplemente muestrea la voz. A partir del PCM se desarrollaron el DPCM y el

ADPCM, que fueron propuestos como estándar por la CCITT (International Consultative Committee for Telephone and Telegraph).

Gracias a la flexibilidad de los sistemas digitales, se pudo experimentar con formas más sofisticadas de representación de la voz. Fant, a finales de los 50, trabajó en el modelo de producción de voz lineal.

El surgimiento de la tecnología VLSI, tecnología de muy baja escala de integración, durante los 60 y 70 permitió nuevas soluciones al problema de la codificación de la voz. Así, por ejemplo, Flanagan y Golden propusieron una solución basada en la Transformada de Fourier.

Durante los 80 y 90, la investigación ha ido encaminada a conseguir codificadores que utilicen un ancho de banda cada vez menor mientras que la calidad de la voz sea cada vez mejor. Con esto se permite utilizar con más eficiencia y eficacia los canales de transmisión, se facilita la encriptación y se aprovechan mejor los sistemas de almacenamiento.

Una de las principales aplicaciones de la codificación de voz es la telefonía móvil. En telefonía móvil, en Estados Unidos se utiliza un estandar de 8 Kbps (VSELP) y otro similar, a 6.7 Kbps, en Japón. En Europa, dentro del sistema GSM, se usa un codificador a 13 Kbps.

OBJETIVOS

El objeto de este proyecto es programar en Matlab un vocoder homomórfico o de análisis cepstral, analizar su comportamiento como codificador-decodificador, y realizar una comparativa de calidad de las señales decodificadas con los resultados obtenidos con un vocoder LPC.

Los 2 vocoders comparados aquí se diferencian en el proceso en el que se obtienen los parámetros y en el número de estos, lo que marca una diferencia clara en el resultado obtenido al decodificar la señal de voz.

Se realizará una comparativa en la que se mida la calidad de ambos vocoders y la cantidad de datos (bits por segundo) que maneja cada uno.

Por último se usará la herramienta creada para el análisis de señal de voz con el fin de determinar si una trama de voz es sonora o sorda y determinar la frecuencia fundamental de las tramas de voz sonoras.

REVISION TEÓRICA

LA VOZ HUMANA

PRODUCCIÓN DE LA VOZ HUMANA

La producción de voz tiene lugar en el aparato fonador. La señal de voz es una onda de presión acústica que se produce por la vibración de las cuerdas vocales. El aparato fonador está formado por cavidades y elementos articuladores.

El aire exhalado de los pulmones es modulado y formado por la vibración en las cuerdas vocales y el tracto vocal. Ese sonido producido por la vibración de las cuerdas vocales es llevado al exterior por el propio aire espirado que ha causado la vibración. Dependiendo del tipo de sonido, las cuerdas vocales pueden vibrar o no y así tendremos una señal cuasiperiódica o aleatoria.

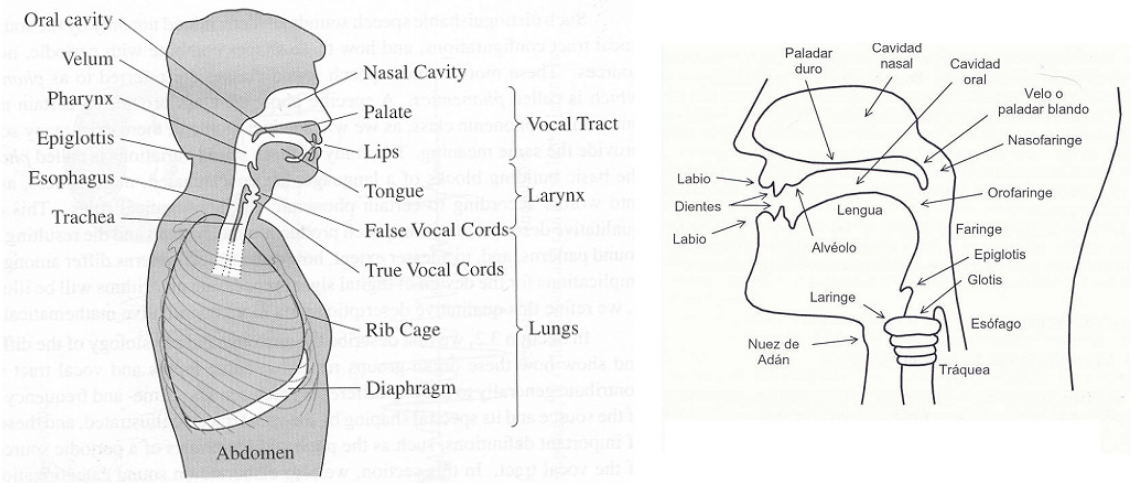


Figura 1: Imagen del aparato fonador y elementos articuladores

Controlando a voluntad los elementos articuladores se pueden modificar los sonidos en un amplio rango mediante 2 mecanismos:

- Filtrado: Modifica el espectro del sonido. Lo llevan a cabo la faringe, cavidad nasal y cavidad oral (tracto vocal). Constituyen resonadores acústicos que enfatizan

determinadas bandas frecuenciales. Estas bandas se denominan formantes o picos de resonancia.

- Articulación: Se trata de la modificación de las formas, posiciones y tensiones de los elementos del aparato de fonación. Esto supone interponer un obstáculo para la circulación del flujo de aire. Se enfatizan diferentes armónicos, pudiendo diferenciarse los distintos fonemas. Los elementos articuladores son los labios, dientes, alvéolos, paladar, lengua y glotis.

PARTICULARIDADES DE LA VOZ HUMANA

La voz humana presenta unas particularidades que nos ayudan a la hora de su análisis.

Las señales de voz se caracterizan por ser no estacionarias y con variaciones lentas en el dominio del tiempo y se procesan normalmente en segmentos de tiempo cortos, entre 20 y 40 ms, siendo típicamente 30 ms. Es por ello que el análisis de la señal de voz se realiza por tramas y no toda la señal a la vez. Si se consideran tramos cortos de la señal de voz, sus propiedades permanecerán semipermanentes. Se puede tomar entonces cada tramo como si hubiese sido generado por la excitación de un sistema lineal invariante temporal. Esta señal de excitación es un tren de pulsos cuasi-periódicos o ruido aleatorio. En la Figura 2 podemos ver una ventana de duración 60 ms de una señal de voz sonora. En ella se puede apreciar la cuasi-periodicidad.

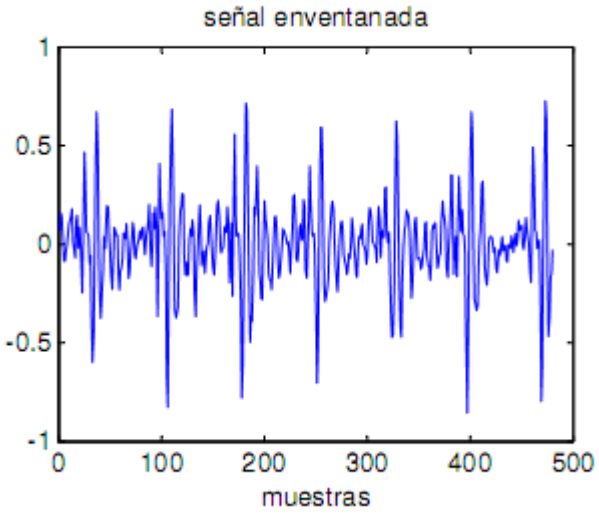


Figura 2: Ventana de 60 ms de una señal de voz

En estos segmentos la señal puede aproximarse a una señal cuasiestacionaria y cada segmento o trama puede ser clasificada como sonora o sorda.

Los sonidos sonoros tienen una naturaleza cuasiperiódica en el dominio del tiempo y una estructura armónica fina en el dominio de la frecuencia, provocada por la vibración de las cuerdas vocales. Además, su espectro decae hacia altas frecuencias. Su energía es alta debido a que el aire encuentra poca obstrucción al pasar por el tracto vocal. Estas características pueden observarse en la figura 3.

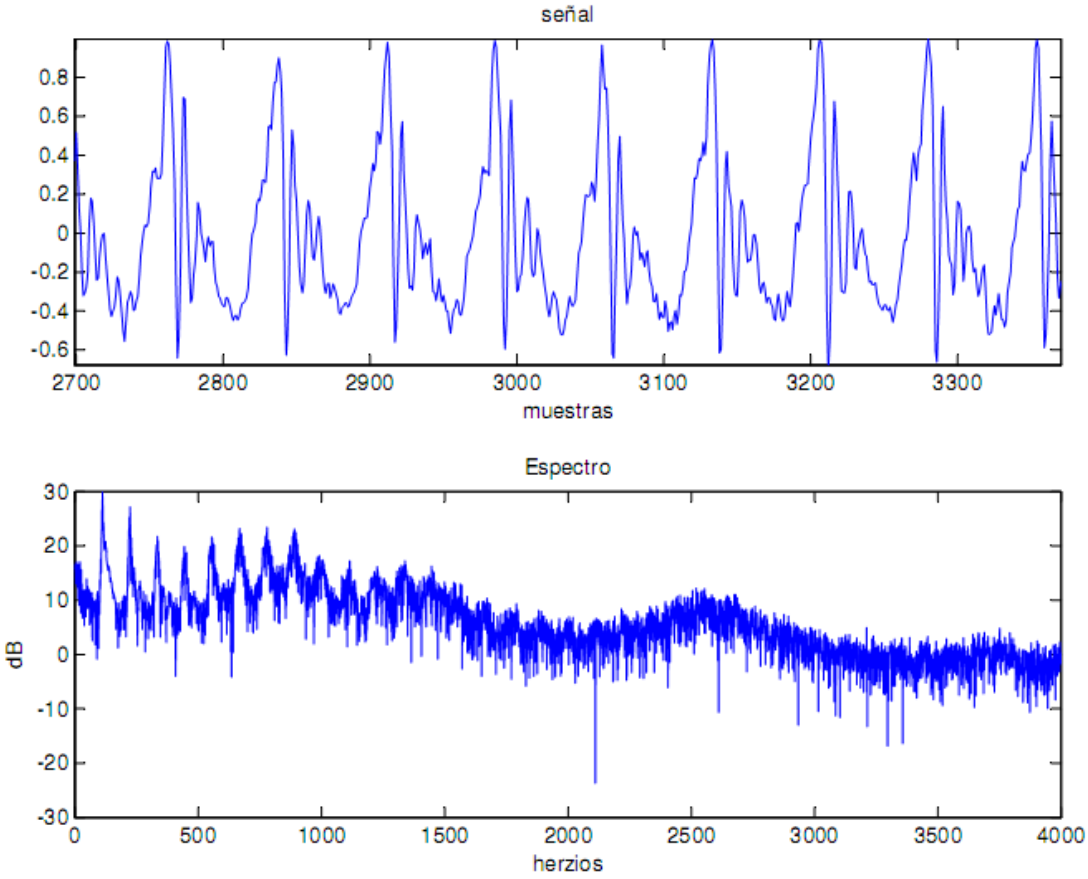


Figura 3: Señal de voz sonora en tiempo y su espectro en frecuencia en dB

Los sonidos no sonoros tienen una estructura típica aleatoria, sin periodicidades marcadas en el dominio del tiempo y un espectro mucho más compensado en frecuencia, sin los picos de energía de los armónicos (Figura 4).

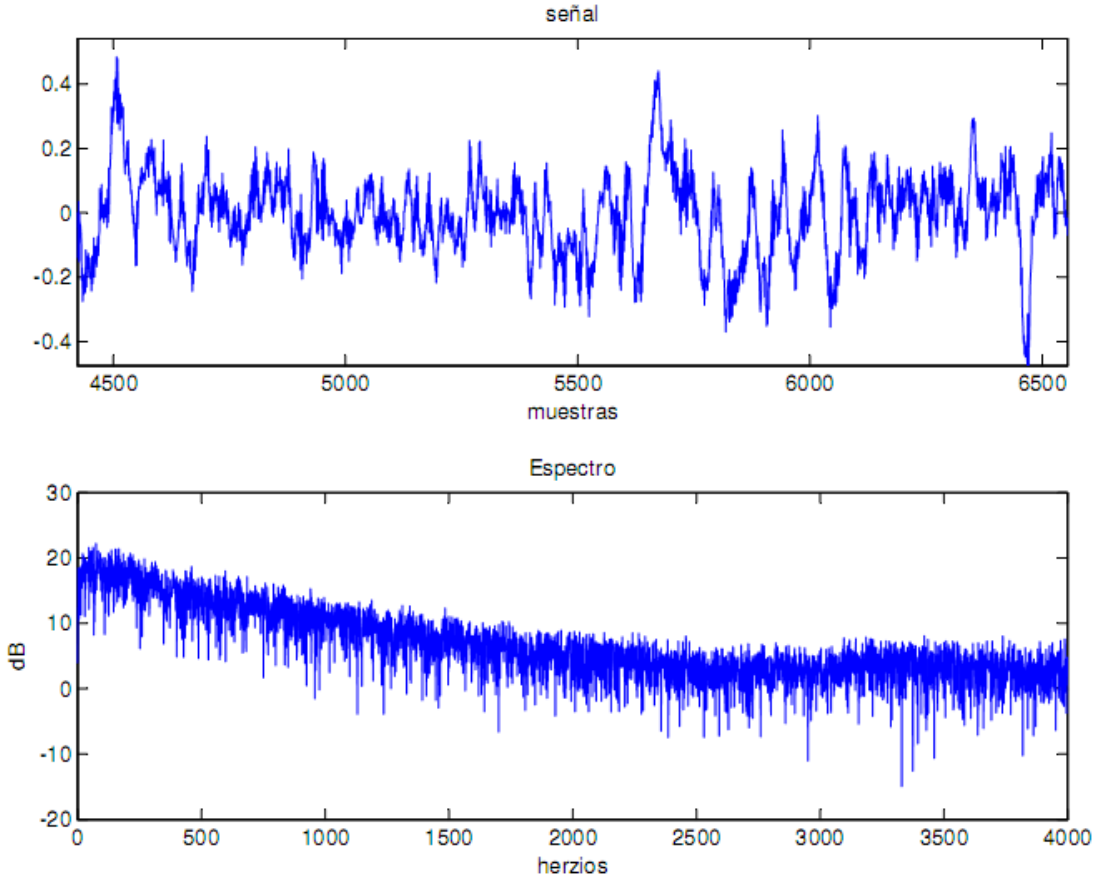


Figura 4: Señal sorda en tiempo y su espectro en frecuencia en escala en dB

Frecuencia de pitch

El periodo de los segmentos sonoros se caracteriza por un pitch o frecuencia fundamental en el dominio de la frecuencia. Este pitch es un parámetro importante para algunos algoritmos de codificación de voz. Se puede identificar como la periodicidad de los picos de la amplitud en la forma de onda y la estructura fina del espectro. Las frecuencias de pitch de hombres y mujeres normalmente se encuentran en el rango 50-250 Hz (4-20 ms) y 120-500 Hz (2-8,3 ms), respectivamente [L. Rabiner, B.H. Juang].

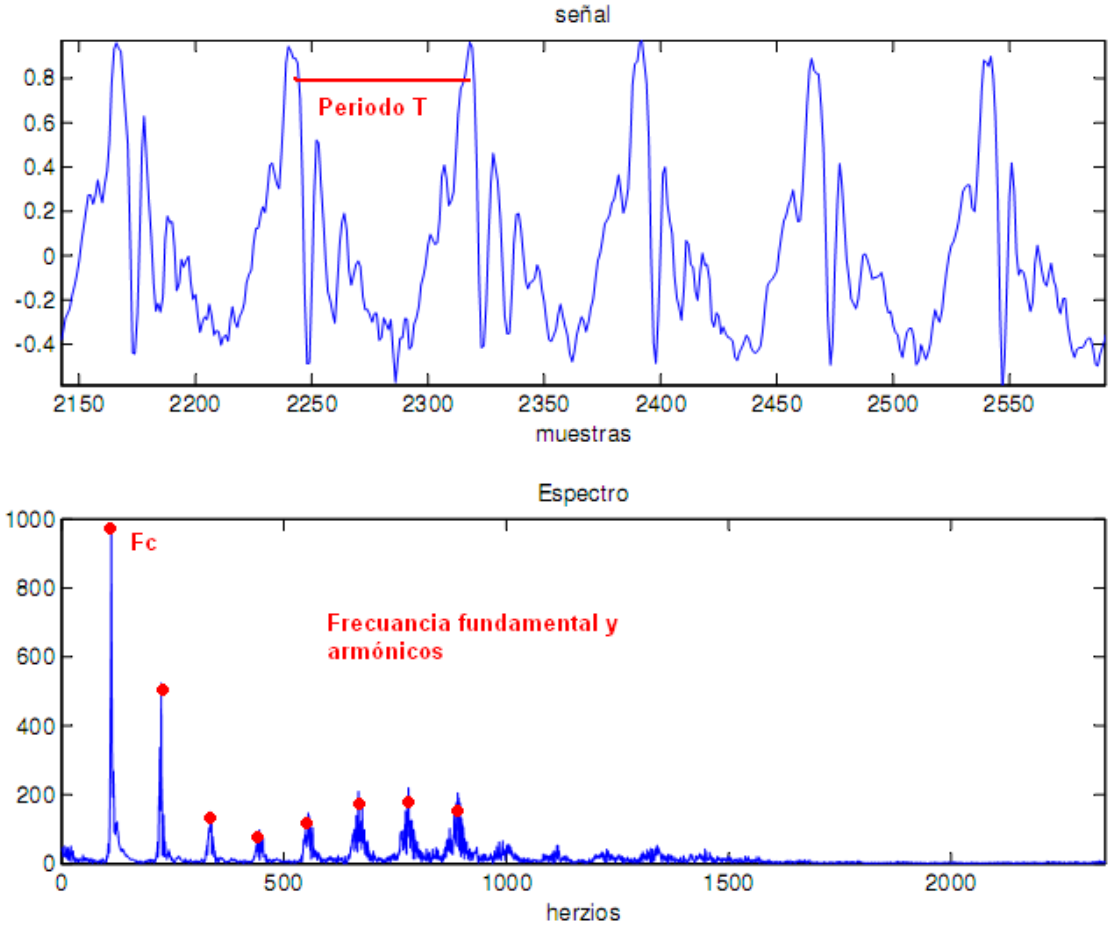


Figura 5: Señal sonora en tiempo y la magnitud del espectro en escala lineal

Formantes

Los sonidos sonoros consisten en una frecuencia fundamental (frecuencia de pitch) y una serie de componentes armónicos de la misma, producidos por las cuerdas vocales. El espectro de la señal de voz varía con el tiempo debido a las variaciones en la forma y en la posición del tracto vocal. Los formantes son las frecuencias de resonancia del espectro, es decir, los picos de la envolvente del espectro de la señal de voz que representan las frecuencias de resonancia del tracto vocal. En señales de voz esas frecuencias dependen del tamaño y de la forma del tracto vocal. Así pues, los formantes caracterizan a un sonido frente a los demás, y son los que nos permiten distinguir a las personas. En un fonema, los más importantes son los 3-4 primeros formantes, que son los que tienen la mayor parte de energía (Figura 6).

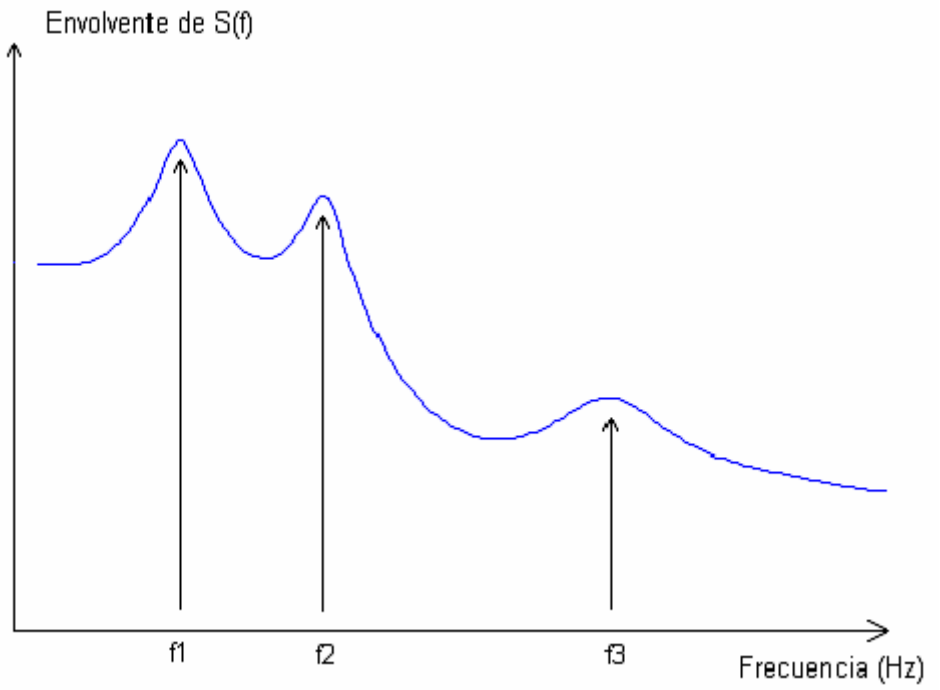


Figura 6: Formantes de la señal de voz en el espectro

CODIFICACIÓN DEL MENSAJE DE VOZ

La información a codificar en el mensaje hablado se puede dividir en 2:

1. El mensaje: La voz puede representarse como una concatenación de fonemas.

2. Información añadida: El mensaje oral no sólo son fonemas. Sino que son también la identidad del hablante, su estado anímico, la velocidad del habla, la intensidad, la entonación, etc. Sirven para que la información sea completa.

Para que estos sistemas funcionen correctamente es importante por una parte que preserven el mensaje para que no se reciba un mensaje distorsionado. Pero por otra existe un compromiso tecnológico, en el que el mensaje debe ser presentado de forma conveniente para su almacenaje, transmisión y manipulación. Los sistemas de transmisión de voz sobre telefonía (por cable o fija) y más concretamente los más modernos de voIP, disponen de BW finito.

Existen 3 grandes grupos de representación de la señal de voz:

1. Representación de forma de onda: Pretende mantener la forma de onda original de la señal analógica. Se basa en procesos de muestreo y cuantificación.

2. Representaciones híbridas: Utilizan un modelo de producción de voz modificado para obtener una calidad intermedia entre los dos anteriores. A este tipo de representación pertenece el vocoder homomórfico creado en este proyecto.

3. Representación paramétrica: Considera la señal de voz como la salida de un sistema de producción de voz, que puede ser representado por un conjunto de parámetros. Es el modelo que se utiliza en los vocoders tipo LPC.

    (a) Parámetros de excitación.

    (b) Parámetros de respuesta del tracto vocal.

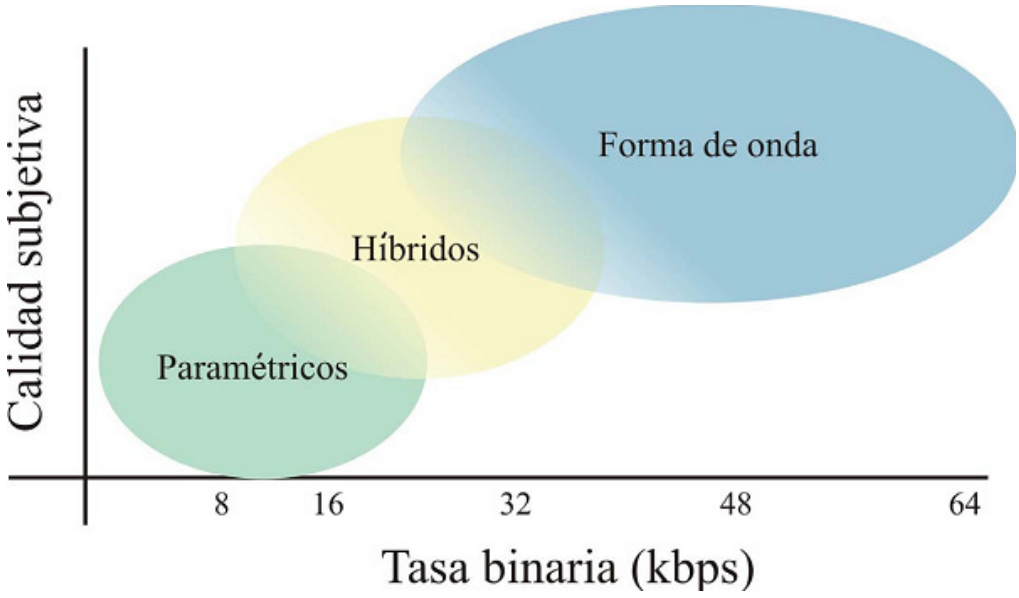


Figura 7: Diferentes tipos de representación en función de la calidad y tasa binaria

Como se puede observar en el gráfico la relación de requerimientos de BW aumentan con la calidad obtenida.

MODELO DE CREACIÓN DE LA VOZ

Se ha acordado que en general existen dos tipos de sonidos, sonoros y sordos. Estos se producen con diferentes excitaciones producidas por la exhalación de aire y modificadas por las características del tracto vocal así como por la posición que adoptan los elementos articuladores.

La generación de voz puede modelarse por un sistema, con una señal de excitación y un proceso de filtrado. Se utiliza por lo tanto un modelo convolucional en el que la señal de excitación que es el aire que sale de los pulmones es filtrada por un filtro que simula el tracto vocal. Ambos, excitación y tracto vocal, son variables en el tiempo. En el caso de los sonidos sonoros la excitación es similar a un tren de pulsos glotales y en el caso de los sonidos sordos es similar al ruido. Los formantes corresponden a los polos de la función de transferencia. En general un modelo sólo polos representa bastante bien la mayoría de las señales vocales. Sin embargo, la teoría acústica dice que los sonidos nasales y sordos, con un contenido mayor en altas frecuencias necesitan polos y zeros para ser representados correctamente [L.R.Rabiner/R.W Schafer].

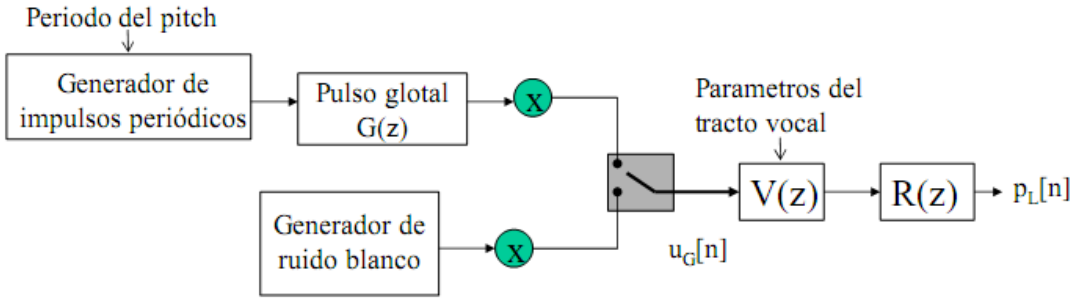


Figura 8: Modelo completo del aparato fonador

No obstante los efectos de los zeros se pueden representar incluyendo más polos.

El modelo solo polos en el que: $H(z) = G(z)\ V(z)\ R(z)$

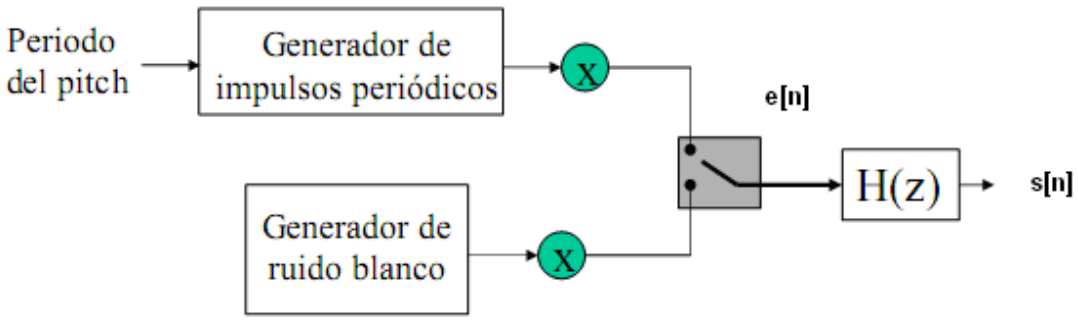


Figura 8: Modelo del aparato fonador sólo polos

Este es el modelo en el que se basan los vocoders paramétricos. La limitación principal de un sistema de codificación paramétrica, como puede ser el LPC, es la mala calidad de la señal de excitación, e[n]. Esta señal se aproxima por pulsos con una frecuencia de pitch, para tramas sonoras y un generador de ruido blanco para tramas sordas. Además para obtener la señal de excitación se usan muestras anteriores. Es decir, se aproxima la señal de excitación por una combinación lineal de muestras anteriores. De esta forma, entre la señal de voz original y la que predecimos se comete un cierto error.

En el caso del vocoder homomórfico la diferencia es que, en lugar de codificar la señal de excitación como una aproximación por tramas sonoras o sordas, esta, se codifica tal cual. Se obtiene directamente de la señal de voz s[n], por lo tanto no es necesario realizar una aproximación. Así, obtenemos una señal de excitación mucho más compleja y rica en matices, más fiel a la original.

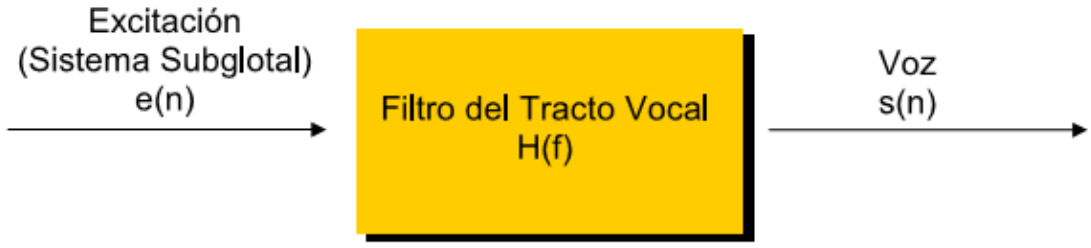


Figura 9: Modelo convolucional del con señal de excitación completa e[n]

Los resultados obtenidos con un decodificador homomórfico son mucho mejores como veremos en las pruebas, a costa de sacrificar necesidad de transmisión pues la tasa de bits aumenta considerablemente.

PROCESADO HOMOMÓRFICO

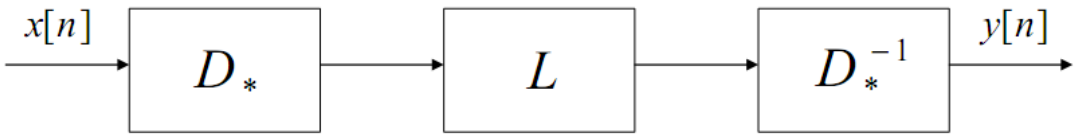
EL DOMINIO CESTRAL

El objetivo del análisis cepstral es separar la señal de voz en señal de excitación y filtro lineal sin saber a priori nada acerca del filtro o la excitación. La señal de voz s[n] es la convolución de la

$$s[n] = e[n] * h[n]$$

Como se ha mencionado anteriormente, el procesado homomórfico permite pasar de una convolución a una suma y extrar la señal señal de excitación e[n] y el filtro lineal h[n]

El esquema general del procesado homomórfico es el siguiente:



Donde x[n] es una trama de voz e y[n] es esa trama de voz en el cepstrum. Se habla de dominio cepstral para diferenciar del tiempo, y el eje horizontal representa Quefrency y no segundos o herzios. *Cepstrum* no es más que una conjugación de spectrum , así como *Quefrency* de frecuency.

Para que este esquema funcione, es necesario encontrar el operador D que transforme dos señales combinadas convolucionalmente en dos señales combinadas aditivamente. Además debe tener un operador inverso D^{-1}.

El operador D, lo vamos a obtener en 3 pasos:

- Transformada de Fourier: Permite pasar de una convolución a un producto:

$$x[n] = x_1[n] * x_2[n] \quad \Rightarrow \quad X(z) = X_1(z) \cdot X_2(z)$$

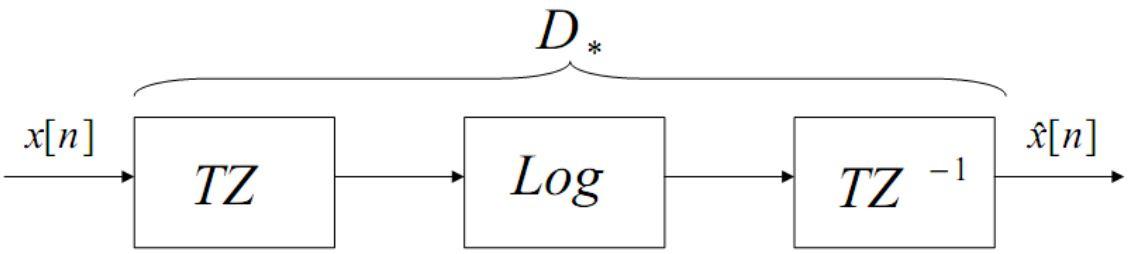
- Logaritmo complejo: Permite pasar de un producto a una suma:

$$\hat{X}(z) = \log(X(z)) = \log(X_1(z) \cdot X_2(z)) = \log(X_1(z)) + \log X_2(z)$$

- Transformada inversa Fourier: Permite la transformación al dominio cepstral.

$$\hat{x}[n] = Z^{-1}\{\log(X(z))\} = Z^{-1}\{\log(X_1(z))\} + Z_{-1}\{\log X_2(z)\} = \hat{x}_1[n] + \hat{x}_2[n]$$

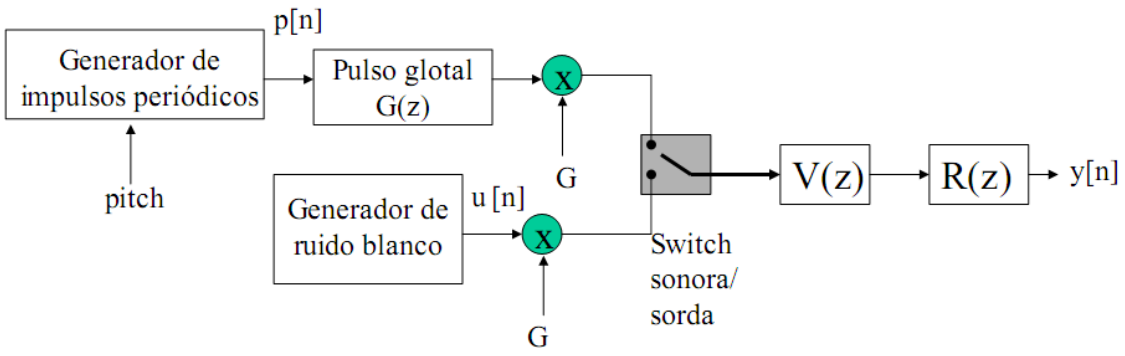
Esquemáticamente este es el proceso:



La señal que se obtiene a la salida se denomina cepstrum complejo.

Cepstrum de la señal de voz

Se ha mencionado más arriba como es el modelo completo que usamos para representar el aparato fonador:



En el que V(z) incorpora la respuesta del tracto vocal y se modela como:

$$V(z) = \frac{Az^{-M}\prod_{i=1}^{M_i}(1-a_k z^{-1})\prod_{k=1}^{M_o}(1-b_k z)}{\prod_{k=1}^{N_i}(1-c_k z^{-1})} \quad , \quad |c_k| < 1$$

Y que según la teoría acústica:

- Para los segmentos nasales es una función solo polos.
- Para los nasales y sordos se usan polos y ceros.

R(z) representa el efecto de la radiación y se modela como un único cero:

$$R(z) = 1 - z^{-1}$$

G(z) incorpora el efecto de la glotis y se modela como un sistema sólo ceros:

$$G(z) = B\prod_{k=1}^{L_i}(1-\alpha_k z^{-1})\prod_{k=1}^{L_o}(1-\beta_k z)$$

El cepstrum complejo es:

$$Z^{-1}\{\log H(z)\} = Z^{-1}\{\log G(z)\} + Z^{-1}\{\log V(z)\} + Z^{-1}\{\log R(z)\}$$

Si analizamos el cepstrum del tracto vocal para un sonido sonoro, sólo polos, será:

$$\hat{v}(n) = Z^{-1}\{\log V(z)\} = Z^{-1}\left\{\log\left[\frac{G}{\prod_{k=1}^{P}\left(1-c_k z^{-1}\right)}\right]\right\} = G\sum_{k=1}^{P}\frac{c_k^n}{n} \quad , \quad |c_k| < 1 \ , \ n > 0$$

El cepstrum complejo del tracto vocal tiene las siguientes propiedades:

$$G\sum_{k=1}^{P}\frac{c_k^n}{n}$$

- La distribución de energía depende de los polos ck.
- La energía decae rápidamente conforma aumenta n.
- La energía por lo tanto esta concentrada alrededor del origen.

La excitación es un tren de pulsos, en espectro un tren de deltas.

$$g(n)=\sum_{r=0}^{M}\alpha_r\delta(n-rN_p) \quad \text{Tren de impulsos} \quad G(z)=\sum_{r=0}^{M}\alpha_r z^{-rN_p}$$

Su cepstrum será:

$$\hat{g}(n)=Z^{-1}\{\log G(z)\}=Z^{-1}\left\{\log\left[\sum_{r=0}^{M}\alpha_r z^{-rN_p}\right]\right\}=$$

$$=\sum_{r=1}^{\infty}(-1)^{r+1}\frac{\alpha_r^r}{r}\delta(n-rN_p)\,,\quad 0<\alpha_r<1\,,\quad r>0$$

Podemos deducir que son unas deltas espaciadas NP.

Las propiedades del cepstrum de la excitación son las siguientes:

- La distribución de energía depende de los coeficientes α_r.
- La energía se concentra en las deltas equiespaiadas N_p.

En el siguiente esquema se puede apreciar como es la representación de una señal en el espectro así como su separación haciendo uso del procesado homomórfico (Figura 10).

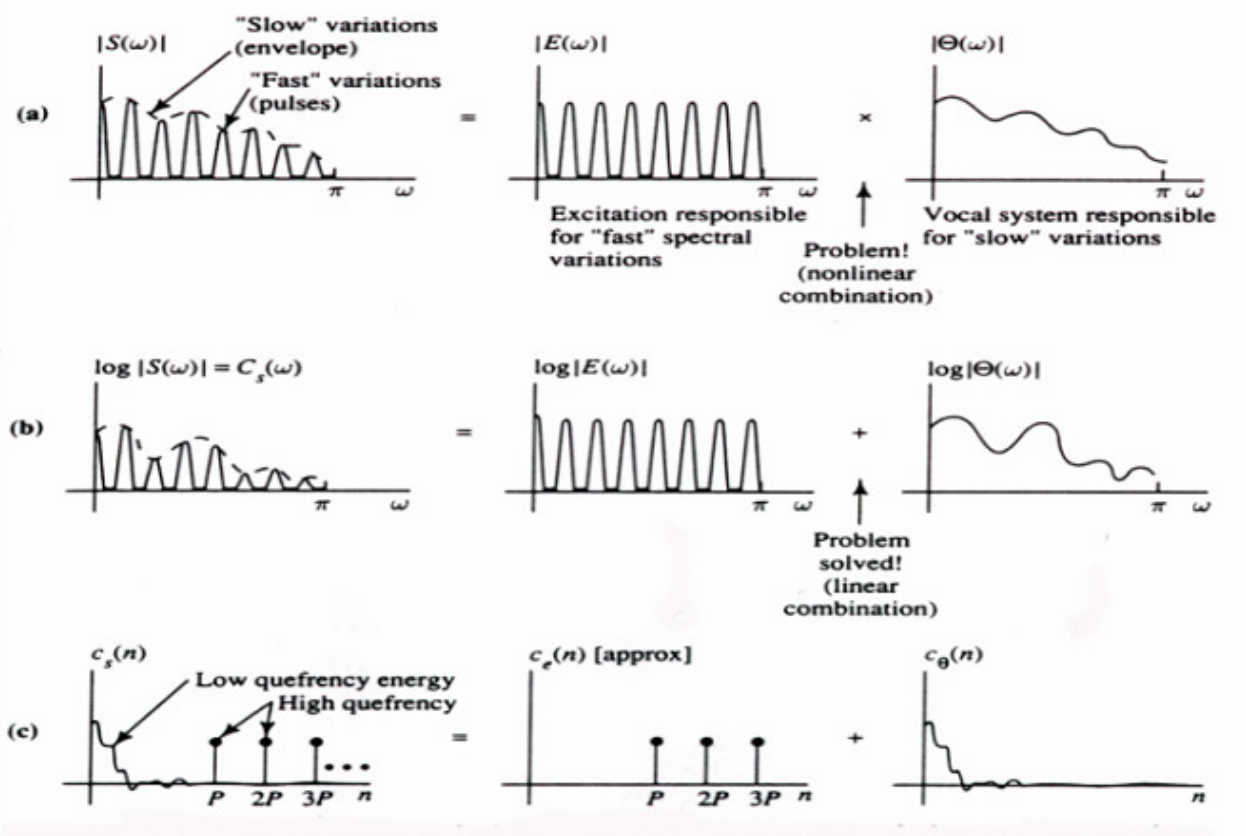


Figura 10: Representación del análisis cepstral

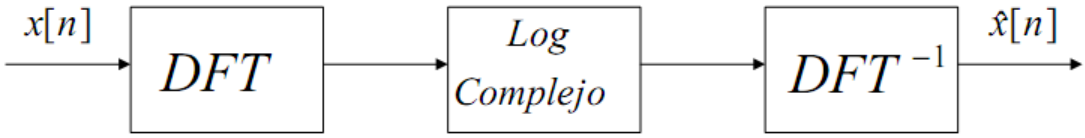
CONSIDERACIONES COMPUTACIONALES

Lo discutido aquí hasta ahora es conocido como el cepstrum complejo. Se computa a partir del valor de magnitud y de fase del espectro. El cepstrum real no tiene en cuenta el valor de la fase, sólo la magnitud. Por lo tanto no es posible la reconstrucción de la secuencia a partir del cepstrum real.

La transformada de Fourier, o Z, es un número complejo. Su logaritmo viene dado por:

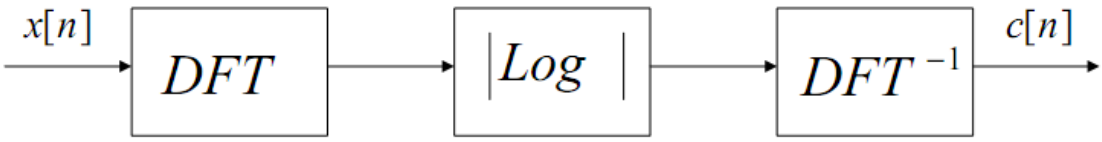
$$\log(X(w)) = \log(|X(w)|e^{j\angle X(w)}) = \log(|X(w)|) + j\angle X(w)$$

Para calcular el cepstrum complejo se utiliza el logarítmo complejo.



También se puede calcular el cepstrum real, tomando sólo la parte real del logaritmo, es decir, sin tener en cuenta la fase.

$$\log(|X(w)|)$$



c[n] es de la forma:

$$c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2}$$

El cepstrum real es útil a la hora de analizar la voz humana pues su representación en Quefrency permite discernir más fácil la información necesaria para el cálculo del pitch y así diferenciar una trama sonora de una sorda o para el análisis de los formantes.

SEPARACIÓN DE LAS COMPONENTES EN EL DOMINIO CEPSTRAL (LIFTERING)

La manera de extraer independientemente las características del tracto vocal y la excitación es aislando cada parte usando una ventana en el quefrency. Este proceso se conoce como Liftering (La conjugación del término filtering).

En líneas anteriores se ha hablado sobre el cepstrum y se ha dicho que el tracto vocal lo compone la información cercana al origen mientras que la excitación esta contenida en la parte alta del cepstrum, cerca de la posición del primer pico y más allá.

Se usarán ventanas para separar o filtrar los componentes del tracto vocal y la excitación, las cuales se multiplican con el cepstrum en el proceso conocido como liftering.

El tracto vocal esta contenido en las primeras muestras de quefrecncy, siempre más abajo del primer pico, sobre las primeras 20-30 muestras. Se usará una ventana rectangular que es nula más allá de n0.

$$\left.\begin{array}{ll} l[n]=1, & |n|<n_0 \\ 0, & |n|>n_0 \end{array}\right\} \quad n_0 < N_p$$

Donde NP es la posición del primer pico.

Para la extracción de la excitación el proceso es el mismo, sólo que la máscara es nula desde el origen hasta n0.

$$\left.\begin{array}{ll} l[n]=0, & |n|<n_0 \\ 1, & |n|>n_0 \end{array}\right\} \quad n_0 < N_p$$

REGRESO AL DOMINIO TEMPORAL

Una vez hemos separado la respuesta del tracto vocal y la señal de excitación, debemos volver de vuelta al dominio temporal mediante en operador inverso D^{-1}.

El operador inverso D^{-1}, que nos permite recuperar la señal original, se realiza también en 3 pasos:

- Transformada Z (o transformada de Fourier):

$$Z\{\hat{x}[n]\} = \hat{X}(z) = \log(X(z))$$

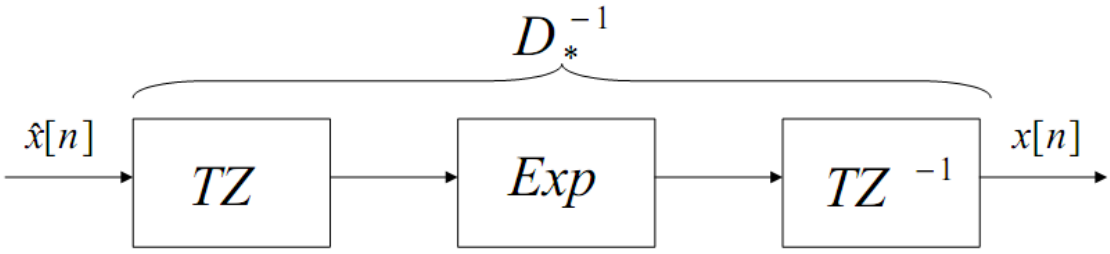
- Función exponencial:

$$\exp(Z\{\hat{x}[n]\}) = \exp(\log(X(z))) = X(z)$$

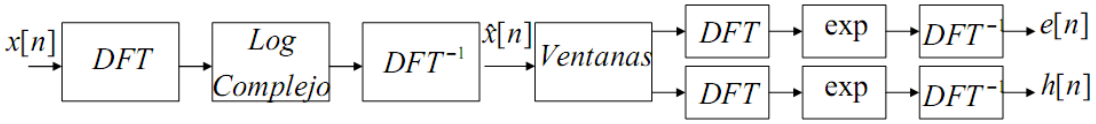
- Transformada Z inversa (o de Fourier):

$$Z^{-1}\{\exp(Z\{\hat{x}[n]\})\} = Z^{-1}\{\exp(Z(\hat{x}[n]))\} = Z^{-1}\{X(z)\} = x[n]$$

El esquema:



Esquema general de todo el procesado homomórfico



DIGITALIZACIÓN DE LA SEÑAL

En esta sección se explicara los motivos por lo se que se ha elegido la frecuencia de muestreo de 8000 Hz con 16 bits de cuantización para las señales que se someterán a las pruebas

MUESTREO

La frecuencia de muestreo elegida para las señales usadas en este proyecto es de 8 KHz. De acuerdo al modelo de producción de voz, la señal de voz no esta inherentemente limitada en banda. La amplitud de las altas frecuencias tiende a caer rápidamente. Se ha observado que para una señal sonora las altas frecuencias están más de 30 dB por debajo del pico del espectro para las frecuencias por encima de 4 KHz. Las señales no-sonoras o sordas no presentan esa caída tan pronunciada hasta los 8 KHz. Los primeros formantes (hasta el 4º) habitualmente se localizan antes de los 3400 herzios. Por lo tanto bastará esa información para reconocer los fonemas y al interlocutor. Además la transmisión telefónica tiene un efecto de paso bajo filtrando las señales a partir de 3,5 KHz. En la siguiente página se representa el espectro de una señal de conversación típica telefónica [L.R.Rabiner/R.W Schafer] (figura 11).

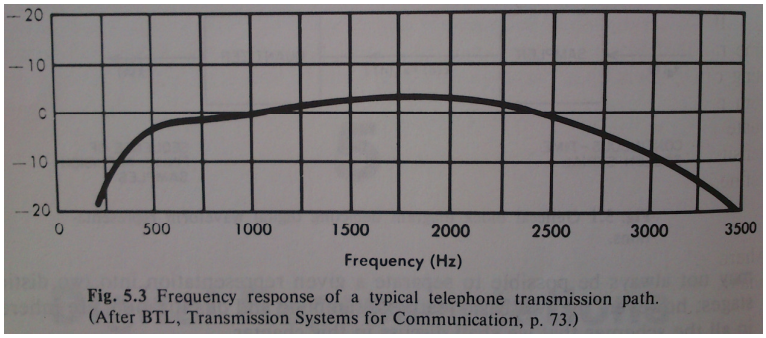


Fig. 5.3 Frequency response of a typical telephone transmission path. (After BTL, Transmission Systems for Communication, p. 73.)

Figura 11: Efecto del filtro paso bajo de la línea telefónica sobre la señal de voz

Por lo tanto, a efectos prácticos y computacionales, se asume que una frecuencia de muestreo de 8 KHz es suficiente para señales de transmisión telefónica, puesto que no compromete la inteligibilidad del mensaje.

Las figuras siguientes muestran el espectro de una señal mixta, 'Universidad Pública de Navarra' (Figura 12).

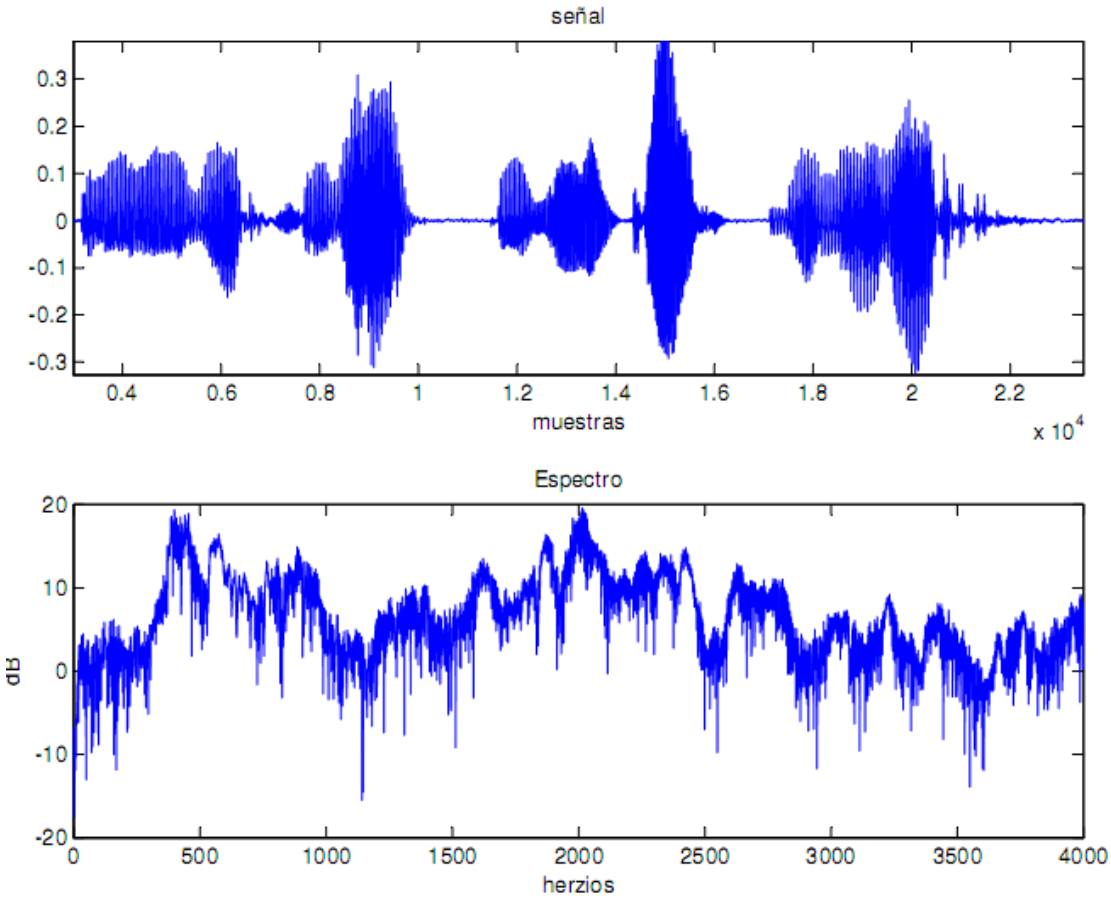


Figura 12: Señal en tiempo y su espectro con escala en dB

Se puede ver como la mayor parte de la energía esta contenida en bajas frecuencias, sobre los 500 Herzios, y en la zona de 2000 Herzios.

CUANTIZACIÓN

En realidad una cuantificación de 8 bits (2^8=256 niveles) resulta pausible pero esta origina una señal con bastante ruido de cuantización, y poco margen dinámico. Para una señal de 8 bits el SNR es de

50 dB, por lo tanto lo usual es usar 16 bits de cuantificación, lo que da 65536 niveles y un SNR de 98 dB. También se emplean técnicas de cuantificación no lineales que dan una calidad excelente.

Por lo tanto la entrada del programa tiene una señal digital a 8 KHz y 16 bits.

DESCRIPCIÓN DE LAS HERRAMIENTAS Y MÉTODO DE EVALUACIÓN

Este es el esquema general del vocoder homomórfico:

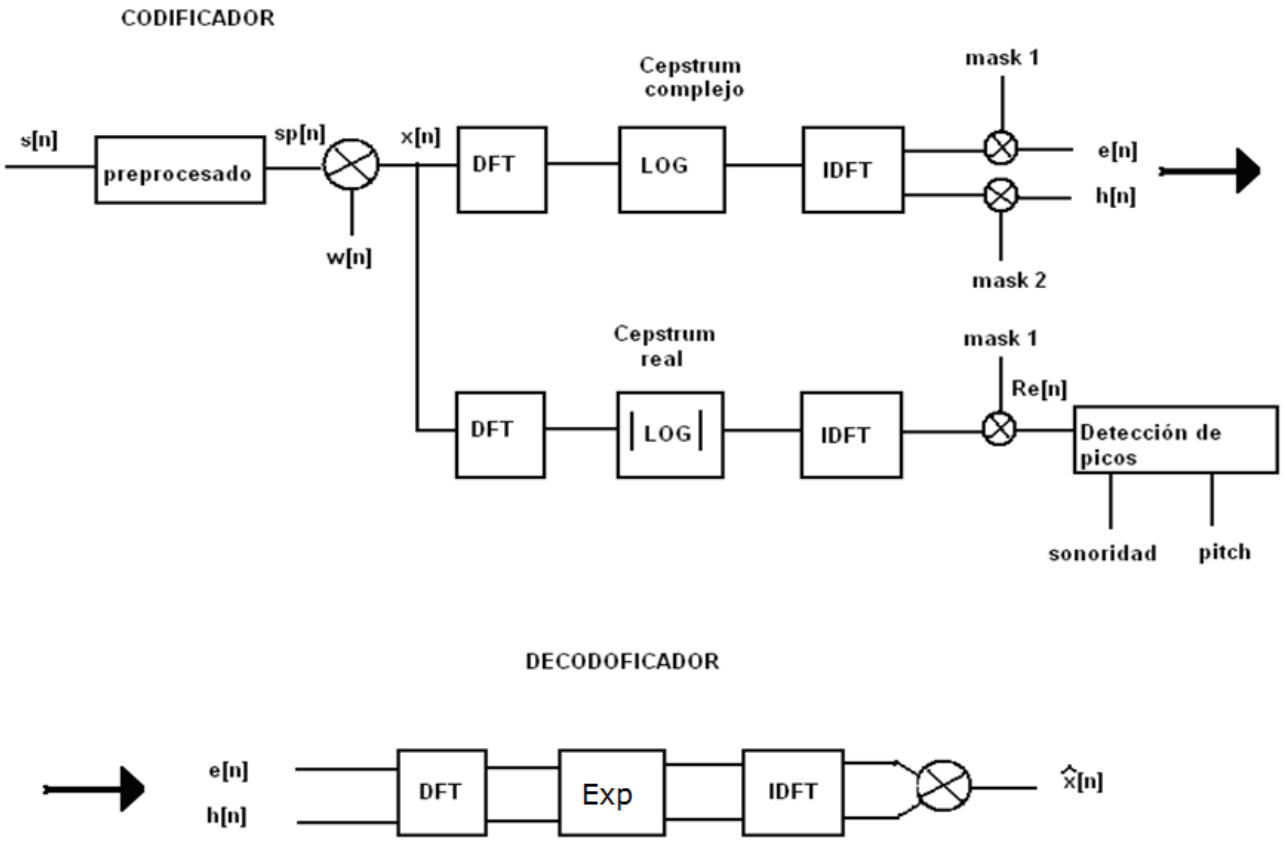


Figura 13: Esquema del vocoder homomórfico

Como se observa la señal original sufre un preprocesamiento previo al enventanado.

Se calcula el cepstrum complejo de la señal y de ahí se saca la señal de excitación y el tracto vocal, que es lo que se usa para reconstruir la señal.

El cepstrum real se usa para calcular la sonoridad y el pitch de la señal, así como para realizar un análisis de los formantes del tracto vocal.

PREPROCESADO

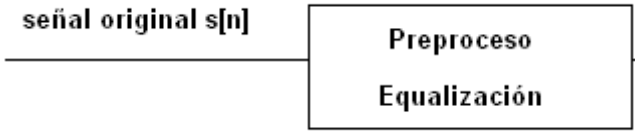


Figura 13: Esquema del vocoder homomórfico

Los pulsos glotales y los labios producen un efecto de pérdida de energía a altas frecuencias, similar a un filtro paso bajo lo que se traduce en una caída sensible de energía a altas frecuencias. Este efecto es consecuencia de la fricción del aire en las paredes del tracto local y sobre todo debido al efecto de la radiación en los labios. La energía disipada debida a la radiación es proporcional a la parte real de la impedancia de la radiación [L.R.Rabiner/R.W Schafer] (Figura 14). Además típicamente, los segmentos de voz tienen una curva espectral negativa. A esto hay que añadir que la audición humana es más sensible en la zona alrededor de los 2,5KHz, y hay que considerar que para la inteligibilidad de la voz esa es una zona crucial. Para contarrestar esta caída se aplicará una enfatización a las frecuencias altas de la señal de voz

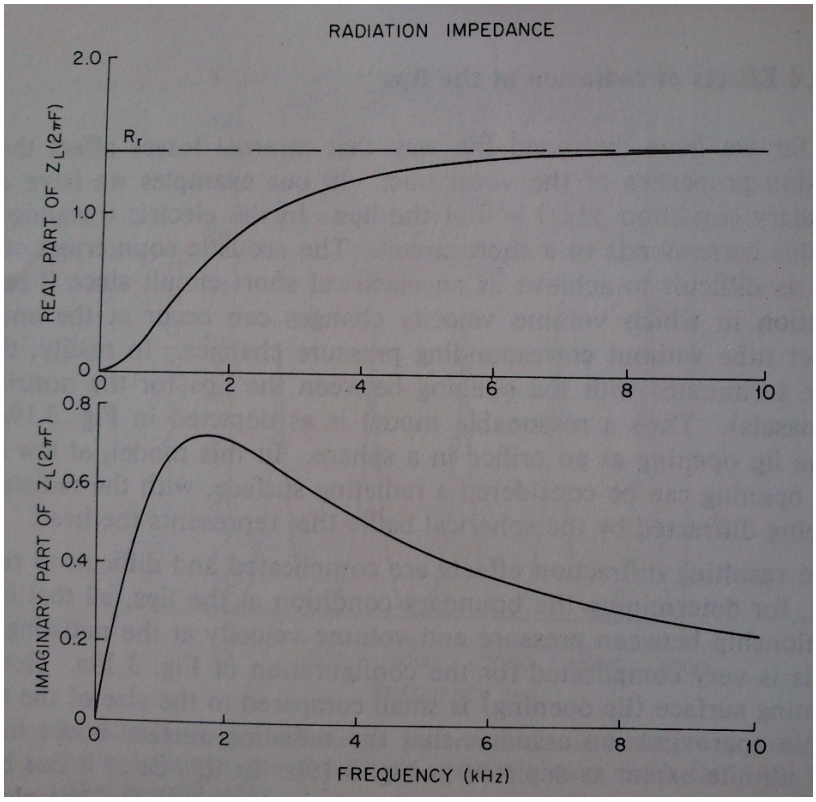


Figura 14: Efecto de impedancia de la radiación en la señal de voz

Para que la señal de excitación obtenida sea más inteligible y compensar esa pérdida en altas frecuencias se le aplica un filtro equalizador, que corrige la caída típica y aplana el espectro. Esta aplicación se realiza después de que la señal de entrada se tiene digitalizada.

Generalmente se usa un filtro digital de primer orden cuya función de transferencia es la siguiente:

$$H(z) = 1 - a \cdot z^{-1}$$

Donde 0,9≤ a ≤ 0,95, valor que se escoge cercano a la unidad a fin de que la parte alta del espectro sea acentuada. La representación en frecuencia del filtro de preénfasis mencionado se muestra en la figura número 15.

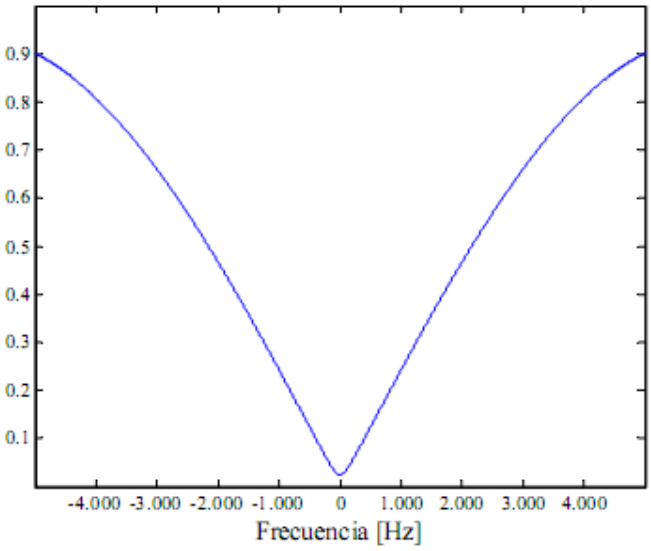


Figura 15: Filtro de preprocesado

En la siguiente figura se observa como el espectro se aplana considerablemente.

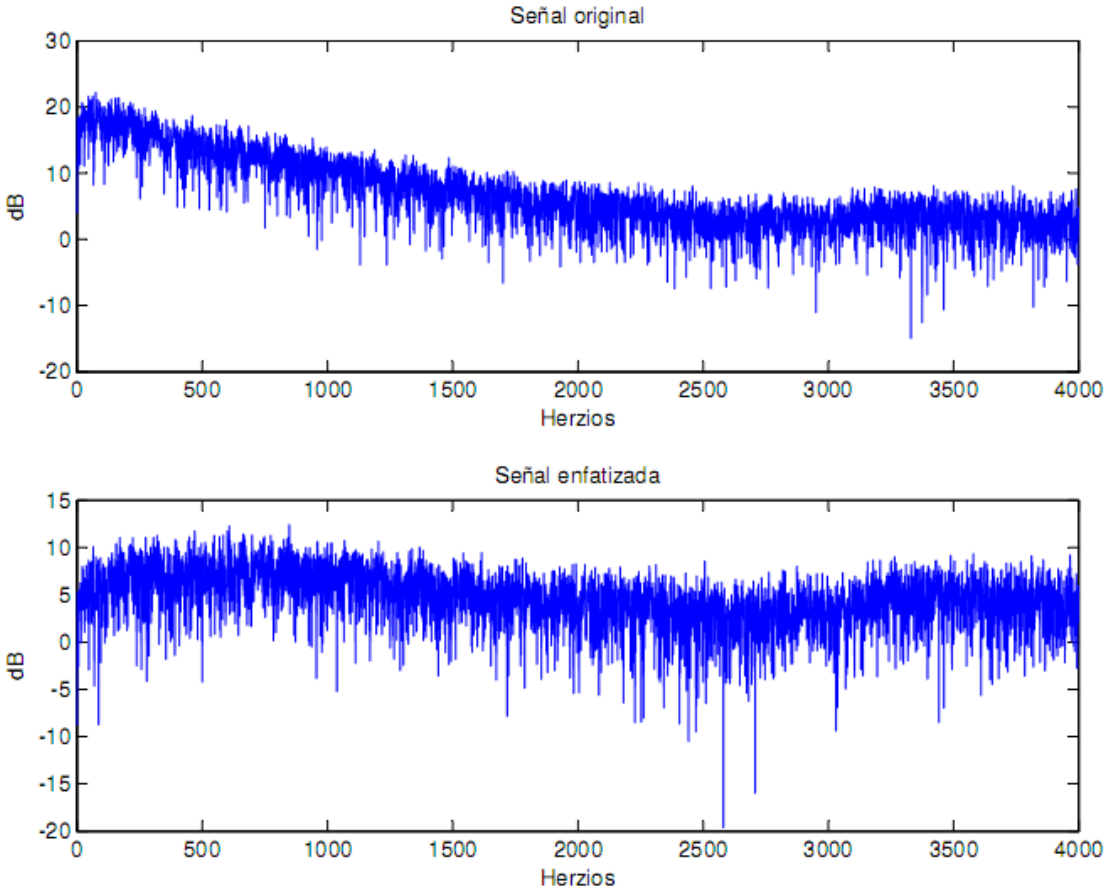


Figura 15: efecto del filtro de preprocesado sobre la señal de voz

Como se observará más adelante, la aplicación de este enfatizador de altas frecuencias mejora los resultados obtenidos en la decodificación.

ENVENTANADO

La señal se analiza en tramas regulares, para que las propiedades de la señal de voz se mantengan cuasi-estacionarias.

El enventanado requiere que cada una de las tramas sea multiplicada por una función limitada en tiempo de tal manera que su valor fuera de ese intervalo sea nulo. De esta forma, el enventanado consiste en agrupar las muestras de la señal x[n] en bloques de N elementos y multiplicarlas por una ventana w[n].

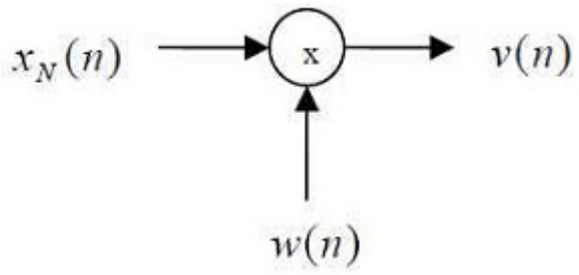


Figura 16: Aplicación de la ventana a la señal de voz

Cuanto más rápidamente cambien las características de la señal, más corta deberá ser la ventana para poder detectar esos cambios en el tiempo. Por otra parte a medida que decrece la longitud de la ventana se reduce la resolución frecuencial, es decir la capacidad de distinguir componentes cercanas en frecuencia.

Por lo tanto aparece un compromiso en la selección de la longitud de la ventana entre la resolución en tiempo y en frecuencia. Queda recogido en líneas anteriores que las características de la voz permanecen semi-estacionaias en periodos de tiempo cortos, normalmente comprendido entre los 10 y los 60 ms. Se trata por lo tanto de elegir una duración temporal de la ventana que mantenga las propiedades de la señal de voz estacionarias.

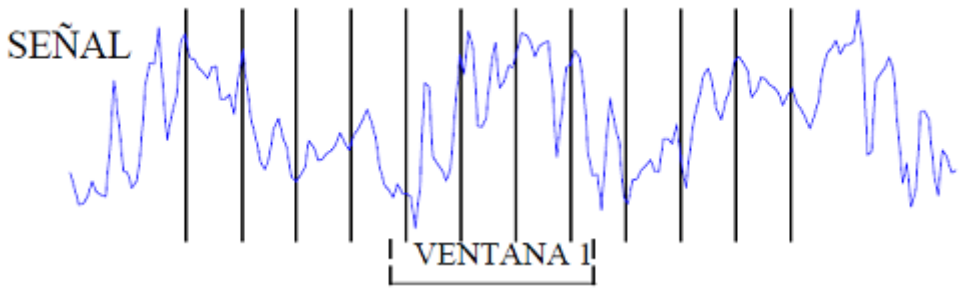


Figura 17: Señal de voz y las diferentes ventanas

Existen diferentes tipos de ventanas temporales. La más típica es la ventana rectangular. Otras ventanas son curvas, suavizándose en los bordes. Cada una de ellas tiene un efecto en el dominio espectral. Al multiplicar la señal por una ventana, la representación frecuencial se verá alterada, y lo que en tiempo es una multiplicación, en el espectro se convierte en una convolución.

$x[n]\, w[n] \rightarrow X(\Omega) * W(\Omega)$

La representación espectral de la funciona ventana rectangular es una sinc.

$$\mathrm{sinc}_N(x) = \frac{\sin(\pi x)}{\pi x}$$

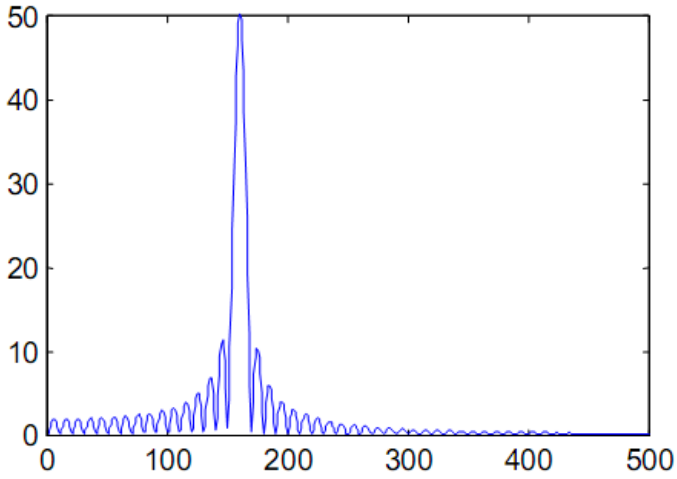


Figura 18 : Representación en módulo de un seno convolucionado con un sinc.

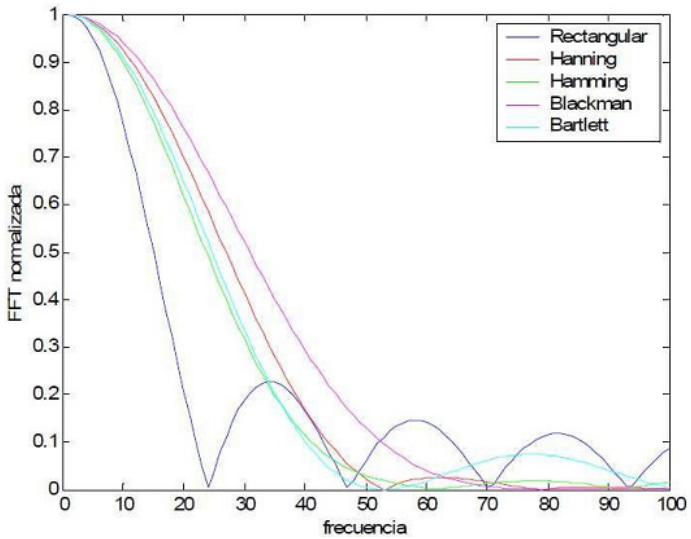


Figura 19: Espectro de las diferentes ventanas

Por lo tanto la representación frecuencial de la señal se verá alterada por el rizado que produce la sinc. Los demás tipos de ventanas tienden a suavizar el rizado a costa de aumentar la anchura del lóbulo central. En el diseño planteado esto no es crucial puesto que el análisis no se centra tanto en el

dominio frecuencial si no en el cepstral. Sin embargo, es importante tener en cuenta el efecto que produce el enventanado de la señal.

CEPSTRUM COMPLEJO

En este paso es donde se hace todo el proceso de transformación de la señal al dominio cepstral, liftering, y extracción por separado de la señal de excitación y del tracto vocal. Se seguirá el proceso descrito en la sección 3.

CODIFICADOR

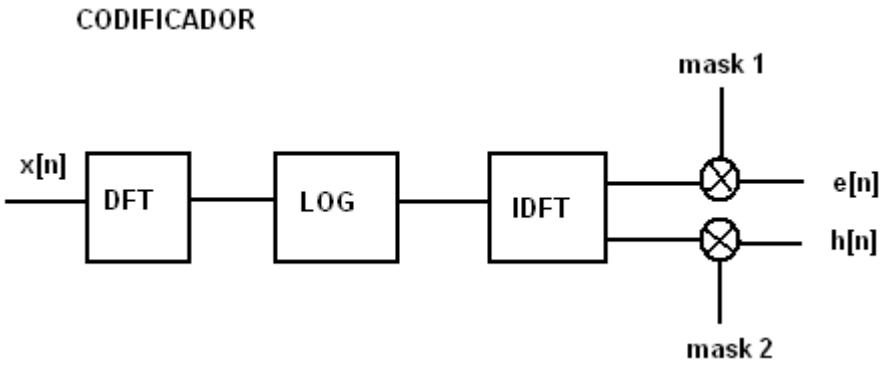


Figura 20: Codificador homomórfico

La transformada inversa de Fourier permite pasar de tiempo a frecuencia. En este caso como la DFT ha sido alterada por el logaritmo la representación es en el dominio cepstral. Además como el cepstrum proviene del logaritmo de un espectro, es también simétrico. Por lo tanto sólo interesa una parte, motivo por el que se ha dibujado la parte derecha del cepstrum.

Para este ejemplo se usará una señal de voz sonora, la vocal 'a'.

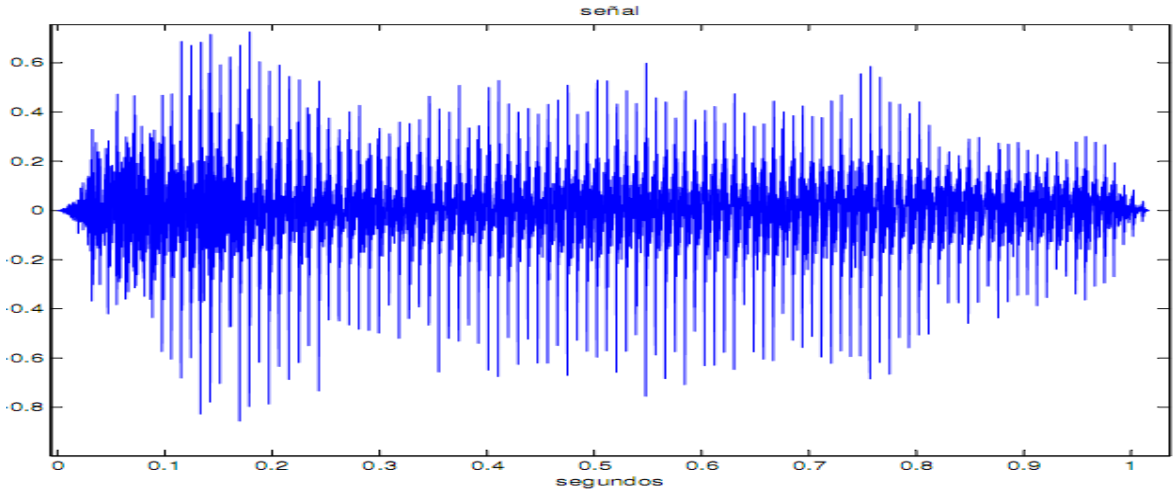


Figura 21: Señal de voz en tiempo 'a'

El primer paso es enventanar la señal. Para el ejemplo se usará una ventana de 60 ms, que a la frecuencia de muestreo de 8 KHz da ventanas de 480 muestras.

Figura 22: Esquema del vocoder homomórfico

Se observa claramente la periodicidad de la señal.

En la siguiente figura se observa el módulo de la DFT de la ventana. Al ser una señal sonora el espectro presenta mayor energía en bajas frecuencias.

Figura 23: Esquema del vocoder homomórfico

Se aplica el logaritmo al espectro para el posterior paso al cepstrum.

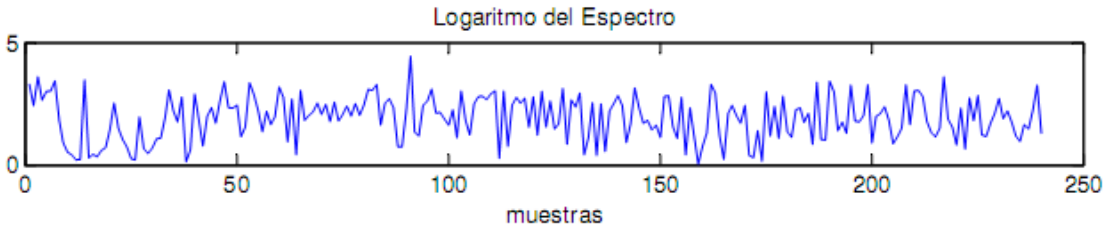


Figura 24: Esquema del vocoder homomórfico

El logaritmo es una versión comprimida del espectro, en el que han desaparecido los picos dominantes y la energía ya no se concentra en una zona concreta.

El último paso es aplicar una DFT inversa. Al aplicar el logaritmo al espectro no volvemos a tiempo sino al cepstrum

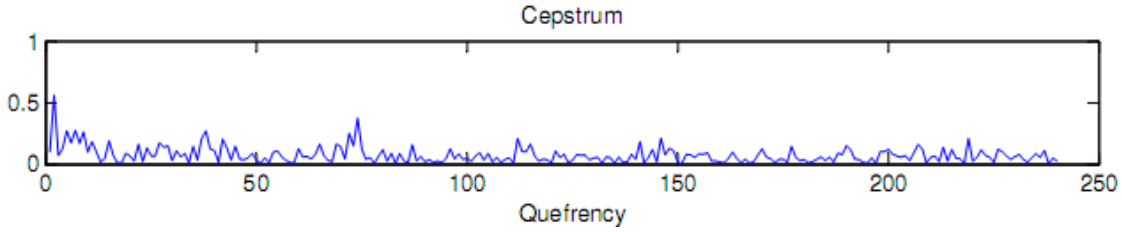


Figura 25: Esquema del vocoder homomórfico

En la última gráfica se pueden apreciar las características del cepstrum de la señal de voz. Estas características son las siguientes:

- La información del tracto vocal está contenido en las muestras cercanas al origen.
- La excitación son los picos equiespaciados, de los que sólo el primero es fácilmente reconocible en la muestra 75 más o menos.

Una vez con la señal en el dominio cepstral, hay que crear unas ventanas con las que se hará el proceso de liftering y se separaran el tracto vocal y la señal de excitación.

En el ejemplo, el tracto vocal se ha declarado como las primeras 30 muestras, así que se ha creado una ventana que es nula a partir de la muestra 31. Si se usan más muestras para el tracto vocal, se entra en una zona del cepstrum donde la señal se empieza a mezclar con la señal de excitación. Se pueden usar menos muestras para caracterizar el tracto vocal, pero cuantas más información para modelar el filtro H(Z) mejor. Por lo tanto se ha de llegar a un compromiso y finalmente se ha decidido usar 30 muestras para el cepstrum, por que es un estándar

Se multiplica la máscara creada con el cepstrum (Figura 26)

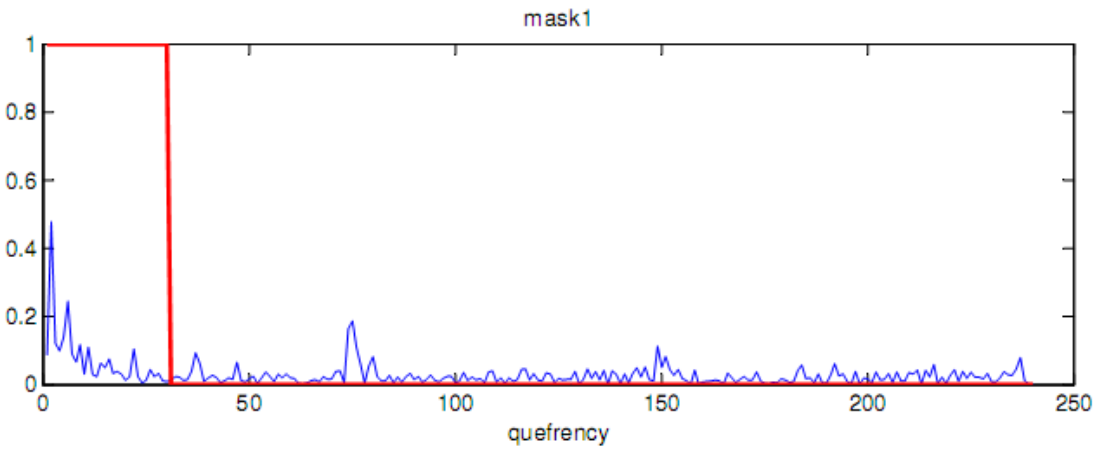


Figura 26: Aplicación de la ventana del tracto al cepstrum de la señal de voz (Módulo)

Y este es el tracto vocal extraído del cepstrum real (Figura 27)

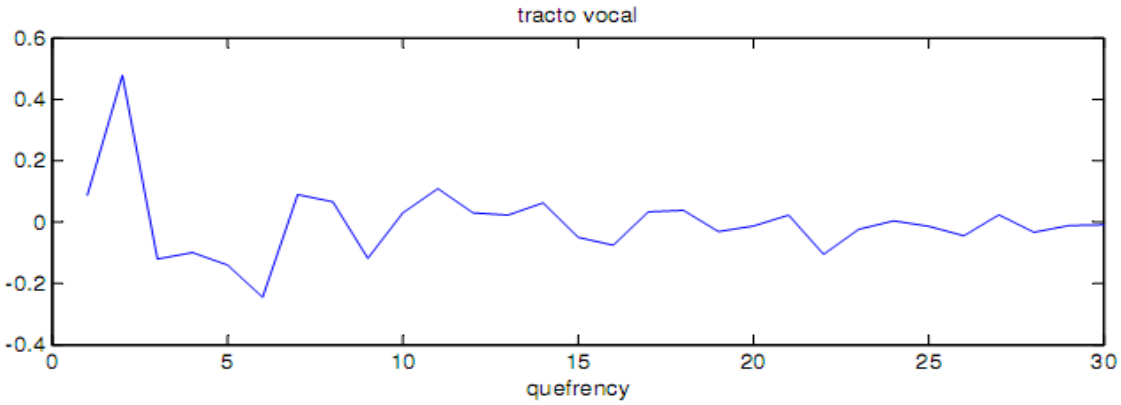


Figura 27: Cepstrum del tracto vocal

Para la extracción de la señal de excitación se crea una máscara con valores nulos para las primeras 30 muestras.

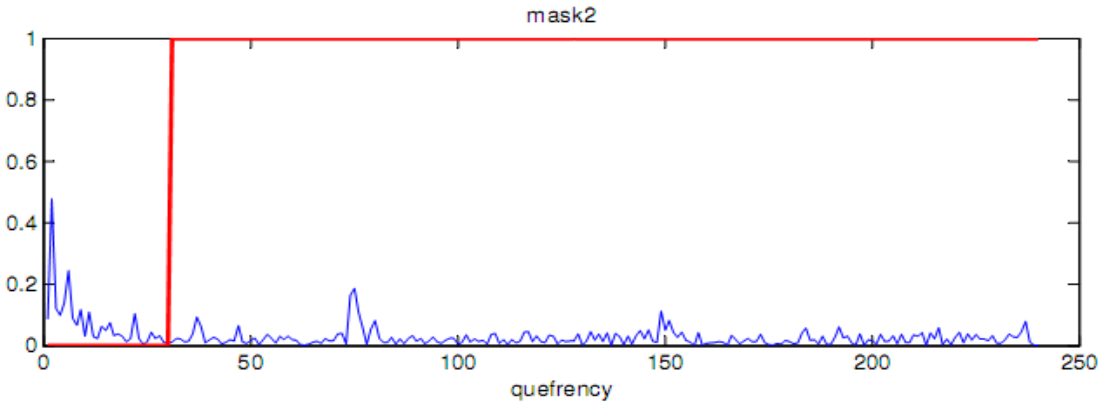


Figura 26: Aplicación de la ventana para la extracción del cepstrum (Módulo)

Y este es la excitación extraído del cepstrum real (Figura 29)

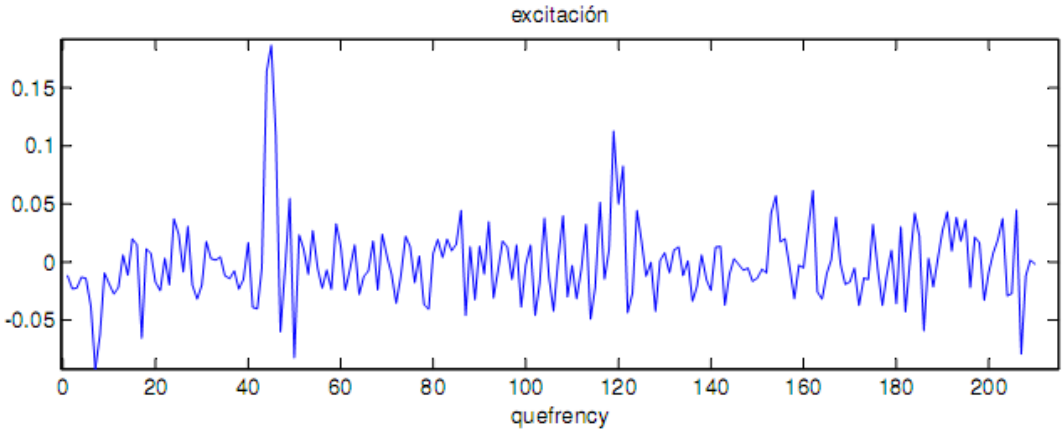


Figura 29: Excitación en el cepstrum

Con la señal de excitación y tracto vocal por separado en el dominio cepstral, se realizará el proceso inverso para llegar con las señales por separado al punto de partida, en el dominio tiempo.

DECODIFICADOR

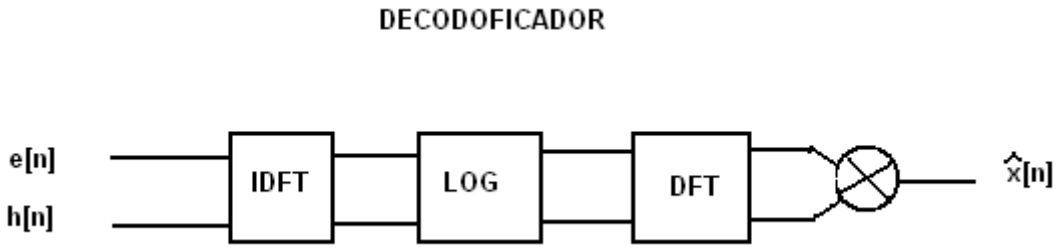


Figura 36: Esquema del decodificador homomórfico

Más adelante se presenta el ejemplo de una señal sonora (vocal 'a') y su deconvolución en las componentes tracto vocal y excitación así como su representación en cepstrum, frecuencia, log de espectro y tiempo.

La componente del **tracto vocal** correspondiente a las primeras 30 muestras del cepstrum.

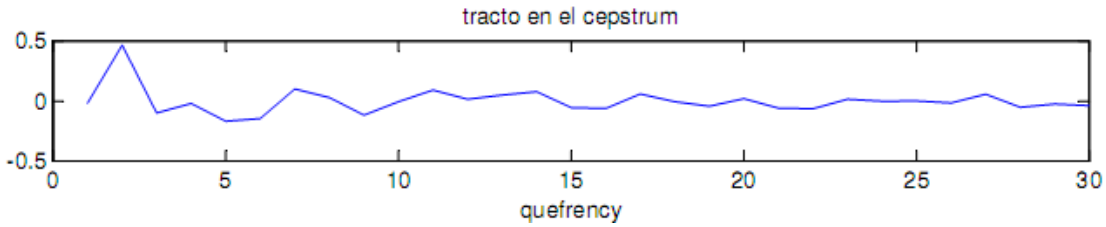


Figura 30: Esquema del vocoder homomórfico

Se aplica la DFT para obtener el logaritmo del espectro del tracto vocal. Es interesante apreciar que este logaritmo del espectro es una versión suavizada del logaritmo el espectro de la señal, y puede servir para localizar los formantes.

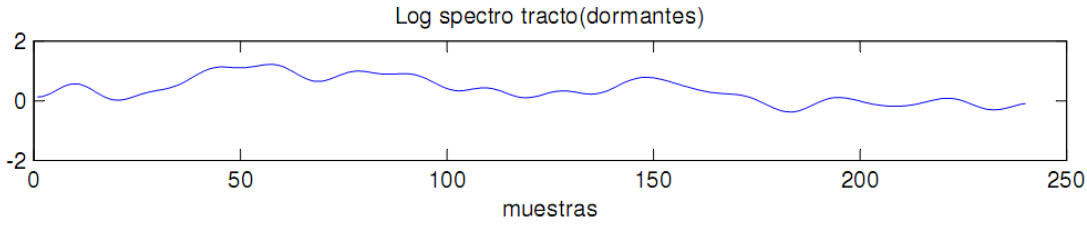


Figura 31: Esquema del vocoder homomórfico

Para eliminar la transformación logarítmica usamos la función exponencial. Se puede observar que el espectro no ha variado mucho. El logaritmo tiene menos rango dinámico.

Y por último se aplica la transformada inversa para obtener la respuesta al impulso.

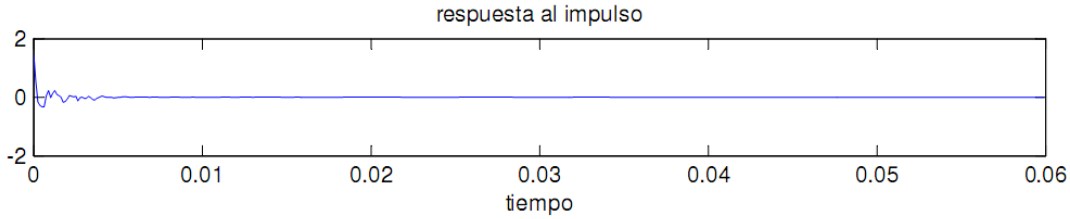


Figura 32: Esquema del vocoder homomórfico

Se hará el mismo proceso pero con la componente de **excitación** del cepstrum.

Aquí se obtiene la componente de la excitación en el cepstrum. Se puede observar como la parte que falta es la correspondiente al tracto vocal.

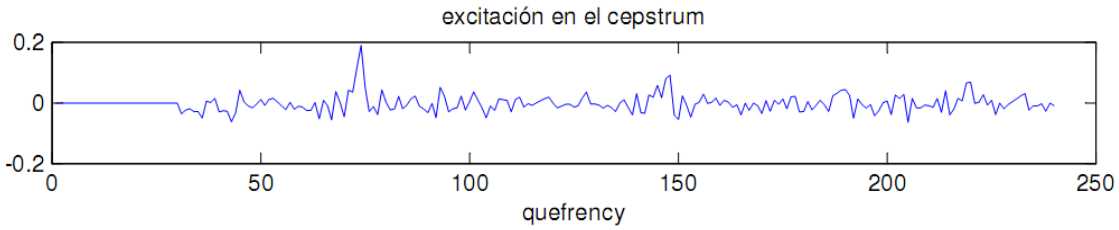


Figura 33: Esquema del vocoder homomórfico

Este es el logaritmo del espectro de la señal de excitación. Se observa fácilmente la diferencia con el tracto vocal. La información se concentra alrededor de los picos.

Y aquí se muestra el espectro de la señal de excitación tras aplicar la función exponencial. Se puede ver que es un espectro de una señal sonora, su frecuencia fundamental y los armónicos.

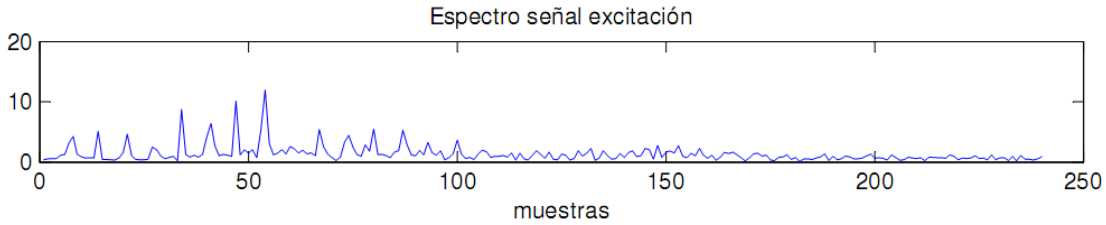


Figura 34: Esquema del vocoder homomórfico

Y finalmente se obtiene la señal de excitación en el dominio temporal.

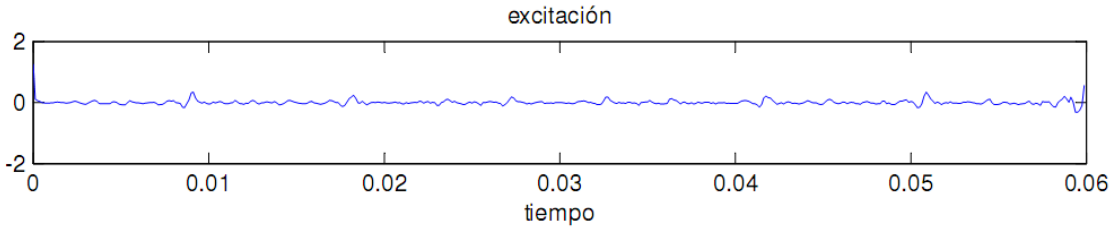


Figura 35: Esquema del vocoder homomórfico

Este es el proceso completo para separar la excitación del tracto vocal y volver con las señales independientes al dominio del tiempo. El proceso para la obtención de la señal de voz es filtar la señal de excitación con el tracto vocal.

CEPSTRUM REAL

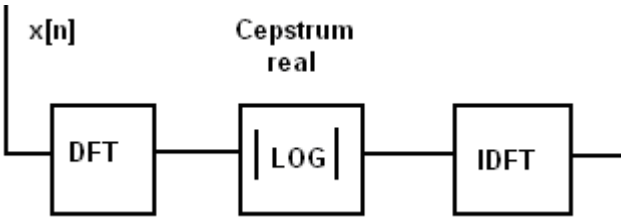


Figura 36: Esquema de la obtención del cepstrum real

Anteriormente se ha comentado que para el cálculo del pitch y la sonoridad era preferible analizar el cepstrum real. A continuación se explica por qué.

En la Figura 37, se realiza la representación cepstral de la señal en tiempo enventanada mostrada en la primera gráfica. Esta primera gráfica corresponde a la ventana de la señal en tiempo. La segunda corresponde al cepstrum complejo y la tercera al cepstrum real.

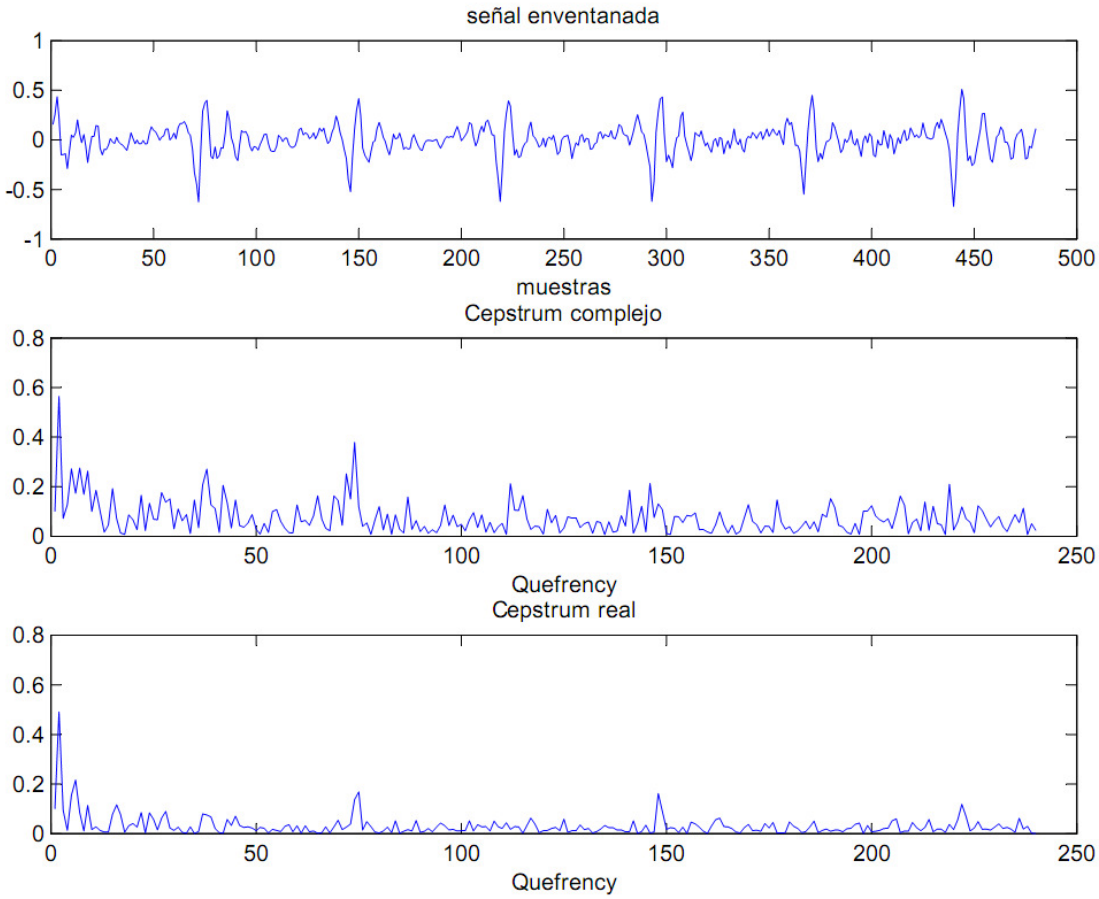


Figura 37: Ventana temporal de la señal de voz, su cepstrum complejo y el real

En ambas gráficas se puede observar la gran cantidad de información cerca del origen y más alejado los impulsos aislados que caracterizan el tono (pitch), separados del origen la a la distancia de pitch. Los coeficientes de orden bajo (cerca del origen) son debidos a las características del tracto vocal y proveen información sobre la envolvente. Las rápidas variaciones de la parte superior del cepstrum representan las características de la excitación de la trama de voz. El primer pico es el que da la información de pitch y es indicativo de la sonoridad de la trama. Los siguientes, que no tienen por que aparecer siempre, son los múltiplos de la fundamental y están a la misma distancia entre ellos.

En una secuencia sorda el cepstrum no muestra ese pico a la frecuencia fundamental, lo que nos indica que la trama no es sonora, no hay una excitación periódica y por lo tanto carece de frecuencia fundamental o pitch.

SONORIDAD Y FRECUENCIA DE PITCH

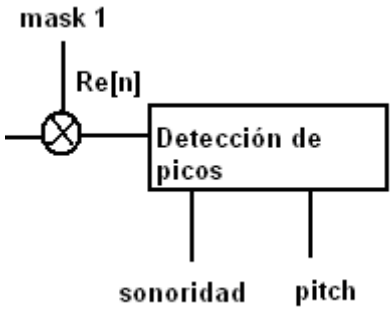


Figura 32: Detección de picos en el cepstrum real para la obtención del pitch y la sonoridad

Con la señal de excitación real se calcula el pitch de la señal buscando la posición del primer pico. Si lo hay, se califica la trama como sonora y se calcula la frecuencia de pitch. Si no lo hay, la trama se califica como sorda y no se calcula la frecuencia de pitch.

En la figura 31 se observa que la trama de voz sonora tiene un pico cerca del origen. Pico que no aparece en la trama sorda. Por lo tanto estas propiedades del cepstrum pueden usarse para determinar la sonoridad o no de una trama de voz. La distancia del pico al origen sirve para determinar la frecuencia de pitch.

La forma de estimar la sonoridad y el pitch es fácil. Se busca un pico en el cepstrum en la zona de la excitación, que está a unas 50 muestras más allá del origen (recordar que más o menos las primeras 30 corresponden al tracto vocal). Si el pico esta por encima de un umbral es pausible pensar que la trama será sonora. La posición de este sirve para determinar la frecuencia fundamental.

No obstante la ausencia de este pico no es un fuerte indicativo de que la señal no es sonora. La fuerza e incluso la existencia de este pico dependen de muchos factores, incluyendo la longitud de la ventana y la estructura de los formantes de la señal. La amplitud máxima de este pico es la unidad. Este caso sólo se da cuando hay dos periodos de la señal idénticos, situación que nunca se dará en la señal de voz.

Para voces masculinas, las cuales son más graves y por consiguiente con un periodo mayor, es necesaria una duración de la ventana mayor, sobre los 40 ms. Para voces con pitch más alto es posible utilizar una ventana más corta, lo que favorece que las características de la voz permanezcan cuasi-estacionarias en la trama.

También influye el tipo de ventana que se utilice. Por ejemplo es más fácil que una ventana rectangular capture dos periodos enteros antes que una ventana Hamming, que por naturaleza tiende a suavizar los bordes.

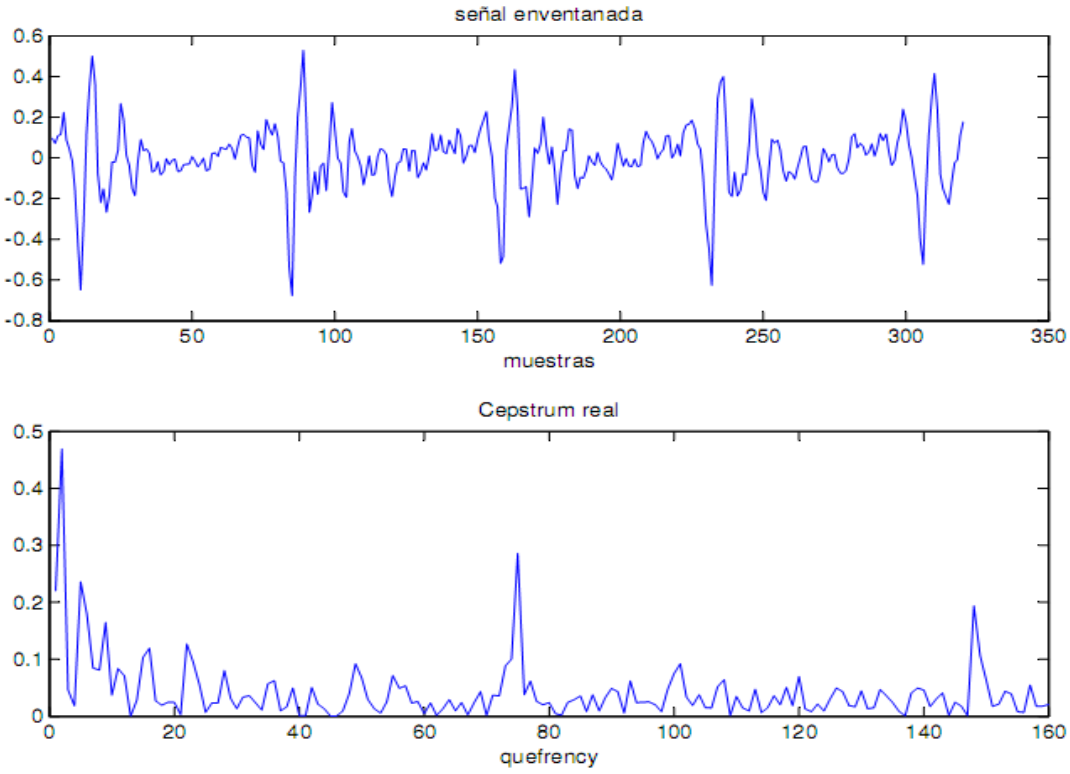


Figura 38: Señal sonora con ventana de 40 ms y su cepstrum real

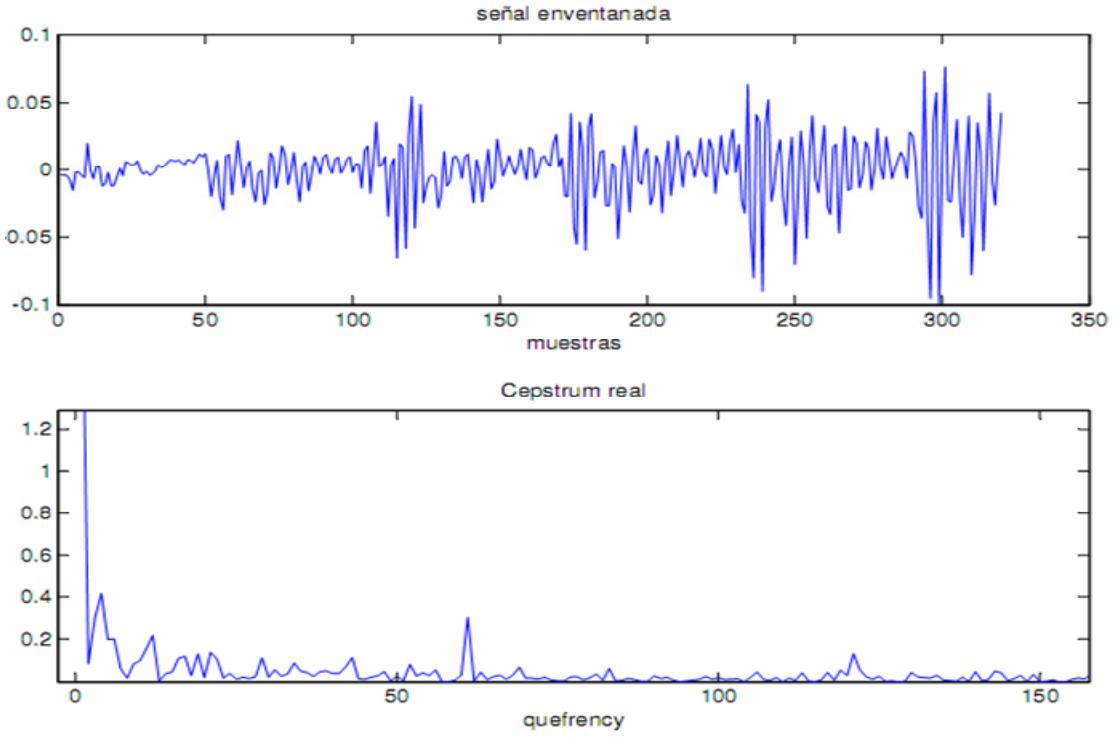


Figura 39: Señal sonora con ventana de 40 ms y su cepstrum real

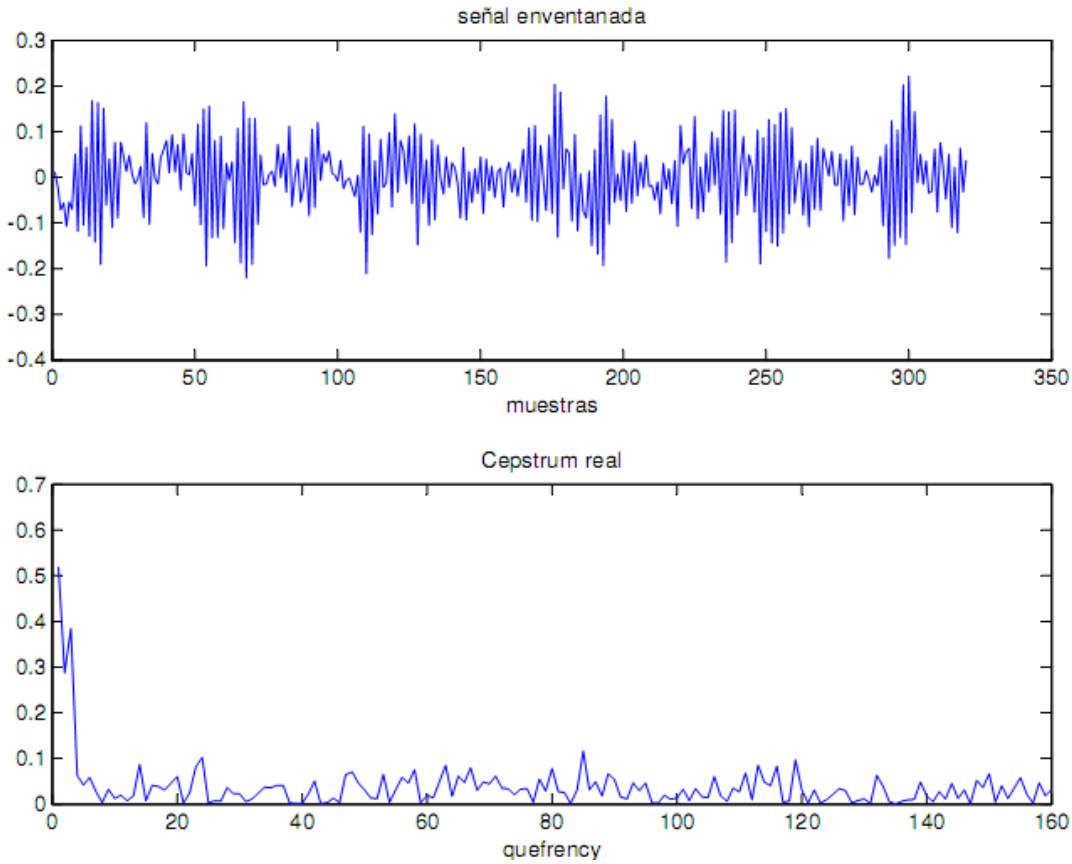


Figura 40: Señal sorda con ventana de 40 ms y su cepstrum real

RECONSTRUCCIÓN DE LA SEÑAL

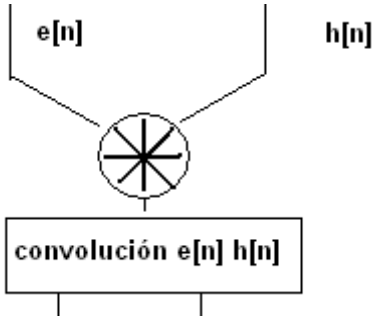


Figura 41: Convolución entre e[n] y h[n] de la señal enventanada

Para reconstruir la señal original es necesario realizar la convolución de la señal de excitación con el tracto vocal. Este proceso se realiza trama a trama, y después se concatenan todas las ventanas para obtener la señal decodificada. Esto permite evaluar y comparar los resultados trama a trama y con la señal entera.

MÉTODO DE EVALUACIÓN

Un método sencillo de evaluación de la señal decodificada es por comparación con la señal original, pero este método, aunque sencillo, es un método subjetivo, por lo que necesitamos evaluar la señal de manera que obtengamos un valor que nos permita comparar los resultados de un vocoder y otro.

Como método de evaluación usaremos la relación existente entre las 2 señales atendiéndonos al ratio señal-señal en dB (RSS dB) , que nos da un valor con el cual poder hacer comparaciones objetivas.

$$RSS\ dB = 10\log\left|\frac{\Sigma_n^N x[n]^2}{\Sigma_n^N (x[n]-\hat{x}[n])^2}\right|$$

- Análisis de la tramas convolucionadas:

Figura 41: Obtención del SRR ventana a ventana

Se realiza el análisis SRR de cada trama y se dibujan en un gráfico en tiempo. Así se puede apreciar que evolución sigue la calidad de la señal decodificada obtenida trama a trama.

- Análisis de la señal decodificada:

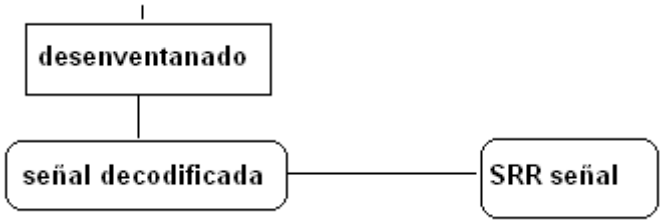


Figura 42: Obtención de la señal decodificada y su SRR

En este paso se concatenan las tramas y se obtiene la señal decodificada. Se compara con la original y se realiza la evaluación de SRR para obtener un parámetro de calidad objetivo.

RESULTADO EXPERIMENTAL

INTRODUCCIÓN

Una vez explicadas las técnicas y las herramientas utilizadas se procede al análisis de los resultados obtenidos. Además se procederá al estudio comparativo entre el vocoder lpc y el homomórfico.

Se pretende comparar y analizar:

- Calidad subjetiva
- El ratio señal-señal (SRR) de la señal decodificada
- Análisis trama a trama de SRR
- Cantidad de ancho de banda requerido en bits
- Impacto que produce el preprocesado de la señal
- Capacidad de detección de sonoro/sordo y cálculo de pitch

El método utilizado será el siguiente: Se han grabado 10 frases cortas en las mismas condiciones por dos locutores, un varón y una mujer. Se aplicará a cada frase ambos vocoders y analizaremos 3 de las frases para ver el comportamiento de cada vocoder.

Además se elaborará una lista con los resultados de SRR para la señal completa para poder hacer una valoración global.

Para poder hacer las comparaciones se usará el mismo tipo de ventana en ambos vocoders, esto es una ventana rectangular. Para la comparativa de las 20 señales la duración de la ventana se establecerá en 30 ms para ambos vocoders.

SEÑALES UTILIZADAS EN EL ESTUDIO COMPARATIVO

Como se ha comentado son 20 señales. 10 frases cortas grabadas por una voz masculina y por una femenina.

En ellas hay una mezcla de frases con una sonoridad más o menos marcada, para que sea una batería de pruebas heterogénea. Estas son las señales:

1. 'a e i o u'
2. Diccionario enciclopédico
3. Sentimientos diferentes
4. Estroboscópico
5. Ingeniería técnica de telecomunicaciones
6. Mañana soleada
7. Producción digital de voz y audio
8. Sentimiento diferentes
9. Universidad publica de Navarra
10. Vocoder homomórfico

De las 10 señales se han elegido 3 que serán las que se analicen con detalle.

PRIMERA SEÑAL

"A e i o u"

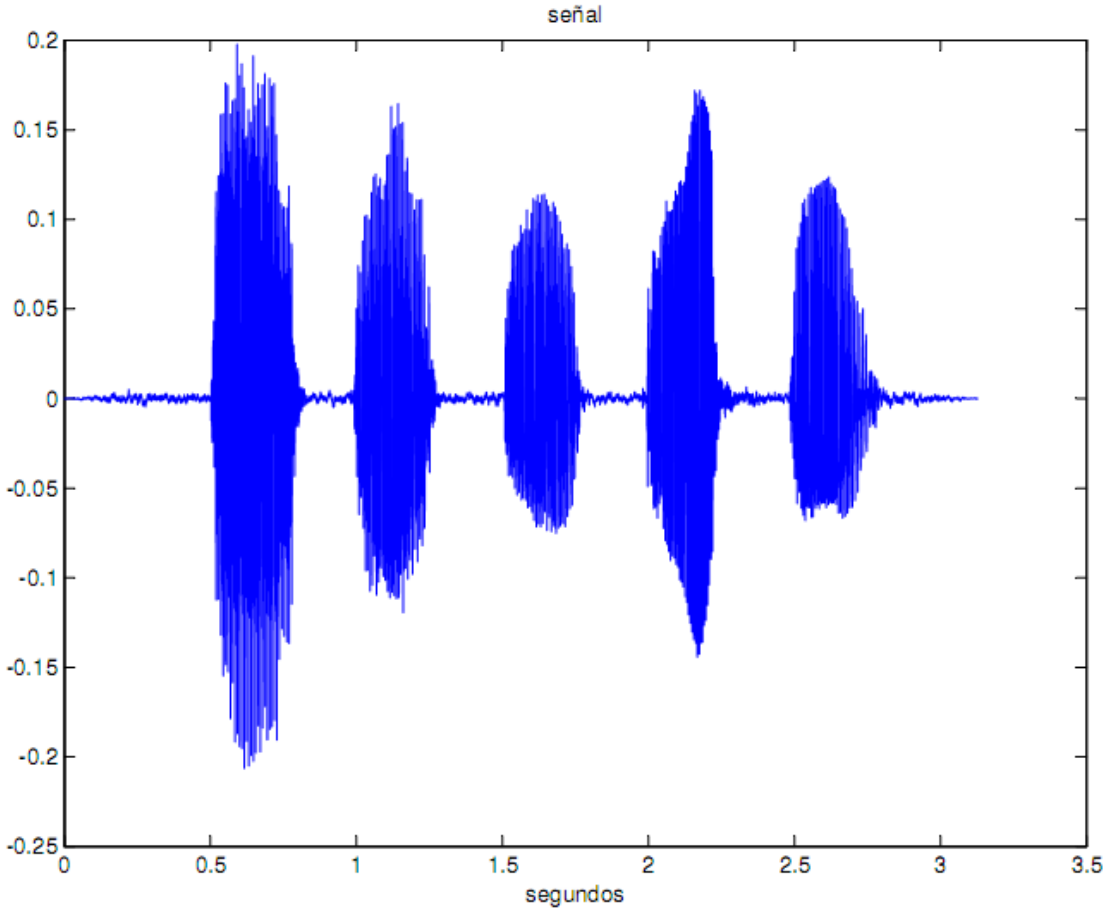


Figura 43: Señal 'a e i o u'

El análisis comenzará con una señal muy simple, son las cinco vocales. Todas ellas señales sonoras. Se puede observar la envolvente de cada una de ellas. El ataque pronunciado, y un decaimiento corto, sobre todo la 'o'.

SEGUNDA SEÑAL

"Vocoder homomórfico"

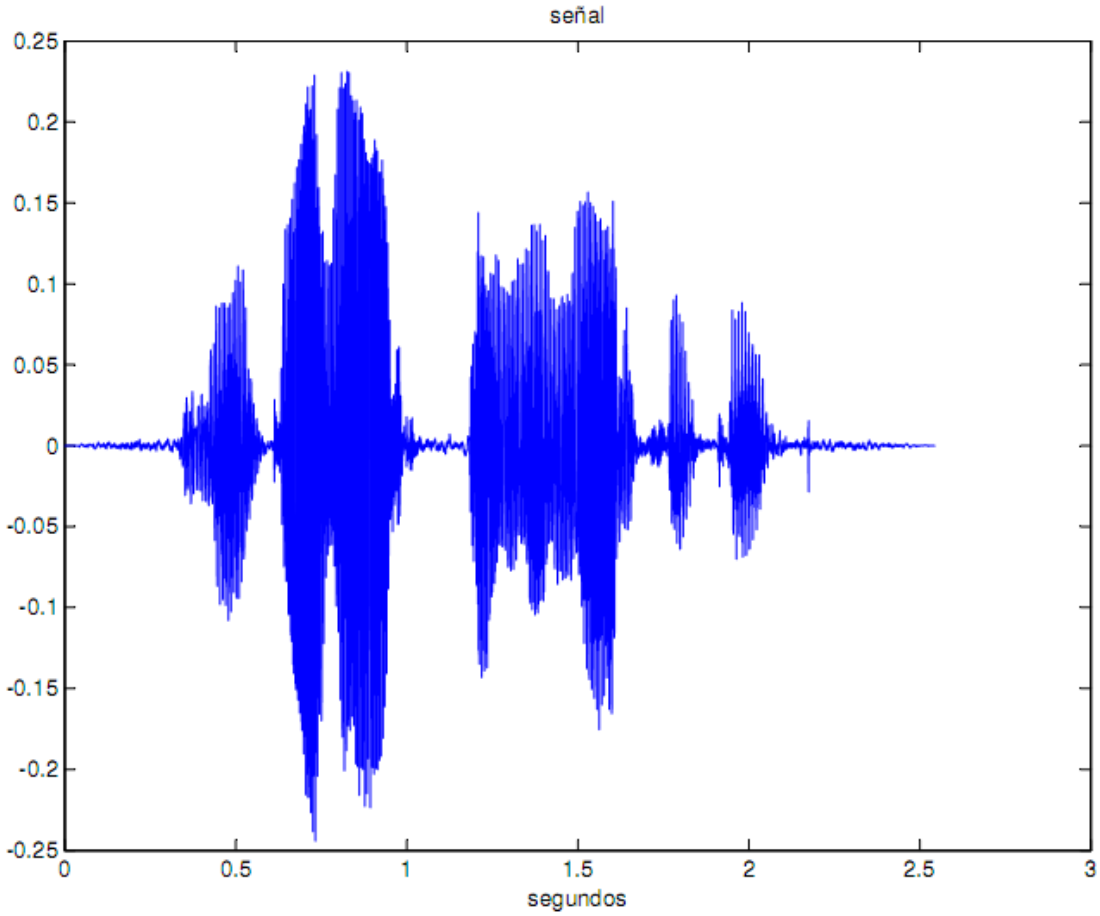


Figura 44: 'vocoder homomórfico'

En esta gráfica también se puede ver la separación de las silabas, y sobe todo como antes de una consonante oclusiva como es la 'c' (segundo 0'7), o fricativo como la 'f' (segundo 1'8), hay un pequeño espacio por el cambio que hay que hacer en el modo de articulación de la lengua y los labios, así como el ataque pronunciado que tienen.

También es posible distinguir la entonación, pues las sílabas acentuadas tienen mayor energía.

TERCERA SEÑAL

"Mañana soleada"

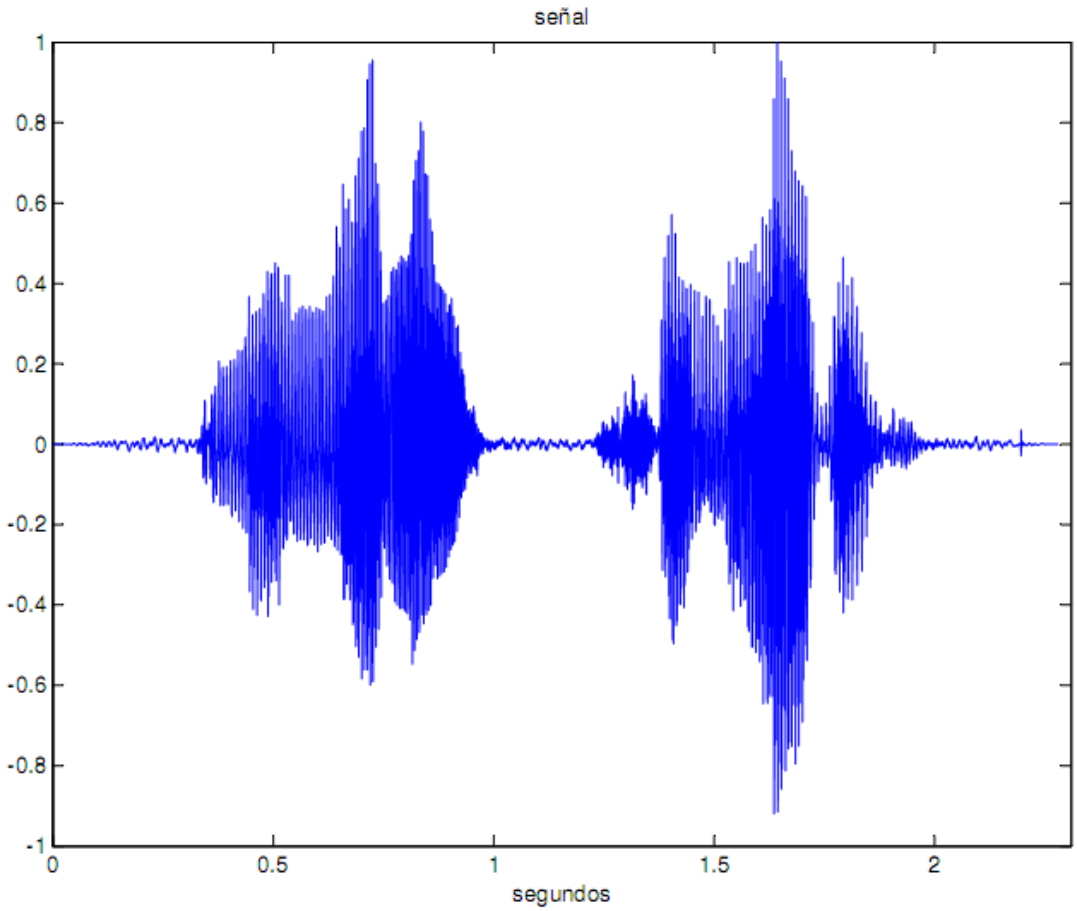


Figura 45: Señal mañana soleada.

La última señal es una frase en la que se mezclan vocales con bastantes consonantes sordas (oclusivas 't' y fricativas 's').

En la gráfica de tiempo se diferencian fácilmente las dos palabras y también se puede apreciar como el principio de soleada al ser sonido sordo 's' tiene menor energía y no tiene un ataque pronunciado, como por ejemplo la vocal que viene después 'o'.

CALIDAD SUBJETIVA

En este primer análisis se han escuchado tanto las señales originales como las decodificadas, y se ha realizado una primera valoración subjetiva.

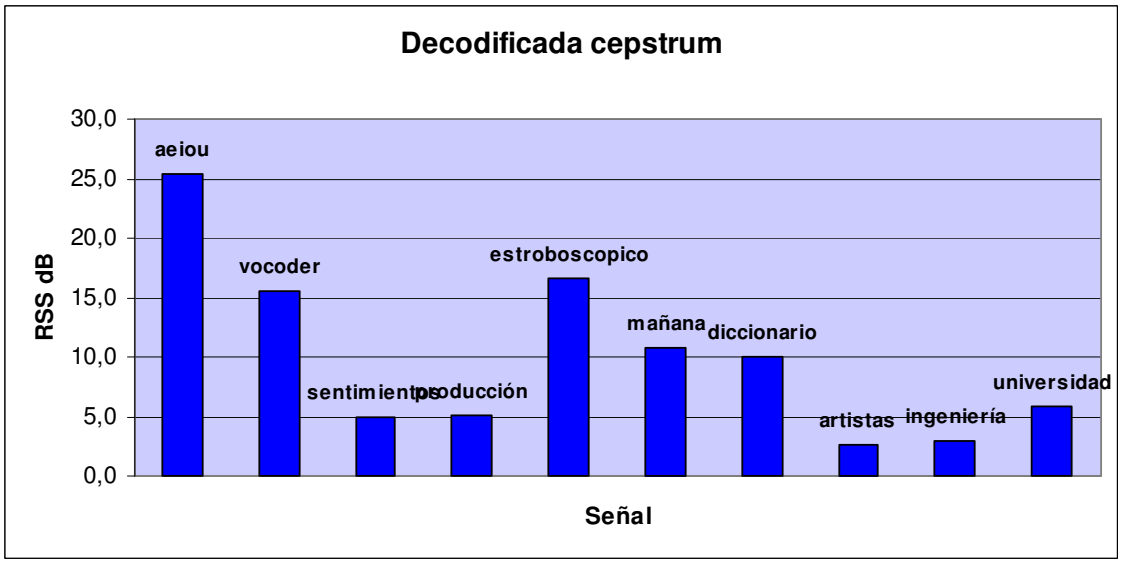
Sin ninguna duda se puede decir que la balanza se decanta a favor del decodificador homomórfico. Las pruebas arrojan unos resultados objetivos que no admiten duda alguna. Las señales procesadas con el Vocoder lpc aunque se entienden, suena latosas, como si las pronunciase una máquina y además la señal tiene una distorsión que hace complicado la identificación del locutor. Sin embargo la reproducción de las tramas sonoras y sordas se mantiene en líneas generales como en la original, exceptuando algunas tramas. Se puede decir que las frases son medianamente inteligibles.

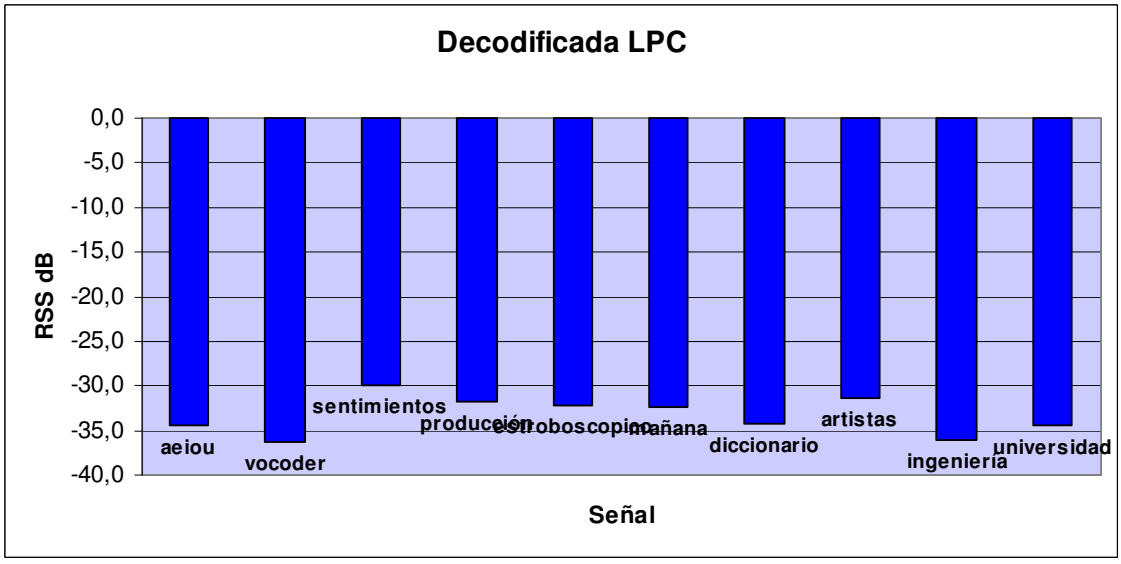
El Vocoder homomórfico es muy superior en cuanto a calidad se refiere. Las frases decodificadas se mantienen prácticamente fieles a la original. La inteligibilidad es muy buena, y se puede identificar fácilmente al locutor. Solamente se escucha un ruido que suena como un crack cíclico, pero que apenas molesta.

ANÁLISIS DEL SRR DE LAS SEÑALES DECODIFICADAS

VOZ MASCULINA

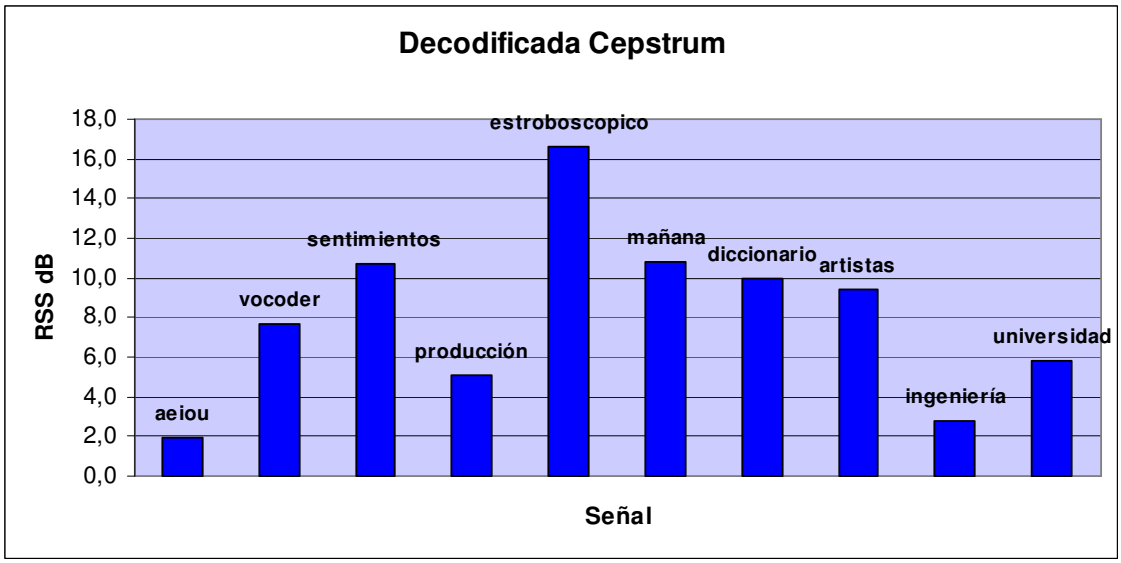
Señal	Decodificada cepstrum	Decodificada LPC
aeiou	25,4	-34,5
vocoder	15,6	-36,2
sentimientos	5,0	-30,0
producción	5,1	-31,9
estroboscópico	16,6	-32,3
mañana	10,8	-32,4
diccionario	9,9	-34,2
artistas	2,6	-31,4

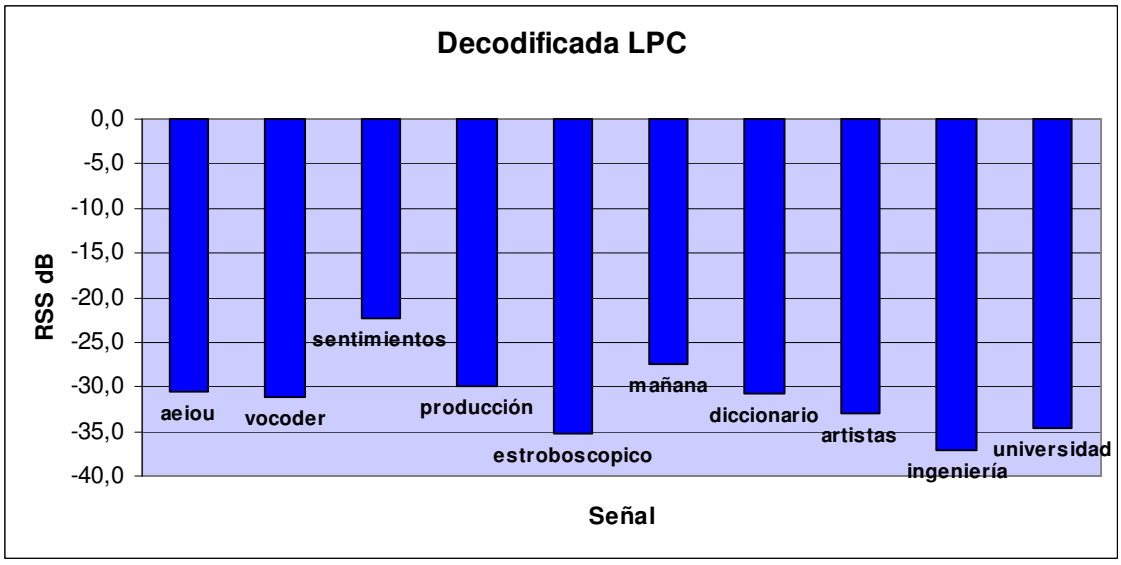




VOZ FEMENINA

Señal	Decodificada Cepstrum	Decodificada LPC
aeiou	1,9	-30,5
vocoder	7,7	-31,1
sentimientos	10,7	-22,4
producción	5,1	-30,0
estroboscópico	16,6	-35,2
mañana	10,8	-27,6
diccionario	9,9	-30,8
artistas	9,4	-33,1
ingeniería	2,8	-37,2
universidad	5,9	-34,6





Los números dejan claro que el vocoder que mejor calidad obtiene es el homomórfico, pues mantiene en todas las señales un valor de SRR positivo.

Como era de esperar los valores SRR del vocoder lpc son negativos en todas las señales. No obstante es curioso como los resultados del vocoder lpc, aunque malos, mantienen unos valores constantes en todas las señales analizadas, hecho que no ocurre en el vocoder homomórfico, donde vemos una mayor variabilidad entre los resultados obtenidos.

Tras ver los resultados de SRR para toda la señal se va a proceder al análisis del valor SRR para las 3 señales trama a trama.

ANÁLISIS SRR TRAMA A TRAMA

En este apartado se analiza el comportamiento trama a trama. En los siguientes gráficos se muestran las 3 señales y el valor SRR de cada trama con el vocoder homomórfico y con el lpc

PRIMERA SEÑAL

Decodificador lpc

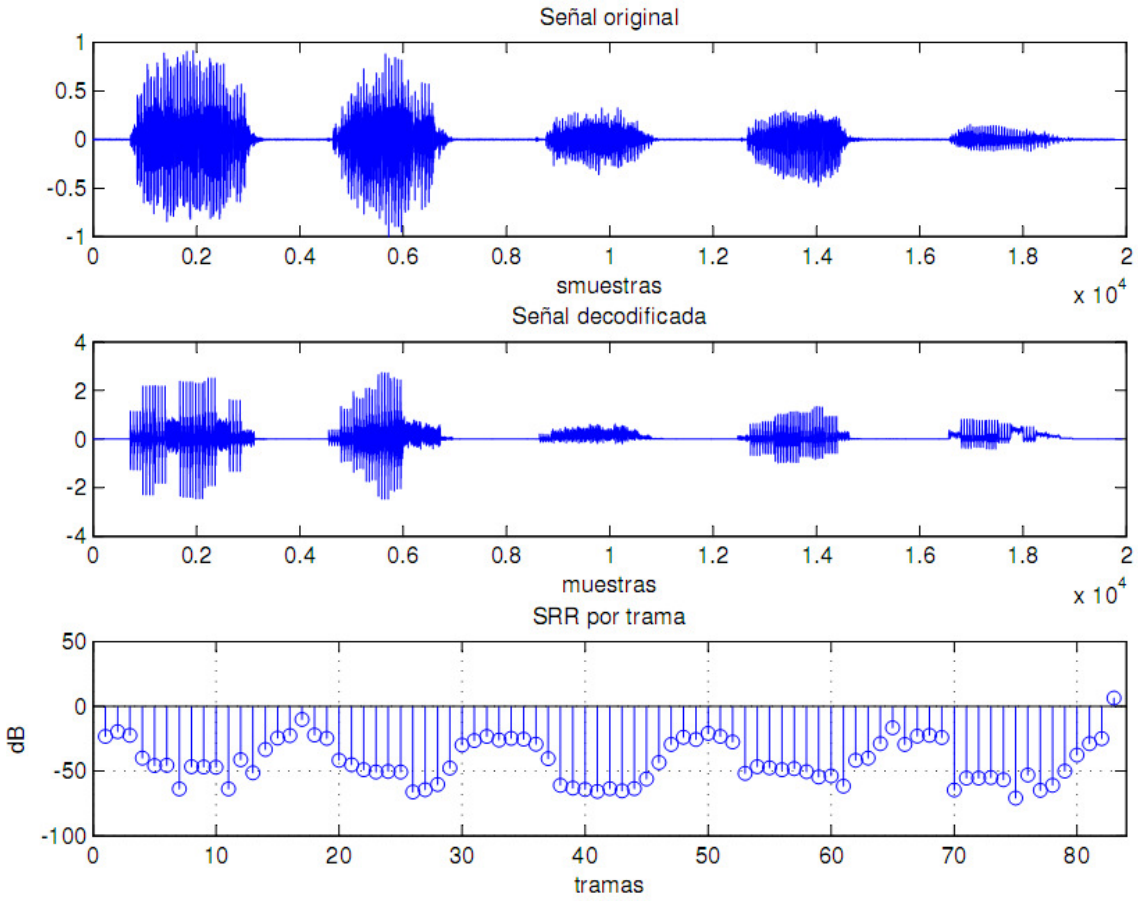


Figura 46: 1ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

Decodificador homomórfico

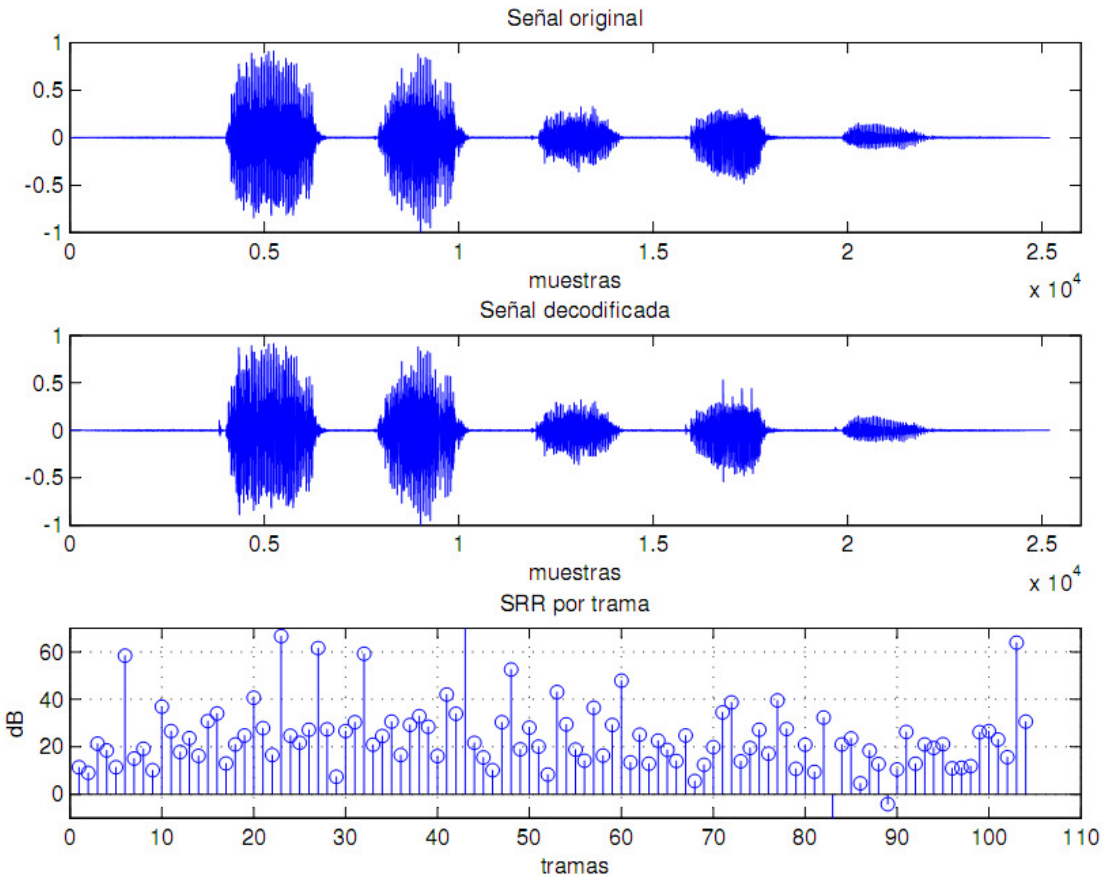


Figura 47: 1ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

SEGUNDA SEÑAL

Decodificador lpc

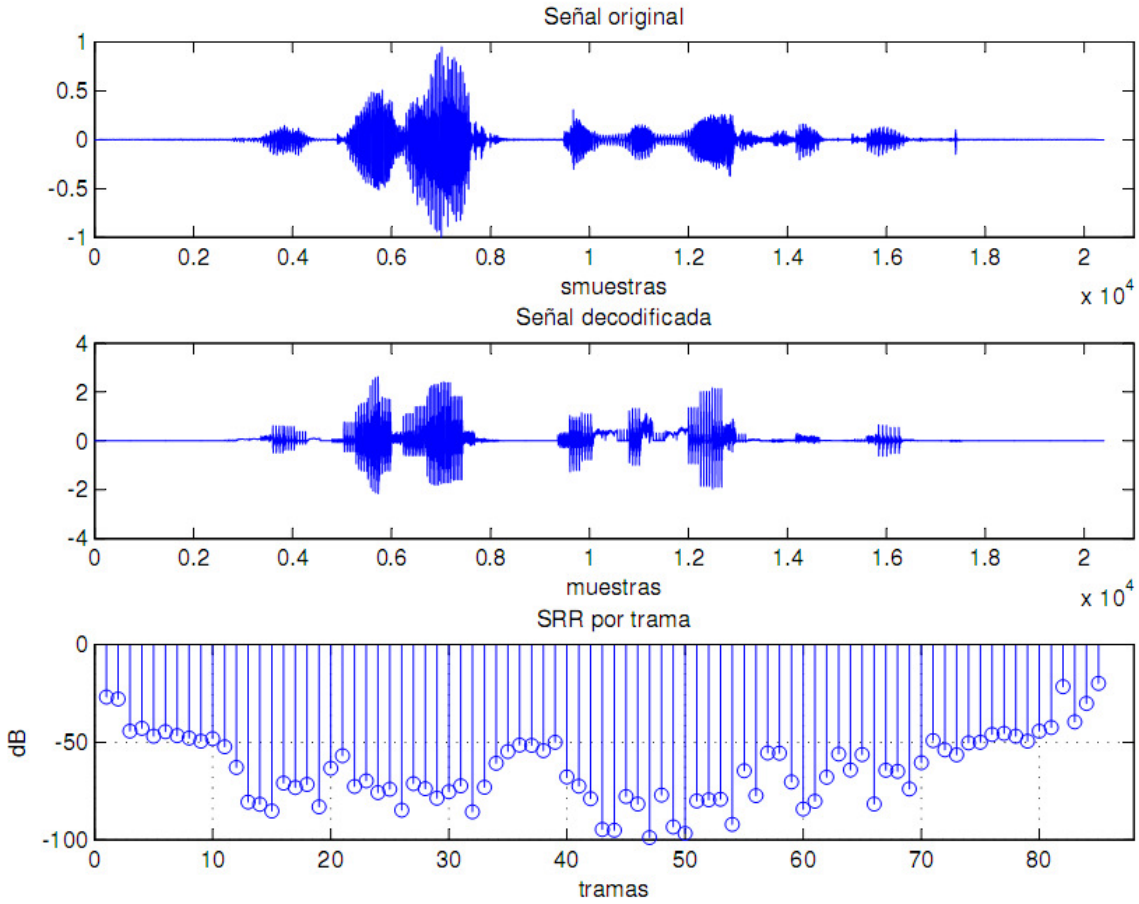


Figura 48: 2ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

Decodificador homomórfico

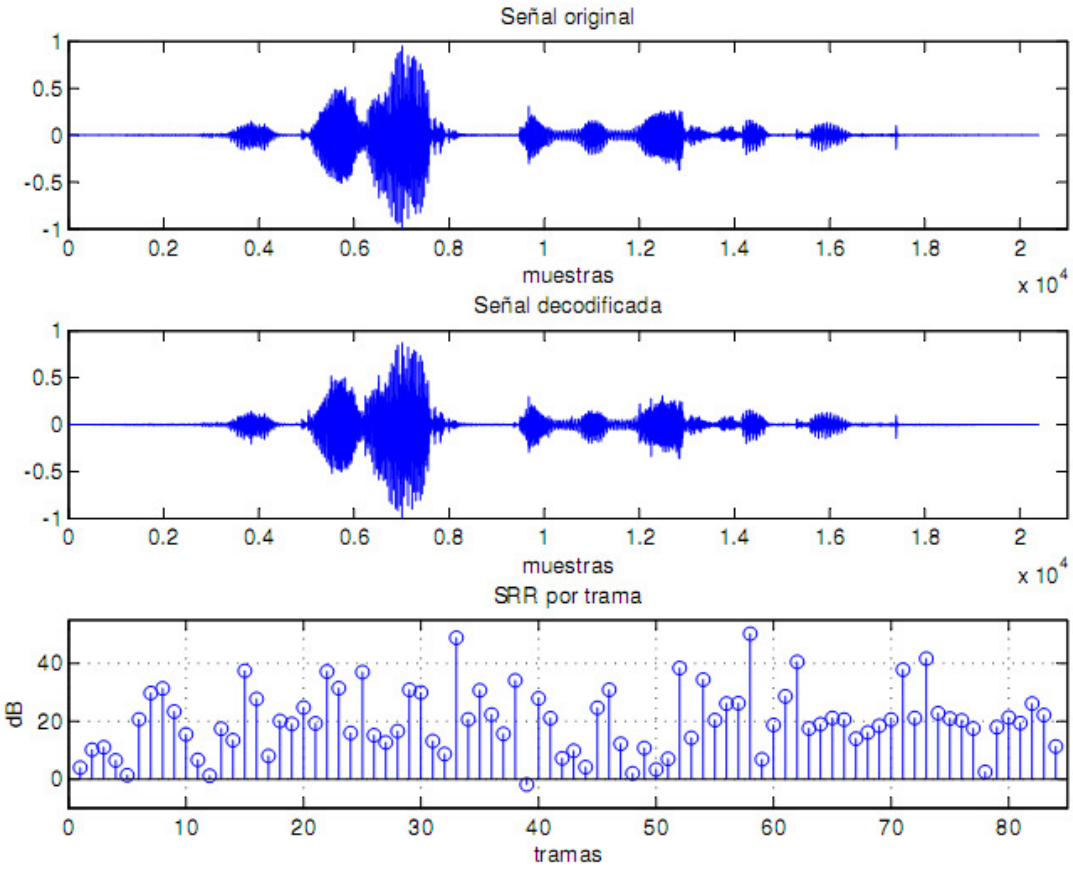


Figura 49: 2ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

TERCERA SEÑAL

Decodificador lpc

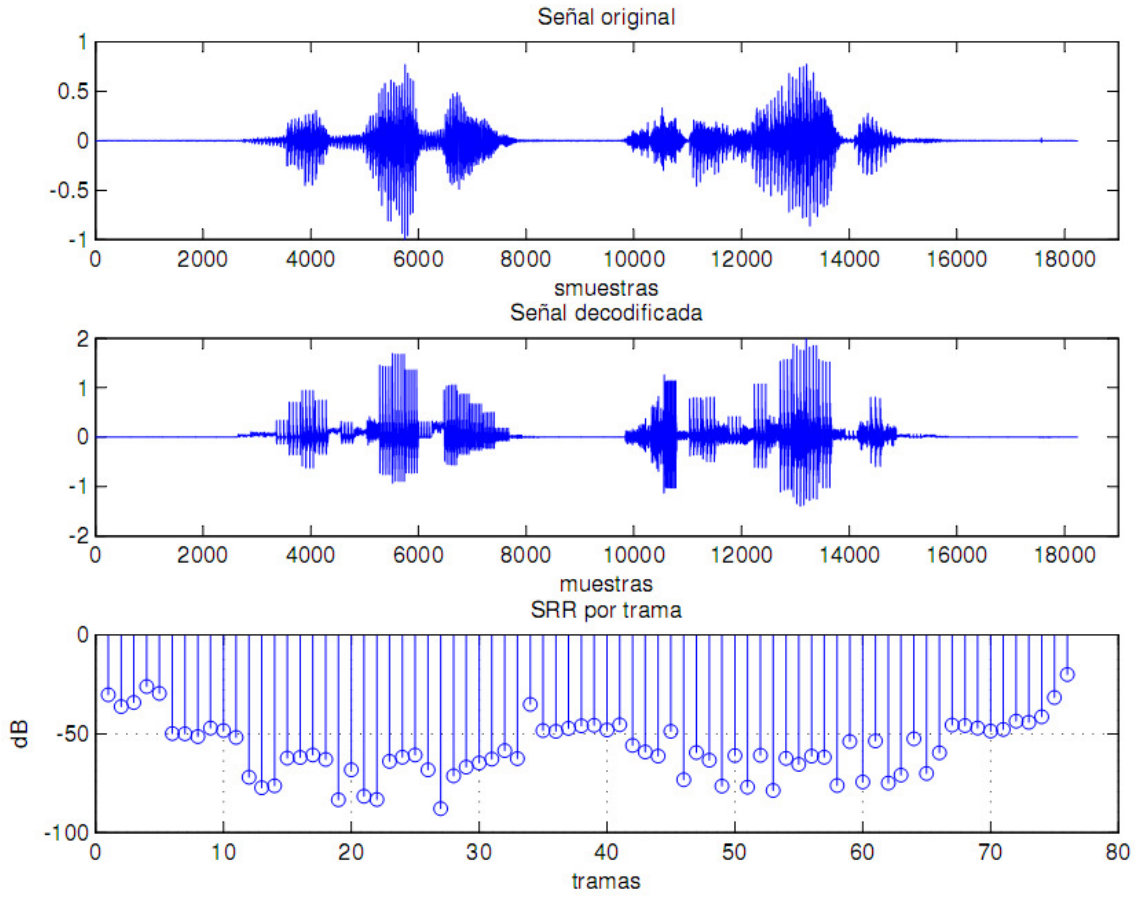


Figura 50: 3ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

Decodificador homomórfico

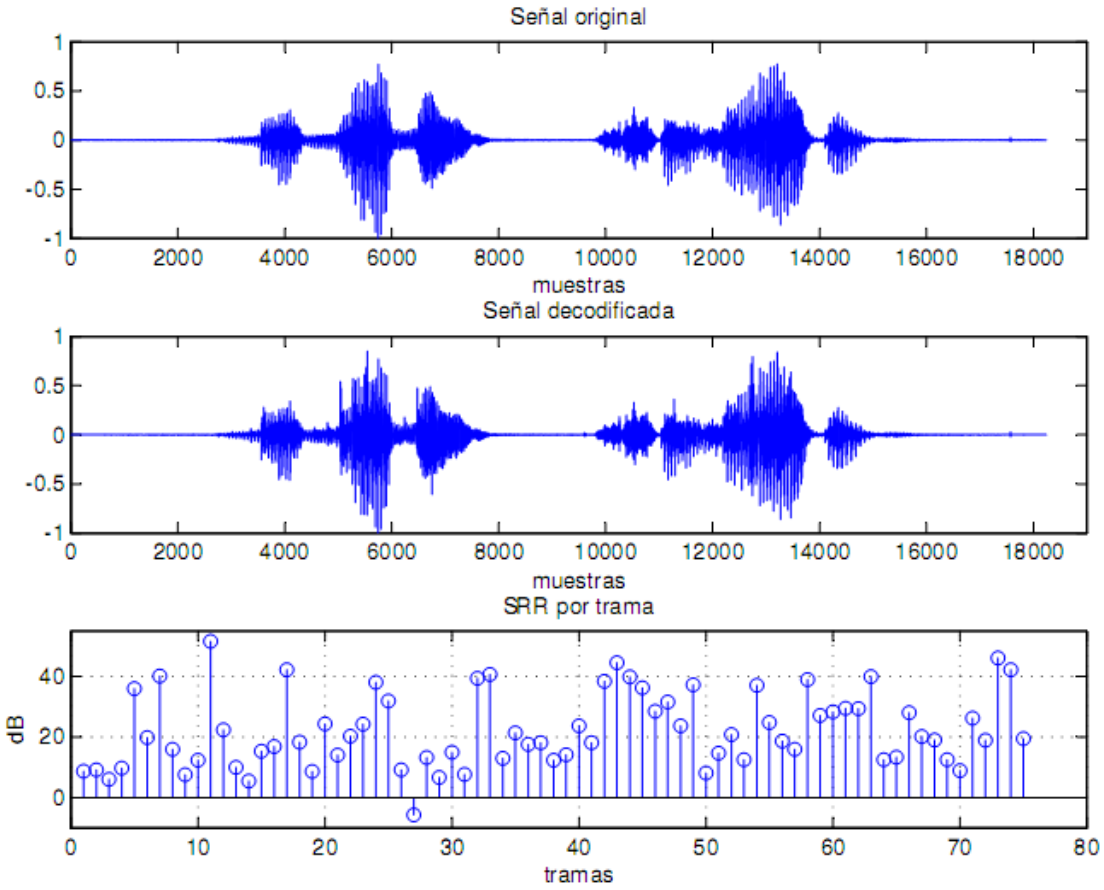


Figura 51: 3ª Señal en tiempo, la señal decodificada y el valor SRR de cada trama

En líneas generales también se ve la superioridad del vocoder homomórfico, puesto que analizando trama a trama, el SRR es siempre positivo, exceptuando una o dos tramas, el vocoder lpc por el contrario en ninguna trama llega a tener un resultado SRR positivo.

Sin embargo hay una aspecto en el que el vocoder lpc es mejor, y es que mantiene el SRR constante trama a trama, sin grandes saltos en su valor. El vocoder homomórfico aunque obtiene mejores resultados siempre, no es muy lineal trama a trama. Vemos tramas en los que el valor SRR obtiene máximos que no se vuelven a repetir en toda la señal, así como valores negativos que sólo se dan en una o dos tramas.

Lo que resulta fácil ver es que la señal decodificada con el vocoder homomórfico es muy parecida a la original, mientras que la decodificada del vocoder lpc, guarda menos parecido, y es normal pues la calidad es mucho peor. Al fin y al cabo, la señal de excitación del decodificador lpc se compone se una serie de trenes de pulsos glotales para las tramas consideradas sonoras y ruido aleatorio para las tramas consideradas sordas. Esto resulta muy fácil de ver en las gráficas. Se diferencian perfectamente

los tramos que han sido generados con una señal de excitación compuesta por un tren de pulsos y los que han sido generados con ruido.

Sin embargo, aunque la señal decodificada con el vocoder homomórfico parece idéntica, no es así. Para verlo resulta más útil el análisis del error cometido en la señal decodificada, en la que podemos ver que diferencias existen entre la señal original y la decodificada.

ANÁLISIS DEL ERROR

PRIMERA SEÑAL

Decodificador lpc

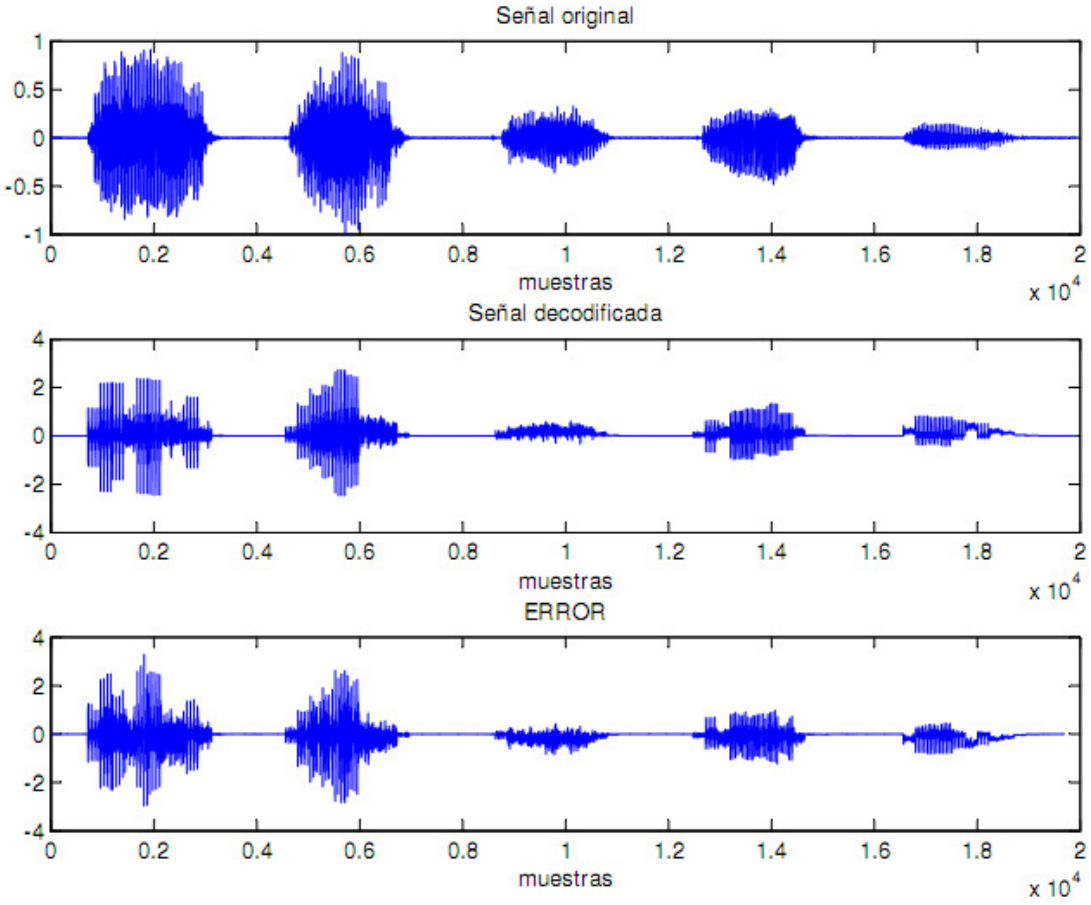


Figura 52: 1ª Señal en tiempo, la señal decodificada y el error cometido

Decodificador homomórfico

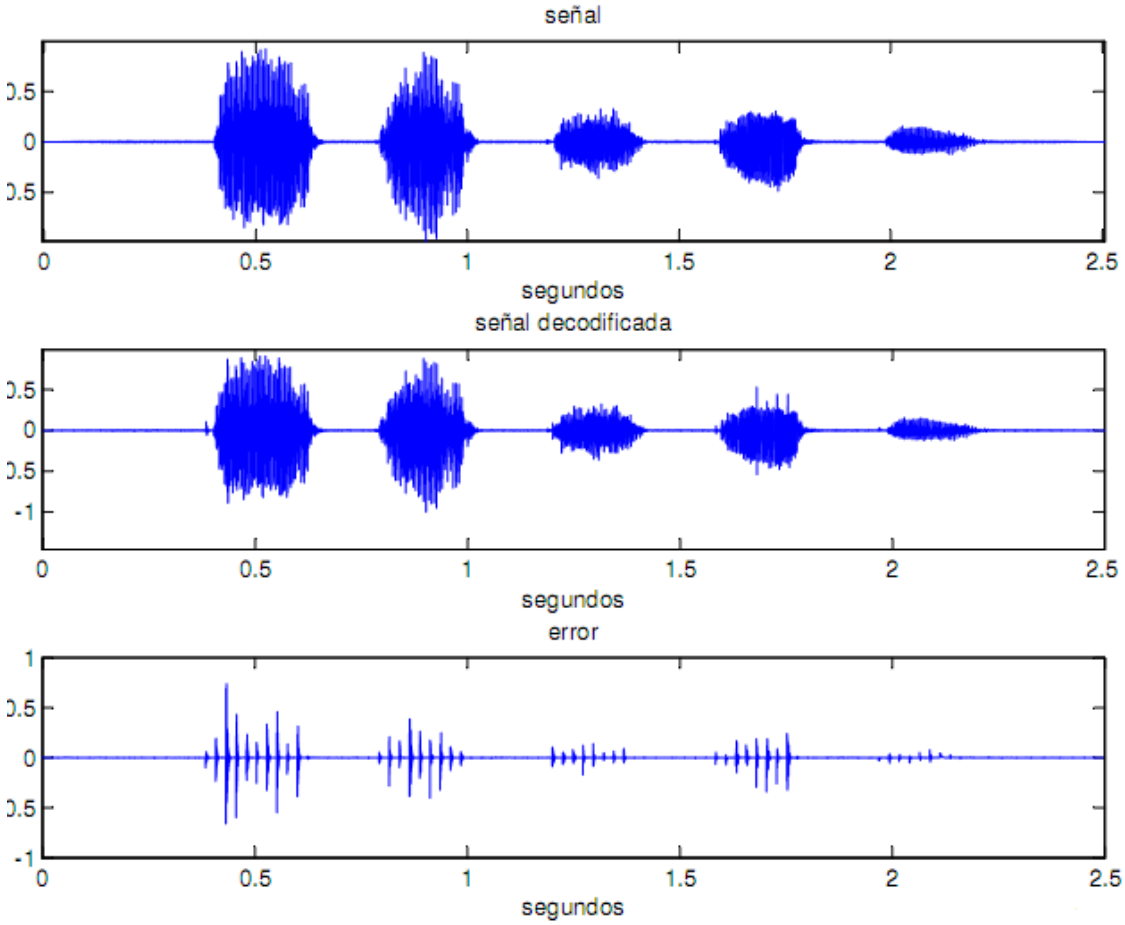


Figura 53: 1ª Señal en tiempo, la señal decodificada y el error cometido

SEGUNDA SEÑAL

Decodificador lpc

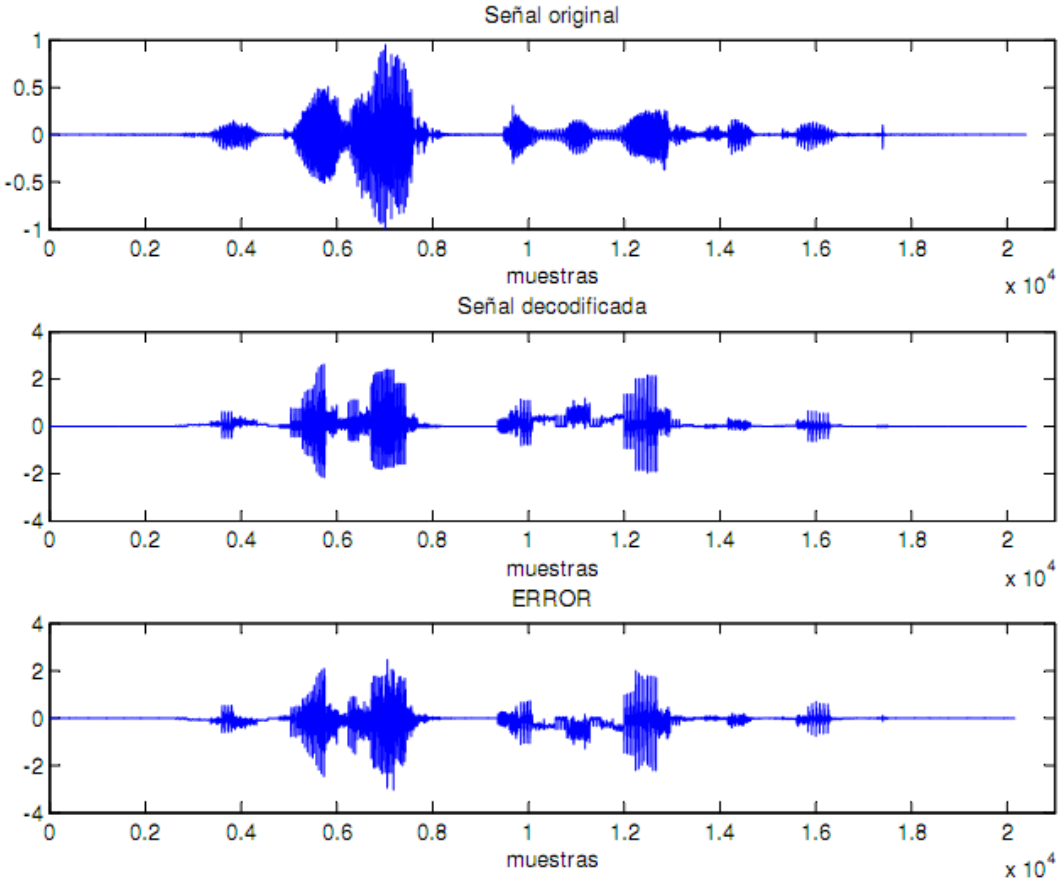


Figura 54: 2ª Señal en tiempo, la señal decodificada y el error cometido

Decodificador homomórfico

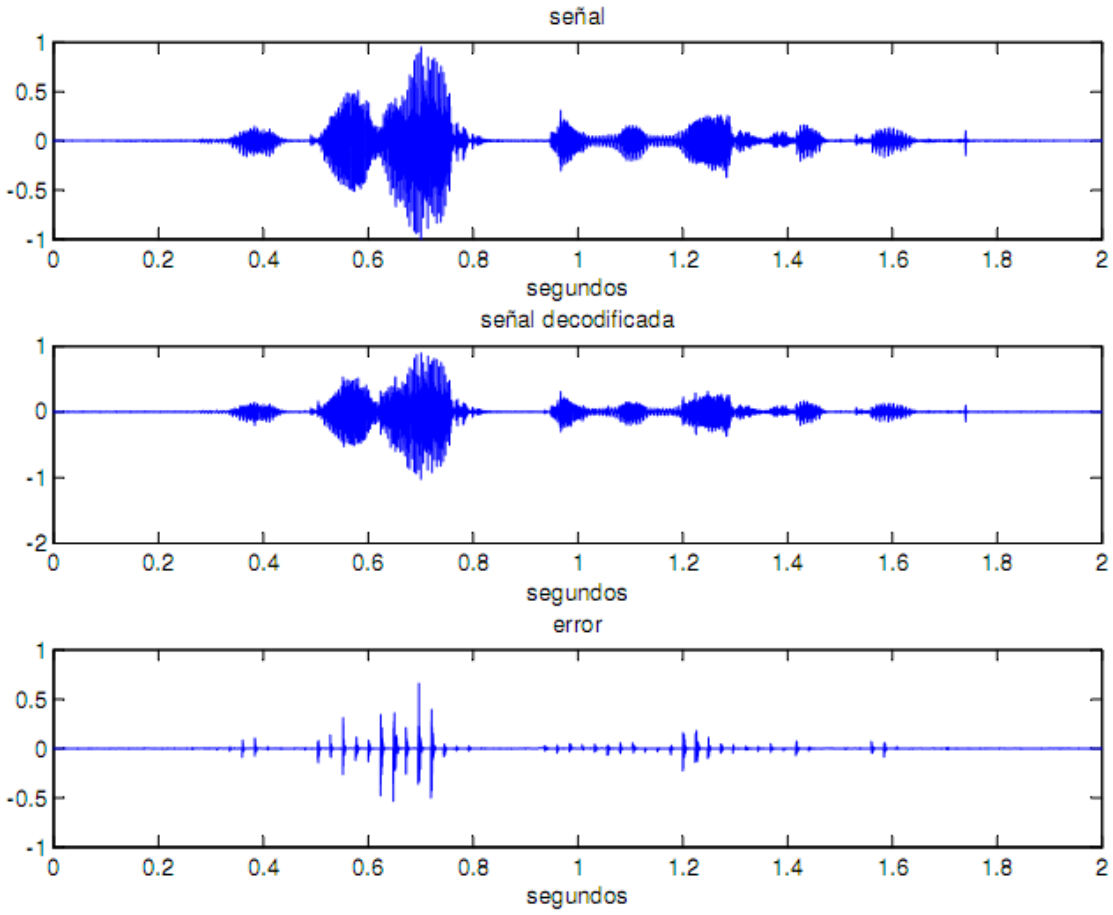


Figura 55: 2ª Señal en tiempo, la señal decodificada y el error cometido

TERCERA SEÑAL

Decodificador lpc

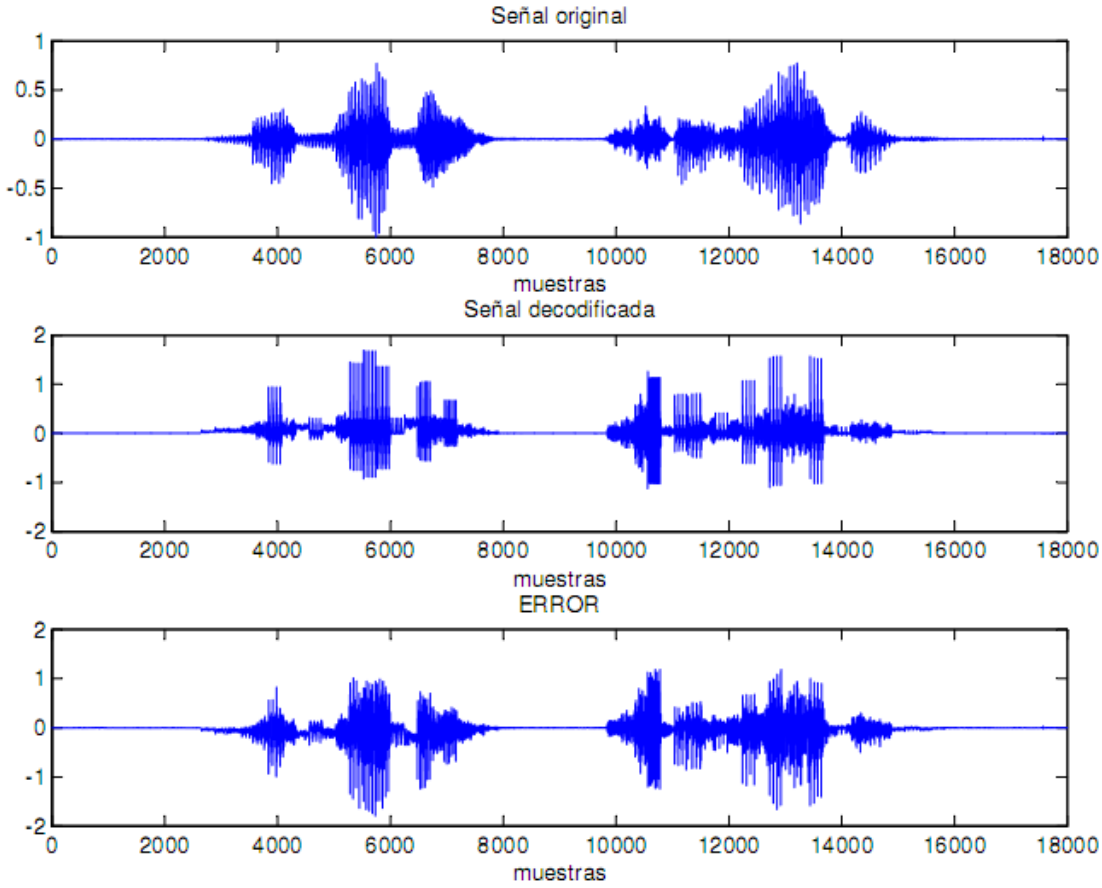


Figura 56: 3ª Señal en tiempo, la señal decodificada y el error cometido

Decodificador homomórfico

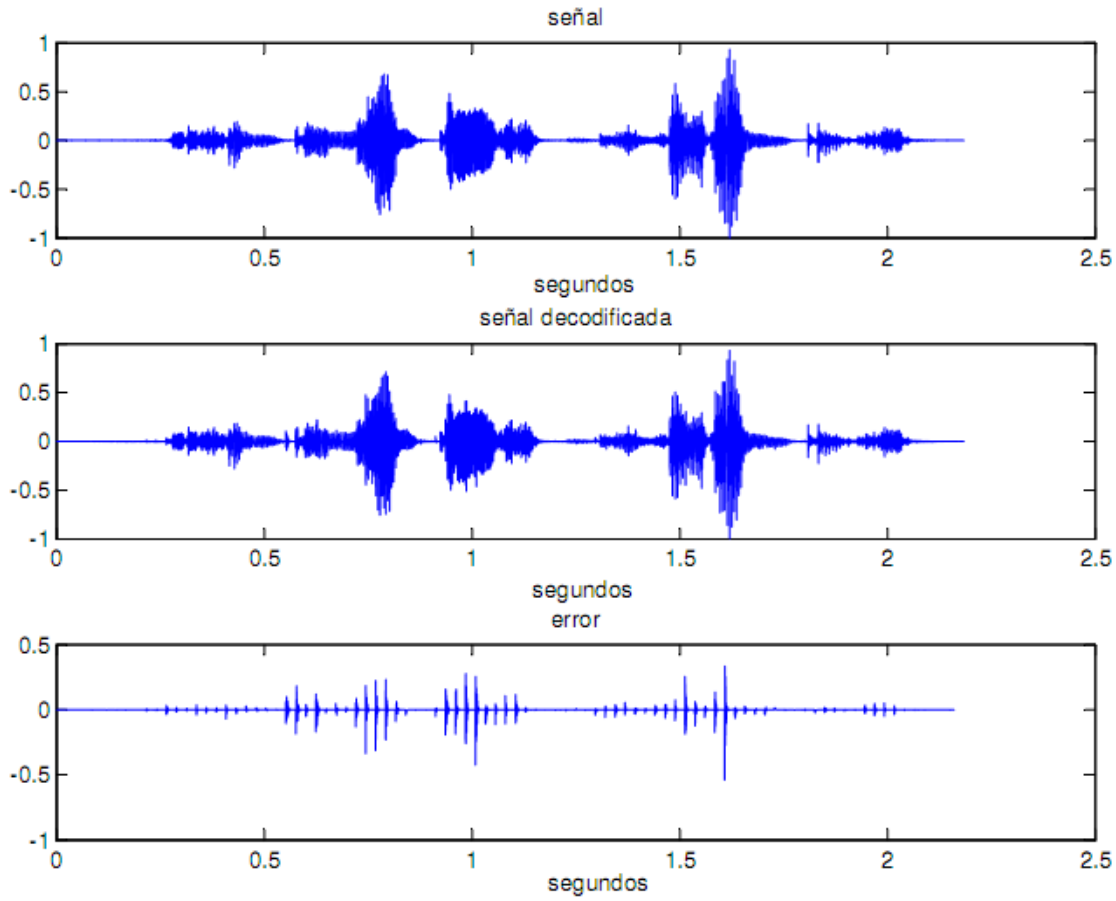


Figura 57: 3ª Señal en tiempo, la señal decodificada y el error cometido

En estas gráficas se puede ver que aunque la señal decodificada con el vocoder homomórfico es más fiel a la original tiene unas diferencias sutiles pero visibles. Se pueden ver unos picos en el error que mantienen una periodicidad. Esto es ese crack cíclico que se escucha en la señal decodificada.

Es muy difícil una separación total en el dominio cepstral de la señal de excitación y el tracto vocal, por que ambas se acaban mezclando en el punto de separación como se vio en las gráficas de las máscaras. Por eso la señal de excitación no es exactamente igual a la original, y por lo tanto se acaba cometiendo un pequeño error.

En la señal decodificada con el vocoder lpc se ve que la señal de error es parecida a la señal original, sin embargo lo que aparece en el error son los detalles rápidos en tiempo que no presenta la señal decodificada. En todo caso la señal decodificada es de baja calidad y por lo tanto es normal que el error sea bastante mayor que el del vocoder homomórfico.

La señal decodificada usando el vocoder lpc es de baja calidad incluso cuando apenas hay señal, es decir en los silencios. En la siguiente gráfica se muestra la diferencia existente entre la señal original y la decodificada en una zona donde apenas existe señal alguna (Figura 55).

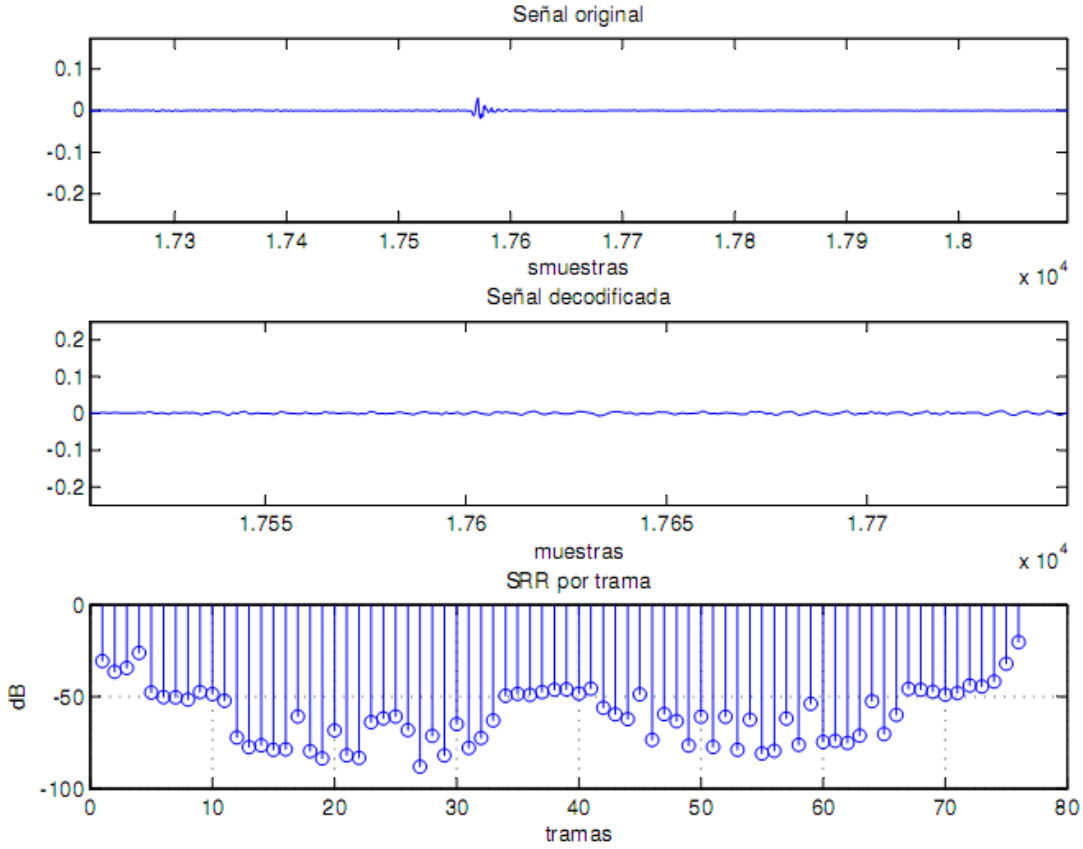


Figura 58: Diferencias entre la señal original y la decodificada con el vocoder lpcy el valor de RSS de cada trama

Se observa como la gráfica de la señal original se mantiene plana excepto un pequeño pulso.

La señal decodificada sin embargo no dibuja el pulso correctamente y además introduce más ruido en forma de ondulaciones de baja amplitud donde la señal original no lo tiene. Esta trama corresponde con la nº 73 en la gráfica del SRR (17600/240 ≈ 73).

En el siguiente caso, se aprecia como en el límite que separa una trama de otra, cada una de ellas se reconstruye con señales de excitación totalmente diferentes. La primera trama considerada sorda y la siguiente sonora, y así, se forman unos cambios muy notables en la señal decodificada (Figura 56).

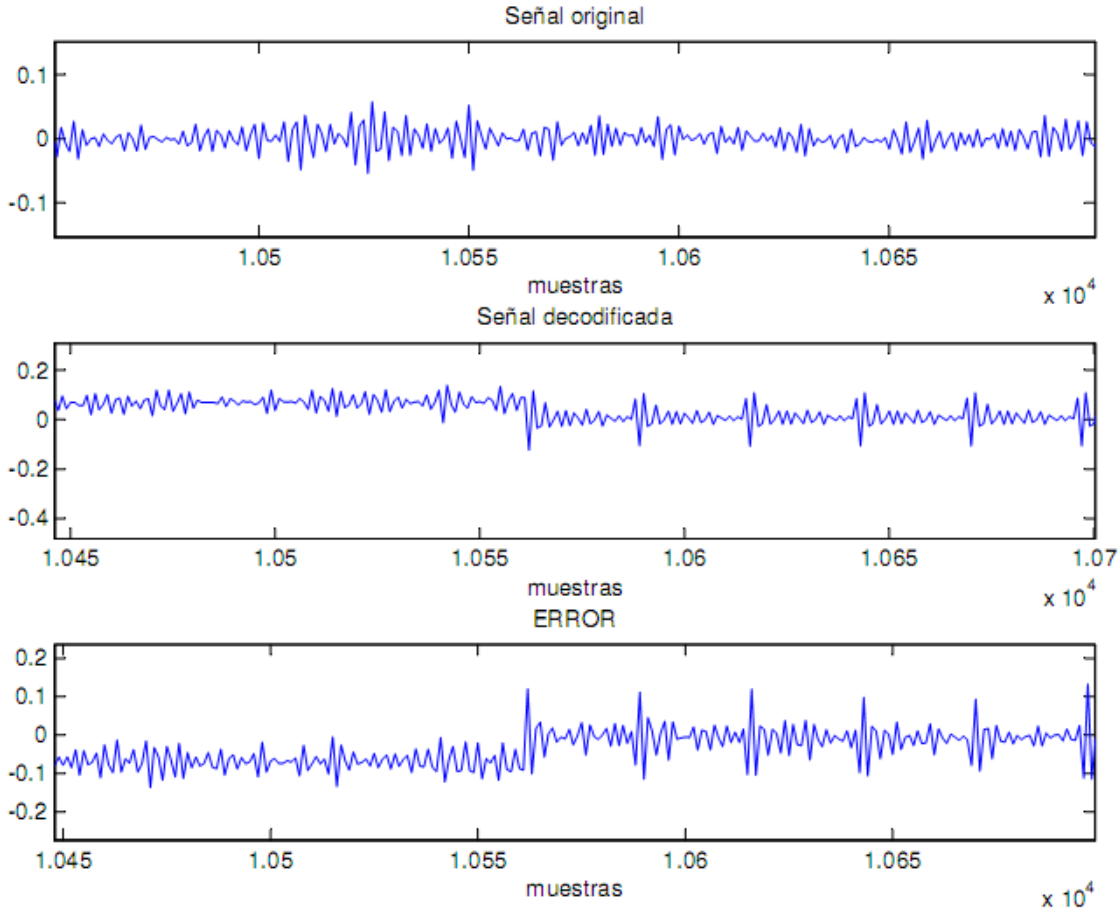


Figura 59: Diferencias entre la señal original y la decodificada con el vocoder lp, y el error de esta.

En este otro ejemplo se ve como la señal original no tiene una periodicidad muy marcada, de hecho pocas tramas guardan una periodicidad total. Como el vocoder lpc sólo decodifica la señal usando un tren de pulsos o ruido rosa, este el resultado de una mala caracterización de la señal de excitación (Figura 57).

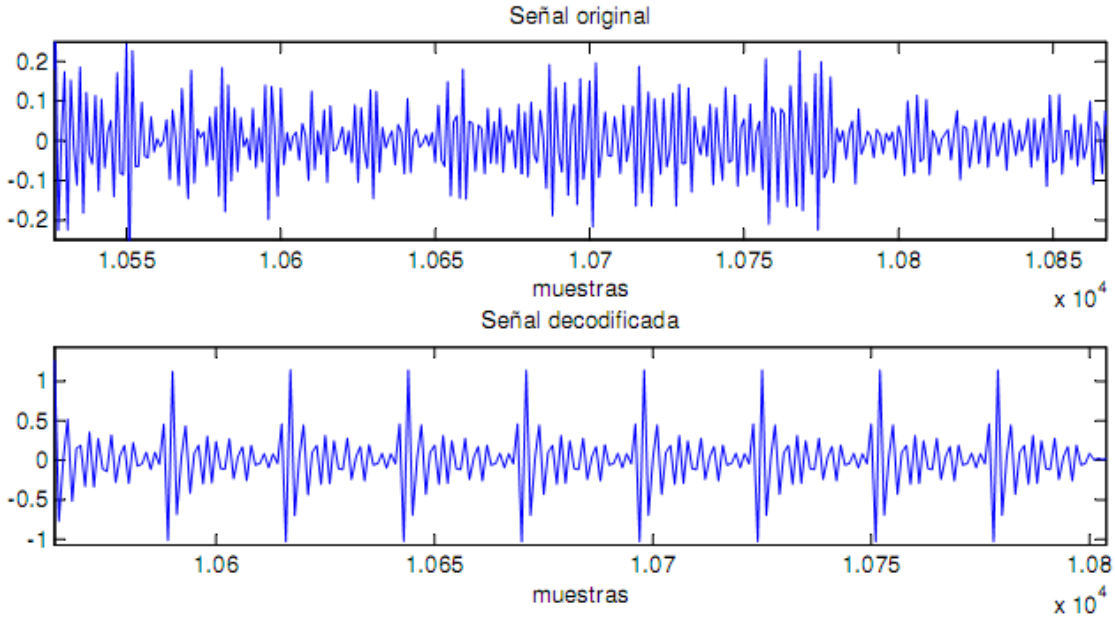


Figura 60: Diferencias entre la señal original y la decodificada con el vocoder lpc

En la parte superior la señal original y debajo la decodificada con el vocoder lpc. Se puede ver que la trama decodificada es perfectamente periódica, creada por un tren de pulsos glotales con una frecuencia de pitch, al haber sido considerada la trama sonora.

El vocoder homomórfico no tiene la limitación del lpc a la hora de generar la señal de excitación en el decodificador, pues el decodificador recibe la señal de excitación entera, y no se limita a generar un tren de pulsos glotales o ruido aleatorio, obteniendo así unos resultados mejores.

Sin embargo en las gráficas de SRR trama a trama vemos que algunas de ellas obtienen un resultado negativo. Si hacemos zoom sobre las señales apreciaremos ligeras diferencias (Figura 61).

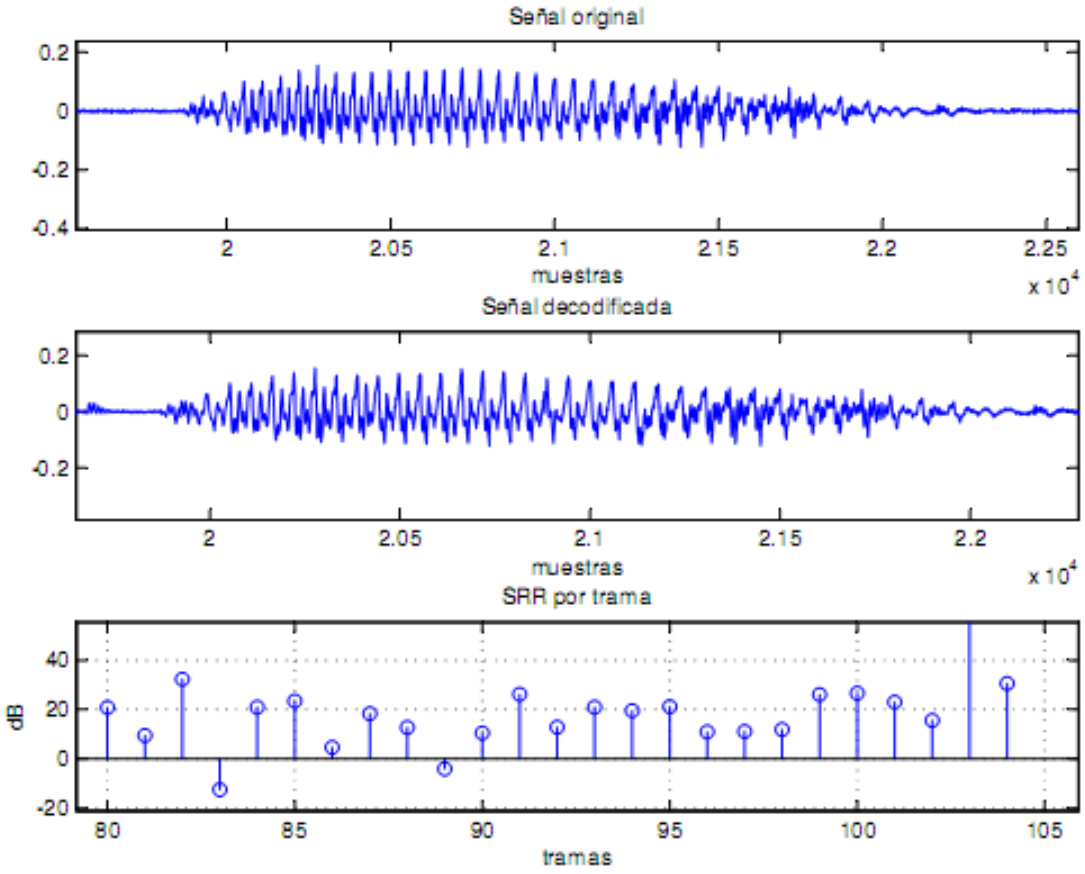


Figura 61: Análisis de varias tramas decodificadas con el vocoder homomórfico

La trama nº 83 corresponde aproximadamente con las muestras anteriores a la nº 20000 (20000/240 ≈ 83).

La trama nº 89 que corresponde con las muestras entre la 20000 y la 25000 (89*240 = 21360), en la que si miramos detenidamente se observan diferencias sutiles.

IMPACTO DEL PREPROCESADO EN LA SEÑAL

En la programación del vocoder homomórfico se ha usado un preprocesado que funciona parecido a un equalizador actuando positivamente en la zona de agudos. Aplana el espectro.

Por las pruebas con resultados positivos realizadas en la programación del vocoder homomórfico, el preproceso se introdujo en el código. Sin embargo el vocoder lpc no implementaba este preproceso, por que el código es anterior a este proyecto. No obstante para la comparativa se ha decidido integrarlo en el código y que ambas pruebas sean con la misma señal. El resultado es impresionante, la calidad de la señal decodificada mejora muchísimo en el vocoder lpc.

Para comprobar los resultados se mostraran a continuación las gráficas y los resultados de la señal 'vocoder homomórfico' decodificadas con y sin preprocesado con ambos vocoders.

Se muestra primero el resultado con el vocoder homomórfico.

Señal decodificada preprocesada

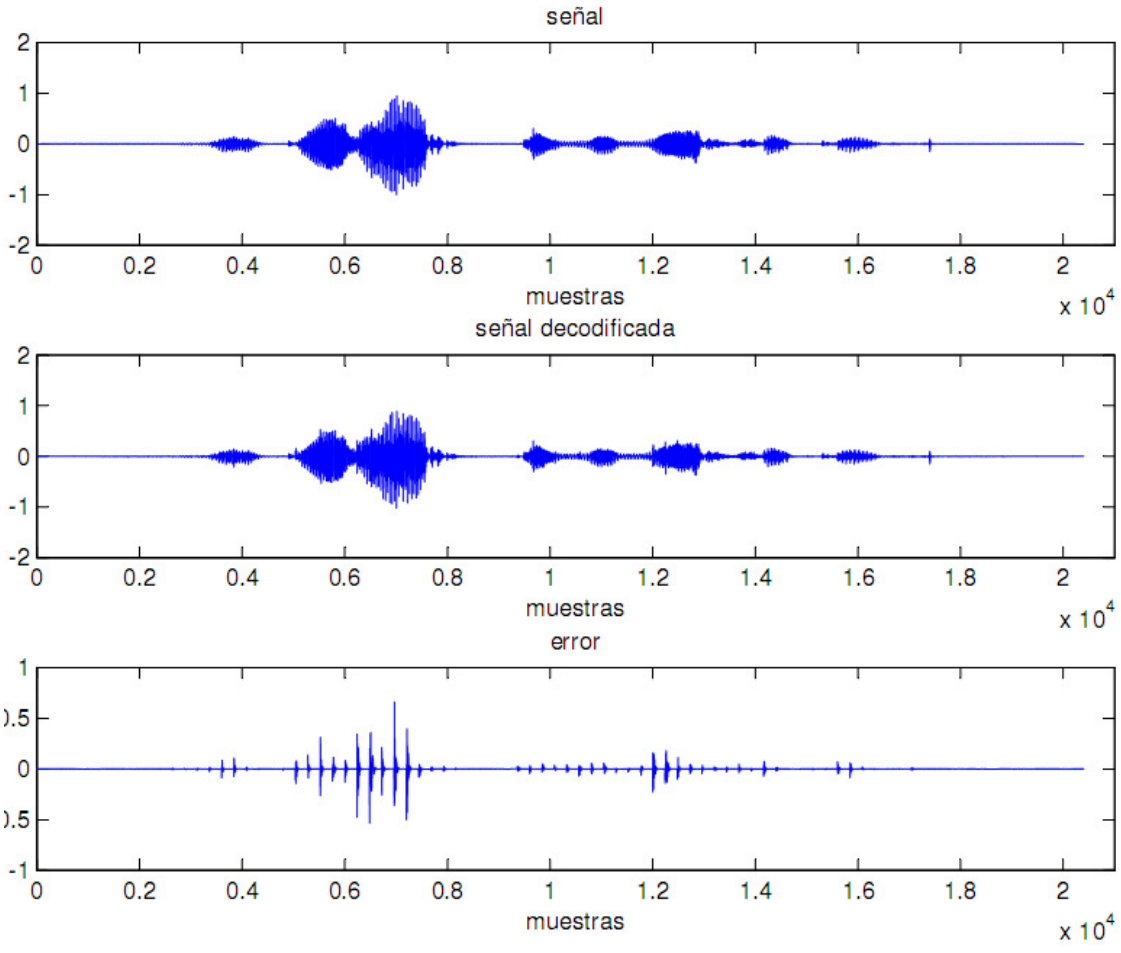


Figura 62: 1ª Señal en tiempo, la señal decodificada y el error cometido

Esta, Es la señal decodificada y el error cometido respecto a la original con el vocoder homomórfico usando el preproceso. Como se puede ver, es prácticamente igual, salvo zonas en las que el error es mayor.

Debajo se muestra la misma señal decodificada y su error pero sin ser preprocesada la señal con anterioridad:

Señal decodificada sin preprocesar

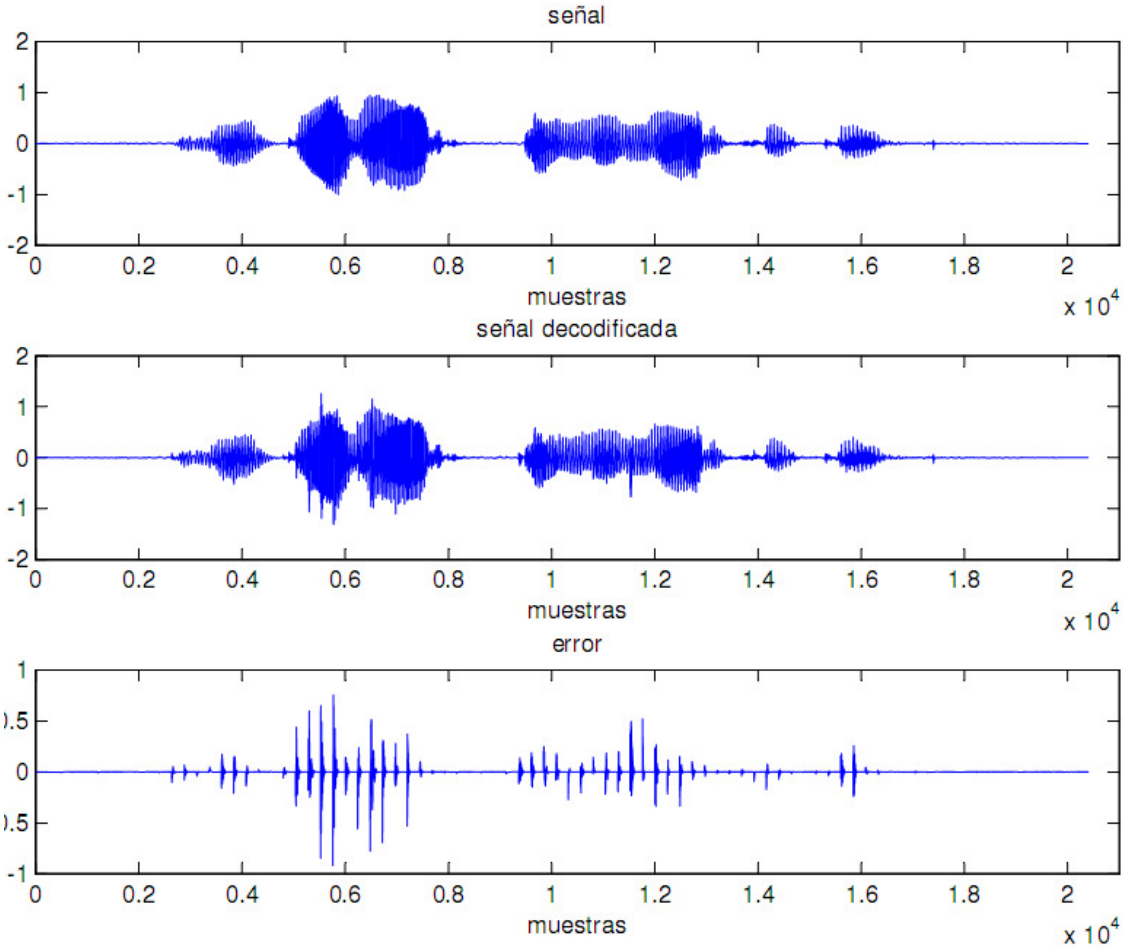


Figura 63: 1ª Señal en tiempo, la señal decodificada y el error cometido

Hay una mayor diferencia entre la señal original y la decodificada, y así se observa en el error, que es mayor también. Se puede decir que sin aplicar el preproceso a la señal original, la señal decodificada obtenida muestra mayores diferencias.

Seguidamente se encuentran los resultados con el vocoder lpc.

Señal original 'vocoder homomórfico'

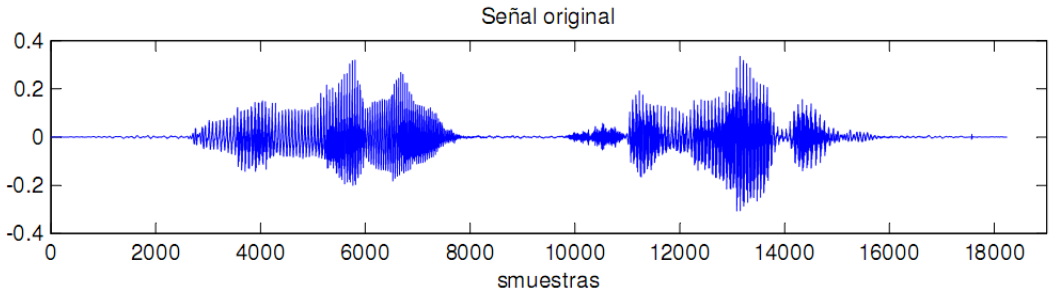


Figura 64: Señal original

Señal decodificada con vocoder lpc aplicando el preproceso

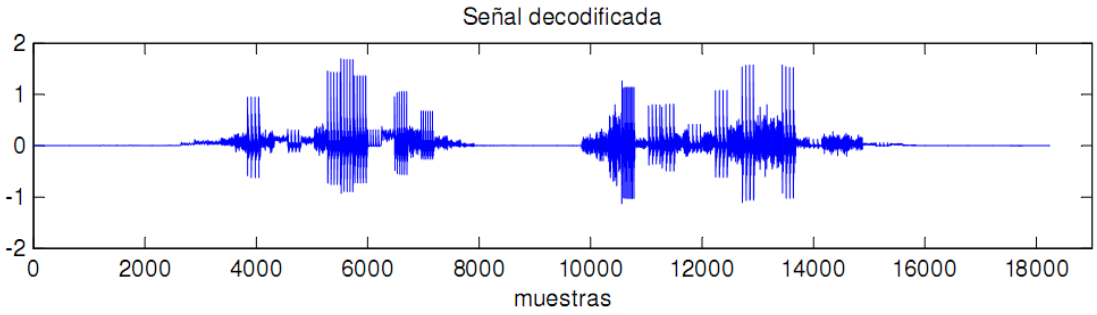


Figura 65: Señal decodificada con vocoder lpc de la señal sin preprocesar

Señal decodificada con vocoder sin aplicar preproceso.

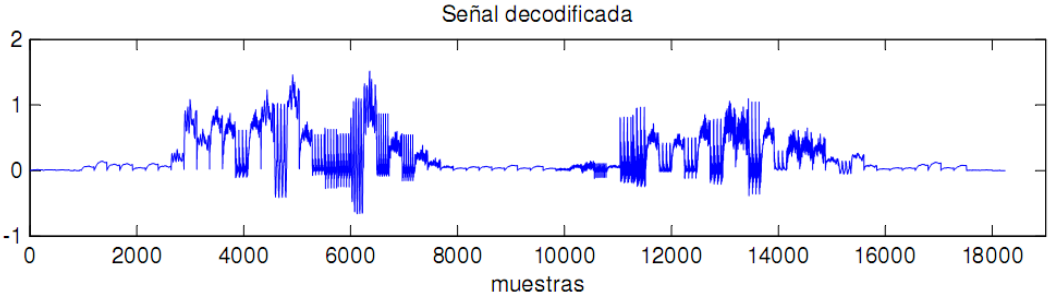


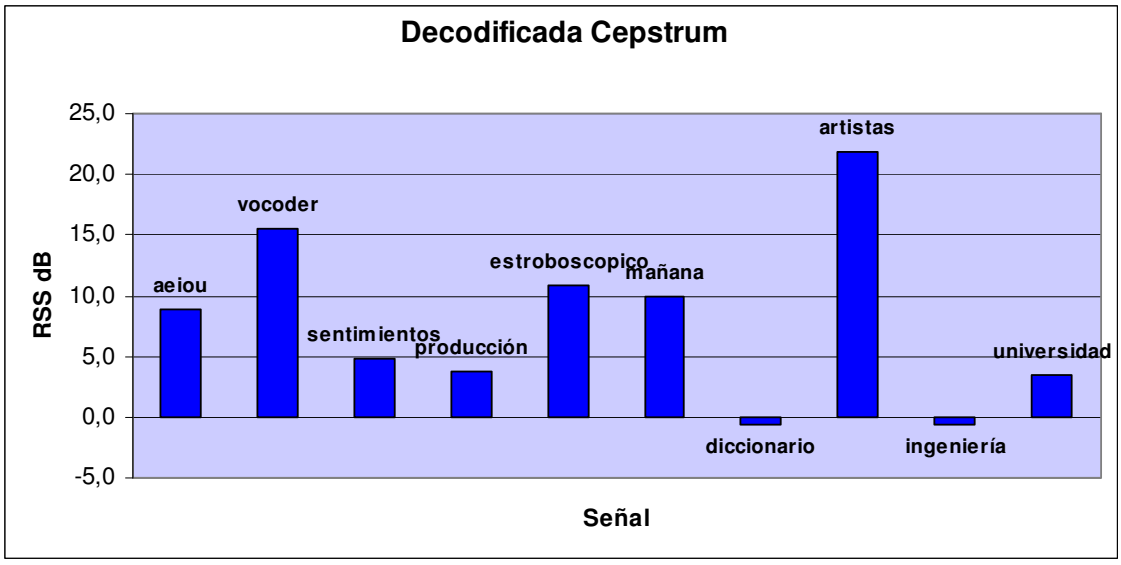
Figura 66: Señal decodificada con vocoder lpc de la señal preprocesada

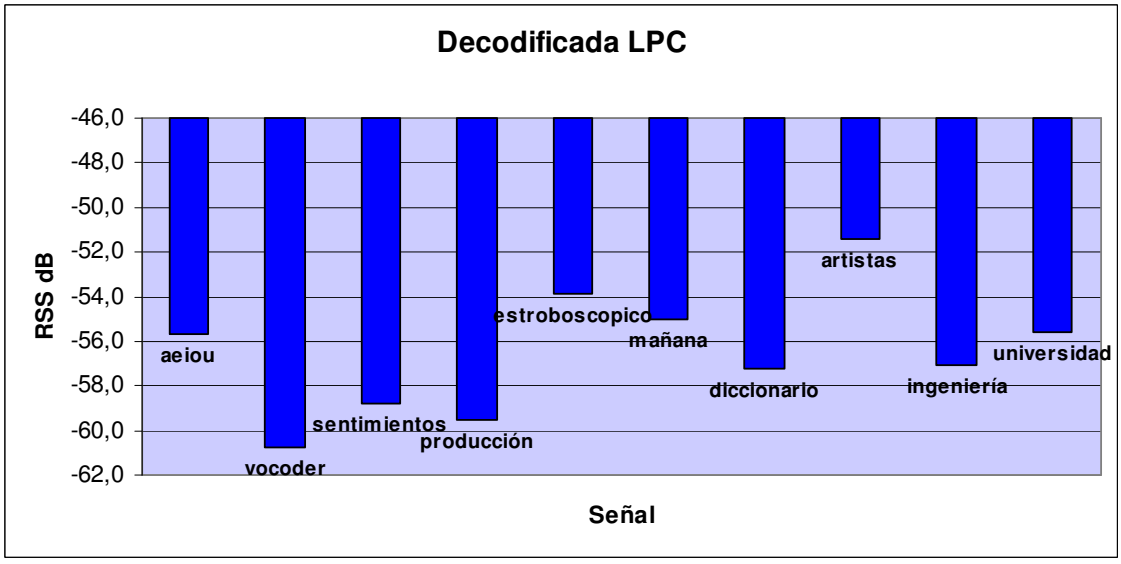
No es necesario realmente ver el error cometido, puesto que la diferencia es muy notable. La señal decodificada sin preprocesar no mantiene la integridad de la señal. Sólo esta compuesta por una especie de pulsos cuadrados, y la parte negativa casi desaparece. Sin embargo al aplicar el preproceso, la señal decodificada, aún siendo todavía bastante diferente a la original, empieza a tomar un parecido mayor. Se sigue manteniendo la limitación del lpc, en forma de claras diferencias entre tramas sordas o sonoras para la señal de excitación.

Y para finalizar se mostrarán los resultados SRR de las señales decodificadas con ambos vocoders con la señal original sin preprocesar.

VOZ MASCULINA

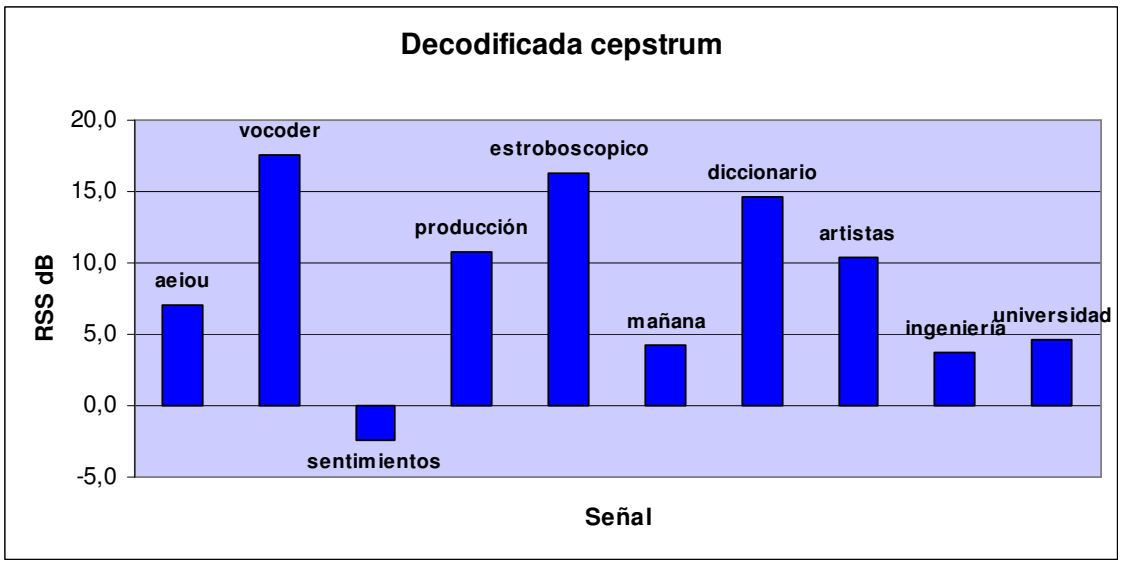
Señal	decodificada cepstrum	decodificada LPC
aeiou	8,8	-55,7
vocoder	15,5	-60,8
sentimientos	4,8	-58,8
producción	3,7	-59,6
estroboscópico	10,8	-53,9
mañana	9,9	-55,0
diccionario	-0,6	-57,3
artistas	21,8	-51,4
ingeniería	-0,6	-57,1
universidad	3,5	-55,6

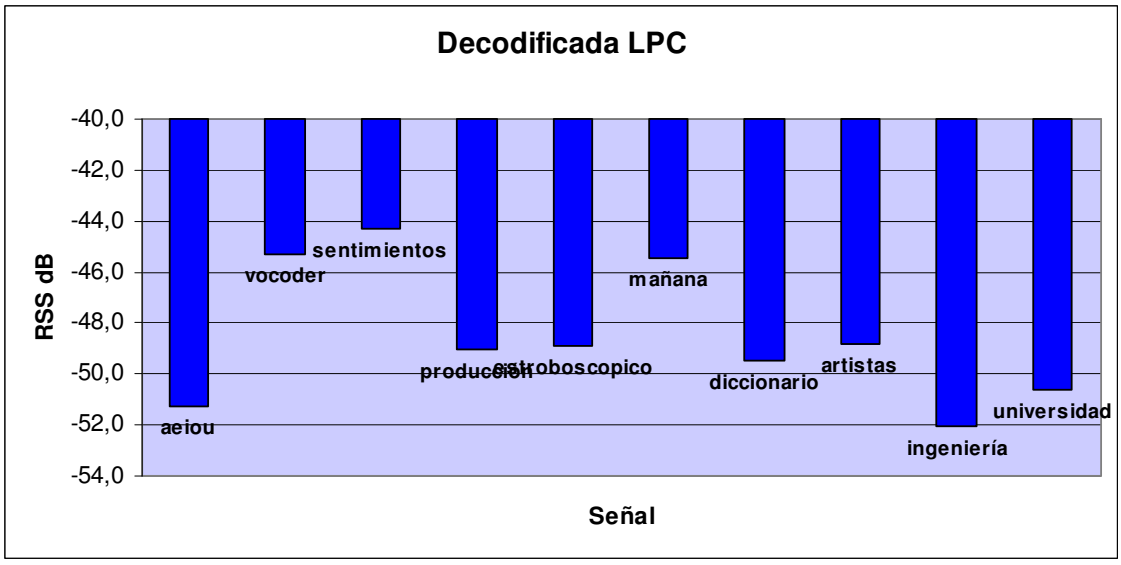




VOZ FEMENINA

señal	decodificada cepstrum	decodificada LPC
aeiou	7,1	-51,3
vocoder	17,5	-45,3
sentimientos	-2,4	-44,3
producción	10,7	-49,1
estroboscópico	16,3	-48,9
mañana	4,2	-45,5
diccionario	14,6	-49,5
artistas	10,4	-48,9
ingeniería	3,7	-52,0
universidad	4,7	-50,6





Se observa que la pérdida en el vocoder lpc es asombrosa, pues pasa de estar en unos valores entorno a -30 dB a obtener -40 dB en el mejor de los casos y más de 60 dB negativos en el peor.

Con las señales del vocoder homomórfico los cambios no son tan acusados, y además no se observa una relación general, pues algunas señales obtienen resultados peores y otras sin embargo peores.

Lo que esta claro es que este proceso que no tiene coste computacional, ni impacto en la cantidad de información enviada, mejora de forma general el proceso de codificación-decodificación.

También resulta interesante remarcar que el proceso de equalización realza las frecuencias donde se centra la inteligibilidad de la voz, entorno a los 2500 Hz.

Esa zona es donde más sensible es a la audición del oído humano; es algo genético, Los humanos son más sensibles a las frecuencias del habla humana para poder entender y oír mejor a otras personas. Hay que apuntar que se pierde un poco de volumen, pues la voz humana concentra gran parte de la energía a 500 Hz, cosa fácilmente solucionable. Para finalizar: en la comunicación es más importante la inteligibilidad, algo que el preprocesado ayuda a mejorar.

CANTIDAD DE INFORMACIÓN REQUERIDA PARA CODIFICAR LA SEÑAL

El vocoder lpc ofrece unos resultados muy pobres, pero en contrapartida los requerimientos de ancho de banda son muy escasos. Para codificar una trama de voz son necesarios sólo 13 parámetros:

Solamente 3 para la señal de excitación

- Sonoridad: sonora/sorda.
- Pitch, en el caso de que sea sonora.
- Energía de la señal.

Y 10 para parametrizar el filtro del tracto vocal

- Respuesta al impulso del filtro lineal.

En el vocoder homomórfico se usan muchos más.

- 30 son para modelizar el tracto vocal (los 30 primeros del cepstrum).
- N-30 para la señal de excitación, donde N es dependiente de la ventana.

En los ejemplos mostrados con una ventana de 30 ms y señales a 8 KHz se obtienen ventanas de 240 muestras -30 del tracto vocal = 210. La diferencia es notable; se dedican 3 parámetros para la señal de excitación en el vocoder lpc, por unos 200 en el vocoder homomórfico.

En el tracto vocal la diferencia no es tanta, y además se pueden utilizar más de 10 parémtros en el vocoder lpc para el filtro H(z), como también se puede reducir el nº de muestras representativas del tracto vocal en el cepstrum, a 25 o incluso 20.

No obstante no es esto lo que marca la diferencia, sino la señal de excitación usada en cada uno de los sistemas. El vocoder lpc esta muy limitado por la mala calidad de la señal de excitación que se genera en el decodificador: Solamente hay dos posibilidades para cada trama, o un tren de pulsos glotales o ruido aleatorio, y la ganancia de cada una de ellas. Con los pocos parámetros disponibles no es posible aumentar la señal de excitación. Habría que recurrir a otras técnicas.

En el vocoder homomórfico es posible recuperar la señal de excitación prácticamente intacta pues son pocos los valores que se pierden en la zona del cepstrum en la que se mezclan tracto vocal y excitación, y así, la señal de excitación es de mucha mayor calidad, y por consiguiente la señal decodificada.

CAPACIDAD DE DETECCIÓN DE SONORIDAD Y PITCH

Ambos vocoders usan técnica diferentes para determinar la sonoridad o no de una trama.

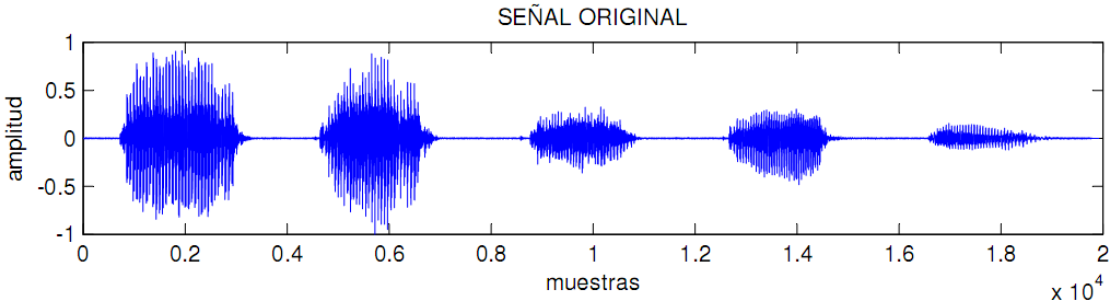
El vocoder lpc se basa en calcular la auto-correlación de la tama, y el vocoder homomórfico determina la sonoridad a partir del cepstrum real.

Se ha realizado una prueba comparativa con varias señales, y se verá como diferencian las tramas sonoras de las sordas y como evalúan el pitch. Las primera señal probada es la más sencilla de nuestro banco de pruebas, la señal 'a e i o u'. Es una señal totalmente sonora, en la que será fácil ver cuando detectan una trama sonora y cuando no.

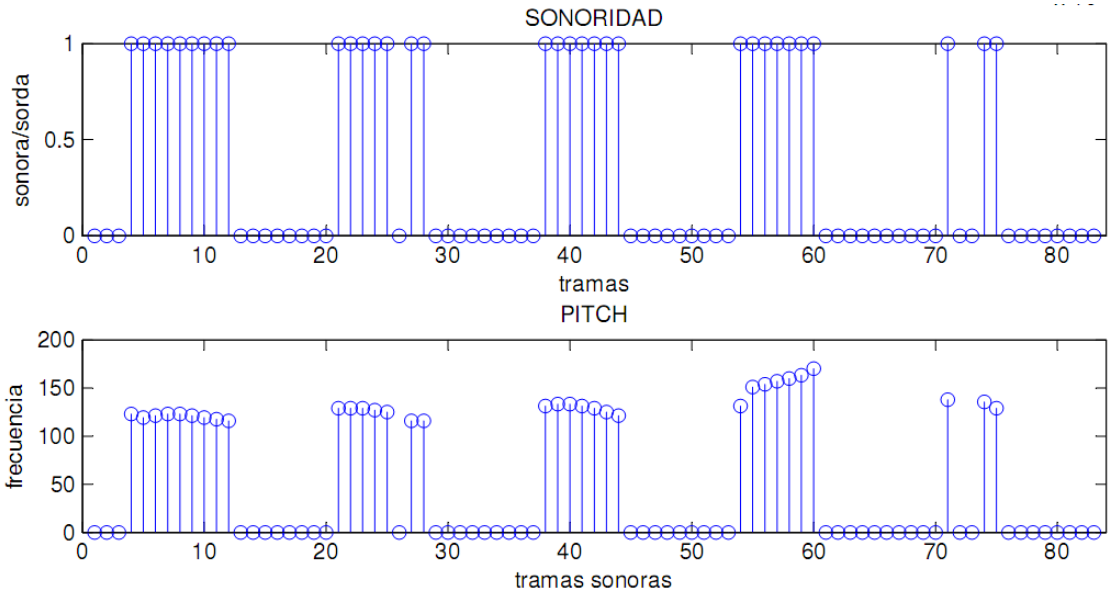
La segunda señal será la frase 'mañana soleada', que tiene una 's' al principio de la segunda palabra, y las demás tramas deberían ser sonoras.

Y la última señal es 'Universidad Pública de Navarra', que en general en una frase sonora con una 's' en medio de la primera frase, y 'bl' en la mita de la segunda.

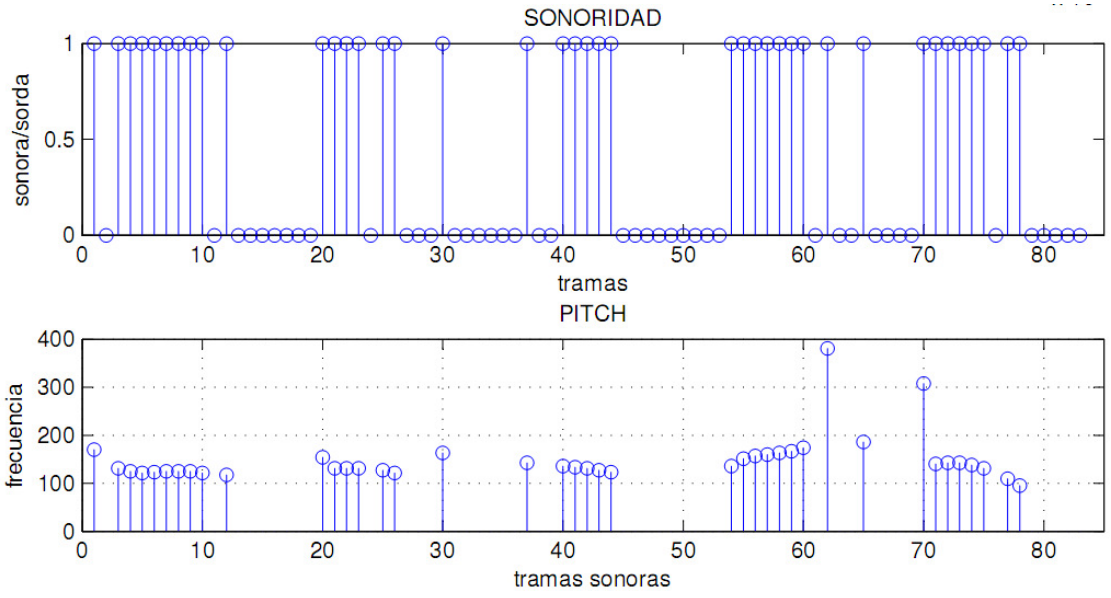
SEÑAL Nº 1 'AEIOU'



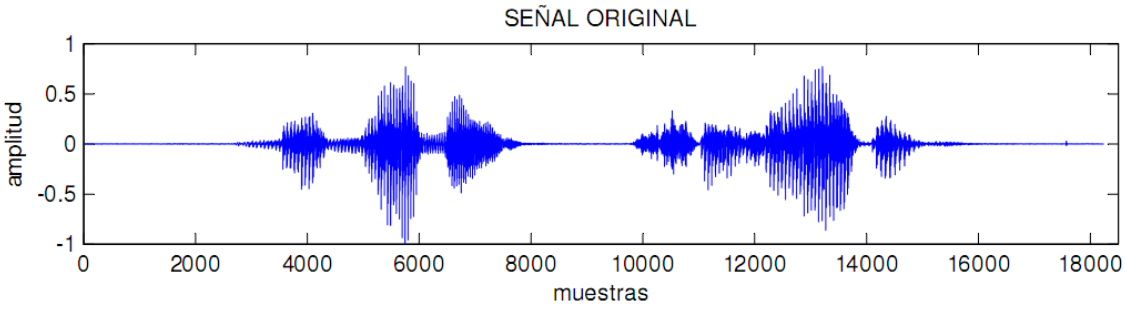
- Picth cepstrum



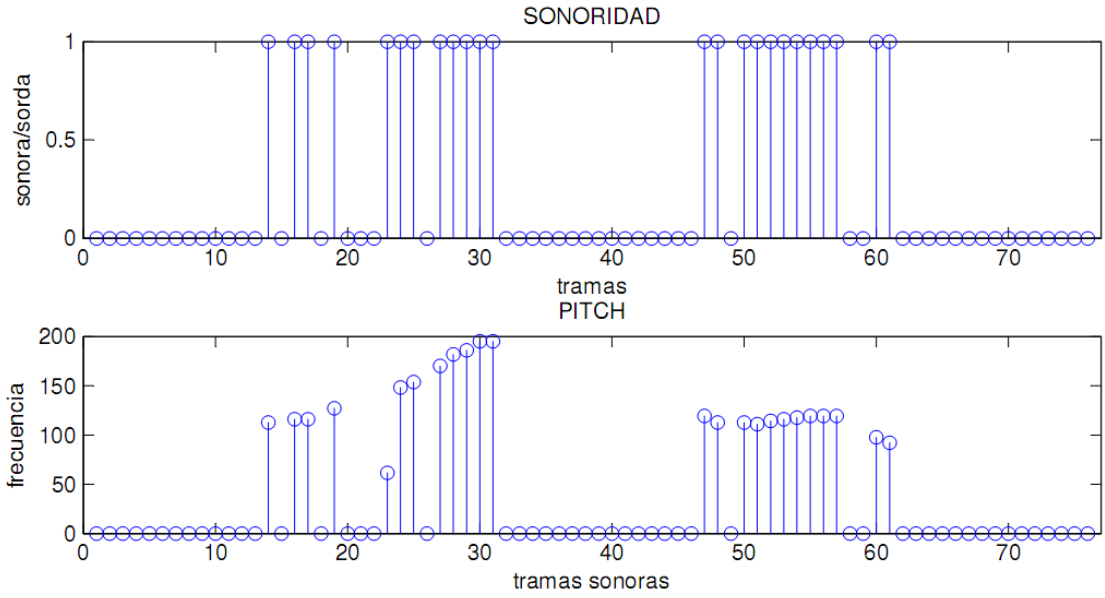
- Pitch lpc



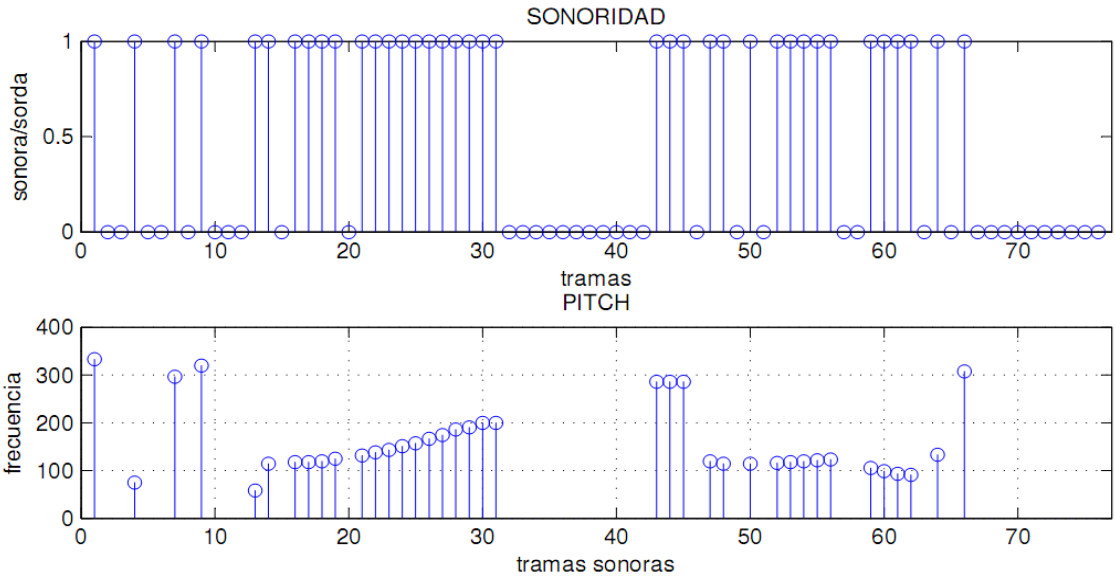
SEÑAL Nº 2 'MAÑANA SOLEADA'.



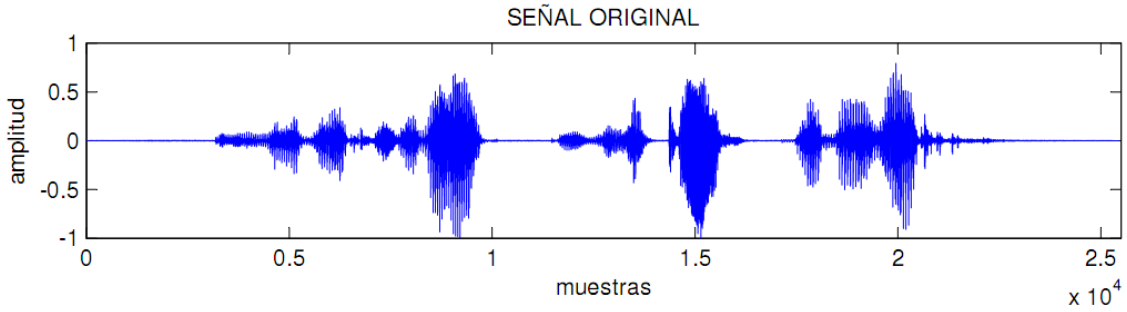
- Pitch cepstrum



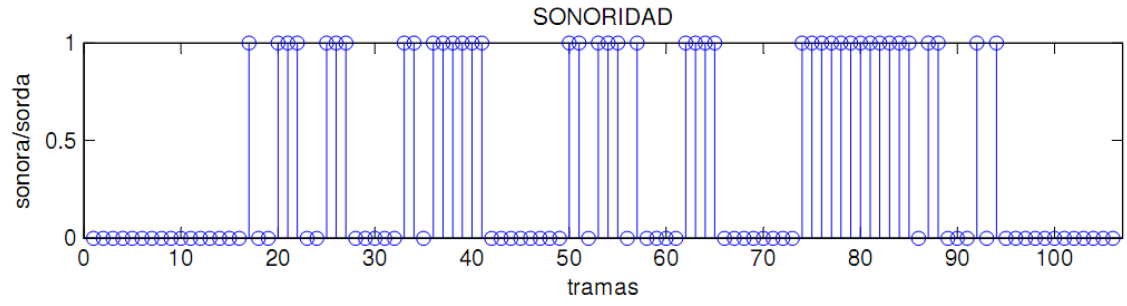
- Pitch lpc

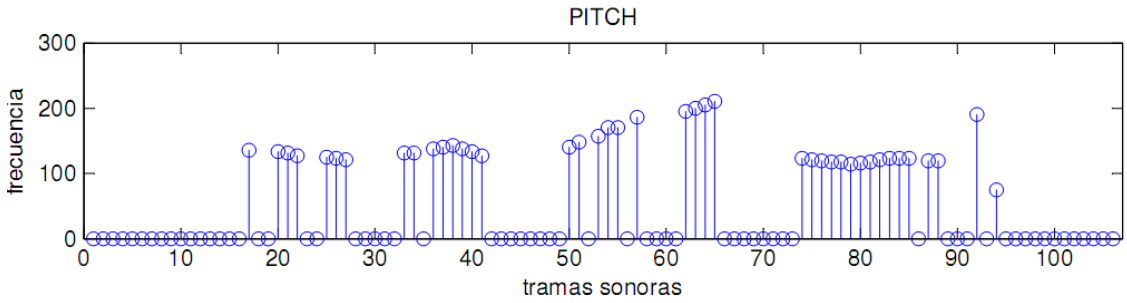


SEÑAL Nº 3 'UNIVERSIDAD PÚBLICA DE NAVARRA'

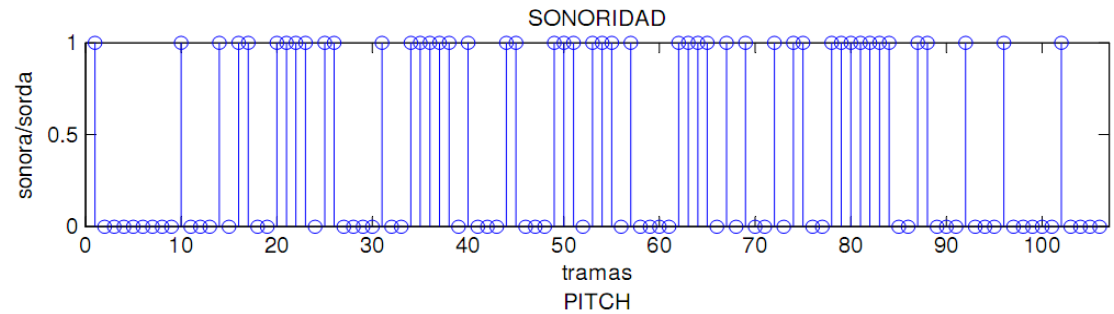


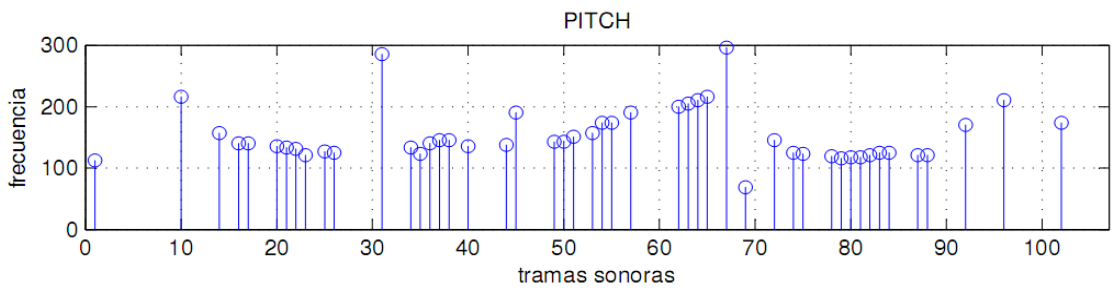
- Pitch cepstrum





- Pitch lpc





En la primera señal se ve como el vocoder homomórfico recoge muy bien los espacios entre vocales como sordas, las vocales las interpreta como sonoras, exceptuando alguna trama de la vocal ‘u’, quizá por el bajo volumen de esta.

El vocoder lpc aunque en líneas generales interpreta bien las vocales como tramas sonoras, tiene problemas en señalar los silencios como tramas sordas, de ahí que algunas frecuencias de pitch sean algo raras, como la que marca en la trama 62 como 400 Hz, cuando posiblemente esa trama sea sorda, pues pertenece al final de la ‘o’.

La segunda señal es más complicada, y aquí ya se pueden ver las deficiencias de la auto-correlación, pues donde el vocoder homomórfico detecta muy bien los silencios y la ‘s’ de ‘soleada’ como tramas sordas, el vocoder lpc no, y marca muchas tramas como sonoras cuando no lo son. En cálculo del pitch, las tramas en las que coinciden ambos vocoders en calificar como sonoras, el resultado es igual.

Y en la tercera señal, más larga y por ello más compleja, el vocoder homomórfico sigue diferenciado muy bien los silencios de la señal, y tanto la ‘s, como ‘bl’ las interpreta como sordas, y el vocoder lpc tiene los mismo problemas que antes, confusión con los silencios.

En líneas generales, el vocoder homomórfico, valiéndose del cepstrum real, comete menos errores que el lpc. Lpc usa la auto-correlación de la trama como método de distinción sorda/sonora, método que tiene problemas para declarar correctamente las tramas sordas. Esto se debe a que la autocorrelación compara la señal consigo mismo, y en una trama en la que hay silencio, la forma de onda es muy básica y no presenta las aleatoriedades del ruido o una señal sorda, sino más bien pequeña ondulaciones que pueden crear una falsa trama sonora.

CONCLUSIONES Y LÍNEAS FUTURAS

CONCLUSIONES

Al margen de los objetivos concretos del presente proyecto, cualquier estudio o proyecto de investigación tiene como objetivo principal llegar a ciertas conclusiones sobre el tema estudiado. Recordando los objetivos de este proyecto, se buscaba diseñar un vocoder basado en el análisis cepstral, que mejorara la calidad de las señales decodificadas, respecto a un vocoder lpc, y realizar una comparativa entre ambos. Además se intentó ver si el análisis cepstral resultaba una herramienta útil en la clasificación clásica de tramas en sonoras o sordas y en la determinación del pitch.

Se ha dicho que la señal de voz se puede expresar como la convolución de una señal de excitación y un filtro que modela el trato vocal, y ambos cambian en el tiempo.

La posibilidad que brinda el procesado Homomórfico junto a las propiedades de la señal de voz, permiten realizar una transformación de la señal en tiempo al dominio cepstral y volver al tiempo mediante el proceso inverso, sin que la señal sufra degradación, pues es una transformación lineal. En el dominio cepstral se ha visto que el tracto vocal se concentra en la zona más cercana al origen, y tiende a desaparecer con *n* 'grandes', mientras que la señal de excitación se compone de unos picos que aparecen en el cepstrum más allá de las primeras muestras.

La separación en el dominio cepstral mediante ventanas de la señal de excitación y el tracto vocal, y su posterior convolución ha demostrado ser una técnica que consigue unas señales decodificadas de mucha mayor calidad que lo que consigue el vocoder lpc. La principal razón es que la señal de excitación no hay que generarla en el decodificador, pues gracias al análisis cepstral conseguimos sacarla de la propia señal de voz.

Así pues en lugar de tener una señal de excitación simple, tenemos toda la señal de excitación con todos sus matices. Al filtrar la señal excitación con el filtro que modela el tracto vocal (al volver a convolucionar) obtenemos una réplica muy buena de la señal original. Esto conlleva que la demanda en cuanto a ancho de banda sea mucho mayor también, pues con el vocoder lpc, sólo se necesitan 3 parámetros para reconstruir la señal de excitación, y en vocoder homomórfico se la señal entera.

No obstante aunque la tendencia general es positiva, los resultados obtenidos muestran muchas diferencias entre las diferentes señales de prueba utilizadas, hecho que no curre con el vocoder lpc, donde los valores de SRR son malo, pero igual de malos para todas las señales.

También cabe destacar que el comportamiento del vocoder homomórfico no se ve alterado por una señal de voz de pitch más alto como puede ser la de una mujer. Los resultados con la señal de voz femenina siguen un patrón parecido.

Uno de los descubrimientos de este proyecto ha sido el filtro de preprocesado de la señal. El vocoder homomórfico no necesariamente se beneficia de él, puesto que sin aplicar preprocesado a la señal de entrada, los resultados ya son buenos. Sin embargo donde más se nota este proceso en el vocoder lpc. Los resultados obtenidos con el vocoder con la señal preprocesada han mejorado notablemente haciendo la señal decodificada más inteligible y además no incrementa las necesidades de ancho de banda.

Por último el cepstrum real ha demostrado ser una herramienta mejor para la clasificación de las tramas en sonoras o sordas, por que el método de la autocorrelación tiende a clasificar algunas tramas que son silencio como sonoras. De querer usarse este método, sería conveniente acompañarlo con algún criterio más como el cálculo de la energía, para descartar aquellas tramas en las que hay silencio pero que han sido consideradas como sonoras.

LÍNEAS FUTURAS

En este proyecto se ha analizado el procesado homomórfico para la separación de la señal de excitación y tracto vocal de la propia señal de voz como técnica de codificación-decodificación de esta. Se ha visto que resultados se obtienen del análisis cepstral y se ha realizado una comparativa con un vocoder lpc.

Se ha podido comprobar el potencial del procesado homomórfico y las posibilidades que éste brinda, y a la vista está que la calidad obtenida es muy buena, sin embargo, sigue habiendo ruido y clicks en la señal decodificada, ruidos, que posiblemente se deben a que la separación en el dominio cepstral del tracto vocal y la señal de excitación no es del todo limpia. Para mitigar estos problemas se podría

trabajar en una mejora del proceso de separación, por ejemplo usando ventanas diferentes a la rectangular. En la presente memoria se ha hablado del enventanado y los efectos que tiene sobre la señal, así como las diferentes ventanas existentes, y quizá aplicando alguna ventana que reduzca el peso de las muestras d los extremos se pueda lograr que la s interferencias entre tracto vocal y señal de excitación en el cepstrum desaparezcan.

En lo que respecta al ancho de banda necesario, es un aspecto importante en cualquier vocoder y por lo tanto siempre es importante intentar reducir la cantidad de información pero con la menor pérdida posible en la calidad. En ese aspecto se puede trabajar en el dominio cepstral, en la parte que ocupa la excitación, reduciendo la cantidad de muestras, pues la señal de excitación es la que mayor cantidad requiere para ser transmitida, mucho más que el tracto vocal

Una reducción de la cantidad de muestras de la excitación traerá consigo una reducción también en la calidad final, por lo tanto hay que valorar con cuantas muestras podemos trabajar y donde esta el límite que compromete la calidad. Para este fin se puede implementar la técnica de análisis por síntesis, partiendo por ejemplo de la mitad o un cuarto de las muestras totales e ir comprobando hasta que punto sale rentable añadir más muestras de la excitación en el cepstrum y así reducir la cantidad de información en el codificador.

Otra técnica que se puede probar en la dirección de intentar reducir la cantidad de información transmitida es aplicar la teoría de la repuesta psicoacústica. La audición humana es sabido que no es lineal con la frecuencia, y además debido al enmascaramiento un sonido deja de percibirse debido a la presencia de otro, por lo tanto se podría probar estas técnicas que al fin y al cabo consisten en la compresión de la señal sin pérdida aparente de calidad.

BIBLIOGRAFÍA

L. Rabiner, R.W. Schafer, “Digital Processing of Speech Signals”. Prentice-Hall, 1978.

T. F. Quattieri, “Discrete-Time Speech Signal Processing. Principles and Applications”. Prentice-Hall, 2002.

A. Spanias, T. Painter, V. Atti, “Audio Signal Processing and Coding”. Wiley, 2006.

R. Goldberg, L. Riek, “A practical handbook of speech coders”. CRC Press, 2000.

D. O’Shaughnessy, “Speech Communications. Human and Machine (2nd. Ed.)”. IEEE Press 2000.

B. Gold, N. Morgan, “Speech and audio signal processing”. John Wiley & Sons, 2000.

L. Rabiner, B.H. Juang, “Fundamenatls of speech recognition”. Prentice-Hall, 1993.

A. V. Oppenheim, A.S. Willsky, S. Hamid Nawab, “Signals and Systems” Second Edition. Prentice Hall. 1997.

S. S. Soliman, M.D. Srinath, “Continuous and Discrete Signals and Systems” Second Edition, Prentice-Hall, 1998.

S. Haykin, B. Van Veen, “Señales y Sistemas” Limusa Wiley, 2001.

F. G. Stremler, “Introducción a los sistemas de comunicaciones” Adisson-Wesley, 1990.

J. G. Proakis, D . Manolakis, “Tratamiento Digital de Señales”, Tercera Edición. Prentice-Hall, 1998.

A. V. Oppenheim, “Discrete-time Signal Processing”, Prentice-Hall, 1989.