**Análisis de la señal de voz**

**Introducción.**

Técnicas clásicas en RAH: Son técnicas que se utilizan en forma previa al posterior reconocimiento automático del habla.

* Acondicionamiento, preproceso y ventaneo.
* Análisis de energía por bandas (Fourier e integración).
* Coeficientes de predicción lineal (LPC): Forma de representación de la señal de voz.
* Coeficientes cepstrales (en escala de mel).
* Métodos de estimación de F0 y formantes.
* Métodos de cuantización vectorial (HMM discretos).
* Otros: PLP, Rasta, modelos auditivos.

Acondicionamiento, preproceso y ventaneo.

Existen métodos relativamente sencilos para “preparar” la señal, independientemente de su análisis posterior:

* **Remoción de la media:** La remoción de la media permite tener una señal con media 0 que facilita el análisis posterior con los otros métodos.
* **Filtrado para eliminación de ruido:** Si el espectro de la señal no está solapado con el de la señal del habla se suele hacer una filtración de ruido. Se pueden realizar filtrados más complejos posteriormente si es necesario.
* **Pre-énfasis:** . Con el pre-énfasis se busca hacer un filtrado sencillo de primer orden con un sistema auto regresivo de un solo coeficiente con un valor aproximado de 0.97. Hace una especie de derivada pesada que enfatiza algunos contenidos de alta frecuencia que suelen facilitar la aplicación de métodos de estimación lineal, por ejemplo.
* **Perturbación:** . Se agrega una pequeña perturbación a la señal con un valor aleatorio (números frecuentemente con orden del 1\*1000). Este pequeño agregado de ruido mejora la estimación a lo que se refieren los métodos de estimación de modelos estadísticos.
* **Ventaneo:** Por último, se trabaja a nivel de ventana, entonces la señal se corta en trozos mediante ventanas particulares (una de las más utilizadas es la de Hamming que como se ve en la imagen anterior se usan funciones trigonométricas para tomar una ventana suave que nos permite particionar la señal y aplicar los métodos por separado a cada parte).

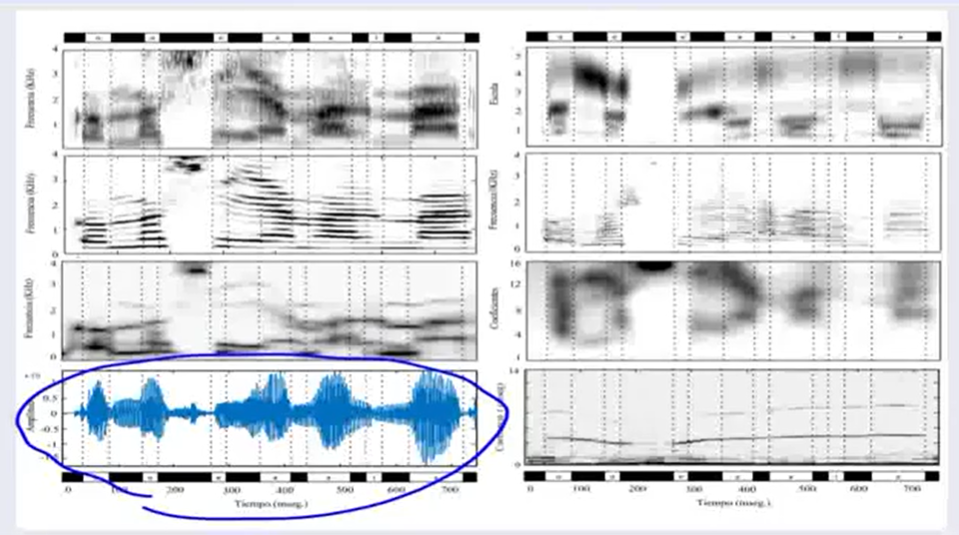
Ventaneo: Análisis por tramos.

Este análisis por tramo supone que la velocidad de cambio del aparato fonador es suficientemente lenta en relación con el ancho de la ventana. O sea que para cualquier análisis que hagamos podemos suponer que la señal es estacionaria en esa ventana pequeña, que en las señales del habla que trabajamos se suele usar entre 10-20ms para el ancho de la ventana.

Tipos de análisis de señales

* **LTI (Basados en Fourier)**: En donde suponemos que el sistema que produce la señal de voz es lineal e invariante en el tiempo.
* **Lineales no estacionarios**: Usamos la transformada de Fourier de corta duración o la de onditas. En el caso del habla si analizamos toda la señal no cumple que sea LTI, pero al trabajarla en ventanas pasa a ser estacionaria.
* Otros:
* Específicos señal habla:
* Cepstral, LPC, PLP, RASTA, modelos auditivos.
* F0 y formantes.
* **Robustez:** Cómo el análisis que vamos a hacer enfatiza las pistas acústicas en forma más o menos independiente del ruido de fondo.

Cada tipo de análisis nos va a permitir encontrar algunas características particulares de la señal de voz:



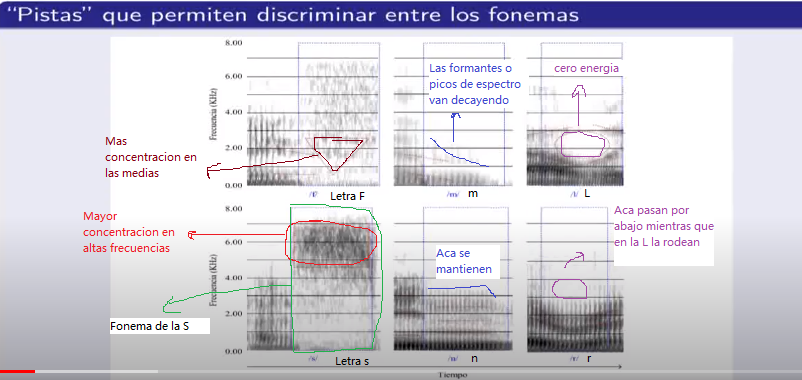
En la parte inferior izquierda está la señal de voz. Los otros cuadros son distintos tipos de análisis por tramos aplicados a la misma señal. Inmedianemente arriba está el espectro estimado a partir de los coeficientes de predicción lineal por tramos. Los dos superiores a éste son dos espectogramas; abajo: espectograma de banda angosta; arriba: espectograma de banda ancha.

**Esta representación de los espectogramas cambia solamente cambiando el ancho de la ventana:** Ventanas más largas nos dan menor resolución temporal pero mayor resolución frecuencial. Jugando con estos parámetros podemos hacer que se haga más evidente por ejemplo, las frecuencias formates, frecuencia fundamental o algún otro parámetro imporante de interés o pistas acústicas para la señal de voz.

A la derecha de arriba hacia abajo están: Análisis de transformada onditas (Escalograma); representación basada en la distribución de Binner-Bill; Coeficientes de predicción lineal perceptual (PLP); Coeficientes Ceptrales.

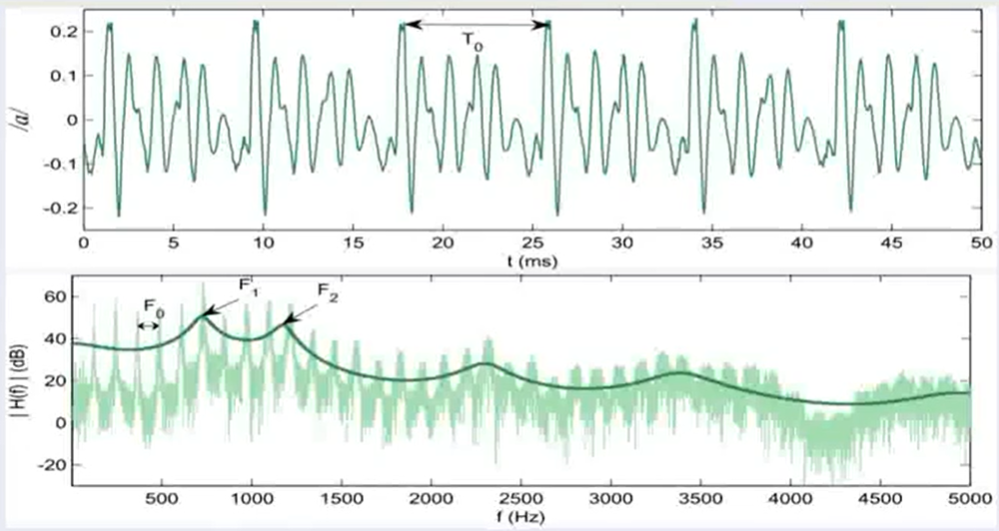
Pistas acústicas.

Lo que le importa al sistema automático que pretende clasificar fonemas, sonidos y palabras. Entonces lo que busca el sistema es encontrar las pistas acústicas que permiten discriminar entre los distintos fonemas.



Vocal sostenida – Período y Frecuencia fundamental (F0) – Formantes.

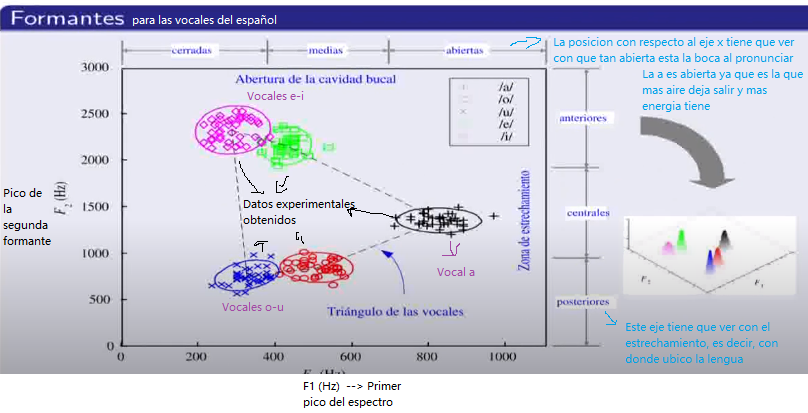
Sea una vocal sostenida durante un tiempo corto. Se realiza de manera artificial para el análisis.



La señal que representa la a sostenida es de tipo cuasi-periódicas, donde la señal es muy parecida en cada período, pero con pequeñas diferencias. Sin embargo, este cuasi período nos permite encontrar el período fundamental de la señal (T0).

La **frecuencia fundamental F0** corresponde a la frecuencia glótica que está asociada a los pulsos que se dan en la glotis, la cual está presente en los fonemas sonoros, y **es una componente importante de la entonación del habla**.

La curva verde de abajo representa el espectro suavizado estimado a través de los coeficientes de predicción lineal. Los picos de este espectro corresponden a las **frecuencias formantes (F1, F2, F3,…)**, que **permiten discriminar entre las vocales**. Su variación temporal posibilita también diferenciar entre los diferentes fonemas sonoros.

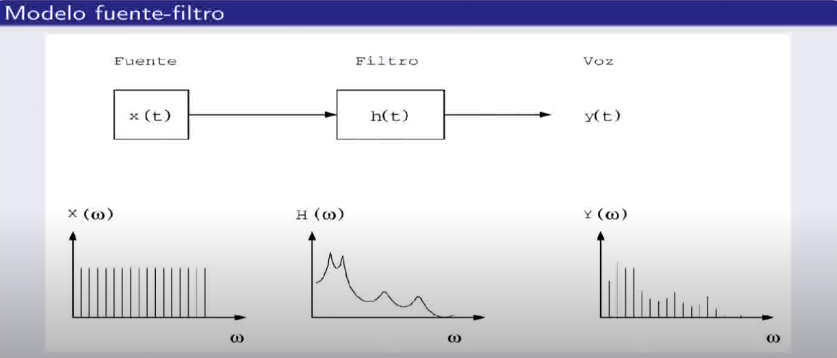


**Coeficientes de predicción lineal.**

Modelo lineal del aparato fonador.

También conocido como modelo fuente-filtro. Acá suponemos que la señal de voz está producida a través del aparato fonador mediante la transformación de la energía muscular en energía acústica. Esto puede modelizarse como un sistema de filtros que modifica a una o más fuentes sonoras.

Como trabajamos suponiendo que el sistema es lineal, hay un filtro con una función de transferencia H(ω) que corresponde al tracto vocal, y una función X(ω) que es la fuente de excitación (perturbación acústica de la corriente de aire pulmonar, tiene que ver con si es una perturbación de tipo cuasi-periódica, o un ruido cuando hay una constricción fuerte). El espectro de la señal resultante puede representarse como Y(ω) = H(ω) X(ω).



**Mecanismo de excitación del tracto vocal.**

El sistema respiratorio es la principal fuente de energía para producir sonidos en el aparato fonador, en forma de flujo de aire y presiones.

Existen tres mecanismos de excitación:

* Generación de pulsos cuasi-periódicos: las cuerdas vocales modulan el flujo de aire.
* Generación de ruido de banda ancha: el flujo procedente de los pulmones pasa por una constricción del tracto vocal. Esto genera un sonido del tipo ruidoso, por ejemplo la S o la F que podemos asociar a un ruido blanco.
* Excitación del tipo plosivo: el flujo produce una presión en un punto de oclusión total del tracto vocal y se libera rápidamente, por ejemplo, la P o la T.

**Función fonatoria de la laringe.**

La frecuencia fundamental de oscilación en la laringe es la frecuencia de los pulsos glóticos. Cada uno de estos pulsos glóticos, no son simplemente del tipo Delta de Dirac, sino que tienen una forma particular. Allí intervienen las cuerdas vocales, los cartílagos en los que se insertan y los músculos laríngeos intrínsecos.

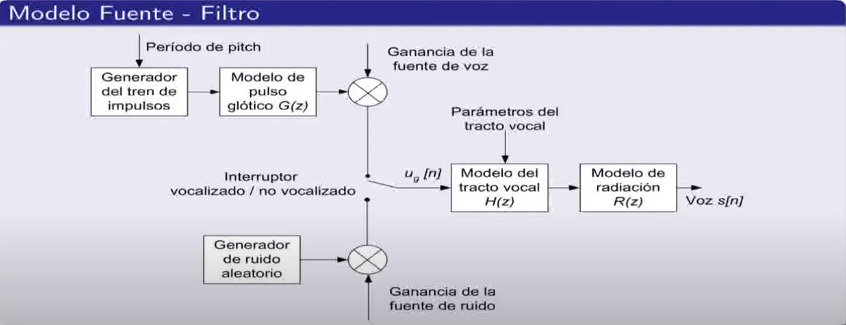
La forma de onda de los pulsos generados puede representarse en forma simplificada como una onda triangular. Un modelo más adecuado es el de Rosemberg, donde estos pulsos tienen la siguiente forma:

y son los parámetros que me dan la forma particular de cada pulso, que son propios del hablante:

* En el hombre, la frecuencia de esta onda varía entre 100 y 170 Hz, en las mujeres entre 180 y 280 Hz y en los niños puede superar los 300 Hz.
* Los valores de estga frecuencia glótica se modifican en forma voluntaria y son los responsables de la frecuencia fundamental F0 que se percibe como “tono”.

**Función transferencia H(ω):**

Producción de la señal de voz.



Modelo autoregresivo del tracto vocal:

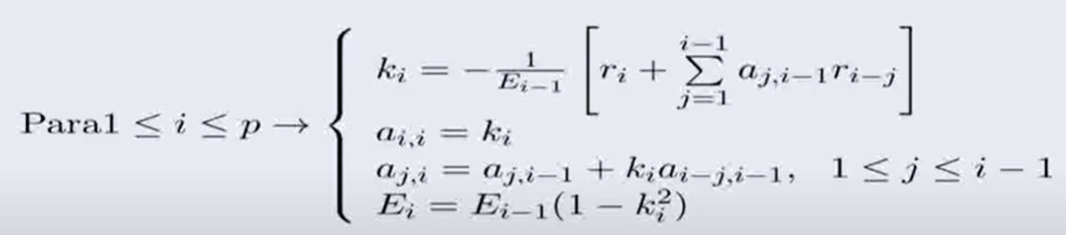
**Análisis por tramos.**

Tomamos ventanas de la señal idénticamente cero fuera de es una ventana, como por ejemplo la ventana de Hamming ya vista.

Estimamos: , esto es la **autocorrelación.**

*Solución:* Se resuelve hallando la solución de la ecuación que propone la función de autocorrelación.

Una método eficiente es el de Levinson-Durbin ya visto:



Donde los coeficientes {ai} se calculan recursivamente a partir de un conjunto de coeficientes auxiliares {ki] que pueden ser interpretados como los *coeficientes de reflexión* de un tubo acústico equivalente y el error de predicción *E* se inicializa a partir de *r0*.

**Escalas perceptuales.**

Se sabe que la forma en la que se percibe la frecuencia es altamente no lineal; por ejemplo, si nos muestran dos tonos puros de 10 Hz y 50 Hz, percibimos la diferencia, pero si nos muestran una señal de 2010 Hz y 2050 Hz, ya es mucho más difícil que percibamos la diferencia. Es decir que la resolución en frecuencia de nuestro sistema auditivo varía con la frecuencia de una manera no lineal. Esto hace que haya información frecuencial más importante que otra, a fin de discriminar los distintos fonemas.

Frecuencia y Pitch.

A menudo confundimos en la literatura, el pitch no es igual a la frecuencia fundamental. La frecuencia, intensidad y las propiedades espectrales de un sonido interactúan en formas muy complejas para dar una percepción de pitch que puede ser un reflejo muy pobre de la F0. El pitch percibido cambia la intensidad.

El pitch se refiere a un atributo perceptual del sonido, generalmente en una escala continua. La frecuencia es un atributo físico de las señales La American Standars Association define al pitch como aquel atributo de la sensación auditiva por la cual los sonidos pueden ordenarse según una escala musical.

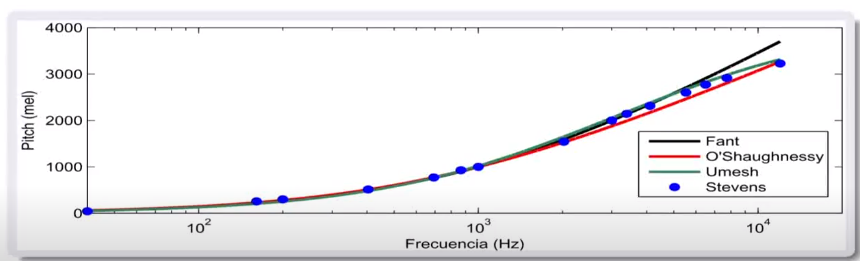
**Pitch: Percepción de la frecuencia fundamental.** Fenómeno totalmente subjetivo de como percibimos esa frecuencia.

Mel.

La unidad del pitch percibido de un tono puro es el **mel**. No se corresponde literalmente con la frecuencia física del tono. Stevens estableció que 1000 Hz = 1000 mel.

Pero años más tardes otros físicos establecieron más relaciones entre la frecuencia percibida y la física:

* **Fant (1973) (la más utilizada):**
* **O’Shaugnessy (1987):**
* **Umesh (1999):**



**Coeficientes ceptrales.**

Procesamiento Homomórfico.

**Sistemas homomórficos.**

Los sistemas homomórficos obedecen a un “principio de superposición generalizado”.

Decimos que un sistema es homomórfico para la convolución si obedece a un principio de superposición donde la adición se reemplaza por la convolución:

El sistema del aparato fonador puede verse como un sistema que cumple con este tipo de propiedad. A partir de allí se puede calcular algunos coeficientes basados en este principio.

La idea es tratar de convertir una convolución (supongamos un modelo lineal o fuente-filtro) por una suma que nos va a permitir separar las características de la fuente de las características del filtro.

Cepstrum.

**Cepstrum Real:** donde es la transformada discreta de Fourier de la señal x[n]: .

**Vocabulario:** Analogía entre el espectro y este nuevo dominio que surge de hacer la transformada y la inversa.

* Spectrum -> Cepstrum
* Frequency -> Quefrency (tiempo)
* Filter -> Lifter
* Phase -> Saphe
* Amplitude -> Gamnitude
* Harmonic -> Rahmonic
* Period -> Repiod

**Cepstrum Complejo:** donde es la transformada discreta de Fourier de la señal x[n] y .

La ventaja del cepstrum complejo es que se puede invertir. Los coeficientes ceptrales puedo convertirlos para volver a obtener la señal original. Mediante que en el cepstrum real esa parte se pierde.

La desventaja es que calcular un algoritmo complejo implica mayor costo computacional.

**Preguntas:**

* **¿Es posible revertir el proceso y volver a obtener x[n]?** Si, usando el cepstrum complejo
* **¿Cómo haría un liftrado?** Cuando hablábamos de filtrado era en el dominio de la frecuencia, ahora con el liftrado estamos en el dominio de la quefrencia, el cual consistiría en dejar algunas componentes y eliminar otras, pero en este dominio.
* **¿Que podríamos hallar con el liftrado?** Podemos separar las características que pertenecen al tracto vocal de las que son de las fuentes de excitación.
* **¿Elegiría el cepstrum real o el complejo?** Usaríamos el real si queremos una representación más rápida de calcular y tomaríamos el complejo cuando queremos invertir el proceso y tener la señal original.

Coeficientes de predicción lineal (LPC).

Estos dan cuenta de la relación que hay en los distintos instantes del tiempo con lo que sería un predictor lineal en la señal de voz.

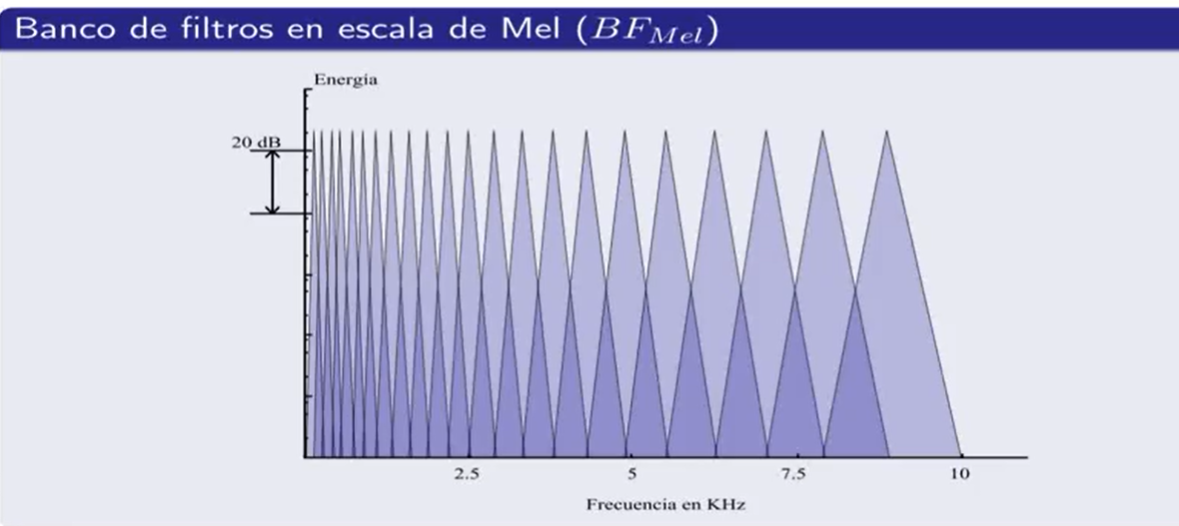
**Cálculo de cepstrum a partir de los LPC.**

* El cepstrum de una señal se calcula realizando alguna transformación frecuencial a partir del logaritmo del espectro.
* En el caso de LPC-Cepstrum el espectro utilizado es el estimado a partir de los coeficientes LPC que se puede obtener a partir de la FFT de los coeficientes.
* Sin embargo, se puede demostrar que es más eficiente calcularlo por la siguiente recursión:
* La principal ventaja de los coeficientes cepstrales es que generalmente son decorrelacionados.

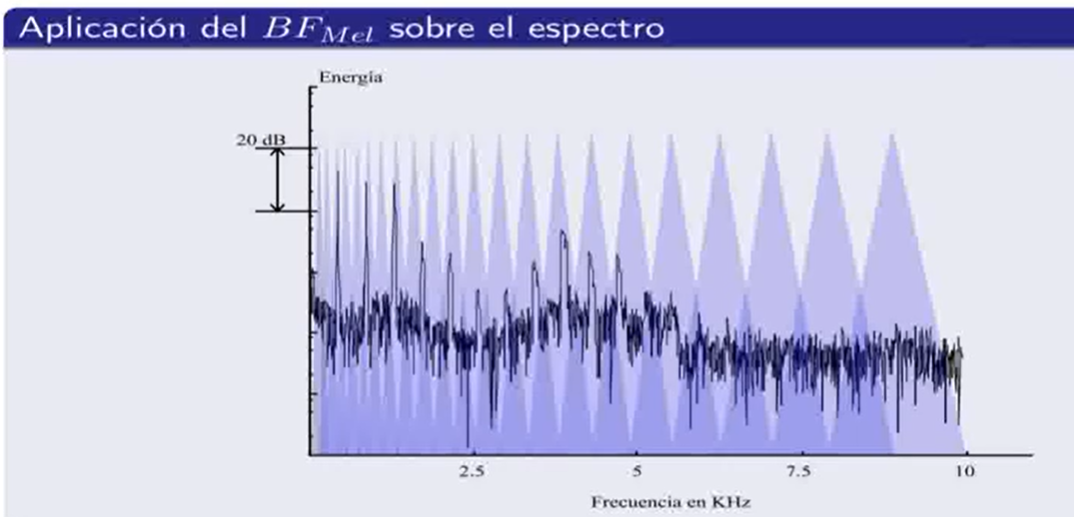
Mel Cepstrum.

Hasta acá hablamos de aplicar algún tipo de transformación frecuencial. Habíamos visto que la relación entre la frecuencia física y la percibida no es lineal. Entonces la idea sería **¿Para qué utilizar información de todo el rango de frecuencias de la misma forma si sabemos que hay frecuencias para las cuales nuestro oído es más sensible que para otras? Entonces podríamos tratar de incorporar sólo aquella información que permita discriminar entre los distintos fonemas tal cual lo hace nuestro sistema auditivo.**

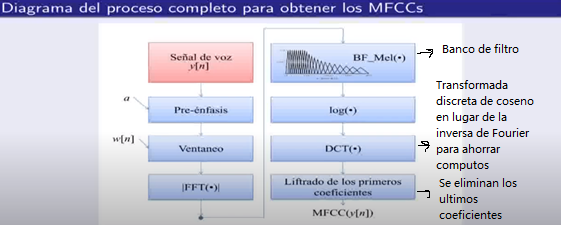
Para eso se utilizan **banco de filtros.** Éstos tratan de agrupar información en los rangos de frecuencias en los cuales el oído tiene menos discriminación.



Vemos cómo los filtros de las escalas más altas son más anchos y los de las escalas más bajas, mas angostos. Esto hace que la resolución en frecuencias bajas sea mucho mayor a la resolución en frecuencias altas, tal como lo hace el oido.



Acá vemos un esepctro en negro y superpuesto el banco de filtros. Lo que se hace es una sumatoria pesada por cada uno de los valores del filtro y termina como salida de cada uno de los elementos del banco de filtro un único coeficiente que sería una especie de espectro modificado psicoacústicamente.



MFCCs: Coeficientes cepstrales en escala de MEL. (Vector)

La idea es que en estos primeros coeficientes de Cepstrum nosotros tengamos la informació más importante respecto a la transferencia del tracto vocal, que es en definitiva lo que estaría dándonos una idea de qué tipo de fonema se pronunció en esa señal.

Deltas y aceleraciones.

Hasta ahora todo el análisis que se hace es a partir de ventanas. Es decir, para cada ventana, se calculan los MFCCs. Esta información es de tipo **estática**, o sea no guarda relación con qué es lo que pasó con una ventana respecto de la anterior, y sin embargo sabemos que hay una correlación muy importante porque el aparato de fonador tiene una velocidad de cambio lenta y entonces el espectro de una ventana es similar al espectro de la ventana siguiente (salvo que haya un cambio abrupto, por ejemplo, un cambio de fonema.

Para incorporar esta información que tiene que ver con la dinámica temporal, se utilizan unos coeficientes derivados de, por ejemplo, los coeficientes cepstrales, que se denominan **coeficientes delta y coeficientes de aceleración.** Para calcularlos, se agrega información de la derivada:

Por ejemplo, se calcula el Cepstrum en una de las ventanas y lo mismo en la ventana siguiente. Suponiendo que tenemos 13 coeficientes cepstrales en cada una, luego hacemos la diferencia entre los 2 coeficientes. Esto nos da otro vector de 13 coeficientes que se agregarían a los 13 originales, es decir que esta representación tendría un total de 26 coeficientes. Esto nos daría información del cambio que hubo entre las dos ventanas.

Además, esto se puede hacer también con una segunda derivada o *coeficientes de aceleración*, donde se ven 2 coeficientes delta sucesivos y se saca la diferencia.

**Información acerca de los cambios.**

* El desempeño de los sistemas de RAH mejora mucho agregando información acerca de las derivadas temporales de todos los análisis “estáticos” vistos (en cada tramo).
* Pra el análisis cepstral puede demostrarse que este tipo de información provee cierta robustez a cambios relacionados con el canal de transmisión o el tipo de micrófono empleado.
* Puede agregarse también en forma explícita información acerca de la energía de la señal (y su correspondiente cambio).
* Esto aumenta la cantidad total de dimensiones del patrón obtenido para cada tramo
* Los coeficientes “delta” son calculados utilizando la siguiente fórmula de regresión: donde es el coeficientes delta en el tiempo t calculado en términos de los coeficientes estático correspondientes a . El valor de se ajusta convenientemente (por ejemplo 3).
* La misma fórmula se aplica sobre los coeficientes delta para obtener los coeficientes de aceleración.
* En los bordes se replican el primer y el último tramo para poder calcular los coeficientes.

C son 13 coeficientes ceptrales, E es 1 coeficiente de energía que suele agregarse.

dt son 13 coeficientes delta y ddt son 14 de aceleración.

**Estimación de frecuencia fundamental F0 con el cepstrum.**

Consideremos un análisis por tramos o ventanas de una señal de vos x[n]. Sea G(ω) la TDF de la señal de excitación del tracto vocal y H(ω) la respuesta en frecuencia de dicho tracto. Sean cx[n], cg[n], y ch[n] sus cepstra respectivamente. Entonces cx[n] = cg[n] + ch[n]

cg[n] y ch[n] ocupan partes diferentes del eje de cuefrencias. Podemos separar la parte que varía rápidamente (correspondiente a la excitación del tracto vocal) de la que varía lentamente (la respuesta en frecuencia del tracto).

Otra técnica que también es utilizada es el método de autocorrelacion. El fundamento de la técnica de extracción de F0 basada en la autocorrelaccion es que si la señal es periódica, entonces su función de autocorrelaccion muestra un pico en el retardo correspondiente al periodo valiendo menos en todos los otros retardos (salvo en cero).

