UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE MUSICA

Síntesis de voz cantada por concatenación de sílabas en español, utilizando el algoritmo TD-PSLA

T E S I S

Que para obtener el grado de:

Maestro en Música

En el campo de

Tecnología musical

Presenta:

Alejandro Ramos Amézquita

ENM
UNAM

Asesor: Dr. Abel Herrera Camacho

"Por mi raza hablará el espiritú"

México, D. F. Abril de 2010

A mi mamá,

la persona más avanzada espiritualmente,

que conozco.

A mi papá,

quien ha aprendido a confiar en ambos,

y que rejuvenece con mi propia juventud.

A mis hermanos:

Sandra, Alberto, Rodrigo y Natalia;

fuentes inagotables de motivación para ser mejor.

A Guadalupe,

con quien un presente de paz y felicidad verdadera,

es posible.

Agradecimientos

Agradezco especialmente al Dr. Felipe Orduña Bustamante, por tomar como propia la empresa de otros.

Agradezco al Dr. Pablo Padilla Longoria, por siempre tratarme como colega y amigo, antes que como subalterno. Su confianza se refleja en mi crecimiento.

Agradezco a la Mtra. Gabriela Pérez Acosta, por un pasado y futuro de risas y trabajo.

Agradezco al señor Director, Mtro. Mijael Gutiérrez López por sus continuas muestras de solidaridad.

Agradezco al Mtro. Jordi Iñaki Senosiain, por creer en mi con tal libertad.

Agradezco al Dr. Damián Hernández Herrán, por continuar a mi lado, a pesar de mi mismo.

Agradezco al Dr. Aurelio Campos R., y al Lic. Octavio Sierra R. por enseñarme que el pasado y el futuro no existen.

A mis alumnos Alux, Tona, Victor Hugo, Rodrigo, Tere, Samuel, Jerónimo, José, y tantos otros, por haber sido a su vez, mis maestros.

Agradezco al Dr. Eduardo Castro-Sierra, por mostrarme que nada es permanente.

Agradezco a mi tutor y sinodales.

Dedico todo mérito que se haya generado mediante éste trabajo, para el beneficio y felicidad duradera de aquellos que me rodean; para que la mente altruista aparezca donde no lo ha hecho y se incremente donde ya exista. Dedico especialmente el mérito, para que la UNAM pueda una vez más, resurgir de las cenizas y ayudar a éste país a transformar la crisis en oportunidad.

Índice

Resumen

La presente tesis para obtener el grado de Maestría en Tecnología Musical cumple con dos objetivos primordiales. El primero de ellos, es presentar un documento comprensivo sobre la historia y desarrollo de la Sintesis de Voz, desde aquella que es hablada hasta la cantada en Español. Asimismo, se presenta el resultado de la modificación de un programa de síntesis de voz basado en la aplicación de un algoritmo tipo PSOLA a difonemas, a uno aplicado a sílabas pregrabadas en español, en la plataforma de programación Matlab. Como productos de la tesis, se incluye un análisis a fondo y comentario del programa de síntesis de voz hablada en español perteneciente y desarrollado por la UNAM (Matlab), así como resultados de la evaluación cualitativa de la aplicación del algoritmo TD-PSOLA a grabaciones de vocales y sílabas grabadas por una cantante soprano, para modificar exclusivamente *tono* y *duración*, sistemáticamente, a una reducida base de datos, de acuerdo a una partitura tradicional de voz cantada. Para la generación de la base de datos, se eligió, para fines de esta tesis, la voz de una soprano (Guadalupe Caro Cocotle), y las sílabas fueron grabadas en dos tonos (*F4* y *C5*), y elegidas en base a dos canciones populares.

Introducción

El presente trabajo constituye apenas uno de los eslabones iniciales en la investigación sobre la síntesis de voz cantada en español en la UNAM. Es, hasta donde el autor conoce, la primera aplicación de algoritmos tipo PSOLA a un sistema de Síntesis de Voz Cantada por concatenación de segmentos pre-grabados (sílabas en Español mexicano), y es el resultado de la modificación de un sintetizador de voz hablada por concatenación de difonemas, originalmente escrito en lenguaje C y más tarde implementado en Matlab. Se escogió esta última como plataforma de desarrollo, debido a la experiencia que el autor tenía con la misma.

El trabajo presenta objetivos prácticos modestos, pero esenciales:

- Un análisis a fondo y comentario de un programa de síntesis de voz hablada en español perteneciente y desarrollado por la UNAM (Matlab).[1]

- Resultados de la evaluación cualitativa de la aplicación del algoritmo TD-PSOLA a grabaciones de vocales y sílabas grabadas por una cantante soprano, para modificar exclusivamente *tono* y *duración*, sistemáticamente, a una reducida base de datos, de acuerdo a una partitura tradicional de voz cantada.
Para la generación de la base de datos, se eligió, para fines de esta tesis, la voz de una soprano (Guadalupe Caro Cocotle),[2] y las sílabas fueron grabadas en dos tonos (*F4* y *C5*), y elegidas en base a dos canciones populares (*Las mañanitas* y *Oda a la Alegría*).

Para lograr una síntesis efectiva de voz se debe tener una comprensión fundamental de la física de la producción de voz, de las constricciones linguísticas que caracterizan un lenguaje dado, y de cómo implementar tal conocimiento. Dado que el presente esfuerzo representa un primer acercamiento al tema, es importante introducir al objeto de estudio desde todos los ángulos posibles. Así, esta tesis presenta también, los antecedentes de la investigación en síntesis de voz cantada. En términos de estructura se encuentra dividida en 6 partes y dos apéndices; el *Capítulo I,* presenta una revisión de la historia y las técnicas más utilizadas para sintetizar voz cantada a lo largo de la historia, así como ciertos fenómenos interesantes de las voces de cantantes. Se presenta de manera relajada y poco profunda con el objetivo de "empapar" al lector del universo de la voz cantada y su síntesis. El *Capítulo II* es de naturaleza introductoria, que presenta los fundamentos del lenguaje, la fisiología y acústica del tracto vocal, y la fonética del Español. El *Capítulo III*, constituye también un capítulo introductorio, pero relativo a temas de Procesamiento Digital de Señal. Los *Capítulos IV* y *V*, son propiamente el cuerpo de la tesis, ya que en ellos se presentan, además de una revisión a profundidad de los métodos de síntesis, y aspectos técnicos sobre Prosodia y su modificación (sobre lo que se experimentó en esta tesis), una descripción del sistema que, siendo laxos en la definición de "canto", podemos llamar de voz cantada. Finalmente se presentan los *Análisis de Resultados* y *Conclusiones* del trabajo. Es muy probable que algunos temas, e incluso capítulos enteros, se antojen alejados del lenguaje y área del conocimiento que cada sinodal maneja individualmente. Empero, todos y cada uno de ellos, son constitutivos globales de una investigación y desarrollo tecnológico emergente; en opinión del autor, se debe poseer algún

---

1 Para ello, se ha trabajado sobre el sistema de voz hablada en español, desarrollado en Matlab por Fernando del Río como proyecto de maestría en Ingeniería Eléctrica, bajo la dirección del Dr. Abel Herrera Camacho, del posgrado de ingeniería de la UNAM.

2 Guadalupe Caro Cocotle concluyó los estudios de Licenciatura en Música, con especialidad en voz, el año del 2004 en la Universidad de Manitoba, Canadá.

nivel de conocimiento en todos ellos, si se desea incursionar en el desarrollo de la *Síntesis de Voz Cantada en Español*.

Debe mencionarse, que como primer acercamiento a la síntesis de voz cantada, nunca se pretendió desarrollar un sistema exhaustivo o muy flexible (y mucho menos que sonara natural). Para lograr los objetivos modestos ya mencionados, se debía estudiar a fondo el sintetizador de voz hablada y comprender su funcionamiento. Ello representó el aspecto más costoso en términos de tiempo dedicado, y se refleja en el comentario del Sistema de Voz Hablada, (que originalmente contaba con aproximadamente mil líneas de programación, pero sólo cuatro de ellas estaban comentadas), que se incluye en los *Apéndices I* y *II*.

Las unidades básicas de concatenación, escogidas para el Sistema de Voz Cantada fueron las *sílabas*, ya que se sostiene la hipótesis de que, mientras los difonemas han demostrado funcionar como las unidades con un mejor compromiso entre inteligibilidad y flexibilidad del lenguaje hablado para sistemas de síntesis de voz hablada, correspondientemente deberán ser las sílabas, las unidades básicas para lograr sintetizar Voz Cantada en Español, esto considerando las sólidas reglas lingüísticas para su formación, el hecho de que incluyen por completo el dominio de la co-articulación y reducen los problemas de borde; ventajas de la que carecen otros idiomas, como el inglés, cuyas reglas de silabización son mucho más ambiguas. La hipótesis de trabajo, obliga la pregunta, de si las unidades básicas del canto (en Español), no son de hecho las sílabas. El hecho de que toda partitura vocal lírica (en español), contiene una línea melódica donde hay un correspondencia uno a uno entre nota (*tono* y *duración*) y sílaba (excepto en los *melismas*), es un buen indicativo.

El punto clave de unión de las investigaciones en síntesis de voz hablada con aquellas de la voz cantada, se llama *Prosodia*. La prosodia de la voz hablada y de la voz cantada, incluyen un gran número de coincidencias en cuanto a vocabulario, siendo en ambos casos, prominentemente de índole musical. El estrés, ya sea en notas (nivel dinámico) o debido a acentuación; el ritmo y la métrica, la duración y el tono, son todos temas de interés tanto para la voz hablada como la voz cantada. Una melodía asume su carácter, de su estructura rítmica, su contorno, su construcción tonal y su contenido interválico, por lo que para generar voz cantada, la correcta manipulación del *tono* y la *duración*, es escencial, y la manipulación de la amplitud y otros controles, es secundaria. Afortunadamente, es posible alterar el *tono* y *duración* en el dominio del tiempo, por separado.

Los resultados obtenidos muestran que puede lograrse una voz cantada en español bastante razonable en cuanto a lo que la *inteligibilidad* se refiere, con un mínimo de requerimientos y control. Lo anterior sugiere como un acierto el haber elegido la *sílaba* como la unidad fonética básica; sin embargo, la naturalidad de la voz resultante es, sin lugar a dudas, comprometida, como era predecible de la aplicación de un algoritmo de modificación del dominio del tiempo. Aún así, la severidad en esta alteración, varía dramáticamente de intervalo a intervalo, ya sea en frecuencia, o en duración, así como por la ejecución misma en las grabaciones de la base de datos. Así pues, la diferencia en naturalidad entre un segmento grabado con vibrato, comparado con otro sin vibrato es notoria, haciendo muy variable el intervalo en tiempo máximo que un segmento pueda ser variado sin agregar ruido. Se encontró que hasta una transposición de segmento de una cuarta justa es posible sin agregar errores digitales muy notorios. Ningún mecanismo de articulación entre segmentos modificados fue implementado, que son disparados a intervalos regulares. El escucha puede reconocer fácilmente la tonada que el sintetizador reproduce, lo cual es en sí mismo un resultado positivo,

y una vez más, confirma el acierto de la elección de las sílabas, reafirmando que se está en un buen camino para generar un sistema de síntesis de voz cantada en español, con un vocabulario ilimitado. Sin duda, una labor más minuciosa en cuanto al mecanismo de concatenación de los segmentos (es decir, un mecanismo para agregar o disminuir silencios inter-sílabas) y una investigación más amplia en cuanto a las combinaciones posibles de variabilidad de las marcas consecutivas de periodo, mejorarían extraordinariamente la naturalidad de la ejecución. Asimismo, una combinación de técnicas de síntesis, permitirían más opciones de control y las consecuentes mejoras en la naturalidad. Sin embargo, se debe mantener en mente que tales mejoras escapan de los objetivos de esta tesis.

Capítulo I

Antecedentes de la Síntesis de Voz Cantada en Español

Dentro de la rama de los sistemas de comunicación, la depuración de algoritmos de codificación y el subsiguiente desarrollo de la electrónica especializada en este campo, ha traído como consecuencia un incremento de las capacidades en las arquitecturas de redes de comunicación. Una de las mayores ventajas que tal desarrollo presenta en la actualidad, es la posibilidad de incrementar la comprensión de fenómenos de naturaleza física de altísima complejidad, como lo puede ser la voz humana, a través de los aprendizajes que el desarrollo de tales arquitecturas permiten. Nuestra habilidad para manejar y analizar la enorme cantidad de información que implica cualquier lenguaje humano, se debe en gran medida a la "superación" de las complicaciones técnicas computacionales de capacidad de almacenaje y velocidad de procesamiento. Por ello resulta ahora factible no sólo el análisis a mayor profundidad de los mecanismos que brindan naturalidad al lenguaje fonético humano, sino también el desarrollo de aplicaciones en las cuales podamos "imitar" tales mecanismos a nuestra conveniencia.

Las diferencias esenciales entre la voz cantada y la voz hablada aún no son comprendidas en su totalidad y existen razones para pensar que difieren en aspectos básicos como podría ser la unidad fonética base. Por otro lado, al distinguirse los distintos idiomas humanos tanto en fonemas, grafías, pronunciación y reglas ortográficas, debe desarrollarse un conocimiento de las características acústicas y fonéticas de los diferentes idiomas individualmente. Es decir un sistema de síntesis de voz, cantada o hablada en español, diferirá sustancialmente de uno en inglés y constituirá por tanto una investigación separada.

La meta que es común entre quienes desarrollan aplicaciones de voz cantada parece ser el lograr motores de síntesis que puedan sonar tan natural y expresivamente como un cantante real, y cuyas entradas puedan ser únicamente: la partitura y la letra de la canción. Aunque algunos esfuerzos orientados a la síntesis de voz cantada existen en la literatura[1], ésta continúa siendo un campo abierto para la exploración. La arquitectura generalmente propuesta para un sistema de estas características, incluye una generalización de la partitura tradicional, que puede incluir cualquier información simbólica requerida para el control del sintetizador; una sección destinada a convertir los controles de entrada en acciones de interpretación de bajo nivel, otra para crear las trayectorias paramétricas que expresen apropiadamente los caminos dentro del *espacio sonoro* del instrumento, y un módulo que contenga el motor de síntesis que produce la señal de salida concatenando una secuencia de muestras transformadas que aproximen la trayectoria de ejecución. La base de datos sobre la que se trabaja no sólo incluye las grabaciones interpretativas sino también modelos y mediciones que se relacionan con el espacio interpretativo y que brindan información relevante para el proceso de conversión de una representación de alto nivel (partitura) al sonido de salida (Bonada y Serra: 2007, p. 68-69).

---

1 Un ejemplo puede encontrarse en la literatura bajo el nombre de "Burcas", un sistema de síntesis de MIDI a voz cantada por concatenación simple aplicado al idioma sueco, desarrollado en el departamento de lingüística y fonética de la universidad de Lund, por Marcus Uneson.

Breve Historia de la Síntesis de Voz Cantada

La *Síntesis de Voz Cantada,* emana de campos bien investigados: la codificación y síntesis del habla, y la síntesis de instrumentos musicales. La *Síntesis de Voz* fue estudiada desde finales del siglo XVIII (1791), cuando Wolfwang von Kempelen construyera su sintetizador mecánico.[2] El primer sintetizador eléctrico bien conocido fue creado por Leon Theremin en la década de los veinte, mientras que el primer *Sintetizador de Voz Cantada* fue el Vocoder (VOice Coder) creado por Homer Dudley (1939), para el que, moviendo parámetros de los canales era posible producir voz cantada. En otro esfuerzo temprano, en los tardíos 1950s, Kelly y Lochbaum hicieron un modelo de tubo acústico del tracto vocal que podía producir voz cantada (1962) (Siivola: 2002, p. 6)

Ejemplos de *Sistemas de Síntesis de Voz Cantada* existen y su desarrollo está ligado, como ha sido señalado, a la evolución de la síntesis de sonidos, rama en la cual una de las metodologías más "exitosas" ha sido sin duda la basada en lo que se denomina *muestreo* (a veces denominado "*sampleo*")[3], o que son el resultado de concatenar secuencialmente muestras del *corpus* de una base de datos. Con todo rigor, ello no constituye una técnica de síntesis pero desde una perspectiva práctica resulta conveniente tratarla como un tipo de *Modelo de Síntesis*. El éxito del método radica en su sencillez, y del hecho de que captura la naturalidad del sonido al extraer los sonidos de su contraparte *real.* El método ha sido aplicado para reproducir prácticamente todo tipo de sonidos, resultando particularmente exitoso con instrumentos que tienen controles de excitación discretos, como percusiones o instrumentos de teclado. Sin embargo, la falta de flexibilidad y expresión son sus dos más grandes problemas lo que ha significado que aún no se cuente con el nivel de calidad que un músico profesional espera que presente un instrumento. Para estos instrumentos, es factible llegar a un nivel aceptable de calidad utilizando grandes bases de datos y *muestreando* una porción suficiente del *espacio sonoro* producido por un instrumento dado. Ello es mucho más difícil para el caso de instrumentos continuamente excitados. Para estos instrumentos, los parámetros de control son numerosos y tienen muchas maneras de *atacar, articular,* o *tocar* cada nota. Los parámetros de control cambian constantemente y el *espacio sonoro* cubierto por un intérprete puede ser considerado mucho más amplio que para instrumentos discretamente excitados (Bonada y Serra: 2007, p. 67-68).

Los ejemplos de Sistemas de Síntesis de Voz hablada que pueden encontrarse de código abierto no son pocos.[4] De los encontrados que manejan el Español, o que están basados en el tipo de interpolación MBROLA[5], muchos están destinados a trabajar con difonemas como unidad de concatenación del lenguaje, y por supuesto, están siempre orientados a la voz hablada. Resulta redundante, pero categórico, mencionar que la mayoría de los sistemas de síntesis de voz que se pueden encontrar están orientados a la voz hablada, en idiomas diferentes al Español aplicando métodos distintos a los planteados en ésta tesis (que incluyen diferencias en los algoritmos y en las bases de datos), o sencillamente no se pueden encontrar sus códigos de manera abierta.

Modelando la voz cantada

---

2 Una revisión detallada sobre los sintetizadores de voz se presenta en el Capítulo II.

3 Que proviene del término comúnmente utilizado: *"samplers"*.

4 Festival, CSLU Speech Toolkit, ModelTaker, MBROLA, EPOS, SAM.

5 Véase Capítulo III

Dependiendo de dónde se focaliza la atención, los modelos de síntesis pueden ser clasificados en modelos espectrales y modelos físicos. Los primeros se basan fundamentalmente en mecanismos de percepción del escucha; mientras los segundos se basan en modelar los mecanismos de producción de las fuentes de sonido. El beneficio de utilizar estos últimos es que los parámetros utilizados en el modelo están relacionados de manera cercana a los que un cantante utiliza para controlar su propio sistema vocal. En consecuencia, algún conocimiento de los mecanismos del mundo real debe ser integrado en el diseño. Sin embargo, tales sistemas a menudo presentan un gran número de parámetros, y el *mapeo* de esos controles intuitivos de los mecanismos de producción a la salida del modelo no es tarea trivial. Los controles están relacionados a los elementos del aparato vocal (apertura de quijada, forma de la lengua, tensión de pliegues vocales, etc.). Los modelos físicos están evolucionando rápidamente tornándose más y más sofisticados y proveyendo mayor control y realismo; mientras que las configuraciones físicas de diferentes órganos durante la producción vocal están siendo estimadas con gran detalle combinando diferentes acercamientos (fMRI, CT, EEG, entre otros) (Bonada y Serra: 2007, p. 70).

Existe también una amplia variedad de modelos pseudofísicos en los cuales el modelo se descompone en una *fuente* y el tracto vocal. Un ejemplo clásico de ello es la predicción lineal,[6] donde las resonancias del tracto vocal son modeladas como *polos* de un filtro, y el *error residual* se considera como una *señal*. El problema de este acercamiento es que la modificación del filtro no produce los resultados esperados, ya que en la realidad, hay mucho más que sólo la excitación glotal en la señal fuente. Esto es, parcialmente, resultado de las no linealidades del tracto vocal que este tipo de método lineal no puede modelar. Es probablemente más correcto clasificar la predicción lineal como un método espectral, ya que es común en estos que intenten modelar el *espectro* producido por el cantante. Por otro lado, los Vocoders dividen el espectro del habla en canales para los cuales se aproximan parámetros de *ganancia y fuente.* Asimismo, los sintetizadores por formantes tienen un parecido cercano a la predicción lineal por tener opción a dos fuentes distintas: *sonoras y sordas,* y el tracto vocal es modelado como un conjunto de filtros formantes (Siivola: 2002, p. 6). Las funciones de onda formantes y la frecuencia modulada, constituyen métodos espectrales ligeramente distintos. Las funciones de onda formantes modelan la respuesta al impulso de los formantes en el dominio del tiempo. Cada una de estas funciones pueden ser excitadas a la frecuencia fundamental requerida para producir voz cantada.

Otra manera de modelar la voz cantada es aprender directamente de muestras grabadas, y después modificar tales muestras para producir la voz cantada requerida. Los métodos PSOLA[7] (Pitch Synchronous Overlap and Add – Superposición y Suma de Tono Sincronizado) realizan esto en el dominio del tiempo. El problema radica en que no es trivial transformar la forma de onda en el dominio del tiempo de manera que permanezca sonando natural. Debido a que el funcionamiento del oído del humano está basado en el dominio frecuencial también es posible aprehender el espectro de tiempo corto de la voz cantada y transformarlo, de esta manera es más sencillo aplicar transformaciones que suenen naturales a la señal. El problema con este modelo es que la forma en la que los parámetros de control afectan el sonido no es siempre directa. (Siivola: 2002, p. 7)

---

6 Véase Capítulo II

7 Vease el Capítulo III

Modelos físicos

Una forma directa de construir un sintetizador de voz, es el modelar los órganos de voz de un humano. En tales modelos, el tracto vocal es usualmente de-compuesto en la fuente y el tracto vocal, y las propiedades cinemáticas y biomecánicas de la glotis son modeladas. El modelaje completo requeriría el resolver la *ecuación de onda*:

$$\Delta^2\psi = \frac{1}{v^2}\frac{\partial^2\psi}{\partial^2 t^2}$$

Donde Ψ representa la presión aérea. Asumiendo para simplificar que las paredes del tracto vocal son duras (Ψ es 0, dentro de la pared), de todas maneras el resultado es una ecuación muy compleja. Modelando el tracto vocal simplemente como un tubo largo, se puede resolver la ecuación diferencial por el método de d' Alembert. Ello brinda una solución de la forma:

$$\psi(x - vt) = f(x + vt) + g(x - vt)$$

Donde *f* y *g* pueden ser funciones cualesquiera y representan dos formas de onda viajando en direcciones opuestas. Por supuesto, el modelar el tracto vocal como un tubo sencillo y largo es una sobre-simplificación. En lugar de ello se puede modelar el tracto vocal como tubos acústicos de diferentes radios interconectados. Resolviendo la ecuación para este tipo de sistema brinda ecuaciones similares a las de la impedancia de un circuito eléctrico, donde parte de la forma de onda atraviesa la unión y una porción es reflejada al hacerlo. Esto es llamado el modelo de *Guía de Onda Digital*. Los pulsos glotales pueden ser generados de templetes hechos a mano.

Los métodos de síntesis basados en modelos físicos tienen la ventaja de tener la parametrización adecuada para ser controlados como un instrumento real y por lo tanto, poseen la flexibilidad y el potencial para tocar *expresivamente*. Uno de los problemas esenciales relacionado al control de tales modelos es el cómo generar las acciones físicas que excitan al instrumento, mientras que en el "sampling" tales acciones están incluidas en los sonidos grabados (Bonada y Serra: 2007, p. 68). El *Modelo de Síntesis Articulatoria Física del Canto* (Spasm-Singing Physical Articulatory Synthesis Model) es una interfaz gráfica para generar diferentes formas de tracto vocal y pulsos glotales. El tracto nasal es modelado como un tubo acústico paralelo unido al tracto vocal principal. Efectos más sutiles son modelados, como la radiación a través de una pared de garganta. El *software* LECTOR, puede convertir texto y partituras a parámetros. El *software* de Síntesis de Voz Cantada puede leer tal traducción de parámetros y generar voz cantada.

Modelos Espectrales: Vocoders, Síntesis por formantes, FOF, FM

El núcleo de tecnologías que se han desarrollado hasta ahora está fundamentalmente basado en el procesamiento espectral, y al paso de los años se han agregado acciones de ejecución y restricciones físicas para convertir el acercamiento de uno de *muestreo* a una tecnología más flexible y expresiva que mantenga su naturalidad inherente (Bonada y Serra: 2007, p. 68).

Más que considerar la síntesis de *muestreo* como una forma de capturar y reproducir

el sonido de un instrumento, debe ser considerada una forma de modelar el *espacio sonoro* producido por un intérprete con un instrumento, ello garantizará un cambio conceptual en los objetivos a alcanzar, ya que asegura que se desea ser flexible en la elección de controladores de entrada, mientras simultáneamente se busca la posibilidad de usar controles de alto nivel. Esto también tiene implicaciones sobre la conceptualización del timbre, ya que el *espacio sonoro* es definido ambos por el sonido mismo y por el control ejercido sobre el instrumento por el intérprete. Por lo tanto, el *espacio sonoro* de un intérprete refinado será mayor que el *espacio sonoro* de uno no refinado. A partir de un *espacio sonoro* dado, más los controles de entrada apropiados, el motor de síntesis deberá ser capaz de generar cualquier trayectoria en el espacio, produciendo, por lo tanto, cualquier sonido contenido en él. La aproximación de la fuerza bruta es realizar un *muestreo* extensivo del espacio y realizar interpolaciones simples para moverse en él. En el caso de la voz cantada, el espacio es tan inmenso y complejo que ésta aproximación se queda lejos de cubrir una porción del espacio suficiente, siendo un ejemplo claro de por qué un "sampler" no es completamente adecuado, y de que cierta parametrización de los sonidos es requerida. Es necesario entender las dimensiones relevantes del espacio y encontrar una parametrización con la cual se pueda mover en estas dimensiones por medio de interpolación o transformación de sonidos existentes (Bonada y Serra: 2007, p. 68).

Los modelos espectrales están relacionados con algunos aspectos del mecanismo perceptual humano. Los cambios en los parámetros de un modelo espectral pueden ser mapeados fácilmente al cambio de sensación en un escucha. Sin embargo, los espacios que estos sistemas proveen, no son necesariamente los más naturales para la manipulación. Los controles típicos serían: el tono, la forma espectral de los pulsos glotales, frecuencias formantes, anchos de banda de formantes, etc. Un típico ejemplo de la combinación de dos acercamientos es el sistema CHANT[8], considerado un modelo pseudofísico, ya que aunque está compuesto esencialmente de modelos espectrales, utiliza la descomposición fuente/filtro, que considera a la voz como el resultado de la forma de onda de la excitación glotal filtrada por un filtro lineal (Bonada y Serra: 2007, p. 70). Los métodos espectrales pueden producir una voz cantada convincente si no se requiere texto cantado (Siivola: 2002, p. 17). Los mismos presentan también la ventaja de descomponer la voz en armónicos y residuales modelados respectivamente como sinusoidales y ruido blanco filtrado. Los resultados que pueden obtenerse en secciones sostenidas *sonoras*, en secciones transitorias y consonantes (especialmente en fricativas *sonoras*) son muy buenos, salvo que resultan *transientes* significativamente "embarrados". Al buscar modificar frecuencias armónicas, por ejemplo, en una transformación de transposición, las *fases* armónicas, usualmente muy coherentes, no se preservan, produciendo un efecto de *falta de presencia* y una calidad reverberante como si se hubiera grabado en un cuarto pequeño, audible especialmente en sonidos de tono grave (Bonada y Serra: 2007, p. 71).

Vocoders

El Vocoder original fue desarrollado en los laboratorios Bell. La idea era dividir el espectro de frecuencias en bandas y codificar la energía de cada banda. Las bandas eran usualmente colocadas de manera que contuvieran información auditiva igual. Además de la potencia de la señal, el analizador de la fuente debía decidir si la fuente era *sonora* o *sorda*, y en el primer

8 Para mayores referencias consultar: http://www.jstor.org/pss/3679810

caso también la frecuencia fundamental. La señal del habla podía ser transformada cambiando los parámetros de control. Los parámetros de *ganancia* controlaban la envolvente espectral y los parámetros de la *fuente*, el tono y la cantidad de *fonación*.[9] Para de hecho sintetizar voz cantada con el Vocoder, todos los parámetros de control debían ser especificados, haciendo que el producir una voz que sonara natural fuera un trabajo difícil. El Vocoder ha sido utilizado también para crear efectos especiales, como vacas que hablan o guitarras cantantes. En estos casos, la señal de excitación es tomada de otra fuente, y tiene por nombre Cross-Vocoding (Siivola: 2002, p. 8-9).

Como ya se ha dicho, la técnica espectral basada en el Vocoder de fase fija, es una donde el espectro se segmenta en regiones, cada una de las cuales contiene un pico espectral armónico y sus alrededores. En esta técnica, cada región es representada y controlada por este pico espectral, de tal forma que la mayoría de las transformaciones tratan con armónicos y se computa el *cómo* sus parámetros deben ser modificados (de forma similar a los modelos sinusoidales).[10] Estas modificaciones son aplicadas uniformemente a la región, preservando la forma del espectro alrededor de cada armónico. El resultado de aplicar este método puede ser que cuando el tono es modificado, la envolvente de fase desenvuelta se escala de acuerdo al factor de transposición. La calidad del sonido puede ser mejorada en términos de la fase pero no lo suficiente como para preservar exactamente la relación entre formantes y su fase (Bonada y Serra: 2007, p. 71).

Uno de los temas centrales de la síntesis de voz, tanto hablada como cantada, es el desarrollo de algoritmos que consideren la coherencia de fase. La mayoría de estos se basan en la idea de definir tiempos de entrada y salida de tono en sincronía, y reproducir en la salida la relación de fase existente en la señal original que para la entrada. Esta metodología, de no implementarse correctamente puede no reproducir las relaciones de fase de las formantes, y agregar una característica tímbrica de "aspereza" no natural (Bonada y Serra: 2007, p. 71).

Las técnicas basadas en modelos espectrales y Vocoders de fase, se basan en modificar las características del dominio frecuencial de las muestras de voz. Buenos resultados se pueden obtener en transformaciones de transposición y modificación tímbrica. Sin embargo, ciertas transformaciones, especialmente las relacionadas a irregularidades en la secuencia de pulsos de voz, son difíciles de alcanzar ya que implican agregar *sub-armónicos* y su control. Tales irregularidades son inherentes a la "*aspereza*" y a la "*voz quebrada*" (creaky voice) de ciertas voces, y aparecen frecuentemente en el canto como recursos expresivos como el *gruñido* (growling). Algunas soluciones han sido sugeridas para este problema, modelando la salida como una secuencia de pulsos de voz filtrada linealmente por el tracto vocal, y aproximando el espectro como uno obtenido por interpolación de los picos armónicos cuando la ventana de análisis es centrada en el inicio de un pulso de voz. Finalmente, una vez obtenido el pulso de voz filtrado se puede reconstruir el sonido de la voz sobre-poniendo varios de estos pulsos arreglados como en la secuencia original. De esta manera la introducción de irregularidades a la secuencia de pulsos de voz sintetizados es directa (Bonada y Serra: 2007, p. 72).

---

9 Véase Capítulo I.

10 Veáse M. Macon *et al.*, "A Singing Voice Synthesis System Based Sinusoidal Modeling", *Proc. Int. Conf., Acoustics, Speech, Signal Processing*, Munich, Germany, vol. 1, 199, pp. 435-4387.

Sintetizadores por formantes

Los sintetizadores por formantes son el resultado de excitar al tracto vocal modelado por un banco de filtros con resonancias formantes. Usualmente contienen diferentes fuentes para sonidos *sonoros* y *sordos*. Los sonidos vocales son creados al sintonizar las resonancias formantes, mientras que el tono puede ser cambiado simplemente al modificar el ritmo de excitación de la fuente. Los sintetizadores por formantes guardan un fuerte parecido a los Vocoders (Siivola: 2002, p. 9).[11]

Síntesis por función de onda formante (FOF-Forme d'Onde Formatique)

La síntesis por funciones de onda formante está basada en la idea de que se puede modelar la respuesta impulso de cada formante individualmente. Es posible también, después sumar la salida de cada función de onda formante para formar la voz cantada. Cada una de las funciones son excitadas una vez que se localizan al inicio del periodo tonal. Un modelo razonable para una función de onda formante es (Siivola: 2002, p. 11):

$$s(k) = 0, k < 0$$

$$s(k) = \frac{1}{2}((1 - \cos[\beta k]) \cdot e^{-\alpha k} \sin[\omega k + \Phi]), 0 \le k \le \frac{\pi}{\beta}$$

$$s(k) = e^{-\alpha k} \sin[\omega k + \Phi], k > \frac{\pi}{\beta}$$

Aquí, el término coseno de la ecuación controla la velocidad de inicio del formante, y el término exponencial controla el decaimiento. Este método de síntesis es una herramienta del dominio del tiempo, y por lo tanto la precisión de los cálculos no es tan crítica. Estas funciones son computacionalmente económicas y pueden ser reducidas a simples *lecturas de una tabla*. La implementación de síntesis de onda formantes llamada CHANT tiene un control extensivo de la voz que incluye controles de vibrato, apertura de garganta, esfuerzo vocal y jitter (Siivola: 2002, p. 10-11).

Frecuencia Modulada (FM)

La modulación de frecuencia pretende generar un espectro que se aproxime al espectro de la voz cantada modulando una onda base (*carrier*) por una sinusoidal (Siivola: 2002, p. 11):

$$s(t) = A \sin[2\pi f_c t + I \sin(2\pi f_m t)]$$

Aquí, f_c es la frecuencia base y f_m la frecuencia modulante. La relación de estas frecuencias determinan el espaciado relativo entre las frecuencias componentes en la señal modulada. Si $f_c = Nf_m$, la señal modulada será una serie armónica donde la fundamental es igual a f_m . El

11 Véase Capítulo II.

índice de modulación *I* determina cuánto de la energía es transferida de la frecuencia base a las bandas laterales, controlando también cuántas bandas se generan.

En los trabajos de Chowning (1989), los formantes de la voz cantada eran generados con tres pares paralelos FM. La frecuencia fundamental f_m era la misma para todos los osciladores FM, y la frecuencia base f_c determinaba el centro del formante. El índice de modulación controlaba el ancho de un formante. Un sintetizador FM capaz de sintetizar tres formantes puede ser descrito con (Siivola: 2002, p. 12):

$$s(t) = A_1^{0.5} \sin[2\pi f_{c1} t + I_1 \sin(2\pi f_0 t)]$$
$$+ A_2^{1.5} \sin[2\pi f_{c2} t + I_2 \sin(2\pi f_0 t)]$$
$$+ A_3^{2} \sin[2\pi f_{c3} t + I_3 \sin(2\pi f_0 t)]$$

El *esfuerzo* del cantante es modelado elegantemente así: a medida que el esfuerzo incrementa, el contenido en frecuencias altas de la señal incrementa, lo que es modelado tomando una potencia mayor en la amplitud de los formantes altos. Para hacer la voz sintetizada sonar más natural, *jitter* y *vibrato*[12] deben ser agregados a f_0. Aunque la síntesis por FM es ámpliamente utilizada en la síntesis musical, no ha ganado mucha popularidad en la síntesis de voz cantada (Sivola: 2002, p. 11-12).

Síntesis Concatenativa

La síntesis concatenativa está basada en la idea de que es posible, con una base de datos suficientemente grande, elegir los sonidos relevantes y colocar uno tras el otro para así obtener el sonido sintetizado deseado. Por supuesto, en la práctica, una imposibilidad es el almacenaje de tal base de datos. En su lugar, se puede encontrar una forma de cambiar un sonido de la base de datos en otro sonido, de manera que el sonido nuevo también suene natural. Usando transformaciones con suficiente flexibilidad, ello se puede realizar con una base de datos razonable (tal metodología proviene de las investigaciones de voz hablada). Para lograrlo, podemos almacenar formas de onda exactas en la base de datos como sucede en el método PSOLA;[13] otra forma es parametrizar las muestras almacenadas de manera que las transformaciones futuras sean más sencillas. La codificación sinusoidal trata de representar la señal como la suma de algunas señales sinusoidales. Este tipo de codificación permite formas relativamente sencillas de transformación de una señal (como son los cambios tonales y los cambios en la escala de tiempo) (Siivola: 2002, p. 13).

PSOLA-Pitch Syncronous Overlap Add

PSOLA, es un método de dominio en el tiempo que descompone la señal en formas de onda elementales, cada una correspondiendo a un periodo de tono. Cuando estas formas de onda son sumadas con superposición, la señal original es reconstruida. Una modificación temporal puede ser lograda mediante la repetición o eliminación de periodos de tono. La modificación

---

12 Más adelante se analizan tales fenómenos.

13 Un análisis más profundo se encuentra en el Capítulo III.

de frecuencias se realiza cambiando el tiempo entre funciones de onda elementales. El PSOLA funciona solamente para las secciones sonoras de la señal, debido a que debe existir un periodo fundamental para que el modelo funcione. Las acciones *sordas* de la señal son almacenadas como si tuvieran un periodo de tono constante. Durante la re-síntesis las muestras *sordas* son concatenadas en orden aleatorio, lo cual reduce la posibilidad de correlaciones accidentales que pueden resultar en reverberación o un efecto de *flanger*. Este método trabaja bien con transformaciones pequeñas pero para cambios mayores la cualidad vocal sufre o se altera. El método es parecido de algún modo a métodos de síntesis por lectura de tabla de onda (Siivola: 2002, pp. 13-14).

Análisis por Síntesis (ABS-Analysis By Synthesis)

Los métodos ABS dividen la señal en ventanas superpuestas. Para cada ventana, el algoritmo trata de encontrar un número de sinusoidales de frecuencia constante que modelen de mejor manera la ventana. Las sinusoidales son encontradas por un método iterativo que trata de minimizar el error cuadrático en cada iteración.

La señal puede ser representada por (Siivola: 2002, p. 14):

$$x[n] = \sigma[n] \sum w_s[n - kN_s] s_k[n - kN_s]$$

Aquí, σ controla la envolvente del sonido producido, w_s es la función de la ventana (usualmente triangular) y s_k es la suma de sinusoidales que pertenecen a esta ventana. La suma de sinusoidales puede ser escrita como (Siivola: 2002, p. 15):

$$s_k = \sum_{i=0}^{\ell-1} A_i * k \cos(w_i kn + \phi_i^k) kN$$

La re-síntesis de la señal es directa, ya que sólo involucra una transformada de Fourier[14] inversa y una suma con sobreposición. La complejidad del algoritmo radica en encontrar sinusoidales, lo que puede ser realizado al construir la base de datos. La síntesis por éste método es computacionalmente ligera (Siivola: 2002, pp. 1-15).

Modificación Escala-Frecuencia y Escala-tiempo

Una manera sencilla de expandir la longitud de una muestra sería hacer cada ventana más larga, pero de esta manera, la envolvente de ganancia σ[n] tendría también que ser estirada. Cambiar la escala de frecuencias sólo significaría escalar las frecuencias de sinusoidales, mientras que cambiar el tono requeriría remover las características espectrales del tracto vocal, cambiando las frecuencias y re-escalando con la envolvente espectral.

Desafortunadamente, debido a que en la práctica los componentes sinusoidales no

14 Véase Capítulo II.

están relacionados armónicamente de manera estricta, tales métodos llevan a problemas de coherencia de fase. Para secciones sonoras, la coherencia de fase puede ser restaurada alterando la fase de las sinusoidales, pero para las *sordas* el problema es más difícil. Para evitar sonidos tonales o reverberantes en la síntesis, la fase de la señal en las secciones *sordas* debe ser aleatoria.

Las sinusoidales re-escaladas simples pueden traer secciones ruidosas bajo los formantes, resultando en una calidad de voz pobre. Esto puede evitarse sintetizando la señal de un espectro re-muestreado de la señal almacenada (Siivola: 2002, p. 15).

Excitación más resonancias (EpR- Excitation Plus Resonances)

El modelo E_PR está basado en una extensión de la bien conocida aproximación fuente/filtro. El modelo comienza con una síntesis de modelado espectral, que difiere un poco de la codificación sinusoidal (presentado previamente). En lugar de intentar encontrar una codificación óptima para cada ventana de análisis por separado, trata de optimizar el problema sobre todas las ventanas de manera que se pueda monitorear cómo cambia cada sinusoidal. El cambio puede ser uno de frecuencia al inicio o al final. El modelo también considera la señal residual e (t) (Siivola: 2002, p. 16):

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\phi_r(t)] + e(t)$$

La respuesta del tracto vocal es estimada para muestras almacenadas. Las sinusoidales y residuales se someten a un filtrado inverso antes de ser guardadas en la base de datos. Ello ayuda en la re-síntesis debido a que la fuente y los parámetros del tracto vocal están separados y pueden ser transformados individualmente. El filtro fuente modela el esfuerzo del cantante, ya que es responsable de la inclinación (*tilt*) general del espectro. El camino de la señal sintetizada en el EpR está dividido en tres canales paralelos y es excitado con sinusoidales. La excitación residual sonora atraviesa los mismos filtros que la excitación principal, pero tiene que ser escalada en el dominio frecuencial debido a que mantiene algunas características del sonido de donde fue aprendido. Los sonidos *sordos* son almacenados directamente, siendo únicamente sometidos a filtrado inverso con el filtro fuente estimado. Tanto la excitación residual como la excitación sorda son aprendidas de un gran número de muestras y almacenadas en la base de datos (Siivola: 2002, p. 16-17).

Modelando al cantante

El modelo del intérprete (cantante), en un sistema de *Síntesis de Voz Cantada*, es el que está a cargo de generar las acciones de bajo nivel, responsable de incorporar conocimiento específico de interpretación al sistema. Es aquí donde la línea de investigación sobre síntesis de voz cantada explota en posibilidades para nuevos trabajos ya que el entendimiento de las capacidades para simular el proceso de interpretación musical es aún uno de los problemas en música más abiertos. Tales temas son diversos y van desde la teoría de la música a la cognición y problemas de control motor, habiéndose obtenido hasta ahora sólo respuestas parciales mediante las aproximaciones actuales. Las metodologías más efectivas han estado basadas en reglas de desarrollo interpretativo usando una aproximación de *Análisis por*

Síntesis, y recientemente, técnicas de aprendizaje de máquinas usadas para generar automáticamente tales reglas. Otro método útil está basado en la utilización de *software* de notación musical que permite el juego interactivo y ajustes al funcionamiento sintético (Bonada y Serra: 2007, p. 73). A la fecha, los sistemas que modelan la interpretación, han tenido objetivos muy acotados en su desarrollo; el *tempo*; la ocurrencia del *vibrato*; la desviación de la nota de un valor estándar; amplitud del *volumen* nota por nota; y el *ataque*, duración y liberación (*release*) en términos de articulación musical.

La voz humana cantada

La teoría acústica clásica sobre la producción de voz hablada parece ser también aplicable a la voz cantada. Tal teoría puede ser brevemente resumida en lo siguiente: los sonidos *sonoros*, son producidos por la vibración de los pliegues vocales que cortan el flujo aéreo proveniente de los pulmones, produciendo así un tren de pulsos de aire. Este flujo de aire pulsante es equivalente a un tono complejo con un espectro armónico y una frecuencia fundamental igual al ritmo de vibración de los pliegues vocales. La señal, en su estado de *fuente sonora*, es modificada cuando pasa por el tracto vocal, eso es, la faringe y la boca. El tracto vocal es un tubo resonador que para generar vocales abre el extremo de los labios y está prácticamente cerrado del extremo de los pliegues vocales. Cuando el *velo nasal* se encuentra abierto, se producen sonidos nasalizados, complementando el tracto con la cavidad nasal. El tracto vocal es un filtro que impone una curva de frecuencias sobre el sonido radiado desde la apertura de los labios. Esta curva de frecuencias está caracterizada por picos, o formantes, que corresponden a las resonancias del tracto vocal y sus valles entre picos. Los cuatro o cinco primeros formantes son los relevantes perceptualmente.

En términos de niveles de presión sonora, sorprendentemente, los cantantes y no cantantes no parecen diferir apreciablemente con respecto al sonido más "*fuerte*" posible que pueden producir. Sin embargo, estudiantes de canto que entrenan sus voces en escuelas, universidades y conservatorios reportan niveles mayores SPL (Sound Pressure Level) tras una educación vocal "exitosa". La razón probable es que los estudiantes de voz tienden a evitar el tipo de fonación que los no cantantes utilizan cuando producen el sonido más "fuerte" posible, debido fundamentalmente, a que suenan como gritos (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1687).

Mientras en el habla la presión sub-glotal generalmente se asume constante durante la pronunciación de una palabra, en la voz cantada varía de acuerdo a patrones bien formados. Esta variación de presión en el canto necesita ser muy precisa ya que la presión afecta el tono, y por lo tanto, fallas en alcanzar una presión objetivo resulta en un error en la frecuencia fundamental, es decir, una desafinación (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1687-88).

La fuente vocal puede ser variada de diferentes maneras resultando en variaciones de frecuencia fundamental, amplitud y la fuerza relativa de la fundamental. La fuente vocal humana puede operar en diferentes modos de oscilación llamados registros. En la voz masculina existen al menos tres registros llamados: *de máscara* (*pulse* o *vocal fry*), *de pecho* (*modal* o *chest*) y *falsetto o falsete* (*loft*). En la primera de ellas, que típicamente ocurre en los finales de frase en el habla neutra, los pliegues vocales están gruesos y laxos, los pulsos de aire glotales aparecen en grupos de tres o más, o arriban con intervalos de tiempo largos.

Como consecuencia, la frecuencia fundamental se vuelve bastante grave, generalmente bien por debajo de los 100 Hz para voces masculinas. En el registro modal, los pliegues están menos laxos; pulsos glotales adyacentes aparecen en intervalos de tiempo constantes y tienden a ocupar 50% o más del periodo. En el *falsetto*, los pliegues son delgados y estirados, y raramente cierran la glotis por completo, siendo los pulsos más largos que en el registro modal (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1689-90).

Mientras en el habla normal, los tres registros son ocasionalmente utilizados, sólo el registro modal es usado por cantantes varones. Sin embargo, los contra-tenores parecen cantar en el registro del *falsetto*. En las voces femeninas esta situación es menos evidente. Existen razones para asumir que las sopranos y altos utilizan ambos registros de pecho y falsete, este último siendo referido como un registro *de cabeza* en el caso de la voz femenina (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1690-91).

Una característica de la voz entrenada para cantar, es el *vibrato.* Parece que existen al menos dos tipos de vibrato que son producidos y suenan diferente. El vibrato utilizado en el canto operístico es producido por pulsaciones en el músculo cricotiroide que puede elevar el tono. Este vibrato puede ser llamado *vibrato frecuencial*. En el canto de música popular, un vibrato de clase distinta ocurre por variaciones en la presión sub-glotal, imponiendo una modulación de amplitud de fuente vocal (AM-Amplitud Modulada). Este puede ser llamado *vibrato de intensidad* (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1690).

El *vibrato frecuencial* corresponde a una modulación lenta y casi sinusoidal de la frecuencia fundamental. Normalmente el ritmo se encuentra entre 5 y 7 ondulaciones por segundo, y la profundidad de modulación varía entre ±50 y ±150 *cents* (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1691). Como efecto secundario, el *vibrato frecuencial* es acompañado de una modulación de amplitud. La razón es que los parciales cercanos a una frecuencia formante varían en amplitud dependiendo de su distancia al formante. La característica de esta modulación parece ser de menor significancia perceptual. Similarmente, el *vibrato de intensidad* contiene una modulación de frecuencia como un producto de la modulación de la presión sub-glotal. La razón es el efecto de la presión sub-glotal en la frecuencia fundamental mencionada. El *vibrato frecuencial* es generado por pulsaciones en las señales de control neural a los músculos laríngeos. El flujo aéreo a veces varía en sincronía con el vibrato, sugiriendo pulsaciones en los músculos regulando la *aducción* glotal (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1691).

Fenómenos interesantes en la voz de un cantante

El estudio de los fenómenos presentes en las voces de cantantes, resulta de importancia para el avance de la modelación de los mismos. A continuación se enlistan y ejemplifican brevemente algunos de ellos. Para iniciar, se sabe por ejemplo que entre mayor sea la apertura de la mandíbula, mayor la frecuencia del primer formante. Por otro lado, entre más cercano esté el punto de constricción entre la lengua y la parte superior de la boca a la parte frontal del tracto vocal, mayor será la frecuencia del segundo formante; y entre mayor sea el grado de curvatura trasera de la punta de la lengua, más bajo será el tercer formante. Finalmente, redondear los labios y bajar la laringe causa que todos los formantes decrementen.

En términos del esfuerzo, el de un cantante no puede ser modelado simplemente escalando la amplitud, ya que esto sólo tendría el efecto perceptual de que el cantante se encuentre más cerca o lejos. Cuando se canta "intensamente", la potencia en la porción final del espectro sube significativamente comparada con aquélla de la región baja del espectro. Esto puede ser modelado variando la inclinación de la envolvente espectral. Este efecto es también audible en notas largas, en donde la energía de la señal es más o menos constante, pero el esfuerzo del cantante se incrementa hacia el final de la señal (Siivola: 2002, p. 4).

Otro efecto interesante que puede ser producido es lo que se conoce en inglés como "Jitter", que es la variación del tono a ritmos mayores que la frecuencia típica del vibrato. Este efecto tiene como consecuencia que un cantante no pueda mantener su voz completamente estable, y es causado por disparos neuronales al azar en la cadena auditiva de retroalimentación. El cambio gradual de entonación es llamado en inglés "drift" o deriva y es mucho más controlable que el "Jitter" (Siivola: 2002, p. 5). Debido a que el tracto vocal tiene masa, posee cierta inercia, respondiendo a la primera ley de Newton. Un cambio en tono no ocurre instantáneamente sino que el cantante debe de cambiar de tono al final de la nota precedente, de modo que la nueva pueda comenzar en el tono correcto (Siivola: 2002, p. 5).

Para agregar expresividad a una interpretación, primeramente debe encontrarse la estructura del fraseo. Ello significa que el inicio de la frase usualmente contiene un *acelerando* y al final un *retardando*. Especialmente la nota final es comúnmente alargada. Adicionalmente, la estructura armónica demanda énfasis en ciertas notas. Así pues, el significado del cambio armónico debe ser encontrado para dar énfasis a los pasajes que deben tenerlos. Algunas técnicas, como el *marcatto*, pueden ser utilizadas para dar énfasis a ciertas notas. Esta técnica consiste en rebasar con claridad la frecuencia que se pretende alcanzar al inicio de la nota, antes de regresar al tono correcto (Siivola: 2002, p. 5).

El Formante del cantante y formas especiales de cantar

El término formante puede ser entendido en dos maneras: en la primera, un formante equivale a una resonancia del tracto vocal, y es manifestado acústicamente como un pico en la envolvente del espectro. Por lo tanto, las frecuencias formantes prácticamente igualan las frecuencias de estos picos que son controlados por la forma del tracto vocal. Por otro lado, un formante es un "intervalo de frecuencias del espectro de un sonido dentro del cual los parciales tienen amplitudes relativamente grandes". Mientras la frecuencia fundamental sea baja, como en el habla neutra, las dos definiciones son casi iguales. Sin embargo, a una frecuencia fundamental de 880 Hz, que está dentro del intervalo soprano, los cuatro parciales más bajos es muy probable que se encuentren relativamente "fuertes". Por lo tanto, de acuerdo con la segunda definición, el intervalo entero de frecuencias de 0-4 kHz es un formante. Esta confusión es evadida en la primera definición, esto es, una resonancia del tracto vocal (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1691).

Las frecuencias de resonancia de un tubo, dependen de su longitud y forma, los cuales, en el caso de la voz, son controlados por: la posición de los labios, la quijada, la laringe, el velo, y las paredes laterales paralingeales; es decir, por la articulación. Las frecuencias formantes son decisivas para el timbre. Existen dos aspectos distintivos sobre el timbre de sonidos vocales. Uno es la *cualidad vocal*, que determina cuál vocal es percibida, y

el otro aspecto es la *cualidad de voz*, que es una característica personal. La longitud del tracto vocal, y por lo tanto las frecuencias formantes dadas por una vocal, varían también entre grupos de hombres, mujeres y niños. Tales diferencias explican mucho sobre las variaciones de timbre de voz entre individuos. Se ha mostrado que para una vocal dada, los cantantes tenores tienden a presentar frecuencias formantes más elevadas que los cantantes bajos. Las frecuencias formantes relativamente elevadas, entonces, pertenecen a las características típicas del timbre de voz tenor (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1691).

Una característica mostrada por bajos, tenores y altos, es un pico de envolvente espectral inusualmente alto, que ocurre entre 2 y 3.5 kHz. Aparecen todos los sonidos sonoros y puede ser explicado como consecuencia de un agrupamiento de la tercera, cuarta y quinta formante. Este pico, ha sido llamado el *formante del cantante*. Tal agrupamiento de formantes es compatible con la teoría acústica de la producción de voz, con los pliegues vocales constituyendo el fondo de una cavidad en forma de tubo que es la laringe. La entrada a éste tubo está localizada en la faringe, a unos 2 cm sobre su extremo terminal. Si el tubo laringeal tiene una entrada considerablemente más estrecha que la faringe, actúa como un resonador separado, cuya frecuencia de resonancia no es influida por el resto del tracto vocal. Puede asumirse que el tubo laringeal está asociado con el cuarto y quinto formante, por lo que para producir el *formante del cantante,* parece esencial incrementar el ancho de la faringe, lo que puede ser obtenido bajando la laringe. Este comportamiento es típicamente observado en cantantes varones (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1691-92).

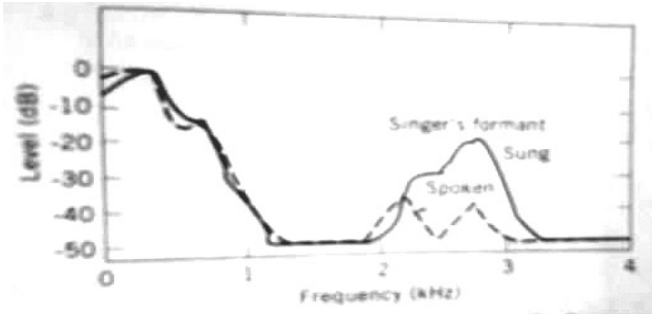


Imagen 1. El Formante del Cantante

El *formante del cantante*, en un intervalo de frecuencias particularmente sensible para el oído humano, contribuye a hacer el timbre de voz más distintivo y "brillante" (*shinny*). Una ventaja importante, es que este efecto puede ser alcanzado sin la necesidad de esfuerzo vocal excesivo. Sin embargo, un incremento de volumen vocal, aumenta el nivel relativo del *formante del cantante* debido al efecto de tales incrementos sobre la fuente de voz. La frecuencia central del *formante del cantante* parece significativa para la cualidad de voz. La cualidad de una voz típica de *bajo* puede ser obtenida de un sintetizador por formantes, si la frecuencia central es cercana a 2.2 kHz, y la de un timbre tenor con una frecuencia central cerca de los 2.9 kHz. Las voces *Alto* parecen tener una frecuencia central aún más elevada, lo que parece reflejar diferencias en la longitud del tracto vocal y forma faringeal. Esta faringe ancha, necesaria para el agrupamiento de formantes elevados, también trae cambios en los dos primeros formantes (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1692).

Dependiendo de la vocal, el valor normal de F_1 (el primer formante) varía entre 250

Hz y 1000 Hz. Por lo tanto, puede ocurrir que la frecuencia fundamental sea más elevada que F_1. Por ejemplo, a tonos muy altos, las sopranos tienden a ensanchar sus quijadas cuando están elevando el tono, eso casi sin distinción de qué vocal se trate. Esto parece mantenerse al menos hasta frecuencias cercanas a 880 Hz. La apertura de la quijada es particularmente influyente sobre el primer formante. Todo apunta a que el propósito de las aperturas de quijada dependientes del tono, es el afinar el primer formante a una frecuencia un poco más elevada que la frecuencia fundamental. La adaptación de la forma del tracto vocal a la frecuencia fundamental afecta todas las frecuencias formantes. La razón para los cambios en la frecuencia formante es probablemente acústica. Como resultado, la amplitud de la fundamental es incrementada por resonancia, y el nivel sonoro del sonido radiado es aumentado de acuerdo a ello. Probablemente no sólo las sopranos aplican esta estrategia; dependiendo de la frecuencia fundamental y el valor normal de F_1 de la vocal cantada, la misma estrategia es aplicada también por barítonos, tenores y altos. Para vocales diferentes que /*a*/, pequeños cambios de F_1 se logran probablemente al disminuir el grado de constricción de la lengua sobre el tracto vocal, pero para cambios mayores los cantantes parecen utilizar un incremento de la apertura de la quijada. La cualidad de la vocal está determinada sobre todo por el primer y segundo formante, por lo que se podrían esperar consecuencias negativas relacionadas a la inteligibilidad de la vocal en estos casos. Cuando debe decidir entre entre tonos inaudibles con una cualidad vocal normal, o tonos audibles con una cualidad vocal extraña, los cantantes probablemente toman una decisión óptima. La cualidad de vocales sostenidas sobrevive a tales correcciones soprendentemente bien, dependientes al tono de las frecuencias formantes, excepto para frecuencias fundamentales sobre 700 Hz. Sobre tal frecuencia ninguna combinación de formantes parece ayudar, y debajo de la misma la cualidad vocal no sería mejor, si frecuencias formantes normales fueran escogidas. La cantidad de inteligibilidad de texto que ocurre en tonos muy elevados radica casi exclusivamente en las consonantes que rodean a las vocales (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1693).

Dado que la inteligibilidad de las palabras en una canción no es tan importante como en la voz hablada, ésta a veces es comprometida para producir una voz cantada más poderosa. Como se ha dicho, sobre todo para cantantes femeninas, el primer formante de una vocal está a menudo por debajo de la frecuencia fundamental; son las cantantes femeninas clásicas quienes a menudo mueven el primer formante para que coincida con la fundamental, y el segundo formante a algún múltiplo de la fundamental, lo cual crea fuertes resonancias en algunos de los armónicos de la nota, creando a su vez un sonido mucho más poderoso. El efecto correspondiente en los hombres es el *formante del cantante*, lo que colorea la voz de forma que es fácilmente reconocible de, por ejemplo, las voces de apoyo coral (Siivola: 2002, p. 5).

Aunque la mayoría de la investigación sobre *Voz Cantada* ha sido dedicada al canto operístico occidental, en recientes años también otros tipos de canto han sido investigados. En el canto en *Sobretonos*, que puede encontrarse en varias culturas asiáticas, el órgano de voz es utilizado de manera especial. La frecuencia fundamental es mantenida constante mientras que el segundo y el tercer formante son sintonizados a *sobretonos* espectrales específicos. Al hacerlo, patrones melódicos emergen en donde los tonos corresponden a los mejorados *sobretonos*. Por otro lado, el canto en la *Ópera China* es completamente diferente al canto de la Ópera Occidental. En la primera, los cantantes masculinos cantan en el intervalo frecuencial de los *altos* y *sopranos*; la posición de la laringe es elevada y las voces también poseen el *formante del cantante*. Por otra parte, el *Belting* es un tipo de canto usado en el teatro musical (Broadway), en el que el sonido es intenso y el color de las vocales es

más similar al usado en el habla normal que en el canto operístico. Esto es producido por una faringe angosta y una laringe elevada. La presión de los pulmones es alta, y la aducción glotal es más forzada que en el canto operístico. La fuente de voz parece ser caracterizada por una gran cantidad de sobretonos y su uso regular es considerado dañino a la función de voz (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1693-94).

El canto coral, es probablemente el más utilizado. Los cantantes corales muestran menos componentes de alta frecuencia (*formante del cantante)* lo cual probablemente promueve la mezcla de las voces corales. En este contexto, el canto coral se parece más a la voz hablada normal. Para una vocal dada, un cantante coral parece dispersar menos F_1 y F_2 que cuando se pronuncian en la voz hablada. Además se tiende a reducir F_3 y F_4 en las vocales a través de ecualización. La coincidencia en la frecuencia fundamental de un coro es muy alta, siendo posible entre los *bajos*, que los promedios de estas difirieran por solo 12 *cents*. Las ejecuciones de un mismo coro, en diferentes lugares, muestran una posición elevada de la laringe cuando se encuentra en un cuarto con una acústica pobre, es decir con poco tiempo de reverberación y muchas reflexiones de frecuencias graves (Sundberg, *Encyclopedia of Acoustics*: 1997, p. 1694).

La voz Soprano

La cantante soprano usa tanto los tonos del extremo superior del rango de la voz humana como la relación entre las frecuencias relativamente altas y los formantes que utiliza en el habla; lo que distingue su técnica de aquella utilizada por los hombres al cantar.

Como característica esencial de la voz soprano, el balance de la energía espectral se carga hacia los armónicos más graves, siendo F_0 (frecuencia fundamental) el componente con mayor intensidad del espectro, lo cual apoya la teoría de Sundberg en la cual las formantes más bajas "siguen" al periodo de la altura tonal (Castro: 1994, 75). Por otro lado, existen picos secundarios en el espectro (2 o más) que corresponden a las resonancias del tracto vocal o formantes superiores. Tales formantes no necesariamente se mantienen a una frecuencia, sino que suben y bajan de acuerdo con la vocal que se esté emitiendo. Las formantes superiores decrecen en energía más rápidamente cuando un tono se canta con intensidad decreciente, y mientras únicamente la formante más baja goza de prominencia en los dinteles de amplitud del ataque y el decaimento, las formantes superiores se marcan únicamente cuando la señal se aproxima a un estado *quasi*-estable (Castro: 1994, p. 75).

Un fenómeno particularmente interesante, y donde el resultado para el escucha es un color tonal "admirable", es el rápido ajuste de los formantes de, por ejemplo, una vocal a un componente armónico del espectro de la voz para notas que se extienden un poco. Tal "afinación" no parece poder tomar lugar de inmediato, pero sí rápidamente dependiendo de la habilidad de la cantante. Para ello, la cantante debe alterar la lengua, quijada, y posición de los labios un poco para hacer que la primer formante suba hasta alcanzar el componente fundamental de su sonido sonoro (como puede verse en la *Imagen 2* /oo/). Con algunos pequeños ajustes, se puede lograr que el segundo formante coincida con el segundo parcial (como se ve en la imagen 2 /ah/). En sus observaciones, John Sundberg describió la manera en que una soprano colocaba sus formantes cantando varias vocales, mostrando experimentalmente que las cantantes pueden alinear sus frecuencias formantes de manera

casi exacta o exacta (Benade: 1990, p. 382). A continuación se muestran las curvas de influencia de los harmónicos de la voz sobre las formantes de varias vocales.

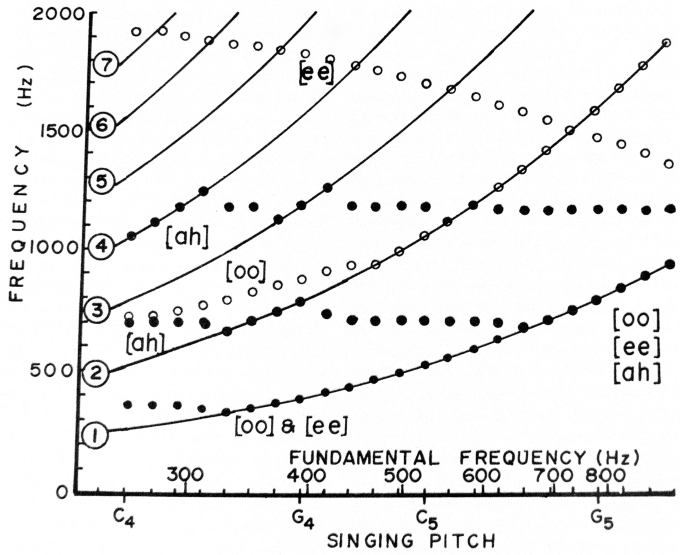


Imagen 2. Influencia de los armónicos de la voz sobre los Formantes de varias vocales cantadas

En ellas, se muestra lo que una soprano podría hacer para realizar afinaciones de sus propias formantes a las frecuencias de voz requeridas en una circunstancia musical. Se presentan marcas de una escala cromática, entre las notas C_4 y $A\#_5$, (sobre el eje de las *abcisas*), con indicaciones de las frecuencias fundamentales que pertenecen a tales notas; el eje de las *ordenadas* se encuentra marcado con una escala frecuencial para indicar las frecuencias de las parciales de la voz y aquellas de varias formantes. Las líneas sólidas que se elevan hacia la derecha, muestran la tendencia de la frecuencia fundamental y de sus armónicos a medida que se canta la escala. Los números muestran a qué armónico se hace referencia. La secuencia de puntos a lo largo de la parte baja de la gráfica muestra la manera en que la frecuencia del formante varía con la vocal en la escala cromática. Para una gran porción del rango de una soprano, ésta puede fortalecer el primer parcial dejándolo montar sobre la primer formante de /ee/ ú /oo/.

Otros Ejemplos

Algunos otros ejemplos de esfuerzos orientados a la síntesis de voz, ya sea hablada o cantada, existen en la literatura, y se recomienda su revisión para la profundización en el tema de Voz Cantada en Español.

Una amplia gama de acercamientos a la emoción en la voz se han intentado con algunos buenos resultados. La Síntesis de Voz con emociones ha sido investigada por varios autores lográndose la creación de *Intérpretes de voz sintéticos, con poca naturalidad pero*

"vivos" que generan empatía.[15] Por otro lado, la Universidad Politécnica de Madrid ha aproximado la síntesis de la voz en Español con emociones[16], mientras que en la Universidad Pompeu Fabra, de Barcelona, se han desarrollado investigaciones sobre los controles para la expresividad.[17] El estudio de emociones en la voz hablada deberá dar luz a la rama de *expresividad* de la voz cantada sintética.

Desde la rama de la investigación de los sistemas *Texto-a-voz*[18], una estrategia de *múltiples dominios* ha sido propuesta, cuya estructura podría ser de beneficio para los sistemas de voz cantada ya que permitiría flexibilidad y calidad.[19] La investigación sobre *Segmentación Automática* podría ayudar a decidir sobre las unidades fonéticas más convenientes para la síntesis de voz, en particular en español.[20] Otros sistemas de síntesis de voz hablada de otros idiomas pueden encontrarse en la literatura. [21]

Por último, se recomienda escuchar un ejemplo de una implementación comercial acertada de la síntesis de voces corales, con un sistema de control y construcción para el usuario. Aunque muy bueno, el sistema es, sin embargo, perfectible:

http://www.soundsonline.com/Symphonic-Choirs-PLAY-Edition-pr-EW-182.html (Última revisión, 2-Noviembre-2009)

Debido a que el sistema es comercial, aún no se ha podido probar de manera extensa, sin embargo, a juzgar por los *demos* accesibles en línea, aunque el prototipo presenta la posibilidad de elegir las sílabas a cantar y se puede controlar la frecuencia y duración mediante la utilización de un teclado MIDI, la inteligibilidad es severamente comprometida y sólo se incluye el idioma inglés y latín. Es incierto el núcleo silábico utilizado, ya que aunque los controles sugieren que funciona a base de ciertos fonemas y todas las vocales, de hecho incluye una base de datos de frases más que de sílabas. No obstante la posibilidad de elegir entre voces masculinas, femeninas y una mezcla de ambas, no parece ser posible elegir una sola voz, ni un conjunto particular de ellas. A pesar de sus limitaciones, el sistema logra cierta naturalidad de las voces y el resultado bien puede pasar como un verdadero coro. La aplicación incluye también algoritmos de reverberación, y una aplicación llamada *"Coros Apocalípticos"*, para los cuales se modifica el ataque y liberación de las palabras, y tiene la posibilidad de ejecutar melismas.

Antes de pasar a las especificaciones técnicas del Sinstema de Síntesis generado, se presenta en los *Capítulos* (*I* y *II*), una introducción de los temas escenciales que constituyen una investigación sobre *Sintesis de Voz Cantada*.

---

15 Cristofer Newell and Alistar Edwards ,"Unnatural but lively voice synthesis for empathic, synthetic performers",

16 José Manuel Pardo,"Nuevas fronteras de la tecnología del habla: Síntesis de voz con Emociones", *Foro Computlense, Fundación General UCM. ONCE*. Dpto de Ingeniería Electrónica. UPM, http://www.funcacionucm.es/www.once.es

17 Jordi Janer, Jordi Bonada, "Performance driven control for sample based Singing voice synthesis", *Proc. Of the 9th Int. Conf. On Digital Audio Effects (DAFx-06) Montreal, Canada,* Sept. 18-20, 2006, pp. 41-44.

18 Los principios de los sistemas *Texto-a-voz* se revisarán a detalle en el *Capítulo III*.

19 Frances Alías *et al*., "Towards High-Quality Next-Generation Text-to Speech Synthesis: A Multidomain Approach by Automatic Domain Classification", *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 7, septiembre 2008, pp. 1340-1354.

20 H. M. Torres, J. A. Gurlekian, "Acoustic Speech unit segmentation for concatenative synthesis", *Computer Speech and Language,* no. 22, 2008, pp. 196-206.

21 Marcus Uneson, "Outlines of Burcas-a Simple Concatenation- Based MIDI-to- Singing Voice Synthesis System", *TMH-QPSR*, *Fonetik,* Vol. 43, 2002, pp. 1-4; Varvara Kyritsi *et al.,* " A Score-to-Singing Voice Synthesis System for the Greek Language", *Proc. Of the Computer Music Conference*, 27-31 agosto, 2007, pp. 216-223.

Capítulo II

I. Fundamentos del Lenguaje

El Lenguaje hablado

Hay quien sostiene la idea de que el lenguaje hablado es el resultado de la búsqueda por un medio de comunicación con suficiente capacidad y velocidad para poder transmitir los pensamientos y conceptos desarrollados por los humanos (Flanagan, *Encyclopedia of Acoustics*: 1997, p. 1557). Así, una vez establecido que el lenguaje a través de señas no era suficiente para este objetivo, se pasó a utilizar el sistema respiratorio para la generación de tal medio de comunicación, ello debido a que la gran diversidad de sonidos que pueden generarse, cada uno con una huella distintiva en patrones de frecuencias audibles, presentaba las ventajas buscadas contando con suficiente capacidad de transmisión para convertir los pensamientos humanos en conceptos (Flanagan, *Encyclopedia of Acoustics*: 1997, p. 1557). Al establecerse convenciones sobre el significado de tales patrones, cuando se colocan en secuencia, se constituyen las bases del lenguaje hablado.

Es posible esquematizar tanto los pasos de la generación del habla como los de percepción como funciones cualitativas (Flanagan, *Encyclopedia of Acoustics*: 1997, p. 1558):

Generación del Habla		**Percepción del habla**
Formulación del mensaje		Sonido
Código del lenguaje		Movimiento de la membrana basilar
Acciones neuro-musculares		Transducción Neural
Fuente sonora (pliegues)		Código de lenguaje
Sistema Acústico (tracto vocal)		Comprensión del lenguaje
Sonido		

Tabla I. Funciones Cualitativas del habla en humanos (ordenadas y no correspondientes)

El proceso de generación del habla comienza en el cerebro, con un pensamiento y la intención de comunicarlo. Esto activa los movimientos musculares que producen los sonidos del habla. El escucha recibe tales sonidos por medio del sistema auditivo, tras lo cual los procesa para su conversión en las señales neurológicas que el cerebro puede comprender. Así, el proceso de producción del lenguaje hablado, inicia con el mensaje de naturaleza *semántica* en la mente de una persona, que debe ser transmitido a quien lo escucha por la vía del habla. La contra-parte computacional del proceso de formulación del mensaje es la aplicación semántica que crea el concepto a ser expresado. Tras la creación del mensaje, el siguiente paso es la conversión del mismo en una secuencia de palabras. Cada palabra consiste en una secuencia de fonemas que corresponden a la pronunciación de las palabras. Cada oración contiene además un patrón *prosódico* que denota la duración de cada fonema, la entonación de la frase, y el volumen de los sonidos. Una vez que el sistema del lenguaje finaliza el mapeo, quien emitirá la voz ejecuta una serie de señales neuromusculares. Las ordenes neuromusculares realizan ahora un mapeo articulatorio para controlar los pliegues vocales, labios, quijada, lengua y paladar, para así producir la secuencia sonora como salida final.

El proceso de la comprensión del lenguaje opera de manera inversa. Primero, la señal pasa a la cóclea en el oído interno, la cuál ejecuta un análisis frecuencial como un conjunto de filtros. Un proceso de transducción neural sigue y convierte la señal espectral en señales de actividad en el nervio auditivo, que corresponden a las cualidades de los componentes de extracción. Actualmente, no es claro el cómo es que la actividad neural es mapeada al sistema del lenguaje, y cómo es que la

comprensión del mensaje es alcanzada en el cerebro. Las señales de voz están compuestas de patrones audibles análogos que sirven como la base para una representación discreta y simbólica del lenguaje hablado. La producción e interpretación de estos sonidos están gobernados por la sintaxis y la semántica del lenguaje hablado (Huang, Acero y Hon: 2001, pp. 19-20).

En el ámbito de la síntesis y el reconocimiento del lenguaje a través de máquinas, cada uno de los procesos anteriores presenta una contraparte que se menciona a continuación:

Generación del Habla		**Percepción del habla**
Mensaje escrito		Transmisión eléctrica
Secuencia de fonemas- Conversión Pro		Análisis Acústico-Espectral
Conversión Digital-análogo		Extracción de Parámetros (re-Codificación)
Movimiento Articulatorio		Conversión Análogo-Digital
Transformación Eléctrica		Fonemas-palabras-prosodia (sintaxis)
Sonido		Significado (semántica)

Tabla II. Funciones Cualitativas del habla en máquinas (ordenadas y no correspondientes)

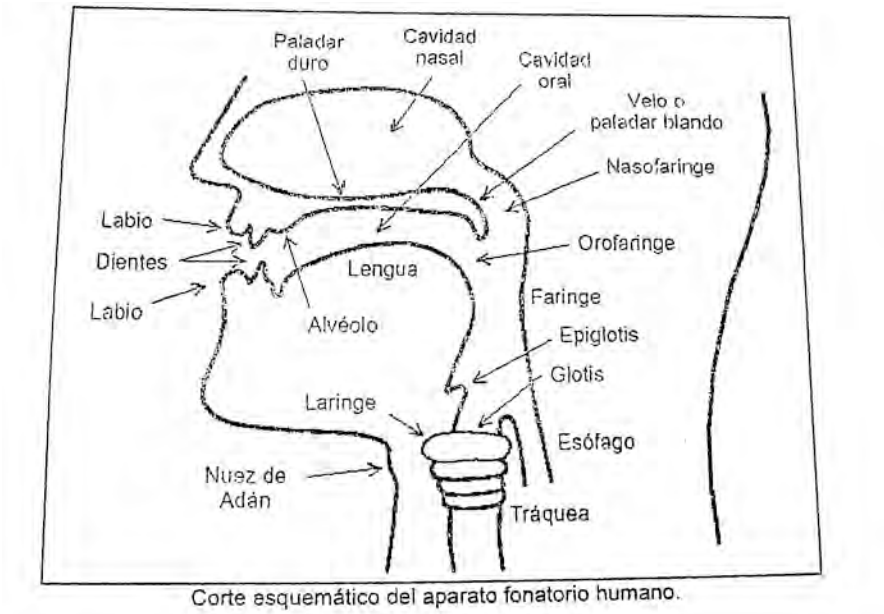
El estudio de la percepción de la voz se encuentra en la intersección de una variedad de disciplinas que incluyen a la acústica, la psicología experimental y la lingüística. El interés fundamental es el entender cómo es que el escucha analiza la señal acústica producida por quien habla, a fin de identificar las unidades lingüísticas que componen las palabras individuales del lenguaje. Así pues, cinco cuestiones experimentales (de extraordinaria relevancia), caracterizan ésta línea de investigación: (1) la naturaleza de las unidades de representación que subyace en el procesamiento del lenguaje al nivel del análisis; (2) la complejidad del mapeo entre la señal acústica y la unidad de representación (la falta de segmentación específica y el problema de especificaciones múltiples); (3) el fenómeno de percepción categórica; (4) las bases biológicas de la percepción de voz, con un enfoque en animales, la especialización hemisférica para la percepción, y la percepción de la voz por infantes; y (5) el papel de los factores *léxicos* superiores sobre la percepción de la voz (Miller, *Encyclopedia of Acoustics:* 1997, p. 1579).

El problema central de la percepción del habla, radica en cómo se mapea la señal acústica, al escucharla, sobre representaciones lingüísticas. En este sentido, dos unidades históricamente propuestas continúan teniendo valor: a) El segmento fonético (o fonemas), que responde a las vocales y consonantes individuales del lenguaje. La idea detrás de éstos segmentos es que el escucha busca, percibe y analiza las propiedades que especifican la secuencia de consonantes y vocales de cada palabra, que son las unidades lexicológicas. Este proceso es esencialmente el de un mapeo de las características acústicas relevantes lingüísticamente sobre representaciones segmentadas. b) El carácter fonético. De acuerdo a la teoría lingüística, los segmentos fonéticos de un idioma pueden ser definidos en términos de constelaciones de pequeños conjuntos de características fonéticas, definidas en términos de atributos articulatorios y/o acústicos. En los últimos cincuenta años, una variedad de técnicas experimentales han dado evidencia de que el mapeo entre señal acústica de voz y la estructura fonética del lenguaje es extraordinariamente compleja y que los seres humanos están bien adaptados para extraer tal información sin importar la complejidad. El siguiente paso es el explicar la naturaleza de los mecanismos de percepción subyacentes que permiten al humano tales capacidades (Miller, *Encyclopedia of Acoustics:* 1997, pp. 1585-87).

II. Fundamentos de Fisiología

El sistema generador de voz

La voz humana tiene la capacidad de retener cierto timbre o *color* al enunciar palabras en distintos tonos, lo que nos permite distinguir que pertenece a cierta persona en particular. Tal *color* se encuentra conectado, como en cualquier otro instrumento musical, a la estructura y construcción del instrumento, y por ende a la forma en que vibra el mismo (Benade: 1990, p. 361) .



Corte esquemático del aparato fonatorio humano.

Imagen 3. Corte esquemático del aparato fonatorio humano

La fuente de energía de la producción de la voz son los pulmones. Cuando los músculos reducen el volumen de los pulmones, el aumento en la *presión* aérea escapa a través de la glotis. Si el sonido producido es *sonoro*, el cantante usa los músculos de las cuerdas vocales para posicionar a las mismas para que vibren. Tal vibración lleva a una presión u onda triangular de la glotis, por lo que el espectro generado es de naturaleza armónica. Así pues, el proceso del habla comienza, como ya se dijo, en los pulmones, forzando aire a través de la tráquea, después a través de la glotis (donde se encuentran las *cuerdas vocales*), y luego por la cavidad de la faringe hasta la boca o cavidad bucal (V. Siivola: 2002, p. 3). Aunque la vibración de la glotis y el tamaño y forma de la faringe afectan las características acústicas de la voz, es la lengua, mas que cualquier otro elemento del tracto vocal, la que crea la articulación del habla (Dodge, Jerse: 1997, p. 221). Los labios radian el sonido de la voz al espacio circundante, y al mismo tiempo su posición determina ciertos sonidos. Cuando el *velum* (paladar suave) desciende, las fosas nasales contribuyen a esta irradiación, creando los sonidos nasales del habla. Cada uno de estos articuladores: la lengua (punta, zona media, y parte posterior), los labios, la mandíbula, el paladar y la laringe, son controlados por separado. Utilizados en conjunto producen los fonemas, que son los varios sonidos del habla.

Existen diferentes fuentes de excitación para la voz. Los sonidos *sonoros* son causados por las vibraciones *quasi*-periódicas de la glotis. Una vez que la onda de presión encuentra las otras partes del tracto vocal, las resonancias del tracto le dan forma al espectro de tal manera que se encuentran picos energéticos (o *formantes*) en ciertas regiones del espectro, esenciales en los fonemas *sonoros*.

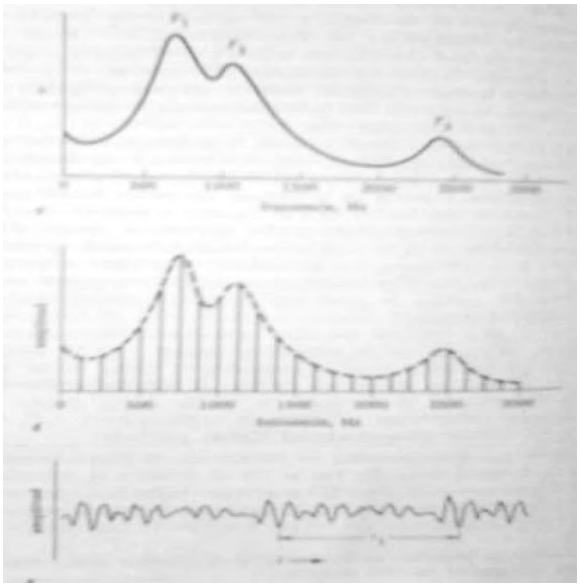


Imagen 4. Imagen de Formantes del sonido /ah/

Para los fonemas que llamamos *sordos*, la excitación es la turbulencia del aire (Dodge, Jerse: 1997, pp. 221-222). La localidad de la turbulencia es controlada regulando el tracto vocal en lugares apropiados como pueden ser garganta, dientes o labios. Los sonidos fricativos son causados por la turbulencia creada por una constricción en algún punto del tracto vocal (h, s, f, etc.); y los sonidos oclusivos por la súbita liberación de presión contenida detrás del cierre de una parte del tracto vocal (Siivola: 2002, p. 4).

El sistema respiratorio

El instrumento de la voz humana se separa en tres partes funcionales fácilmente identificables, y que pueden ser estudiadas por separado.

a) Tracto Pulmonar o respiratorio

Es formado por los pulmones y la tráquea. Los pulmones generan aire comprimido que es conducido por la tráquea. Estos órganos controlan la amplitud de los sonidos, y su única contribución "audible" son los silencios *inter* y *entre* palabras. Los pulmones, que son una masa esponjosa de gran área y con capacidad de 4 a 5 litros en un adulto, están contenidos en una cámara de aire, *la pleura*, a su vez contenida lateralmente por las costillas e inferiormente por el diafragma. El diafragma es un músculo en forma de domo, que cuando se contrae, se extiende hacia fuera, haciendo que el volumen de la pleura se incremente y que el aire entre a los pulmones por un diferencia de presiones interna y externa. Cuando el diafragma se relaja, su extensión se contrae y el proceso es el inverso. El volumen de una primera aspiración a la siguiente, durante la producción del habla, está usualmente entre los 500-1000 cm^3. Para la mayoría de los sonidos del habla, las vías aéreas fijas ubicadas por debajo de la laringe no influyen significativamente al sonido irradiado. Sin embargo, las resonancias de tales regiones pueden observarse en el habla de algunos individuos, particularmente para sonidos producidos con la glotis relativamente abierta y las frecuencias típicas son observadas alrededor de 700, 1500 y 2200 Hz para adultos (Stevens, *Encyclopedia of Acoustics*: 1997 p. 1566). La producción de sonidos requiere de una presión por parte de los pulmones del orden de 4 cm H_2O[1]

[1] Las pulgadas o cm H_2O son unidades de presión derivadas del sistema inglés, que generalmente se

para sonidos muy suaves, y aproximadamente 20 cm H_2O para sonidos muy fuertes y de altas frecuencias. La generación de voz consiste de inhalaciones largas o cortas, así como de exhalaciones controladas, mientras que en la respiración éstas son regulares y aproximadamente de igual longitud (Herrera: 2006, pp. 9-10).

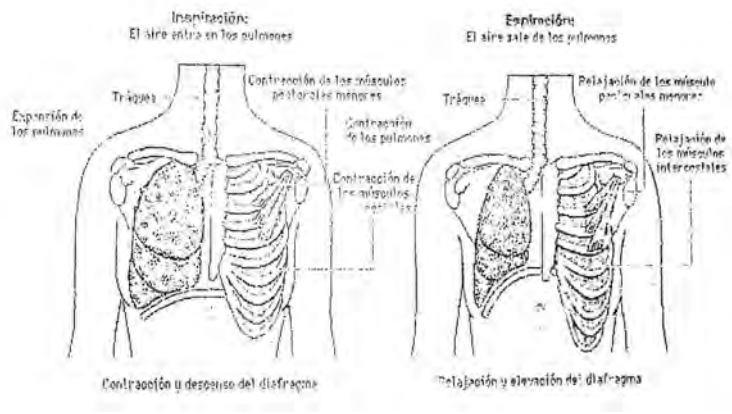


Imagen 5. Inspiración y Espiración

b) *La laringe y las cuerdas vocales*

Es el área situada superiormente a la tráquea e inferiormente a la faringe, y es donde se generan los sonidos. Después de la generación de la presión aérea por parte de los pulmones, la laringe realiza la función de excitación, que tiene por resultados la *fonación*, el *susurreo,* la *fricación*, la *compresión* o la *vibración* (conceptos revisados más adelante). Está formada por: los cartílagos cricoide, tiroides y aritenoides (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1566), un conjunto de músculos, y las cuerdas o pliegues vocales. Estos últimos constituyen la fuente de generación de sonidos, además de que logran el cierre de la tráquea para proteger el tracto pulmonar de objetos externos y permitir la formación de presión dentro del tórax y el abdomen. Los pliegues vocales son un tejido sólido (pliegues de músculo) con dobleces entre el frente y la parte posterior de la laringe, que vibran generando una apertura alargada y estrecha por donde el aire atraviesa. La longitud de los pliegues vocales es de alrededor de 1.0 cm para una mujer adulta, hasta 1.6 cm para un hombre adulto; y tan pequeños como 1 mm para un recién nacido. Para un adulto, el grosor de los pliegues está entre 2 y 3 mm, pero tal grosor decrece cuando los pliegues se estiran (Setevens, *Encyclopedia of Acoustics:* 1997, p. 1566). Pueden adoptar una gran variedad de formas, generando a su vez, una gran variedad de aperturas. Es cuando tales partes se abren o cierran parcial o totalmente, de manera rápida y secuencial, que se producen los sonidos de la exhalación (Herrera: 2006, pp. 9-10).

utilizan como unidad de medida de presiones pequeñas y tiene su equivalente en unidades más comunes como Pascales y Atmósferas, siendo un pie columna de agua, equivalente a 2,989 kilo pascales (kPa)

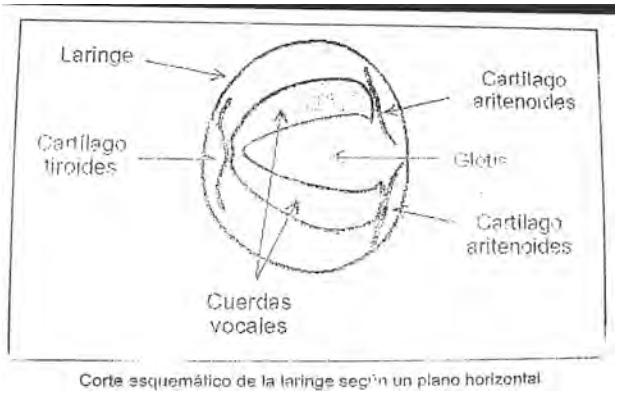


Imagen 6. Corte esquemático de la laringe según un plano horizontal

Los pliegues vocales pueden ser manipulados de dos maneras principales por medio de la contracción de conjuntos de músculos conectados a los varios cartílagos de la laringe. Un tipo de ajuste cambia la rigidez de los pliegues modificando la frecuencia de la vibración. La otra manipulación cambia la separación entre los pliegues. La vibración de los pliegues solo ocurre para ciertas combinaciones de rigidez y separación. Cuando los pliegues vocales son posicionados apropiadamente, y la presión se eleva en las vías aéreas debajo de la glotis, los pliegues vibran, y el flujo de aire que atraviesa la glotis se modula periódicamente. El espectro de tal modulación, es rico en armónicos, y el flujo periódico constituye una fuente que provee la excitación para las vías aéreas sobre la laringe como un *monopolo acústico* (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1566). Cuando las partes terminales de los dobleces están separadas, los pliegues están abiertos, lo que constituye la posición para respirar. Cuando se genera una abertura estrecha es posible generar murmullos compuestos de ruido blanco (*hiss*), y cuando las partes terminales están juntas, las cuerdas están cerradas y se puede deglutir.

La *frecuencia* de la vibración de los pliegues durante el habla normal se encuentra usualmente entre 170-340 Hz para mujeres adultas, 80-160 Hz para hombres adultos, y 250-500 Hz para niños menores. Sin embargo, tales frecuencias pueden extenderse para la voz cantada dependiendo del registro del(a) cantante (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1566).

Imagen 7. Cuerdas vocales abiertas y cerradas

c) Tracto Vocal

Es la zona formada por la faringe y las cavidades bucal y nasal, donde se modulan los sonidos provenientes de la laringe para producir los sonidos resultantes. La forma del impulso de la glotis puede variar enormemente dependiendo de la fuente, y el esfuerzo vocal. La longitud y forma de un tracto vocal particular determinan las resonancias en el espectro de una señal de voz. En promedio, un tracto vocal es aproximadamente de 17 cm (entre 14-18 cm) de longitud y cuando se encuentra en reposo, los formantes están equiespaciados encontrándose el primero alrededor de los 500 Hz, el segundo alrededor de los 1500, el tercero por los 2500, el cuarto alrededor de 3500, el quinto por 4500, etc. (Dodge y Jerse: 1997, p. 222). El volumen de la vía aérea está en general, entre 40-90 cm^3

con un promedio de área transversal de 3.5 cm^2. A medida que el tracto vocal adquiere la forma para producir diferentes vocales y consonantes, el área transversal puede variar desde cero a 10 cm^2 o más. Tales dimensiones hacen que la propagación del sonido en el tracto vocal sea aproximadamente unidimensional en el intervalo de frecuencias hasta de 5 kHz, lo que permite modelar al tracto vocal (sobre tal intervalo) como un tubo con área transversal variable en el cuál ondas unidimensionales se propagan (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1567). Sin embargo, uno de los problemas de intentar modelar el tracto vocal, es que no se puede hacer exactamente como si fuera un tubo acústico lineal simple, ya que cada parte de este está constituido por tejido suave, el cual absorbe la vibración de manera diferente que si las paredes fueran rígidas.

El re-posicionamiento de los articuladores altera la forma del tracto vocal, cambiando así las frecuencias de los formantes, particularmente los más graves. El cambio en la forma de las vías aéreas, tiene dos funciones específicas en el habla. La primera es la de formar una constricción angosta en algún punto de las vías para producir un sonido consonante. La formación o liberación de la constricción produce un tipo particular de discontinuidad o modulación abrupta en las características del sonido. La segunda función es la de dar forma a las vías para lograr frecuencias naturales particulares, para que la fuente sonora en, o sobre la glotis sean filtradas por tales resonancias (siendo las frecuencias de los dos o tres formantes más graves las claves necesarias para la diferenciación fonémica de las vocales).

Hay cinco principales estructuras anatómicas que pueden ser manipuladas para cambiar la forma del tracto vocal: 1) Un conjunto de músculos constrictores que rodea la zona paralingeal del tracto, cuya contracción forma una constricción de las vías en tal zona. 2) El cuerpo de la lengua, que puede ser desplazado vertical y horizontalmente contrayendo varios grupos musculares. 3) El filo de la lengua, que forma una extensión del cuerpo de la lengua y puede ser deformada o desplazada hacia arriba para formar una constricción en la región que se extiende de los dientes al paladar duro. 4) Los labios, que pueden ser deformados y desplazados para formar una constricción angosta o para aumentar el largo efectivo de la vía aérea de la glotis a la apertura labial. 5) Finalmente, el paladar suave, que puede ser elevado o descendido para variar el área transversal del pasaje de la faringe superior a la cavidad nasal, o para cerrar tal paso por completo (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1567). A continuación se presenta un diagrama esquematizado de las secciones revisadas que constituyen la totalidad del Sistema Generador de Voz:

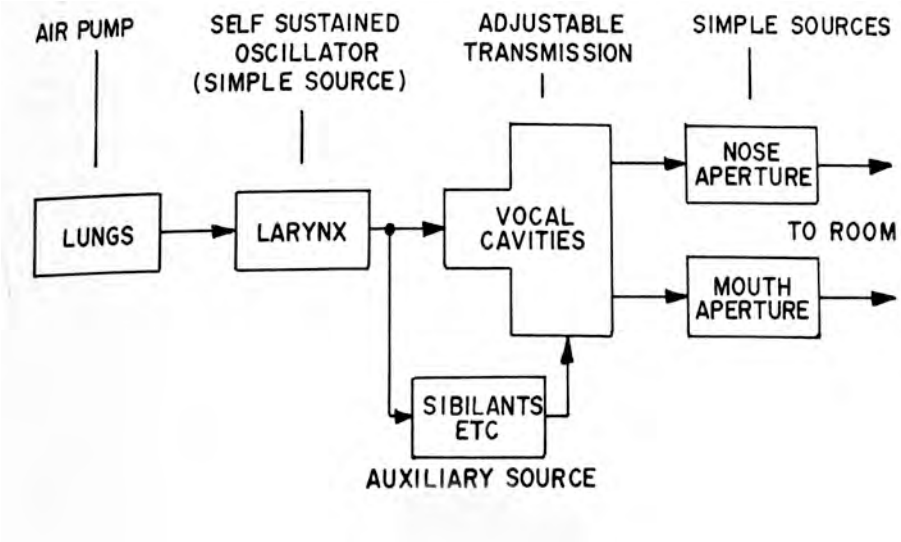


Imagen 8. Diagrama esquematizado del Sistema generador de Voz

El la *Imagen 8*, se representa la estructura del tracto vocal, siendo el recuadro titulado, "sibilantes", no una sección física sino un recordatorio de que existen algunos sonidos que requieren una fuente aleatoria auxiliar que puede estar localizada casi en cualquier lugar de la cavidad vocal (Benade: 1990, p. 362). La laringe puede ser considerada una fuente sencilla, que alimenta la señal hacia una cavidad elongada y estrecha, de estructura compleja. La presión sonora en cualquier punto de la cavidad vocal dependerá drásticamente tanto de la frecuencia de excitación como del punto de observación, y la respuesta acústica dependerá de si los componentes frecuenciales de la fuente coinciden o no con los *modos característicos de vibración* de la cavidad. Sobre la apertura de la boca, el *flujo oscilatorio* de aire, dependerá de la relación entre la frecuencia de excitación de la laringe y las varias resonancias de la cavidad vocal, fungiendo como la fuente de sonidos así como pueden escucharse en el cuarto (Benade: 1990, p 362).

III. Fundamentos de *Acústica*

El Sonido y la Voz

El sonido es, comúnmente, el resultado de la vibración de alguna fuente material;[2] una *onda* de presión *longitudinal* formada por compresiones y rarefacciones de las moléculas y partículas que se propaga tanto en gases, como el aire, así como en sólidos y líquidos en una dirección paralela a aquella sobre la que se aplicó la fuerza.[3] Los sonidos del habla son producidos mediante la modulación del flujo de aire a través de constricciones en las vías aéreas entre la laringe y los labios.

---

[2] Algunos sonidos (sordos), son una consecuencia de la turbulencia en el flujo y por lo tanto de la generación de ruido turbulento. Por otro lado, fuentes de sonidos transientes pueden ser generadas al incrementar o disminuir la presión detrás de un bloqueo en vías aéreas y después rápidamente abrir tal bloqueo, causando un cambio abrupto en la presión.

[3] Aunque las configuraciones alternantes de compresión y rarefacción de moléculas de aire sobre el camino de la fuente de energía, a veces son descritas por una gráfica sinusoidal, donde las crestas de la curva corresponden a los momentos de máxima compresión y los valles a los de máxima rarefacción, su uso es únicamente una conveniencia de notación para graficar variaciones de presión local contra tiempo, ya que el sonido no forma una onda transversal, y las partículas de aire solo se encuentran oscilando en su sitio sobre la línea de aplicación de fuerza.

Como se revisó en la sección anterior, la distinción fundamental entre tipos de sonidos en el habla son los sonidos sonoros y los sordos. Los primeros surgen de la vibración de los pliegues vocales, lo cual a su vez genera cambios quasi-periódicos en el espacio entre los pliegues (la glotis), modulando a su vez el flujo de volumen a través de la glotis. Tienen en su estructura temporal y espectral un patrón más o menos regular, que los segundos no tienen. Los pliegues vocales vibran con ritmos desde 60 Hz para un hombre grande, hasta 300 Hz o más para una mujer pequeña o niño. El ritmo de apertura y cierre de los pliegues vocales en la laringe durante la fonación de sonidos sonoros es conocida como *frecuencia fundamental,* debido a que marca la base periódica sobre la que se construyen todos los componentes armónicos superiores con los que las resonancias de las cavidades orales y faringeales contribuyen. La frecuencia fundamental también contribuye más que cualquier otro factor a la percepción del *tono* en el habla (Huang, Acero y Hon: 2001, pp. 25-26).

La *forma de onda* de las variaciones de presión creadas por el ciclo glotal, puede ser descrita como un flujo periódico en centímetros cúbicos por segundo. Como se muestra en la *Imagen 9*, durante el tiempo que toma un ciclo no hay flujo de aire en la porción cerrada inicial. Después de ello, la glotis se abre y el volumen de flujo de aire se incrementa. Después de un pico corto, el pliegue recupera su posición inicial y el flujo de aire declina hasta que el cierre total es alcanzado iniciándose el siguiente ciclo. Una medida común es el número de tales ciclos por segundo o Hertz, lo que es igual a *la frecuencia fundamental* (Huang, Acero y Hon: 2001, pp. 26-27).

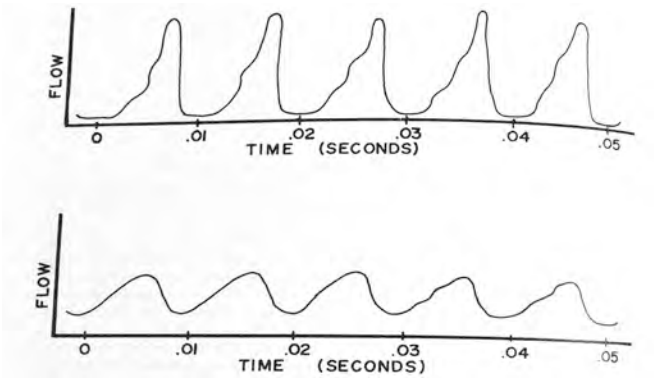


Imagen 9. Variaciones de presión creadas por el ciclo glotal

El instrumento musical: Configuración mecánica, flujos aéreos, presiones y fuentes sonoras en el tracto vocal

La voz cantada tiene una considerable pertinencia musical. Es claro que la voz humana puede producir señales acústicas con ritmos de repetición que pueden variar sobre un amplio intervalo. Además, el hecho de que un(a) cantante pueda enunciar diferentes sonidos de manera sostenida mientras mantiene el tono sugiere que las amplitudes de los componentes sinusoidales están sujetos

a control independiente. Para comprender cómo es que el aire que pasa por los pliegues vocales puede mantener sus oscilaciones, se deben mantener en mente algunos hechos y consecuencias del comportamiento de fluidos y de su movimiento, que se mencionan a continuación.

Todo fluido *fluye* de regiones de alta presión a regiones donde la presión es baja. En caso de que se presente un incremento en la velocidad de flujo durante el movimiento de un punto a otro, se puede deducir que la presión en el punto de mayor velocidad debe ser menor que en el punto de menor velocidad de donde el fluido llegó. Esto es una consecuencia de que no es posible acelerar ningún objeto material, si no existe un exceso de fuerza aplicado al mismo (2a *ley de Newton*). Por otro lado, cuando un fluido viaja por un ducto continuo y largo, se espera que la velocidad de flujo sea mayor en las zonas angostas que en las partes anchas, ya que un fluido que corre en un ducto que no permite pérdidas, siempre presentará un volumen fijo de fluido pasando por cualquier punto, por unidad de tiempo. Debido a que en zonas donde la sección transversal es grande, muchos pequeños "pedazos" de material fluyendo lentamente se acumularán, mientras que en las secciones más angostas correrán rápidamente a través de la constricción en una sola línea, la presión del fluido será entonces, menor en las secciones angostas y mayor en las secciones anchas (*Teorema de Bernoulli de flujo constante*). Por último, la presencia de fricción debido a la viscosidad en el fluido, o entre el fluido y las paredes que lo contienen, reduce el contenido total del fluido que pasa, por unidad de tiempo a través del sistema bajo la influencia de alguna presión sobre la fuente (Benade: 1990, p. 364).

Ahora bien, en el tracto vocal, el flujo de aire proveniente de los pulmones, fluye a través de un ducto de diámetro amplio que corresponde a la tráquea, el aire entonces atraviesa una constricción (los pliegues vocales) saliendo en seguida a una porción alargada del ducto que constituye el inicio del tracto vocal. En la *Imagen 10*, se puede apreciar un símil mecánico de tal sistema (Benade: 1990, p. 365), donde la frontera superior de la constricción consiste esencialmente de una masa M acoplada a un resorte que tiene un coeficiente de rigidez S. La masa es libre de oscilar suavemente hacia arriba y abajo sobre una guía perfectamente a la medida. El sistema masa-resorte es la representación de uno de los pliegues vocales (el segundo se movería simétricamente con respecto al primero). El amortiguamiento viscoso, D, es provisto por la grasa que sella la guía sobre la que la masa se monta, y evita que el aire escape. Si no se envía aire por el análogo mecánico de la laringe, entonces se puede demostrar que la *frecuencia natural de oscilación* de la masa M será proporcional a la cantidad $(S/M)^{1/2}$, y que si se jala y suelta la masa, las oscilaciones decaerán con un tiempo de vida media proporcional a M/D. Ésta es la frecuencia natural que un(a) cantante cambia al moverse de un *tono musical* a otro (Benade: 1990, p. 365).

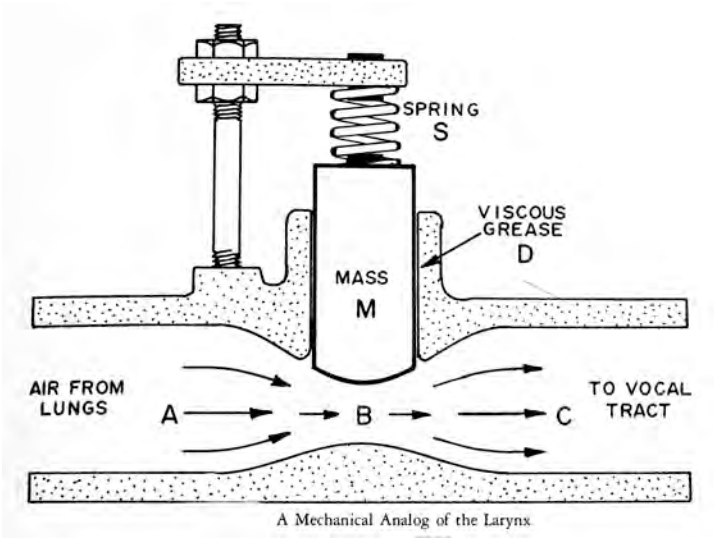


Imagen 10. Un análogo mecánico de la laringe

Por otro lado, al haber una presión subglotal fija, el flujo de aire en el tracto vocal es controlado mediante ajustes en la configuración de la laringe y/o de algunas estructuras supralaringeales. La caida en presión Δ*P*, depende del área de corte transversal, del flujo de aire *U,* y una constante que depende de la geometría de la constricción, obedeciendo a la relación:

$$\Delta P = k\frac{\rho U^2}{2A^2}$$

Donde *A* es el área de corte transversal de la constricción, ρ es la *densidad* del aire, y *k* es una constante cercana a la unidad, pero que varía de 10 a 20% dependiendo de la forma de la constricción (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1567). Cuando el flujo de aire se inicia, la presión en el punto de la constricción se reducirá con respecto a la presión en el ducto de ambos lados de tal constricción. Si la masa se mueve hacia abajo, cerrando aún más la constricción, entonces sucederán dos efectos opuestos: al cerrar la apertura la *velocidad de flujo* aumentará en la constricción, reduciéndose más la presión, lo cuál tenderá a succionar la masa aún más. La *fricción* agregada por reducir la apertura, disminuirá si la *presión* de los pulmones se mantiene igual, es decir, el volumen total de aire que pasa por segundo, haciendo que la presión dependiente del flujo no cambie como se espera (Benade: 1990, p. 365). La presencia del flujo de aire causa que la masa M experimente un *fuerza aerodinámica* con dos componentes, una hacia dentro del sistema, más una fuerza que fluctúa mientras la masa vibra hacia adentro o hacia fuera. Mientras que la masa, al actuar la fuerza estable sobre ella, adquiere una nueva posición de equilibrio (con una reducción de la apertura), la fuerza oscilatoria actúa en contra del resorte como una fuerza adicional que tiende a jalar a la masa alejándola aún más de la nueva posición de equilibrio. Esto puede simplificarse a la acción de un sólo resorte que tiene un coeficiente de dureza ligeramente menor. Se puede concluir entonces, que la frecuencia natural de oscilación del símil del pliegue vocal, disminuye ligeramente por la existencia de un flujo de aire pasando.

Para explicar el inicio o mantenimiento de oscilaciones considerando, en el símil mecánico, la acción del amortiguamiento de la grasa sobre la guía, debemos recordar que la fuerza de amortiguamiento actúa en todo momento en contra de la dirección del movimiento de oscilación de los pliegues. Para mantener cierta oscilación, es necesario entonces aplicar una fuerza periódica que actúe predominantemente en la dirección del movimiento. En el caso de un flujo no constante, el

Teorema de Bernoulli de flujo constante no se mantiene, ya que debido a la *inercia* del aire que se mueve, la velocidad de flujo no puede instantáneamente reajustarse a medida que la apertura cambia. Esto quiere decir que la variación sinusoidal en la apertura, determinada por la oscilación de la masa, cambia debido al paso de un flujo aéreo cuyas variaciones son retrasadas una cantidad pequeña. Así, la fuerza de *Bernoulli* oscilatoria inducida, alcanza sus máximos y mínimos en instantes de tiempo ligeramente después que los máximos y mínimos en amplitud de la masa. Los instantes de mayor interés dinámico son aquellos en los que la fuerza de *Bernoulli* actúa en la misma dirección que el movimiento, debido a que son los instantes en los que la fuerza contribuye a mantener la oscilación. Tales intervalos son de hecho, más largos que los intervalos en los que la fuerza tiende a disminuir la oscilación.

Ahora bien, si se *modela* el sistema completo, entonces se debe considerar una analogía mecánica donde dos masas superiores acopladas entre sí, y a otras dos masas inferiores, también acopladas entre sí, presentan un elemento de rigidez determinado. Cuando los pliegues vocales son colocados cercanamente, y una presión subglotal es aplicada, éstos vibran. La presión genera una fuerza hacia afuera sobre las masas inferiores que se aceleran en respuesta a la fuerza. Al alcanzar cierto valor de desplazamiento, el acoplamiento mecánico causa que las masas superiores se separen. A medida que esto sucede, continúa el flujo de aire a través de la glotis y la presión de *Bernouilli* cae, teniéndose una disminución de la presión en la zona baja de la glotis. Esta caída de presión hace que la fuerza hacia afuera ejercida sobre las masas inferiores, decremente y éstas vuelvan a juntarse, haciendo eventualmente que también las masas superiores corten abruptamente el paso del aire.

El espectro de tal flujo periódico tiene una serie de armónicos en los múltiplos de la frecuencia fundamental. Durante frecuencias por encima de 1kHz, estas amplitudes decrecen a un ritmo de 6 dB/octava y a veces a un ritmo mayor. La frecuencia de vibración de los pliegues vocales es controlada contrayendo los músculos que causan un cambio de tensión en los pliegues vocales, y alguna variación en la forma de onda es lograda variando la configuración de reposo de los pliegues. Cambios en la presión subglotal causan un incremento o decremento en la amplitud de los pulsos glotales (pulsos *quasi*-periódicos como los que se muestran en la *Imagen 9*).

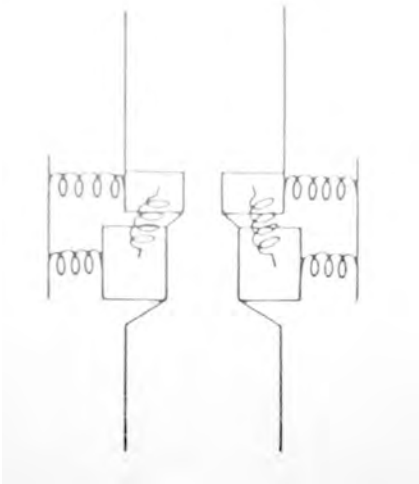


Imagen 11. Analogía mecánica completa

Existen diferentes clases de fuentes de excitación para la voz, de las cuales la *Fonación*, se refiere a la oscilación de las cuerdas o pliegues vocales de los cartílagos aritenoides. La apertura y cierre de los pliegues secciona la salida del aire en llamados pulsos glotales, provistos de una frecuencia fundamental. La forma de onda es aproximadamente triangular (ver *Imagen 9*). Como

consecuencia a su forma, las altas frecuencias disminuyen su amplitud a 12 dB/octava. Su naturaleza paso-bajas proporciona un espectro con una fuerte fundamental y armónicos progresivamente débiles.

Por otra parte, cuando la velocidad del flujo aéreo a través de una constricción en el tracto vocal o en la laringe es suficientemente elevada, se genera turbulencia y puede resultar en la generación de ruido, usualmente como consecuencia del golpeteo de una superficie o un obstáculo por parte del flujo de aire. El ruido se puede considerar como generado por fuerzas fluctuantes ejercidas por el obstáculo en la vía. Esta situación puede ser modelada como un *dipolo* y fuente de presión sonora localizada en la vía aérea en la región donde la *turbulencia* ocurre. El espectro de esta fuente depende de la configuración de la constricción y del obstáculo. La *amplitud* de tal fuente incrementa aproximadamente como el cubo de la velocidad de flujo (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1568). El *Susurreo* y los sonidos provenientes de él, son generados por la laringe. En ellos, los pliegues vocales están juntos, gracias al cartílago aritenoides, pero en lugar de sellar completamente la glotis existe una pequeña apertura triangular. El aire genera turbulencias que ocasionan ruido de banda ancha, el cual sirve como señal excitadora. Los sonidos del susurreo, son más débiles que las fonaciones ya que implican menos volumen de aire, y tienen mayor energía en altas frecuencias.

Otra fuente, que ya se ha mencionado, es la usada en el habla durante los sonidos transientes generados cuando cierta cantidad de presión se acumula detrás de un bloqueo en el tracto vocal y súbitamente se libera. Estos sonidos también pueden ser generados creando vacío parcial en un espacio cerrado, como el formado por la lengua y los labios, y luego liberando el vacío. Este mecanismo es utilizado para generar “clicks” (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1568). Asimismo, la *Vibración* constituye una fuente sonora alternativa *quasi*-periódica y que puede ocurrir en muchos lugares del tracto vocal. Como ejemplo, la “r” vibrante involucra la vibración de la lengua contra el paladar. Estas vibraciones pueden ocurrir con o sin fonación.

La producción del habla se puede dividir en dos fases distintas: (1) la producción de un sonido audible y cuyos mecanismos abordamos en la presente sección, y (2) el control ejercido sobre éste sonido para producir un fonema concreto, cuyas particularidades revisaremos en la siguiente sección.

IV. *Fundamentos de Fonética*

Características de la Articulación y Fonética del Español

La voz humana constituye una fuente sonora que puede producir señales cuyos ritmos de repetición varíen en un amplio intervalo. Al hablar o cantar, una persona continuamente cambia la forma de su cavidad vocal, asociando a cada sonido una forma bien definida, y consecuentemente, también a un patrón particular de respuestas suaves o intensas a los varios componentes sinusoidales del flujo aéreo controlado por las cuerdas vocales (Benade: 1990, p. 362). Una vez que la onda de presión encuentra las otras partes del tracto vocal; la faringe, y las cavidades oral y nasal, las frecuencias características y resonancias de la cavidad en su conjunto, le dan forma al *espectro* y generan los picos energéticos, o *formantes*, esenciales para formar los diferentes sonidos que constituyen al lenguaje.

Por ejemplo, para emitir el sonido /ah/, la lengua adopta una posición situada hacia adelante y hacia abajo, al mismo tiempo que se eleva el paladar, lo que genera frecuencias formantes a 730, 1090 y 2440 Hz aproximadamente. Nuestro oído aprecia el fonema representado por /ah/ como el correspondiente a una /*a*/ larga y de entonación grave. Al entrar la ondas sonoras producidas por las cuerdas vocales en la cavidad del habla, la cavidad empieza a resonar. Cada *formante* tiene una *curva de respuesta***,** y así, aunque ninguna de las frecuencias del espectro de la onda sonora sea exactamente igual a cualquiera de las frecuencias formantes, la cavidad resuena a frecuencias cercanas a las formantes. La siguiente imagen muestra la curva de respuesta combinada de las tres formantes que se crean al emitir el sonido /ah/.

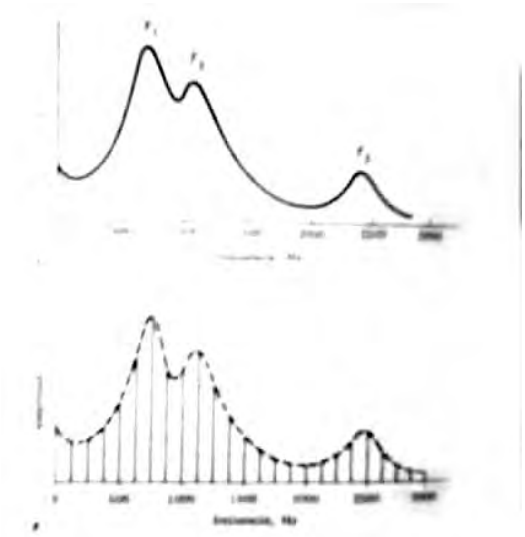


Imagen 12. Formantes y forma de onda /ah/

Por regla general, tres o cuatro armónicos se hayan suficientemente cerca de cada formante para aumentar su valor por resonancia. Ello se puede ver de la figura inferior en la *Imagen 12*, que es la superposición del espectro del sonido (*Imagen 13*) con el perfil de formantes resultado de la configuración adoptada.

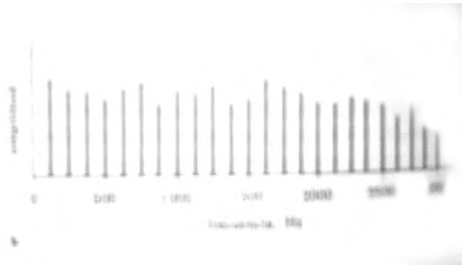


Imagen 13. Espectro del sonido

Si la fundamental f_1 es 125 Hz, el armónico más próximo a la primera formante (F_1= 730 Hz) en /ah/ es f_6 = 6 x 125 Hz = 750 Hz. Éste tendrá una gran amplitud en la cavidad. Los armónicos contiguos, f_5 = 625 Hz y f_7 = 875 Hz, aumentarán su valor pero en menor medida. Un análisis similar es necesario para con todos los sonidos existentes en un idioma, y su clasificación es un primer paso hacia conocer todas sus características acústicas[4] (Cromer, *Encyclopedia of Acoustics:* 1997, p. 321).

Clasificación de fonemas

1) Acción de los Pliegues Vocales

Algunos fonemas pueden ser distinguidos unos de otros dependiendo de si los pliegues vocales vibran o no durante el intervalo de tiempo en que la constricción que genera el sonido está en su lugar. La clasificación debida a la acción de las cuerdas vocales se divide en:

- *Fonemas sonoros que* se producen por la fonación o vibraciones en las cuerdas vocales, para los que existen diferentes modos de vibración, llamados registros. Los sonidos resultantes de la fonación,

[4] Para conocer más sobre las características acústicas específicas para los sonidos del idioma español ver Bernal et al, *Reconocimiento de voz y fonética acústica*, Madrid: Alfaomega ra-ma, 2000.

causados por las vibraciones *quasi*-periódicas de la glotis, se llaman ***sonoros*** y son ejemplos de ellos las vocales.

- *Fonemas sordos*, para los que no existe vibración en las cuerdas vocales. Para éstos, la excitación es la turbulencia de aire. Las consonantes como /f/, /s/, /p/ y /k/ son sonidos sordos. La localidad de la turbulencia es controlada regulando el tracto vocal en lugares apropiados como pueden ser garganta, dientes o labios.

2) Punto de Articulación

Los sonidos fonéticos pueden clasificarse a su vez, en relación al punto de articulación. Es importante mencionar, sin embargo, que tal clasificación presenta variaciones de autor a autor, ya que debido a los fonemas adyacentes, tal punto puede o no alcanzarse perfectamente y sufrir variaciones topológicas o temporales. A continuación se presenta la clasificación de sonidos por punto de articulación:

a. Los bilabiales: Se pronuncian con los labios.

b. Los labiodentales: La punta de la lengua hace contacto con la parte posterior del diente incisivo superior.

c. Los interdentales: La lengua se sitúa entre los dientes.

d. Los alveolares: La punta de la lengua se acerca o toca la punta alveolar en el techo de la boca.

e. Los palatares: La lengua se apoya en el paladar.

f. Los velares: La lengua toca el velo del paladar.

V. Los ladrillos de construcción del español

Los fonemas son usualmente clasificados en 4 clases: vocales, semivocales, diptongos, y consonantes. A continuación se revisan las características de cada una de ellas.

1. Las vocales

Las vocales son sonidos sonoros para los cuales los articuladores del tracto vocal asumen una posición fija, donde el tracto está bien abierto y la única función de la boca es una variación del timbre (Dodge y Jerse: 1997, pp. 222-223). Son acústicamente caracterizadas por sus frecuencias formantes y se distinguen porque sus estructuras de formantes son claras, fruto de la emisión del flujo de aire por el conducto bucal sin mucha resistencia, y con las cavidades resonadoras potenciando los armónicos distintos de cada vocal (Bernal et al: 2000, p.43). Las vocales son normalmente producidas por una fuente en la glotis y las vías sobre la glotis de tal forma que se generan adelgazamientos sobre la vía que no son suficientes para lograr un acumulamiento de presión. El comportamiento acústico de las vocales es especificado por el *área de sección transversal* del tracto como función de la distancia de la glotis. Su función de transferencia es la relación entre las amplitudes complejas de la *velocidad del volumen* en los labios, U_o, y en la glotis, U_g [$T(f)=U_o/U_g$]. Para un tubo de longitud l, esta función de transferencia esta dada por:

$$T(f) = \frac{1}{\cos(2\pi f l / c)}$$

Esta función, contiene solo *polos* a las frecuencias $(2n+1)c/4l$, (siendo n=0,1,2…), y constituyen las formantes o frecuencias naturales del tracto vocal cuando está cerrada del lado de la glotis. La velocidad del sonido c, a temperatura corporal es de 35,400 cm/s. Si se incorporan las perdidas acústicas y la impedancia debida a las paredes del tracto, viscosidad y conducción de calor, los polos resultan perturbados ligeramente. Datos experimentales, así como análisis teóricos muestran que los anchos de banda de las prominencias en $|T(f)|$ se encuentran entre 60-100 Hz para el primer formante, 80-150 Hz para el segundo y progresivamente mayores para formantes más altos (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1570). Las amplitudes de las prominencias espectrales que corresponden a las formantes decrecen a medida que aumenta la frecuencia. El nivel general de presión sonora de las vocales es determinada sobre todo por la amplitud espectral de la primera prominencia. Las formantes cambian cuando el tubo no es uniforme, y pueden ser calculadas de la resolución de la *ecuación de onda* para un tubo no-uniforme usando el *método de perturbaciones* donde la distribución de *energía potencial* y *cinética* en la *onda estacionaria* para cada

uno de los modos, es calculada.

Son las posiciones del cuerpo de la lengua las que dan lugar a los patrones en las formantes características de las vocales, y estas son: el cuerpo de la lengua hacia adelante y elevado para la /i/, un cuerpo de la lengua para atrás y hacia abajo para la /a/, y el cuerpo de la lengua elevado y hacia atrás en conjunto con una extensión y adelgazamiento de los labios para la /u/. Existen diferencias en las amplitudes relativas de las prominencias espectrales para las diferentes vocales como consecuencia de las interacciones entre las contribuciones de formantes individuales en la función de transferencia todo-polo. De hecho, diferentes idiomas difieren en la cantidad y tipos de vocales, existiendo también diptongos y vocales nasales y no nasales. Para producir una vocal nasal se crea una apertura entre la cavidad oral y la nasal, desplazando hacia abajo el paladar suave. Esta configuración es modelada por medio de tubos interconectados como se muestra en la imagen:

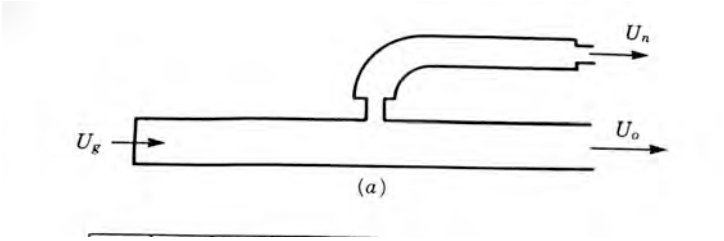


Imagen 14. Acoplamiento de las cavidades nasal y oral

Este acoplamiento causa una modificación de la función de transferencia de la fuente glotal hasta la salida, que ahora es la suma de las velocidades de volumen de la nariz, U_n, y la boca, U_o. La modificación principal es la introducción de *polos* (resonancias) y *ceros* (antiresonancias) adicionales en la función de transferencia, seguido de cambios de la frecuencias de los formantes relativas a aquellos de la configuración no nasal. El polo o cero adicional más bajo es el más consistente y perceptualmente es el más saliente. Si la mayor parte de la energía es radiada por la boca, los ceros en la función de transferencia son las frecuencias para las cuales la impedancia viendo hacia la cavidad nasal es cero (Stevens, *Encyclopedia of Acoustics:* 1997, pp. 1570-1573).

En la lengua española existen 5 vocales principales (/a/, /e/, /i/, /o/, /u/), que pueden ser desdoblados en alófonos orales y nasales según la región geográfica o posición en la palabra. Un alófono nasal se produce cuando su correspondiente fonema vocálico se encuentra entre una pausa y una consonante nasal, o entre dos consonantes nasales y tienen la característica de que la intensidad de su primer formante se reduce (Bernal et al: 2000, p. 43). Las vocales pueden clasificarse por su punto de articulación como palatales (anterior a la boca, /e/, /i/), centrales (central a la boca, /a/) y velares (posterior a la boca, /o/, /u/), o por la apertura o modo de articulación de la boca en máxima (/a/), media (/e/, /o/), y mínima (/i/, /u/).

Recapitulando, desde el punto de vista articulatorio, las vocales se clasifican o caracterizan de acuerdo a tres parámetros articulatorios:

1. La altura de la lengua: las vocales son altas (o cerradas), medias o bajas (o abiertas).

2. La posición de la lengua con respecto al eje antero-posterior de la cavidad bucal: las vocales pueden ser anteriores, centrales o posteriores.

3. La acción de los labios distingue entre vocales redondeadas (o labializadas) y no redondeadas (o no labializadas).

Los dos primeros parámetros dan lugar al triángulo de HELWAG, (representación vocálica que muestra el esquema articulatorio) de las vocales:

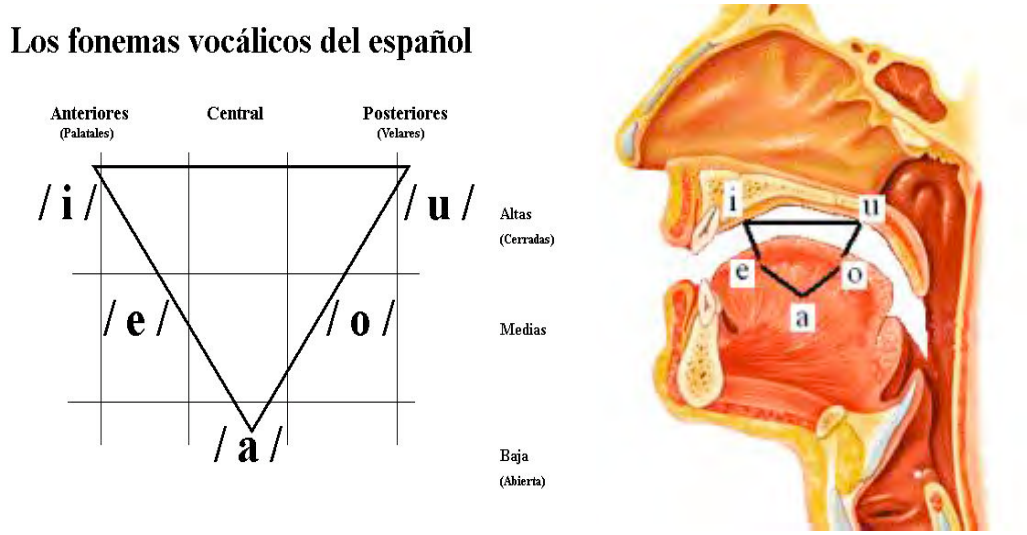


Imagen 15. El triángulo de Helwag

Es evidente que de alguna u otra manera, la voz humana cuenta con cierto mecanismo de control independiente de las amplitudes de sus sinusoidales, ya que es posible, manteniendo el tono igual, enunciar las distintas vocales, que constituyen probablemente, los elementos musicales del canto de mayor relevancia.

2. Las semivocales

Las semivocales son transitivas; sonidos sonoros que carecen de caracterización fija de la estructura resonante. En lugar de ello, la semivocal adquiere las características de resonancia de la vocal de la cual viene o va (Dodge y Jerse: 1997, p. 221).

3. Los diptongos

Es una sucesión de dos sonidos vocales y tiene una transición continua de una vocal a la otra. La articulación de un diptongo comienza en o cerca de la posición de una de las vocales y procede suavemente hacia la otra (Dodge y Jerse: 1997, p. 223). La vocal con la mayor abertura vocal se llama núcleo silábico, mientras la vocal con menor abertura se llama sílaba marginal. Un diptongo puede ser creciente o decreciente. El primero existe cuando el núcleo silábico precede al margen silábico, y el segundo cuando sucede lo contrario. Para diptongos crecientes, al margen silábico se le llama semi-consonante habiendo dos semi-consonantes crecientes [j] y [w] y ocho diptongos crecientes. Para diptongos decrecientes el margen silábico se llama también semi-vocal. Las semi-vocales son distintas a las semi-consonantes por su posición en el diptongo y la forma de articulación. Existen dos semi-vocales [i] y [u], y seis diptongos decrecientes (Herrera: 2006, cap. 2, p. 16).

Existen tres tipos de diptongos (Ríos: 1999, 5.5.2):

1) crecientes: inician en una vocal débil, seguida de una fuerte (/ua/, /ue/, /uo/, /ia/, /ie/, /io/)
2) decrecientes: inician en una vocal fuerte, seguida de una débil (/ai/, /ei/, /oi/, /au/, /eu/, /ou/)
3) homogéneos: formados por dos vocales débiles (/iu/, /ui/)

4. Las consonantes

Los fonemas consonantes son producidos mediante una constricción relativamente angosta localizada sobre la vía aérea sobre la laringe, mientras que las vocales se consiguen con una vía más abierta. Pueden ser clasificadas dependiendo del articulador responsable de la formación de la constricción, el grado de la constricción, el estado de la glotis y los pliegues vocales cuando se forma la constricción, así como si es que se acumula presión detrás de la constricción o no (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1573). En éste sentido, existen cinco diferentes clases de consonantes: oclusivas, nasales, vibrante simple, vibrante múltiple, fricativas, y lateral africativa. Cada una de ellas es el resultado de un arreglo distintivo y el uso de articuladores, pero todas incluyen un bloqueo parcial o total del tracto vocal. Cuando las consonantes son aspiradas, las resonancias asumen las posiciones de la vocal que le sigue, ya que no hay cambio en la posición de ningún articulador que cambie entre los dos fonemas (Ríos: 1999, 5.5.4). En caso de que la constricción sea suficientemente angosta, se genera una caída de presión atravesando la constricción cuando una presión subglotal es aplicada, esto asumiendo que no existe una salida alternativa del flujo, como podrían ser los pasajes nasales. Dentro de ésta clase de consonantes, las que pueden llamarse consonantes plosivas, se producen cuando existe un bloqueo total, y el flujo de aire a través de la constricción se reanuda sólo después de que la obstrucción se libera. A éstas se les conoce como consonantes oclusivas**,** y en ellas tanto el tracto vocal en el punto de articulación y el pasaje se encuentran cerrados. En el momento de la liberación, una corta explosión de ruido turbulento se produce por el rápido flujo a través de la constricción (Stevens, *Encyclopedia of Acoustics:* 1997, p. 1573), y una exhalación cortante con característica de respiración transitoria es generada (Ríos: 1999, 5.5.4). Si el transitorio es abrupto y limpio, el sonido es una oclusiva, /p/; y si es gradual y turbulento, el sonido cae en uno parecido al fricativo como /j/. Las consonantes "oclusivas" también pueden ser sonoras o sordas. Tanto las consonantes "oclusivas" sonoras, como las sordas, son distinguidas por la ubicación de la constricción hacia el frente, en medio y hacia atrás de la boca. Para producir "oclusivas" sonoras, la glotis vibra mientras se genera la presión y se libera la constricción. Para formar las "oclusivas" sordas, la "oclusiva" es seguida por la fricada (Bernal et al: 2000, p. 45). Los sonidos oclusivos sordos presentan una zona de silencio seguida por una breve barra de explosión vertical, que tiene mayor duración temporal en el sonido.

En el caso de las *consonantes fricativas*, el tracto vocal se encuentra abierto parcialmente, con el velo cerrado y se genera ruido en el punto de articulación (Ríos: 1999, 5.4.5). La fricación es similar al "susurreo" ya que el aire turbulento genera ruido de banda ancha, pero existe un lugar de articulación adicional en el tracto vocal. Puede ocurrir con o sin fonación. El lugar de articulación se encuentra cerca de los labios y sólo una pequeña parte del tracto vocal está entre la fuente de excitación y el aire de salida; por lo tanto, la modulación producida por el tracto está limitada en extensión y complejidad. Mientras algunas son creadas por una constricción cerca de los labios, algunas otras lo son por una constricción cerca de los dientes. Algunas incluso pueden ser creadas por constricción al centro del tracto vocal, mientras que otras en la parte posterior de la boca. La

fricación es de menor amplitud que la fonación y tiene una proporción mucho más alta de altas frecuencias, sin embargo los sonidos fricativos son más sonoros que los susurros. Un ejemplo de *consonante fricativa* es el sonido /s/, para la generación del cual, el flujo de aire es dirigido hacia los dientes incisivos inferiores, y la principal fuente de ruido de turbulencia está localizada en la vecindad de estos dientes. Un espectro similar al de /s/ se obtiene de la breve explosión sonora en la liberación de las consonantes oclusivas, /d/ y /t/, que se producen con una constricción en un lugar similar que para /s/ (Stevens, *Encyclopedia of Acoustics:* 1997, pp. 1573-1574). Las *consonantes fricativas* pueden ser tanto *sonoras* como *sordas*. Los puntos de constricción de las fricativas sonoras son los mismos que para sus contrapartes sordas. Existen dos fuentes de excitación simultáneas para cada una de las fricativas sonoras: la glotis y la turbulencia creada en el sitio de la constricción. Asimismo, las realizaciones *fricativas* de los fonemas oclusivos pueden clasificarse también dentro de éste grupo de sonidos (Bernal et al: 2000, p. 48).

En español, existen fonemas *africados*, donde aparece un cierre inicial del tracto vocal seguido de una expiración gradual que produce turbulencia (Ríos: 1999, 5.4.6). Son formadas conectando un fonema oclusivo y uno fricado (Bernal et al: 2000, p. 47).

Las *consonantes nasales* (p.e. */m/, /n/*) se caracterizan porque el tracto vocal está cerrado y el velo abierto (Ríos: 1999, 5.4.2), y son conseguidas bajando el paladar suave de tal manera que la onda glotal resuene en el tracto nasal, además de en el tracto vocal, y es irradiada primordialmente por las fosas nasales. La nasalización aumenta en gran medida la longitud del sistema resonante e introduce ceros (o anti-resonancias) en el espectro del habla nasalizada (Dodge y Jerse: 1997, p. 225). El primer formante nasal aparece mucho más alto que la barra de sonoridad de otras consonantes, y ésta es una buena indicación de la nasalidad (Bernal et al: 2000, p. 47).

El grupo que completa el alfabeto español es el de las *consonantes líquidas* (*p.e. /l/*), que se producen al pasar el aire por la cavidad bucal con una oclusión central o lateral, de manera que estas consonantes se encuentran acústicamente entre las vocales y las demás consonantes. Debido a la poca resistencia a la salida del aire que existe en las consonantes laterales, acústicamente presentan formantes similares a los sonidos vocálicos. Las vibrantes (*/r/*) se producen por medio de interrupciones a la salida del aire. La vibrante simple presenta una breve oclusión, mientras que en la múltiple se producen varias oclusiones seguidas (Bernal et al: 2000, p. 50). Hacia la liberación (*release*) de una consonante, mientras los articuladores se mueven hacia la configuración apropiada para la vocal que sigue, las frecuencias de las formantes sufren cambios. Tales transiciones proveen información adicional sobre la localidad en el tracto vocal donde la constricción es ubicada.

Sonidos de la voz en oraciones, palabras y sílabas

Los fonemas pueden considerarse como “ladrillos básicos de construcción”. Para que éstos puedan contribuir al significado del lenguaje, deben ser organizados en cúmulos cohesionados, para que luego, tales cúmulos puedan ser combinados en patrones característicos que adquieran significado. Estos patrones son las sílabas, palabras y oraciones del lenguaje (Huang, Acero y Hon: 2001, p. 51). Cuando los sonidos de la voz son pronunciados en un contexto más complejo que el de simplemente sílabas consonante-vocal, vocal-consonante, los movimientos articulatorios implementados para generar una secuencia de sonidos son a menudos modificados con respecto a los movimientos que ocurren para sílabas simples, en particular cuando quien habla lo hace de manera casual. Los

modelos de producción de voz que incorporan y consideran los fenómenos que ocurren a nivel de oración aún se encuentran en etapa de desarrollo (Stevens, *Encyclopedia of Acoustics*: 1997, pp. 1575-1576). Sin embargo, puede mencionarse que los espectrogramas y contornos de frecuencia fundamental a nivel oración, muestran patrones de producción de voz que se extienden en intervalos de tiempo que van más allá que los niveles del fonema y la sílaba. Un ejemplo de ello es que existen diferencias sustanciales en las duraciones de las vocales, que pueden variar desde 20 ms hasta 140 ms. Aunque algunas de éstas diferencias son inherentes a ciertas vocales en particular, la posición de una vocal en una palabra u oración también puede alterar la duración. Para el ingles, por ejemplo, las vocales en palabras con muchas sílabas tienden a ser más cortas que las vocales en palabras de una sola sílaba, y las vocales en palabras que se encuentran cerca del final de ésta, tienden a ser más largas que las vocales en la sección media de una palabra (Stevens, *Encyclopedia of Acoustics*: 1997, p. 1578).

Las *sílabas* en ocasiones son consideradas como una unidad interpuesta entre los fonemas y la palabra. Tal construcción conceptual tiene implicaciones en términos de producción y percepción. Las sílabas a menudo se centran alrededor de vocales. Para separar una por completo se requiere hacer juicios sobre afiliación de consonantes con las vocales en la sílaba. La pregunta sobre si tales juicios deben estar basados en criterios de percepción o en criterios de articulación, y su aplicación rigurosa, continúa sin resolverse en algunos idiomas como el inglés.

Los centros silábicos pueden ser considerados *picos* en sonoridad (secciones periódicas y de gran amplitud en la forma de onda del habla). Tales *picos* sonoros tienen afiliados *hombros (shoulders)* cuya sonoridad no aumenta. Mientras las condiciones de sonoridad sean cumplidas, la afiliación exacta de una consonante dada, que teóricamente pueda ser afiliada a cada lado, podría ser ambigua, a menos de que se determine por un orden superior de estructura de palabra, el cual puede hacer bloques de afiliación (Huang, Acero y Hon: 2001, pp. 51-52).

Las sílabas en el español

Para el idioma español, los ordenes superiores de estructura son dados por diez reglas de afiliación que se presentan a continuación (Ríos: 1999, 6.2):

1) En las sílabas siempre debe haber al menos una vocal. Sin vocal no hay sílaba.

2) Existen conjuntos de consonantes que deben ser mantenidas juntas y pertenecen siempre a la misma sílaba: br, bl, cr, cl, dr, fr, fl, gr, gl, kr, ll, pr, pl, tr, rr, ch, sh.

3) Cuando una consonante se encuentra entre dos vocales, ésta se une a la segunda vocal. Ejemplo une: u-ne

4) Cuando hay dos consonantes entre dos vocales, cada vocal se une a una consonante excepto si son consonantes consideradas inseparables (2). Ejemplo componer: com-po-ner. Aprender: a-pren-der

5) Si son tres las consonantes colocadas entre dos vocales, las dos primeras consonantes se asociarán con la primera vocal y la tercera consonante con la segunda vocal excepto si la segunda y tercera consonantes están dentro del grupo de inseparables. En las combinaciones de cuatro consonantes adyacentes, la frontera silábica se situará entre la segunda y la tercera

consonantes y ambos grupos deben pertenecer a la regla (2). Ejemplos Transporte: trans-por-te. Cumple: cum-ple. Inscripción: ins-crip-ción.

6) Las palabras que contienen una "h" precedida o seguida de otra consonante, se dividen separando ambas letras, excepto en aquellas combinaciones que pertenezcan a la regla (2). Ejemplo anhelo: an-he-lo

7) El diptongo es la unión inseparable de dos vocales. Se pueden presentar tres tipos de diptongos:

    a) Una vocal abierta + una vocal cerrada
    b) Una vocal cerrada + una vocal abierta
    c) Una vocal cerrada + una vocal cerrada

    Son diptongos sólo las siguientes parejas de vocales: ai, au, ei, eu, io, ou, ia, ua, ie, ue, oi, uo, ui, iu, ay, ey, oy. Ejemplo jaula: jau-la

    La unión de dos vocales abiertas o semiabiertas no forma diptongo, es decir, deben separarse en la segmentación silábica. Pueden quedar solas o unidas a una consonante. Ejemplo aéreo: a-é-reo.

8) La "h" entre dos vocales, no destruye un diptongo. Ejemplo ahuyentar: ahu-yen-tar

9) La acentuación sobre la vocal cerrada de un diptongo provoca su destrucción. Ejemplo María: Ma-rí-a

10) La unión de tres vocales puede formar un triptongo. La única disposición posible para la formación de triptongos es la siguiente:

    Vocal cerrada + vocal abierta o semiabierta + vocal cerrada

    En caso de ser una combinación diferente a ésta, el grupo debe ser separado en dos sílabas. Solo las siguiente combinaciones de vocales forman un triptongo: iai, iei, uai, uei, uau, iau, uay, uey.

    De acuerdo con éstas reglas existen solo cuatro tipos de sílabas:

    a) V -> vocal (1 ó 2)
    b) VC -> vocal (1 ó 2) + consonante (1 ó 2)
    c) CV -> consonante (1 ó 2) + vocal (1, 2, ó 3)
    d) CVC -> consonante (1 ó 2) + vocal (1, 2, ó 3) + consonante (1 ó 2)

La sílaba, a veces es considerada como el dominio primario de la co-articulación, esto es, que los sonidos dentro de una misma sílaba influyen unos a otros a su realización más que los mismo sonidos separados por una frontera silábica (Huang, Acero y Hon: 2001, p. 52).

Capítulo III

I. *Procesamiento Digital de Señal*

Señales: "La señal de voz"

A pesar de que la naturaleza física de las distintas señales que se ven relacionadas con las diversas ramas de la ciencia pueden ser totalmente distintas unas de otras, coinciden en que son funciones de una o más variables independientes, contienen información sobre la naturaleza o comportamiento de algún fenómeno (Oppenheim, Willsky: 1994, p.1), y cualquiera que sea la naturaleza de la señal que se está estudiando; la información dentro de ésta se encuentra contenida en un patrón de variaciones.

En el caso particular del sistema de voz humana se tiene que las fuentes de sonido causan excitación acústica de las vías aéreas, y el filtrado de las fuentes por el tracto vocal da lugar a prominencias espectrales en el sonido radiado por la nariz y la boca, consecuencias de los modos naturales del tracto. El resultado de tal proceso es la señal de la voz que se representa de forma matemática por la presión acústica como una función del tiempo; a la cual, a menudo se está interesado en procesarla e incluso en alterarla. El proceso acústico involucrado en la producción del habla puede ser modelado mediante una o más fuentes, cuyo espectro es $S(f)$, que genera una excitación para un sistema acústico con una función de transferencia $T(f)=U_o(f)/S(f)$, donde $U_o(f)$ es el espectro de la velocidad de volumen acústico en la boca o nariz, y con una característica de radiación $R(f)=p_r(f)/U_o(f)$, donde $p_r(f)$ es el espectro de la presión sonora a una distancia r de los labios, resultando (Stevens: 1997, p. 1565):

$$p_r(f) = S(f) \cdot T(f) \cdot R(f)$$

Imagen 16. Modelo del proceso acústico de generación del habla

A tal señal se le conoce como forma de onda (waveform). Diferentes sonidos corresponden a diferentes variaciones de presión acústica, y el sistema vocal humano produce una voz inteligible generando secuencias particulares de estos patrones (Oppenheim, Willsky: 1994, p.7). Como se ha dicho, la información dentro de la señal se encuentra contenida en un patrón de variaciones. El mecanismo vocal humano produce el habla mediante la creación de fluctuaciones en la presión acústica, y el registro de esta variación puede lograrse mediante un micrófono que transforma las variaciones de presión en una *señal eléctrica***.** Los grandes avances en la tecnología del cómputo digital y circuitos integrados, han logrado que se desarrolle el procesamiento de señales de tiempo continuo, representándolas con muestras espaciadas a intervalos iguales en tiempo, es decir, convirtiéndolas en señales de tiempo discreto. La digitalización (o discretización) de la señal de voz, se logra mediante lo que se conoce como una *interfaz o tarjeta de audio*, encargada de *muestrear* la señal de variación de *voltaje* que se recibe desde un micrófono,

que a su vez ha sido obtenida de la variación de presión aérea que representa el sonido, a manera de analogía, convirtiéndose entonces en una secuencia de números. El término *Procesamiento Digital de Señal* se refiere a métodos para la manipulación de secuencias de números $x[n]$ en una computadora digital (Huang, Acero y Hon: 2001, p. 202).

II. Sistemas: El *Procesamiento de la Señal*

Existen dos marcos paralelos para el análisis de las señales, así como el estudio y desarrollo de los sistemas que responden y procesan estas señales; uno para fenómenos y procesos que son descritos en *tiempo continuo,* y otro para los descritos en *tiempo*

discreto. Aunque ambos están muy relacionados teórica, histórica, filosófica y prácticamente, su desarrollo ha sido muy diferente. Las señales y sistemas de tiempo continuo están asociados a problemas de la física, los circuitos eléctricos y las comunicaciones; mientras que los de tiempo discreto encuentran sus raíces en el análisis numérico, las estadística y el análisis de series de tiempo como el análisis de datos económicos y demográficos (Oppenheim, Willsky: 1994, p.4).

Así pues, una señal de tiempo discreto puede representar un fenómeno para el cual la variable independiente es inherentemente discreta o puede representar *muestras* sucesivas de un fenómeno subyacente para el cual la variable independiente es continua. El procesamiento de voz requiere el uso de una secuencia discreta que represente los valores de la señal de voz de tiempo continuo en puntos discretos del tiempo (Oppenheim, Willsky: 1994, p.12). El hecho de que sea posible mostrar en su totalidad una señal de tiempo continuo mediante un conjunto de muestras en particular hace del tema de las señales de voz humana uno necesariamente dualista, ya que naturalmente se presenta en tiempo continuo, y por necesidad, su análisis y procesamiento debe darse en el tiempo discreto.

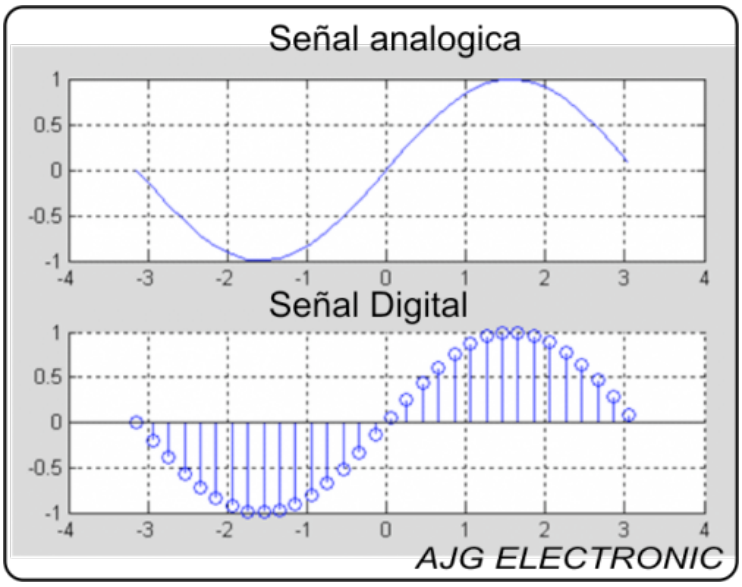


Imagen 17. Señal añalógica vs. Señal digital

En la descripción teórica que sigue, se presentará apenas una introducción a la teoría tanto del procesamiento de señales y sistemas de tiempo continuo, así como del de tiempo discreto, con notación referente al tiempo continuo y al discreto como se muestran a continuación, respectivamente:

$$x(t) \rightarrow y(t) \quad (1)$$
$$x[t] \rightarrow y[t] \quad (2)$$

Recordando que, sin importar el origen de los datos, las señales $x[t]$ están definidas sólo para valores enteros de t.

Propiedades de los Sistemas

Un sistema es cualquier proceso que produce una transformación de una señal, por lo que debe tener una entrada, una salida, y una relación entre ellas a través de tal transformación. Los sistemas tienen una amplia variedad de propiedades básicas, las cuales tienen tanto interpretaciones físicas como matemáticas. Sin embargo, para efectos de esta tesis, sólo se elaborará sobre los conceptos de *causalidad*, *invariancia en el tiempo* y *linealidad* por ser característicos de los sistemas de interés.

Causalidad

Se dice que un sistema es causal cuando su salida en cualquier instante de tiempo depende sólo de los valores de entrada en tiempo presente y en el pasado, dado que el sistema no anticipa valores futuros de la entrada. Una de las consecuencias de tal propiedad es que las salidas de un sistema causal deberán ser idénticas, en caso de que las entradas sean iguales (Oppenheim, Willsky: 1994, pp.42-43).

Invariancia en el tiempo

Cuando un desplazamiento en tiempo de la señal de entrada causa un desplazamiento en tiempo de la salida, entonces el sistema es invariante en el tiempo. Específicamente, si $y[n]$ es la salida de un sistema de tiempo discreto invariante en el tiempo cuando $x[n]$ es la entrada, entonces, $y[n-n_0]$ es la salida cuando se aplica $x[n-n_0]$. Asimismo, para señales de tiempo continuo, con una salida $y(t)$ correspondiente a una entrada $x(t)$, un sistema invariante en el tiempo tendría $y(t-t_0)$ como salida, cuando $x(t-t_0)$ sea la entrada (Oppenheim, Willsky: 1994, pp. 44-45).

Linealidad

Un sistema lineal es aquel en donde puede aplicarse la *propiedad de la superposición*: si una entrada consiste de la suma ponderada de varias señales entonces, la salida es sólo la superposición o suma ponderada de las respuestas del sistema a cada una de estas señales (Oppenheim, Willsky: 1994, pp. 45-46).

La Representación de Convolución:
Representación de señales en términos de impulsos

Un tipo de modelo de sistemas lineales e invariantes en el tiempo es aquel basado en la operación de convolución, que es de hecho un caso especial del *operador* de entrada y salida. Aplica a sistemas en los que puede definirse un *estado cero,* para el que no existe energía inicial en el sistema en el tiempo t_0, y las salidas $y_i(t)$ son todas cero para $t > t_0$ cuando las entradas $x_i(t)$ son todas cero para $t > t_0$. El operador de entrada y salida describe cómo es que el sistema opera sobre señales de entrada para producir señales de salida (Kamen: 1990, p.17).

Los *impulsos unitarios* de tiempo continuo y de tiempo discreto, pueden ser usados, cada uno, como señal básica para construir una amplia variedad de señales. Se puede lograr la descomposición de una señal, $x[n]$, de tiempo discreto en una suma de impulsos ponderados y desplazados, donde para cualquier valor de n, sólo uno de los términos de la expresión es diferente de cero y el escalamiento sobre tal término es precisamente $x[n]$. La expresión compacta de tal construcción es:

$$x[n] = \sum_{k=-\infty}^{k=\infty} x[k]\delta[n-k] \tag{3}$$

Ello corresponde a la representación de una secuencia arbitraria como una *combinación lineal* de impulsos unitarios desplazados $\delta[n-k]$, donde los pesos en esta combinación lineal son $x[k]$.

Para el caso del tiempo continuo, se puede desarrollar una representación análoga. Se puede expresar la aproximación de "escalera" para una señal de tiempo continuo *x(t)*, como una combinación lineal de pulsos retrazados, definiendo:

$$\delta_\Delta(t) = \begin{cases} \frac{1}{\Delta}, 0 < t < \Delta \\ 0, cualquierotro \end{cases} \tag{4}$$

tenemos que, dado que $\Delta\delta_\Delta(t)$ tiene amplitud unitaria, entonces la aproximación a la señal, $\hat{x}$:

$$\hat{x}(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\delta_\Delta(t-k\Delta)\Delta \tag{5}$$

De nueva cuenta, sólo un término en la sumatoria de la ecuación es diferente de cero. A medida que Δ se aproxima a 0, la aproximación dada por (5) es cada vez mejor y en el límite se iguala a $x(t)$. Asimismo, la expresión se aproxima a una integral.

$$x(t) = \lim_{\Delta \to 0} \sum_{k=-\infty}^{+\infty} x(k\Delta)\delta_\Delta(t-k\Delta)\Delta \tag{6}$$

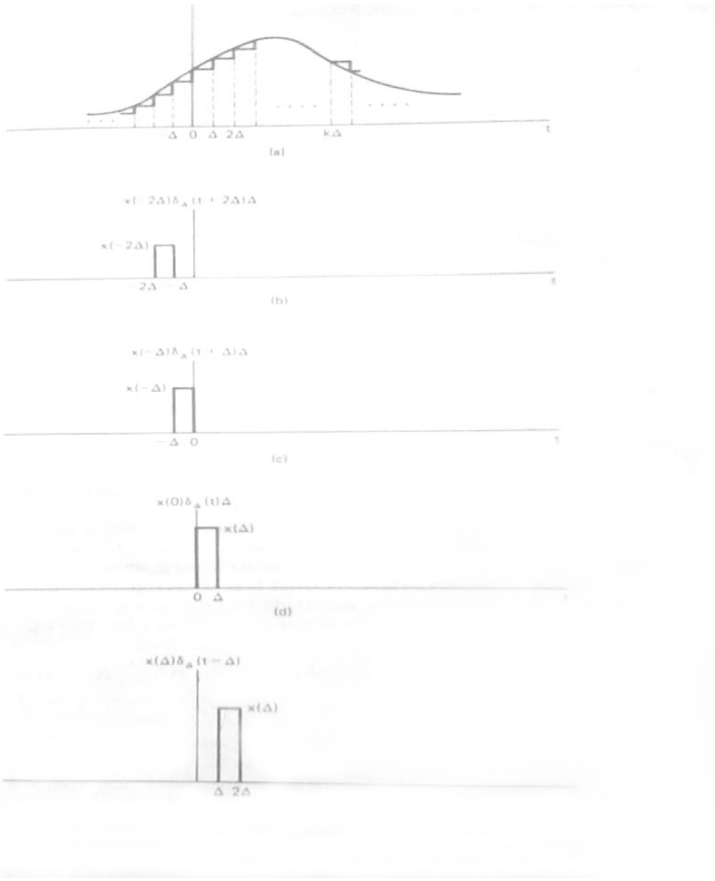


Imagen 18. Aproximación de escalera a una señal de tiempo continuo

En la *Imagen 19*, se representan gráficamente las señales y su producto. La región sombreada tiene un área que se aproxima al área bajo la curva del producto (siendo $k\Delta=\tau$) cuando Δ tiende a cero. El área sombreada es igual a $x(m\Delta)$, donde $t - \Delta < m\Delta < t$. Para éste valor de t, sólo el término con $k = m$ es diferente de cero en la sumatoria de la ecuación (6), y entonces el lado derecho de esta ecuación es también igual a $x(m\Delta)$. Por lo tanto, x(t) es igual al límite, cuando Δ tiende a cero (lo cuál hace que $\delta_\Delta(t)$ sea el impulso unitario), del área bajo $x(t)\,\delta_\Delta(t-\tau)$. Por consiguiente

$$x(t) = \int_{-\infty}^{+\infty} x(\tau)\delta(t-\tau)d\tau \qquad (7)$$

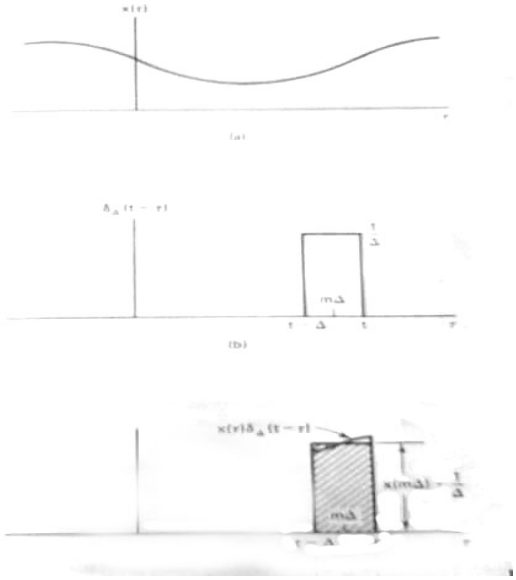


Imagen 19. Interpretación gráfica de la ecuación (6)

Para el caso de la función escalón:

$$u(t) = \int_{-\infty}^{+\infty} u(\tau)\delta(t-\tau)d\tau = \int_{-\infty}^{+\infty} \delta(t-\tau)d\tau \qquad (8)$$

Ahora, bien, considerando un sistema lineal de tiempo discreto y una entrada arbitraria $x[n]$, podemos expresar esta entrada como una combinación lineal de muestras unitarias desplazadas en el tiempo (3). Debido a la propiedad de superposición, la salida $y[n]$ se puede expresar como una combinación lineal de las respuestas, $h_k[n]$ del sistema a muestras unitarias desplazadas d[n-k]:

$$y[n] = \sum_{k=-\infty}^{+\infty} x[k]h_k[n] \qquad (9)$$

Si se conoce la respuesta de un sistema lineal para el conjunto de muestras unitarias desplazadas, entonces se puede reconstruir la respuesta de una entrada arbitraria. La respuesta de un sistema lineal en el instante n, es solo la superposición de las respuestas ocasionadas por cada uno de los valores sucesivos de la entrada. No es necesario que las respuestas $h_k[n]$ estén relacionadas unas con otras para diferentes valores de k, pero si el sistema lineal además es invariante en el tiempo, entonces:

$$h_k[n] = h_0[n-k] \qquad (10)$$

por lo que

$$y[n] = \sum_{k=-\infty}^{+\infty} x[k]h[n-k] \qquad (11)$$

Este resultado se conoce como la *sumatoria de convolución,* o *sumatoria de superposición*, donde a la operación del lado derecho de la ecuación se conoce como

convolución de las secuencias $x[n]$ y $h[n]$, y se representa de manera simbólica como $y[n]=x[n]*h[n]$; y expresa la respuesta de un sistema lineal invariante en el tiempo, a una entrada arbitraria en términos de su respuesta al impulso unitario. Un sistema de esta naturaleza, está completamente caracterizado por tal respuesta. La respuesta debida a la entrada $x[k]$, aplicada en tiempo k, es $x[k]h[n-k]$, la cual es una versión desplazada y escalada de $h[n]$. La salida real será entonces la superposición de todas estas respuestas, lo cual consiste en la suma de todos los valores de k de los números $x[k]h[n-k]$ (Oppenheim, Willsky: 1994, pp. 76-83).

Sistemas descritos por ecuaciones diferenciales y ecuaciones en diferencias.

Una amplia variedad de sistemas y fenómenos físicos, son descritos a través de ecuaciones diferenciales lineales. Su contraparte en los sistemas de tiempo discreto son las ecuaciones en diferencias lineales. La ecuaciones de este tipo se utilizan para describir el comportamiento secuencial de diversos procesos, como podría ser una señal muestreada de voz como la respuesta del tracto vocal humano a la excitación de las cuerdas vocales. Tales ecuaciones aparecen a menudo en la especificación de sistemas de tiempo discreto diseñadas para efectuar operaciones específicas deseadas sobre la señal de entrada, como los filtros.

Una ecuación en diferencias lineal de orden N con coeficientes constantes;

$$\sum_{k=0}^{N} a_k y[n-k] = \sum_{k=0}^{M} b_k x[n-k] \tag{12}$$

se puede resolver encontrando la solución $y[n]$, que puede escribirse como la suma de una solución particular de la ecuación y de una solución de la ecuación homogénea:

$$\sum_{k=0}^{N} a_k y[n-k] = 0 \tag{13}$$

La ecuación (12) no es especifica totalmente la salida en términos de la entrada, ya que para hacer esto se deben definir algunas condiciones auxiliares. El sistema descrito por la ecuación es lineal si las condiciones auxiliares son cero. Únicamente en el caso de tiempo discreto y de ecuaciones en diferencia, la ecuación puede ser re-ordenada de la siguiente manera:

$$y[n] = \frac{1}{a_0}\left\{\sum_{k=0}^{M} a_k x[n-k] - \sum_{k=0}^{N} a_k y[n-k]\right\} \tag{14}$$

En ésta forma, expresa la salida en el tiempo n en términos de los valores previos de la entrada y de la salida. De ésta expresión se observa la necesidad de las condiciones auxiliares, ya que para poder calcular $y[n]$, es necesario conocer $y[n-1],\ldots, y[n-N]$. Si se especifica la entrada para toda n, y un conjunto de condiciones auxiliares tales como $y[-N]$, $y[-N+1],\ldots, y[-1]$, la ecuación (14) puede resolverse para valores sucesivos de $y[n]$. Una ecuación de esta forma es llamada una ecuación *recursiva* ya que especifica un procedimiento recursivo para determinar la salida en términos de la entrada y las salidas

previas. Cuando N =0, la ecuación se reduce a:

$$y[n] = \sum_{k=0}^{M} \left(\frac{b_k}{a_0} \right) x[n-k] \tag{15}$$

En este caso, $y[n]$ es una función explícita de los valores presentes y previos de la entrada y es llamada ecuación *no recursiva*, ya que no se usan los valores de la salida ya calculados para determinar el valor presente de la salida, por lo que no necesita valores auxiliares para determinar $y[n]$. Además, la ecuación describe un sistema lineal invariante en el tiempo y por cálculo directo se puede ver que la respuesta al impulso del sistema es:

$$h[n] \begin{cases} \frac{b_k}{a_0}, 0 \le n \le M \\ 0, para cualquier otro valor \end{cases} \tag{16}$$

La respuesta al impulso, en éste caso, tiene duración finita, o lo que es lo mismo, es diferente de cero sólo dentro de un intervalo finito de tiempo. A los sistemas que tienen tal propiedad se les llama *respuesta al impulso finita (FIR)*.

Para el caso recursivo son necesarios los valores auxiliares ($N \ge 1$). Un sistema descrito por una ecuación en diferencias recursiva es llamado con frecuencia *sistema de respuesta al impulso infinita (IIR).*

Análisis de Fourier

La suma de convolución tiene como punto de partida la representación de una señal de entrada de tiempo discreto en un sistema lineal invariante en el tiempo, como una suma ponderada de impulsos unitarios desplazados. Puede realizarse una representación análoga para una señal de tiempo continuo como una integral ponderada de impulsos desplazados, para obtener la *Integral de convolución.* A partir de tales desarrollos, se pueden analizar con detalle, muchas de las propiedades de los sistemas lineales invariantes en el tiempo, y relacionarlas con las características equivalentes de las respuestas al impulso de tales sistemas. Ahora bien, existe una representación alternativa para estas señales y sistemas, en donde el punto de partida del desarrollo, de igual manera que en la sección anterior, es el de una representación de señales como sumas e integrales ponderadas de un conjunto de señales básicas. La diferencia radica en que se usan las funciones *exponenciales complejas* como señales básicas, resultando la serie y la transformada de *Fourier* de tiempo continuo y discreto.

De tales representaciones, y gracias a la propiedad de superposición de los sistemas lineales, las respuestas de un sistema lineal invariante en el tiempo a cualquier entrada formada por una combinación lineal de señales básicas, es la *combinación lineal* de las respuestas individuales a cada una de estas señales básicas. La respuesta de un sistema lineal invariante en el tiempo a una exponencial compleja, tiene también una forma particularmente simple, lo que proporciona otra representación conveniente para éstos sistemas, y por lo tanto una forma alternativa de análisis. La herramientas del análisis de Fourier de tiempo continuo, junto con sus contrapartes de tiempo discreto,

constituyen la base de la discusión sobre filtrado, modulación y muestreo, que a su vez, representan zonas de conocimiento necesario para la síntesis de la voz y el desarrollo de esta tesis.

El desarrollo de las técnicas de análisis de Fourier en tiempo continuo y discreto, es en gran medida paralelo; y las conclusiones en extremo similares, aunque la historia de los desarrollos ha sido, como se mencionó, paralela y ajena ya que el origen del análisis de Fourier en tiempo continuo es atribuido a las investigaciones sobre física matemática en el s. XVIII, mientras que las de tiempo discreto tienen raíces distintas, siendo fundamentales para el análisis numérico y eran investigadas desde tiempos de Newton en el s. XVII. El incremento en el uso y capacidad de las computadoras digitales en el siglo pasado, es decir, de sistemas de tiempo discreto para el procesamiento de señales muestreadas de tiempo continuo, hicieron que creciera el traslape de los dominios de aplicación de las técnicas de tiempo continuo y de tiempo discreto, y se proporcionó una conexión natural entre las dos metodologías, desarrolladas hasta entonces por separado (Oppenheim, Willsky: 1994, pp. 174-178). A finales de los años 60 se desarrolló un algoritmo conocido como la *Transformada Rápida de Fourier* (FFT), que demostró ser adecuado para una implementación digital eficiente, lo cual hizo prácticas muchas ideas y herramientas hasta ese entonces imprácticas. Ya que existen varias similitudes entre las técnicas del análisis de *Fourier* de tiempo discreto y continuo, se presenta parte del desarrollo de ambas, haciendo énfasis en las diferencias importantes, como puede ser el hecho de que la representación en serie de *Fourier* de una señal periódica de tiempo discreto es un serie *finita*, en oposición con una serie infinita requerida por la señal periódica de tiempo continuo, propiedad que se manifiesta intrínsecamente en la implementación de la *Transformada Rápida de Fourier*.

Análisis de Fourier para señales de tiempo continuo
Respuesta de Sistemas de tiempo continuo y discreto, lineales invariantes en el tiempo, a exponenciales complejas.

La importancia de la exponenciales complejas en el estudio de éste tipo de sistema, proviene del hecho de que la respuesta de estos sistemas a una entrada de aquel tipo, es la misma exponencial compleja modificada sólo en amplitud, es decir:

$$e^{st} \rightarrow H(s)e^{st} \tag{17}$$

donde el factor complejo de amplitud $H(s)$ será en general una función de la variable compleja s. Una *función característica* del sistema, es una señal para la cuál la salida es igual a la entrada multiplicada por una constante, y el valor característico es el factor de amplitud. Empleando la *Integral de convolución* se puede demostrar que cualquier exponencial compleja es una función característica de un sistema lineal invariante en el tiempo, y su respuesta es de la forma:

$$y(t) = H(s)e^{st} \tag{18}$$

De ello se desprende que si se conocen los valores característicos $h(s_k)$, la respuesta a una combinación lineal de exponenciales complejas se puede construir de manera directa.

La motivación del desarrollo de una representación en el caso del tiempo discreto es idéntica al caso del tiempo continuo. Las secuencias de exponenciales complejas son funciones características de sistemas de tiempo discreto, lineales invariantes en el tiempo. Ahora bien, si un sistema de esta índole, que tiene una respuesta al impulso $h[n]$, tiene como entrada la señal:

$$x[n] = z^n \tag{19}$$

donde z es un número complejo, entonces las salida del sistema puede determinarse a partir de la suma de convolución como

$$y[n] = h[n] * x[n] = \sum_{k=-\infty}^{+\infty} h[k]x[n-k] = \sum_{k=-\infty}^{+\infty} h[k]z^{n-k} = z^n \sum_{k=-\infty}^{+\infty} h[k]z^{-k} \tag{20}$$

De aquí se puede ver que si la entrada $x[n]$ es la exponencial compleja dada por la ecuación (19), entonces la salida es la misma exponencial compleja multiplicada por una constante que depende del valor de z, esto es:

$$y[n] = H(z)z^n \tag{21}$$

donde,

$$H(z) = \sum_{k=-\infty}^{+\infty} h[k]z^{-k} \tag{22}$$

Aquí $H(z)$ es el valor característico asociado con la función característica z^n. La ecuación (21) junto con la propiedad de superposición, implica que la representación de señales en términos de exponenciales complejas conduce a una expresión conveniente para la respuesta de estos sistemas. Si la entrada a un sistema es:

$$x[n] = \sum_k a_k z_k^n \tag{23}$$

entonces la salida será

$$y[n] = \sum_k a_k H(z_k) z_k^n \tag{24}$$

lo que significa que la salida puede representarse como una combinación lineal de las mismas señales exponenciales complejas, y cada coeficiente en la representación de la salida se obtiene como el producto del correspondiente coeficiente a_k de la entrada y el valor característico $H(z_k)$ del sistema asociado con la función característica $z^n{}_k$.

Representación de Fourier de señales periódicas

Una señal $x(t)$ es periódica con periodo T, si:

$$x(t+T) = x(t) \tag{25}$$

para toda t, $-\infty < t < \infty$, donde T es el número positivo menor para el cuál ésta relación se satisface. El valor $2\pi/T$ se conoce como la frecuencia fundamental. Se debe recordar que las funciones de forma sinusoidal, $x(t) = cos\ \omega_0 t$, así como las exponenciales complejas, $x(t) = e^{j\omega t}$ son periódicas con frecuencia fundamental ω_0 y periodo fundamental $2\pi/\omega_0$. Asociado a ésta última función, están las exponenciales complejas relacionadas armónicamente:

$$\phi_k(t) = e^{jk\omega_0 t}, k = 0, \pm 1, \pm 2, \ldots \tag{26}$$

y una combinación lineal de exponenciales complejas relacionadas armónicamente tiene la forma:

$$x(t) = \sum_{k=-\infty}^{+\infty} a_k e^{jk\omega_0 t} \tag{27}$$

teniendo también periodo T. En ésta expresión, los componentes para $k=+N$ y $k=-N$, representan los componentes de la N-esima armónica, y a tal representación se le conoce como la serie de Fourier. Una forma alterna de representación es mediante señales periódicas reales, como las sinusoidales, esto cuando la señal es real.

Si suponemos que una señal periódica se puede representar con la serie de la ecuación (27), debemos conocer la manera de determinar los coeficientes a_k, para ello multiplicamos de ambos lados de la ecuación por $e^{-jn\omega_0 t}$ y luego integramos de ambos lados de 0 a T, es decir, sobre un periodo. Para determinar coeficientes tenemos que:

$$a_n = \frac{1}{T_0}\int_0^{T_0} x(t)e^{-jn\omega_0 t}dt = \frac{1}{T_0}\int_{T_0} x(t)e^{-jn\omega_0 t}dt \tag{28}$$

El par de ecuaciones (27) y (28) definen la serie de Fourier para una señal periódica, la primera conociéndose como la ecuación de síntesis y representando la expresión de la señal como combinación lineal de exponenciales complejas, y la segunda como la ecuación de análisis ya que permite el cálculo de los coeficientes $\{a_k\}$ de tal combinación lineal, conocidos como coeficientes de la *serie de Fourier,* o coeficientes espectrales, que miden la porción de la señal $x(t)$ que está en cada armónica de la componente fundamental.

Para el caso en el tiempo discreto, tenemos que una señal es periódica si para algún valor positivo de N:

$$x[n] = x[n+N] \tag{29}$$

Sabemos que el conjunto de todas las señales exponenciales complejas discretas, periódicas en N, están dadas por:

$$\phi_k[n] = e^{jk(2\pi/N)n}, k = 0, \pm 1, \pm 2, \ldots \tag{30}$$

y están relacionadas armónicamente ya que sus frecuencias son múltiplos de la fundamental. Ahora bien, mientras en el caso del tiempo continuo, todas las señales en la ecuación (26) son distintas, en el caso discreto solo existen *N* señales diferentes, ya que las exponenciales complejas que difieren por un múltiplo de 2π son idénticas, y como consecuencia cuando *k* se cambia por cualquier múltiplo entero de *N*, se genera la misma secuencia:

$$\phi_k[n] = \phi_{k+rN}[n] \tag{31}$$

La representación de secuencias periódicas como combinación lineal de las secuencias $\Phi_k[n]$ tiene la forma:

$$x[n] = \sum_k a_k \phi_k[n] = \sum_k a_k e^{jk(2\pi/N)n} \tag{32}$$

La suma únicamente incluye sumas sobre el rango de *N*, por razones previamente discutidas.

$$x[n] = \sum_{k=\langle N \rangle} a_k \phi_k[n] = \sum_k a_k e^{jk(2\pi/N)n} \tag{33}$$

A esta ecuación se le conoce como la *serie de Fourier* en tiempo discreto, y los coeficientes a_k son los coeficientes de la *serie de Fourier*, misma que es finita como consecuencia de la ecuación (31).

Ahora bien, para la representación de señales periódicas mediante la *serie de Fourier,* si se tiene una secuencia *x*[*n*] periódica con periodo *N*, se buscan determinar los valores a_k, para determinar una solución para el conjunto de *N* ecuaciones lineales que se obtienen de evaluar la ecuación (33) para valores sucesivos de *n*.

$$\begin{aligned} x[n] &= \sum_{k=\langle N \rangle} a_k \\ x[1] &= \sum_{k=\langle N \rangle} a_k e^{jk(2\pi/N)} \\ &\vdots \\ x[N-1] &= \sum_{k=\langle N \rangle} a_k e^{jk(N-1)2\pi/N} \end{aligned} \tag{34}$$

Al ser tal conjunto de ecuaciones linealmente independiente, pueden resolverse para obtener los coeficientes a_k en términos de los valores dados de *x*[*n*]. Se puede obtener una expresión para tales coeficientes en términos de la secuencia de valores de *x*[*n*], considerando que:

$$\sum_k e^{jk(2\pi/N)n} = \begin{cases} N, k = 0, \pm 1, \pm 2, \ldots \\ 0, otrovalor \end{cases} \tag{35}$$

si la sumatoria se realiza en cualquier intervalo de extensión *N, y*a que la suma sobre un periodo de los valores de una exponencial compleja periódica es cero, a menos de que la

misma sea constante. Al multiplicar la ecuación (33) de ambos lados por $e^{jk(2\pi/N)n}$, sumar los *N* términos e intercambiar los ordenes de la sumatoria:

$$\sum_{n=<N>}^{N-1} x[n]e^{-jr(2\pi/N)n} = \sum_{n=<N>} \sum_{k=<N>} a_k e^{j(k-r)(2\pi/N)n} \tag{36}$$

Por el análisis previo, la suma interior sobre *n* es cero, a menos de que *k*-*r* sea cero o múltiplo entero de *N*. Por lo que si se escogen valores de *r* sobre el mismo rango sobre el cual *k* varía en la sumatoria exterior, la suma interior es igual a *N,* si *k = -r, ó* 0 si *k* es diferente de *r*, por lo que la sumatoria se reduce a Na_k*:*

$$a_r = \frac{1}{N} \sum_{n=<N>} x[n]e^{-jr(2\pi/N)n} \tag{37}$$

Expresión que proporciona una expresión para obtener los *coeficientes de Fourier* de tiempo discreto. Esta ecuación y la ecuación (33), juegan el mismo papel para las señales de tiempo discreto, que (27) Y(28) para tiempo continuo. Los coeficientes espectrales de *x*[*n*], especifican su descomposición en una suma de *N* exponenciales complejas relacionadas armónicamente, siendo además la *serie de Fourier* de tiempo discreto una serie finita que posee muchas propiedades útiles. En particular, la *serie de Fourier* no tiene problemas con la convergencia debido a que cualquier serie periódica de tiempo discreto está totalmente especificada por un número finito *N* de parámetros o de valores de la secuencia sobre un periodo. La ecuación de análisis únicamente transforma tal conjunto de *N* parámetros por los valores de los *N coeficientes de Fourier*, mientras que la ecuación de síntesis muestra como recuperar los valores de los valores de la secuencia original en términos de una serie finita. En contraste, una señal periódica de tiempo continuo, toma una continuidad de valores sobre un periodo, lo que significa que se requiere un número infinito de *coeficientes de Fourier* para representarla, por lo que en general ninguna de las sumas parciales finitas producirá los valores exactos de la señal original. Surge además un problema de convergencia cuando se busca evaluar el límite cuando el número de términos se aproxima al infinito.

Una de las similitudes del tratamiento de la *serie de Fourier* de tiempo discreto y continuo, es el hecho de que las exponenciales complejas son las funciones características de los sistemas lineales invariantes en el tiempo en ambos casos. Por esto, si la entrada *x*[*n*], a un sistema de tiempo discreto de estas características es periódica con periodo *N*, y si *h*[*n*] es la respuesta al impulso del sistema, se tiene una expresión conveniente para los *coeficientes de Fourier* de la salida en términos de los de la entrada:

$$y[n] = \sum_{n=<N>} a_k H\left(\frac{2\pi k}{N}\right) e^{jr(2\pi/N)n} \tag{38}$$

donde de (22):

$$H\left(\frac{2\pi k}{N}\right) = \sum_{n=-\infty}^{\infty} h[n]e^{-jk(2\pi/N)n} \tag{39}$$

Representación de Señales no periódicas para el tiempo continuo y el tiempo discreto.

Los resultados anteriores pueden extenderse para desarrollar una representación de señales no periódicas como una combinación lineal de exponenciales complejas, lo cuál representa una de las contribuciones más importantes de Fourier. La manera de desarrollar sus ideas, fue la de pensar en una señal no periódica como el límite de una señal periódica cuando el periodo se hace arbitrariamente grande, y examinar el comportamiento en el límite de la representación de la serie de Fourier para ésta señal. Para ello se debe considerar una señal no periódica general $x(t)$ de duración finita a partir de la misma se puede construir una señal periódica para la cual $x(t)$ es un periodo. Como se escoge un periodo T_o grande, la nueva señal periódica es idéntica a la no periódica original sobre un intervalo largo, y conforme T_o se aproxima a infinito, la señal periódica será idéntica a la original para cualquier valor finito de t. Las ecuaciones que se presentan a continuación se conocen como el par de transformadas de Fourier con la función: $X(\omega)$ conocida como la transformada o integral de Fourier de $x(t)$, mientras que la otra es la ecuación de la transformada inversa de Fourier.

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\omega)e^{j\omega t}d\omega$$
$$X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}d\omega \qquad (40)$$

Para señales no periódicas las exponenciales complejas ocurren en una sucesión continua de frecuencias y tienen una amplitud $X(\omega)(d\omega/2\pi)$.

En el caso de tiempo discreto, para desarrollar la representación de la transformada de Fourier para secuencias no periódicas, se debe considerar una secuencia general $x[n]$ no periódica de duración finita, y se construye, como en el caso continuo, una secuencia periódica para la cuál $x[n]$ es un periodo. Cuando se hace que el periodo sea más extenso, ambas son idénticas sobre un intervalo más grande, y conforme N se aproxima a infinito entonces son idénticas para cualquier valor finito n. Tal señal periódica tendrá la representación en serie de Fourier:

$$\tilde{x}[t] = \sum_{k=<N>} a_k e^{jk(2\pi/N)n} \qquad (41)$$

$$a_k = \frac{1}{N}\sum_{k=<N>} \tilde{x}[t]e^{-jk(2\pi/N)n} \qquad (42)$$

Se establece el periodo como intervalo de la sumatoria en la ecuación (42), de manera que pueda remplazarse la señal periódica por la original no periódica, ya que sobre tal intervalo son idénticas, y fuera de éste la función es cero:

$$a_k = \frac{1}{N}\sum_{n=-N_1}^{N_1} x[n]e^{-jk(2\pi/N)n} = \frac{1}{N}\sum_{n=-\infty}^{+\infty} x[n]e^{-jk(2\pi/N)n} \qquad (43)$$

Se define $X(\Omega)$ como:

$$X(\Omega) = \sum_{k=-\infty}^{\infty} x[n]e^{-j\Omega n} \tag{44}$$

y se obtiene que los coeficientes están dados por

$$a_k = \frac{1}{N} X(k\Omega_0) \tag{45}$$

donde Ω_0 se usa para denotar el espaciamiento de las muestras $2\pi/N$. Los coeficientes son entonces proporcionales a muestras equiespaciadas de tal función $X(\Omega)$.

Al igual que en el caso continuo, conforme N tiende a infinito, la señal periódica y la señal no periódica se igualan para cualquier valor finito de n, y Ω_0 tiende a cero, convirtiendo a la ecuación en una integral cuyo intervalo total de integración siempre tendrá un ancho de 2π, ya que la sumatoria en la ecuación se lleva a cabo sobre N intervalos consecutivos de ancho $\Omega_0 = 2\pi/N$. Por lo que:

$$x[n] = \frac{1}{2\pi} \int_{2\pi} X(\Omega)e^{j\Omega n} d\Omega \tag{46}$$

que junto con

$$X(\Omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\Omega n} \tag{47}$$

son la transformada de Fourier de tiempo discreto (47) y su inversa (46). Aunque el argumento que se dio fue elaborado suponiendo que $x[n]$ era de duración arbitraria pero finita, estas ecuaciones siguen siendo válidas para una amplia gama de señales de duración infinita. Las principales diferencias con su contraparte en el tiempo continuo, son la periodicidad de la transformada de tiempo discreto y el intervalo finito de integración de la ecuación de síntesis; tales diferencias se dan debido a que las exponenciales complejas de tiempo discreto que difieren en frecuencia por un múltiplo de 2π son idénticas. Tal afirmación tiene como consecuencia, para las señales periódicas, que los coeficientes sean periódicos y que la representación en serie de Fourier sea una suma finita, mientras que para las señales periódicas las implicaciones análogas son la periodicidad de $X(\Omega)$, y que la ecuación de síntesis involucra una integración de 2π.

La transformada de Fourier de tiempo discreto

Para indagar cuál es la salida de un sistema lineal invariante en el tiempo con respuesta al impulso $h[n]$, cuando la entrada es una exponencial compleja, se sustituye en la ecuación (11), $x[n] = e^{j\omega_0(n)}$ y usando la propiedad conmutativa de la convolución se obtiene:

$$y[n] = \sum_{k=-\infty}^{+\infty} h[k] e^{j\omega_0(n-k)} = e^{j\omega_0 n} \sum_{k=-\infty}^{+\infty} h[k] e^{-j\omega_0 k} = e^{j\omega_0 n} H(e^{j\omega_0}) \tag{48}$$

que es otra exponencial compleja de la misma frecuencia y amplitud multiplicada por la cantidad compleja $H(e^{j\omega})$ dada por:

$$H(e^{j\omega}) = \sum_{k=-\infty}^{+\infty} h[k] e^{-j\omega_0 k} \tag{49}$$

Debido a que la salida de un sistema lineal invariante en el tiempo a una exponencial compleja es otra exponencial compleja, se dice que las exponenciales complejas son *vectores propios* de sistemas lineales invariantes en el tiempo, con la cantidad compleja $H(e^{j\omega_0})$ como su valor propio.

La cantidad $H(e^{j\omega_0})$, es definida como la transformada discreta de Fourier de $h[n]$, y es claro que es una función periódica de ω con periodo 2π y por lo tanto solo es necesario preservar un periodo para describirla por completo; típicamente $-\pi<\omega<\pi$.

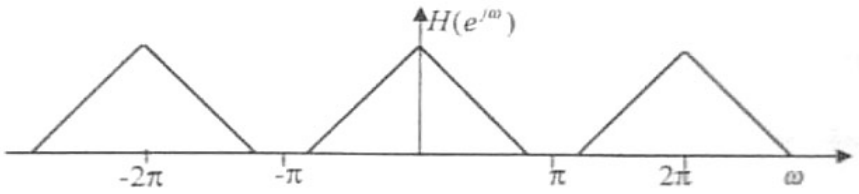


Imagen 20. $H(e^{j\omega_0})$ ***es una función periódica de ω.***

Si el valor $|H(e^{j\omega_0})|>1$, el sistema lineal invariante en el tiempo amplificará tal frecuencia, y de la misma manera la atenuará, o *filtrará*, si $|H(e^{j\omega_0})| < 1$, razón por la cuál estos sistemas también son llamados filtros. La transformada de Fourier $H(e^{j\omega})$ de un filtro $h[n]$ es llamada la *respuesta frecuencial* o *función de transferencia* del sistema. (Huang, Acero y Hon: 2001, p. 210).

La transformada discreta de Fourier (DFT) y la transformada rápida de Fourier (FFT)

Si una señal $x_N[n]$ es periódica con periodo N, entonces $x_N[n]= x_N[n+N]$ y la señal es representada por N muestras consecutivas. La *Trasformada Discreta de Fourier* (DFT) de una señal periódica $x_N[n]$ es definida como:

$$X_N[k] = \sum_{n=0}^{N-1} x_N[n] e^{-j2\pi nk/N}, 0 \le k < N \tag{50}$$

$$x_N[n] = \frac{1}{N} \sum_{n=0}^{N-1} X_N[k] e^{j2\pi nk/N}, 0 \le n < N \tag{51}$$

que son pares transformados. La ecuación (51) también se conoce como una expansión en *series de Fourier*. Bajo tal transformación, al retener un menor número de sinusoides (o coeficientes de *Fourier*) se puede mantener una aproximación apropiada para una función periódica.

La trasformada rápida de Fourier

Existe una familia de algoritmos rápidos para computar la transformada discreta de Fourier (DTF). Mientras el cómputo directo de la DFT requiere N^2 operaciones, asumiendo que las funciones trigonométricas han sido pre-calculadas, el algoritmo de la transformada rápida de Fourier requiere del orden de solo $Nlog_2N$ operaciones, por lo que es ampliamente usada para el procesamiento de voz (Huang, Acero y Hon: 2001, p. 222).

La Transformada Z

Para un sistema lineal de tiempo discreto e invariante en el tiempo, con respuesta al impulso $h[n]$, la respuesta $y[n]$ del sistema a una entrada exponencial compleja de la forma z^n es:

$$y[n]=H(z)z^n \tag{52}$$

donde

$$H(z)=\sum_{k=-\infty}^{+\infty} h[k]z^{-k} \tag{53}$$

Para $z=e^{i\Omega}$ con Ω real (es decir, con |z|=1), la sumatoria en la ecuación (53) corresponde a la *transformada de Fourier* de tiempo discreto de $h[n]$. Es decir, cuando |z| no está restringido a la unidad, la sumatoria en la ecuación (53) se conoce como la *transformada z* de $h[n]$. Al igual que con la *transformada de Fourier*, tanto para señales de tiempo continuo como de tiempo discreto, la transformada *z* juega un papel importante como una transformación aplicada a secuencias en general, ya sea que representen o no la respuesta al impulso de un sistema.

La *transformada z* de una secuencia $x[n]$ se define como:

$$X(z)\triangleq\sum_{n=-\infty}^{+\infty} x[n]z^{-n} \tag{54}$$

donde *z* es una variable compleja. Por conveniencia, la *transformada z* de $x[n]$ algunas veces se denota como $Z\{x[n]\}$, y la relación entre $x[n]$ y su *transformada z* se indica como:

$$x[n]\overset{Z}{\longleftrightarrow}X(z) \tag{55}$$

Existen un número importante de relaciones entre la *transformada z* y la *transformada de Fourier*. Para explorar tales, se debe expresar la variable compleja *z* en forma polar como sigue:

$$z = re^{j\Omega} \tag{56}$$

siendo *r* la magnitud de *z,* y Ω como el ángulo de *z*. En términos de *r* y Ω, la ecuación (54) pasa a ser

$$X(re^{j\Omega}) = \sum_{n=-\infty}^{+\infty}\left\{x[n]r^{-n}\right\}e^{-j\Omega n} \tag{57}$$

A partir de la ecuación (57) se puede ver que $X(re^{i\Omega})$ es la transformada de Fourier de la secuencia $x[n]$ multiplicada por una exponencial real r^{-n}, esto es,

$$X(re^{i\Omega}) = \Im\{x[n]r_n\} \tag{58}$$

La exponencial r^{-n} puede ser creciente o decreciente al incrementarse *n*, dependiendo de si *r* es mayor o menor que uno. Se debe notar en particular que para r=1 o |z|=1, la *transformada z* se reduce a la *transformada de Fourier*, esto es:

$$X(z)\big|_{z} = \Im\{x[n]\} \tag{59}$$

Ello quiere decir que la *transformada z* se reduce a la transformada de Fourier sobre un contorno del plano *z* complejo correspondiente a un círculo con radio unitario. Para que la *transformada z* converja, es necesario que la *transformada de Fourier* de $x[n]r^{-n}$ converja.

Filtros Digitales y Ventaneo

Uno de los objetivos primordiales del análisis del habla por computadora ha sido el de reducir la cantidad de información requerida para representar una señal de voz. Durante el análisis, el habla es seccionado en segmentos llamados “ventanas”. Para cada uno de estos segmentos el algoritmo de análisis determina los atributos del habla. Tales características representan el sonido del habla y pueden ser usados más adelante para la recreación del sonido analizado. Una de las principales funciones del análisis del habla es la de determinar las características de resonancia de tracto vocal en lo que dura tal “ventana”.

Las técnicas más usadas en el diseño de filtros IIR (filtros de respuesta al impulso infinita), o funciones de valores pasados de entrada y salida, están basadas en las transformaciones de sistemas IIR de tiempo continuo en sistemas IIR de tiempo discreto. En contraste, los filtros FIR (filtros de respuesta al impulso finita), o filtros cuya respuesta al impulso tiene un número limitado de coeficientes diferentes de cero, están casi completamente restringidos a implementaciones en tiempo discreto. El método más sencillo para diseñar filtros FIR es justamente el método de ventana, que generalmente comienza con una respuesta en frecuencia ideal deseada que puede ser representada por:

$$H_d(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h_d[n]e^{-j\omega n} \quad (60)$$

donde $h_d[n]$ e la secuencia de la respuesta al impulso correspondiente, que puede ser expresada en términos de $H_d(e^{j\omega})$ como:

$$h_d[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_d(e^{j\omega})e^{j\omega} \quad (61)$$

La manera más simple de obtener un filtro FIR causal a partir de $h_d[n]$ es definiendo un sistema con respuesta al impulso $h[n]$ dada por:

$$h_n = \begin{cases} h_d[n], 0 \le n \le M \\ 0, c\,c. \end{cases} \quad (62)$$

y de manera más general podemos representar $h[n]$ como el producto de la respuesta al impulso deseada y una ventana de duración finita $w[n]$:

$$h[n] = h_d[n]w[n] \quad (63)$$

La elección de la ventana obedece al deseo de tener una de tan corta duración como sea posible con el fin de minimizar los cálculos en la implementación del filtro siendo la transformada de Fourier de la ventana aproximadamente un impulso, esto es, que esté altamente concentrada en frecuencia para que se reproduzca la respuesta en frecuencia deseada. Mientras las funciones de ventana *triangulares, Kaiser, y Barlett* entre otras ocasionalmente aparecen en sistemas de procesamiento digital de voz, son las ventanas *rectangular, Hanning y Hamming* las que más son utilizadas. Éste tipo de ventanas poseen la propiedad deseada de que sus transformadas de Fourier están concentradas alrededor de ω=0, y además tienen una forma simple que permite calcularlas fácilmente (Herrera: 2006, p. 29). Para efectos de esta tesis, únicamente se revisará la ventana de *Hanning* debido a que es la única utilizada en el desarrollo.

Hanning $$w[n] = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{M-1}\right), 0 \le n \le M \\ 0, c\,c. \end{cases} \quad (64)$$

Análisis espectral (del habla)

Existen dos tipos de representaciones espectrales en el ámbito del análisis del habla: *amplitud vs. frecuencia* e *intensidad-de-frecuencia vs. tiempo*. La primera, es un tipo de análisis que se realiza a cuadros de corta duración (3-30 ms), y por lo tanto es asociado con una una sección espectral de localidad muy específica en el tiempo. Un espectro promedio de tiempo largo es el resultado de extender la duración de la porción del habla; este análisis tiene la ventaja de ofrecer características específicas de quien está hablando.

Para visualizar un sonido del habla con lujo de detalle, uno debe realizar un análisis de tiempo corto renovando los cálculos espectrales a intervalos cortos de tiempo (1 ms), tal es la base del espectrograma de intensidad-de-frecuencia contra tiempo (Fant, *Encyclopedia of Acoustics:* 1997, p. 1589).

En un espectrograma, la intensidad espectral es representada por la densidad o el color de las marcas en un espectrograma dentro de las áreas de mínima resolución tiempo-frecuencia, las cuales están conceptualmente definidas por una extensión T en tiempo y B en frecuencia, siendo el producto BT una constante del orden de 1. Ello implica que, necesariamente, si en el análisis se busca una alta resolución en frecuencia, esta se perderá en tiempo y viceversa. Este detalle es especialmente importante para cuando se requieren observar armónicos individuales, caso en el que es necesario que B sea más pequeño que la frecuencia fundamental (la cual varía de 60-200 Hz en hombres y menos de una octava más para mujeres). Por el contrario, al ser B mayor que la fundamental, y tener una incrementada resolución temporal, puede observarse en el análisis un dominio del patrón espectral por los formantes, esto hace posible el seguimiento de variaciones rápidas en el espectro. En conclusión, el análisis angosto, $B<Fo$, es aplicable a la representación armónica, mientras que el análisis de banda ancha $B>Fo$, extrae una visión en el dominio temporal de sonidos sonoros, como la suma y secuencia de frecuencias amortiguadas que representan la respuesta de los modos resonantes del tracto vocal a cada excitación impulsiva sucesiva de la fuente glotal. Para seccionar las vocales, el primer acercamiento es utilizado, mientras que el segundo permite generar espectrogramas de tiempo-frecuencia-intensidad, y son útiles en el análisis por partes de fonemas fricativos y de compresión. Un camino intermedio tiene como consecuencia una ambigüedad entre formantes y armónicos en una señal (Fant, *Encyclopedia of Acoustics*: 1997, p. 1589-1590).

El análisis espectral puede realizarse por medio de un proceso de filtrado (a través de un banco de filtros) o por un análisis digital de Fourier de la forma de onda basado en la transformada discreta de Fourier, que usualmente es implementada por la transformada rápida de Fourier. La duración efectiva de la respuesta impulso, que es del orden de $1/B$, funge como un equivalente funcional de la duración T de la ventana de una Transformada Discreta de Fourier definida por:

$$V_n = \sum_{i=0}^{N-1} v_i \exp\left(\frac{-j2\pi ni}{N}\right)$$

donde N es el número de muestras, v_i es el número de la muestra i, y V_n el número del componente espectral n. Una definición mejorada del espectro, se consigue agregando una ventana de muestras con amplitud cero, antes del cómputo.

Se debe poner cuidado en la elección de los parámetros [que incluyen frecuencia de muestreo, número de muestras usadas en la transformada rápida de Fourier (N), intervalo de tiempo cubierto en ms (T)], de acuerdo a la si el análisis corresponde a sonidos de vocales, fricativas u otras. Asimismo, se debe mantener en mente el efecto que puede tener el emplear ventanas de análisis de naturaleza distinta, ya que la duración efectiva de una ventana de Hanning o Hamming, es alrededor de 40% menor que el valor nominal de una ventana rectangular, en otras palabras, enfatiza las sección intermedia de la ventana (Fant, *Encyclopedia of Acoustics*: 1997 pp. 1589-1590).

Procesamiento digital de señales analógicas (codificación de la Voz)

La transmisión de voz usando redes de datos requiere que la señal de voz este codificada de manera digital. El almacenaje digital de señales de audio, que pueden resultar en mayor calidad y menor tamaño que su contraparte análoga, es común en discos compactos, video-discos digitales y archivos MP3. En algunos sistemas de lenguaje hablado usan voz codificada para una comunicación más eficiente (Huang, Acero y Hon: 2001, p. 337). Las técnicas para la codificación de la voz están diseñadas para convertir la forma de onda de la voz en códigos digitales con un mínimo de pérdida de información. La voz constituye una señal con una gran cantidad de información redundante, lo cual se explota para minimizar el ritmo de información necesario para reproducir la señal a un nivel de precisión específico (Atal, *Encyclopedia of Acoustics*: 1997, p. 1599).

El ritmo de información de un canal digital es expresado en *bits por segundo.* La conversión de una señal analógica a forma digital consiste de dos pasos: muestreo y cuantización. El muestreo, como se ha revisado, es el proceso de conversión de una función continua en el tiempo, en una secuencia discreta que representa a la función en intervalos de tiempo regularmente espaciados, mientras que la cuantización convierte una amplitud continua en valores discretos (Atal, *Encyclopedia of Acoustics*: 1997, p. 1599).

El propósito esencial de la codificación de la señal de voz es el reducir el número de bits necesarios para representarla. El flujo de bits puede ser reducido eliminando la redundancia en la señal, con lo que aún puede recuperarse exactamente la señal original. Existen, sin embargo, algunas compresiones que implican cierto nivel de pérdida que no puede ser recuperada. La calidad de la señal recuperada es un atributo fundamental de un codificador de voz. Una medida de la calidad es la relación señal-ruido en dB´s. Todos los codificadores tienen cierto *retraso* (*delay*), que se pueden dividir en *retraso algorítmico* (por la operación en bloques de muestras), *retraso computacional* (el tiempo que tarda en procesar el cuadro), y el *retraso por transmisión* (debido al tiempo que toma al cuadro correr por el canal); en exceso, los retrasos pueden modificar la dinámica de una comunicación en dos vías.

Las técnicas de codificación de voz pueden ser divididas en dos clases: codificación de forma de onda que busca reproducir ésta de manera tan fidedigna como sea posible, y los llamados ***vocoders*** que pretenden preservar sólo las propiedades espectrales del habla en la señal codificada. Los primeros son capaces de producir voz de alta calidad con *tasa de transmisión* suficientemente elevados, mientras que los segundos logran voz inteligible a ritmos de bitaje mucho más bajos (Atal, *Encyclopedia of Acoustics*: 1997, p. 1599).

Representaciones de Señales de Voz

Existen gran cantidad de representaciones de señales de voz útiles en la codificación de voz, síntesis y reconocimiento. El tema central es la descomposición de la señal de voz haciendo las veces de fuente que pasa por filtros lineales variantes en el tiempo. Tales

filtros pueden ser derivados de modelos de producción de voz basados en teorías acústicas donde la fuente representa el flujo de aire a través de los pliegues vocales, mientras que los filtros representan las resonancias del tracto vocal, mismas que cambian en el tiempo (Huang, Acero y Hon: 2001, p. 275). Tal modelo fuente-filtro se ilustra en la *Imagen 21*. Algunos otros métodos son inspirados en modelos de percepción del habla.

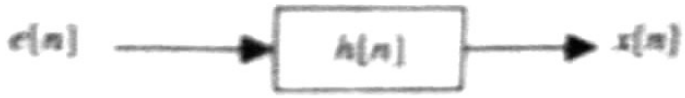


Imagen 21: Modelo Fuente-Filtro básico

Una vez que el filtro ha sido estimado, la fuente puede ser obtenido haciendo pasar la señal de voz por el filtro inverso. La separación de la fuente y el filtro es uno de los mayores retos en el procesamiento de voz. La clasificación de fonemas es dependiente en su mayor parte, de las características del filtro. De hecho, tradicionalmente los reconocedores de voz estiman las características del filtro e ignoran la fuente. Muchas técnicas de síntesis utilizan modelos fuente-filtro debido a la flexibilidad que permite para alterar el tono y el filtro. Al permitir también una *tasa de transmisión* baja, también es muy usado en los *vocoders*.

Muestreo

Para poder utilizar los métodos de procesamiento digital de señal, es necesario convertir la señal de voz $x(t)$, naturalmente analógica, a una señal digital $x[n]$. El *muestreo* de una señal de tiempo continuo en tiempos iguales a nT, donde T es el intervalo de muestreo, resulta en una secuencia de muestras de tiempo discreto.

Uno de los resultados más importantes dentro de la teoría de muestreo es el que se conoce como el *teorema del muestreo* (Oppenheim, Willsky: 1994, p. 547). Tal teorema forma un puente entre las señales de tiempo continuo y las señales de tiempo discreto, ya que proporciona un mecanismo para representar una señal de tiempo continuo mediante una señal de tiempo discreto. En éste, una secuencia de muestras $\boldsymbol{x[n]}$ es obtenida a partir de unas señal de tiempo continuo $\boldsymbol{x_c(t)}$ de acuerdo a la relación:

$$x[n] = x_c(nT) \qquad \text{- infinito < n < infinito}$$

donde $\boldsymbol{T}$ es el periodo de muestreo, y su recíproco $\mathbf{f_s = 1/T}$, es la frecuencia de muestreo.

El procesamiento de señales de tiempo discreto, es más flexible debido al desarrollo de sistemas digitales de tiempo discreto de bajo costo, ligeros, programables y fácilmente reproducibles, y permiten transformar una señal de tiempo continuo en tiempo discreto, para más adelante retransformarla en una de tiempo continuo (Oppenheim, Willsky: 1994, p. 547). Sin embargo, tales posibilidades también vienen con algunas implicaciones como se verá en la siguiente sección.

El teorema del muestreo o teorema de Nyquist

Una señal, en ausencia de condiciones o información adicional, no puede especificarse de manera unívoca por una secuencia de muestras igualmente espaciadas. Existen, en general, una cantidad infinita de señales que pueden generar un conjunto dado de muestras. Sin embargo, si una señal es de banda limitada y si las muestras son tomadas lo suficientemente cercanas unas de otras, en relación con la frecuencia más alta presente en la señal, entonces las muestras representan de manera unívoca a la señal y se puede reconstruir perfectamente.

Si de una señal $x_c(t)$ de banda limitada, son extraídos segmentos de tiempo igualmente espaciados, en otras palabras, si se modula en amplitud con un tren de pulsos periódicos, ésta se puede recuperar exactamente mediante un filtrado paso bajas, si la frecuencia fundamental del tren de pulsos modulador es mayor al doble de la frecuencia más alta presente en $x_c(t)$. La habilidad para recuperar $x_c(t)$ es independiente de la duración en tiempo de los pulsos individuales. Conforme tal duración se hace arbitrariamente más pequeña, la *modulación* de la amplitud de pulsos estará efectivamente representando a la señal $x(t)$ mediante muestras instantáneas igualmente espaciadas en el tiempo.

En cualquier sistema práctico de modulación de amplitud de pulsos, se debe mantener constante una potencia promedio en el tiempo para la señal modulada cuando se hace que el ancho de pulso sea pequeño. Conforme este ancho de pulso se aproxima a cero, la señal modulada se convierte en un tren de impulsos en el que los impulsos individuales tienen valores correspondientes a muestras instantáneas de $x_c(t)$ en valores de tiempo espaciados T segundos, o el intervalo de muestreo (Oppenheim, Willsky: 1994, p. 551). Nos referimos a un sistema que implemente esta tarea como un convertidor continuo-discreto ideal (C/D). Para aplicaciones prácticas, la operación de muestreo se implementa con un convertidor analógico-digital (A/D), que pueden considerarse aproximaciones al convertidor (C/D) ideal. Se deben tomar consideraciones importantes que incluyen la cuantización de las muestras de salida, linealidad, la necesidad de circuitos muestreador/retenedor, y limitaciones en la tasa de muestreo (Herrera: 2006, p.4).

Como ya se mencionó, el muestreo es generalmente irreversible, ya que dada una salida discreta, en general no es posible reconstruir la entrada del *muestreador,* ya que muchas señales de tiempo continuo pueden producir la misma secuencia de muestras de salida. Tal ambigüedad inherente es primordial en el procesamiento de señales pero puede ser removida restringiendo la clase de señales de entrada con las consideraciones dadas anteriormente. Matemáticamente es conveniente representar el proceso de muestreo en dos etapas, consistentes en un modulador de tren de impulsos seguido de la conversión del tren de impulsos en una secuencia, en la cual la señal es indexada en la variable **n**, que introduce un normalización en el tiempo, pero que no contiene información sobre la frecuencia de muestreo (Herrera: 2006, p. 4).

El tren de impulso $p(t)$ se conoce como la función de muestreo, misma que tiene a T como periodo de muestreo, y la frecuencia fundamental de $p(t)$, $\omega_s=2\pi/T$ como la frecuencia de muestreo. En el dominio del tiempo se tiene:

$$x_p(t) = x(t)p(t)$$

donde

$$p(t) = \sum_{-\infty}^{+\infty} \delta(\tau - nT) \tag{65}$$

$x_p(t)$ es un tren de impulsos cuyas amplitudes son iguales a las muestras de $x_c(t)$ en intervalos espaciados por T, lo cuál se representa como:

$$x_p(t) = \sum_{n=-\infty}^{\infty} x(nT)\delta(t - nT) \tag{66}$$

Calculando sus coeficientes de la serie de Fourier se obtiene:

$$P(\omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi k}{T}\right) \tag{67}$$

Entonces:

$$P(\omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_0) \tag{68}$$

por lo que:

$$X(\omega) = \frac{1}{T} \sum_{k=0}^{\infty} X(\omega - k\omega_0) \tag{69}$$

Un análisis detallado de esta última, nos dice que $X_p(\omega)$ es una función periódica en el dominio de la frecuencia que consiste en la suma de réplicas de $X(\omega)$ desplazadas y escaladas por $1/T$, donde si $\omega_M < (\omega_s - \omega_M)$, no existe traslape entre réplicas desplazadas de $X(\omega)$, de lo contrario si las habrá. Para el primero de los casos, $X(\omega)$ se reproduce fielmente en múltiplos enteros de la frecuencia de muestreo. Por lo tanto, si $\omega_s > 2\omega_M$, una señal de ancho de banda limitado $x(t)$, con $X(\omega) = 0$ para $|\omega| > \omega_M$, se puede recuperar exactamente a partir de $X_p(t)$ por medio de un filtro paso bajas con ganancia T y un frecuencia de corte mayor que ω_M y menor que $(\omega_s - \omega_M)$, donde $\omega_s = 2\pi/T$, y está determinada unívocamente por sus muestras $x(nT)$, $n = 0, +-1, +-2...$ (Oppenheim, Willsky: 1994, pp. 550-554)

La frecuencia de muestreo ω_s también se conoce como la frecuencia de Nyquist.

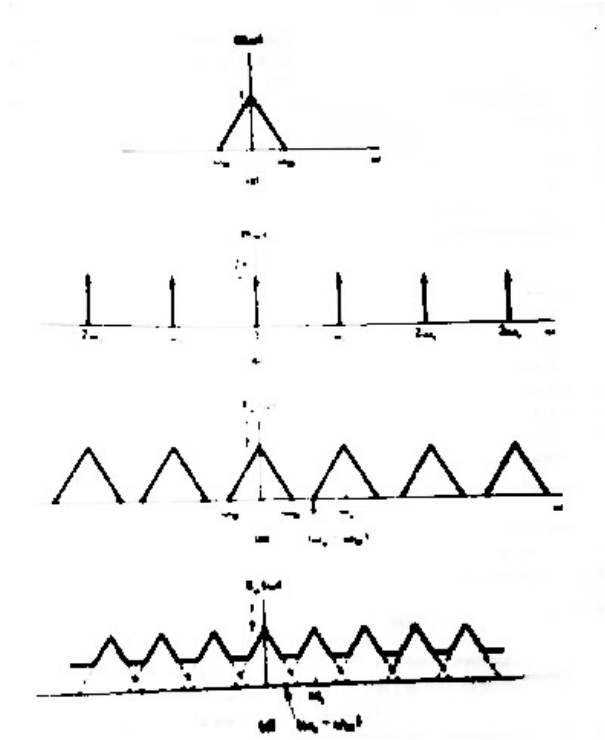


Imagen 22. Efecto en el dominio de la frecuencia del muestreo en el dominio del tiempo

Cuando la desigualdad $\omega_s > 2\omega_M$ no se cumple, ya no se encuentran réplicas del espectro $X(\omega)$ de $x(t)$ y la señal ya no es recuperable con un filtro paso-bajas. Tal efecto, en el que los términos individuales de la ecuación (69) se sobreponen, es conocido como traslape, y que tiene como consecuencia que una frecuencia original ω_0 asuma la identidad o "alias" de una frecuencia inferior (ω_s-ω_0).

Para la transmisión de voz a través de canales tradicionales de teléfono, la frecuencia máxima utilizada es de 4 kHz, por lo que la frecuencia de muestreo es de 8 kHz y el intervalo de muestreo de 125 microsegundos. Las señales de voz hablada pueden ser limitadas en frecuencia a 10 Khz sin pérdida significativa en la percepción, sin embargo, esto es distinto en la voz cantada debido a que el color y timbre de esta es más importante (Atal, *Encyclopedia of Acoustics:* 1997, p. 1599).

Cuantización

La cuantización, es el procedimiento para representar la amplitud de una forma de onda analógica por medio de ciertos valores preasignados. El proceso de cuantización invariablemente introduce errores, o ruido de cuantización en la señal. La relación de potencia de la señal contra potencia de ruido, es usualmente expresada por la relación señal a ruido (SNR-Signal to Noise ratio) y es una medida de gran importancia para la valoración de la calidad de los cuantizadores (Atal, *Encyclopedia of Acoustics*: 1997, p. 1600).

Cuantización escalar y vectorial

La *cuantización* independiente de cada muestra obtenida de una señal continua, lleva por nombre *cuantización escalar*, mientras que la *cuantización* conjunta de un bloque de muestras de una señal es conocida como *cuantización vectorial*. Esta última resulta en una menor cantidad de errores que la *cuantización escalar*, con el mismo número de valores (Atal, *Encyclopedia of Acoustics*: 1997, p 1600).

Cuantización Uniforme

En la *cuantización uniforme*, los niveles de *cuantización* son espaciados a intervalos iguales. Para un *cuantizador* con un paso de longitud Δ, el *ruido de cuantización* se distribuye uniformemente en el intervalo que va desde -Δ/2 a +Δ/2 con una potencia igual a $\Delta^2/12$. Para *cuantizar* un intervalo de valores de señal de $-x_{max}$ a $+x_{max}$ con un cuantizador de *n*-bits, el tamaño del paso se vuelve:

$$\Delta = 2x_{max}/2^n$$

Codificadores de Forma de Onda Escalares: Linear Pulse Code Modulation (PCM)

Los convertidores análogo-a-digital realizan ambos, *muestreo* y *cuantización* simultáneamente. En la *cuantización* se codifica cada muestra con un número fijo de bits. Con *B* bits, es posible representar a 2^B niveles de *cuantización* separados. La salida del *cuantizador* está dada por:

$$\hat{x}[n] = Q\{x[n]\} \tag{70}$$

Linear Pulse Code Modulation (PCM) está basado en el entendido de que la señal discreta de entrada, $x[n]$ *está acotada*

$$|x[n]| \leq X_{\max} \tag{71}$$

y que se usa una *cuantización uniforme* con un paso de *cuantización* Δ, que es una constante para todos los niveles x_i.

$$x_i - x_{i-1} = \Delta \tag{72}$$

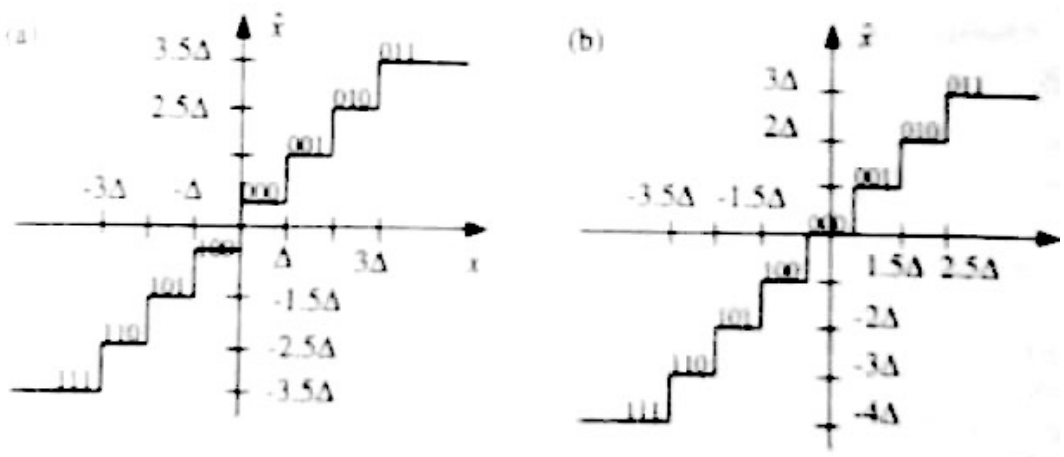


Imagen 23. Características de la Cuantización uniforme de tres bits: (a) mid-riser, (b) mid-tread

Las características de entrada/salida se muestran en la *Imagen 23* para el caso de un *cuantizador* unitario de 3 bits. El *cuantizador* llamado *mid-riser*, difiere del *mid-tread* en que este último tiene un nivel negativo más que los positivos, mientras el primero tiene el mismo número de niveles positivos y negativos. El código $c[n]$ se expresa en dos representaciones complementarias. Para el *mid-riser* la salida puede ser obtenida del código $c[n]$ a través de:

$$\hat{x}[n] = sign(c[n])\frac{\Delta}{2} + c[n]\Delta \tag{73}$$

y para el *mid-tread*

$$\hat{x}[n] = c[n]\Delta \tag{74}$$

Existen dos parámetros independientes para un *cuantizador uniforme*: el número de niveles $N=2^B$, y el tamaño del paso Δ. Asumiendo que la (71) es válida, entonces se tiene:

$$2X_{max}=\Delta 2^B \tag{75}$$

Es conveniente expresar la relación entre la muestra *cuantizada* y la muestra no *cuantizada*:

$$\hat{x}[n] = x[n] + e[n] \tag{76}$$

donde $e[n]$ es el r*uido de cuantización*. Si se escogen Δ y B para que satisfagan la ecuación (75) entonces:

$$-\frac{\Delta}{2} \le e[n] \le \frac{\Delta}{2} \tag{77}$$

Cada bit contribuye con 6 dB de la relación sonido a ruido. El audio digital almacenado en computadoras así como el Compact Disc, usan PCM lineal de 16 bits como su formato principal. Este formato fue inventado en los años sesenta por James T. Russell, y fue lanzado comercialmente en 1982, alcanzando un éxito sin precedente en cuanto a consumo de tecnología electrónica, ya que hacia 1997 existían unos 700 millones de CD players.

Cuantización no uniforme

En la cuantización uniforme, la potencia del ruido se mantiene constante independientemente del nivel de la señal. La potencia en la señal de voz varía sobre un considerable intervalo de valores, por lo que los segmentos de voz con niveles bajos de potencia son cuantizados con una pobre relación señal-a-ruido. Este problema puede ser evitado si los intervalos de cuantización no son uniformes, sino espaciados uniformemente para incrementar con el valor de la señal. La cuantización no uniforme, es equivalente a la compresión de la señal analógica seguida por una cuantización uniforme de la salida.

Existen además, códigos de cuantización por compresión que logran que la potencia del ruido sea decrementada considerablemente durante segmentos de voz de baja amplitud, como es la ley m de compresión que se define como:

$$F(x) = \mathrm{sgn}(x)\, x_{\max} \frac{\log(1 - \mu |x| / x_{\max})}{\log(1 + \mu)}$$

donde x representa la señal de entrada, x_{max} es el valor máximo de amplitud de la señal, y m es el parámetro que controla el grado de compresión, que adquiere valores típicos entre 100 y 200.

Por otro lado, existen códigos de *cuantización* que toman ventaja del hecho de que muestras adyacentes de señales de voz, están altamente correlacionadas y *cuantizan* las diferencias entre la señal y su valor de predicción basado en las muestras previamente *cuantizadas*. El intervalo dinámico de la señal de diferencia resulta mucho más pequeña que la señal de entrada, por lo que necesita un número menor de bits para obtener el mismo ruido de *cuantización*. Tales métodos se conocen como diferenciales adaptativos (ADPCM), y en ellos la señal de diferencia es *cuantizada* con control adaptativo de la dimensión del escalón de *cuantización*.

Códigos de predicción

Los sistemas de predicción lineal de codificación explotan las correlaciones de muestra-a-muestra en la voz para predecir la muestra actual de voz a partir de los valores *cuantizados* pasados y para reducir el número de bits utilizados de la señal de voz codificada. La diferencia entre la muestra de voz y su valor predicho es generalmente mucho más pequeña que la amplitud de la señal de voz, y por ende puede ser codificada con un menor número de bits. En un sistema de predicción, la señal de voz es muestreada,

y el predictor forma un estimado del valor presente basado en las muestras pasadas de la señal *cuantizada* de voz en el transmisor. El valor predicho se sustrae del valor de la señal para formar la diferencia que es *cuantizada*, codificada y transmitida al receptor. La señal transmitida es decodificada en el transmisor y utilizada para predecir la muestra siguiente usando un predictor que es idéntico al empleado en el transmisor. En el receptor, la señal transmitida es decodificada y agregada al valor predicho de la señal para generar muestras de la señal decodificada de voz que son luego filtradas por un paso-bajo para producir una señal análoga de voz. El error entre la señal original y la señal decodificada de voz, es idéntico al introducido por el cuantizador, codificador y decodificador. Por lo tanto, en la voz decodificada, la relación señal-a-ruido excede aquella de la señal de diferencia decodificada, por un factor igual a la relación entre el promedio al cuadrado de la señal de voz de entrada y el promedio cuadrado de la señal de diferencia.

Uno de los métodos más poderosos para el análisis de voz está basado en los *códigos de predicción lineal* (LPC) también conocido como método de auto-regresión. Este método es muy usado debido a que aunque es rápido y sencillo, es una manera efectiva de estimar los parámetros más importantes de las señales de voz.

Códigos de Predicción Lineal (LPC)

Un filtro todo-polos con un número suficiente de polos es una buena aproximación a las señales de voz.

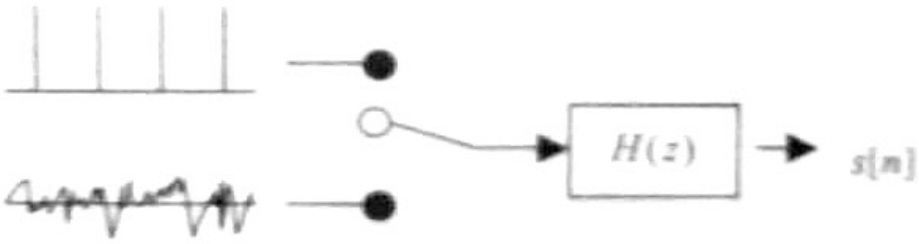


Imagen 24. Modelo Fuente-Filtro para voz fonada y sorda

Por lo tanto, se podría modelar el filtro en la *Imagen 24* como:

$$H(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{78}$$

donde p es el orden del análisis LPC. El *filtro inverso A(z)* está definido como

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{79}$$

Tomando el inverso de la transformada *z* en la ecuación (79) resulta en:

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + e[n] \tag{80}$$

Los códigos de predicción lineal obtienen su nombre del hecho de que predicen la muestra actual como una combinación de las *p* muestras pasadas:

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{81}$$

El error en la predicción cuando se usa esta aproximación es

$$e[n] = x[n] - \tilde{x}[n] = x[n] - \sum_{k=1}^{p} a_k x[n-k] \tag{82}$$

Capítulo IV

I. *Síntesis y el Reconocimiento de voz: Historia*

Es dentro de la investigación sobre codificación de voz en donde se tiene un mayor interés en la transmisión y almacenaje eficiente de la información del habla. La década de los setenta, atestiguó un creciente interés por desarrollar la tecnología que permitiera la interacción por medio de voz entre máquinas sofisticadas y humanos. Ello requería proporcionarles a tales máquinas, una "boca" (para hablar la información), y "oídos" (para escuchar y comprender comandos). El interés en la comunicación humano-máquina a través de voz, usualmente es solidificado por de la investigación en *síntesis de voz* y *reconocimiento de voz* automatizado (Flanagan, *Encyclopedia of Acoustics*: 1997, p. 1557).

La síntesis efectiva de voz está basada en una comprensión fundamental de la física de la producción de voz, así como de las constricciones linguísticas que caracterizan un lenguaje dado. Las partes constitutivas de la rama de síntesis de voz incluyen el mecanismo humano de generación de sonidos, la propagación acústica de las ondas en el tracto vocal, la dinámica y fisiología del movimiento articulatorio, y las convenciones del lenguaje que gobiernan las secuencias sonoras permitidas (Flanagan, *Encyclopedia of Acoustics*: 1997, p. 1561). Tales mecanismos deben ser simulados por una máquina para producir habla sintética. De manera ideal, se busca poder sintetizar irrestrictamente texto impreso, y a tal fin se le conoce como síntesis *texto-a-voz*. La inteligibilidad para la síntesis irrestricta de texto es generalmente buena, sin embargo, la calidad de la señal es aún muy mecanizada. Las fronteras de esta actividad radican en la síntesis de sonidos mas naturales y la duplicación de las variedades de las características de las distintas voces, así como en una variedad de idiomas.

Uno de los primeros intentos de síntesis de voz, se remonta al año de 1779, cuando el científico ruso Kratzenstein construyó un conjunto de 5 resonadores acústicos que al activarse producían imitaciones de las vocales (Herrera: 2006, Capítulo 4, p. 1).

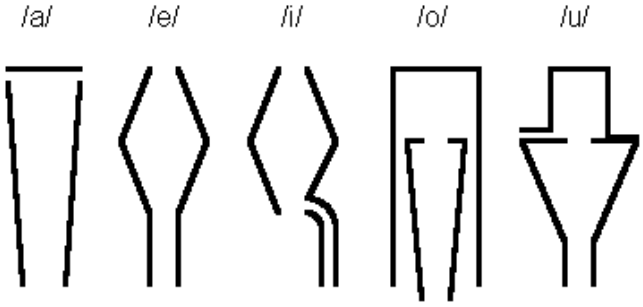


Imagen 25. Resonadores de Kratzenstein (vocales)

En 1791, el húgaro Von Kemplen, construye una máquina capaz de pronunciar palabras y frases completas, que consistía de largos pliegues que proporcionaban un flujo de aire a un ducto que excitaba un tubo de caucho. A éste, se agregaban tubos y silbatos adicionales para imitar sonidos nasales y fricativos.

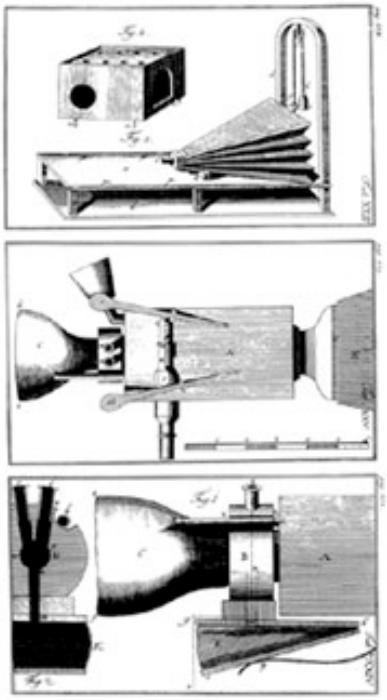


Imagen 26. Sintetizador Mecánico de Von Kemplen

En 1937, Reisz construye un sintetizador de voz mecánico que simulaba el movimiento de los articuladores de la voz presionando teclas que variaban la forma del tracto vocal, y que cuando se operaba hábilmente podía producir voz continua.

Uno de los primeros sintetizadores eléctricos fue el llamado *Voder*, construido en 1938, que modelaba el tracto vocal usando 10 filtros paso-banda contiguos y paralelos que cubrían la banda de frecuencias de la voz y excitados por zumbidos periódicos, o ruido aleatorio, y cuyo tono era controlado por teclas, una barra en la muñeca y un pedal (Herrera: 2006, Capítulo 4, pp. 1-2). Un año después, el *Vocoder* de Dudley permitía la transmisión de la voz vía telefónica reduciendo el ancho de banda (a una décima parte), y consistente de un analizador de diez filtros pasa-banda en paralelo cuyas amplitudes de salida eran medidas continuamente y enviadas a un sintetizador, y cuyos parámetros controlaban las ganancias de un conjunto de filtros pasa-banda idénticos a los del analizador, que eran excitados por una fuente de pulsos o de ruido (Herrera: 2006, Capítulo 4, p. 3). La señal de voz era reconstruida por la suma de las salidas de los filtros.

En los 40´s Dunn daba un salto cualitativo representando al tracto vocal como una línea de transmisión acústica y a cada una de sus secciones mediante una red eléctrica equivalente; y en 1953, Lawrence se aproximaba al problema de la síntesis de voz en el dominio del tiempo, basándose en el hecho de que la respuesta de los sistemas resonantes a una excitación impulsiva es una oscilación sinusoidal atenuada, logrando producir sonidos sonoros de esta manera. Así, los primeros sintetizadores, tanto articulatorios como de terminal análoga, fueron construidos usando circuitos análogos, lo que implicaba que fueran difíciles de controlar, y no producían voz natural. Los desarrollos en la computación digital en los años 60s, revolucionaron la investigación sobre la voz, permitiendo simular nuevos diseños y controlar los sistemas para su mejor evaluación (Herrera: 2006, Capítulo 4, p. 4).

Métodos de Síntesis

Los métodos de síntesis de voz pueden ser categorizados en tres tipos de acuerdo con el modelo utilizado en la generación de voz:

La Síntesis Articulatoria

Una de las mayores dificultades en el diseño de la síntesis *texto-a-voz*, es formular un conjunto de reglas que puedan generar transiciones realistas entre prominencias espectrales y modelar de manera precisa la co-articulación; aspectos esenciales para la producción de voz sintética natural. La resolución a tal problema tal vez se encuentre mediante una aproximación a la síntesis de voz que modele los movimientos detallados de los articuladores de la voz, así como la generación y propagación del sonido dentro del tracto vocal (Herrera: 2006, Capítulo 4, pp. 19-20).

La síntesis articulatoria, utiliza un modelo físico de producción de voz que incluye todos los articuladores. Ésta modalidad, usa parámetros que modelan los movimiento mecánicos y sus distribuciones resultantes de velocidad de volumen y presión sonora en los pulmones, laringe, y tractos vocales y nasales. Debido a que los articuladores humanos no tienen tantos grados de libertad, los modelos por articulación a menudo usan hasta 15 parámetros para manejar un sintetizador por formantes (Huang, Acero y Hon: 2001, p. 803).

Los componentes principales de un sintetizador por articulación son:

a) Un *modelo articulatorio* que transforme un conjunto de 6-10 parámetros que representen posiciones de articulación; la relación entre articuladores y acústica es *muchos-a-uno*. Asumiendo que los articuladores no cambian rápidamente en el tiempo, es posible estimar el área del tracto vocal a partir de las frecuencias de prominencias espectrales. A menudo se utilizan en estos modelos cinco parámetros de articulación: área de apertura de labios, constricción formada por el filo de la lengua, apertura a las cavidades nasales, promedio de área glotal, y ritmo de expansión activa o contracción del volumen del tracto vocal detrás de la constricción, todos ellos en función del área transversal del tracto vocal. A estos cinco parámetros se le adicionan las primeras cuatro prominencias espectrales y la frecuencia fundamental. Los parámetros de área pueden ser obtenidos de voz real a través de rayos X, e imágenes de resonancia magnética, aunque algunas de estas técnicas alteran la voz natural.

b) Un *modelo de tubo acústico*, para el cuál la función del área transversal especifica el área transversal en diferentes posiciones a lo largo de la longitud y determina la "forma" del tubo acústico, que modela el flujo de aire y la propagación del sonido dentro del tracto vocal.

c) Un *modelo de excitación* o *modelo de cuerdas vocales* que controla al tubo acústico, donde se modela el flujo modulado con el detalle deseado (Herrera: 2006, Capítulo 4, p. 20). Si se incorpora la dinámica del sistema en el modelo articulatorio, el control es ejecutado proporcionando una secuencia de parámetros articulatorios objetivos. Así se

logra de manera natural la coarticulación y sin reglas complejas.

Dado que los articuladores se mueven de manera relativamente lenta, la tasa de información para un sintetizador articulatorio, se cree, puede ser tan baja como 50 bits/s, sugiriendo que algún día, esta técnica podría ser la mejor manera de sintetizar voz, suponiendo que se desarrollen técnicas para extraer los objetivos articulatorios de la señal de voz (Herrera: 2006, Capítulo 4, p. 20).

La Síntesis por Formantes y por predicción lineal

La síntesis por formantes es del tipo de terminal análoga, y es generalmente implementada usando redes eléctricas, analógicas o digitales, teniendo respuestas en frecuencias similares a las del tracto vocal. Estas redes son excitadas mediante una fuente eléctrica similar a la fuente de sonido que excita al tracto vocal. Es decir, un generador de pulsos quasi-periódicos en el caso de sonidos sonoros, y un generador de ruido aleatorio en el caso de sonidos sordos. Un modelo fuente-filtro es utilizado, donde el filtro está caracterizado por la variación lenta de las frecuencias formantes. Es posible sintetizar una vocal estacionaria, haciendo pasar una forma de onda de impulso glotal periódico a través de un filtro con la frecuencia formante del tracto vocal. La fuente de sonidos sonoros, se puede obtener al integral una señal cuadrada de salida de un comparador de histéresis, para producir una fuente de señal triangular (como es usualmente la forma de onda producida por la glotis). El control de la frecuencia de la salida se logra al variar un resistor de manera electrónica (Herrera: 2006, Capítulo 4, p. 5), mientras que un potenciómetro puede ser usado para variar la relación marca-espacio de la salida del comparador y por lo tanto la forma de la señal triangular.

Para el caso de la voz *sorda,* puede usarse ruido blanco como fuente, producido cuando la unión base-emisor de un transistor, se encuentra suficientemente polarizada en inversa para que ocurra una ruptura, y amplificándolo posteriormente. En un sintetizador digital por formantes, el ruido aleatorio es obtenido al generar una secuencia pseudo-aleatoria de números, generando un número aleatorio en cada instante del muestreo (Herrera: 2006, Capítulo 4, p. 5). En la práctica, los sonidos de las señales de voz no son estacionarios y por lo tanto es necesario cambiar el tono de la fuente glotal y la frecuencia formante a través del tiempo (Huang, Acero y Hon: 2001, p. 796).

Cada resonancia o formante del tracto vocal es simulada mediante un circuito resonante de segundo orden, con una frecuencia central y ancho de banda variables, siendo los intervalos frecuenciales típicos de los primeros tres formantes de la voz (F_1, F_2, F_3): 100-1000 Hz, 700-2500 Hz, y 1500-3000 Hz, respectivamente. El ancho de banda de cada formante también varía, aunque comparado con la variación de la frecuencia de los formantes, es perceptualmente menos significativa y los anchos de banda de los formantes en un sintetizador, con frecuencia son fijos (Herrera: 2006, Capítulo 4, pp. 5-6). Hoy en día se sabe, que el mínimo cambio porcentual en la frecuencia de los formantes es de 5%, mientras que el respectivo para el ancho de banda es de 40%. Usualmente se utilizan tres resonadores de formantes, pero a menudo se agrega un cuarto para mejorar la calidad de voz. Los resonadores también pueden ser implementados mediante

filtros digitales, con una función de transferencia similar a los electrónicos. Los generadores de formantes pueden ser conectados en serie o paralelo. En un sintetizador serial, la salida de un resonador de formante provee la entrada del siguiente.

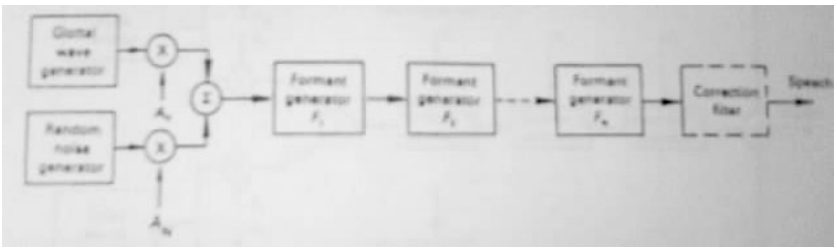


Imagen 27. Sintetizador serial por formantes básico

Normalmente, se utilizan tres filtros de formantes, haciendo que los formantes de alta frecuencia se agrupen en un sólo filtro de corrección de formantes mayores. Este último es un filtro de aplanamiento espectral que compensa la atenuación acumulativa de -12 dB/octava de cada filtro de formante en las frecuencias altas, y sólo se requiere en implementaciones analógicas, ya que en las digitales, la naturaleza de la respuesta en frecuencia del resonador digital por encima de la resonancia es diferente. Éste, cae a una tasa menor en la altas frecuencias, y de hecho es plano a la mitad de la frecuencia de muestreo. Se ha mostrado que los sintetizadores digitales que no poseen corrección para formantes altos proporcionan una aproximación más cercana a la respuesta del tracto vocal que los sintetizadores con corrección de formantes altos (Herrera: 2006, Capítulo 4, p. 8).

Aunque la estructura serial proporciona una aproximación muy cercana a la respuesta del tracto vocal a sonidos vocálicos no nasales o similares, es sin embargo, incapaz de simular de manera precisa las antiresonancias o ceros que ocurren en la producción de sonidos nasales, fricativos o plosivos. Para producir tales dentro del marco de un sintetizador por formantes serial, puede agregarse una red de polos-ceros en serie con los resonadores , o alternativamente la red puede agregarse en paralelo.

Por otro lado, en un sintetizador por formantes en paralelo, los generadores de formantes son conectados en paralelo y sus salidas son pesadas y sumadas para producir la señal de voz sintética. En el proceso, se especifican las amplitudes de los formantes, así como sus frecuencias, lo que tiene por consecuencia favorable, el generar más parámetros a controlar, permitiendo un mayor control sobre la forma del espectro de frecuencias sintetizado. Bajo el rubro de sintetizadores por formantes paralelos, se usa normalmente un resonador paso-bajas de segundo orden para la primera formante, y resonadores paso-banda de segundo orden para los formantes mayores. La función de transferencia de una conexión en paralelo de resonadores contiene tanto polos como ceros, lo que presenta ventajas en la producción de sonidos nasales y fricativos, pero puede conducir a mínimos profundos entre los picos de los formantes de vocales y sonidos afines. Estos mínimos, pueden ser eliminados si se alternan los signos de las amplitudes de los formantes (Herrera: 2006, Capítulo 4, pp. 8-10).

Con el fin de explotar las mejores características de ambos tipos de sintetizadores por

formantes (en serie y en paralelo), el investigador de voz Denis Klatt desarrolló en los años 80s, un sintetizador por formantes serial/paralelo, caracterizado por utilizar una conexión en serie de resonadores para producir sonidos sonoros no nasales, y una conexión en paralelo para producir nasales, fricativas y plosivas, obteniendo muy buenos resultados.

La denominada *síntesis por regla* o, *por formantes,* se refiere a un conjunto de reglas sobre cómo modificar el tono, las frecuencias formantes y otros parámetros de un sonido a otro, manteniendo la continuidad presente en sistemas físicos. Si éstos parámetros son lo suficientemente precisos entonces, se puede obtener una voz sintética de buena calidad (de hecho, virtualmente indistinguible de la voz natural), a partir de la cual se obtuvieron los parámetros. Algunos de tales sistemas han sido exitosos en reproducir con fidelidad una grabación natural, mientras los parámetros hayan sido manualmente escogidos como el sistema de Klatt (Huang, Acero y Hon: 2001, p. 797). Sin embargo tales logros son costosos en términos de tiempo y esfuerzo, al requerir una gran cantidad de sintonización manual de valores de parámetros, haciendo de esta forma de síntesis un problema difícil (Herrera: 2006, Capítulo 4, pp. 10-11).

Síntesis por predicción lineal

El proceso de la predicción lineal es similar a la síntesis por formantes, en el sentido de que es un modelo fuente-filtro de producción de voz en el que la señal de error representa la señal de excitación, y el tracto vocal es representado por el filtro todo polos H(z). Conociendo la señal de error y los coeficientes de predicción lineal, se puede reconstruir la señal original de voz aplicando la señal de error al filtro digital. Este método constituye tanto un método de síntesis como de análisis de voz.

La estructura del sintetizador de predicción lineal, es una en la que la señal de error es "estilizada" como un generador de muestras unitarias periódicas, en la frecuencia del tono en el caso de la voz sonora, o como generador de números aleatorios en el caso de la voz sorda.

La Síntesis por Copia, por Fonemas y por Concatenación

Un problema significativo en la síntesis de voz, es el cómo obtener el conjunto de parámetros de control para producir una locución particular. Una de las maneras más evidentes para la realización de tal tarea, es el obtener los parámetros de control a partir de un análisis de la voz real. La técnica de producción de voz sintética, mediante el análisis de voz real, es a menudo llamada *analisis/síntesis de voz* o simplemente *síntesis por copia*. En los sistemas donde tal técnica encuentra aplicaciones, la señal de voz es codificada en los parámetros de control del sintetizador y almacenada ya sea en una memoria de semiconductor o en un medio magnético. La ventaja de esta técnica sobre el almacenamiento directo de la señal de voz original en el dominio del tiempo, subyace en la economía de almacenamiento, lográndose una compresión por un factor aproximado de 27, que es similar a la lograda por voz codificada por medio de códigos de predicción lineal. El principal problema de la síntesis por copia, era que no podía ser usada cuando se requería una cantidad grande o ilimitada de vocabulario o cuando debían generarse

una amplia variedad de mensajes. La técnica requiere además una sesión de grabación y de análisis (Herrera: 2006, Capítulo 4, pp. 12-13).

Una forma alternativa para obtener los parámetros de control del sintetizador es generarlos a partir de la transcripción fonética de la locución o *síntesis por fonemas*. Para ello, la cadena de fonemas que representa la locución a ser sintetizada constituye la entrada de un programa de computadora que genera la secuencia de parámetros de control de la salida. Ello se realiza mediante un conjunto de reglas para la conversión de información fonética en información acústica, por lo que a esta técnica se le conoce también como *síntesis por regla*. Este tipo de sistemas tienen además un método de asignación de información prosódica (entonación y tensión) a la locución, ya que de no contar con el mismo, sonaría completamente antinatural. Esto se logra normalmente insertando marcadores prosódicos o símbolos en la cadena de fonemas de entrada. Ahora bien, aunque a nivel lingüístico, la voz puede ser vista como una secuencia de segmentos fonéticos, la manifestación acústica de tales unidades es mucho más compleja en particular debido al mecanismo de co-articulación, que da como resultado sonidos agrupados e incide en la naturalidad de la voz. La co-articulación genera movimientos complejos de la frecuencia y los valores de amplitud de los formantes entre un sonido y el siguiente. Existen además, muchas variaciones acústicas para cada fonema (*alófonos*), que dependen del contexto o posición en una locución. El programa de síntesis por regla, para producir voz natural, debe simular estos efectos de la mejor manera posible. Un sistema de síntesis por regla contiene cuatro componentes básicos (Herrera: 2006, Capítulo 4, pp. 13):

1. Una tabla de búsqueda o valores de amplitud y frecuencia en estado estable. Puede consistir de una serie de valores de frecuencia y amplitud de los formantes en estado estable para los tres primeros formantes de cada fonema, como el generado por Ainsworth en 1974, que además especificaba un cuarto formante fijo.

2. Información y reglas para generar transiciones de formantes entre sonidos vecinos.

3. Información y reglas para permitir variaciones alofónicas de los sonidos dependiendo de la naturaleza del sonido circundante.

4. Un mecanismo de asignación de patrones prosódicos a la locución.

Desafortunadamente, el procedimiento de generación de parámetros es incapaz, por sí mismo, de producir voz sintética de alta calidad debido al gran número de casos de excepción a considerar, uno de los cuales es que la sección estable de los sonidos plosivos es un silencio, y que son las transiciones de las formantes las que distinguen uno de otro. Las transiciones lineales entre formantes no son una buena aproximación, y a menudo son modeladas como la salida de sistemas de primer orden críticamente amortiguados con constantes de tiempo que pueden ser diferentes para cada formante, o que varían según el fonema. Pero incluso si los efectos de co-articulación y variaciones alofónicas son bien modelados, la voz sintética no sonará natural, a menos que exista un mecanismo adecuado para imponer información prosódica. Ello puede lograrse mediante (Herrera: 2006, Capítulo 4, pp. 15):

a) Un contorno adecuado para el tono o entonación.

b) Duración del fonema que abarca en gran parte el alargamiento de las vocales que se tensionarán en la locución.

c) Inserción de pausas.

Varios métodos se han desarrollado para asignar patrones prosódicos a una cadena de fonemas de entrada. El valor del tono y duración para cada fonema eran, en un principio, individualmente especificados, lo que constituía una tarea difícil e incómoda. Actualmente, los algoritmos prosódicos utilizan alguna forma de notación en la que los números, marcadores y delimitadores de la cadena de fonemas de entrada, indican el contorno del tono, las sílabas acentuadas y el ritmo total de la locución. Además, los símbolos fonéticos estándares del *IPA* (*International phonetic alphabet*) son codificados en una o dos letras para adecuarlos a los teclados de computadora convencionales. La duración de fonemas es calculada de acuerdo con la teoría de la *base isócrona* del ritmo de la voz. La palabra *base,* en este contexto, se refiere al intervalo de tiempo entre sílabas acentuadas sucesivas. Por otro lado, la teoría isócrona establece que la duración de cada sílaba es una base ajustada para hacer que la duración de cada base sea aproximadamente constante. Ninguna de las otras sílabas es acentuada; la locución es dividida por signos de puntuación en grupos tónicos y la forma del contorno de entonación es especificada por un número al inicio de cada uno. La inserción de pausas se logra mediante el uso de altos completos, y el carácter "^" denota un punto de respiración (Herrera: 2006, Capítulo 4, p. 16).

La síntesis de voz a partir de la entrada de fonemas es una representación conveniente y muy económica en cuanto a almacenamiento. Entre cuarenta y cincuenta fonemas, más unos cuantos marcadores prosódicos, pueden ser codificados usando 6 bits y, con una tasa normal de locución de alrededor de 12 fonemas por segundo, la tasa de información es 72 bits/s. Por ende, sólo 8 bytes de memoria pueden almacenar aprox. 15 minutos de voz (Herrera: 2006, Capítulo 4, p. 16).

Debido a que las características acústicas son relativamente invariantes en el centro de un fonema, y los efectos de co-articulación son capturados en la transición entre dos fonemas, para superar el difícil problema de escribir reglas para simular la co-articulación en la síntesis por fonemas, una aproximación ha sido el usar unidades fonémicas más largas, como la *Demi-sílaba* (definida a partir del inicio de una sílaba a la mitad de la vocal, o bien de la mitad de la vocal al final de la sílaba); los *difonemas* (unidades que van del centro de un fonema al centro del siguiente); las sílabas, e incluso las palabras, que poseen los efectos de co-articulación inherentes entre fonemas. Se requieren al menos 1000 difonemas para sintetizar voz con una calidad razonable, pero no de alta calidad. La interpolación de los valores de los parámetros en los límites de las Demi-sílabas es relativamente directa, ya que la co-articulación en tales regiones es débil. Para su generación, el tracto vocal está parcialmente abierto en el punto de articulación y no se produce turbulencia.

La ventaja del acercamiento *concatenativo* es que no requiere reglas ni afinación manual, además de que cada segmento es completamente natural, lo que mejora su salida (Huang, Acero y

Hon: 2001, p. 804). Sin embargo, existe una discontinuidad en el color de segmento a segmento. En caso de concatenarse dos segmentos de voz que no son adyacentes unos a otros, existirán diferencias espectrales y de prosodia. Estas discontinuidades espectrales, suceden en donde los formantes en el punto de concatenación no coinciden. Lo anterior hace que al desarrollar síntesis de voz por concatenación se consideren los siguientes puntos (Huang, Acero y Hon: 2001, pp. 804-805):

a) El tipo de segmento a utilizar.

b) El diseño de un inventario acústico, o conjunto de segmentos de voz (cuales, cuantos, etc.).

c) La mejor selección de segmentos de voz de cierta librería, dadas las características fonéticas y en prosodia.

d) Una manera de alterar la prosodia de un segmento para que empalme mejor a una prosodia de salida deseada.

En cuanto al cómo escoger las unidades a grabar, los requerimientos son similares a los de las unidades usadas en el reconocimiento de voz (Huang, Acero y Hon: 2001, p. 806):

a) La unidad deberá generar una baja *distorsión de concatenación.* Una buena manera de disminuir tal distorsión, es teniendo una menor cantidad de concatenaciones y por lo tanto utilizar segmentos más largos. Sin embargo, debido a que algunas concatenaciones son inevitables, también es deseable utilizar unidades que lleven discontinuidades "pequeñas" en los puntos de concatenación. El tener varias instancias por unidad es una alternativa a tener unidades muy largas, para poder ofrecer elección en las instancias con baja distorsión de concatenación.

b) La unidad debe tener una baja *distorsión prosódica*. Aunque no es crucial el contar con unidades con la misma prosodia que el objetivo deseado, el remplazo de una unidad con un tono ascendente por uno con tono descendente, puede resultar en un sonido poco natural. La alteración del tono y duración de un segmento es posible, a expensas de distorsión adicional.

c) La unidad debe ser *generalizable*, si se requiere transformación irrestricta de texto a voz.

d) La unidad debe ser *entrenable*. Los datos de entrenamiento deben ser suficientes para estimar la totalidad de las unidades.

En la práctica, se obtienen suficientes muestras de cada una de las unidades relevantes, para después analizarlas y almacenar una representación paramétrica de las mismas en el sistema (formantes, coeficientes LPC, etc). La voz es entonces sintetizada mediante la concatenación en secuencia de estos parámetros, usando reglas simples para la interpolación de los valores de los parámetros en los límites de la unidad. Como en los otros sistemas, se requiere

un mecanismo que asigne el patrón prosódico adecuado, por lo que el tono y duración de la unidad original son "desechados" (Herrera: 2006, Capítulo 4, p. 16). En general, tales sistemas sufren una gran variabilidad en calidades, y un reto práctico es el balancear la selección de los criterios presentados.

Síntesis de Texto a voz

La tarea de un sistema *texto-a-voz*, puede ser considerada como el inverso del reconocimiento de voz – el proceso de generar una maquinaria sin restricciones que puede generar un sonido natural de voz "humana", a partir de cualquier entrada de texto. La conversión de palabras en forma escrita a voz es un trabajo no trivial, ya que aunque fuera posible incluir un diccionario con la mayoría de las palabras habituales del idioma, aún se tendría que lidiar con millones de nombres, y acrónimos, por no mencionar que la entonación de las oraciones deben estar apropiadamente generadas para que suene natural (Huang, Acero y Hon: 2001, p. 6).

La historia de los convertidores *texto-a-voz* se remonta a los años 30´s, con el *Voder,* desarrollado en los laboratorios Bell y presentado en la feria mundial. El avance desde entonces ha sido vertiginoso, sobre todo utilizando las ventajas del incremento en potencia y almacenaje computacional.

Para su funcionamiento, los sistemas de síntesis *texto-a-voz* requieren dos tareas adicionales a las descritas para la síntesis por fonemas. El texto debe ser primero traducido en fonemas, para después determinar la prosodia directamente de su representación textual. Desafortunadamente, tales tareas no son fácilmente separables, ya que la acentuación y pronunciación se encuentran muy relacionadas, sobre todo si se consideran los casos en los que, para una misma palabra, la pronunciación depende de la sección de la palabra que recibe acentuación, que a su vez distingue la función de la palabra en el contexto de la oración y que requiere un conocimiento semántico del texto. Algunos exitosos (aunque perfectibles) representantes de este tipo de síntesis se han elaborado, entre el que destaca, por su relevancia a ésta tesis, el desarrollado en el posgrado de Ingeniería Eléctrica de la UNAM (Fernando del Río, 2006) sobre la plataforma *Mathlab*.

A continuación se presenta el esquema básico de los sistemas de *texto-a-voz*.

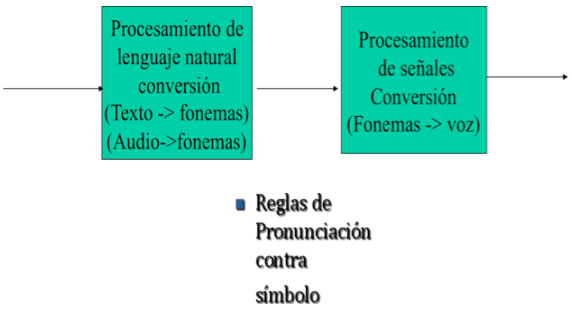


Imagen 28. Esquema básico de un sistema texto-a-voz

Los componentes básicos de un sistema *texto-a-voz* son: 1) Un mecanismo de entrada de texto crudo que se introduce a (2) una sección de análisis textual, en el que se incluye una sección de detección de la estructura del documento, normalización del texto y análisis lingüístico del que sale texto marcado que luego ingresa a (3) una sección de análisis fonético que incluye reglas de conversión de grafos a fonemas y cuya salida son fonemas marcados que constituyen la entrada a (4) una sección de análisis prosódico donde se incluyen parámetros de control del tono y la duración que permiten la salida a (5) una sección de síntesis de voz que tiene por objetivo el generar voz (Huang, Acero y Hon: 2001, p. 7). [1]

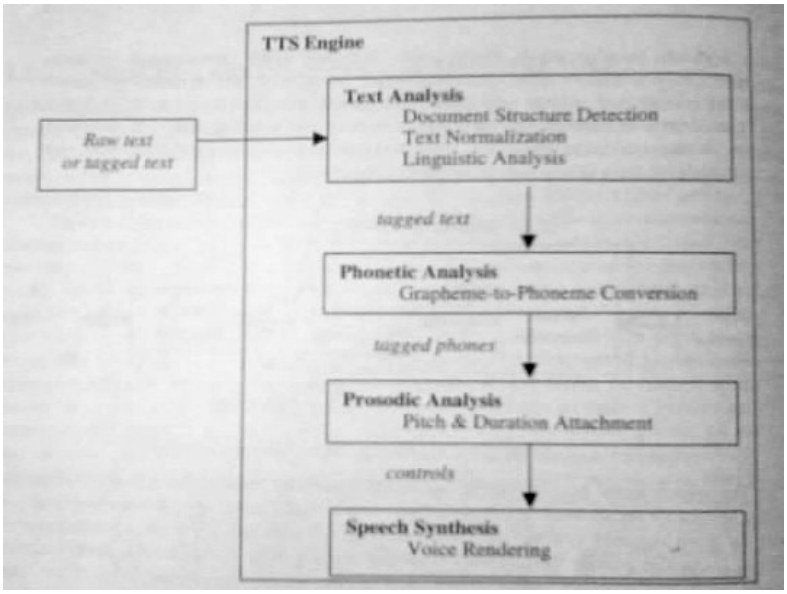


Imagen 29. Arquitectura básica de un sistema texto-a-voz

El componente de análisis de texto normaliza a éste de una forma apropiada para que pueda hacerse hablado. La función del normalizador de texto es procesar cualquier caracter no alfabético y su salida será un texto plano en forma de una secuencia de caracteres alfabéticos y signos de puntuación. Puede usarse texto marcado para asistir el análisis de texto, el fonético y prosódico. El módulo de análisis de sintaxis y prosodia, utiliza algún tipo de algoritmo de análisis para segmentar el texto de tal forma que se le pueda asignar una entonación y ritmo significativo, lo que involucra un análisis gramatical en algún nivel. En este módulo se agregan marcadores al texto que pueden indicar las sílabas acentuadas, los puntos de acentuación tónica en un patrón de entonación, así como los tipos de patrones de entonación a ser usados durante la locución. La salida del módulo (texto y marcadores) es la entrada del módulo de pronunciación, habiendo convertido el texto procesado en la secuencia fonética correspondiente, a lo cual sigue un análisis prosódico para asignar la información de tono y duración apropiados a la secuencia fonética. La mejor manera de hacer esto es almacenando cada palabra del lenguaje junto con su equivalente fonético, entonces la pronunciación puede realizarse usando un simple diccionario de búsqueda. Sin embargo, la gran cantidad de palabras constituye una restricción importante. Los sistemas

[1] Los *Apéndices I* y *II*, preséntan un comentario comprensivo del Sistema de Síntesis de Voz hablada desarrollado por Fernando del Río, donde se pueden identificar claramente los componentes básicos del sistema, ejemplo de *texto-a-voz*, de acuerdo a la clasificación aquí presentada.

prácticos tienden a usar reglas de pronunciación o reglas *letra-a-sonido* que especifican el equivalente fonético de las letras individuales o de grupos de letras tomando en cuenta el contexto. Este proceso a menudo es asistido por el uso de un pequeño diccionario de excepciones que almacena aquellas palabras y sus pronunciaciones que se usan recurrentemente, y sobre las que las reglas *letra-a-sonido* pueden fallar. La salida del módulo de pronunciación es una secuencia de fonemas junto con marcadores de sintaxis/prosodia. Finalmente el componente de síntesis de voz toma los parámetros de la secuencia fonética marcada para generar la forma de onda correspondiente.

Sistemas *texto-a-voz* han sido logrados a escala comercial, ya que desde 1991, es posible obtener a un costo razonable, sistemas que aceptan texto ordinario en inglés para convertirlo en una voz altamente inteligible, con entonaciones y ritmo razonables, pero que aún carecen de naturalidad. En algunos casos incluso se pueden tener opciones de voces, estando basados algunos en fonemas y otros en difonemas, o bien usando sintetizadores de formantes o LPC como salidas. Un ejemplo es el actual sistema de síntesis de voz de sistema operativo Mac-OSX de Macintosh.

Codificación de voz

Hoy en día, se sabe que el oído humano no requiere de la preservación de la forma de onda para mantener una percepción de alta calidad, ya que debido a las propiedades de análisis de frecuencias del sistema auditivo, no es necesario mantener los detalles de fase. Además, debido al fenómeno de enmascaramiento en el sistema auditivo, un componente espectral intenso puede tornarse imperceptible a uno más débil. Ello quiere decir que, al transportar la generación y percepción del habla al ámbito de las máquinas mucha información puede ser omitida accediéndose a un procesamiento mucho más veloz. En el dominio digital, la búsqueda por la eficiencia radica en encontrar la manera de obtener la mayor calidad de señal con la menor cantidad de bits y esto puede ser realizado por medio de dos acercamientos: 1) el primero implica codificar con un criterio de fidelidad apropiado para el oído humano, y 2) codificar en términos de parámetros que describen una clase específica de señales, como el habla. Se ha encontrado que únicamente el segundo acercamiento se puede representar con precisión. Los parámetros que han sido centrales en el diseño de "*vocoders*" eficientes son aquellos que se extraen de la producción del habla humana y resonancias del tracto vocal (formantes), frecuencia fundamental (obtenida por la vibración de los pliegues vocales), sonidos sordos (turbulencia pasando por una constricción) entre otros parámetros asociados (Flanagan, *Encyclopedia of Acoustics*: 1997, pp. 1557-1560).

Uso óptimo de cadenas: El proceso de decodificación

El objetivo del proceso de decodificación, es el escoger las unidades óptimas para formar la cadena fonética, y que se ajuste adecuadamente a la prosodia deseada. Se debe arribar a una función objetiva que logre cierta calidad sonora y que permita escoger la mejor cadena. La calidad de las unidades de cadena está típicamente dominada por las discontinuidades espectrales y tonales en las fronteras de las unidades, mismas que pueden ocurrir por:

1. Diferencias en los contextos fonéticos. Una unidad pudo haber sido obtenida de un contexto fonético diferente que el deseado para la unidad.

2. Segmentación incorrecta. Los errores de segmentación pueden causar discontinuidades espectrales aún cuando tengan el mismo contexto fonético.

3. Variabilidad acústica: Las unidades pueden tener el mismo contexto fonético y estar segmentadas de manera apropiada, pero la variabilidad de una repetición a la siguiente puede causar pequeñas discontinuidades. Una unidad, pronunciada rápidamente, es generalmente distinta a la que se pronuncia de manera lenta o normal. Además, condiciones distintas de grabación, como amplitud, tipo de micrófono o tarjeta de audio, pueden causar además discontinuidades espectrales.

4. Cambios en la prosodia. Una discontinuidad entre fronteras de las unidades también causa degradación.

La severidad de tales discontinuidades en general decrece a medida que el número de unidades incrementa (Huang, Acero y Hon: 2001, p. 810).

Reconocimiento de voz

En cuanto al reconocimiento del habla, se han establecido técnicas en las que se utilizan templetes de reconocimiento incorporados. Estos han probado su eficiencia en sistemas en los que se establece un comando desconocido el cual es medido, y su espectro es comparado con cada una de las entradas de vocabulario guardado. La comparación se realiza usando un procedimiento de alineación automática (dynamic time warp). Cierto avance se ha logrado incorporando modelos estadísticos de las secuencias sonoras que van más allá del acercamiento por templetes archivados. Estos modelos estadísticos proveen estimados estadísticos de patrones de palabras. Una de las técnicas más utilizadas actualmente, es la de modelos ocultos de Marcov donde una palabra es modelada en términos de estados, probabilidades de estados, y probabilidades de transición entre estados. Tales modelos son deducidos de secuencias espectrales observadas en el habla.

II. *La Prosodia hablada (y su Modificación)*

La prosodia es un complejo telar de efectos físicos fonéticos que se emplean para expresar actitud y atención, así como para dar a entender algunos aspectos que se dan por sentados; puede pensarse en ellos como un canal paralelo en la comunicación hablada del día a día. El contenido semántico de un mensaje hablado o escrito es conocido como *denotación*, mientras que los efectos de emoción y atención que un interlocutor expresa a otro forma parte de la *connotación* del mensaje. La prosodia tiene un papel importante en guiar a quien escucha a recuperar los mensajes básicos (*denotación*), así como un papel principal en la comprensión de la connotación, o la actitud general de quien emite el mensaje hacia el receptor y hacia el evento comunicativo en su totalidad (Huang, Acero y Hon: 2001, p. 739). Desde el punto de vista del escucha, la prosodia consiste en una percepción y recuperación sistemática de las intenciones de un interlocutor basado en:

- Pausas: que indican frases y evitan quedarse sin aire.

- Tono: frecuencia fundamental. Ritmo de repetición del ciclo de los pliegues vocales.

- Duración relativa: duración de fonemas, tiempos y ritmo.

- Intensidad: amplitud relativa/volumen.

El *tono,* es el más expresivo de los fenómenos prosódicos. A medida que se emite el habla, existe una sistemática variación de la frecuencia fundamental para expresar los sentimientos acerca de lo que se dice, o para dirigir la atención de quien escucha a pasajes especialmente importantes del mensaje hablado. Su determinación es muy importante para muchos algoritmos de procesamiento de voz, y un método comúnmente usado para estimarlo, está basado en la detección del valor más grande de la función de *autocorrelación* en la región de interés, que debe excluir el máximo absoluto de la función de correlación. Si un párrafo es hablado con un tono constante y uniforme, sin pausas, o con pausas uniformes entre palabras, entonces sonará poco natural. En algunos lenguajes, la variación del tono se encuentra constreñida por convenciones sintácticas o lexicológicas. Estos casos se entienden como un uso gramatical y lexicológico del tono. Sin embargo, cada lenguaje permite algún rango de variación de tono que puede ser explotado para propósitos emotivos y de atención. El uso de algunos efectos prosódicos para indicar emoción, estado de ánimo y atención, son universales incluso en lenguajes que usan el tono para señalizar identidad de palabras.

Es conveniente presentar el análisis de las pausas, generación de tono y duración de manera separada, ya que cuando se construyen sistemas de síntesis de voz, éstos corresponden a módulos separados. Sin embargo, debe mantenerse en mente que todas las cualidades de la prosodia mantienen una alta correlación entre ellas en el lenguaje humano. El efecto de la intensidad no es uno tan importante como los demás factores, cuando se desea sintetizar el habla, y por ello en general su discusión no es abordada, además de que para sistemas basados en la concatenación, se encuentra en general incluido en el segmento de habla.

La prosodia simbólica

La prosodia simbólica o abstracta, representa la unión entre la infinita multiplicidad de cualidades pragmáticas, semánticas y sintácticas de una expresión, y las relativamente limitadas frecuencias fundamentales, duración de fonemas, energía y cualidad de voz. La prosodia simbólica trata con:

- El rompimiento de la oración en frases prosódicas, posiblemente separadas por pausas.

- La asignación de etiquetas tales como el énfasis a diferentes sílabas o palabras dentro de cada frase prosódica.

Normalmente, la palabras en el lenguaje hablado son pronunciadas continuamente, a menos de que exista una razón lingüística específica que señalice una discontinuidad. El término *coyuntura* (*juncture*) se refiere al fraseo prosódico, es decir, al lugar donde las palabras tienen coherencia, y donde los cortes prosódicos como pausas y movimientos tonales especiales ocurren. Los efectos de coyuntura, que expresan el grado de cohesión o discontinuidad entre palabras adyacentes, son determinadas por la fisiología, la fonética, sintáctica, semántica y el pragmatismo. Los medios fonéticos primarios para señalizar una coyuntura son:

- La inserción de silencios

- Movimientos característicos en la sílaba final de la frase

- El alargamiento de ciertos fonemas en la sílaba de la frase final

- Irregularidad en la cualidad vocal como lo que se conoce en inglés como "*vocal fry*" que es la vibración inicial de los pliegues vocales.

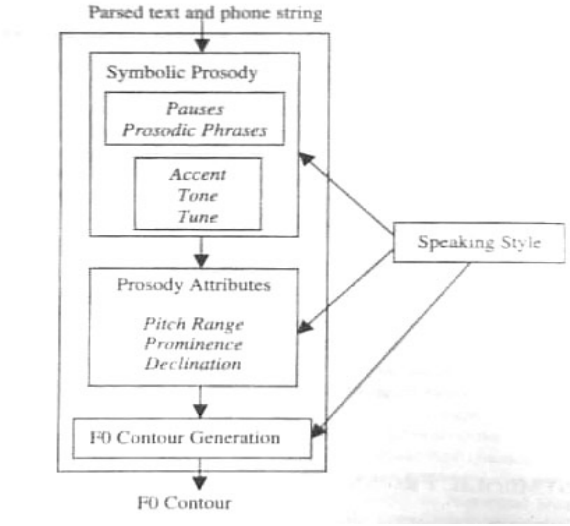


Imagen 30. Generación de tono de-compuesto en prosodia simbólica y fonética.

La estructura prosódica abstracta, especifica los elementos que incluyen: pausas y frases

prosódicas, acento, tono y tonada. Los tipos de acentos son elegidos de un inventario de tonos. La secuencia de acentuación y tonos de *coyuntura* en una estructura prosódica puede tener cierta coherencia que hace que surjan efectos de tonada que tienen alguna interpretación holística semántica. Aunque en principio, en implementaciones de sistemas de voz, el módulo de atributos prosódicos aplica a todas las variables de la prosodia, es en su mayoría utilizado prácticamente para la generación de frecuencias fundamentales.

Pausas

En una oración larga, en voz hablada, normalmente se pausa cierto número de veces. Aunque tales pausas han sido tradicionalmente tratadas en relación a la estructura sintáctica, es más apropiado pensarlas como marcadores de la estructura de la información. También pueden ser motivadas por un pobre entendimiento de las idiosincrasias del lenguaje por parte de quien lo habla, o por restricciones físicas (de respiración por ejemplo), y en el habla espontánea también pueden funcionar para expresar duda y confusión. En los sistemas, el mejor indicador de pausas es, por supuesto, la puntuación (excluyendo los símbolos asociados a abreviaciones). En lo que a símbolos de puntuación se refiere, cada uno corresponde a una frontera de frase prosódica y se les puede dar movimientos tonales especiales en su punto final (Huang, Acero y Hon: 2001, pp. 747).

En la predicción de pausas, a pesar de que se deben considerar ambas, ocurrencia y duración, la simple presencia o ausencia de un silencio (mayor que 30 ms) es la decisión más significativa, y su duración exacta es secundaria. Existen muchos sitios razonables donde pausar, pero solo algunos donde es crítico el *no* pausar. Un sistema de voz hablada debe, por tanto, evitar colocar pausas *donde sea*, y así evitar ambigüedades y malas interpretaciones (por no mencionar una absoluta falta de inteligibilidad). Casi cualquier escritura apropiada, incorpora puntuación de acuerdo a su métrica (donde sea necesario para agregar a la interpretación y no entre cada palabra).

Frases prosódicas

Mientras un punto de final de oración puede generar un cambio de tono muy acentuado hacia lo grave, una coma que termina una frase prosódica puede exhibir una pequeña elevación de tono que obedece a una sugerencia de que hay más por venir. Algunos efectos de intervalo tonal sobre una expresión también pueden estar basados en la puntuación, como el uso de paréntesis o signos de exclamación. Las coyunturas prosódicas que están claramente señalizadas por silencios, y usualmente por movimiento tonal característico, también son llamadas *frases de entonación.* Por el contrario, las coyunturas prosódicas que no están señalizadas por silencios, sino únicamente por un movimiento tonal característico, son llamadas *frases fonológicas* y pueden ser más difíciles de evaluar.

Al analizar en el habla espontánea la naturaleza y extensión de la señalización del movimiento tonal se encuentra que éstas pueden variar de un interlocutor a otro. Para discutir los tipos de coyunturas lingüísticas significativas y movimiento tonal es útil contar con un

vocabulario estándar sencillo que los consideren. Existen para el español hablado esfuerzos para establecer *Índices de Pausa y Tono*(ToBI, por sus siglas en inglés)[2], en donde se busca transcribir simbólicamente la entonación del español. Los índices de pausas especifican un listado de números que expresan la fuerza de una coyuntura prosódica, y son marcados para cualquier expresión, con la notación de índices de pausas alineados temporalmente con una representación de la fonética del habla y una pista tonal. La asociación prosódica de palabras en una expresión se muestra marcando el final de cada palabra para la fuerza subjetiva de su asociación con la siguiente palabra, en una escala de 0 (la conjunción más intensa percibida) a 4 (la más disociada).

Acentuación

En la lengua española, las palabras se pueden clasificar de acuerdo con la sílaba en la que recae la acentuación en cuatro tipos diferentes: agudas, graves, esdrújulas y sobre-esdrújulas. El acento es la señalización de una prominencia semántica por medios fonéticos. Típicamente se realiza a través de un cambio en el tono (más alto o bajo que la tendencia general) y posiblemente una extensión de duración fonética. El español es un idioma muy afortunado en el tema de la acentuación, ya que existen reglas muy claras sobre donde pueden encontrarse. Esto contrasta con el inglés, que requiere un análisis minucioso en clases y orden de introducción de conceptos, y que por si fuera poco, no presenta acentos ortográficos o escritos. Recordemos además que una ventaja del español con respecto a, por ejemplo, idiomas como el francés, es que solo existe un tipo de acento escrito (o pronunciado) mientras que el francés cuenta con al menos cuatro. En el idioma español, si una palabra lleva acento escrito, ya no es necesario hacer nada más para encontrar el acento tónico. En caso contrario, solo es necesario conocer la letra final de la palabra para conocer la sílaba que lleva el acento. Una vez que se obtiene esta letra se puede proceder a obtener la última o la penúltima sílaba de la palabra. Finalmente, y para el caso de los sistemas de voz, se escribe el acento para facilitar el procesamiento de los demás módulos.

Tono

Los *tonos* pueden ser entendidos como señalizaciones de niveles perceptuales de prominencia o movimientos de la frecuencia fundamental en sílabas. Los niveles tonales y movimientos sobre sílabas acentuadas de fronteras de frase pueden mostrar gran diversidad basados en las características de hablante, la naturaleza del evento en el que se habla, y la expresión misma. Es útil tener un inventario de tipos básicos de tonos abstractos que en principio podrían funcionar para expresar contrastes significativos lingüísticos. Los distintos hablantes podrían realizar tales tonos de forma ligeramente distinta de acuerdo con su fisiología, estado de animo, contenido en la expresión y la ocasión en la que se inserta el habla. Sin embargo la variedad entre las formas tonales y contrastes entre unos y otros se mantienen en un nivel predecible en cierto rango (Huang, Acero y Hon: 2001, p. 756).

Las categorías lingüísticas abstractas deben ser correlacionadas, o ser marcadas para

---

2 Vease: http://www.ling.ohio-state.edu/~mbeckman/Sp_ToBI/Sp_ToBI_Jul29.pdf

expresar contraste en significado. Como se ha hecho notar, las *coyunturas* están típicamente marcadas con movimientos tonales perceptibles que son independientes del acento. Las especificaciones ToBI también muestran las combinaciones de las primitivas que señalan fronteras de frase, cláusula, y expresión; tales son llamadas frases tonales. Las especificaciones ToBI, señalan además, que dado que las entonaciones de las frases están formadas de una o más frases intermedias más un tono de frontera, una frontera de entonación de frase tiene dos tonos finales. Sin embargo, las transcripciones simbólicas ToBI no son suficientes para generar un contorno completo de frecuencia fundamental (Huang, Acero y Hon: 2001, p. 756).

Tonada

Algunos contornos tonales pueden ser inmediatamente reconocibles e interpretables emocionalmente, ello independientemente del contenido en cuanto al léxico. Una pregunta válida, en cuanto al español hablado, es si la idea de *tonada* estilizada, que pueda ser descompuesta en tonos, puede ser aplicada a la entonación del habla ordinaria. Quizá es en ésta pregunta más que en cualquier otra, donde convergen los estudios en cuanto a síntesis de voz cantada con aquellos de voz hablada. En cuanto a esta última, se puede decir que el uso de las marcas de acentos tonales ToBI son usados idealmente como elementos primitivos en descripciones prosódicas holísticas, análogas al papel de fonemas en las palabras. Un diccionario de contornos de significado, descritos de manera abstracta por símbolos para permitir una realización de variabilidad fonética, constituiría una teoría del significado de la entonación. Idealmente tales significados se desprenderían composicionalmente de sus constituyentes de acento tonal y tonos de frontera. Los métodos de estilización de contorno los describen de manera holística. En tales métodos, se asignan índices en base al tipo de expresión (usualmente basados en una tipología sintáctica) como por ejemplo una pregunta, una orden, etc. En general, aunque la representación holística de contornos puede ser una buena aproximación, la categorización mediante descripciones sintácticas basadas estrictamente en puntuación son muy criticables, ya que no parece haber una correspondencia uno a uno entre actos del habla y tipos sintácticos. La prosodia en el habla espontánea es mediada a través del contexto del uso pragmático y el manejo de una expresión. Por ello, la descripción completa de un evento del habla es lo que mejor garantiza la calidad, más que ciertas inferencias sobre el contenido del texto. Es por ello que casi se puede asegurar que el siguiente paso en sistemas de voz hablada, es el de *concepto-a-habla* (Huang, Acero y Hon: 2001, p. 758).

Para sistemas de *texto-a-voz* comerciales, que deben inferir estructura de texto crudo, existen algunos patrones fragmentarios de tono característicos que pueden ser utilizados como tonadas y aplicados a segmentos especiales de expresiones, que incluyen:

- entonación en una lista

y

- citas

Duración

El tono y la duración no son enteramente independientes, y muchos de los factores semánticos de alto nivel que determinan los contornos de tono también pueden incluir los efectos de duración. De acuerdo con Huang *et al*, la relación entre tono y duración es una para la cual apenas algunas exploraciones iniciales han sido desarrolladas. Sin embargo muchos sistemas tratan la duración y el tono de manera independiente debido a consideraciones prácticas (Huang, Acero y Hon: 2001, p. 762).

Numerosos factores, incluyendo semántica y condiciones pragmáticas, pueden influir en la duración de fonemas. Algunos factores que son típicamente ignorados incluyen (Huang, Acero y Hon: 2001, p. 762):

- El ritmo del habla relativo a la intensión, estado de ánimo, y emoción de quien habla.

- El uso de ritmo y duración para posiblemente señalizar la documentación de la estructura sobre el nivel de frase u oración, como por ejemplo un párrafo.

- La falta de una definición práctica y coherente de fonos, de tal manera que las fronteras puedan ser fácilmente localizada para su medición.

Generación de Tono

Como se dijo anteriormente, el tono, o la frecuencia fundamental, es probablemente la más característica de todas las dimensiones prosódicas. La calidad de un módulo destinado a la prosodia está dominada por la calidad de sus componentes diseñados para la generación de tono. Debido a que la generación de contornos tonales es un problema extraordinariamente complicado, la generación de tonos a menudo se divide en dos niveles; el primero computa la prosodia simbólica, y el segundo de hecho genera los contornos tonales partiendo de tal prosodia simbólica. Ésta división resulta un tanto arbitraria, debido a que una gran cantidad de fenómenos prosódicos importantes no caen limpiamente en uno u otro nivel, sino que parece involucrar aspectos de ambos (Huang, Acero y Hon: 2001, p. 764). A menudo es conveniente agregar otros atributos del contorno tonal, antes de su generación. El contorno tonal está caracterizado no solo por su prosodia simbólica, sino también por atributos como el intervalo tonal, gradiente de prominencia, declinación y microprosodia. Tales atributos son conocidos en el campo de la prosodia fonética (término que es una analogía a la fonología y la fonémica). A continuación se enlistan las características de tales atributos:

– *Intervalo tonal*: se refiere a los límites superiores e inferiores dentro de los cuales todos los acentos y tonos de frontera deben ser realizados. Es típicamente especificado en Hertz. Pueden ser considerados en términos de límites estables, específicos en cuanto a quien habla, así como en términos de una expresión o pasaje. Para un sistema *texto-a-voz,* cada voz típicamente tiene un intervalo tonal característico que representa algún promedio de los extremos tonales en expresiones de prueba. Este intervalo, específico a una voz, puede usarse como punto de partida para el carácter de la voz, y tales límites pueden ser

cambiados por una aplicación.

La variación del tono, que se correlaciona con la emoción u otros aspectos del evento hablado, a veces se denomina *para-lingüístico*. El uso lingüístico y *para-lingüístico* del intervalo tonal incluye aspectos tanto de prosodia simbólica como fonética. Debido a su naturaleza cuantitativa, es ciertamente una propiedad fonética del contorno de la frecuencia fundamental de una expresión. Además, parece ser que la mayoría de los contrastes lingüísticos que involucran acentos tonales, tonos de frontera, etc. pueden ser realizados en cualquier intervalo tonal. Los valores, pueden ser estimados del habla natural, para propósitos de investigación, calculando el promedio y varianza de la frecuencia fundamental sobre una expresión, o simplemente adoptando la medida máxima y mínima (Huang, Acero y Hon: 2001, p. 764).

Aunque el intervalo tonal es una propiedad fonética, puede ser sistemáticamente manipulada para expresar estado de ánimo y sentimientos. El intervalo tonal interactúa con todos los atributos prosódicos que se han examinado, y ciertos valores del intervalo tonal pueden ser característicos de estilos o expresiones particulares. El intervalo tonal no puede ser considerado un atributo arbitrario o fisiológico, ya que es manipulado directamente para efectos de la comunicación. En la investigación prosódica, ha sido difícil el distinguir el uso emotivo del tono, de aquel estrictamente lingüístico. La variación en intervalo tonal parece incluir expresión emotiva, lingüística y fonética. El intervalo tonal lingüístico puede ser reducido o ampliado, y la zona de variación tonal puede ser colocada en cualquier lugar del intervalo físico tonal de una persona. Un sistema práctico *texto-a-voz* debe permanecer dentro de los límites del intervalo. Al mismo tiempo deberá buscar la maximización de la explotación de tales límites.

- *Gradiente de Prominencia*: Se refiere a la fuerza relativa de un acento con respecto a sus vecinos y los valores de intervalo tonal. La altura relativa de los acentos puede alterar de manera fundamental el contenido de información de un lenguaje hablado determinando el foco de la atención, contraste y énfasis. Es deseable que tal contenido lingüístico estuviera determinado por la presencia o ausencia de acentos simbólicos, sin embargo no hay garantía de existencia de un nivel mínimo de prominencia para la detección perceptual de acentos. El darse cuenta de la prominencia de un acento es sensible al contexto, y dependiente de los valores de intervalo tonal establecidos. Hasta ahora, parece que la prominencia relativa está relacionada al estatus de información que las palabras acentuadas conllevan, y por lo tanto es de carácter lingüístico, y al no existir ninguna teoría de categorías de prominencias, no hay posibilidades de abstracción (Huang, Acero y Hon: 2001, p. 766)

- *Declinación*: es la tendencia (a largo plazo) de acentuar alturas en una oración declarativa neutral semánticamente, en un estilo típico de lectura.

- *Microprosodia*: Se refiere a aspectos del contorno tonal que son fonéticos (fuera de cualquier ambigüedad) y que pueden involucrar alguna interacción con los sonidos que llevan la sensación de velocidad.

Modificación Prosódica del Habla: Métodos PSOLA

Un problema de la concatenación de segmentos, es que no es posible generalizarlos adecuadamente en contextos que no están incluidos en el proceso de entrenamiento, en parte porque la variabilidad prosódica es muy amplia. Por ello, existen técnicas que permiten la modificación de la prosodia de una unidad para igualarla a la prosodia que se desea. Tales técnicas, aunque en cierta medida degradan la calidad de la voz sintetizada, presentan beneficios de flexibilidad mucho mayores que la distorsión introducida por su uso.

El objetivo de la modificación prosódica, es el cambio de amplitud, duración y tono de un segmento de voz hablado. La modificación de la amplitud puede ser fácilmente logrado por multiplicación directa, sin embargo la duración y el tono presentan mayor dificultad (Huang, Acero y Hon: 2001, p. 818). Los métodos PSOLA (de suma y superposición de tono sincronizado), originalmente desarrollados en Telecom (Francia), no constituyen en sí un método de síntesis, pero permiten que muestras pregrabadas de voz hablada sean concatenadas de manera suave, y provee buenos controles de tono y duración, los cuales han sido aplicados a sistemas comerciales como ProVerbe y HADFIX.

Existen varias versiones del algoritmo PSOLA, y todas funcionan de manera similar. El TD-PSOLA constituye el algoritmo en el dominio del tiempo, y es el más usado debido a su eficiencia computacional. El algoritmo básico consiste en tres pasos: 1) El paso de análisis, donde la señal original es dividida por primera vez en señales de análisis de tiempo corto separadas, pero que a menudo se superponen. 2) La modificación de cada señal de análisis para generar la señal de síntesis, y 3) el paso de síntesis, en el cual tales segmentos son recombinados a través de la sobreposición y suma (Lemmety: 1999, pp. 34-35). Las señales de tiempo corto son obtenidas de la forma de onda digital, multiplicando la señal por una secuencia de ventanas de análisis de tono sincronizado. Las ventanas son usualmente del tipo Hanning, usualmente centradas alrededor de instantes sucesivos llamados marcas de periodo, o tono. Tales marcas son colocadas a un ritmo sincronizado (en lo que respecta al tono) sobre las secciones *sonoras* de la voz y a un ritmo constante en las secciones *sordas*. La longitud de la ventana es proporcional al periodo tonal local y el factor de ventana va usualmente de 2 a 4. Los marcadores de tono son determinados ya sea manualmente por inspección de la señal de voz, o automáticamente por algún método de estimación. La recombinación de segmentos en el paso de síntesis, se realiza después de definir una nueva secuencia de marcas de tono (Lemmety: 1999, p. 35).

La manipulación de la frecuencia fundamental se logra mediante el cambio de los intervalos de tiempo entre los marcadores de periodo, o tono. La modificación de la duración se logra ya sea a través de la repetición o eliminación de segmentos de voz. En principio, la modificación de frecuencia fundamental también implica modificación de duración, sin embargo se puede solucionar tal problema.

Otras variaciones del algoritmo son el FD-PSOLA (*dominio frecuencial*), y el LP-PSOLA (*predicción lineal*), que son aproximaciones teóricas más apropiadas para la modificación del tono, ya que presentan control independiente sobre la envolvente espectral de la señal de síntesis. Uno de los problemas del uso de los métodos PSOLA, son que el tono sólo puede ser determinado para sonidos *sonoros,* y que aplicado a sonidos *sordos* puede generar ruido tonal

(Lemmety: 1999, p. 36). Por ello, durante la re-síntesis, las muestras de sonidos *sordos* se concatenan de forma aleatoria para reducir la posibilidad de correlaciones accidentales, que podrían resultar en reverberación, efecto de *flanger*, o tono. El método realiza de manera adecuada pequeñas transformaciones, sin embargo para cambios más grandes, la calidad de la voz sufre. Guarda, a su vez, algunas similitudes con los métodos de síntesis por lectura de *tabla de onda* (Siivola: 2002, pp. 13-14).

Sincronización y Suma Superpuesta (SOLA-Synchronous Overlap and Add)

La técnica OLA (overlap-and-ad) es una en la que se modifica la escala temporal de la voz, mediante ventanas de análisis y ventanas de síntesis. Dada una ventana de Hanning de longitud *2N,* y un factor de compresión de *f*, las ventanas de análisis están espaciadas *fN*. Cada ventana de análisis multiplica la señal de análisis, y en el momento de la síntesis éstas son sobrepuestas y sumadas. Las ventanas de síntesis están espaciadas cada *N* muestras. El uso de una ventana tipo Hanning, permite la perfecta reconstrucción cuando *f* es igual a 1.

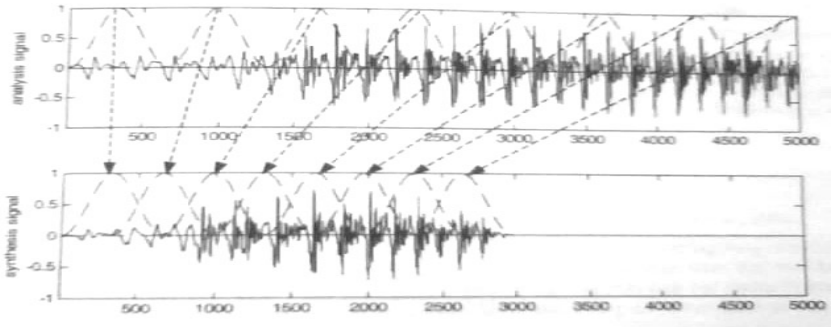


Imagen 31. Método para la compresión del tiempo OLA, superposición y suma

En la *Imagen 31*, se puede observar que, tras la aplicación del método, pareciera que algo de la apariencia original de la señal se pierde, además de que se presentan algunos periodos de tono irregulares. Para resolver tal problema, la técnica de sincronización y suma sobrepuesta (SOLA), permite el posicionamiento flexible de la ventana de análisis a través de la búsqueda de la ventana de análisis *i* alrededor de *fNi*, de tal manera que la región sobrepuesta presente una correlación máxima. El algoritmo SOLA produce una compresión temporal de alta calidad (Huang, Acero y Hon: 2001, p. 819).

Típicamente, los algoritmos de compresión operan a un ritmo uniforme, sin embargo, algunos de ellos han sido usados a ritmos no uniformes para tomar en consideración la percepción humana, de manera tal que las transiciones rápidas son comprimidas levemente, los sonidos estables son comprimidos un poco más, y las pausas son las más comprimidas. Mientras un factor de compresión de 2.5 se logra con la implementación de compresión uniforme (sin la degradación de la inteligibilidad), se puede lograr un factor de 4, si se aplica una compresión no uniforme (Huang, Acero y Hon: 2001, p. 819)

Sincronización y Suma Superpuesta de Tono (PSOLA-Pitch Synchronous Overlap and Add)

Los métodos OLA y SOLA, a pesar de realizar modificación a la duración, no pueden realizar modificación de tono, e incluso operan sin conocimiento del tono de la señal. El método más utilizado para realizar modificaciones de tono es el PSOLA (Sincronización y Suma Superpuesta de Tono). Tal método requiere el conocimiento del tono de la señal. El proceso, que resulta indistpensable para la comprensión del trabajo presente, se ilustra a continuación.

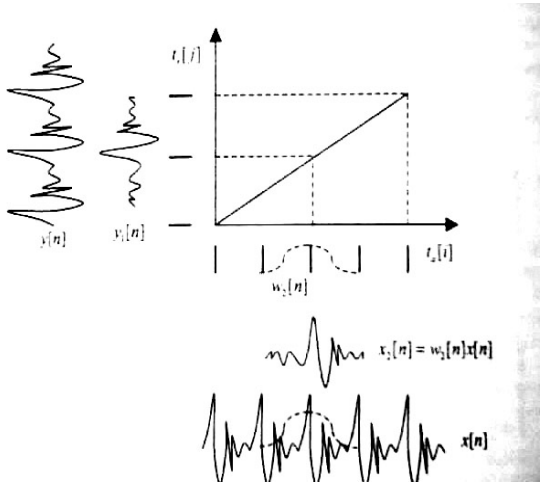


Imagen 32. Mapeo entre 5 periodos o ciclos de tono a 3 de síntesis.

Asumiendo que la señal de entrada $x[n]$ es *sonora*, por lo que puede ser expresada como función de periodos o ciclos de tono $x_i[n]$

$$x[n] = \sum_{i=-\infty}^{\infty} x_i[n - t_a[i]] \tag{83}$$

donde $t_a[i]$ son periodos particulares de tiempo de la señal (donde el tono se mantiene cuasi-constante), de manera tal que la diferencia entre periodos adyacentes $P_a[i]=t_a[i]-t_a[i-1]$ es el periodo del tono en el tiempo $t_a[i]$ en muestras. El ciclo del tono es una versión ventaneada de la entrada

$$x[n] = w_i[n]x[n] \tag{84}$$

que requiere que las ventanas $w_i[n]$ cumplan con la condición:

$$\sum_{i=-\infty}^{\infty} w_i[n - t_a[i]] = 1 \tag{85}$$

condición que puede lograrse con una ventana de Hanning, o una ventana trapezoidal que abarca dos periodos de tono.

El objetivo es el de sintetizar una señal $y[n]$, que tiene las mismas características que $x[n]$ pero con una duración y/o tono diferente. Para lograrlo, se reemplazan la secuencia de periodos temporales de análisis $t_a[i]$ con los periodos sintetizados $t_a[j]$, y los ciclos de tono de análisis $x_i[n]$ con los ciclos de tono sintetizados $y_j[n]$:

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_s[j]] \tag{86}$$

Las periodos temporales de síntesis son computados de forma que cumplan con un contorno de duración y tono específico, como se muestra en la *Imagen 32*. Ello es equivalente a un tren de impulsos espaciados de manera variable que conducen a un filtro variable en el tiempo $x_t[n]$ que es conocido para $t=t_a[i]$, como se muestra en la *Imagen 33*. El ciclo de tono sintetizado $y_j[n]$ es obtenido a través del mapeo del ciclo de tono de análisis más próximo correspondiente $x_i[n]$ (Huang, Acero y Hon: 2001, p. 821).

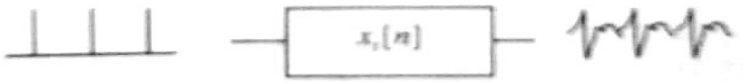


Imagen 33.Técnica PSOLA como un tren de pulsos variables alimentados a un filtro variable

El término superposición-y-suma, se deriva del hecho de que se utilizan ventanas sobrepuestas que se suman. El aspecto de la *sincronización* resulta del hecho de que las ventanas están separadas por un periodo de tono y son de dos periodos de tono de longitud. En el caso de la voz *sorda*, un conjunto de marcas temporales que están uniformemente espaciados funcionan bien en la práctica, siempre y cuando el espaciamiento sea menor que 10 ms. Si los segmentos deben ser estirados de manera tal que las formas de ondas características son repetidas, una periodicidad artificial aparecería. Para evitar esto, la forma de onda característica que debe ser repetida es volteada en el tiempo (Huang, Acero y Hon: 2001, p. 821), o como se dijo anteriormente, se colocan las señales de tiempo corto de forma aleatoria para evitar correlación.

Aunque el procedimiento es muy sencillo, conduce a una modificación prosódica de alta calidad, siempre y cuando la decisión de voz *sonora/sorda* sea la apropiada y la secuencia de marcas temporales sea precisa. Para realizar la modificación prosódica vía PSOLA, se requiere mantener la forma de onda del segmento de voz y su conjunto correspondiente de periodos temporales, o marcas en el tiempo (Huang, Acero y Hon: 2001, p. 822).

Comportamiento espectral del PSOLA

Si se considera el caso en que la señal de voz $x[n]$ que es exactamente periódica, con periodo T_0 y puede ser creada pasando un tren de impulsos a través de un filtro con respuesta al impulso $s[n]$:

$$x[n] = s[n] * \sum_{i=-\infty}^{\infty} \delta[n - iT_0] = \sum_{i=-\infty}^{\infty} s[n - iT_0] \tag{87}$$

Si la respuesta al impulso $s[n]$ es conocida, entonces se puede cambiar el tono modificando T_0. El problema real radica en estimarlo de $x[n]$. Asumiendo que se quiere construir un estimado multiplicando $x[n]$ por una ventana $w[n]$:

$$\tilde{s}[n] = w[n]x[n] \tag{88}$$

La transformada de Fourier de $x[n]$ es la ecuación (87), y está dada por:

$$X(\omega) = \frac{2\pi}{T_0} S(\omega) \sum_{k=0}^{T_0-1} \delta(\omega - k\omega_0) = \frac{2\pi}{T_0} \sum_{k=0}^{T_0-1} S(k\omega_0)\delta(\omega - k\omega_0) \tag{89}$$

donde $\omega_0=2\pi/T_0$. La transformada de Fourier del estimado puede obtenerse de las ecuaciones (88) y (89)

$$\tilde{S}(\omega) = \frac{1}{2\pi} W(\omega) * X(\omega) = \sum_{k=0}^{T_0-1} S(k\omega_0) \frac{W(\omega - k\omega_0)}{T_0} \tag{90}$$

En caso de que la ventana $w[n]$ esté sincronizada en tono, una ventana rectangular con longitud T_0 o una ventana de Hanning con longitud $2T_0$ por ejemplo, entonces el estimado es exacto en los armónicos, debido a que los términos de fuga son cero en los armónicos. Entre armónicos, la transformada de Fourier del estimado es una interpolación usando la función de transferencia de la ventana W(ω). Si se usa una ventana rectangular, los valores de $S(\omega)$ entre $S(k\omega_0)$ y $S((k+1)\omega_0)$ no son determinados sólo mediante esos dos armónicos, debido a que la fuga de los otros armónicos no es despreciable. El uso de una ventana de Hanning atenúa drásticamente tal fuga, de manera que la envolvente espectral es mejor. Lo que PSOLA hace es obtener un estimado de la envolvente espectral por medio de la utilización de una ventana sincronizada en tono (Huang, Acero y Hon: 2001, p. 822).

Debido a que es matemáticamente imposible recuperar $S(\omega)$ para una función periódica, es razonable el rellenar los valores restantes por medio de interpolación con los lóbulos principales de la transformación de la ventana. Tal aproximación funciona particularmente bien si los armónicos forman un muestreo denso de la envolvente espectral, que es el caso de las voces masculinas. En el caso de la voz de mujer, para el cual los armónicos se encuentran más espaciados, la envolvente espectral estimada por medio de la interpolación a través de armónicos, pudiera ser muy distinta a la envolvente real (Huang, Acero y Hon: 2001, p. 823).

Cálculo de los periodos temporales de síntesis

En la práctica, lo que se desea es generar un conjunto de periodos de síntesis $t_s[j]$, dado un tono objetivo $P_s(t)$. Si el periodo tonal que se desea $P_s(t)=P$ es constante, entonces las marcas están dadas por $t_s[j]=jP$. En general el periodo tonal deseado $P_s(t)$ es una función del tiempo. De manera intuitiva se pueden obtener $t_s[j+1]$ en términos de las marcas previas $t_s[j]$ y el periodo tonal en ese momento:

$$t_s[j+1] - t_s[j] = P_s(t_s[j]) \tag{91}$$

Ésta es solo una aproximación, pero funciona bien si $P_s(t)$ cambia lentamente a través del tiempo.

Una ecuación exacta puede ayudar a entender las modificaciones en la escala de tiempo y tono. Las marcas de periodo $t_s[j+1]$ pueden ser calculadas de forma que la distancia entre marcas de periodo adyacentes $t_s[j+1]$-$t_s[j]$ sea igual al promedio del periodo tonal en la región $t_s[j] \le t < t_s[j+1]$ entre ellos, como lo muestra la imagen.

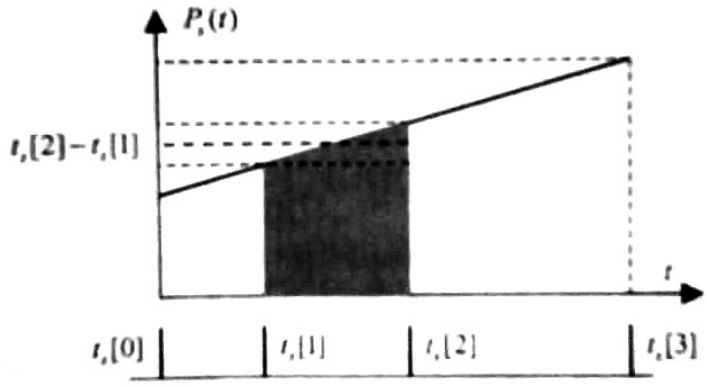


Imagen 34. El periodo tonal deseado es una función lineálmente creciente del tiempo

Ello puede lograrse con la expresión:

$$t_s[j+1] - t_s[j] = \frac{1}{t_s[j+1] - t_s[j]} \int_{t_s[j]}^{t_s[j+1]} P_s(t)dt \tag{92}$$

Es útil considerar el caso en el que $P_s(t)$ es lineal con t en el intervalo:

$$P_s(t) = P_s\left(t_s[j]\right) + b(t - t_s[j]) \tag{93}$$

de tal forma que la integral en la ecuación (92) está dada por

$$\int_{t_s[j]}^{t_s[j+1]} P_s(t)dt = \delta_j \left[P(t_s[i]) + b\frac{\delta_j}{2} \right] \tag{94}$$

donde se ha definido a δ_j como

$$\delta_j = t_s[j+1] - t_s[j] \tag{95}$$

Insertando las ecuaciones (94) y (92), obtenemos

$$\delta_j = P(t_s[i]) + b\frac{\delta_j}{2} \tag{96}$$

lo cual, usando (95), brinda una solución para el periodo $t_s[j+1]$ como

$$t_s[j+1] - t_s[j] = \delta_j = \frac{P_s(t_s[j])}{(1 - b/2)} \tag{97}$$

de la marca anterior $t_s[j]$, el tono objetivo en tal marca $P_s(t_s[j])$, y la pendiente b. Vemos que la ecuación (91) es una buena aproximación a la ecuación (97) si la pendiente b es pequeña. Evaluando la ecuación (93) para $t_s[j+1]$ resulta en una expresión para $P_s(t_s[j+1])$

$$P_s(t_s[j+1]) = P_s(t_s[j]) + b\left(t_s[j+1] - t_s[j]\right) \tag{98}$$

Las ecuaciones (97) y (98) pueden ser usadas de manera iterada. Es importante notar que la ecuación (97) requiere $b<2$ para que se obtengan resultados significativos. En la práctica este siempre es el caso.

Cuando se sintetizan excitaciones para síntesis de voz, es conveniente especificar el periodo tonal de síntesis $P_s(t)$ como una función lineal del tiempo en pedazos. En éste caso, la ecuación (97) es aún válida, siempre y cuando $t_s[j+1]$ caiga dentro del segmento lineal. De otro modo, la integral en la ecuación (94) tiene dos componentes, y es necesario resolver una ecuación de segundo orden para obtener $t_s[j+1]$.

Cálculos de las Marcas de periodo para la modificación de la escala tonal

A veces, en lugar de generar una secuencia de periodos dados por una función $P_s(t)$, se quiere modificar la secuencia de una señal de análisis con segmentos $t_a[i]$ cambiando su tono, mientras se mantiene la duración intacta. A ello se le conoce como modificación de la escala tonal. Para obtener los periodos correspondientes de síntesis, se asume que el periodo tonal $P_a(t)$ de la forma de onda de análisis al tiempo t es constante y es igual a la diferencia entre periodos temporales.

$$P_a(t) = t_a[i+1] - t_a[i] \tag{99}$$

Ello se ejemplifica en la *Imagen 35:*

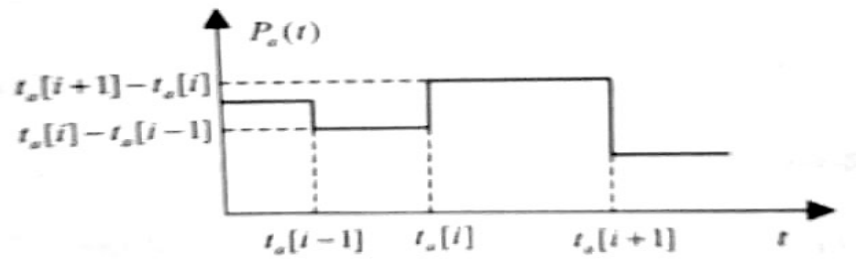


Imagen 35. Periodo tonal de la forma de onda de análisis como función del tiempo. Es una función del tiempo constante por partes.

El periodo tonal de la forma de onda $P_s(t)$ en el mismo tiempo t, cae entre las marcas j y $j+1$:

$$t_s[j] \le t < t_s[j+1] \tag{100}$$

con $t_s[j]$ siendo el instante de tiempo de la marca *j´esima* de la forma de onda sintetizada. Se define una relación entre el análisis y la síntesis de periodos tonales:

$$P_s(t) = \beta(t) P_a(t) \tag{101}$$

donde $\beta(t)$ refleja el factor de modificación de escala tonal, que, en general, es una función del tiempo.

Según la derivación de la sección anterior, se calculan los periodos de síntesis $t_s[j+1]$ de forma que:

$$t_s[j+1] - t_s[j] = \frac{1}{t_s[j+1] - t_s[j]} \int_{t_s[j]}^{t_s[j+1]} \beta(t) P_a(t) \tag{102}$$

lo que refleja el hecho de que el periodo tonal de síntesis en el tiempo t es un periodo tonal promedio de la forma de onda de análisis por un factor de modificación de escala tonal. Debido a que $\beta(t)P(t)$ es lineal por pedazos, se puede usar el resultado de la sección anterior para resolver $t_s[j+1]$. En general, debe ser resuelta de manera recursiva, lo cuál resulta en una ecuación de segundo orden si $\beta(t)$ es una constante, o una función lineal de t.

Cálculos de periodos para la modificación de la escala de tiempos

La modificación de la escala temporal requiere el cambio de duración de un segmento de voz, manteniendo el tono intacto. Para ello se define un mapeo o una función de transformación temporal, $t_s=D(t_a)$, entre la señal original y la señal modificada. Es útil definir el ritmo de modificación de duración $\alpha(t)$ de la cuál tal función puede ser derivada:

$$D(t) = \int_0^t \alpha(\tau) d\tau \tag{103}$$

Si el ritmo de la modificación de la duración es constante, entonces el mapeo es lineal. Si la constante es mayor que uno, lo que se hace es lentamente alentar el habla, mientras que si es menor a uno, entonces se está acelerando. Si consideramos el tiempo *t* en los periodos *i* e *i+1* de forma que $t_a[i] \leq t < t_a[i+1]$

$$D(t_a[0]) = 0$$
$$D(t) = D(t_a[i]) + \alpha(t - t_a[i]) \quad (104)$$

De forma que la relación entre los periodos del análisis y la síntesis están dados por

$$P_s(D(t)) = P_a(t) \quad (105)$$

Para resolver esto es útil definir una cadena de instantes virtuales $t'_a[j]$ en la señal de análisis relacionados a los instantes de tiempo de síntesis por

$$t_s[j] = D(t'_a[j]) = \alpha t'_a[j] \quad (106)$$

como se muestra en la *Imagen 36*.

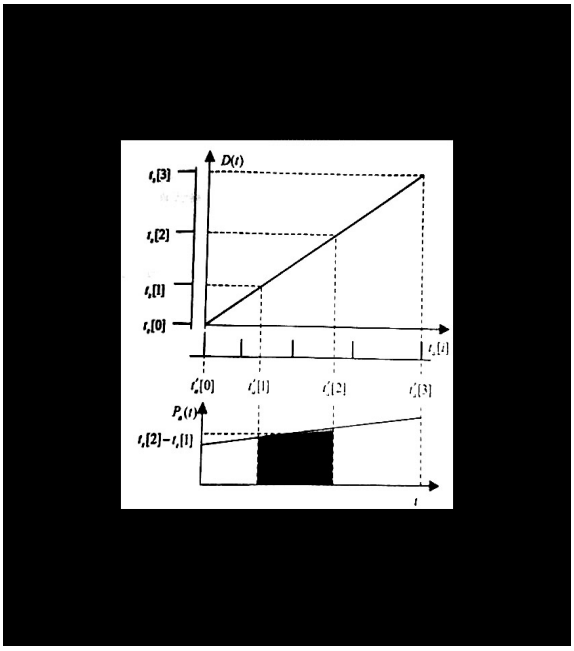


Imagen 36. Escala de tiempo de la modificación de voz

Ahora se busca determinar $t_s[j+1]$ de manera que $t_s[j+1]$-$t_s[j]$ sea igual al promedio del periodo tonal en la señal original de tiempo entre $t_a'[j]$ y $t_a'[j+1]$:

$$t_s[j+1] - t_s[j] = \frac{1}{t_a[j+1] - t_a[j]} \int_{t_s[j]}^{t_s[j+1]} P_a(t)dt \qquad (107)$$

que resulta en, usando (106):

$$t_s[j+1] - t_s[j] = \frac{\alpha}{t_a[j+1] - t_a[j]} \int_{t_s[j]/\alpha}^{t_s[j+1]/\alpha} P_a(t)dt \qquad (108)$$

lo cuál una vez más resulta ser una ecuación de segundo grado si $P_a(t)$ es constante por pedazos o lineal en t.

En el caso de que se desee hacer transformaciones tanto de tono como de tiempo, se obtiene, combinando las ecuaciones (102) y (108):

$$t_s[j+1] - t_s[j] = \frac{\alpha}{t_s[j+1] - t_s[j]} \int_{t_s[j]/\alpha}^{t_s[j+1]/\alpha} P_a(t)dt \qquad (109)$$

Detección de periodos

Utilizando el método PSOLA, los periodos de análisis $t_a[i]$ deben ser conocidos. En la práctica, estos deben ser estimados de la señal de voz. Pueden surgir errores, como sonido áspero, ruidoso, etc. si el periodo tonal no es estimado correctamente. Esta resulta la parte más sensible de la utilización del método PSOLA. Resulta sencillo el conocer el tono a partir de los periodos $t_a[i]$, ya que $P(t)=t_a[i+1]-t_a[t]$ para $t_a[i] < t < t_a[i+1]$. Sin embargo, a partir del tono, la localización de tales periodos no está determinada unívocamente, debido a que el origen en el tiempo no está especificado.

Algunos errores comunes, que incluyen doblarlo o dividirlo en dos, o errores en decisiones de voz *sonora* o *sorda*, resultan en una voz sintetizada áspera. Aunque el marcaje a mano de las marcas de periodo o tono, puede resultar muy eficiente, toma mucho tiempo y no está exento de errores.

Uno de los métodos más exitosos incluye el uso de laringógrafo, sin embargo, para efectos de este trabajo, los periodos fueron calculados de manera automática mediante un algoritmo descrito por Vladimir Goncharoff y Patrick Gries, de la Universidad de Illinois en Chicago. En éste, Goncharoff et. al desarrollan un algoritmo que marca automáticamente la locación de pulsos de tono en habla continua, logrando marcas de periodo espaciadas regularmente, incluso durante silencios o secciones de voz *sorda*, por lo que decisiones entre voz *sonora/sorda* no son necesarias, y las marcas se asignan en la señal entera. En su trabajo, bajo la consideración de que para la correcta aplicación del método PSOLA se requiere el conocimiento tanto del tono como de la fase de la señal, definen la localidad deseable para colocar las marcas de tono de la siguiente manera:

1. Durante los segmentos de voz *sonora* (quasi-periódica) las marcas de tono se colocan en

uno de los puntos de mayor amplitud (positivo o negativo) de cada ciclo tonal, resultando una secuencia de marcas que varían sutilmente.

2. Durante los segmentos que no tienen una forma de onda periódica, las marcas de periodo, ó "tono", se colocan en lugares en donde una secuencia sutilmente variable de marcas se desarrollen hasta llegar a un segmento de voz *sonora*.

Como puede distinguirse, tal método no hace ningún intento por discriminar entre voz *sonora* y *sorda*. El algoritmo de tono consta de dos bloques independientes, uno de estimación de periodo tonal, y otro de estimación de fase. En el primero, el periodo tonal es calculado de manera un poco burda, sin reparar en la posición de la marca de tono durante los segmentos de voz sonora, e interpolada como una función continua sobre segmentos sordos. En el segundo, éste estimado burdo se refina y las marcas de tono son colocadas en los picos de amplitud de la forma de onda para completar el procedimiento. Para el logro de estos bloques, una rutina dinámica para encontrar rutas es implementada, que a continuación se describe.

Programación dinámica de estimación de ruta

Tal programación es utilizada para calcular la ruta óptima de la primera a la última columna de una matriz rectangular, bajo condiciones de restricción de pendiente máximas. Dada una matriz **A,** *N x M,* se define un vector **p** el cual especifica el índice de la columna para cada columna de **A**. La suma de los valores de los coeficientes sobre la ruta a través de la matriz es:

$$E = \sum_{m=1}^{M} A(p(m), m) \tag{110}$$

Lo que la programación realiza es el cálculo de **p** para el cual se logra la maximización de *E*, con la condición de que:

$$|p(m) - p(m-1)| \leq S_{max} \tag{111}$$

para toda *m* entre 2 y M, y donde S_{max} es la pendiente máxima permitida de la ruta.

Estimación del periodo tonal: algoritmo de Goncharov et al.

El primer paso es la estimación de la energía de tiempo corto, logrado a través del filtrado paso-bajos del cuadrado de la señal de voz. Con ello, se logra producir un contorno de energía que tiene aproximadamente un pico por periodo tonal durante segmentos de voz sonora (des-enfatizando los componentes frecuenciales sobre el intervalo típico de frecuencia fundamental para todas las voces). Para ello se convolucionan las muestras de la señal de voz al cuadrado con una ventana de 12.8 ms de suavización, misma que es encontrada como producto de la convolución de una ventana de Hanning de 6.4 ms de ancho consigo misma. El contorno de energía resultante tiene picos de alta amplitud que están espaciados regularmente durante los

segmentos de voz sonora, y picos de baja amplitud espaciados irregularmente durante los segmentos de voz sorda. Una manera de calcular el periodo tonal de tal contorno de energía es medir el espacio entre picos. Sin embargo, tal método no es confiable ya que ocasionalmente aparecen más de un pico dentro de un periodo tonal y, como ya se dijo, típicamente los picos de energía se encuentran espaciados irregularmente durante los segmentos de voz sorda. Es por ello que para obtener una estimación acertada se aplica el algoritmo de programación dinámica.

Siendo e_m y t_m el valor de la energía y el índice del muestreo del pico m-ésimo en el contorno de energía, respectivamente, se inicializa la matriz **A** para que tenga R renglones y C columnas, y donde R = p_{max}, es decir el valor máximo permitido para el periodo tonal, y C es el número de cuadros de análisis en la señal de voz (aproximadamente de 40 ms por cuadro). Si f_m es el índice de los cuadros cuyos tiempos centrales son los más cercanos al pico m-ésimo en el contorno de energía, **A** es inicializada en cero, y:

1. Si $p_{min} \leq (t_m - t_{m-1}) \leq p_{max}$, entonces e_{m-1} se guarda en $\mathbf{A}(t_m\text{-}t_{m-1}, f_m)$;

2. si $p_{min} \leq (t_{m-1} - t_m) \leq p_{max}$, entonces e_{m+1} se guarda en $\mathbf{A}(t_{m*1} - t_m, f_m)$.

Ello, carga dos estimados tonales por pico de energía en una columna de la matriz **A**: ambas distancias entre cualquier pico y sus vecinos son guardadas en la columna que corresponde a tal pico. Las constantes p_{min} y p_{max}, especifican el intervalo permisible de valores enteros de periodos tonales. En su desarrollo, Goncharov et al, establecieron tales valores como el redondeo de:

p_{min}= ($f_{muestreo}$/400) muestras tiempo
p_{max}= ($f_{muestreo}$/60) muestras tiempo

Goncharov et al, establecen que para una frecuencia de sampleo de 8 Khz, S_{max}=3. El periodo tonal estimado para la muestra-tiempo f_m es simplemente la muestra m-ésima de la ruta óptima calculada. Los estimados de periodos tonales, encontrados solo en los tiempos centrales de los cuadros de análisis, son interpolados para ser definidos para cada punto de muestra de la señal de voz original. El algoritmo de ruta óptima, intenta pasar a través de picos de alta energía y no de baja energía, lo cual resulta en mayor precisión donde es necesaria. Durante los segmentos de voz sorda, básicamente genera un contorno tonal que busca sincronizarse con los pulsos tonales de alta energía de los segmentos de voz sonora que lo rodean.

Localización de las Marcas de Periodo

El arreglo denominado *pitch(n)* es uno que guarda los estimados de los periodos tonales para cada punto de la señal de voz original $s(n)$, de *Npts* puntos totales. El algoritmo que a continuación se presenta establece un estimado inicial de localidad de marca de periodo (basado en la colocación arbitraria de la primera marca en *n=1*):

Se inicializa: *mark_array* $\leftarrow$ *Npts zeros; n=1*
while ($n \leq Npts$)

$mark\_array(n)=1$
$n = n + pitch(n)$
end while

Este ciclo, produce marcas de periodo que generalmente siguen a los pulsos tonales de la voz, aunque no están sincronizados para quedar localizados en el pico máximo de amplitud de la forma de onda. Si ahora, $c(k)$ guarda tales localidades de pulso, entonces se define una matriz **B** en la cuál cada columna k guarda los valores:

$$\left|s(c(k)) + [-p_{\max}, \ldots, p_{\max}]\right| \times hanning(w)$$

o un segmento ventaneado de las magnitudes de la voz tomadas de un intervalo de muestras centradas en la localidad de las marcas de periodo $n=c(k)$. Lo aquí calculado puede ser considerado una afinación del estimado original del periodo tonal (corregido de la desviación de la ruta óptima), junto con cálculos de una constante de fase para todas las marcas de periodo (para establecer la posición correcta de la primera marca).

Problemas con el uso del PSOLA

PSOLA resulta muy efectivo en cambiar el tono y duración de una señal, ello si los periodos son calculados de manera precisa. Sin embargo, incluso asumiendo que no existen errores en el marcaje de periodo tonales, pueden existir problemas con la concatenación de los diferentes segmentos:

1. Incompatibilidad de fase: Ello puede causar errores en la salida. El método MBROLA es una técnica que busca resolver tal problema, usando el PSOLA en dominio del tiempo como método de modificación de la prosodia, pero los ciclos tonales son pre-procesados de forma que tienen una fase fija. La ventaja radica en que la suavización espectral puede realizarse interpolando directamente los ciclos tonales en el dominio del tiempo sin agregar complejidad extra, y debido a que la fase es una constante, el algoritmo es más robusto ante errores de fase en la detección y marcaje de periodos. Sin embargo, tal metodología agrega ruido.

2. Incompatibilidad de tono: Ello puede ocurrir aún cuando no existan errores de tono o fase en la sección de análisis. Puede suceder que dos segmentos tengan la misma envolvente espectral pero diferente tono, por lo que el estimado del envolvente espectral es distinto y ocurren discontinuidades. Aún cuando se produce el mismo sonido en el mismo contexto fonético, un tono muy distinto seguramente resultará en una envolvente espectral muy distinta. Como ya se ha señalado, tal efecto es especialmente acentuado en cantantes de ópera, quienes tienen la capacidad de mover los formantes de su voz para que los armónicos caigan cerca de los valores de los formantes, y de tal forma producir una salida más poderosa (Huang, Acero y Hon: 2001, p. 830).

3. Incompatibilidad de amplitudes: Una diferencia de amplitud entre unidades puede ser corregida con una amplificación apropiada, pero el cómputo de tal factor no es evidente. Asimismo, el timbre del sonido puede cambiar con diferentes niveles de amplitud (Huang, Acero

y Hon: 2001, p. 830).

4. Fricativas sonora: la aproximación PSOLA no es capaz de manejar fricativas sonoras que se estiran considerablemente debido a que adiciona ruido (ello debido a que la repetición de cuadros induce periodicidad en la altas frecuencias que no estaban presentes en la señal original), o atenuación de la componente aspirada (en los cuadros que son interpolados) (Huang, Acero y Hon: 2001, p. 830).

III. *Diferencias entre la voz cantada y la voz hablada: De la prosodia hablada a la prosodia cantada*

Convenciones básicas: lírica

En la mayoría de las partituras de canciones la melodía y los acordes son acompañados de los líricos o letra. Al ser la voz un instrumento de naturaleza melódica, y particularmente para el caso del Español, tradicionalmente la letra se escribe debajo de cada línea de música, mostrando exactamente cuales *sílabas* se cantan junto con cada nota musical.

Los dos elementos básicos de la música que definen la melodía son el tono y el ritmo. La melodía es una sucesión de tonos que siguen cierto ritmo, y es usualmente la parte más memorable de una canción; la sección que quien la escucha recordará y será capaz de repetir. Al suceder la música a través del tiempo, su control es el que hace a la misma. El tiempo musical se mide en pulsos, la métrica y el ritmo. El pulso, es una serie indistinguible de pulsaciones, mientras que la métrica es el número de pulsaciones que existen entre acentos regularmente recurrentes. En otras palabras, la métrica es el agrupamiento de pulsaciones. El ritmo es el elemento estructural más importante; es siempre cambiante, y sin embargo, puede ser medido por su relación a la métrica. A un nivel micro, controla la estructura interna de cada frase y la relación entre las frases individuales; a un nivel macro, controla la forma de la composición musical completa. El estrés métrico siempre tiene un efecto de suma en cualquier nivel de la división o subdivisión de pulso en el cual tome lugar. El factor más importante en la determinación de cuanto estrés recibe cierta nota, es la relación del ritmo con la métrica, y las divisiones o subdivisiones rítmicas que ocurren dentro de la métrica. Otros factores que contribuyen a cuanto estrés recibe una nota son 1) la duración, 2) el tono, 3) acentuación y 4) el nivel dinámico. Debido a que generalmente, en la música lírica se utilizan palabras reales y se habla en oraciones, la mayoría de las letras de canciones tienen algún parecido con la estructura regular de las oraciones, aunque en otros casos simplemente se arrojan frases o imágenes, haciendo que quien lo escucha trabaje en entender la historia o situación (Wayne Whadams: 2001, pp. 21-22).

Una gran cantidad de elementos de análisis pueden aplicarse a las letras de las canciones, desde el balance en el número de frases, la duración de las frases y el control de la velocidad, la estructura y forma general, hasta qué tan bien se acoplan las sílabas a la melodía (*prosodia musical*). El elemento rítmico, debe ser entendido desde sus ladrillos más básicos de construcción, que son las sílabas, que como se ha mencionado, están constituidas por un sonido vocálico y, cero, uno o más sonidos consonantes. Los diptongos, al contener dos sonidos vocálicos, pueden ser musicalizados por los compositores con dos notas, mientras que típicamente una sílaba con sólo un sonido vocálico, se musicaliza con una sola nota (las reglas de división de palabras en sílabas en español fueron revisadas en capítulos pasados y no se repetirán aquí). Un aspecto importante es el énfasis o acento que se pone en cada sílaba, lo cuál da a las palabras de dos o más sílabas, una "figura" sonora que ayuda a nuestros oídos a escuchar grupos de sílabas que van juntas. Una sílaba enfatizada o acentuada difiere de las sílabas no enfatizadas en tres aspectos primordiales: una sílaba acentuada, comparada con las sílabas a su alrededor es (Pat Pattison: 1991, p.20):

1. más alta en tono 2. más sonora 3. más larga.

Las palabras de más de dos sílabas tienen una pequeña melodía asociada, presentando la sílaba acentuada sobre el *"tiempo fuerte*" (tal melodía, nos brinda una forma adicional para reconocer palabras durante las conversaciones, sin que quien habla tenga que pausar entre palabras). Para el canto en inglés, además del énfasis primario, existen énfasis secundarios, así como *énfasis por importancia*.[3] Una vez que se colocan las sílabas en patrones es cuando se realiza la mezcla entre el elemento rítmico con el tamaño de las frases y el número de éstas. Las sílabas en la letra de una canción están dispuestas de manera que encajen con las notas, en este sentido la letra está "casada" con la música (Pat Pattison: 1991, p. 23), y la música es rítmica por naturaleza. De tal forma que las sílabas deben estar arregladas en patrones rítmicos, ya sea para prepararlas para la música o para hacerlas coincidir con música que ya ha sido escrita. No hace falta mencionar que normalmente, se colocan las sílabas con mayor énfasis en posiciones rítmicas "fuertes", ya que las frases líricas y las musicales deben presentarse en sintonía.

Una frase melódica, de manera similar a lo que sucede con una oración o cláusula en el lenguaje verbal, usualmente incluye una declaración musical completa; se define a si misma por medio de ciertas pausas, o notas sostenidas, o llegando a cierto punto de resolución (rítmicamente y/o tonalmente). Especialmente en la música vocal, ello está relacionado directamente con las áreas naturales de la respiración. Las frases cortas usualmente se agrupan para formar una frase de mayor longitud. Existen dos tipos de *movimiento melódico*, el *conjunto*, el cual procede por paso de un grado de una escala al siguiente (por intervalos de segunda), y el *desasociado*, que procede mediante saltos (intervalos mayores a una segunda). Una melodía asume su carácter de: su estructura rítmica, su contorno, su construcción tonal y su contenido interválico. La mayoría de las melodías vocales consisten en un movimiento conjunto, el cual es el más natural y sencillo de cantar. Sin embargo, son justo los saltos de intervalo los que otorgan carácter a la melodía y por los cuales adquiere un perfil memorable.

Los términos "masculino" y "femenino" son usados en análisis poéticos y líricos. Una palabra que termina en una sílaba acentuada, se dice que tiene una terminación "masculina", mientras que una terminación "femenina" la presenta una palabra que termina en una sílaba no acentuada. Se debe considerar el problema en prosodia que generaría el colocar una sílaba no acentuada, o con estrés secundario, en el primer pulso de un compás, debido a que no coincidirían el acento musical con el acento silábico. Debe también considerarse la posición de los líricos dentro del compás, cuanto espacio es necesario entre frases, la prolongación y acentuación de ciertas palabras que necesitan énfasis, y la elección del tono de cada sección de la melodía. Tales factores, entre otros, deben considerarse en la conformación de la letra de una canción para lograr que las sutilezas emocionales puedan ser expresadas, por lo que naturalmente resultan relevantes en el diseño y operación de un sistema de síntesis de voz cantada, al que se busca algún nivel de aplicación musical lírica.

Se puede decir, en cuanto al "esfuerzo" que imprime un cantante, que no puede ser

[3] Existen en la literatura, escasos análisis de la aplicación de la prosódia de la voz cantada, y el autor se atreve a decir que ninguno aplicado al canto del Español, lo que motiva y obliga a la inclusión de referencias de otros idiomas más estudiados, como lo es el inglés.

modelado simplemente escalando la amplitud, ya que hacerlo solamente daría la sensación al escucha de que el cantante se encuentra localizado más lejos. Cuando se canta "fuerte", la potencia en la parte alta del espectro se eleva significativamente con respecto a la sección baja, y por lo tanto es más adecuado *modelarlo* variando la inclinación de la envolvente espectral. Tal efecto es audible también en notas de larga duración, donde la energía de la señal es más o menos constante, pero el esfuerzo del cantante crece hacia el final de la señal (Siivola: 2002, p. 4).

El papel del entendimiento

La buena prosodia depende del entendimiento del orador o lector, y del significado del mensaje. A la fecha, la mayoría del trabajo realizado sobre prosodia para sistemas *texto-a-voz,* se ha centrado exclusivamente sobre el contenido literal del mensaje. Es decir, un sistema *texto-a-voz* aprende lo que puede aprender de una representación textual individual de una oración o frase para ayudar en la generación de la prosodia. Típicamente, estos sistemas se basan en: identidad de la palabra, lugar o parte en el discurso que una palabra toma, puntuación, largo de una oración o frase y otras características superficiales. De manera más sofisticada pueden considerarse también propiedades más profundas que incluyen el contexto del documento o discurso. Hasta ahora, la cualidad vocal de un humano es generalmente superior a las voces sintetizadas, y la voz humana natural es mas placentera al escucha. El método más comúnmente utilizado para derivar prosodia en los sistemas *texto-a-voz* está basado en la distinción entre clases cercanas de funciones de palabras, como pueden ser los *determinantes* y las *preposiciones*, que en general es considerado que reciben menos énfasis que, por ejemplo, los *sustantivos*. En la prosa ordinaria la definición y recuperación del *significado* se mantiene como una interrogante ya que el grado de *entendimiento* del contenido de un mensaje que se requiere para una rendición prosódica convincente es aún desconocido. Aunque es claro que entre más *conoce* el humano, o la máquina lectora mejor será la ejecución prosódica pero parte del conocimiento más importante para tal fin es sorprendentemente superficial y accesible.

No existe una especificación o definición rigurosa de *significado*. El significado de la ejecución prosódica misma es más importante que el significado inherente del texto si es que existe. En la voz hablada son primariamente las metas de quien habla y de quién escucha las que determinan el significado de la ejecución prosódica. Mientras son los atributos textuales como las convenciones métricas, sintaxis, morfología, semántica lexicológica, tema, etc. quienes contribuyen a ambos tipos de significado; el significado de la ejecución prosódica incorpora elementos contextuales y pragmáticos de mayor importancia, como las metas del evento comunicativo, así como la identidad de quién habla y actitud de proyección (Huang, Acero y Hon: 2001, pp. 740-742).

Parece inevitable que el énfasis y análisis del entendimiento sea radicalmente distinto para la voz cantada que para la voz hablada. De hecho, al ser la inteligibilidad de las palabras en una canción menos importantes que en la voz hablada, es incluso común que se comprometa la misma para producir características deseables, como una voz cantada más poderosa (Siivola: 2002, p. 5). Dependiendo del estilo de la música el *significado* del texto cantado tendrá mayor o menor peso.

Se debe mencionar que, en lo referente a este trabajo sólo se abordarán los temas de la modificación del tono y duración relativa, debido a que tales son los temas centrales en la investigación y el desarrollo presente, que se enfoca a la *altura* y *duración* (y en menor medida a la *inteligibilidad*) de la voz cantada. No por ello se quiere decir que en lo que respecta a la voz cantada los demás aspectos de la prosodia resultan de menor importancia cuando son trasladados al ámbito de la *melodía* sino que esta investigación se remite exclusivamente a los dos mencionados como punto de inicio de la generación de un sintetizador de voz cantada en español y como objetivo de la tesis de maestría.

Capítulo V

I. *Modificación de un Sintetizador de Voz hablada en Español (basado en difonemas) en uno de Voz cantada por concatenación de sílabas en Español.*

El Sintetizador de voz cantada (en Español) desarrollado para esta tesis, está basado en un sistema de *texto-a-voz* desarrollado en el posgrado de Ingeniería Eléctrica de la UNAM por Fernando del Río, 2006, y cuya tesis de maestría lleva por nombre: "Diseño de un sintetizador de voz en español usando el método TD-PSOLA". Aunque originalmente el sintetizador fue escrito en lenguaje C, la implementación del algoritmo PSOLA fue realizado en *Matlab*, lo que constituyó, en conjunto con la experiencia previa en programación, la razón por la cual se eligió esta última plataforma. Por lo anterior, un apropiado entendimiento del programa original era requerido como primer paso hacia la construcción del sistema, ya que del conocimiento de la arquitectura y flujo de información originales se propondrían las modificaciones pertinentes para lograr la aplicación del sistema a voz cantada en español.

Como se mencionó al inicio de esta tesis, el sistema está basado en la síntesis por concatenación de sílabas cantadas, pre-grabadas. La hipótesis de que la *sílaba* es la unidad básica estructural del canto (*popular,* en el idioma español, hablado en México), está basada en varios *hechos*, sin embargo el más importante es que toda partitura vocal lírica (en español), contiene una línea melódica donde hay un correspondencia uno a uno entre nota (*tono* y *duración*) y sílaba (excepto en los *melismas*). El presente autor no pudo hallar trabajos relacionados específicamente con este tema, aunque en muchos de los que pueden hallarse sobre voz cantada,[1] se hace referencia a la *silabación*, como un aspecto de la metodología del canto.

Presumiblemente, un sistema basado en unidades del tipo "sílabas", disminuirá dramáticamente su propensión a discontinuidades generadas por la concatenación. Sus características intrínsecas de articulación (revisadas en capítulos anteriores). Como se ha dicho, las sílabas incluyen la co-articulación en el lenguaje, y constituyen una unidad fonética clara y sin efectos de borde; asimismo el hecho de que las sílabas siempre incluyen una vocal (al menos en el canto popular en español) facilitan la implementación PSOLA, ya que el *tono* de la nota a la cual la sílaba estará asociada, siempre relaciona la frecuencia fundamental a la vocal cantada. Los algorítmos tipo *PSOLA*, pueden extraer información sobre el *tono*, por lo que controlan el aspecto *sonoro* más importante de las *sílabas,* pero aún más importante es que pueden modificarlo**.** El algoritmo asociado a la extracción de marcas de clase es sin duda el aspecto central de los métodos *PSOLA.*

Como primer acercamiento a la síntesis de voz cantada, el presente trabajo nunca pretendió desarrollar un sistema exhaustivo o muy flexible, planteándose objetivos muy modestos dirigidos a modificar *tonos* y *duraciones* de las sílabas, para lo cual se debía estudiar un sintetizador de voz hablada y comprender el funcionamiento de su módulo de modificación prosódica. La mayor parte del tiempo dedicado a este trabajo fue asociado al comentario de aquel sistema, que en un principio incluía más de mil líneas de programa, y sólo cuatro de ellas estaban comentadas. El *Apéndice I* y *II*, incluyen el comentario a fondo y

[1] Vease: http://www.consuperiorsal.com/documentos/Asignaturas/lengua%20inglesa%20aplicada%20al%20canto.pdf

completo del Sistema original.

Un comentario sobre el sintetizador de voz hablada.

En términos de lo planteado en el esquema general de los sistemas *texto-a-voz* (*Imagen 28*), la sección del sintetizador original que fue utilizada, era la encargada del procesamiento de señales (conversión Fonemas a Voz) en la que la *función* **avoz** generaba el archivo de sonido mientras que la sub-función ***modifica,*** constituía el sub-módulo de análisis de sintaxis y prosodia. Ésta última incluye al algoritmo TD-PSOLA.

Imagen 37. La función* avoz *convertía cadenas de fonemas en voz hablada

Función* avoz*: concatenación de difonemas.

Para la *función* **avoz**, debido a que los difonemas constituían las unidades fonéticas básicas, las variables de entrada resultaban **l1, l2, l3, P1** y **S1**, donde:

p1 - fonema previo al difonema a modificar

l1 y **l2** - primer y segundo fonemas del difonema a modificar

s1 - fonema siguiente al difonema a modificar

ant - última muestra de la señal (valor, no posición) de salida generada

Y los parametros de salida:

y - Difonema actual modificado

El objetivo de la *función* **avoz**, era el finalmente alimentar a la *función* modifica sus parámetros de entrada. En el programa original (en Mathlab con el algoritmo TDPSOLA), para la generación de la voz hablada, se forzaba la coincidencia entre el segundo ciclo de un difonema con el ultimo ciclo del difonema anterior. La modificacion de frecuencia se distribuía de forma lineal a lo largo del 2º difonema para que la parte final de éste no fuera modificada. La modificación podía darse en cualquier porcentaje de 50 (=.5) a 200% (=2). En caso de que lo que se deseara era que no se modificara la velocidad al inicio del 2º difonema (es decir, para variar únicamente la frecuencia), se conserva el mismo número de muestras. Una vez con la misma frecuencia, se emparejaban entonces amplitudes, modificando todas las muestras del

2 difonema, y finalmente se obtenían las nuevas marcas de periodo. En la implementación del algoritmo, se eliminaban las orillas de la señal modificada, ya que éstas son más propensas a contener distorciones y elementos no deseados, quitando del 2o difonema, el primer y último ciclo. Si la señal no comienzaba en silencio, entonces todo el 2o difonema era escalado para que fuera proporcional en amplitud, mejorando la unión del 1er y 2º difonema.

Function **pitch_marks:** *Marcas de periodo*

La *function* **pitch_marks** = find_pmarks(**speech**, **fs_in**), que era la encargada del cálculo de marcas de periodo, tiene como entradas:

speech: señal de audio a analizar

fs_in: frecuencia de muestreo

Y como salida:

Pitch_marks: posición de las marcas de periodo

Su funcionamiento puede ser dividido y explicado en cuatro pasos:

1. La generarción del contorno de energía de la señal. Para ello se obtiene la longitud de la ventana de Hanning, de acuerdo a la frecuencia de muestreo, y el numero de muestras que corresponda a 6.4ms de señal. Se duplica la señal (spch) y se obtiene su longitud (xsamp). La señal de audio se eleva al cuadrado y se convoluciona dos veces con una ventana de Hanning.

2. Obtención de la primera aproximación de marcas de periodo a partir de los máximos locales del contorno de energía. En este paso se debe obtener el número de la muestra donde se localizó cada máximo local, y la longitud en muestras de cada periodo obtenido.

3. Adición de marcas de periodo en las secciones no periódicas de la señal, ajustándolas a las marcas de las secciones periódicas. Aquí se divide la señal en bloques de acuerdo a la frecuencia de muestreo (aprox. 1.5 centésimas de segundo) y se acomodan en una matriz indicando la posición de las marcas. La matriz se inicializa en 0 y se llena de la siguiente manera:

$$(\text{renglón}, \text{columna}) = \text{muestra}(n)$$

donde n es el valor de la muestra. El número de columnas, es el número de ventanas analizadas, y el número de renglones corresponde al número de muestras de la señal. La matriz contendrá valores sólo en la posición que corresponde a las muestras donde se localizaron las marcas.

4. Ajuste de marcas de periodo para que se ajusten a máximos de la señal. Para ello se ubica, en cada columna, una ventana alrededor de la muestra donde se localizó la marca de periodo. Se recorre la posición de las marcas para que coincidan con los máximos locales de cada

ventana, obteniéndose una nueva ruta. De acuerdo a las distancias entre las marcas, se obtiene el número de muestras donde se deben localizar las marcas (pitch_marks).

Función **path:** *Cálculo de la Ruta Óptima*

La *función* **path** = rridern(**MAT**,**N**) tiene por entradas:

MAT: la matriz a traves de la cual se desea obtener la ruta óptima (donde la sumatoria de los valores de los renglones por los que pase sea la máxima posible sin tener una pendiente mayor a N).

El mecanismo que se sigue es el recorrer la matriz, añadiendo marcas de periodo en las zonas donde no existan, generando una ruta óptima, tocando el mayor número posible de marcas encontradas sin exceder una pendiente máxima (N = 1 salto de máximo 3 renglones por columna). Se obtiene entonces la matriz "*best_paths*", que contiene para cada renglón y columna, la pendiente que se requiere para tocar el punto de mayor valor accesible desde esa posición. Después se modifica la matriz, añadiendo el error acumulado mínimo que se ha obtenido para cada posición, recorriendo columna por columna. Empezando por la última columna, se obtiene la *ruta (path)* en la que el error acumulado es menor de entre todas las posiciones accesibles. Se almacena el renglón que corresponda a cada columna de la ruta óptima, y se realiza una interpolación lineal de la ruta de longitud igual al número de muestras. Finalmente, se obtiene la nueva posición de las marcas.

Función **tdpsola:** *modificación prosódica*

La *función* **y** = **tdpsola**(**s, fs, pscale, pscale2, tscale, tscale2**), encargada de presentaba los siguientes parametros de entrada:

s - 2º difonema o difonema a modificar

fs- frecuencia de muestreo de la señal

pscale - porcentaje a escalar la frecuencia al inicio del (2º) difonema **s**

pscale2- porcentaje a escalar la frecuencia al final del (2º) difonema **s**

tscale - porcentaje a escalar en duración al inicio del (2º) difonema **s**

tscale2 - porcentaje a escalar en duración al final del (2º) difonema **s**

Para la modificación prosódica, se calculaba la posición de las nuevas marcas de periodo, obteniéndose la longitud de cada marca de periodo. El porcentaje de modificación del periodo se obtenía a partir de los porcentajes de variación inicial y final, aplicándose una variación lineal. La longitud de la nueva marca de periodo a la salida estaba dada por:

- longitud actual * % de modificación

Así pues, se obtenía el número de marcas de periodo modificadas, para luego calcular el número de muestras a agregar o eliminar para el primer periodo. Para cada una de las marcas existentes, se aumentaba el número de muestras a agregar o eliminar, en caso de que se deseara aumentar la duración de la señal, entonces se debían duplicar periodos hasta que el numero de muestras requeridas fuera menor que la duración del periodo. En caso de que se estuviera disminuyendo la duración de la señal; entonces se debían eliminar periodos hasta que el número de muestras a eliminar fuera menor que la duración del periodo.

Al duplicar un ciclo, se debía obtener el número de periodos de la nueva señal modificada. Para obtener la posición de la muestra inicial y final de la primera ventana, era necesario poner en cero el número de muestras de la nueva señal, mediante un vector auxiliar. Para cada una de las nuevas ventanas, se debía obtener la posición de la muestra inicial y final pre-calculadas. Luego se obtenía la longitud de la ventana, generando una ventana de Hanning con longitud igual a la de la ventana. Al duplicar ventanas, las copias pares (2,4,6...) eran invertidas para evitar generar efectos de periodicidad en señales no periódicas. A la señal de salida generada hasta ese momento, se debía sumar la nueva ventana multiplicada por la ventana de Hanning. Esta señal podía traslaparse con muestras anteriores, esto es, calculando el factor de normalización para disminuir la distorsión generada por la ventana de Hanning. A cada muestra, se debía dividir la muestra por el factor de distorsión si este era diferente de 0.

El sintetizador de Voz cantada y la Sesión de Grabación

Tras el análisis minucioso del sintetizador de voz hablada, se llegó a la conclusión de que la única función necesaria, era:

1) **tdpsola** (que incluye la función ***find_pmarks***)

De los parámetros de entrada, únicamente la señal debía ser cambiada, sustituyendo difonemas en formato (*.pcm*), por sílabas pre-grabadas en formato (*.wav*). Se eligió para la construcción de una muy pequeña base de datos, la voz de una soprano, Guadalupe Caro Cocotle, quien terminó la licenciatura en Música como cantante en la Universidad de Manitoba, Canadá. Para la grabación se realizó un calentamiento de media hora (12:30 pm-1:00 pm) y se grabó utilizando un sistema Protools 7.4 LE y una tarjeta de audio MBOX (Digidesign) a 48 kHz de frecuencia de muestreo y 24 bits de profundidad. Asímismo, se utilizó un micrófono AKG 414 (aplicación: voz, locución y piano) en modalidad cardioide. La sesión de grabación de la base de datos se llevó a cabo el 18 de agosto del 2008, en un estudio de grabación casero acondicionado por el autor con material absorbente y difusor, en un cuarto de 5 x 5 metros aproximadamente.

Considerando que el intervalo vocal de los vocalistas populares, se le requirió a la cantante que las sílabas se grabaron sin *vibrato* y a dos tonos diferentes, Fa-4 y Do-5, a tempo = 60, correspondiendo cada sílaba a una *blanca* (sin embargo cierta cantidad de vibrato fue ineludible). La longitud de las sílabas fue de 1 segundo (± 0.2 s). Se realizó una sesión de calentamiento de 15 minutos que incluía: ligados de tres notas, escala cromática con rango máximo de octava y media, y arpegios de hasta octava y media. El registro de voz de la cantante (Guadalupe Caro) es de dos octavas, Do-4 a Do-6.

Resultados: El Sintetizador a Prueba

Considerando los modestos objetivos del programa, no se desarrolló ningún protocolo específico de análisis de los resultados de modificar el *tono* y *duración* de las *sílabas* grabadas. En cambio, la presente tesis se limitó a dar un reducido número de tareas al *Sintetizador,* relacionadas directamente con nociones melódicas como son: 1) El *Tono y duración y,* 2) La escala mayor (intervalos)*.* En particular, la evaluación del *desempeño* del *Sintetizador*, se realizó de manera cualitativa, mediante la ejecución de una canción popular mexicana en Español: *Las mañanitas,* elegida debido a su estructura interválica y rítmica, así como a su texto que presenta un número limitado de sílabas. Una segunda melodía cantada en español fue construída, de manera que facilitara la comparación y evaluación del uso de la misma base de datos (sílabas pre-grabadas) para dos canciones distintas.

A continuación se presentan los espectrogramas de las sílabas de la melodía "Las Mañanitas", antes (columna izquierda) y después (columna derecha) de ser modificadas para su comparación. La modificación obedece a la primera versión generada de dicha tonada, y cuya programación es el archivo *"mananitas.m"* (matlab). Todos los espectrogramas fueron generados con ventanas Hanning de 1024 muestras, y con un traslape de 512 muestras, parámetros con los que la mayoría de ellos pueden leerse con facilidad. El eje temporal en todos, está dado en segundos y el título del gráfico se refiere a la sílaba y el tipo de modificación a la que se sostuvo. Las sílabas marcadas "(C)", son aquellas grabadas originalmente en el tono Do-5, el resto fueron grabadas originalmente a Fa-4. Los espectrogramas se presentan en la secuencia silábica que la partitura marca, hasta el compás 9. Se usó la versión coloquial que sustituye "…tu cumpleaños…" por "…día de tu santo…". Asimismo, se traspuso a Do la partitura.

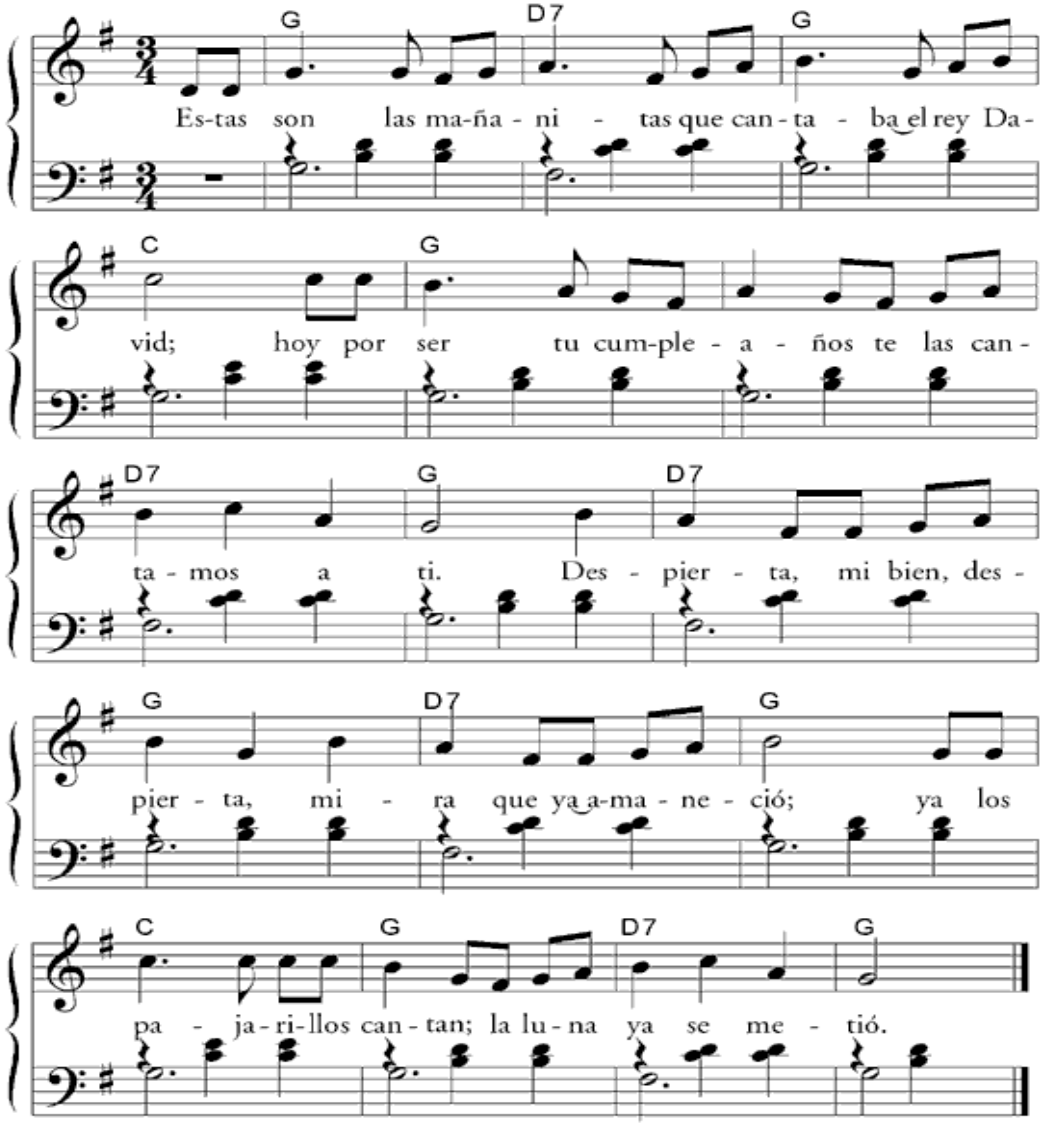
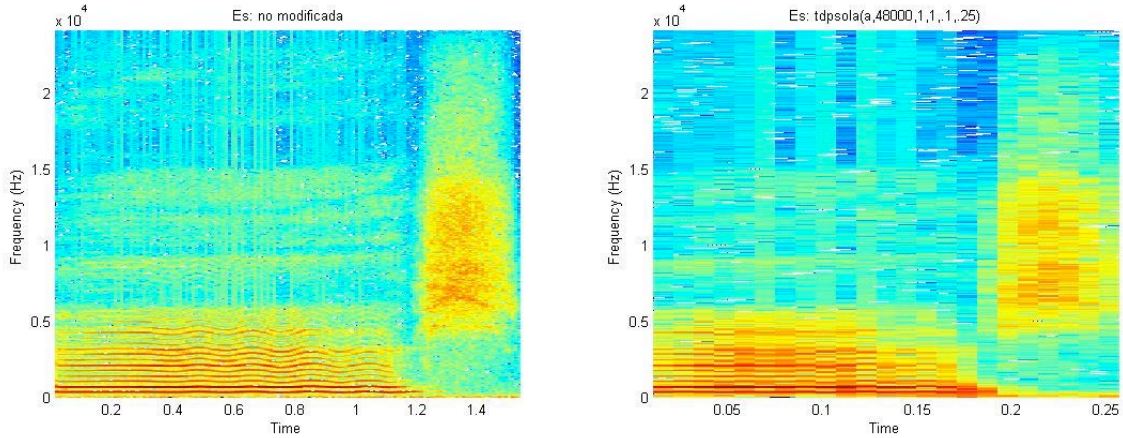
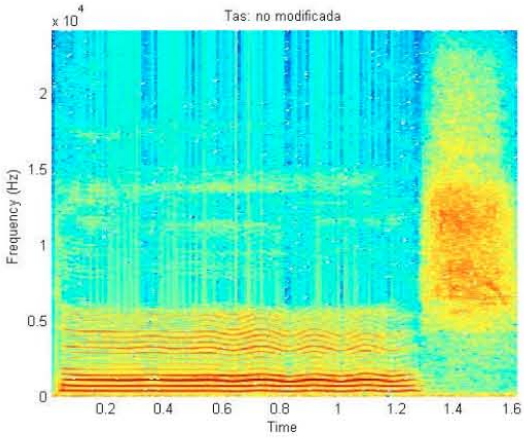
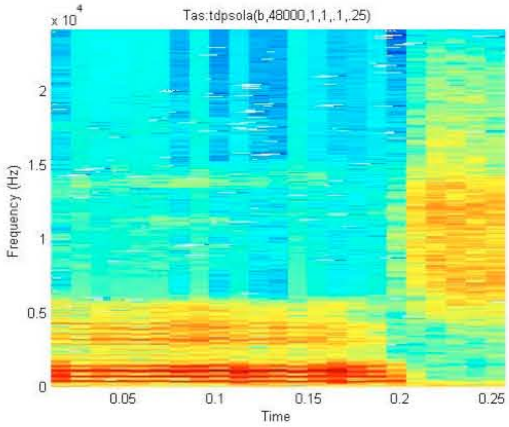
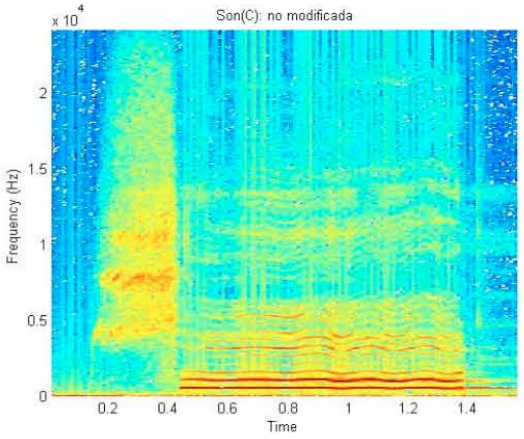


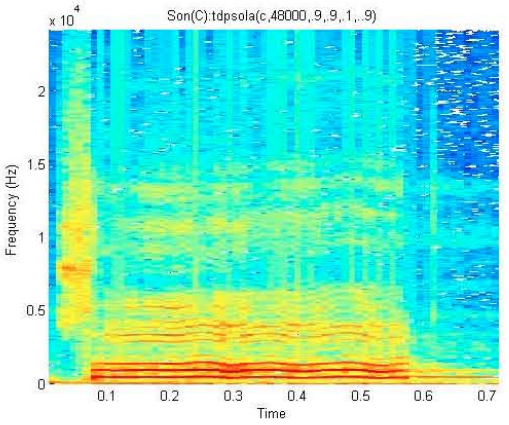
Imagen 38. "Las mañanitas" (canción popular)

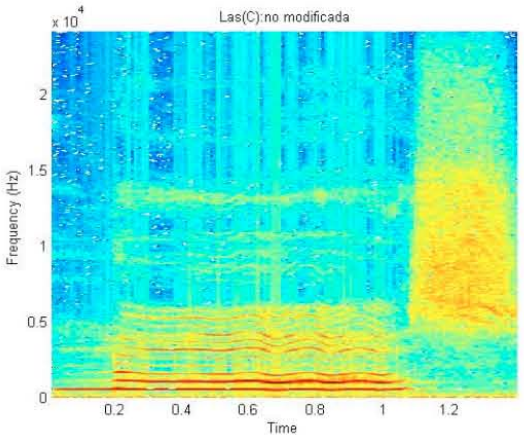


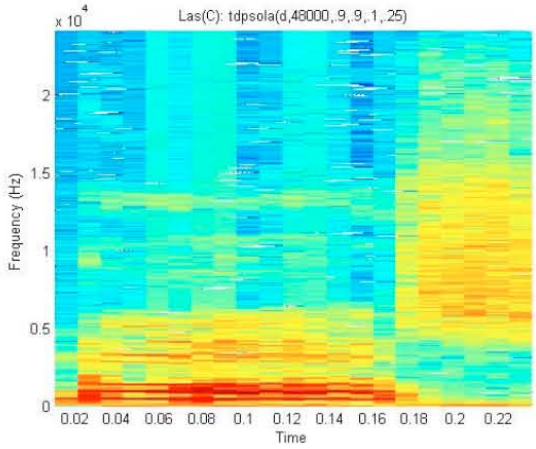

Tas: no modificada
x 10^4
Frequency (Hz)
Time

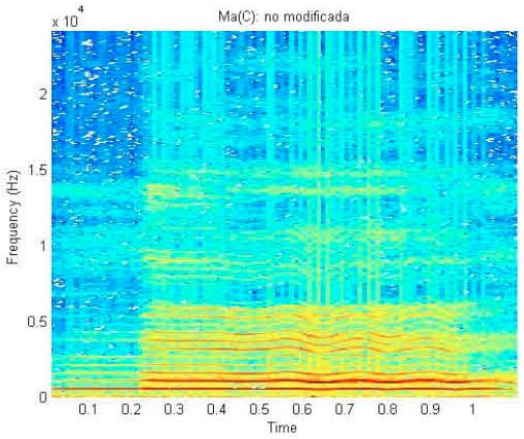

Tas:tdpsola(b,48000,1,1,.1,.25)
x 10^4
Frequency (Hz)
Time

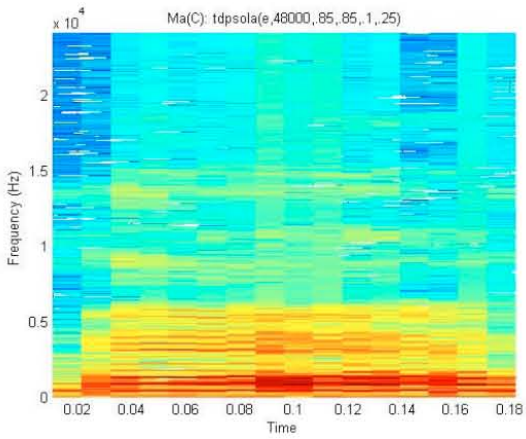

Son(C): no modificada
x 10^4
Frequency (Hz)
Time

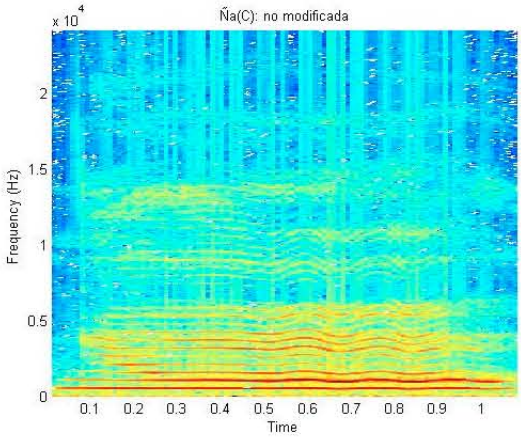

Son(C):tdpsola(c,48000,.9,.9,.1,..9)
x 10^4
Frequency (Hz)
Time

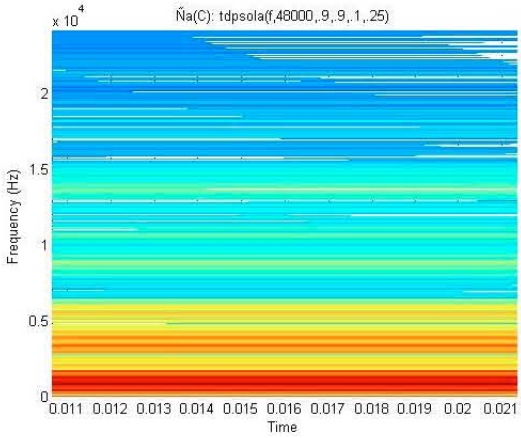

Las(C):no modificada
x 10^4
Frequency (Hz)
Time

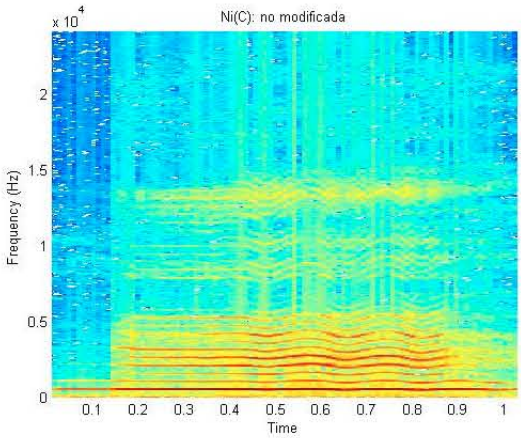

Las(C): tdpsola(d,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

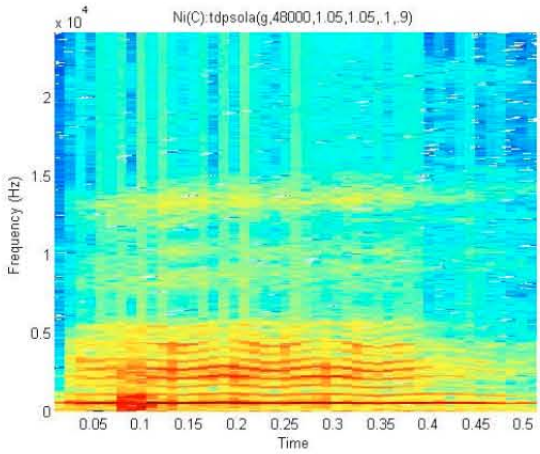

Ma(C): no modificada
x 10^4
Frequency (Hz)
Time

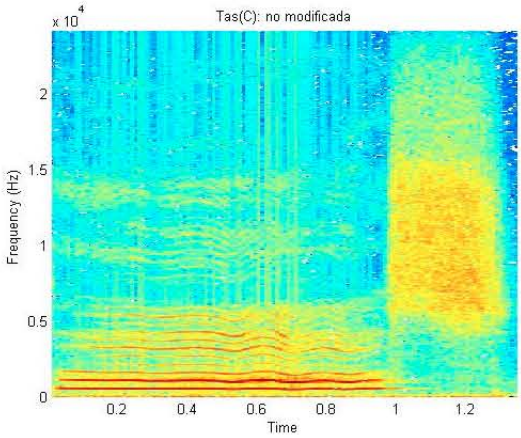

Ma(C): tdpsola(e,48000,.85,.85,.1,.25)
x 10^4
Frequency (Hz)
Time

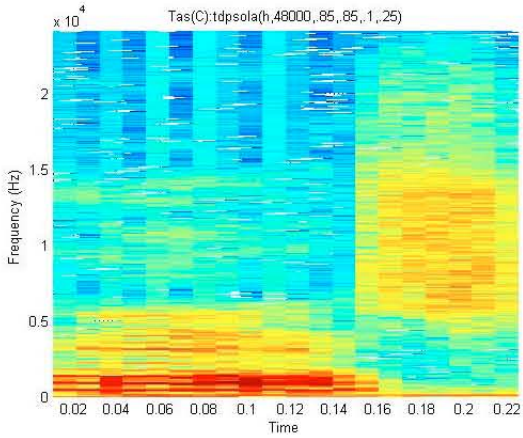

Ña(C): no modificada
x 10^4
Frequency (Hz)
Time

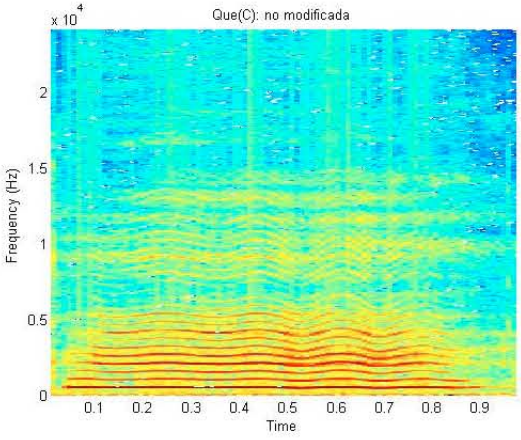

Ña(C): tdpsola(f,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

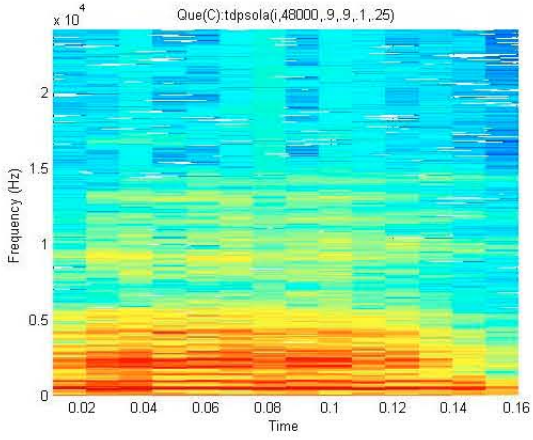

Ni(C): no modificada
x 10^4
Frequency (Hz)
Time

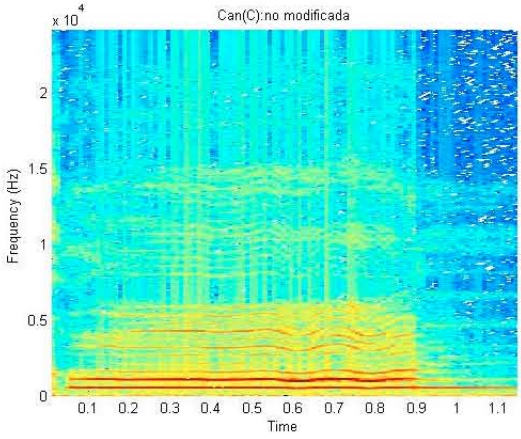

Ni(C):tdpsola(g,48000,1.05,1.05,.1,.9)
x 10^4
Frequency (Hz)
Time

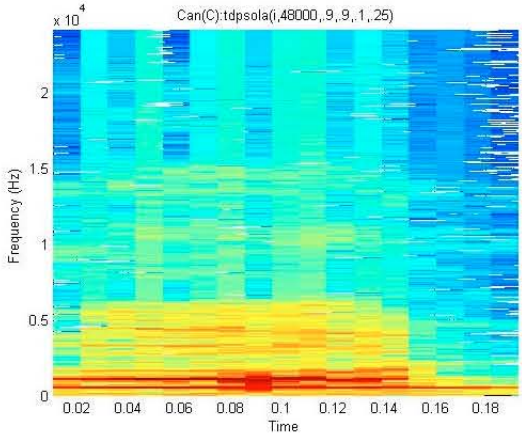

Tas(C): no modificada
x 10^4
Frequency (Hz)
Time

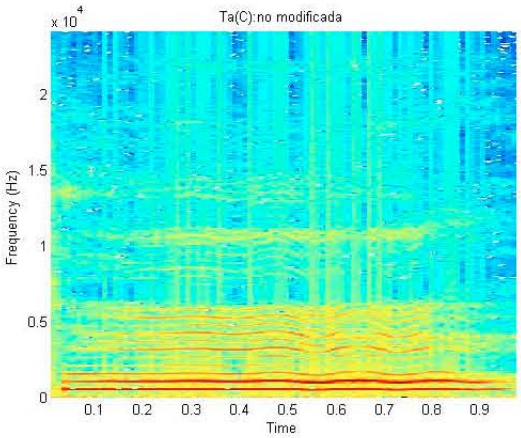

Tas(C):tdpsola(h,48000,.85,.85,.1,.25)
x 10^4
Frequency (Hz)
Time

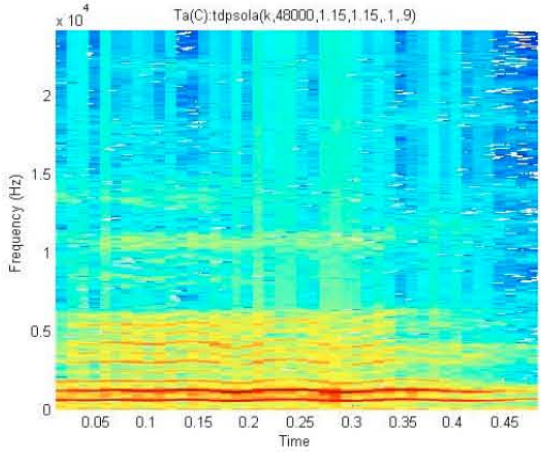

Que(C): no modificada
x 10^4
Frequency (Hz)
Time

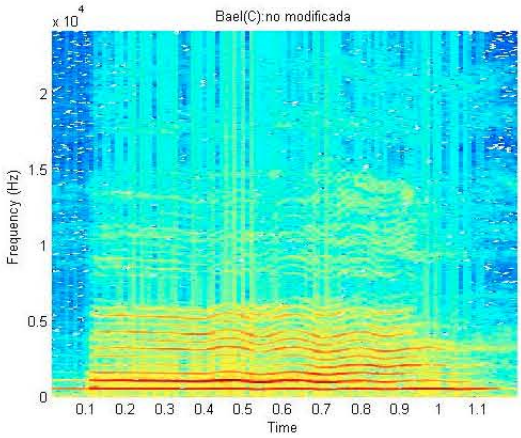

Que(C):tdpsola(i,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

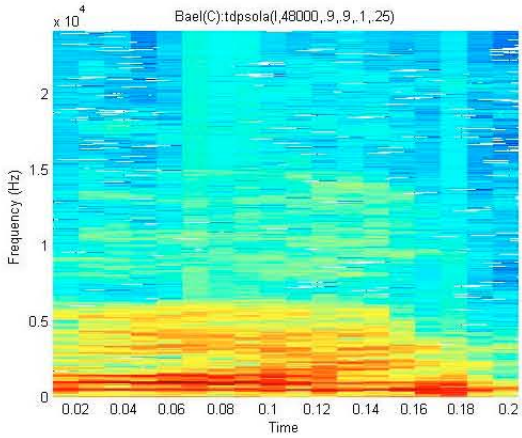

Can(C):no modificada
x 10^4
Frequency (Hz)
Time

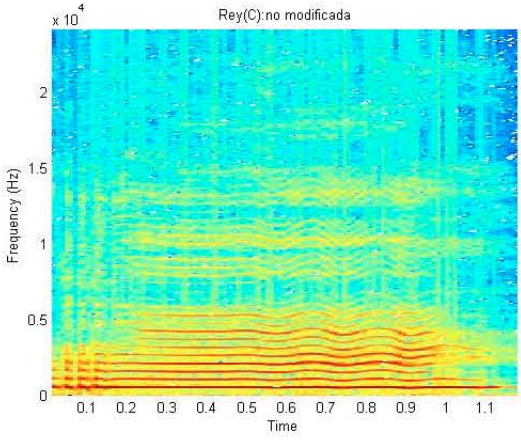

Can(C):tdpsola(i,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

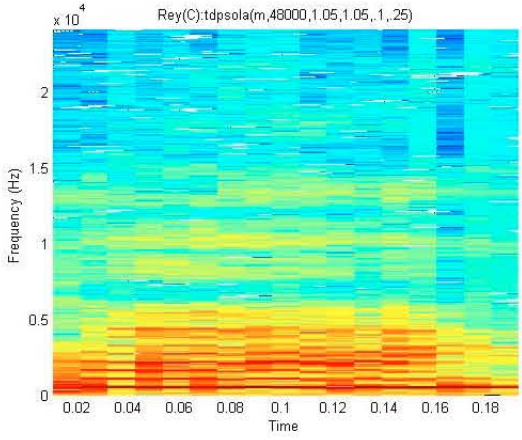

Ta(C):no modificada
x 10^4
Frequency (Hz)
Time

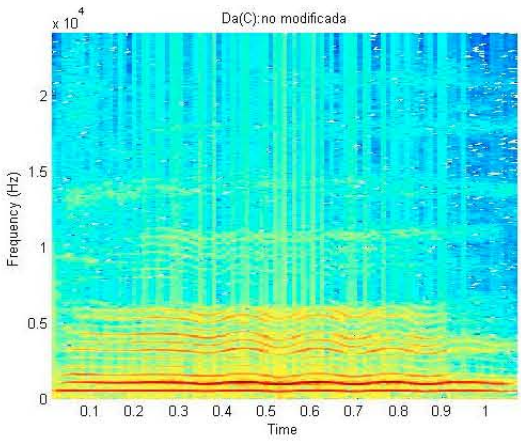

Ta(C):tdpsola(k,48000,1.15,1.15,.1,.9)
x 10^4
Frequency (Hz)
Time

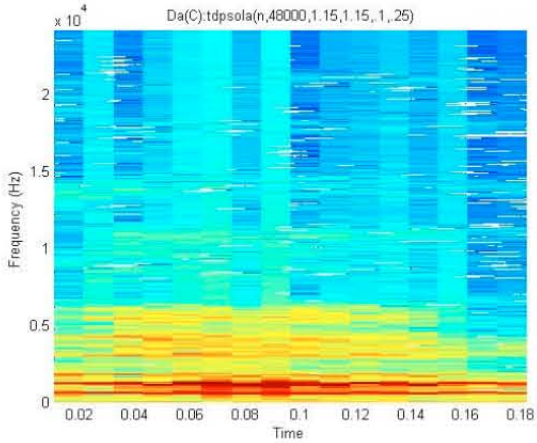

Bael(C):no modificada
x 10^4
Frequency (Hz)
Time

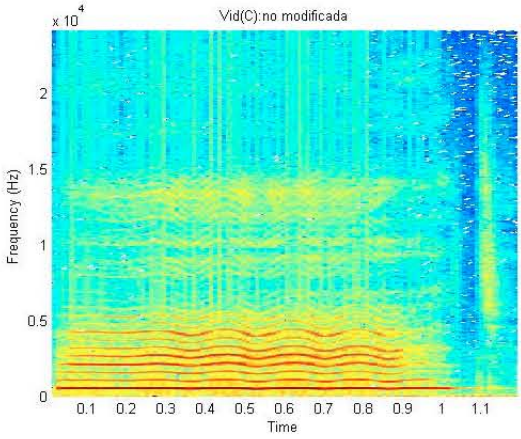

Bael(C):tdpsola(l,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

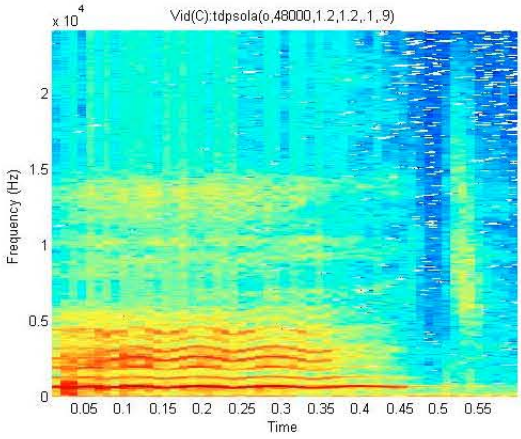

Rey(C):no modificada
x 10^4
Frequency (Hz)
Time

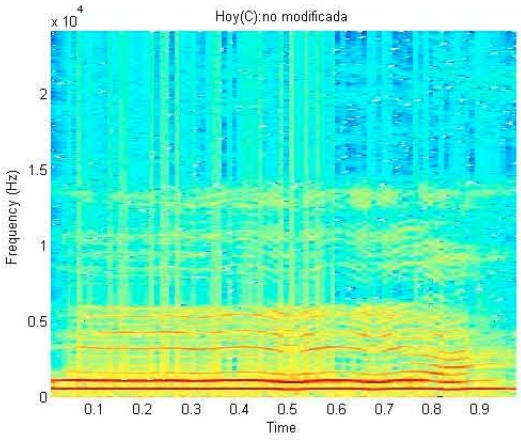

Rey(C):tdpsola(m,48000,1.05,1.05,.1,.25)
x 10^4
Frequency (Hz)
Time

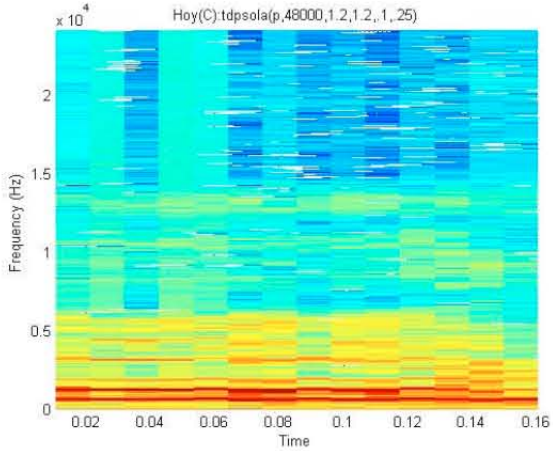

Da(C):no modificada
x 10^4
Frequency (Hz)
Time

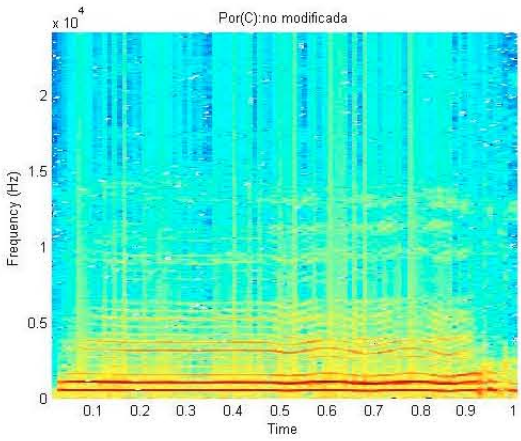

Da(C):tdpsola(n,48000,1.15,1.15,.1,.25)
x 10^4
Frequency (Hz)
Time

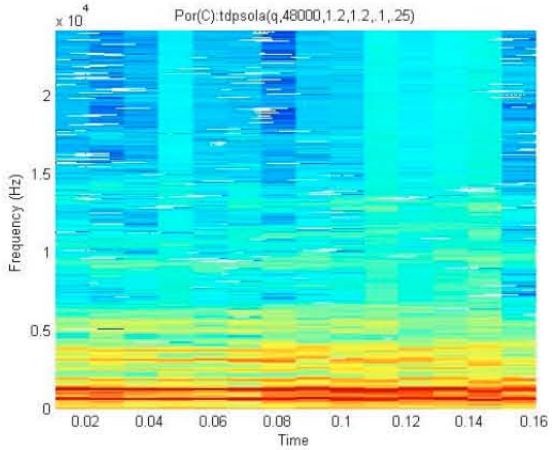

Vid(C):no modificada
x 10^4
Frequency (Hz)
Time

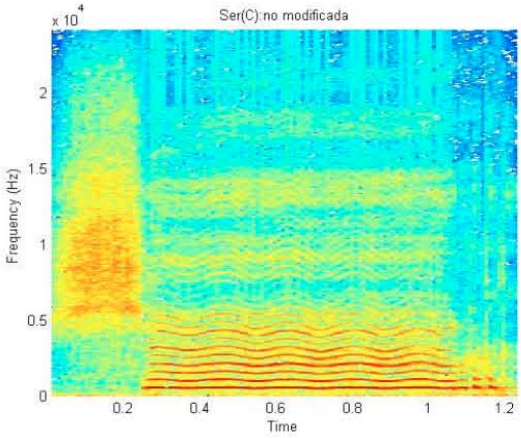

Vid(C):tdpsola(o,48000,1.2,1.2,.1,.9)
x 10^4
Frequency (Hz)
Time

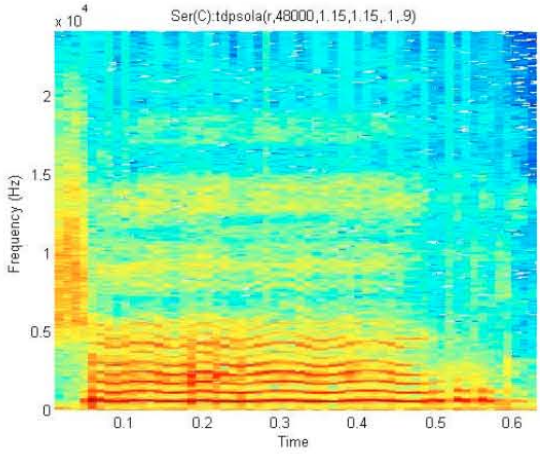

Hoy(C):no modificada
x 10^4
Frequency (Hz)
Time

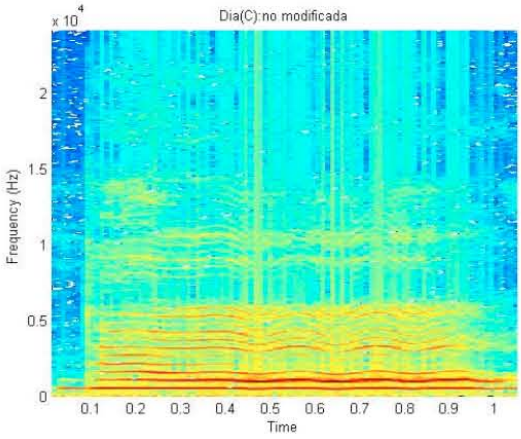

Hoy(C):tdpsola(p,48000,1.2,1.2,.1,.25)
x 10^4
Frequency (Hz)
Time

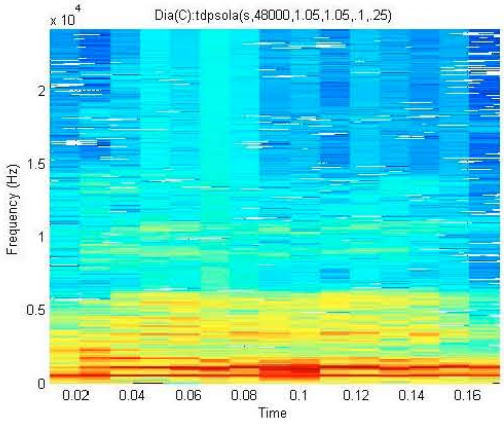

Por(C):no modificada
x 10^4
Frequency (Hz)
Time

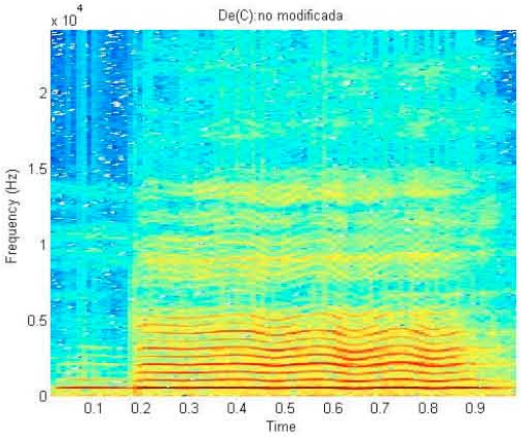

Por(C):tdpsola(q,48000,1.2,1.2,.1,.25)
x 10^4
Frequency (Hz)
Time

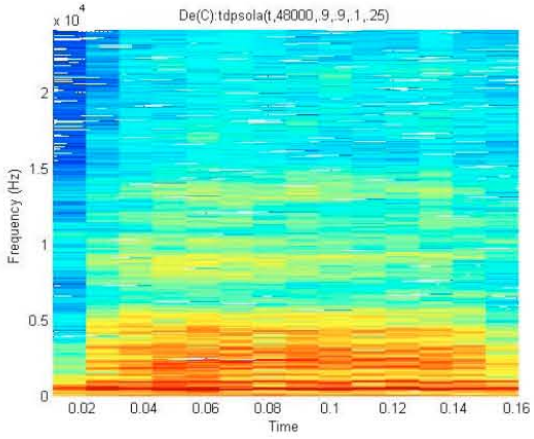

Ser(C):no modificada
x 10^4
Frequency (Hz)
Time

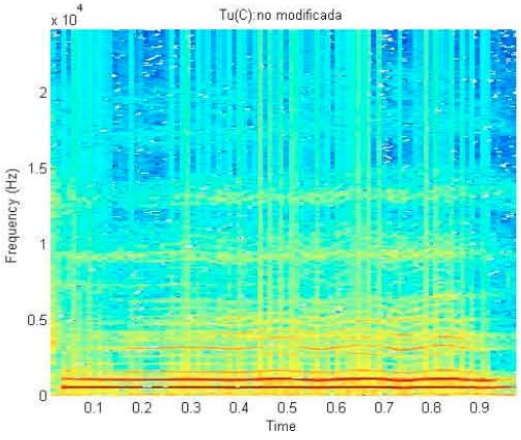

Ser(C):tdpsola(r,48000,1.15,1.15,.1,.9)
x 10^4
Frequency (Hz)
Time

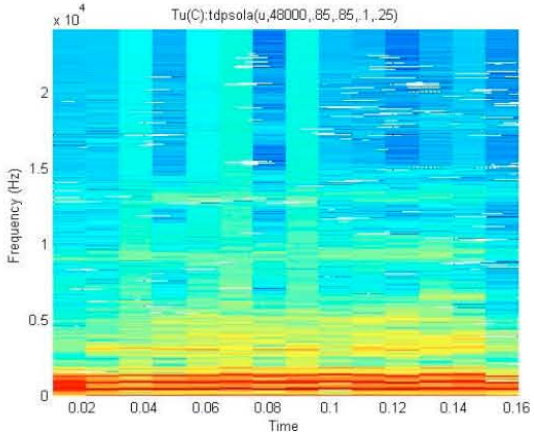

Dia(C):no modificada
x 10^4
Frequency (Hz)
Time

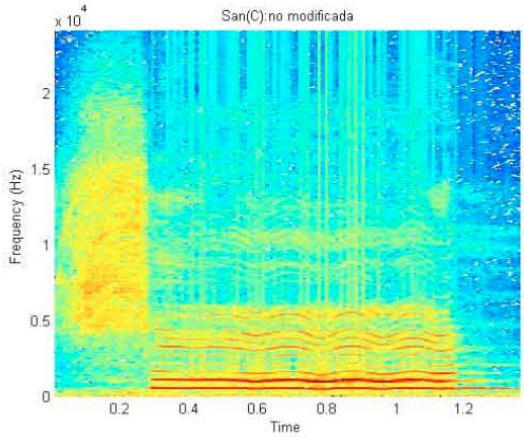

Dia(C):tdpsola(s,48000,1.05,1.05,.1,.25)
x 10^4
Frequency (Hz)
Time

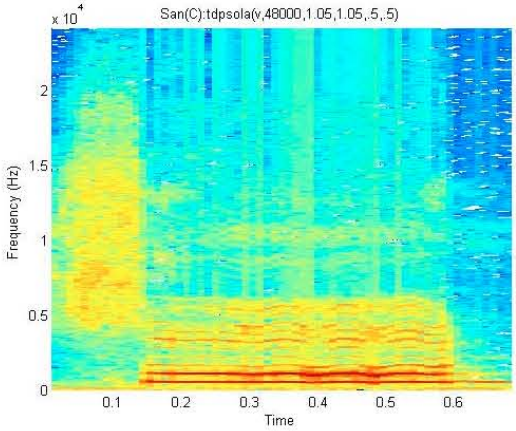

De(C):no modificada
x 10^4
Frequency (Hz)
Time

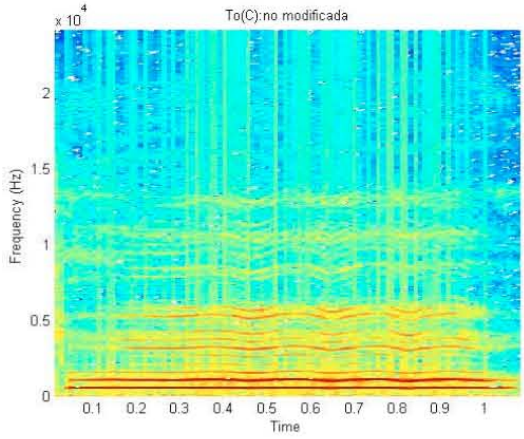

De(C):tdpsola(t,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

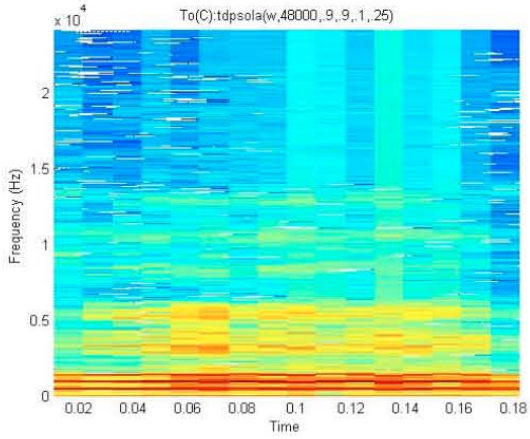

Tu(C):no modificada
x 10^4
Frequency (Hz)
Time

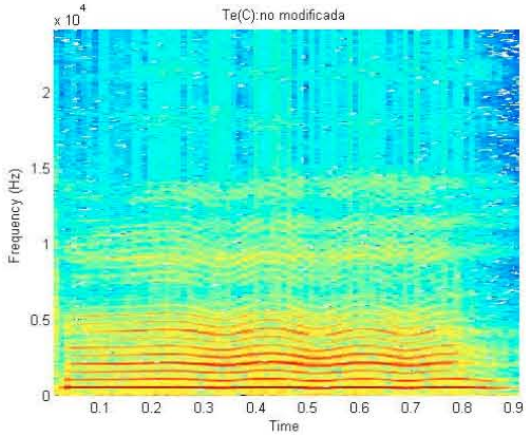

Tu(C):tdpsola(u,48000,.85,.85,.1,.25)
x 10^4
Frequency (Hz)
Time

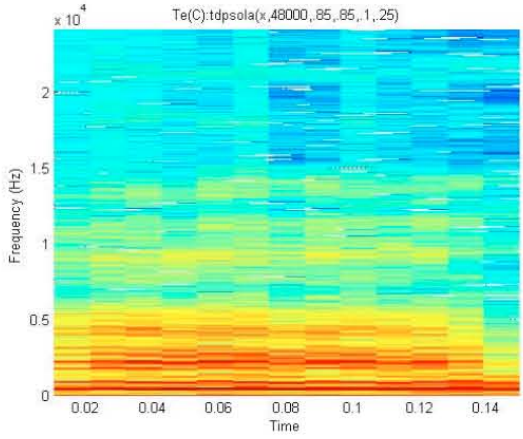

San(C):no modificada
x 10^4
Frequency (Hz)
Time

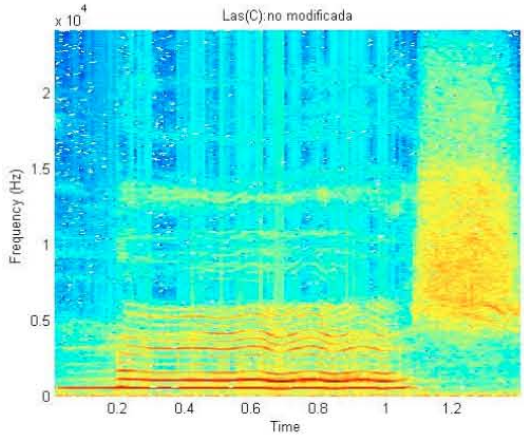

San(C):tdpsola(v,48000,1.05,1.05,.5,.5)
x 10^4
Frequency (Hz)
Time

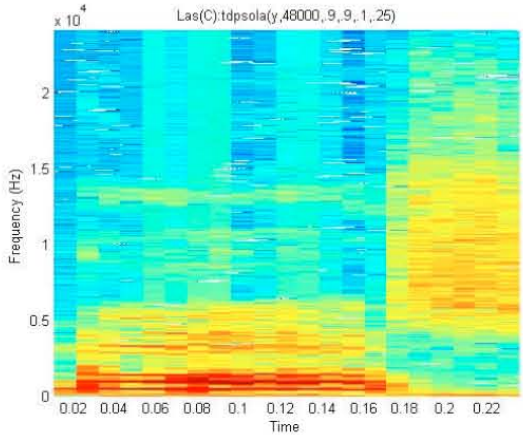

To(C):no modificada
x 10^4
Frequency (Hz)
Time

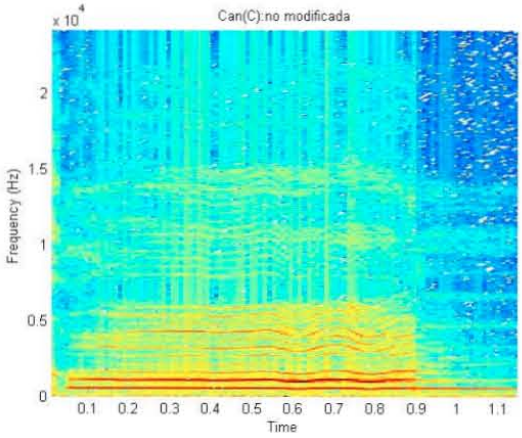

To(C):tdpsola(w,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

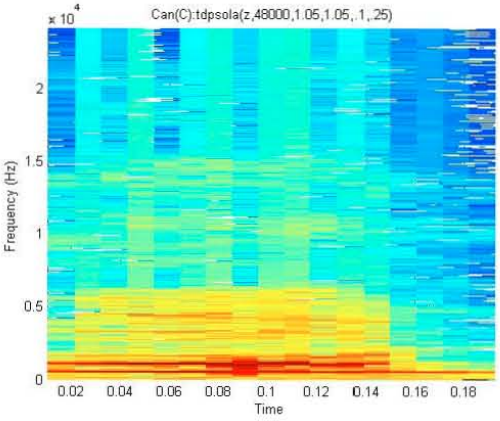

Te(C):no modificada
x 10^4
Frequency (Hz)
Time

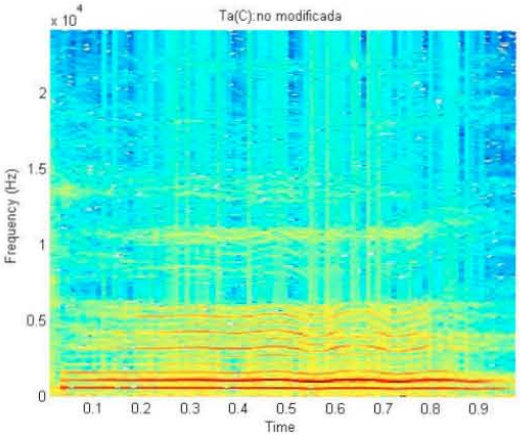

Te(C):tdpsola(x,48000,.85,.85,.1,.25)
x 10^4
Frequency (Hz)
Time

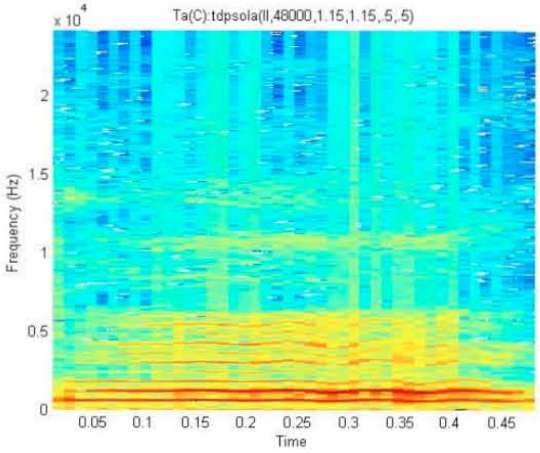

Las(C):no modificada
x 10^4
Frequency (Hz)
Time

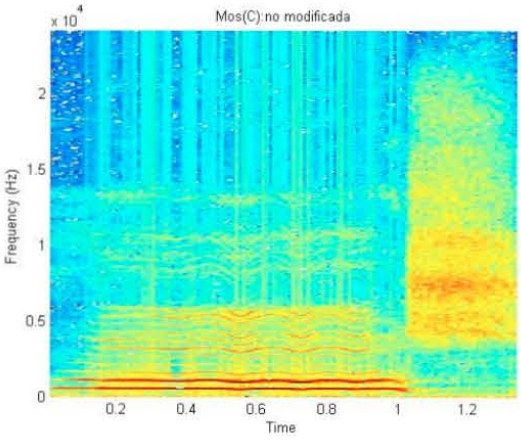

Las(C):tdpsola(y,48000,.9,.9,.1,.25)
x 10^4
Frequency (Hz)
Time

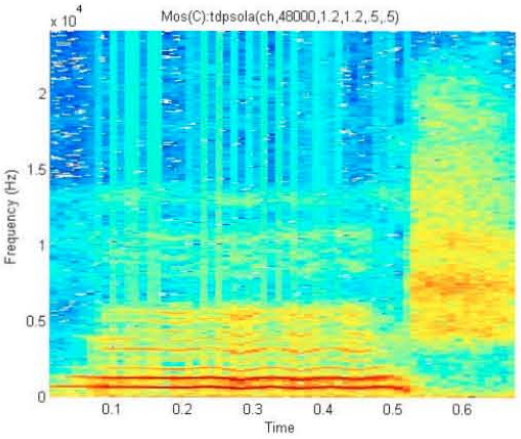

Can(C):no modificada
x 10^4
Frequency (Hz)
Time

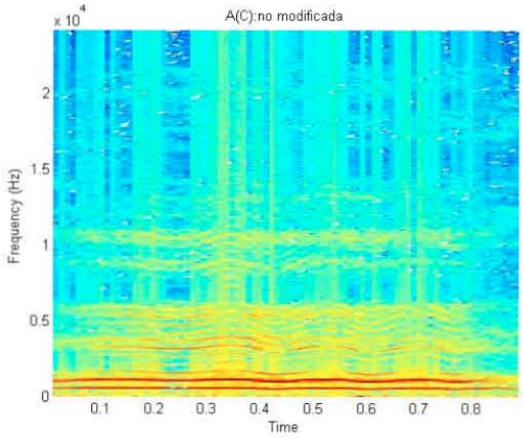

Can(C):tdpsola(z,48000,1.05,1.05,.1,.25)
x 10^4
Frequency (Hz)
Time

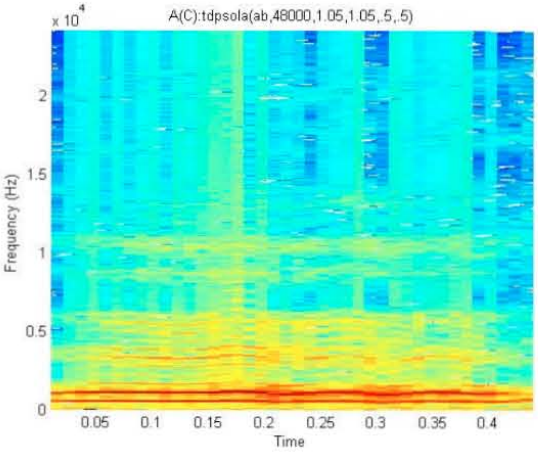

Ta(C):no modificada
x 10^4
Frequency (Hz)
Time

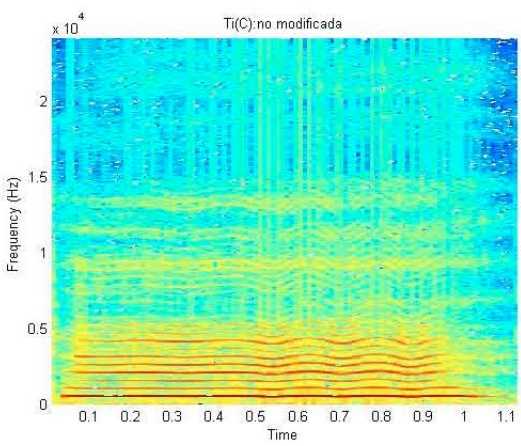

Ta(C):tdpsola(ll,48000,1.15,1.15,.5,.5)
x 10^4
Frequency (Hz)
Time

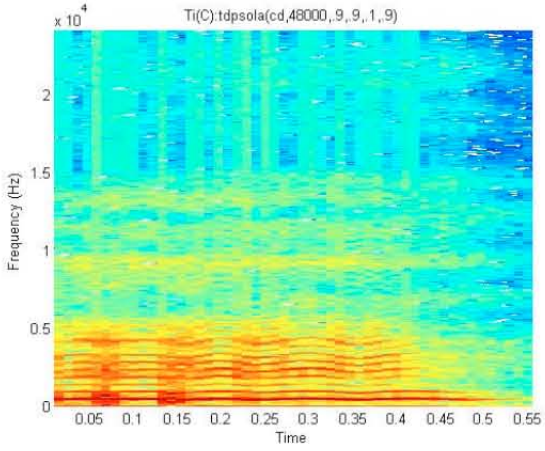

Mos(C):no modificada
x 10^4
Frequency (Hz)
Time


Mos(C):tdpsola(ch,48000,1.2,1.2,.5,.5)
x 10^4
Frequency (Hz)
Time


A(C):no modificada
x 10^4
Frequency (Hz)
Time


A(C):tdpsola(ab,48000,1.05,1.05,.5,.5)
x 10^4
Frequency (Hz)
Time


Ti(C):no modificada
x 10^4
Frequency (Hz)
Time
0
0.5
1
1.5
2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1


Ti(C):tdpsola(cd,48000,.9,.9,.1,.9)
x 10^4
Frequency (Hz)
Time
0
0.5
1
1.5
2
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

II. *Análisis de Resultados*

Frente a las muy variadas capacidades de la voz humana como instrumento musical, los objetivos de esta tesis son bastante modestos. Resulta difícil encontrar mecanismos para efectivamente evaluar los resultados debido a que no existen aún parámetros para hacerlo. Sin embargo, los resultados presentados, relativos a la variación del *tono* y *duración* de *sílabas* pre-grabadas y su aplicación sobre una canción popular, son suficientes para imaginar la amplia gama de posibilidades musicales explotables a través de esta línea de trabajo.

Los resultados muestran que puede lograrse una voz cantada en español bastante razonable en cuanto a lo que la *inteligibilidad* se refiere, con un mínimo de requerimientos y control. Lo anterior sugiere como un acierto el haber elegido la *sílaba* como la unidad fonética básica; sin embargo, la naturalidad de la voz resultante es, sin lugar a dudas, comprometida, como era predecible de la aplicación de un algoritmo de modificación del dominio del tiempo. Aún así, la severidad en esta alteración varía dramáticamente de intervalo a intervalo ya sea en frecuencia, o en duración, así como por la ejecución misma en las grabaciones de la base de datos. Así pues, la diferencia en naturalidad entre un segmento grabado con vibrato, comparado con otro sin vibrato es notoria, haciendo muy variable el intervalo en tiempo máximo que un segmento pueda ser variado sin agregar ruido. Se encontró que hasta una transposición de segmento de una cuarta justa es posible sin agregar errores digitales muy notorios. Considerando la explicación de la implementación del algoritmo TD-PSOLA, se encontró que una variación en la escala de frecuencias de 0.05, corresponde a una variación de un semitono. El *Sintetizador* en cuestión es capaz de realizar glisandos en caso de que se indique una variación en frecuencia inicial y final, sin embargo la tonada en cuestión no los requiere, por lo que no fueron probados a profundidad. El resultado sonoro de la primera versión (mananitas.m) es más parecido a una recitación que a una ejecución cantada, e incluye algunos errores de afinación en la voz, como puede observarse de los resultados presentados. Sin embargo, las pruebas subsiguientes han resultado mucho más satisfactorias, debido no solo a una revisión de la modificación frecuencial, sino a que, contrario a lo que se muestra en los resultados, los cambios temporales requeridos al *sintetizador* fueron regulares, es decir, se requiere que la modificación de la primera marca de periodo sea la misma que la última. Ello agrega mucha mayor naturalidad al resultado, haciendo mucho más melódica la ejecución reduciendo el efecto de *legato* que se obtenía en los primeros intentos. Ello sugiere que en efecto, el cantante trata, al menos temporalmente, a la sílaba como a una unidad del canto. Ningún mecanismo de articulación entre segmentos modificados fue implementado, mismos que son disparados a intervalos regulares. El escucha puede reconocer fácilmente la tonada que el sintetizador reproduce, lo cual es en sí mismo un resultado positivo, y una vez más, confirma el acierto de la elección de las sílabas, reafirmando que se está en un buen camino para generar un sistema de Síntesis de Voz Cantada en Español, con un vocabulario ilimitado.

III. *Conclusiones*

La síntesis efectiva de voz, está basada en una comprensión fundamental de la física de la producción de voz, así como de las constricciones linguísticas que caracterizan un lenguaje dado. La síntesis efectiva de voz cantada deberá seguir el mismo principio -entre mejor se comprendan los mecanismos físicos que generan y limitan la voz cantada, mejor se podrán reproducir- . Sin embargo el logro de la implementación de tales conocimientos resultan aún muy costosos en términos de tiempo y esfuerzo. La imitación del mundo material, al requerir una gran cantidad de sintonización manual de valores de parámetros, hace del *Modelo Físico* (como forma de síntesis), un problema difícil. Por otro lado, los métodos espectrales pueden producir una voz cantada convincente si no se requiere texto cantado, dejando de lado al universo de la voz cantada lírica o popular.

Sin ser realmente una técnica de *síntesis*, la aproximación de *sampler*, ha logrado satisfacer a un gran número de instrumentos. Sin embargo, la falta de flexibilidad y expresión, para el caso de instrumentos continuamente excitados (para los que los parámetros de control son numerosos y tienen muchas maneras de *atacar, articular,* o *tocar* cada nota), son los mayores problemas de este modelo, lo que ha significado que aún no se cuente con el nivel de calidad que un músico profesional espera que presente un instrumento. Para estos instrumentos, es factible llegar a un nivel aceptable de calidad, utilizando grandes bases de datos y *muestreando* una porción suficiente del *espacio sonoro* producido por un instrumento dado. El modelaje del *espacio sonoro* (producido por un intérprete con un instrumento), busca ser flexible en la elección de controladores de entrada, mientras simultáneamente busca la posibilidad de usar controles de alto nivel. A partir de un *espacio sonoro* dado, más los controles de entrada apropiados, el motor de síntesis debe ser capaz de generar cualquier trayectoria en el espacio, produciendo, por lo tanto, cualquier sonido contenido en él. La aproximación de la fuerza bruta es realizar un *muestreo* extensivo del espacio y realizar interpolaciones simples para moverse en él. Considerando las limitaciones en uno y otro sentido, esta aproximación requiere de la búsqueda de un equilibrio, estableciendo límites en las sesiones de grabación, pero considerando que el *espacio sonoro* de un intérprete refinado será mayor que el *espacio sonoro* de uno no refinado.

En el caso de la voz cantada, el espacio es tan inmenso y complejo que esta aproximación se queda lejos de cubrir una porción del espacio suficiente, por lo que es claro que el modelo de "*sampler*" sencillo no es completamente adecuado, ya que cierta parametrización de los sonidos es requerida. Para poder avanzar, es necesario entender las dimensiones relevantes del espacio y encontrar una parametrización con la cual se pueda mover en estas dimensiones por medio de interpolación, o transformación de sonidos existentes. Esta tesis, responde a ese llamado, específicamente para dos controles: el *tono*, y la *duración*. Como se dijo, la arquitectura generalmente propuesta para un sistema de síntesis de voz cantada, incluye una sección destinada a crear las trayectorias paramétricas que expresen apropiadamente los caminos dentro del *espacio sonoro* del instrumento, y un módulo que contenga el motor de síntesis que produce la señal de salida concatenando una secuencia de muestras transformadas que aproximen la trayectoria de ejecución. Al desarrollar síntesis de voz por concatenación se deben considerar los siguientes puntos: el tipo de segmento a utilizar (sílaba), el diseño de un inventario acústico, o conjunto de segmentos de voz (que para este trabajo respondía a la generación de dos melodías populares), la mejor selección de segmentos de voz de cierta base de datos, y una manera de alterar la prosodia de un segmento para que empalme mejor a una prosodia de salida deseada (algorítmo PSOLA de dominio en el

tiempo).

El objetivo de la modificación prosódica, es el cambio de *amplitud*, *duración* y *tono* de un segmento de voz. *Pausas*, *tono*, *duración relativa* e *intensidad*, son temas de interés tanto para la Síntesis de Voz Hablada, como para la Cantada. El *tono,* por ejemplo, puede ser entendido como señalización de niveles perceptuales de prominencia, o bien, como movimientos de la frecuencia fundamental en sílabas. El *tono* y la *duración* no son enteramente independientes, y muchos de los factores semánticos de alto nivel que determinan los contornos de *tono*, también pueden incluir los efectos de *duración*. No obstante, es una conveniencia técnica el presentar el análisis de las *pausas*, generación de *tono* y *duración* de manera separada. Debe mantenerse en mente que todas las cualidades de la prosodia mantienen una alta correlación entre ellas en el lenguaje humano. El efecto de la *intensidad* no es uno tan importante como los demás factores cuando se desea sintetizar el habla, por lo que su discusión no es abordada, además de que para sistemas basados en la concatenación sus efectos se encuentran en general incluidos en el segmento grabado del habla. Aunque en principio, en implementaciones de sistemas de voz, el módulo de atributos prosódicos aplica a todas las variables de la prosodia, es en su mayoría utilizado prácticamente para la generación de frecuencias fundamentales. El método PSOLA de dominio en el tiempo, realiza de manera adecuada pequeñas transformaciones de *tono* y *duración*, sin embargo para cambios más grandes, la calidad de la voz sufre. Asimismo, al ser la coherencia de fase uno de los temas centrales de la síntesis de voz, tanto hablada como cantada, se debe mencionar que el conjunto de algoritmos PSOLA presentan soluciones ingeniosas en ese respecto.

En términos de las limitaciones impuestas a la grabación de segmentos, además de los factores de procesamiento de señal (modificación de *tono* y *duración*), se atendió a aquellos de naturaleza acústica (co-articulación), y lingüística (como el espacio total de palabras que pueden generarse). La síntesis de voz a partir de la entrada de fonemas es una representación conveniente y muy económica en cuanto a almacenamiento. Entre cuarenta y cincuenta fonemas, más unos cuantos marcadores prosódicos, pueden ser codificados usando 6 bits y, con una tasa normal de locución de alrededor de 12 fonemas por segundo, la tasa de información es 72 bits/s. Por ende, sólo 8 bytes de memoria pueden almacenar aproximadamente 15 minutos de voz. Empero, la co-articulación en el habla, genera movimientos complejos de la frecuencia y los valores de amplitud de los formantes entre un sonido y el siguiente. Existen además, muchas variaciones acústicas para cada fonema (*alófonos*), que dependen del contexto o posición en una locución. Se deben escoger apropiadamente las unidades óptimas para formar la cadena, y ajustar adecuadamente a la prosodia deseada. Se debe arribar a una función objetiva que logre cierta calidad sonora y que permita escoger la mejor cadena. Para superar el difícil problema de escribir reglas para simular la co-articulación en la síntesis por fonemas, la mejor aproximación es el usar unidades fonémicas más largas. Al considerar que el *núcleo silábico* es siempre una vocal, con una altura tonal bien definida, y que la *sílaba* incluye el dominio de la co-articulación (por lo que sus características acústicas son más compactas y definen sin ambigüedad el inicio y fin de una nota) se eligieron éstas como unidad base. Sus propiedades ayudan a reducir el número de problemas surgidos de una metodología de concatenación, por lo que mejoran en la inteligibilidad de las voces logradas, lo que por definición implica cierto logro en cuanto a *naturalidad*. En el acercamiento *concatenativo,* cada segmento es completamente natural, por lo que tiene una mejor salida. A través de éste trabajo se ha constatado que las sílabas cumplen cabalmente con 3 de 4 de los requerimientos para escoger las unidades a grabar para

un sistema de voz cantada (una discusión más profunda debe sostenerse para definir si son *Entrenables* o no puesto que involucra una re-definición del concepto de *timbre*):

a) Generan una baja *distorsión de concatenación,* al ser segmentos más largos. El tener varias instancias por unidad (varias grabaciones de la misma *sílaba*), permite ofrecer elección en las instancias con más baja distorsión de concatenación.

b) Contienen una baja *distorsión prosódica*. La calidad de las unidades de cadena está típicamente dominada por las discontinuidades espectrales y tonales en las fronteras de las unidades. Éstas pueden ocurrir por diferencias en los contextos fonéticos, segmentación incorrecta, variabilidad acústica, o cambios en la prosodia. Sin embargo, las variaciones en prosodia, de la voz cantada, de uno a otra grabación, son menores que para la voz hablada. Un cantante con una voz entrenada, es capaz de minimizar tal variación, a menos de que se le requiera lo contrario durante la grabación. Las discontinuidades entre fronteras de unidades, también causan degradación; los errores de segmentación pueden causar discontinuidades espectrales, aún cuando tengan el mismo contexto musical. Asimismo, la variabilidad de una repetición a la siguiente causa discontinuidades pequeñas. Una unidad, grabada con vibrato, generalmente responde de manera distinta a una alteración de *tono* y *duración*, que no lo presenta. Además, condiciones distintas de grabación, como amplitud, tipo de micrófono o tarjeta de audio, pueden causar además discontinuidades espectrales. El resultado sonoro de este trabajo, presenta hasta cierto punto, discontinuidades en el color de segmento a segmento, por lo que un primer e importante paso para su mejora, es poner más atención en la base de datos, y por lo tanto a la sesión de grabación y ejecución del cantante.

c) Son *generalizables.* Algunos trabajos arrojan como resultados preliminares la cifra de 2, 240 *sílabas en español*, corrientes totales existentes; 242 sílabas existentes sólo atribuibles a morfemas verbales; 45 a plurales; 77 a palabras compuestas; 42 a nombres propios, apellidos, gentilicios y topónimos; y 12 a palabras latinas del DRAE (diccionario de la Real Academia Española).[2] Considerando que son 1000 los difonemas necesarios para lograr vocabulario irrestricto, en un Sistema de Voz Cantada, no parece fantasiosa la idea de grabar la totalidad de sílabas.

Uno de los mayores logros de el presente intento, es el descubrimiento de la enorme posibilidad que implica la aplicación de un algoritmo variable de modificación de marcas de periodo inicial y final del segmento a modificar, ya que ello significa cierto nivel de investigación sobre *acelerandos* y *retardandos* sintéticos, y su efecto directo sobre la *naturalidad*. Una modificación inmediata al Sistema, podría ser el sustituir la variación lineal, por una no-lineal, y escuchar los resultados. Por otro lado, una labor más minuciosa en cuanto al mecanismo de concatenación de los segmentos (es decir, un mecanismo para agregar o disminuir silencios inter-sílabas) y una investigación más amplia en cuanto a las combinaciones posibles de variabilidad de las marcas consecutivas de periodo, mejorarían extraordinariamente la *naturalidad* de la ejecución. Asimismo, una combinación de técnicas de síntesis, permitirían más opciones de control y las consecuentes mejoras en la naturalidad. Por ejemplo, el *esfuerzo* del cantante es modelado elegantemente por la técnica de *Sintesis FM*, por lo que un sistema híbrido podría generar una salida mejor. Una solución para la generalización de la partitura tradicional, es también deseable como siguiente paso para poder incluir cualquier información simbólica requerida para el control del sintetizador.

[2] Vease: http://www.cuadernoscervantes.com/art_36_silabas.html#

IV. *Recomendaciones para trabajo futuro*

Para completar el trabajo, el inventario acústico, o conjunto de segmentos de voz necesario (Cuáles, cuántos, etc) para cubrir en su totalidad al idioma español, está directamente relacionado con la cantidad de sílabas en el Español. Afortunadamente, trabajos en ese sentido ya se han elaborado y se tiene una base que, aunque es sustancialmente mayor que la de difonemas, es finita y es posible generarla. La base de datos a generar, deberá contener al menos dos grabaciones de cada sílaba, para tener la posibilidad de escoger la mejor selección de segmentos de voz, dadas las características fonéticas, en prosodia y melodía.

Como siguiente paso, el sistema de síntesis de voz, debe ser trasladado a una plataforma de utilización más "amigable" para su uso musical, como por ejemplo MAX/MSP, que permite además la construcción de nuevos objetos en los lenguajes de programación JAVA y Javascript, y que además presenta la gran ventaja de haber sido diseñado específicamente para la explotación musical de recursos computacionales y el manejo de señales en tiempo real.

La contribución más importante que esta línea de investigación de desarrollo puede aportar es el establecimiento de las diferencias fundamentales entre la voz cantada y hablada en español, y por tanto, un mayor conocimiento en el ámbito de la fonética. Si se estuviera interesado en continuar alguna de las ramas de investigación, encaminadas a establecer si es que un segmento fonético, o una combinación de ellos, es lo mejor para fines de construcción de un sistema *Partitura-a-Voz* (de *Síntesis por Concatenación*, con la mayor *naturalidad* posible de *Voz Cantada* en *Español*), entonces éste es un buen lugar para iniciar, ya que una gran variedad de trabajos de Maestría y Doctorado pudieran tener esta tesis como referencia inicial. Se debe, asimismo, contemplar la elaboración de una *Interfaz Sistema-Usuario* que simplifique el uso del sistema haciéndolo funcional y práctico.

En el arte de la escritura de música vocal resulta esencial tomar en consideración el sentimiento en la voz, el alcance a notas agudas, y concentrarse en experimentar la relación entre el texto y la melodía. Es especialmente importante considerar el grado de des-asociación de la melodía, ya que demasiados saltos pueden dificultar el canto o incluso imposibilitarlo. Se debe considerar si el vocalista tiene suficiente espacio para respirar entre frases, o si existen frase tan largas que no permiten al vocalista la respiración; además se debe considerar la amplitud del intervalo vocal, para quién está escrito, y si tal intervalo cambia demasiado rápido en la canción. Todos estos factores deben tomarse en cuenta para el desarrollo de un sistema comercial de síntesis de voz cantada.

Por último, la técnica y tipo de voz utilizada presentan otro ámbito al cuál debe ponerse atención. Para ello resulta pertinente tomar en consideración ciertos aspectos de la técnica vocal (Maragliano, 1993), (Mari, 1996); y los trabajos de Joe Wolfe y colaboradores (Joliveau, 2004) sobre las resonancias en el tracto vocal de una voz soprano. Asimismo, las aportaciones realizadas por el Dr. Eduardo Castro-Sierra a la Psicoacústica y fisiología auditiva y de la voz: su aplicación a la música y al canto (Castro-Sierra, 1994), y sus investigaciones relacionadas al análisis de las voces sopranos (Castro-Sierra, 2004) resultarán a largo plazo de mucha utilidad en el perfeccionamiento y naturalidad de la voz sintetizada resultante.

Apéndice I

Un sintetizador de voz de la UNAM

Estructura Básica

A continuación se presenta la estructura de un sintetizador de voz hablada en español, usando el método TD-PSOLA", desarrollado para la plataforma *Matlab* (.m), que constituye un sistema de texto-a-voz desarrollado en el posgrado de Ingeniería Eléctrica de la UNAM por Fernando del Río, en el año 2006, y por el cual obtuvo el grado de Maestro en Ingeniería, con la tesis intitulada: "Diseño de un sintetizador de voz en español usando el método TD-PSOLA". El código ha sido comentado a detalle, con excepción del módulo de traducción de números escritos a números hablados. En el apéndice II, se incluye el comentario del módulo encargado de calcular las marcas de tono (pitch_marks).

```
function y=habla(texto) %Inicia el programa: "habla" que contiene las siguentes 5 líneas
afonemas(texto) %llama la función "afonemas" cuya entrada es "texto"
y=avoz(afonemas(texto)); %introduce la salida de "afonemas" como entrada de la función "avoz"
y=y./max(y); %línea de normalización de la salida "y"
plot(y); % gráfica de la salida "y"
sound(y,11025); % Genera sonido con la salida "y"

%-----------------------------------%
%----CONVERSION TEXTO A FONEMAS----%
%-----------------------------------%
function y=afonemas(cadena) % Aquí inicia la función afonemas; el "texto" ingresado es la "cadena"
cadena=strcat(cadena,'-'); %Agrega un guión al final de la "cadena"
cadena=lower(cadena); % convierte la "cadena" de entrada en una de minúsculas
l=length(cadena); %calcula la longitud "l" de la "cadena" tomando en cuenta espacios y guión
ctemp=''; %define "ctemp" como variable, está vacía al principio
salida=''; %define "salida" como variabñe, está vacía al principio
tipo=0; %define "tipo" como variable, y es igual a cero al principio
for i=1:l %se inicia un ciclo en el cuál "i" toma valores iniciando en 1, y termina cuando se alcanza la longitud "l"
   l1=cadena(i); %cada entrada "i" de la "cadena" (es decir, cada letra o espacio) se asigna como "l1". Éste irá cambiando
   l2=''; %se define "l2" como variable, esta vacía al principio
   l3=''; %se define "l3" como variable, está vacía al principio
   if(i<l) %en caso de que "i" no haya alcanzado a "l" (al final de la cadena)---
      l2=cadena(i+1); %se define "l2" como la entrada "i+1"
   end %fin del "if" (i<1)
   if(i<l-1) %en caso de que "i" aún no haya alcanzado a la penúltima entrada---
      l3=cadena(i+2); %se define "l3" como la entrada "i+2"
   end %fin del "if" (i<l-1)
   if (( (l1>='a' & l1<='z') | (l1>='á' & l1<='ú') | l1=='ñ' | l1=='ü') & (tipo==1 | tipo==0))
% en caso de que "l1" esté entre "a" y "z", ó entre "á" y "ú", ó entre "ñ" y "ü" (listados de letras y sus tipos), y la variable "tipo" sea 1 o 0 (redundante porque siempre será así al inicio)
      ctemp=strcat(ctemp,l1); %entonces se re-define "ctemp" como la concatenación de "ctemp-l1" (al principio "ctemp" está vacío y solo es "l1", luego se suman las "l1" del ciclo)
      tipo=1; %y se define el "tipo" como 1 (palabra)
   elseif ( (l1>='0'& l1<='9') & (tipo==0 | tipo==2) ) %por otro lado, en caso de que "l1" esté entre 0 y 9, y el "tipo" sea 0 o 2 (redundante también)
      ctemp=strcat(ctemp,l1); %entonces se re-define "ctemp" como la concatenación de "ctemp-l1",
      tipo=2; %y se define el "tipo" como 2 (número)
   elseif ((l1=='.' | l1==',') & ( l2>='0' & l2<='9') & tipo==2) %por otro lado, si "l1" es una puntuación, y "l2" está entre 0 y 9, y es del "tipo" 2
      ctemp=strcat(ctemp,l1); %se re-define "ctemp" como la misma concatenación de antes, pero no el "tipo"
   else % si ha termina éstos casos, se examina
      if (tipo==1) % si es del "tipo" 1
         salida=strcat(salida,palabra(ctemp)); %se envía "ctemp" como entrada de la función "palabra" y se re-define la "salida" como la concatenación de la salida de "palabra" y "salida" (originalmente vacía), de aquí debe verse
"palabra"
      end %fin del "if" (tipo==1)
      if (tipo==2) % si es del "tipo" 2
         salida=strcat(salida,numero(ctemp)); %se hace lo mismo, pero enviando "ctemp" a la función "número"
      end %fin del "if" (tipo==2)
      ctemp=''; %ésta orden vacía el "ctemp" cuando se llega a un espacio lo que hace que reinicie el ciclo y renueve el envío a "palabra" y "número"
      tipo=0; %se vuelve a definir el "tipo" como 0 (para reiniciar el ciclo)
      if((l1>='a' & l1<='z') | (l1>='á' & l1<='ú') | l1=='ñ' | l1=='ü' )
         ctemp=l1; %ésta sección repite la clasificación en palabras y números.
         tipo=1;  %es redundante y considerando que la sección de números no será utilizado
      end
      if(l1>='1' & l1<='9')
         ctemp=l1;
         tipo=2;
      end
   end %fin del "if" (( (l1>='a' & l1<='z') | (l1>='á' & l1<='ú') | l1=='ñ' | l1=='ü') & (tipo==1 | tipo==0))
end %fin del "for" (i=1:l)
y=salida; %se define "y" como lo que contenga "salida", que será el resultado de la salida de "palabra" y "número"

%-----------------------------------%
%----CONVERSION FONEMAS A VOZ----%
%-----------------------------------%
function y=avoz(salida) %Ésta función recibe la "salida" de la función "afonemas" (cadena de fonemas y casos especiales)
fentra=0; %se define la variable "fentra" como cero
j=length(salida); %se calcula el tamaño de la "salida" y se define como "j"
j=j-1; %se le resta una unidad a "j" (por el guión)
cadena=''; %se define "cadena", y se deja vacío
tam=0; %se define la variable "tam" como cero
for i=1:j %se inicia un ciclo para el índice "i" que va de 1 a "j"
p1=''; %se define "p1" como variable ahora vacía
if i>1 %si el índice "i" es mayor a 1 (es decir es ya el segundo)
   p1=salida(i-1); %entonces "p1" se define como la entrada i-1 (o la anteriór a "i")
end %fin del "if" (i>1)
l1=salida(i); %se define "l1" como la entrada "i" de  "salida" (o la cadena de entrada) a la función
l2=salida(i+1); %se define "l2" como la entrada "i+1" de "salida"
l3=''; %se define "l3" como variable vacía actualmente
if i+2<j %si "i+2" es menos a "j" (max)
   l3=salida(i+2); %entonces "l3" se define como la entrada (i+2) de "salida"
```

```
end %fin del "if" (i+2)<j
ctemp=l1; %se define "ctemp" como "l1" (que va cambiando)
ctemp2=l2; %se define "ctemp" como "l2" (que va cambiando)
switch l1 %aquí se revierten los casos especiales sobre "l1" pero a través de una variable temporal (ignorar ésta sección hace que la traducción no sea precisa)
case 'C' %en caso de que se identifique "l1" como "C"
ctemp='ch'; %se hace "ctemp" (al final "l1") como "ch"
case 'L' %en caso de que se identifique "l1"  como "L"
ctemp='y'; %se hace "ctemp" (al final "l1") como "y"
case 'R' %en caso de que se identifique "l1" como "R"
ctemp='rr'; %se hace "ctemp" (al final "l1") como "rr"
case 'S' %en caso de que se identifique "l1" como "S"
ctemp='sh'; %se hace "ctemp" (al final "l1") como "sh"
end %fin del "switch" (casos) sobre "l1"
switch l2 %se cubren los casos sobre "l2" a través de una variable temporal para completar los casos especiales
case 'C' %en el caso de que se identifique "l2" como "C"
ctemp2='ch'; %se hace "ctemp2" (al final "l2") como "ch"
case 'L' %en el caso de que se identifique "l2" como "L"
ctemp2='y'; %se hace "ctemp2" (al final "l2") como "y"
case 'R' %en el caso de que se identifique "l2" como "R"
ctemp2='rr'; %se hace "ctemp2" (al final "l2") como "rr"
case 'S' %en el caso de que se identifique "l2" como "S"
ctemp2='sh'; %se hace "ctemp2" (al final "l2") como "sh"
end %fin del "switch" (casos) sobre "l2"

switch p1 %aquí se revisan los casos sobre "p1" y se aplican los cambios a la misma variable "p1"
case 'C' %en el caso de que se identifique "p1" como "c"
p1='ch'; %se hace "p1" como "ch"
case 'L' %en el caso de que se identifique "p1" como "L"
p1='y'; %se hace "p1" como "y"
case 'R' %en el caso de que se identifique "p1" como "R"
p1='rr'; %se hace "p1" como "rr"
case 'S' %en el caso de que se identifique "p1" como "S"
p1='sh'; %se hace "p1" como "sh"
end %fin del "switch" (casos) sobre"p1"

switch l3 %aquí se revisan los casos sobre "l3" y se aplican los cambios a la misma variable "l3"
case 'C' %en el caso de que se identifique "l3" como "C"
l3='ch'; %se hace "l3" como "ch"
case 'L' %en el caso de que se identifique "l3" como "L"
l3='y'; %se hace "l3" como "y"
case 'R' %en el caso de que se identifique "l3" como "R"
l3='rr'; %se hace "l3" como "rr"
case 'S' %en el caso de que se identifique "l3" como "S"
l3='sh'; %se hace "l3" como "sh"
end %fin del "switch" (casos) sobre "l3". Se entiende que cada caso se revisa para "l1","l2","l3", y "p1" ya que entra la información por cuartetos (repetida)

fsale=modifica(p1,ctemp,ctemp2,l3,fentra(length(fentra))); %se define como "fsale", la salida de la función "modifica", cuyos parámetros se especifican. La salida es ya numérica
fl=length(fsale); %se define "fl" como la longitud (en muestras) de "fsale"
if fl>1 %si "fl" es mayor que uno
fentra=cat(1,fentra,fsale); %entonces se define "fentra" como la concatenación de los arreglos numéricos 1,fentra y fsale (suma?)
tam=tam+fl; %se redefine "tam" como el valor anterior más "fl" (lo que lo haría de la longitud de "fsale" al principio)

else %de lo contrario (es decir si es vacío)
fsale=modifica(p1,ctemp,'-',ctemp,fentra(length(fentra))); %sustituye a la entrada de la función "modifica", "ctemp2" por "-"
fl=length(fsale); %vuelve a definir "fl" como la longitud de la salida "fsale"
fentra=cat(1,fentra,fsale); %entonces se define "fentra" como la concatenación de los arreglos de numéricos 1, fentra y fsale (vector)
tam=tam+fl; %se redefine "tam" como el valor anterior más "fl"
fsale=modifica(p1,'-',ctemp2,l3,fentra(length(fentra))); %y así sucesivamente para "ctemp"
fsale=openpcm(strcat('-',ctemp2,'.pcm'));
fl=length(fsale);
fentra=cat(1,fentra,fsale);
tam=tam+fl;
end
end
y=fentra;


function y=modifica(p1,l1,l2,s1,ant)
% Los Parametros de entrada de ésta función son:
% p1 - signo del fonema previo a difonema a modificar. Se usará el ej mama,en ma,a-  seria la a su fonema es a,
% sin guión (slatch)pues se espera un solo alfanumerico
% l1 y l2 - Primer y segundo fonemas de difonema a modificar, por ej mama,
% en -m,ma serian m y a, un solo alfanumerico
% s1 - Siguiente fonema de difonema a modificar
% ant - Ultima muestra de señal (su valor más no su posición) de salida generada
% hasta el momento, en -m,ma seria la muestra hasta la m.
% Parametros de salida:
% y - Difonema actual modificado, por ejemplo en -m, ma da el valor de
% muestras del difonema ma

dif1=1;
dif2=1;


%

% se debe leer marcas de periodo (en disco) de difonema previo (éste no se modifica, sino el siguiente igualándolo al primero):
% las marcas se generan en el programa "crea_pmarks" que está en misma carpeta, en -m,ma daria las marcas
% de -m .
man=openpcm(strcat(p1,l1,'.pm')); %Tomará el archivo p1l1.pm, que en el ejemplo seria -m.pm

if length(man)<2 % si no existe por alguna razon (debe ser de al menos 2 entradas), lee un difonema similar (inicio de segundo fonema)
% aunque en el ejemplo omite el archivo, no truena el programa , pero para ma leeria -m
    man=openpcm(strcat('-',l1,'.pm')); %ésta línea abre las marcas de la letra siguiente
% por ejemplo en mamacita al estudiar ci,it se iria a -i de cualquier palabra
end %fin del "if" (length (man)<2)


% ahora debe leer las marcas de periodo del difonema a modificar, en: -m,ma leeria: ma.pm
mac=openpcm(strcat(l1,l2,'.pm')); %una vez más se envía para abrir las marcas del fonema

% desde 181 185 es basura (¿y porqué no se eliminó?)
% leer  marcas de periodo de  de difonema siguiente de disco
% [Ojo: Esta parte no se utiliza en la salida]
msi=openpcm(strcat(l2,s1,'.pm'));
if length(msi)<2
% si no existe leer difonema similar (final de segundo fonema)
   msi=openpcm(strcat(s1,'-.pm'));
end

%si se pudieron obtener  marcas de periodo de difonema anterior se debe obtener el porcentaje de diferencia de la última y
%la primera muestra (punto máximo de señal), en -m,ma seria ultimo pitch de: -m
%y el primero de: ma
if length(mac)>1 & length(man)>1 %si se pudieron obtener  marcas de periodo de difonema anterior se debe obtener el porcentaje de diferencia de la última (---¿Marca?)  y
   a=man(length(man)-1)-man(length(man)-2); %¿Cómo es que se calcula el procentaje de diferencia explicado en 187-188?
```

```
    b=mac(3)-mac(2); %--------Explicar líneas 191, 192 y 193, -------me parece que "length(man)-1" es la penúltima muestra o periodo?...
    if a>0 %----si "a>0" significa que es creciente la forma de onda?
% a es la longitud del ultimo periodo en -m,ma seria de -m, tampoco usa el ultimo periodo (-----¿porqué?)
% b es la longitud del primer periodo de ma
% no usa mac(1) porque podria tener basura o ruido -----(explíquenme porqué?
        dif1=b/a; %diferencia porcentual de los ¿periodos? penúltimo y segúndo
    end %fin del if a>0
end %fin del "if length (mac...)"

% De 202 a 211 es basura. (--------¿Y porqué no lo borran?)

%si se pudieron obtener  marcas de periodo de difonema siguiente obtener % de diferencia longitud de ultima y
%primera marca (punto maximo de señal) [Ojo: Esta parte no se utiliza en la salida]
if length(mac)>1 & length(msi)>1 %si se pudieron obtener  marcas de periodo de difonema siguiente obtener % de diferencia longitud de ultima y
    a=msi(length(msi)-1)-msi(length(msi)-2); %-----no entiendo diferencias de muestras y marcas
    if a>0

  dif2=(mac(3)-mac(2))/a;
end
end
% leer difonema a modificar de disco, en ej leeria el valor de muestras de ma
fsale=openpcm(strcat(l1,l2,'.pcm'));

% esta es para asegurar que lo haya leido, aqui quedamos 17/6/09
fl=length(fsale);
% si el difonema existe
if fl>1
    % modificar  con la funcion tdpsola el difonema
    % para que coincidan el primer ciclo con el
    % ultimo ciclo del difonema anterior
    % La modificacion de frecuencia se va distribuyendo
    % de forma lineal a lo largo del 2o difonema
    % para que la parte final no sea modificada
    fsale=tdpsola(fsale,11025,dif1,1,1,1);
    % fsale contiene todas las muestras del segundo difonema, el indice
    % inicia en uno,
    % 11025 porque es la tasa de muestreo
    % dif1 es el porcentaje en frecuencia de cambio del primer ciclo -----Este paso es vital, aparentemente iguala frecuencias, pero ¿COMO?---- para
    % que coinicida con la frecuencia del ultimo ciclo del difonema ---¿%?
    % anterior
    %1: para que no se modifique el ultimo ciclo del segundo difonema, -----¿en qué se modifica? puede ser cualquier procentaje de .5 a 2
    %1: para que no se modifique la velocidad al inicio del 2o.difonema,
    % esto es varia la frecuencia, -------(ok, pero de donde a donde puede variar?) pero conserva el mismo No de muestras
    %1: para que no se modifique la velocidad al final de 2a difonema
    % hasta aqui ya tienen la misma frecuencia -----¿qué y qué, el 1ero y el 2o?, pero falta emparejar
    % amplitudes, modficando ahora todas las muestras del 2 difonema
    % fsale de regreso, contiene las muestras del 2o difonema
    a=find_pmarks(fsale,11025); % obtiene las nuevas maracas de periodo
    % Eliminar orillas de la señal modificada que pueden contener basura.
    % quita del 2o difonema primer y ultimos ciclos
    fsale=fsale(a(2):a(length(a)-1));
    if ant>0
    % ant es la amplitud de la ultima muestra del 1er difonema
    % si la señal no comienza en silencio,
    % escalar todo 2o difonema para que sea proporcional en amplitud
    d=fsale(1)/ant;
    fsale=fsale.*d;
    % lo hace porque asi se escucha mejor la frase de 1er difonema y 2o
    % difonema
    end
end
% si f1 no es mayor a uno, f1 es cero y no puede hacer nada
%regresar difonema modificado
y=fsale;

%y es el 2o difonema (muestras)
% si f1 esta vacio habra un silencio en la señal de salida, a veces falto.
% suena mal en ese momento

%----------------------------%
%----ANALISIS POR PALABRA----%
%----------------------------%
function y=palabra(p)%una vez que sale una palabra del tipo=1, se envía de la función "afonemas" a aqui convirtiendose la entrada en "p"
p=lower(p); %se convierte a minúsculas toda la cadena
l={'a' 'be' 'ce' 'de' 'e' 'efe' 'ge' 'ache' 'i' 'jota' 'ka' 'ele' 'eme' 'ene' 'o' 'pe' 'ku' 'erre' 'ese' 'te' 'u' 'be' 'doble-u' 'equis' 'i' 'zeta'}; %un listado de las letras en caso de que se quieran leer de manera singular
t=length(p); %se calcula el tamaño de la palabra "p"
if t==1 %en caso de que "t" (tamaño) sea 1, entonces
p=lower(p)-96; %una vez más se convierte a minúsculas, el número tendrá que ver con un protocolo de definición de fonemas
y=strcat('-',l(p),'-'); %se concatenan con dos guiones (al inicio y al final), la letra asociada al listado de la línea 267
else %en caso contrario ("t" no vale 1)
y='-'; %se define "y" como un guión (al que se le concatenará algo)
b=0; %se define "b" como 0
for a=1:t % se inicia un ciclo en el que "a" es el índice que corre de 1 al valor máximo de "t"
    if b==1 %en caso de encontrar en algun punto del ciclo que "b" es 1,
        b=0; %se coloca en cero de nuevo. "b" servirá como bandera de algún evento
    else %de lo contrario (si "b" no es 1)
l1=p(a); %se define "l1" como la entrada "a" de "p", debido a que "a" cambia, entonces "l1" lo hará también
l2=' '; %"l2" se define como vacío (hasta ahora)
l3=' '; %"l3" se define como vacío
p1=' '; %"p1" se define como vacío
if (a<t) % en caso de que el índice "a" en el ciclo aún no haya llegado a "t" (valor máximo)
    l2=p(a+1); %entonces se define "l2" como la entrada (a+1) de "p" (palabra)
end %fin del "if" (a<t)
if (a<t-1) %en caso de que el índice "a" aún no haya llegado a la penúltima entrada
    l3=p(a+2); %entonces se define "l3" como la entrada (a+2) de "p" (palabra)
end %fin del "if" (a<t-1)
if(a>1)   %en caso de que el índice "a" sea mayor a 1 (es decir, desde 2)
    p1=p(a-1); %se define "p1" como la entrada (a-1) de "p" (palabra)
end %fin del "if" (a>1)
switch l1 %Ésta sección trata con casos especiales del idioma (español), sustituyendo la entrada "l1"
    case 'c' % el caso cuando se identifica a "l1" como "c"
    l1='k'; %por "k" (fonema)
    if (l2=='h') %y en caso de que "l2" fuera "h"
        l1='C'; %se sustituye por "C" que representará el fonema "ch"
        b=1 %y se cambia el valor de "b" a 1, como bandera de que algo ha cambiado
    elseif vdebil(l2) %en cambio si "l2" es una vocal debil
        l1='s'; %entonces se cambia "l1" por el fonema "s"
    end %fin del "if" (l2=='h')
case 's' %el caso cuando se identifica a "l1" como "s"
    if (l2=='h') %si "l2" es "h"
        l1='S'; %entonces se sustituye ("l1") por "S" que representará el fonema "sh"
        b=1 %y se cambia el valor de "b" a 1, como bandera de que algo ha cambiado
    end %fin del "if" (l2=='h')
case 'l' %el caso cuando se identifica a "l1" como "l"
    if (l2=='l') %si "l2" es otra "l"
```

```
        l1='L'; %entonces se sustituye ("l1") por L que representará el fonema "ll"
        b=1; %y se cambia el valor de "b" a 1, como bandera de que algo ha cambiado
    end %fin del "if" (l2=='l')
case 'r' %el caso cuando se identifica a "l1" como "r"
    if (l2=='r') %si "l2" es "r"
        l1='R' %entonces se sustituye ("l1") por R que representa el fonema "rr"
        b=1; %y se cambia el valor de "b" a 1, como bandera de que algo ha cambiado
    elseif (1-vocal(p1)) %por otro lado, si lo anterior no es cierto, y si "p1" NO es una vocal
        l1='R'; %entonces se sustituye ("l1") por R que representa el fonema "rr"
    end %fin del "if" (l2=='r')
case 'q' %el caso cuando se identifica a "l1" como "q"
    l1='k'; %entonces se sustituye ("l1") por k (fonema)
    if(l2=='u') %si "l2" es "u"
        b=1 %se cambia el valor de "b" a 1, como bandera
    end %fin del "if" (l2=='u')
case 'v' %en el caso cuando se identifica a "l1" como "v"
    l1='b'; %entonces se sustituye "l1" por el fonema "b" (español mexicano)
case 'z' %en el caso cuando se identifica a "l1" como "z"
    l1='s'; %entonces se sustituye a "l1" poe el fonema "s" (español mexicano)
case 'y' %el caso cuando se identifica a "l1" como "y"
    if(1-vocal(l2)) %si "l2" NO es una vocal
        l1='i'; %entonces se sustituye "l1" por el fonema "i" (no se hace referencia al caso en el que "y" se conporta como "ll")
    end %fin del "if" (1-vocal(l2))
case 'g' %el caso cuando se identifica a "l1" como "g"
    if (l2=='e' | l2=='i') %si "l2" es "e" ó "i"
        l1='j'; %entonces se sustituye a "l1" como "j"
    elseif (l2=='u' & (vdebil(l3))) %por otro lado, si lo anterior no es cierto y "l2" es "u" y "l3" una vocal debil
        b=1; %se cambia el valor de "b" a 1, como bandera pero se mantiene el fonema igual. (misterio cómo elimina la "u")
    end %fin del "if" (l2=='e' ...)
case 'ü' %en el caso cuando se identifica a "l1" como "ü"
    l1='u' %entonces se sustituye a "l1" por "u"
case 'x' %en el caso cuando se identifica a "l1" como "x"
    y=strcat(y,'k') %a "y" (salida) se le concatena un fonema "k"
    l1='s'; %y se cambia "l1" por "s"
case ' ' %en el caso de que la entrada esté vacía
    l1='-'; %entonces se sustituye "l1" por "-"
case 'h' %en el caso de que "l1" sea "h"
    l1=''; %entonces se sustituye "l1" por " "
end   % fin de los casos "switch"
y=strcat(y,l1); %se concatena el guión inicial de "y" con lo que termine siendo "l1"
end  %fin de b==1
end  % fin de for
y=busca_acento(y); %se envía "y" a la función "busca_acento"
end  % fin de length

%--------------------------%
%----ANALISIS DE NUMEROS----%
%--------------------------%
function y=numero(num)
num=strrep(num,',','');
b=length(num);
a=strfind(num,'.');
l1=num(1);
l2='0';
l3='0';
y='';
salida='';
if b>1
l2=num(2);
end
if b>2
l3=num(3);
end
if length(a) > 0
 a=a(1);
salida=strcat( numero(num(1:a-1)) ,'-púnto-', numero(num(a+1:b)) );
b=0;
elseif (b==1)
switch l1
case '1'
      salida='úno-';
case '2'
      salida='dós-';
case '3'
      salida='trés-';
case '4'
      salida='kuátro-';
case '5'
      salida='sínko-';
case '6'
      salida='séis-';
case '7'
      salida='siéte-';
case '8'
      salida='óCo-';
case '9'
      salida='nuéBe-';
case '0'
   salida='séro-';
end %fin case
elseif(b==2)
switch l1
case '1'
      if l2=='0'
      salida=palabra('diez');
      b=0;
      elseif (l2=='1')
      salida=palabra('once');
      b=0;
      elseif (l2=='2')
      salida=palabra('doce');
        b=0;
      elseif (l2=='3')
      salida=palabra('trece');
        b=0;
      elseif (l2=='4')
      salida=palabra('catorce');
        b=0;
      elseif (l2=='5')
      salida=palabra('quince');
        b=0;
      else
      salida='diesi';
      end
case '2'
```

```
        if l2=='0'
            salida=palabra('veinte');
        else
            salida='beinti-';
        end
case '3'
        if l2=='0'
            salida=palabra('treinta');
        else
            salida='treintai-';
        end
case '4'
        if l2=='0'
            salida=palabra('cuarenta');
        else
            salida='kuarentai-';
        end
    case '5'
        if l2=='0'
            salida=palabra('cincuenta');
        else
            salida='sinkuentai-';
        end
case '6'
        if l2=='0'
            salida=palabra('sesenta');
        else
            salida='sesentai-';
        end
case '7'
        if l2=='0'
            salida=palabra('setenta');
        else
            salida='setentai-';
        end
case '8'
        if l2=='0'
            salida=palabra('oenta');
        else
            salida='oCentai-';
        end
case '9'
        if l2=='0'
            salida=palabra('nobenta');
        else
            salida='nobentai-';
        end
if l2=='0'
    b=0;
end
end %fin case
elseif (b==3)
switch l1
case '1'
        if (l2=='0' & l1=='0')
        salida='sién';
            b=0;
        else
        salida='siento-';
    end
case '2'
salida='dosiéntos-';
case '3'
salida='tresiéntos-';
case '4'
salida='kuatrosiéntos-';
case '5'
salida='kiniéntos-';
case '6'
salida='seisiéntos-';
case '7'
salida='setesiéntos-';
case '8'
salida='oCosiéntos-';
case '9'
salida='nobesiéntos-';
end
if (l2=='0' & l1=='0')
    b=0;
end
elseif(b>3 & b<7)
        salida=strcat(numero(num(1:b-3)),'míl-');
b=4;
elseif(b>6 & b<13)
        salida=strcat(numero(num(1:b-6)),'míLónes-');
b=7;
else
salida='un-número-berdadéraménte-gránde-';
b=0;
end%fin if longitud
b=b-1;
if b>0
    c=length(num);
    salida=strcat(salida, numero( num(c-b+1:c) ));
end
    y=salida;

%---------------------------%
%----FUNCIONES DE APOYO----%
%---------------------------%
function y=nsv(a) % función donde se detectan las letras "n", "s" o Vocal en la terminación de la palabra. Llamada por primera vez desde la función "busca_acento". Es igual a "y" para decidir si es nula o no.
x=length(a); % se define "x" como la longitud de la palabra "a"
l1=a(x); %y a "l1" como la entrada "x" (es decir final) de la palabra "a".
if(l1=='n' | l1=='s' | vocal(l1) ) % si "l1" coincide con "n", "s", o "vocal"
y=1; % entonces "y=1" o es no nula la función
else % sino
    y=0; % "y=0" o la función es nula
end %fin de la función "nsv"

function y=vdebil(a) % entra "a" a la función "vdebil", es llamada por primera vez desde la función "palabra" (casos)
if (a=='e' | a=='é' | a=='i' | a=='í') % si la entrada "a" es ya sea "e" o "i" (vocales débiles), acentuada o no
        y=1; % entonces y=1, y va de regreso
else % si no...
        y=0; % y=0
end % fin...es como un switch
```

```
%{
function p1=quitar_acento(p,lugar)
l=length(p)
   if l>=lugar
      if (p(lugar)=='á')
         p(lugar)='a';
         elseif (p(lugar)=='é')
               p(lugar)='e';
         elseif (p(lugar)=='í')
               p(lugar)='i';
         elseif (p(lugar)=='ó')
               p(lugar)='o';
         elseif (p(lugar)=='ú')
               p(lugar)='u';
      end
   end
p1=p;
%}
%la función "quitar_acento" fue solo una buena intensión, ya que el
%sintetizador no corrige errores de acentuación gráfica, lo cuál es
%conveniente para el trabajo, ya que el canto debe respetar los acentos tal
%cual se ingresan

function p1=poner_acento(p,lugar) %función para poner acentos encontrados, recibe (salida, remember) o una palabra y el lugar de la acentuación. Llamada por primera vez desde la función "encontrar_acentuación" (linea 604)
l=length(p); %se define "l" como la longitud de la palabra "p"
   if l>=lugar %si "l" es más grande o igual que el "lugar" (como debería de ser)
      if (p(lugar)=='a') %si la entrada "lugar" de la palabra "p" es "a"
         p(lugar)='á'; %cambiar a "á"
        elseif (p(lugar)=='e') %sino, si la entrada "lugar" de la palabra "p" es "e"
         p(lugar)='é'; %cambiar a "é"
        elseif (p(lugar)=='i') %lo mismo de "i" a "í"
         p(lugar)='í';
        elseif (p(lugar)=='o') %lo mismo de "o" a "ó"
         p(lugar)='ó';
        elseif (p(lugar)=='u') %y de "u" a "ú"
         p(lugar)='ú';
      end %fin del "if" de las entradas "lugar" de "p"
   end %fin del "if" (l>=lugar)
p1=p; %se define "p1" como la palabra "p" (ya acentuada)

function y=palabra_acentuada(palabra) %función que localiza las palabras acentuadas, llamada por primera vez desde la función "busca_acento" (al final de la función "palabra")
l=length(palabra); %de define "l" como un el largo de la "palabra"
y=0; %se define "y=0" o nula, inicio siempre. ¿Tiene ventajas a veces inicializar "y" como "0" y otras veces no?
   for i=1:l %para un índice "i" que va de 1 a l (longitud de la palabra)
      if acento(palabra(i)) % se envían una a una las entradas de la "palabra" indexada por "i" si no es nula la función "acento(a)"
         y=1; %entonces "y=1"
      end %fin del "if"
   end %fin del "for"

function y=vocal(a) % entra "a" a la función "vocal", es llamada por primera vez desde la función "palabra" en los casos
   if (a=='a' | a=='á'  | a=='o' | a=='ó' | a=='u' | a=='ú' | a=='e' | a=='é' | a=='i' | a=='í') % si "a" es alguna vocal (acentuada o no)
       y=1; %entonces y=1
      else % si no...
         y=0; % y=0
   end %fin


function y=acento(a) %función que distinge acentos gráficos, llamada por primera vez desde la función "palabra_acentuada" (en funciones de apoyo)
if (a=='á' | a=='é' | a=='í' | a=='ó' | a=='ú') %si la entrada "a" (justo el índice "i") es alguna vocal acentuada gráficamente
      y=1; %entonces "y=1"
   else %en caso de que no...
      y=0; %entonces "y=0" y la función es nula
end %fin

%-----------------------------%
%----ENCONTRAR ACENTUACION----%
%-----------------------------%
function y=busca_acento(salida) %la entrada de ésta función será "y" (a la que llama salida) que viene de la función "palabra", y su resultado será también "y"
if palabra_acentuada(salida)==0 %si el resultado de la función "palabra_acentuada" es nulo, habiendo recibido la palabra de entrada de ésta función (en otras palabras: si no tiene acento escrito, pasa por aquí)
b=length(salida); %se define "b" como la longitud de "salida" que en éste caso es la "palabra"
found=0; %se define "found=0"
   if (nsv(salida)) %si la función "nsv" no es nula. Es importante notar que cada palabra por separado llega a ésta parte del programa, y después se pega en la salida con el resto
      for a=1:b % para un índice "a" que corre de 1 a la longitud máxima de la "palabra" o "b"
         l2=0; % "l2" se define como cero
         l3=0; % "l3" se define como cero
         p1=0; % "p1" se define como cero
         p2=0; % "p2" se define como cero
         l1=salida(a) % "l1" se define como la salida para el índice "a", donde sea que se encuentre. Esto hará que se genere un ciclo que acaba al término de la palabra

            if(a<b) % si el ínidice "a" aún no alcanza el valor máximo "b" o la longitud total de la palabra
               l2=salida(a+1) % entonces "l2" será la entrada "a+1" de "salida" o en éste caso la palabra.
            end %fin del "if" (a<b)

            if(a<(b-1)) %si el índice "a" aún no alcanza la penúltima entrada de "salida", "b-1"
               l3=salida(a+2) %entonces "l3" será la entrada "a+2" de "salida" o en éste caso la palabra
            end %fin del "if" (a<(b-1))

            if(a>1) %si el índice "a" es mayor que "1" (al menos segunda entrada)
               p1=salida(a-1) %entonces "p1" será la entrada "a-1" de "salida" o de la palabra, o el previo a "l1"
            end %fin del "if" (a>1)

            if(a>2) %si el índice "a" es mayor que "2" (al menos tercera entrada)
               p2=salida(a-2) %entonces "p2" será la entrada "a-2" de "salida" o de la palabra, o el previo a "p1"
            end %fin del "if" (a>2)

            if (vocal(l1) & found==0) % si "l1" es una vocal y el resultado de "found" es nulo (que aún no se utiliza, aún no se sabe qué hará y en un principio se define=0)
               remember=a; %se define "remember" como el índice "a", en donde éste se encuentre (1:b)y que será justamente "l1"
               look=0; %y se define "look=0" (no comprendo exactamente que cosa es "look", ya que cambia mucho)
               if((a<=(b-2)) & (l2=='u' | l2=='i')) %dentro del "if", si "a" es menor o igual a la antepenúltima entrada de "salida" y "l2" es "u" ó "i" ¿No entiendo la relevancia de ésto?
                  look=a+2; %entonces se define "look=a+2" (donde al máximo que llegará es a la última entrada de "salida" o la palabra
               elseif(a<b) %y si "a<b", que deja únicamente la posibilidad de que "a" sea el penúltimo
                  look=a+1; %entonces se define "look=a+1", con lo cuál llegará a la última entrada
               end %fin del "if" (()&())
            l1=0 %"l1" será cero. Con ello se cambió de "l1" a "remember" y "look" y se reinició "l1" (ello quiere decir que "l1" ya no es "salida(a)")

               if look<1  %ahora, si "look<1", es decir, si aún no se satisfacen las dos condiciones del "if" anterior ("l2=u,i), por lo que "look=0" todavía
                  look=1; %entonces hacer "look=1"
               end %fin del "if" (look<1)

               while(look<=b & (vocal(salida(look)))==0) %mientras "look<=b" y la entrada "look=#(donde se encuentre) no sea una vocal
                  look=look+1; %continua sumándole 1´s hasta que alguna de las condiciones se satisfaga
               end %fin del "while"

               if (look>0 & look<=b) %si "look" está entre "1" y "b"
```

```
                l1=salida(look) %"l1" será la entrada "look=#" de "salida"
              end %fin de (0<look<=b)

              if(look>=(b-1) & vocal(l1)) %si "look" es ya sea "b" o "b-1" y "l1=vocal"
                 found=1; %redefinir "found=1"
              end %fin del "if" (look>=(b-1) & vocal(l1))
            end %fin del "if" (vocal(l1) & found==0)
        end %fin del "for" (a=1:b)
    end %fin "if" (nsv(salida) no nula)
    if (found==0) %si "found=0"
       if (b==0 & vocal(salida(0))) %y si "b=0" y la entrada "0" de salida es una vocal???)
          remember=0; %entonces "remember=0"
          found=1; %y "found=1"
       end %fin del "if" (que no entiendo???) parece que ésto nunca va a pasar!!!
       if(b>0) %si "b>0"
          if (vocal(salida(b-1))) %y si la entrada "b-1" de "salida" (palabra) es una vocal
          remember=b-1; %entonces "remember=b-1"
          found=1; %y "found=1"
          elseif(b>1 & vocal(salida(b-2)) & (vocal(salida(b-1))==0 | salida(b-1)=='i' | salida(b-1)=='u') & vocal(salida(b))==0) %sino, si "b>1" y la entrada "b-2" de "salida" es una vocal, y si la entrada "b-1" de "salida" NO es una vocal ó la
entrada "b-1" de salida es "i" o "u", y la entrada "b" de "salida" no es vocal
          remember=b-2; %entonces "remember=b-2"
          found=1; %y "found=1"
          end %fin del "if" (vocal(salida(b-1)))
       end %fin del "if" (b>0)
    end %fin del "if" (found==0)
    if (found==1) %si "found=1"
       salida=poner_acento(salida,remember); %enviar a la función "poner_acento", que tiene por entradas "salida" y "remember"
    end %fin de "if" (found==1)
end %fin de "if" (palabra_acentuada (salida)==0)
y=salida; %"y" será la salida



%-------------------------------%
%----MODIFICACION PROSODICA----%
%-------------------------------%
function y = tdpsola(s,fs,pscale,pscale2,tscale,tscale2);
% Parametros de entrada
% s - 2o difonema se va a modificar
% fs- frecuencia de muestreo de la señal, 11025 Hz
% pscale - % a escalar la frecuencia al inicio del 2o difonema, ya caculado en
% modifica,
% pscale2- % a escalar la frecuencia al final del 2o difonema es 1
% tscale - % a escalar en duracion al inicio del 2o difonema, siempre es 1
% tscale2 - % a escalar en duracion al final de la señal, siempre es 1

pm=find_pmarks(s,11025); % son las mismas del inicio
  %ahora Calcula de posicion de la nuevas marcas de periodo, hasta la linea 758
pm_ps=pm;
pshift=0;
lp=length(pm);% es el numero de marcas

    for i = 1:lp % hasta 701
% obtener longitud actual de cada marca de periodo
       if (i>1)
       T0=pm(i)-pm(i-1);%por ejemplo es 210 menos 14 (ver hoja)
       else
       T0=pm(i)-0;% en la 1era marca T0 es 14
       end
       %obtener % de modificacion de la marca de periodo actual, a partir
       %de %inicial y final.(La variacion es de forma lineal)
       pin=pscale+ ((pscale2-pscale)*i/lp); % es el %de cambio a cada frecuencia
       pshift=pshift-round(T0*(1-(1/pin))); %aqui el ejemplo seria 210x.15, 290x.12
       %obtener longitud de la nueva marca de periodo a la salida,
       %longitud actual * % de modificacion
       pm_ps(i)=pm(i)+pshift; % es la muestra de nueva posicion de la marca
    end
    %obtener numero de marcas de periodo modificadas
    useds=zeros(1,length(pm_ps)-2);
    %Obtener numero de muestras a agregar o eliminar para primer periodo
    tin=tscale+ ((tscale2-tscale)/lp); % es siempre uno para este diseño, lo necseita el prog
    tot=tin*(pm(2)-pm(1))/(pm_ps(2)-pm_ps(1));% es la nueva longitud de toda difonema, en este diseño simepre es igual a la original
    %Para cada una de las marcas de periodo existentes
    for i = 1 : length(useds)
    %Ir aumentando el numero de muestras a agregar o eliminar para cada marca de
    % periodo
       tin=tscale+ ((tscale2-tscale)*i/lp); %sigue valiendo uno siempre
       avg=pm_ps(i+1)-pm_ps(i);%en el ejemplo es 224.7
       new_tscale=tin*(pm(i+1)-pm(i))/avg; %en el ejmplo seria (210-14)/224.7 = .872
       %Si se esta aumentando la duracion de la señal
       if (new_tscale>1)
       %Si el numero de muestras nuevas > duracion del periodo, duplicar
       %el periodo hasta que el numero de muestras requeridas < duracion
       %del periodo
          while(tot>1)
              useds(i)=useds(i)+1; % avisa que ya hay un sobrante en este ciclo, se incrementa cuando el sobrante es mayor o igual a la longitud del ciclo
              tot=tot-(pm_ps(i+1)-pm_ps(i))/avg; %ver figura de hoja 2
          end
       %Eliminar longitud de periodo del numero de muestras nuevas
       %requeridas
       tot=tot+new_tscale;
       else % cuando new_tscale es menor que 1 por ejemplo aqui es .87
       %Si se esta aumentando la duracion de la señal
          useds(i)=1;
       %Si el numero de muestras a eliminar > duracion del periodo,
       %eliminar periodos hasta que el numero de muestras a eliminar < duracion
       %del periodo
          while(tot<1)
              useds(i)=useds(i)-1;
              tot=tot+(pm_ps(i+1)-pm_ps(i))/avg;%
          end
       %Eliminar longitud de periodo del numero de muestras a eliminar
       %requeridas
       tot=tot-(1-new_tscale);
       end
  end
  start=1;
  count=1;
  %Para cada una de las ventanas existentes (de longitud 1 marca de
  %periodo)
  for i=1:length(useds)% hasta linea 758
   %Agregarla el numero de veces requeridas (0 si se elimino) (2 o mas si
   %se duplico)
     if (useds(i)>0)
       final(count,:)=[start pm(i) pm(i+2) 0]; % inicio ventana, pmarks asociadas,  invertir?, shift
```

```
      count=count+1;
      start=start+pm_ps(i+1)-pm_ps(i)+1;
    end
    for j=2:useds(i)% este caso es cuando se duplica un ciclo, cuando voy quitando muestras
      final(count,:)=[start pm(i) pm(i+2) mod(j,2)];
      count=count+1;
      start=start+pm_ps(i+1)-pm_ps(i)+1;
    end
  end
  % Obtener numero de periodos de la nueva señal modificada
  numfrm=size(final,1);
  %Obtener posicion de muestra inicial y final de la primera ventana
  A=final(:,1) + (final(:,3)-final(:,2)+1);
  ylen=max(A); %numero de muestras de la nueva señal
  y=zeros(ylen,1);% pone sus muestras en cero
  w=zeros(size(y));% otro vector auxiliar ahora en ceros
    %Para cada una de las nuevas ventanas
  for i = 1 : numfrm % hasta 788
    %Obtener posicion de muestra inicial y final precalculas
    start=final(i,1);
    %Obtener longitud de la ventana
    len=final(i,3)-final(i,2)+1;
    %Generar ventana de Hanning con longitud igual a la de la ventana
    wgt=hanning(len);
    %Extraer ventana
    frm=s(final(i,2):final(i,3));
    %Cuando se duplican ventanas, las copias pares (2,4,6..) se invierten
    %para evitar generar efectos de periodicidad en señales no periodicas
    if (final(i,4))
      frm=wrev(frm);
    end
    %A la señal de salida generada hasta el momento sumar la nueva
    %ventana multiplicada por la ventana de Hanning.
    %Esta señal puede traslaparse con muestras anteriores.
    y(start:start+len-1)=y(start:start+len-1)+frm.*wgt;
    %Calcular factor de normalizacion para la ventana calculada
    %(para disminuir la distorsion generada por la ventana de Hanning)
    w(start:start+len-1)=w(start:start+len-1)+wgt;
  end
    %Para cada muestra%
    for i=1:ylen
      if w(i)==0
        w(i)=1;
     end
      %Dividir la muestra por el factor de distorsion si este es
      %diferente de 0.
      y(i)=y(i)/w(i);
    end

    function pitch_marks = find_pmarks(speech, fs_in)
%
% function pitch_marks = find_pmarks(speech)
%
%
% This MATLAB function calculates and returns the pitch marks (placed at
% peaks in the short-time energy function) for the input speech, that is
% assumed to be sampled at 8 KHz.
%
%         speech:    the input speech data
%
% W. Goncharoff (goncharo@ece.uic.edu)
% 06/12/97
%
% Modified for any sampling rate
% Oytun Turk - 01.11.2002
%_____________________________________________________________________


p1 = round(fs_in/400);
p2 = round(fs_in/60);

spch = speech(:)';
xsamp = length(spch);
N=1:xsamp;

%calculate the approximate pitch contour based on energy peaks:
wlen = round((p1+p2)/3);
ecurve = conv(hanning(wlen)',spch.^2);
ecurve = conv(hanning(wlen)',ecurve);
ecurve = ecurve((1:xsamp)+wlen);

peaks = ([0,diff(ecurve)]>0) & ([diff(ecurve),0]<0);
index = 1:xsamp;
index(~peaks) = [];
Npeaks = length(index);
pitch = diff(index);
pitch1 = [pitch, pitch(Npeaks-1)];
pitch2 = [pitch(1), pitch];
mat_row1 = max(1,min(p2-p1+1,pitch1-p1+1));
mat_row2 = max(1,min(p2-p1+1,pitch2-p1+1));
z1 = ecurve([index(2:Npeaks),index(Npeaks)]);
z2 = ecurve([index(1),index(1:(Npeaks-1))]);


step_size = round(fs_in/192);
Nbatch2 = ceil(xsamp/step_size);
mat_col = round(1+(index-1)*(Nbatch2-1)/(xsamp-1));
subset = zeros(p2-p1+1,Nbatch2);
for n = 1:length(index)
      subset(mat_row1(n),mat_col(n)) = z1(n);
      subset(mat_row2(n),mat_col(n)) = z2(n);
end
path = rridern(subset,3)+p1-1;
pitch = round(interp1(1:Nbatch2,path,linspace(1,Nbatch2,xsamp)));

array = zeros(1,2*xsamp);
n = 1;
array(1) = 1;
while n < xsamp
      n = n + pitch(n);
      array(n) = 1;
end
peaks = 1:length(array);
peaks(~array) = [];
```

```
Xres = 500;
xpts = round(linspace(1,xsamp,Xres));
M = length(peaks);
N2 = p2;
N = 2*N2;


pointers = max(1,min(xsamp,([1:N]'-N2)*ones(1,M)+ones(N,1)*peaks));
MAT = reshape(abs(spch(pointers)),N,M) .* [hanning(N)*ones(1,M)];
path = rridern(MAT,4);
peaks = round(peaks+path-N2);
pitch_marks = peaks([peaks>=1]&[peaks<=xsamp]);
if (pitch_marks(1)~=1)
  pitch_marks=[1 pitch_marks];
end
if (pitch_marks(length(pitch_marks)~=xsamp))
  pitch_marks=[pitch_marks xsamp];
end

return



function path = rridern(MAT,N)
%
% y = rridern(MAT)
%
% This function traces a path from the first to the last columns
% of MAT, one that does not exceed slope == N (N integer >0) when
% assuming that successive rows are separated by one unit, and that
% successive columns are separated by one unit) and has the maximum
% possible cumulative MAT values along the path.  The output
% path y adheres to the sample points of MAT.
%
% W. Goncharoff  3/17/96


% calculate best-path cumulative errors:
[mrows,mcols] = size(MAT);
sf = mean(mean(MAT));
MAT = [-Inf*ones(N,mcols); MAT; -Inf*ones(N,mcols)];
best_paths = zeros(size(MAT));
range = N + (1:mrows);
T = zeros(1+2*N,mrows);
B = zeros(1+2*N,mrows);
R = zeros(1,(1+2*N)*mrows);
for i = -N:N
      B(i+N+1,:) = ones(1,mrows) * sf/sqrt(1+i*i);
      R(mrows*(i+N)+[1:mrows]) = range + i;
end
for col = 2:mcols
      T = reshape(MAT(R,col-1),mrows,1+2*N)';
      [temp1,temp2] = max(T+B);
      MAT(range,col) = MAT(range,col) + temp1';
      best_paths(range,col) = temp2';
end


% trace the optimal path backwards through the cum. error matrix:
best_paths = best_paths - N - 1;
path = zeros(1,mcols);
[total_error,row] = max(MAT(:,mcols));
path(mcols) = row;
for col = mcols:-1:2
      row = row + best_paths(row,col);
      path(col-1) = row;
end

path = path - N;

return



function W = hanning(N)

   W1 = (1 + cos(pi*linspace(-1,1,N+2)))'/2;
   W = W1(2:(N+1));

return


function y=openpcm(filename) %ésta función es llamada desde "modifica",
fid=fopen(filename,'r'); %abre las marcas de periodo de la combinación de fonemas
if fid>0 %en caso de que se encuentren las marcas de la combinación de fonemas
y=fread(fid,inf,'int16'); %se lee y se asigna a "y"
fclose(fid); %cerrandose "fid"
else %de no encontrarse (fid = -1)
   disp(strcat(filename,' not found')) %leer mensaje de que no encuentra lo que se busca
   y=0;       % y se hace "y" igual a 0 (el primero al ser "p1" (vacio) y "l1", naturalmente se encontrará)
end %fin del "if" (fid>0)


function y=savepm(filename,pm)
fid=fopen(filename,'w');
if fid>0
y=fwrite(fid,pm,'int16');
fclose(fid);
else
   disp(strcat(filename,' not found'))
   y=0;
end
```

Apéndice II

Función "pitch_marks"

```
function pitch_marks = find_pmarks(speech, fs_in)
Entradas:
->speech: señal de audio a analizar
->fs_in:  frecuencia de muestreo
Salida:
Pitch_marks: posición de las marcas de periodo
```

Generar contorno de energía de la señal:

- Obtener longitud de ventana de Hanning de acuerdo a la frecuencia de muestreo (wlen), numero de muestras que corresponda a 6.4ms de señal.

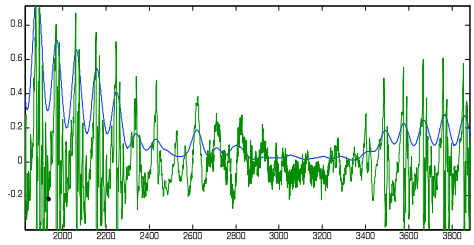
```
p1 = round(fs_in/400);
p2 = round(fs_in/60);
wlen = round((p1+p2)/3);
```

- Se duplica la señal (spch) y se obtiene su longitud (xsamp)

```
spch = speech(:)';
xsamp = length(spch);
N=1:xsamp;
```

- La señal de audio se eleva al cuadrado y se convoluciona dos veces con una ventana de Hanning.

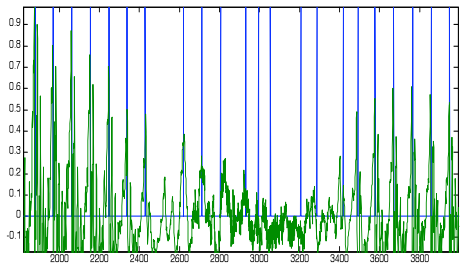
```
ecurve = conv(hanning(wlen)',spch.^2);
ecurve = conv(hanning(wlen)',ecurve);
ecurve = ecurve((1:xsamp)+wlen);
```



(Segmento de ecurve)

Obtener primera aproximación de marcas de periodo a partir de los máximos locales del contorno de energía:

```
peaks = ([0,diff(ecurve)]>0) & ([diff(ecurve),0]<0);
```



(segmento de peaks)

```
-Obtener el # de muestra donde se localizó cada máximo local (index) y la longitud
en muestras de cada periodo obtenido (pitch)
index = 1:xsamp;
index(~peaks) = [];
Npeaks = length(index);
pitch = diff(index);
pitch1 = [pitch, pitch(Npeaks-1)];
pitch2 = [pitch(1), pitch];
```

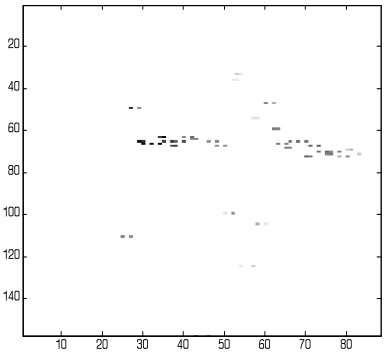
3.**Añadir marcas de periodo en las secciones no periodicas de la señal, ajustandolas a las marcas de las secciones periodicas.**

```
-Obtener el valor de las muestras en las posiciones donde se encontraron las
marcas(z1 y z2)
z1 = ecurve([index(2:Npeaks),index(Npeaks)]);
z2 = ecurve([index(1),index(1:(Npeaks-1))]);
```

- Divide la señal en bloques de acuerdo a la frecuencia de muestreo (aprox. 1.5 centésimas de segundo) y las acomoda en una matriz indicando la posición de las marcas.
- La matriz se inicializara a 0 y se llenera de la siguiente manera:
- Subset(renglón,columna)=muestra(n) donde:
- n:valor de la muestra en la marca encontrada
- el renglón y columna corresponden a la muestra donde se enconro la marca, donde el numero de columnas es el numero de ventanas analizadas (de longitud wlen) y el numero de renglones corresponde a el numero de muestras de la señal.

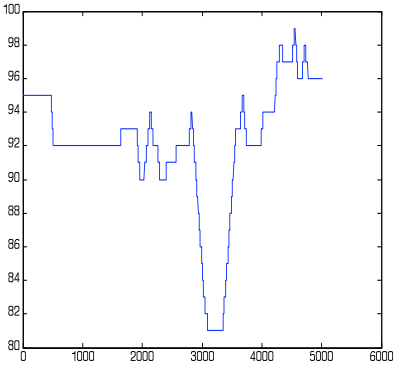
Es decir, la matriz contendrá valores solo en la posición que corresponde a las muestras donde se localizaron las marcas.

```
step_size = round(fs_in/192);
mat_row1 = max(1,min(p2-p1+1,pitch1-p1+1));
mat_row2 = max(1,min(p2-p1+1,pitch2-p1+1));
Nbatch2 = ceil(xsamp/step_size);
mat_col = round(1+(index-1)*(Nbatch2-1)/(xsamp-1));
subset = zeros(p2-p1+1,Nbatch2);
for n = 1:length(index)
        subset(mat_row1(n),mat_col(n)) = z1(n);
        subset(mat_row2(n),mat_col(n)) = z2(n);
end
```



(matriz subset)

```
-Recorre la matriz añadiendo marcas de periodo en las zonas donde no existan
generando una ruta optima de tal manera que se toque el mayor número posible de
marcas encontradas sin exceder una pendiente máxima (salto de máximo 3 renglones
por columna).
-La ruta se almacena en path (se almacena el renglón que sigue la ruta para cada
columna)
-Pitch es una interpolación lineal de la ruta de longitud igual al numero de
muestras.
path = rridern(subset,3)+p1-1;
pitch = round(interp1(1:Nbatch2,path,linspace(1,Nbatch2,xsamp)));
```

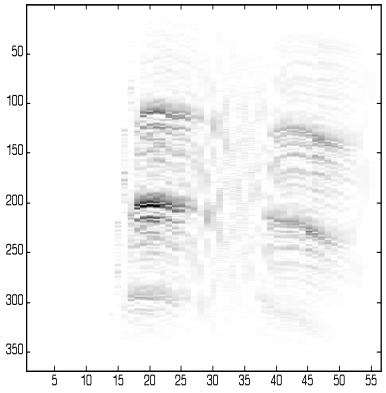


```
-Obtener nueva posicion de las marcas (peaks).
array = zeros(1,2*xsamp);
n = 1;
array(1) = 1;
while n < xsamp
        n = n + pitch(n);
        array(n) = 1;
end
peaks = 1:length(array);
peaks(~array) = [];


Xres = 500;
xpts = round(linspace(1,xsamp,Xres));
M = length(peaks);
N2 = p2;
N = 2*N2;
```

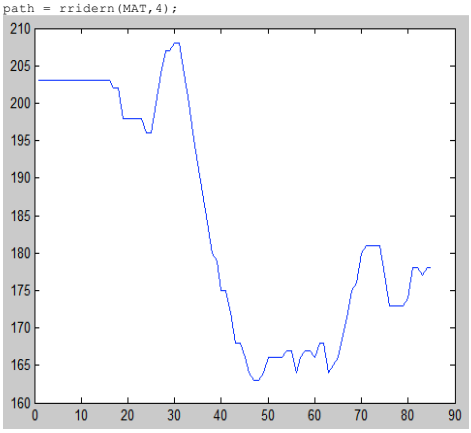
4. Ajustar marcas de periodo para que se ajusten a máximos de la señal

```
-MAT contiene en cada columna una ventana alrededor de la muestra donde se localizo
la marca de periodo.
pointers = max(1,min(xsamp,([1:N]'-N2)*ones(1,M)+ones(N,1)*peaks));
MAT = reshape(abs(spch(pointers)),N,M) .* [hanning(N)*ones(1,M)];
```



-Se obtiene una nueva ruta de tal manera que se recorre la posición de las marcas para que coincidan con los máximos locales de cada ventana.

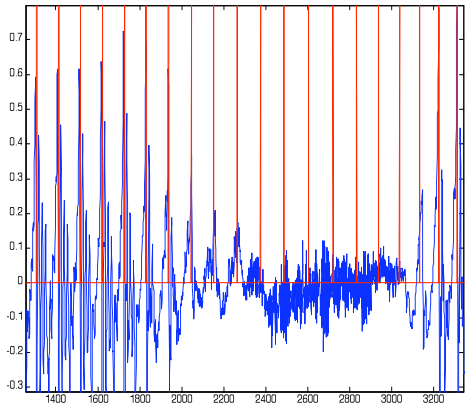
```
path = rridern(MAT,4);
```



-De acuerdo a longitudes de las marcas (path) obtener número de muestras donde se localizan las marcas(pitch_marks).

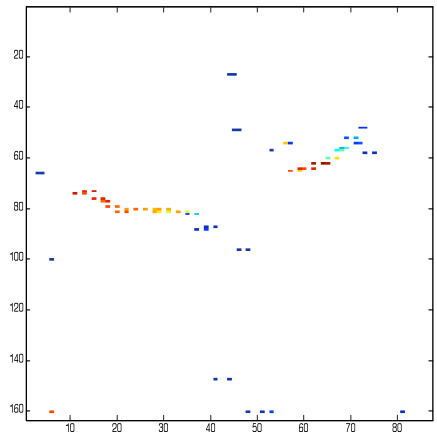
```
peaks = round(peaks+path-N2);
pitch_marks = peaks([peaks>=1]&[peaks<=xsamp]);
if (pitch_marks(1)~=1)
   pitch_marks=[1 pitch_marks];
end
if (pitch_marks(length(pitch_marks)~=xsamp))
   pitch_marks=[pitch_marks xsamp];
end

return
 end
```



```
function path = rridern(MAT,N)
```

1. Entradas:
MAT: matriz a traves de la cual se desea obtener la ruta obtima (donde la sumatoria de los valores de los renglones por los que pase sea la máxima posible sin tener una pendiente mayor a N).

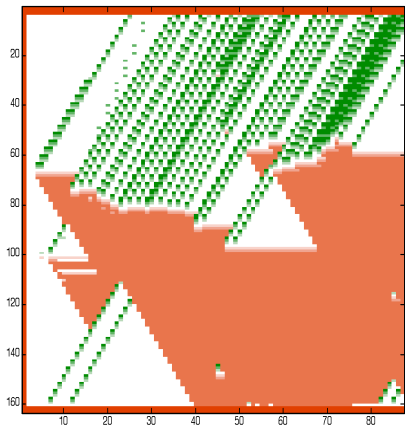


-Inicializar matrices a 0.

```
[mrows,mcols] = size(MAT);
sf = mean(mean(MAT));
MAT = [-Inf*ones(N,mcols); MAT; -Inf*ones(N,mcols)];
best_paths = zeros(size(MAT));
range = N + (1:mrows);
T = zeros(1+2*N,mrows);
B = zeros(1+2*N,mrows);
R = zeros(1,(1+2*N)*mrows);
```

-Obtener la matriz best_paths que contiene para cada renglon y columna la pendiente que se requiere para tocar el punto de mayor valor accesible desde esa posición.
-Modificar la matriz MAT añadiendo el error acumulado minimo que se ha obtenido para cada posición, recorriendo columna por columna.

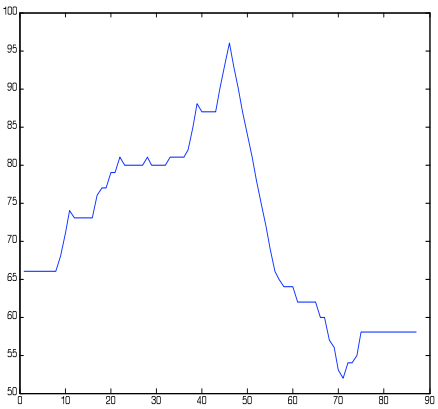
```
for i = -N:N
    B(i+N+1,:) = ones(1,mrows) * sf/sqrt(1+i*i);
    R(mrows*(i+N)+[1:mrows]) = range + i;
end
for col = 2:mcols
    T = reshape(MAT(R,col-1),mrows,1+2*N)';
    [temp1,temp2] = max(T+B);
    MAT(range,col) = MAT(range,col) + temp1';
    best_paths(range,col) = temp2';
end
best_paths = best_paths - N - 1;
```



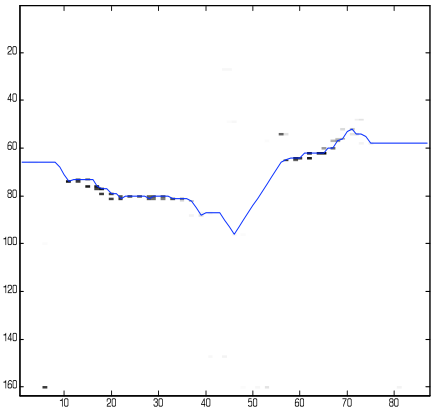
(Verde: pendiente positive / Rojo: pendiente negative)

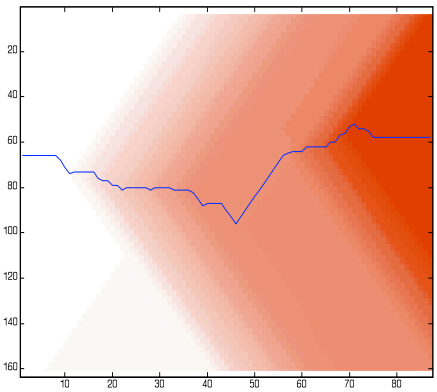
-Empezando por la ultima columna, obtener la ruta en la que el error acumulado sea menor de entre las posiciones accesibles.path contendrá el renglón que corresponda a cada columna de la ruta optima.

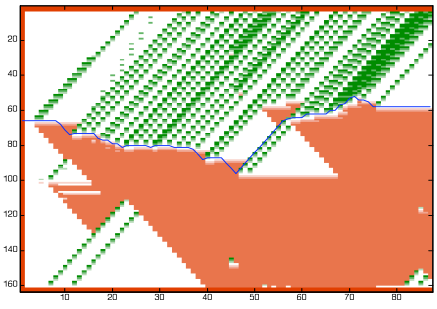
```
path = zeros(1,mcols);
[total_error,row] = max(MAT(:,mcols));
path(mcols) = row;
for col = mcols:-1:2
    row = row + best_paths(row,col);
    path(col-1) = row;
end
path = path - N;
return
end
```



Path montado sobre matrices:







Listado de imágenes

Capítulo I

Capítulo II

Capítulo III

Bibliografía

Alías, Frances *et al*. "Towards High-Quality Next-Generation Text-to Speech Synthesis: A Multidomain Approach by Automatic Domain Classification", *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 7, septiembre 2008, pp. 1340-1354.

Amario Toro, Jerónimo. "Un listado de sílabas en Español", *Cuadernos Cervantes*, <http://www. cuadernoscervantes.com/art_36_silabas.html> Fecha de último acceso: 21 de marzo 2010.

Beckman, Mary E. *Intonation Across Spanish, in the Tones and Break Indices Framework*, <http://www.ling.ohio-state.edu/~mbeckman/Sp_ToBI/Sp_ToBI_Jul29.pdf >, fecha de último acceso 5 de marzo 2008.

Benade Arthur H. *Foundamentals of Musical Acoustics*, New York: Dover, 1990.

Bernal Bermúdez Jesús, Jesús Bobadilla Sancho y Pedro Gomez Vildo, *Reconocimiento de voz y fonética acústica*, Madrid: Universidad Politécnica de Madrid, Alfaomega Ra-ma, 2000.

Bonada J. y X. Sierra, "Synthesis of the Singing Voice by Performance Sampling and Spectral Models", *IEEE Signal Processing Magazine*, vol. 24, 2007, pp. 67-79.

Castro Sierra, Eduardo. *Conceptos Básicos de Psicoacústica y Fisiología Auditiva de la Voz. Su Aplicación a la Música y el Canto*, México: CNCA/INBA, 1994.

_____. "Constant interval distance between resonance peaks of Spanish vowels sung at high pitch by sopranos", *148th ASA Meeting*. San Diego, CA. 2004.

Crocker Malcom J. (ed), Flanagan, Miller, Sundberg, Fant et al. *Encyclopedia of Acoustics*. s. l., Wiley-Interscience, 1997.

De la Vega Segura, Lilia Elena. *Diseño de un sintetizador de voz del idioma español hablado en México, Tesis de Maestría en Ingeniería*. México: Facultad de Ingeniería, División de Estudios de Posgrado, UNAM, 2007.

Del Río, Fernando. *Diseño de un sintetizador de voz en español usando el método TD-PSOLA, Tesis de Maestría en Ingeniería*. México: Facultad de Ingeniería, División de Estudios de Posgrado, UNAM, 2004.

Dodge, Charles y Rhomas A. Jerse. *Computer Music: Synthesis, Composition, and Performance*, New York: Schrimer, 1997.

Herrera Camacho, Abel. *Notas de la clase: Procesamiento Digital de Audio (voz)*,

<http://www.microsoft.fib.unam.mx/PaginasProfesores/AbelHerrera/Pro_dig_audio/frames.html>, fecha de último acceso: 21 de marzo 2010.

Huang X, A. Acero y H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, New Jersey: Prentice Hall, 2001.

Janer, Jordi y Jordi Bonada. "Performance driven control for sample based Singing voice synthesis", *Proc. Of the 9th Int. Conf. On Digital Audio Effects (DAFx-06) Montreal, Canada,* Sept. 18-20, 2006, pp. 41-44.

Joliveau E., J. Smith y J. Wolfe, "Tuning of vocal tract resonances by sopranos", *Nature*, vol. 427, 2004, pp. 116.

Kamen, E. *Introduction to Signals and Systems,* London: MacMillan, 1990.

Kyritsi, Varvara *et al.,* " A Score-to-Singing Voice Synthesis System for the Greek Language", *Proc. Of the Computer Music Conference*, 27-31 agosto, 2007, pp. 216-223.

Lemmety, Sammy. *Review of Speech Synthesis Technology*, <http://www.acoustics.hut.fi/publications/files/theses/lemmety_mst/thesis.pdf>, fecha de último acceso: 21 de marzo 2010.

Macon, M. *et al.*, "A Singing Voice Synthesis System Based Sinusoidal Modeling", *Proc. Int. Conf., Acoustics, Speech, Signal Processing*, Munich, Germany, vol. 1, núm. 199, pp. 435-4387.

Maragliano Mori, Rachele. *Conscienza della voce nella scuola italiana di canto*, Milano: Edizioni Curci-Milano, 1970.

Mari, Nanda. Canto e voce. *Difetti causati da un errato studio del canto,* Milano: Ricordi, 1987.

Newell, C. y Alaistair Edwards. "Unnatural but lively voice synthesis for empathic, synthetic performers", *Social Intelligence and Interaction in Animals, Robots and Agents, Joint symposium on Virtual Social Agents, The University of Hertfordshire*, 2005, pp. 137-143.

Oppenheim, A. y A. Willsky. *Señales y sistemas*, México: Prentice Hall Hispanoamericana, 1994.

Pardo, José Manuel,"Nuevas fronteras de la tecnología del habla: Síntesis de voz con Emociones", *Foro Computlense, Fundación General UCM. ONCE*. Dpto de Ingeniería Electrónica. Madrid: Universiad Politécnica de Madrid, <http://www.funcacionucm.es/www.once.es>

Pattison, Pat. *Songwriting Essential Guide to Lyric Form and Structure. Tools and Techniques for writing better lyrics*, Boston: Berklee Press, 1991.

Rabiner, Lawrence y R. Schafer. *Digital Processing of Speech Signals*, s. l. ,Prentice Hall,

1978.

Rabiner, Lawrence y Juang Biing-Hwang, *Fundamentals of Speech Recognition*, New Jersey: Prentice Hall, 1993.

Ríos Mestre, Antonio. "La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico", *Estudios de Lingüística del Español,* vol. 4, 1999, <http://elies.rediris.es/elies4/> fecha de último acceso 6 de mayo 2009.

Siivola, Veesa. *A Survey of Methods for the Synthesis of the Singing Voice*, <http://www.cis.hut.fi/vsiivola/papers/svs.ps>, fecha de último acceso: 4 de noviembre 2009.

Torres, H. M. y J. A. Gurlekian, "Acoustic Speech unit segmentation for concatenative synthesis", *Computer Speech and Language,* núm. 22, 2008, pp. 196-206.

Uneson, Marcus, "Outlines of Burcas-a Simple Concatenation- Based MIDI-to- Singing Voice Synthesis System", *TMH-QPSR*, *Fonetik,* vol. 43, 2002, pp. 1-4.