Principios científicos del sistema VQR de reconstrucción de la calidad de voz

Ing Oscar Bonello, Solidyne
Fellow Audio Engineering Society,USA

INTRODUCCION

Las comunicaciones telefónicas son la forma más difundida de transporte de señales de audio. Su rango de frecuencias, sin embargo, es muy restringido. Al principio de su historia lo fue debido a las limitaciones técnicas de los micrófonos de carbón y los auriculares con diafragma metálico. A medida que los transductores mejoraron y la tecnología original devino en electrónica, las mejoras de sonido no incluyeron un aumento significativo de su respuesta a frecuencias. Esto fue por razones económicas, para reducir el ancho de banda necesario, un recurso escaso. Si bien está probado que para tener 100% de inteligibilidad es necesario llegar hasta 5.000 Hz (French-Steinberg,1947) la industria telefónica acepta un ligera pérdida de inteligibilidad para poder reducir este ancho de banda a 3.400 Hz y así poder manejar más cantidad de canales en el mismo ancho de banda.
Las frecuencias bajas (graves) también son cortadas por debajo de 300 Hz pues no agregan inteligibilidad a la palabra y simplifican el manejo de la información y el diseño electroacústico de los transductores. Asimismo eliminan la interferencia telefónica debido a maquinarias y ruidos de baja frecuencia.

¿Cómo influye esto en la transmisión de noticias para radiodifusión? Es evidente que el teléfono no fue diseñado como sistema de alta calidad de audio. Por lo tanto las voces transmitidas (aún cuando se ingresa directamente a la línea telefónica, sin pasar por los transductores), son doblemente limitadas en frecuencias, tanto en graves como en agudos. La transmisión de frecuencias entre 300-3.400 Hz tiene limitaciones muy importantes. La voz humana tiene frecuencias desde 80 Hz hasta 10.000 Hz. Por lo tanto se pierden dos octavas en graves y otras dos octavas en agudos. Esto produce una voz decididamente "nasal" (falta de graves) y a la vez totalmente falta de brillo por la carencia de agudos. Esto hace que muchas radios hayan invertido grandes sumas para realizar enlaces de radio montados en camiones de exteriores. Esto sin embargo les resta movilidad a los periodistas e impide que trabajen en forma rápida y los obliga a viajar seguidos de una costosa corte de técnicos.

A las limitaciones en la respuesta de frecuencias, se le suma el reducido rango dinámico de la telefonía. Recordemos que llamamos rango dinámico a la relación entre el sonido más fuerte y el ruido de fondo. El oído humano maneja en condiciones normales de ruido en la vida cotidiana, rangos dinámicos de 90 dB, siendo aceptables para radiodifusión de FM entre 70 y 80 dB. Las transmisiones telefónicas manejan rangos dinámicos de entre 40 dB y 50 dB en el mejor caso.

LAS SOLUCIONES EXISTENTES HOY

Este problema de transmitir audio por línea telefónica ha tenido muchos intentos de solución. Ninguno de ellos totalmente satisfactorio. Los analizaremos someramente:

a) Transladores múltiples de frecuencia Un equipo codificador divide la banda de audio en tres sub-bandas de 300 a 3.400 Hz y las transmite empleando tres líneas telefónicas fijas. En el otro extremo (Estudios) un equipo decodificador las desplaza y une nuevamente. Se obtiene un sonido aceptable (50-7.500 Hz) pero el alto costo y la dificultad de mantener tres líneas

simultáneas para cada evento, así como la imposibilidad de usarlo con telefonía celular, ha hecho caer el sistema en desuso.
b) Extensores de frecuencia El principio es parecido al anterior, pero usa una sola línea. Requiere un equipo codificador y otro decodificador en Estudios. El codificador en transmisión desplaza hacia arriba todas las frecuencias 250 Hz (o un valor similar) De esta manera una voz con graves en 100 Hz es transmitida en 250 + 100 = 350 Hz En los estudios, el decodificador hace el desplazamiento inverso corriendo hacia abajo 250 Hz. En este proceso recuperamos graves pero empeoramos los agudos en 250 Hz que se pierden.
Desventajas: Necesita usar dos equipos (coder + decoder) / No sirve para entrevistas telefónicas (en donde el entrevistado habla por su propio aparato) / No corrige los agudos, los empeora / No mejora el rango dinámico

c) Sistemas ISDN Digitalizan las transmisiones para enviarlas por el servicio telefónico de datos de alta velocidad, disponible en Europa. La calidad es excelente y esta es su ventaja principal.
Desventajas: Solamente es usable en Europa y muy pocos lugares de mundo, en ciudades importantes / No es 100 % portable pues requiere línea física /Alto costo en equipos y por hora de transmisión / Necesita usar dos equipos / No sirve para entrevistas telefónicas (en donde el entrevistado habla por su propio aparato)

d) Sistemas de CODEC Digital En esta categoría entra el Musicam, Patriot Tieline, TELOS, etc. Se trata de un sistema codificador que convierte la señal a digital, la comprime MPEG y la transmite a través de un Modem telefónico, por una línea telefónica fija o con un Modem GSM a través de la red celular. En estudios es reconvertida. La respuesta a frecuencias obtenida es buena, así como la calidad de sonido. Posee *latencia*, es decir un retardo de sonido que en algunos casos (como entrevistas remotas realizadas desde los estudios) puede ser molesta.
En resumen:
Desventajas: / No es 100% portable pues requiere línea física /Alto costo en equipos / Generalmente requiere usar otro equipo en estudios para decodificar / Tiene retardo, lo que impide entrevistas desde Estudios.

e) Sistemas de transmisión DUAL En esta categoría entra el Codec MB2400 de Solidyne que transmite simultáneamente un streaming MP3 y otro por celular para procesado VQR, eliminando las desventajas del Codec Digital convencional

Comparemos con el nuevo sistema VQR

El VQR no requiere un equipo codificador. De hecho mejora la calidad de las entrevistas en que el entrevistado usa su propio teléfono de línea o celular. El VQR no tiene retardo. Reconstruye frecuencias entre 50 Hz – 10.000 Hz. Mejora los graves y los agudos. Aumenta el rango dinámico a 70 dBA (similar a un buen micrófono de estudio) El 95 % de los oyentes creerá que el entrevistado está en el Estudio de la Radio. Es de bajo costo.
Desventajas: La calidad de sonido es buena, pero no perfecta como en el CODEC Digital. Recomendamos escuchar demos en Internet.

LA TECNOLOGIA VQR (Voice Quality Restoration)

La comparación que acabamos de realizar entre VQR y otras soluciones, parecería apuntar a que el VQR es una especie de *solución mágica*. Pero por supuesto esto no es así, pues está basado en sólidos principios científicos.

Solidyne ha trabajado en el tema de la transmisión de exteriores por vía telefónica durante 35 años. Hemos analizado numerosas técnicas matemáticas que permitían resolver este problema. Se construyeron incluso prototipos y se realizaron numerosos ensayos. Pero no nos conformaron.

Nuestras soluciones, aún siendo de menor costo, tenían los mismos problemas ya antes señalados y por eso nunca fueron fabricadas.

En los últimos dos años encaramos el problema desde un punto de vista radicalmente diferente. El punto de vista Psicoacústico y del análisis de los mecanismos de generación de la voz humana. Lo hicimos alentados por el éxito que esta misma idea había tenido en 1988 cuando logramos el primer sistema del mundo de grabación de música y comerciales en Hard Disk, basado en la invención de la compresión de datos PCM (bit compression) usando la teoría psicoacústica del Enmascaramiento de Bandas Críticas. Esto nos permitió construir el primer sistema del mundo para compresión de datos PCM que fue precursor del MPEG, MP3, ATRAC y de todos los otros sistemas hoy en uso por millones de personas en todo el mundo. Pero en 1988 parecía magia.

El sistema VQR, inventado por Solidyne en 2005, opera exclusivamente sobre la voz humana. Veamos entonces qué sabemos acerca de la generación de voz. En la Fig 1 vemos el sistema de fonación

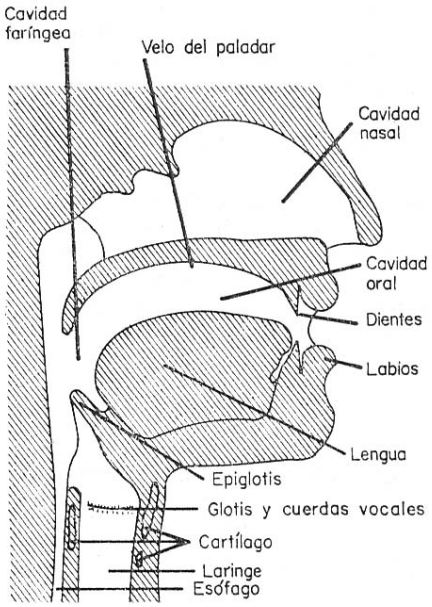

Figura-1

Podemos observar que las cuerdas vocales están al comienzo de toda la cadena de generación, produciendo un sonido fundamental o *pitch* del cual depende el tono de voz de la persona que habla. Observemos que el sonido atraviesa tres cavidades (faríngea, oral y nasal) que se comportan como tres resonadores alterando las componentes armónicas del sonido de las cuerdas vocales que es muy rico en armónicos por ser una onda de tipo triangular asimétrica, con armónicos pares e impares. El movimiento de las cuerdas vocales puede verse en la secuencia de la Fig-2

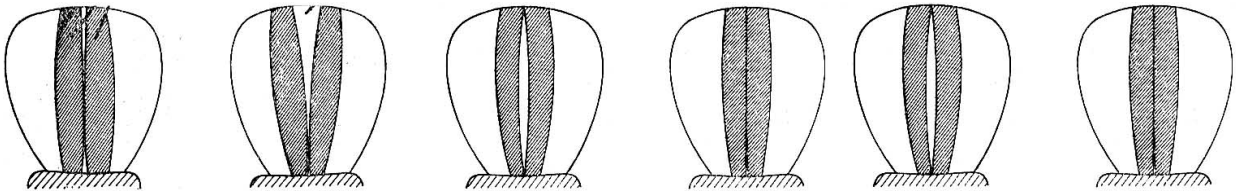

Figura 2 Movimiento de las cuerdas vocales (M.Guirao, 1980)

La acción acústica de las cuerdas vocales ha sido simulada en laboratorio con dos masas conectadas por un resorte (Fant 1970, Veldhuis 1995) Es evidente de la Fig 1 que los ingenieros podemos simular las cavidades con resonadores tubulares, imitando de esta forma el tracto vocal (O'Saughnessy 1987). Veamos por ejemplo la Fig-3

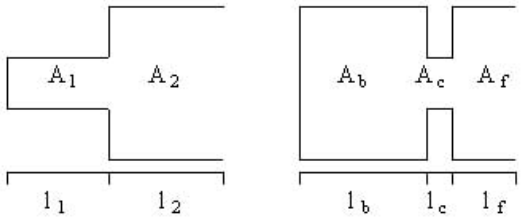

Figura 3 Modelo de tracto vocal con dos y con tres tubos

El sistema de glotis (generación de *pitch*) y las tres cavidades se comportan como un circuito equivalente serie. Esta hipótesis es hoy plenamente aceptada (McAulay 1983, Macon 1996, Klijn 1998)

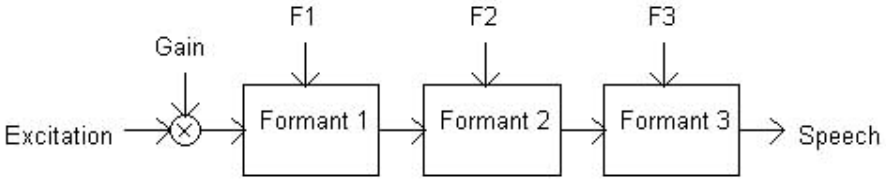

Figura 4 Circuito equivalente de las cuerdas vocales y el tracto vocal

Cada uno de los filtros (son del tipo pasa-banda de segundo orden) es rápidamente variado por los movimientos de la lengua, produciendo distintas resonancias que son denominadas *formantes de la voz.* Esto se define en la ya clásica ecuación de Harvey Fletcher (Allen, 1995) que define la relación entre la presión sonora para un armónico $\mathbf{P_k}$ con relación a la fundamental $\mathbf{P_1}$

$$\frac{P_k}{P_1} = \frac{F_k}{F_1}\sqrt{\frac{\left(\frac{\Delta}{\pi f_1}\right)^2 + \left[1 - \left(\frac{f_0}{f_1}\right)^2\right]^2}{\left(\frac{\Delta}{\pi f_k}\right)^2 + \left[R - \frac{1}{R}\left(\frac{f_0}{f_R}\right)^2\right]^2}}$$

Siendo delta el factor de amortiguamiento

Veamos ahora cuál es el espectro de audio que el mecanismo de fonación de la voz humana presenta (Fant (1960), Flanagan (1965), Fry (1979), Lieberman & Blumstein (1988).
Analicemos la Fig 5

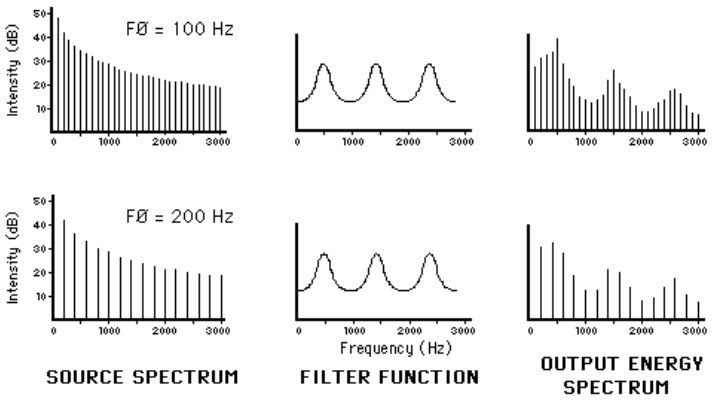

Figura 5 Espectros de audio para *pitch* de 100 y 200 Hz

Vemos en el primer gráfico el sonido de las cuerdas vocales exclusivamente, para el caso de una fundamental de 100 Hz (voz muy grave) y otra de 200 Hz (voz casi femenina). Es interesante observar que existen numerosas armónicas que se extienden hasta más de 3.000 Hz
En el segundo gráfico vemos los tres filtros correspondientes a las tres cavidades, o formantes. Finalmente en el tercer gráfico vemos el resultado final emitido por la voz humana. Contrario a lo que podríamos pensar se trata de un ***espectro discreto*** y no de un espectro continuo. Es decir que está formado por frecuencias individuales espaciadas.
Destacamos este detalle pues es la base de la tecnología VQR para restauración de graves.

Veamos entonces este fenómeno con mayor detalle. Grafiquemos cada armónica como una barra vertical y tendremos la Fig 6 y la Fig 7

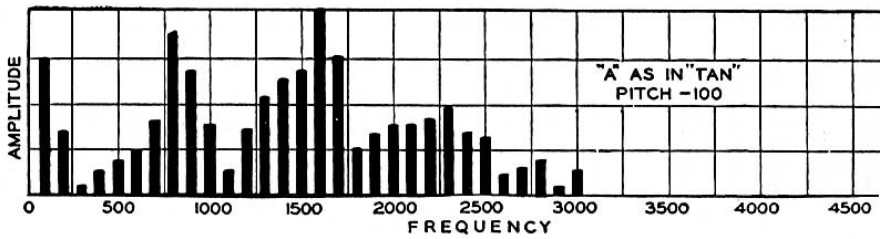

Figura 6; armónicos de la "A" inglesa con voz grave (pitch = 100)

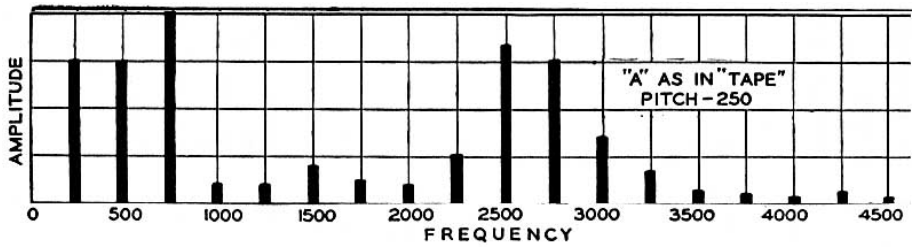

Figura 7; armónicos de la "A" inglesa con voz aguda (pitch = 250)

Aquí es observado con mucho detalle el carácter ***discreto*** del espectro de audio. Esto significa que aún cortando la parte inferior del espectro, debajo de 300 Hz, las fundamentales perdidas pueden en teoría reconstruirse a partir del resto del espectro. Imaginemos que un observador examine ambos espectros en la zona entre 1.000 y 2.000 Hz y no conozca los armónicos por debajo de 300 Hz. ¿Podrá nuestro observador saber cual es la fundamental? Es evidente que puede, **pues midiendo la separación entre dos armónicos consecutivos conoce perfectamente el valor de la fundamental** (en nuestro ejemplo, 100 y 250 HZ).

En esto se basa la sección del VQR que reconstruye los tonos graves de la voz humana, aún escuchados a través del micrófono de un teléfono.

No entraremos en detalles técnicos acerca de la manera de hacer esta operación. Uno de los clásicos métodos de análisis puede verse en la Fig 8

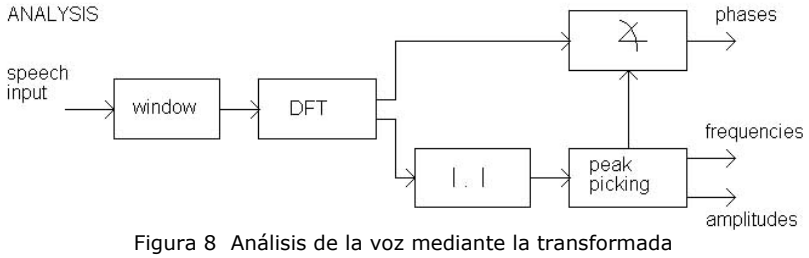

Figura 8 Análisis de la voz mediante la transformada

RECONSTRUCCION DE LOS AGUDOS DE LA VOZ
Y DEL RANGO DINAMICO

Tal como hemos visto, los registros más graves de la voz humana pueden reconstruirse a partir de las componentes discretas que atraviesan el canal telefónico. Este es un notable descubrimiento que cambia nuestras perspectivas acerca de la calidad de una transmisión por vía telefónica.

La reconstrucción de los agudos también es posible, pero menos exacta que la de los graves. Esto es debido a que existen dos mecanismos de generación de agudos. Uno de ellos está dado por las armónicas elevadas de los sonidos *vocalizados*, es decir producidos por emisión de voz mediante las cuerdas vocales. Pero otro grupo muy importante es producido por los sonidos no-vocalizados *o fricativos*. Estos son generados cuando el tracto vocal está cerrado parcialmente en algún lugar y el aire empuja con fuerza en ese mismo lugar produciendo una turbulencia. Ejemplo de sonidos fricativos son las consonantes **f**, **s** y **j**
Afortunadamente esos sonidos tienen su fundamental entre 2.000 y 3.000 Hz por lo que atraviesan los sistemas telefónicos. Pero sus armónicos que llegan hasta 10.000 Hz no pueden hacerlo, perdiendo el brillo característico de la voz grabada en alta calidad.
Sin embargo, un detenido estudio de las componentes armónicas superiores de la voz humana indica que si excitamos con las componentes de la banda de 2 a 3 KHz a un generador armónico basado en ondas triangulares asimétricas (del tipo de las generadas por las cuerdas vocales), se obtiene un espectro armónico bastante similar al sonido real. Si a este espectro se le eliminan las componentes de excitación (banda de 2 a 3 KHz), se obtiene una *banda alta complementaria* que puede sumarse al sonido original para lograr la reconstrucción de los agudos de una forma parecida a la que empleamos con los graves.

Para poder hacer que la calidad de voz transmitida telefónicamente sea lo más similar posible a la calidad que obtendríamos en estudios, es necesario aumentar el rango dinámico desde los 40/50 dBA de las comunicaciones telefónicas hasta unos 70 dBA que es lo que obtenemos en estudio con un buen micrófono. El principio con que actuamos es conocido, pues es el mismo que se emplea en los estudios de grabación de audio digital para lograr que la mezcla de numerosos micrófonos que toman a los instrumentos musicales no sumen sus niveles de ruido de fondo, impidiendo obtener los necesarios 80 a 90 dB de rango dinámico que caracterizan a un buen CD.

La solución es de tipo psicoacústico (¡nuevamente!) pues está basada en el enmascaramiento temporal que una señal intensa produce en el oído humano. El dispositivo que permite esto es denominado *Expansor de Audio*. Puede verse en la Fig 9 la curva de transferencia de un Expansor

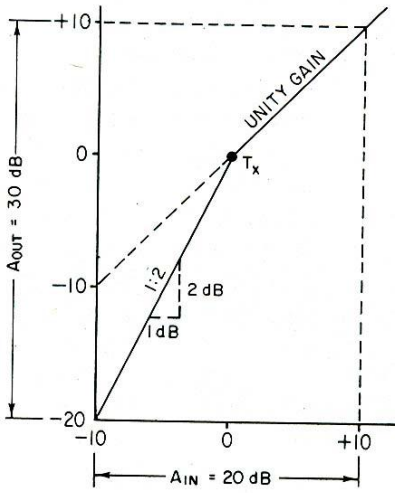

Figura 9 Diagrama de Entrada-Salida de un Expansor de audio

Se trata de un expansor de pendiente 1:2 Este dispositivo, absolutamente lineal para las frecuencias de audio, varía su ganancia en función del nivel de señal. Para ser efectivo y cumplir con las condiciones psicoacústicas, el ajuste de ganancia frente a una señal que aumenta su nivel debe ser muy veloz, típicamente de menos de 5 milisegundos. Cuando la señal disminuye su nivel el tiempo de respuesta deberá ser del orden de 50 a 100 mS

Observemos su acción. En su entrada tenemos una señal que varía su nivel 20 dB. A la salida la misma señal presenta una variación de 30 dB, es decir ha *incrementado su rango dinámico*.

Este mismo principio es aplicado en la tecnología VQR para reducir el ruido de fondo de la comunicación telefónica, para obtener una señal de audio muy limpia y con su espectro balanceado.