Universidad Zaragoza
1542

Trabajo Fin de Grado

Evaluación perceptual de vocoders para síntesis de voz basada en HMM

Autor

Fernando Martín Lana

Director

Luis Vicente Borruel

Escuela de Ingeniería y arquitectura

2015

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. FERNANDO MARTÍN LANA,

con nº de DNI 77131070H en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Grado, (Título del Trabajo) Evaluación perceptual de vocoders para síntesis de voz basada en HMM

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, a 5 de septiembre de 2015

Fdo: Fernando Martín Lana

Agradecimientos

Quisiera dar mi más sincero agradecimiento a Luis Vicente por toda su ayuda y dedicación en la realización de este trabajo. Sin él no hubiera sido posible.

Para mi familia, en especial para mi madre

Nieves y mi hermana Marta, que siempre

me han apoyado y a quien les debo

todo lo que soy

Evaluación perceptual de vocoders para síntesis de voz basada en HMM

RESUMEN

Hoy en día, la síntesis de voz basada en modelos ocultos de Markov es uno de los métodos más utilizados y conocidos para sintetizar voz. Una de las características de este tipo de síntesis es la parametrización de la señal de voz que hace que la síntesis sea más flexible y eficiente.

En este trabajo se da una visión general de tres de los vocoders que se utilizan actualmente en síntesis de voz basada en modelos ocultos de Markov, viendo algunas de sus características básicas. Los vocoders analizados son: SPTK, AHO-coder y STRAIGHT.

Estos vocoders han sido comparados perceptualmente desde el punto de vista de la audición humana para determinar cuál de todos proporciona un mayor nivel de calidad vocal. Para ello se ha implementado en la herramienta matlab la recomendación P.862 de la ITU (más conocida como PESQ), un estándar que define un método objetivo de evaluar la calidad de códecs vocales de banda estrecha. Este método objetivo es una aproximación o predicción de la nota de calidad MOS que se obtendría en un experimento subjetivo de escucha.

El nivel de calidad de cada vocoder ha sido evaluado en resíntesis, es decir, en el proceso en el cual se parametriza la señal de voz y, con dichos parámetros, inmediatamente se sintetiza. Se ha utilizado la base de datos de voz Albayzin para realizar el proceso de evaluación y se ha desarrollado en matlab todo el software necesario para hacer la resíntesis de la base de datos con los diferentes vocoders y obtener el valor de calidad perceptual mediante PESQ.

En la parte final del trabajo se presenta un análisis detallado de los resultados obtenidos, mostrando el valor de calidad promedio para cada vocoder y realizando un análisis detallado por diferentes grupos (por sexo y edad). Finalmente se presentan las conclusiones extraídas del trabajo y se nombran algunas líneas futuras de trabajo.

Índice de contenidos

Índice de figuras

Índice de tablas

Lista de acrónimos

FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HTS	HMM-based speech synthesis system
IIR	Infinite Impulse Response
IRS	Intermediate Reference System
ITU	International Telecommunication Union
LPC	Lineal Predictive Coding
MCC	Mel Cepstral Coefficients
MLSA	Mel Log Spectral Approximation
MOS	Mean Opinion Score
MOS-LQO	Mean Opinion Score, Listening Quality Objective
MVF	Maximum Voice Frequency
PESQ	Perceptual Evaluation of Speech Quality
RAPT	Robust Algorithm Pitch Tracking
SPTK	Speech Signal Processing Toolkit
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum
VAD	Voice Activity Detector

1. Introducción

1.1. Introducción del trabajo

En los últimos años, la síntesis de voz ha experimentado un gran auge debido al aumento de la demanda de tecnología y a la necesidad de que ésta sea más cómoda y accesible para el usuario. Una de las últimas líneas de investigación es la síntesis de voz mediante modelos ocultos de Markov, explicados detalladamente en [1] y [2].

En este tipo de síntesis, no se trabaja con segmentos de señal originales, sino con su versión parametrizada. Esto dirige a una síntesis eficiente y flexible. Es aquí donde intervienen los codificadores paramétricos, también conocidos como vocoders. En este trabajo se presentan 3 vocoders que se utilizan actualmente en síntesis mediante HMM y se realiza una evaluación perceptual de todos ellos. Los vocoders objeto de estudio son el implementado por la herramienta SPTK, además de STRAIGHT y AHO-coder.

Para evaluar de forma objetiva su nivel de calidad, en este trabajo se presenta, se estudia y se implementa el estándar de la ITU-T denominado PESQ. Este estándar proporciona una nota de calidad objetiva equivalente a una nota MOS. Con esta herramienta y una base de datos de voz llamada Albayzin, se realiza el proceso de evaluación perceptual de los 3 vocoders, buscando cuál es el que proporciona una mayor calidad.

1.2. Objetivos principales del trabajo

El objetivo principal de este trabajo es evaluar la calidad de cada vocoder en el proceso de resíntesis. Es decir, en el proceso en el que se obtienen los parámetros de la voz e inmediatamente se realiza la síntesis con dichos parámetros.

Para ello, deben realizarse dos tareas principalmente:

-Llevar a cabo la implementación en la herramienta matlab del estándar PESQ, que permite la evaluación perceptual de los vocoders de forma objetiva.

-Realizar en matlab el software necesario que permita realizar el proceso de resíntesis con todos los vocoders en estudio a una base de datos con señales de voz, y

determinar con la recomendación el nivel de calidad de cada uno. Obtener un amplio conjunto de resultados para determinar cuál es el mejor vocoder.

Aunque algunos codificadores están en lenguaje C, tanto la recomendación como el proceso de evaluación de la base de datos han sido implementados en el entorno de matlab, al ser una herramienta más cómoda y más potente para realizar tareas de procesado de señal.

1.3. Estructura de la memoria

La memoria está dividida principalmente en tres partes. En el apartado 2 se realiza una breve introducción a la síntesis con HMM y a los codificadores paramétricos y se presentan los tres codificadores que han sido objeto de estudio y comparación en este trabajo. En el apartado 3 se presenta una descripción detallada de la recomendación utilizada para determinar el nivel de calidad de los vocoders. En la parte final de la memoria (los apartados 4 y 5) se explica la metodología llevada a cabo para realizar la evaluación de los vocoders y se presentan los resultados obtenidos. Finalmente, se comentan las conclusiones y las líneas de investigación futuras.

2. HMM y los codificadores paramétricos, los vocoders

2.1. Introducción a la síntesis de voz basada en HMM

Un modelo oculto de Markov es un conjunto de estados, en los que para cada estado se da una observación de salida. La salida de cada estado puede ser un escalar ó un vector y tiene una función de densidad de probabilidad. En cada instante de tiempo se transita de un estado a otro con una cierta probabilidad. Se denomina modelo oculto de Markov porque dicho modelo es una extensión de una cadena de Markov pero con un doble proceso estocástico. Por un lado, la probabilidad de transición de un estado a otro y, por otro, la probabilidad de tener una determinada observación a la salida de cada estado. El término oculto se usa porque ante una determinada observación, no podemos asegurar en qué estado del modelo nos encontramos, pero sí estimar cuál es el más probable.

Se ha comprobado que parametrizando la voz y modelándola mediante modelos ocultos de Markov, se alcanza un buen nivel de inteligibilidad de forma eficiente (sin necesidad de enormes bases de datos) y muy flexible (facilidad para cambiar algunos parámetros de la voz, como son la entonación o la velocidad de producción).

El proceso de obtención de los modelos que permiten sintetizar voz se puede observar en la figura 2.1. Se realiza un entrenamiento con una base de datos de voz etiquetada de uno o varios locutores. Dicha base de datos es dividida en unidades fundamentales (fonemas) y, para cada una, se obtiene su representación paramétrica, que es la representación con la que se entrena.

Tras completarse el proceso de entrenamiento estadístico, se obtiene para cada fonema un conjunto de modelos ocultos que en el proceso de síntesis permitirán sintetizar voz. Estos modelos generalmente suelen modelar la duración del fonema, su información espectral y su entonación, aunque puede haber modelos adicionales.

En el proceso de síntesis, la secuencia de texto de entrada es transformada en una sucesión de modelos ocultos de Markov, que representan los fonemas. A partir de ellos se generan los parámetros que posteriormente permiten sintetizar la voz. Es en la parametrización de la voz y en la síntesis a partir de ellos donde entran en escena los vocoders.

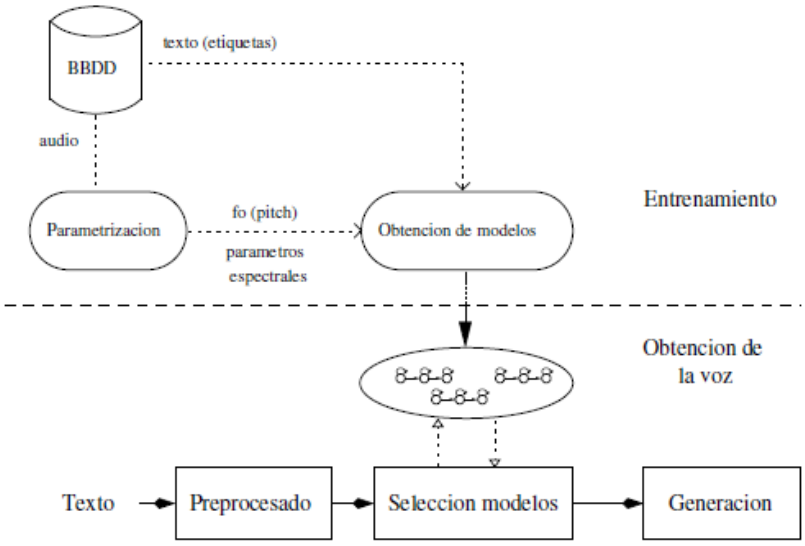


Figura 2.1: Sistema de síntesis mediante HMM

2.2. Los codificadores paramétricos

Los codificadores pueden agruparse en tres grandes tipos: codificadores de forma de onda, paramétricos e híbridos. Los primeros se limitan a cuantificar y codificar muestra a muestra la señal, por lo que la señal decodificada tendrá un gran parecido a la señal original y será de gran calidad. Los codificadores paramétricos o vocoders realizan un análisis de la señal para obtener una serie de parámetros que son suficientes para poder reconstruir de nuevo la señal. Esto permite reducir el ancho de banda de transmisión, aunque la calidad es bastante inferior. Sin embargo, la parametrización proporciona mayor flexibilidad e inteligibilidad (aunque menor calidad y naturalidad) si se realiza una síntesis basada en HMM, en lugar de hacer una síntesis basada en la concatenación de fonemas pregrabados.

Un vocoder está dividido en dos grandes bloques: codificación o análisis y decodificación o síntesis. El modelo de generación de la voz se basa en un esquema de excitación más filtro. En la figura 2.2 puede verse el esquema de generación de la voz de un vocoder clásico, como podría ser el vocoder LPC.

La señal a codificar es enventanada con un cierto tamaño de trama y solapamiento entre tramas. Para cada trama se calcula el período de pitch T_0 y los coeficientes espectrales. Estos parámetros son utilizados para resintetizar la voz tal y como se muestra en la figura 2.2.

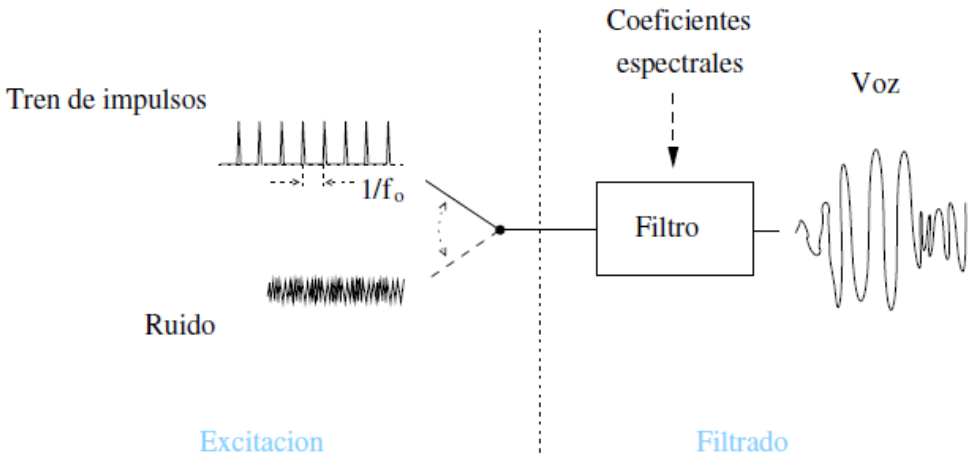


Figura 2.2: Esquema del modelo de generación de la voz de un vocoder clásico

Se genera una señal de excitación consistente en una secuencia de impulsos de periodo T_0 para tramas sonoras, y ruido gausiano para tramas sordas. La frecuencia de pitch es la frecuencia fundamental con la que vibran las cuerdas vocales al pronunciar fonemas sonoros como por ejemplo las vocales. En estos casos el aire circula libremente sin ninguna obstrucción, mientras que para los sonidos sordos se produce una obstrucción del aire que no permite que las cuerdas vocales vibren, asemejándose más a ruido gausiano. En el análisis se toman dos decisiones con respecto al pitch: si una trama es sorda o sonora y, si es sonora, que periodo de pitch tiene. Mientras que las tramas sonoras poseen un alto nivel de periodicidad, las tramas sordas carecen de él. La frecuencia fundamental o pitch se encuentra entre 50 y 250 Hz para los hombres y entre 120 y 500 Hz para las mujeres.

Los coeficientes espectrales modelan la forma que tiene el tracto vocal, que varía con el tiempo. Básicamente, describen la envolvente de la señal de voz, mostrando sus resonancias (denominadas formantes).

La convolución de la señal de excitación con el filtro del tracto vocal originan la señal sintetizada. Los parámetros que permiten realizar la síntesis son hallados en el análisis y hay diferentes métodos para obtener cada parámetro, tanto el pitch como los coeficientes del filtro.

2.3. Presentación de los vocoders comparados

2.3.1. SPTK

SPTK es un toolkit de procesado de voz desarrollado por el grupo de trabajo SPTK del Instituto de Tecnología de Tokyo. Es una herramienta de libre distribución y código abierto escrito en el lenguaje de programación C y puede ser ejecutado tanto en sistemas UNIX como Windows. Puede ser descargado en http://sp-tk.sourceforge.net.

Entre sus múltiples funcionalidades, permite realizar un análisis/síntesis Mel-cepstral, explicado en detalle en [2]. El filtro de síntesis viene determinado por la siguiente ecuación:

$$H(z) = exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m} \tag{2.1}$$

donde $c(m)$ son los coeficientes del cepstrum basados en la escala Mel (MCC) y $\tilde{z}^{-1}$ es la transformación bilineal mediante la cual se aproxima el plano Z en la escala Mel en función del plano Z en escala lineal.

$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\ ,\ |\alpha| < 1 \tag{2.2}$$

α es el parámetro que permite ajustar la relación entre la escala lineal y la escala Mel y depende de la frecuencia de muestreo. Por ejemplo, para 8 kHz, el valor óptimo de α es 0.31, para 16 kHz α es 0.42... La relación que queda entre la pulsación angular lineal y la pulsación angular en la escala Mel es la siguiente:

$$\beta(\omega) = arctan \frac{(1-\alpha^2)sen\omega}{(1+\alpha^2)cos\omega-2\alpha} \tag{2.3}$$

Los coeficientes MCC son obtenidos en el proceso de análisis de la señal mediante la minimización de la siguiente función de coste:

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\{expR(\omega) - R(\omega) - 1\}d\omega \tag{2.4}$$

donde

$$R(\omega) = logI_N(\omega) - log\left|H(e^{j\omega})\right|^2 \tag{2.5}$$

$$I_N(\omega) = \frac{\left|\sum_{n=0}^{N-1}\omega(n)x(n)e^{-j\omega n}\right|^2}{\sum_{n=0}^{N-1}\omega^2(n)} \tag{2.6}$$

$I_N(\omega)$ es el periodograma modificado con $\omega(n)$ la ventana elegida. La minimización de (2.5) conlleva a la minimización de (2.4) y ésta se realiza respecto a los coeficientes

cepstrum. El valor de E es convexo con respecto a los coeficientes, por lo que se halla un único valor de los coeficientes en el cual E es mínimo. Este mínimo se halla utilizando el método de Newton Raphson.

Como el filtro de síntesis $H(z)$ no es directamente realizable, se utiliza la aproximación de Padé para aproximar la función exponencial mediante una función racional. Este filtro resultante se denomina MLSA (Mel Log Spectral Approximation). La obtención de dicho filtro y de sus coeficientes a partir de $H(z)$ puede verse en detalle en [2].

Respecto a la obtención del periodo de pitch, se utiliza el algoritmo RAPT, descrito en [3]. Para cada trama de la señal, los coeficientes MCC y el pitch son interpolados muestra a muestra entre los valores de dos tramas consecutivas, para conseguir un efecto de suavizado que elimina cambios bruscos en los parámetros.

Para finalizar, la señal de excitación es generada mediante una excitación mezclada en el dominio de la frecuencia. Parte de la idea de que un segmento de voz no está compuesto por un tren de deltas o ruido gausiano únicamente, sino que es una composición de ambos. SPTK utiliza el hecho observable de que la componente sonora o periódica decae al aumentar la frecuencia y a partir de una determinada frecuencia de corte puede considerarse que sólo existe componente aperiódica (ruido gausiano). Esto se lleva a cabo en SPTK de la siguiente forma. Para cada trama se genera la señal de excitación en el dominio del tiempo correspondiente a la frecuencia de pitch y se filtra paso bajo en el dominio de la frecuencia. La frecuencia de corte es elegida y es constante. Después se genera ruido gausiano en el dominio del tiempo y es filtrado paso alto en el dominio de la frecuencia con la misma frecuencia de corte. Así pues, la señal de excitación final es suma de ambas contribuciones.

Este modelo de excitación mezclada es utilizado en bastantes vocoders para mejorar el nivel de calidad y también es utilizado por el resto de vocoders analizados.

2.3.2. STRAIGHT

STRAIGHT es un codificador de voz diseñado por Hideki Kawahara, disponible en http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTtrial para matlab. STRAIGHT se basa en el aspecto fundamental del Vocoder, la descomposición de la voz en excitación + filtro, pero se aleja del otro concepto básico de obtener una versión paramétrica de la voz eficiente en cuanto a ancho de banda.

STRAIGHT parte de la idea básica de que un segmento de voz es el resultado de la convolución de una señal de excitación con una envolvente espectral. En el dominio de la frecuencia esto equivale a un producto. Para segmentos sonoros aparecen picos espectrales en la frecuencia de pitch y sus armónicos, mientras que para segmentos sordos se tiene ruido gausiano. La amplitud de estos picos está conformada por la forma del filtro espectral.

Para obtener la envolvente del espectro, STRAIGHT trata de eliminar las interferencias producidas por las secuencias de impulsos periódicos de la excitación en el dominio temporal y frecuencial, eliminando los valles del espectro y obteniendo los picos de los armónicos. Esto se consigue filtrando la señal con un conjunto de ventanas complementarias síncronas a la frecuencia fundamental como se detalla en [4] y [5]. En las figuras 2.3 y 2.4 puede observarse una trama de señal en el dominio de la frecuencia y su envolvente obtenida.

Respecto a la señal de excitación, se utiliza un extractor del pitch basado en la frecuencia instantánea. Se genera la excitación basándose en un modelo de excitación mezclada. Es decir, para cada trama de señal se asume que puede haber tanto componentes periódicas como aperiódicas, y además se hace la distinción en función de la frecuencia. Esto se observa en la figura 2.3, en la que para frecuencias bajas se puede apreciar claramente la naturaleza periódica del pitch, mientras que para altas frecuencias la componente dominante es el ruido. A diferencia de SPTK, STRAIGHT no distingue entre frecuencias bajas periódicas y frecuencias altas aperiódicas, sino que para cada frecuencia existe una contribución de ambas componentes, cumpliéndose que la aperiodicidad aumenta con la frecuencia.

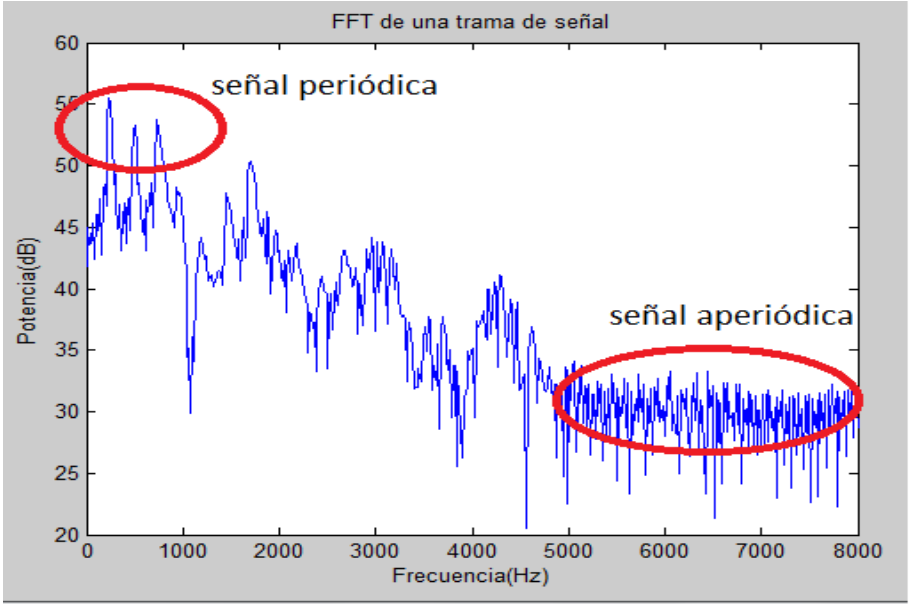


Figura 2.3: Trama de una señal de voz en el dominio de la frecuencia

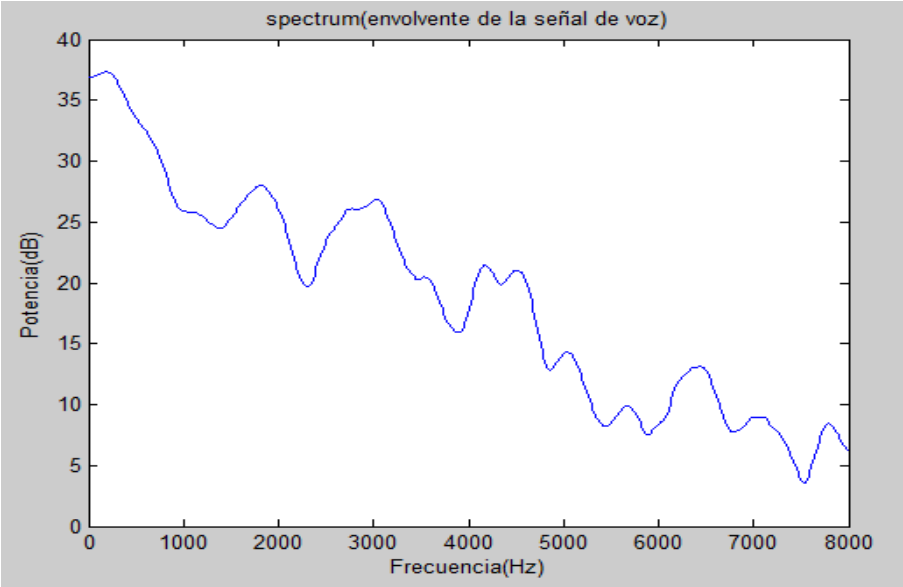


Figura 2.4: Envolvente obtenida con STRAIGHT

Así pues STRAIGHT calcula un coeficiente de aperiodicidad para cada muestra frecuencial, que indica la contribución del ruido gausiano en la señal de excitación (figura 2.5).

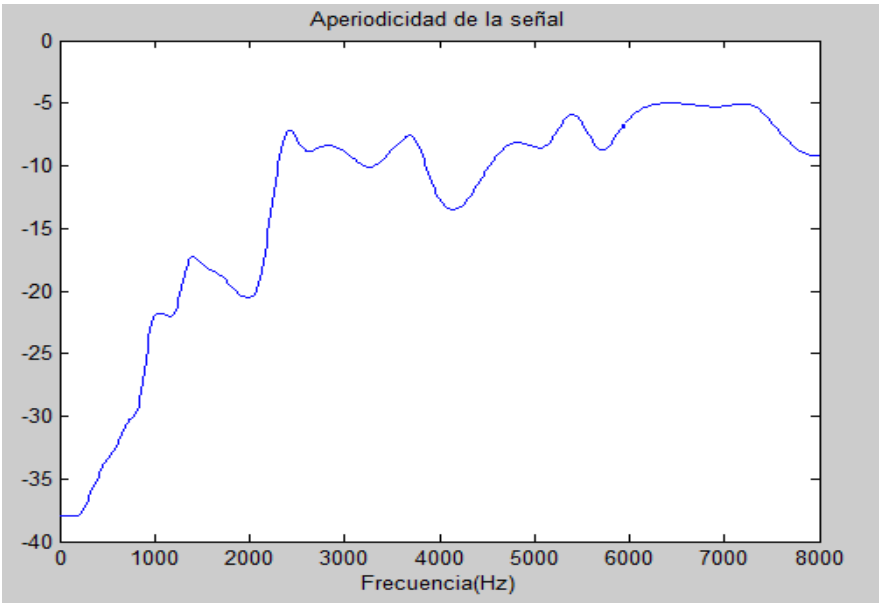


Figura 2.5: Coeficientes de aperiodicidad

Cuando un coeficiente vale 0 dB, indica que esa componente es puramente aperiódica, mientras que si está cercano a -60 dB, indica periodicidad pura. En la figura 2.5 se puede observar como los coeficientes varían de acuerdo a lo que cabría esperar observando la figura 2.3.

Como se ha indicado anteriormente, STRAIGHT no tiene en cuenta consideraciones sobre el número de parámetros (muy alto) sino que se centra en la calidad. Por ejemplo, para 16 kHz, por cada trama STRAIGHT extrae 513 coeficientes espectrales, 513 de aperiodicidad y uno del pitch. Esto hace inviable en la práctica la posibilidad de entrenar con tantos parámetros en sistemas basados en HMM. Por ello, se extraen coeficientes MCC del espectro y los coeficientes de aperiodicidad son agrupados en 5 bandas y promediados en cada banda. Estos parámetros son con los que se entrena y se ha tenido en cuenta a la hora de realizar la resíntesis de STRAIGHT en el proceso de evaluación.

2.3.3. AHO-coder

AHO-coder es un vocoder de voz desarrollado por Daniel Erro, miembro de AHOLAB, laboratorio de procesado de señal de la Universidad del País Vasco. Existe un ejecutable para linux que puede ser descargado en http://aholab.ehu.es/ahocoder. Este vocoder trabaja sólo con señales de 16 kHz y 16 bits y está orientado para ser utilizado conjuntamente con HTS, una herramienta que permite realizar síntesis de voz basada en HMM.

AHO-coder extrae 3 tipos de parámetros en el proceso de análisis. Extrae los MCC, el pitch y un último parámetro denominado MVF (maximum voice frequency). Este último parámetro indica la frecuencia de corte hasta la que se considera que existe voz (componentes periódicas) y se calcula para cada trama. Este tipo de modelo se conoce como HNM (harmonics plus noise model) y comparte la misma idea que la explicada en el vocoder SPTK.

Los algoritmos utilizados para obtener la MVF y el resto de parámetros se explican en detalle en [6], [7] y [8]. Básicamente se sigue la tendencia marcada por STRAIGHT de obtener la envolvente espectral de forma precisa a partir de la señal original.

El primer paso es obtener el pitch mediante un algoritmo de detección preciso. En función de si la trama se considera sorda o sonora se sigue un procedimiento u otro para obtener los MCC. Si la trama es sonora, se obtiene el valor de amplitud en frecuencias armónicas de la frecuencia fundamental. Interpolando dichas muestras se obtiene el valor de la envolvente. Si la trama es sorda, la envolvente simplemente es igual a la FFT. De la envolvente de la señal se obtiene de forma recursiva los MCC.

En el proceso de síntesis se obtiene la envolvente a partir de los MCC. Se genera la señal de excitación a partir del pitch y de la MVF. Si la trama es sorda, se

genera ruido gausiano en todo el espectro de frecuencia. Si es sonora, se genera la señal de excitación a partir de la información de las amplitudes de los armónicos y el periodo de pitch correspondiente y se filtra paso bajo con frecuencia de corte la determinada por la MVF. Después se genera ruido gausiano y se filtra paso alto con la misma frecuencia de corte. Es decir, la generación de la señal de excitación es muy parecida a SPTK con la principal diferencia de que la frecuencia de corte es variable y la marca la MVF.

3. PESQ, un estándar para la evaluación de la calidad vocal

Para realizar una comparación perceptual objetiva de los diferentes códecs de voz, se ha implementado en matlab la recomendación P.862 de la ITU [9]. Este estándar, más conocido como PESQ (perceptual evaluation of speech quality), es un método objetivo para la evaluación de la calidad vocal de extremo a extremo de redes telefónicas y códecs vocales de banda estrecha (hasta 16 kHz). Sirve para medir objetivamente la calidad subjetiva que se percibiría en un experimento de sólo escucha.

Trata los efectos del filtrado, el retardo variable y las distorsiones localizadas para medir principalmente los efectos del ruido y la distorsión de voz unidireccionales sobre la calidad vocal, pero no así otros como la pérdida de sonoridad, el retardo, efectos locales, ecos y otros factores de degradación relacionados con la interacción bidireccional. Es decir, el estándar mide objetivamente cómo de parecidas perceptualmente son las señales inicial y final y qué nivel de calidad percibiremos, pero no tiene en cuenta efectos degradantes como ecos y retardos, que son importantes en comunicaciones en tiempo real.

3.1. Visión general del algoritmo PESQ

La idea principal de PESQ es la comparación de dos señales. Una que se denominará señal de referencia y que es una señal limpia sin ningún tipo de degradación, y otra que se denominará señal degradada, que es el resultado de pasar la señal de referencia por un sistema de comunicaciones, como se ve en la siguiente figura.

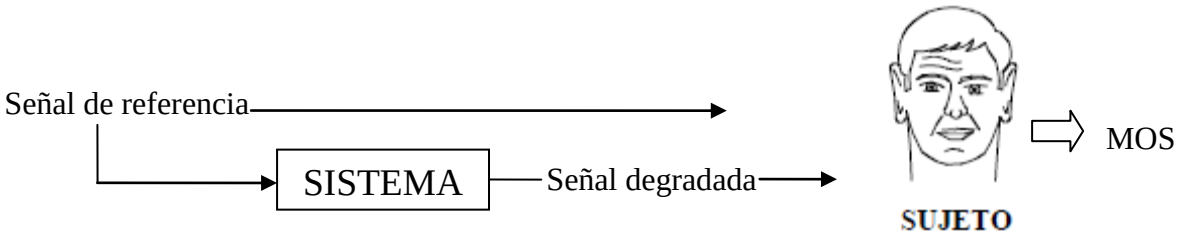


Figura 3.1: Evaluación subjetiva de un sistema de comunicaciones

En un test subjetivo de escucha, los sujetos escuchan ambas señales y dan una puntuación de calidad a la señal degradada respecto a la de referencia. El valor medio de esta puntuación, denominado MOS (mean opinion score) está comprendido entre 1 (calidad pobre) y 5 (máxima calidad). La salida de PESQ es una predicción objetiva, basándose en el sistema de percepción auditivo humano, del valor que se obtendría en una prueba de escucha subjetiva.

El algoritmo acepta señales de voz grabadas con frecuencias de muestreo de 8 ó 16 kHz y puede resumirse en un diagrama de bloques como el de la figura 3.2.

Ambas señales son normalizadas y filtradas. Después, la señal de referencia es dividida en articulaciones y el retardo para cada articulación es calculado. Cada trama de una articulación posee su mismo retardo, con lo que cada trama de la señal de referencia es alineada con su correspondiente segmento de la señal degradada. Una vez que las señales están correctamente alineadas, dos indicadores de distorsión son obtenidos para con ellos obtener la nota PESQ. Esta nota está comprendida entre -0.5 y 4.5. Para obtener la nota MOS-LQO (mean opinion score, listening quality objective), comparable a una nota MOS subjetiva, se utiliza una función de correspondencia optimizada experimentalmente con una gran cantidad de casos. En los siguientes apartados se explica más en detalle el algoritmo PESQ.

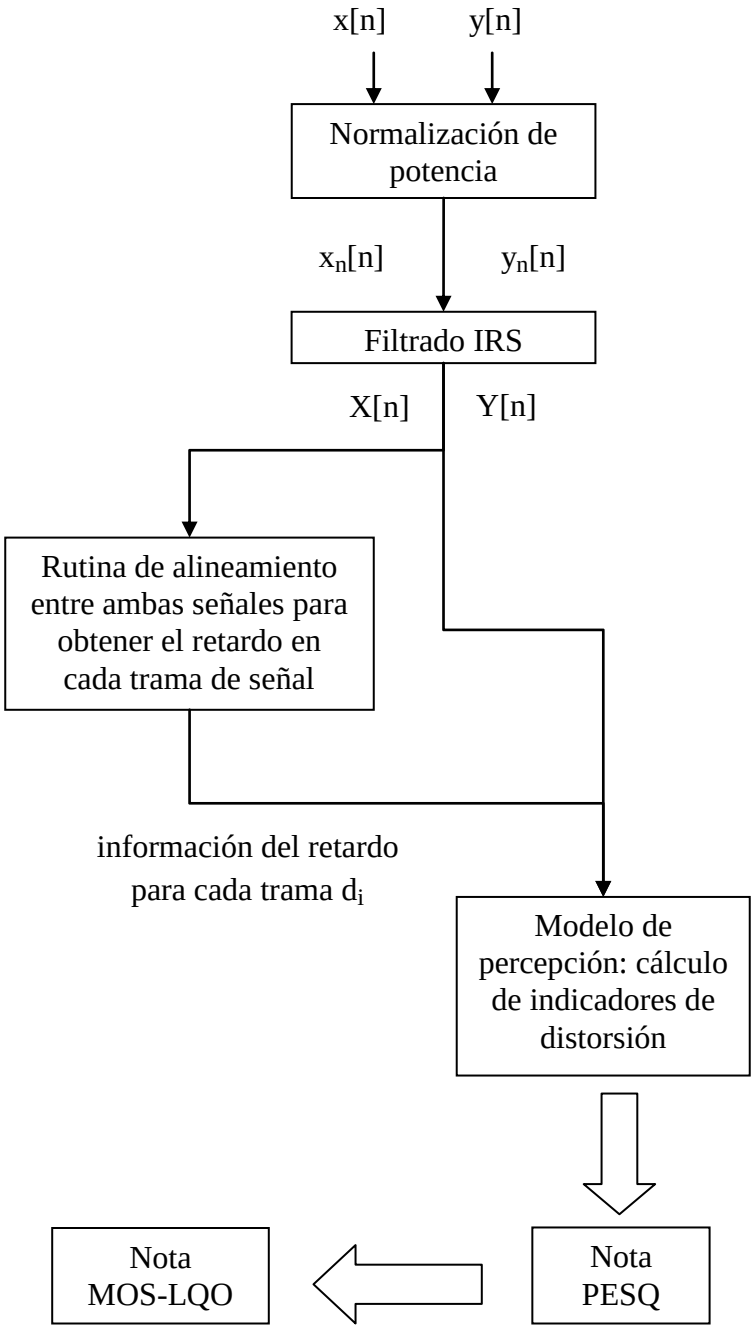


Figura 3.2: Esquema general del algoritmo PESQ

3.2. Normalización de potencia y filtrado IRS

El primer paso del algoritmo es realizar una normalización en potencia de las señales de referencia x[n] y degradada y[n], ya que la señal de referencia puede estar calibrada a cualquier nivel y se desconoce la ganancia del sistema que está siendo sometido a prueba. Ambas señales son filtradas en el dominio de la frecuencia con un filtro que deja pasar sin atenuación las componentes entre 350 y 3250 Hz y atenúa 500 dB las frecuencias por debajo de 300 Hz y por encima de 3500 Hz. Estas señales filtradas se utilizan sólo para obtener el valor de sus potencias, siendo la potencia de una señal el valor promedio de sus muestras al cuadrado. Para normalizar ambas señales se dividen todas sus muestras por la raíz de su potencia obtenida y después se multiplican por la raíz de la potencia deseada. Así ambas señales estarán a un mismo nivel fijo de potencia para ser comparadas.

El segundo paso es filtrar las señales normalizadas $x_n[n]$ e $y_n[n]$ con un filtro de característica IRS (intermediate reference system), es decir, con la respuesta frecuencial correspondiente a un sistema intermedio de referencia. Este sistema intermedio es aquel que se utilizaría en la prueba subjetiva para escuchar las señales. Esto se debe tener en cuenta desde el punto de vista perceptual ya que el sistema de escucha filtra las señales que son finalmente escuchadas por el sujeto. PESQ utiliza un filtro de característica IRS como el de la tabla 3.1, siendo un filtro con una respuesta en frecuencia promedio para hacer al método PESQ lo menos sensible del sistema de escucha utilizado y de su acoplamiento con la oreja del sujeto.

Frecuencia (Hz)	Amplitud (dB)	Frecuencia (Hz)	Amplitud (dB)
0	-200	300	6
50	-40	350	8
100	-20	400	10
125	-12	500	11
160	-6	(600 - 3250)	12
200	0	3500	4
250	4	>= 4000	-200

Tabla 3.1: Valores de la respuesta en frecuencia del IRS

Realizando una interpolación lineal de los valores de la tabla y filtrando en el dominio de la frecuencia se obtienen las señales X[n] e Y[n], que serán las señales de entrada de la rutina de alineamiento y del bloque del modelo de percepción.

3.3. Alineamiento temporal

El alineamiento temporal puede resumirse en un esquema como el de la siguiente figura:

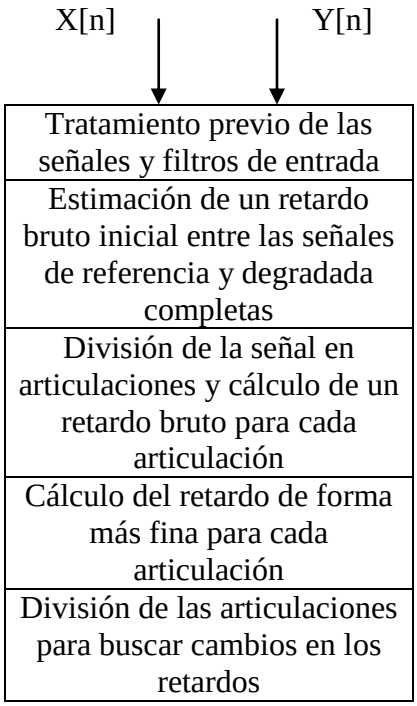


Figura 3.3: Proceso de alineamiento

Inicialmente las señales son pre-procesadas y filtradas. Posteriormente, mediante un detector de actividad vocal, la señal de referencia es dividida en subsecciones denominadas articulaciones, que son segmentos de voz separados por silencios. Se calcula un retardo estimado inicial utilizando ambas señales completas y después un retardo bruto para cada articulación. Finalmente, se obtiene un retardo más fino y se procede a la división de las articulaciones en otras más pequeñas para encontrar cambios bruscos en el retardo dentro de cada articulación. El resultado final del proceso de alineamiento es una información acerca del número de articulaciones encontradas y su inicio, final, retardo... para que ambas señales puedan ser alineadas antes de ser comparadas dentro del modelo de percepción.

3.3. 1. Tratamiento previo y filtros de entrada

Para el proceso de alineamiento, a ambas señales se les sustrae la media y se les somete a un proceso de fade-in y fade-out. Esto es multiplicar el comienzo y el final de las señales por un valor que crece o decrece progresivamente de 0 a 1 para que el comienzo y el final de las señales sea suave. Esto sólo se le realiza a los 4 primeros y últimos milisegundos, que equivale en muestras a 32 y 64 para una frecuencia de muestreo de 8 y 16 kHz respectivamente.

Después ambas señales son filtradas mediante una serie de filtros IIR (Infinite Impulse Response) de orden 3 puestos en cascada. El número de filtros y sus coeficientes depende de la frecuencia de muestreo de las señales, pero la respuesta frecuencial final tiende a ser un filtro paso bajo.

3.3.2. Alineación basada en envolventes

Ambas señales son segmentadas en tramas de 4 milisegundos (32 ó 64 muestras). Para cada señal, su señal envolvente es definida como Log(P_k / thresh), donde P_k es el valor de la potencia de la trama "k" y "thresh" es el umbral de actividad vocal proporcionado por el VAD. Para cada trama donde el valor de la potencia es inferior al umbral, el valor de la envolvente es puesto a cero.

En esta recomendación, la búsqueda del retardo se realiza mediante la función de correlación cruzada. Una primera estimación del retardo bruto se obtiene mediante la correlación cruzada de las envolventes de la señal de referencia y de la señal degradada completas. El índice del máximo de la correlación cruzada indica el retardo entre ambas señales, que a partir de ahora denominaremos "delay1". En la figura 3.4 puede observarse una señal de ejemplo y su correspondiente envolvente, obtenidas con matlab.

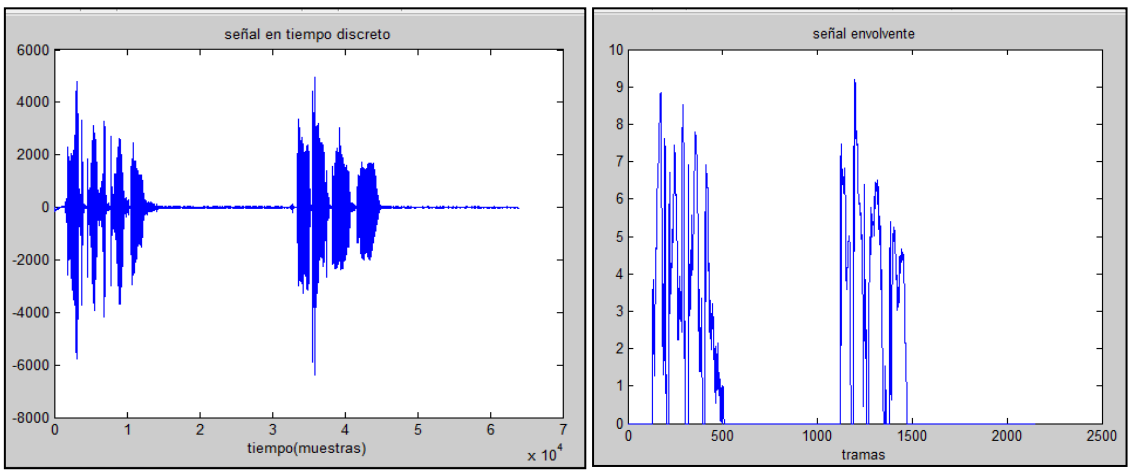


Figura 3.4: Ejemplo de envolvente de una señal

3.3.3. División de la señal de referencia en articulaciones y cálculo del retardo bruto de cada una de ellas

La información proporcionada por el VAD sirve para dividir la señal de referencia en articulaciones. Las articulaciones son segmentos de voz separados entre sí por períodos de silencio. Un segmento de voz debe tener un período de tiempo mínimo durante el cual el valor de su envolvente es no nulo para considerarse una articulación. En esta implementación se considera el caso de un mínimo de 50 tramas de 4 ms consecutivas, es decir, 0.2 segundos. En la señal de la figura 3.4 se pueden apreciar claramente 2 articulaciones.

De cada articulación se guarda la información sobre la trama de inicio y la trama de parada. Cuando todas las articulaciones están perfectamente identificadas, se procede a buscar el retardo entre éstas y los segmentos correspondientes de la señal degradada. Se conoce el inicio y final de cada articulación, y con ello su tamaño. Se coge un intervalo de la señal degradada del mismo tamaño que cada articulación. El inicio y final de dicho intervalo es igual al inicio y final de la articulación más el retardo bruto "delay1" obtenido en el apartado anterior. Una vez identificado el segmento correspondiente a cada articulación en la señal degradada, se realiza la correlación cruzada entre las envolventes de las articulaciones y las envolventes de las correspondientes partes de la señal degradada. De nuevo, el índice del máximo proporciona el retardo, que denominaremos "delay2". El retardo bruto para cada articulación es igual a la suma del retardo "delay2" (que es distinto para cada articulación) más el retardo "delay1" hallado en el apartado anterior. Este retardo está expresado en tramas, por lo que tiene un error de resolución de 32 o 64 muestras, dependiendo de la frecuencia de muestreo. Para obtener el retardo en muestras, basta con multiplicar por el tamaño de trama correspondiente.

3.3.4. Alineamiento fino

Hasta ahora, el retardo de cada articulación ha sido calculado de forma aproximada, basándose en las envolventes de las señales. Además, no teníamos suficiente resolución y no se podía saber la muestra exacta del retardo. Por ello, se procede a realizar una búsqueda más fina y exacta del retardo de las articulaciones.

Para cada articulación disponemos de su retardo bruto, su comienzo y su final. Desde el comienzo hasta el final de cada articulación, se toman tramas de 64 ms (512 ó 1024 muestras), con un solapamiento del 75% entre tramas consecutivas. A estas tramas se les aplica una ventana de Hanning de la misma longitud. El proceso para buscar los correspondientes segmentos en la señal degradada es el mismo que para las articulaciones: el inicio y final de cada segmento es el mismo que el de cada trama más el retardo bruto de la articulación a la que pertenece. Se realiza la correlación cruzada entre las tramas de la señal de referencia con las correspondientes tramas de la señal degradada, hallándose el retardo para cada trama. El máximo de la correlación elevado a 0.125 da una medida de la confianza de la alineación.

Dentro de cada articulación, hay varias tramas, con lo cual hay varias estimaciones del retardo y varias confianzas para cada trama. Se realiza un histograma de estas estimaciones, ponderado por la medida de confianza y se suaviza con la convolución de un kernel triangular simétrico de 1 ms de ancho. El retardo final para cada articulación es igual a la suma del retardo bruto anterior que ya teníamos más el retardo que nos da el índice del máximo del histograma. El máximo del histograma, dividido por la suma total del histograma antes de la convolución con el kernel, da una medida de confianza entre 0 (ninguna confianza) y 1 (plena confianza).

Las articulaciones encontradas hasta ahora solo comprenden los tramos de voz que superan un cierto umbral de potencia, pero también hay tramos de señal silenciosos que hay que asignar a una articulación y que no era necesario incluir anteriormente para correlar las señales. Cada muestra "silenciosa" es asignada a la articulación más cercana y así todas las muestras de la señal de voz pertenecen a una articulación. El resultado final del proceso de alineamiento nos da la información acerca del número de articulaciones y, para cada una de ellas, la muestra de comienzo, la de parada, el retardo y la confianza del retardo para trasladar la información al modelo de percepción y corregir el desalineamiento en cada trama de la señal.

3.3.4. División de las articulaciones

Es posible que algunas de las articulaciones halladas sean muy largas y que se produzcan cambios extremos en el retardo durante su duración. Por ello, se realiza una división iterativa como la mostrada en la figura 3.5. Se comprueba si cada articulación supera una duración umbral (0.8 segundos). En caso afirmativo, se realizan cortes en la articulación, siendo del mismo tamaño los segmentos que quedan entre dos cortes. Se busca el mejor corte de la articulación tal que la alineación de las dos partes con sus homólogas de la señal degradada producen un valor de confianza mayor para cada parte. El proceso de alineación de los apartados anteriores se repite para cada parte. La confianza de cada parte ha de ser superior a la confianza de la articulación entera y se ha de tener una diferencia de retardo entre las dos partes considerable (unos 4 ms). Si todo esto se cumple, se divide la articulación por el punto de corte escogido y el retardo y la confianza de cada parte son los obtenidos en la nueva alineación de las partes. Se repite el proceso de forma iterativa hasta completar todas las articulaciones. El proceso de división de las articulaciones permite tratar los cambios en el retardo tanto en los periodos de habla como en los de silencio.

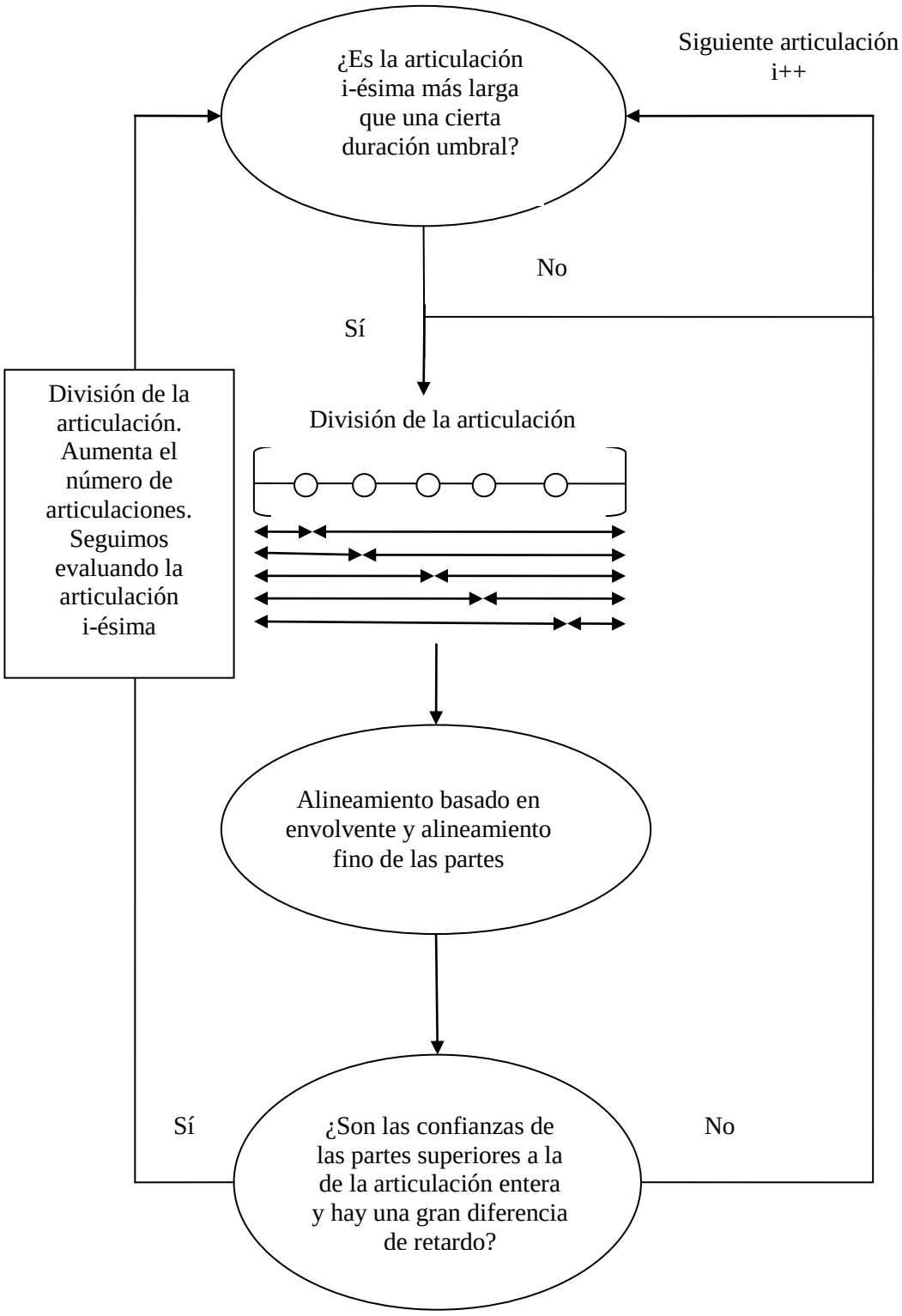


Figura 3.5: Proceso de división de las articulaciones.

3.4. Modelo de percepción

El modelo de percepción utiliza la información de los retardos obtenida en el alineamiento para hacer corresponder a cada trama de la señal de referencia su correspondiente trama de la señal degradada. Después, una distancia es calculada entre ambas señales y expresada mediante la nota pesq. Ésta se pasa por una función monotónica para obtener una predicción de la MOS subjetiva para una determinada prueba subjetiva. A continuación se puede observar un esquema general del modelo de percepción, que se explica más en detalle en los apartados siguientes.

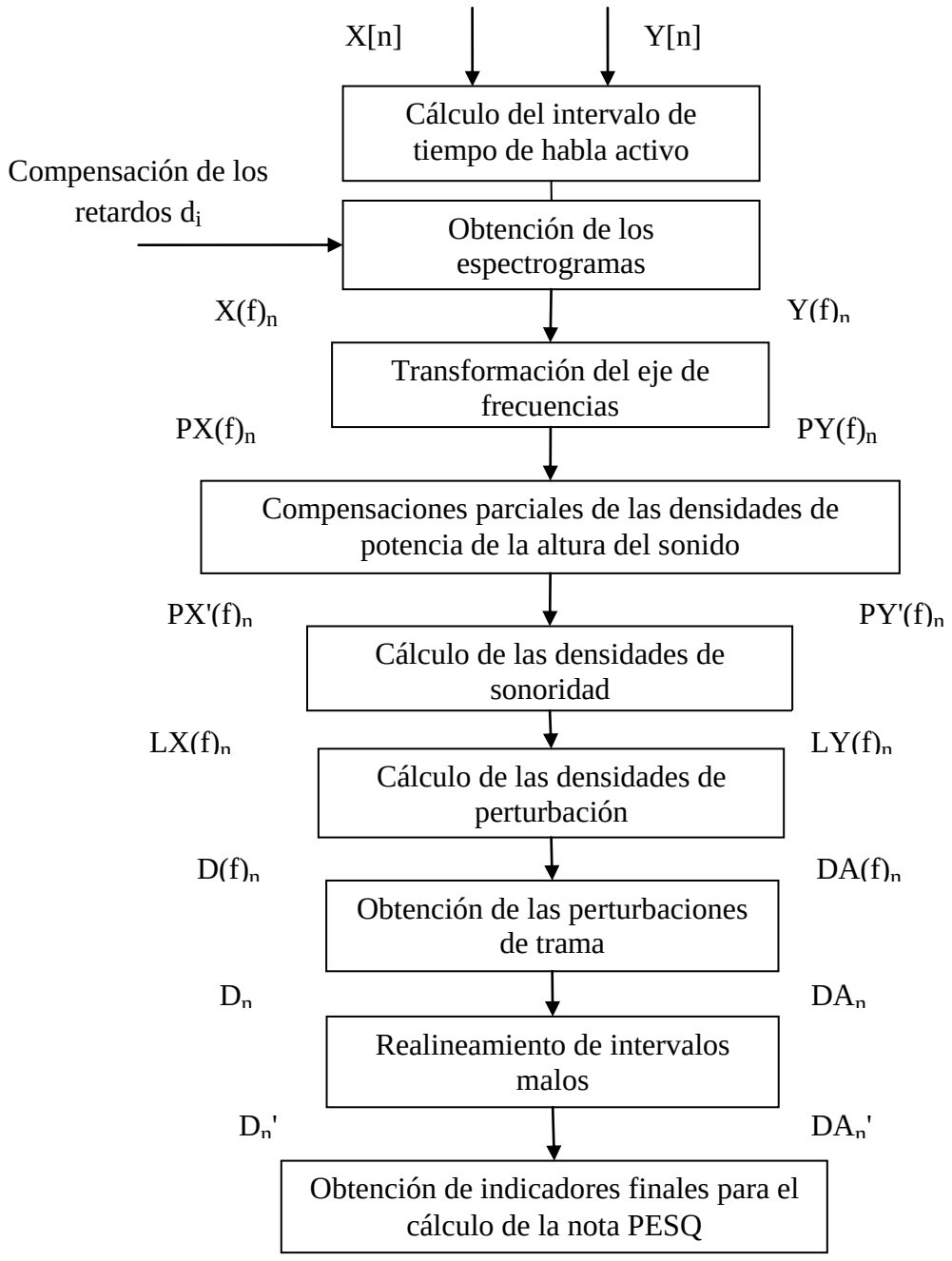


Figura 3.6: Esquema general del modelo de percepción

3.4.1. Cálculo del intervalo de habla activo

En primer lugar, se calcula el intervalo de tiempo de habla activo de la señal inicial. Ésta puede comenzar o finalizar con largos períodos de silencio que podrían afectar al cálculo de los valores promedio de distorsión. Por ello, algunos de los cálculos realizados en el modelo de percepción deben limitarse a este intervalo de habla activo. El comienzo y fin de este intervalo debe ser aquel en el cual la suma del valor absoluto de 5 muestras consecutivas sea superior a 500 (umbral escogido).

3.4.2. Representación tiempo-frecuencia

El oído humano realiza una transformación tiempo-frecuencia. Por ello, se calcula los espectros de ambas señales en función del tiempo (espectrogramas). Esto se realiza enventanando las señales con tramas de 32 ms, es decir, 256 muestras/trama para el caso de 8 kHz y 512 muestras/trama para 16 kHz. El enventanado se realiza con una ventana de Hanning y el solapamiento entre tramas sucesivas es del 50%. Los puntos de comienzo de las ventanas en la señal degradada se desplazan por el valor del retardo correspondiente (cada trama pertenece a una articulación). Se realiza la FFT a cada trama y nos quedamos con el valor cuadrático de la amplitud. La información de fase es descartada ya que el oído humano es mucho más sensible a la información de potencia de un sonido que a su fase. Los resultados se guardan en dos arrays bidimensionales $X(f)_n$ y $Y(f)_n$.

3.4.3. Transformación del eje de frecuencias

Los espectros obtenidos están en una escala frecuencial en hertzios. La escala frecuencial que mejor se corresponde con el sistema de percepción auditivo humano es la escala bark. Esta escala divide las frecuencias en una serie de bandas críticas, las cuales determinan la resolución frecuencial que tiene el oído humano para distinguir sonidos diferentes. Si dos tonos se encuentran dentro de una misma banda crítica serán percibidos como uno solo. El oído humano tiene mayor resolución a bajas frecuencias (las bandas críticas son más estrechas). La escala bark tiene 24 bandas más una adicional para frecuencias mayores de 16 kHz y cada banda equivale a 1 bark.

La transformación de frecuencias de hertzios a barks que se realiza aquí no es exactamente la misma que se puede encontrar en la literatura. Además se divide el rango frecuencial en más de 24 bandas. 42 bandas para el caso de 8 kHz y 49 para el caso de 16 kHz. La figura 3.7 muestra gráficamente la función de conversión y la anchura de cada banda.

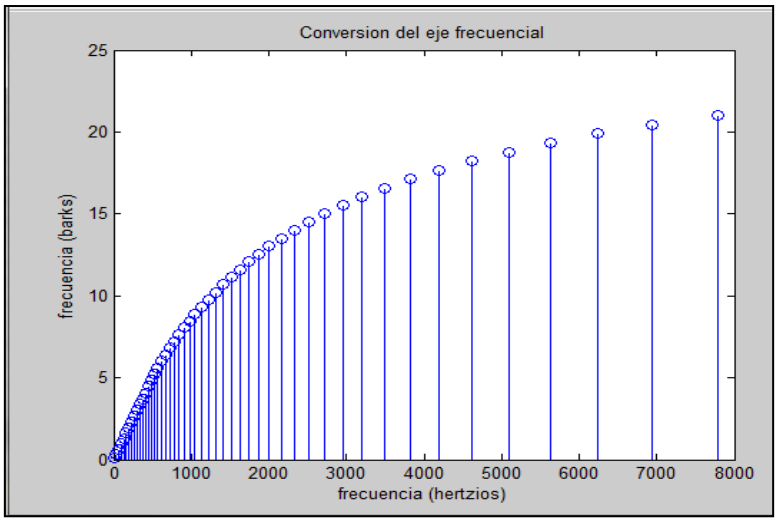
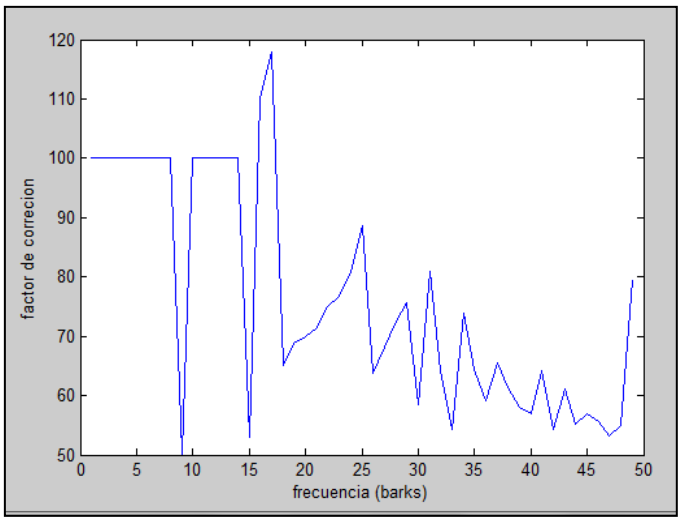


Figura 3.7: Deformación del eje de frecuencias a la escala bark

Las bandas FFT de las señales $X(f)_n$ y $Y(f)_n$ son confinadas dentro de bandas bark, donde la potencia de cada banda es igual a la suma de las bandas que confina. Después, cada banda es multiplicada por dos factores: un factor de corrección que depende de la banda, y un factor de escala de potencia Sp. Este factor es el que se obtiene al calibrar una onda senoidal de 1 kHz y 40 dB SPL de amplitud, para que tras enventanarla, realizar la FFT y confinar sus bandas en bandas bark, la amplitud de dicho tono siga siendo de 40 dB SPL. Ambos factores se pueden observar en la figura 3.8.



Frecuencia de muestreo (kHz)	Sp
8	$2.764344 * 10^{-5}$
16	$6.910853 * 10^{-6}$

Figura 3.8: Factores de corrección en el proceso de deformación del eje de frecuencias

Las señales resultantes se conocen como densidades de potencia de la altura del sonido y se guardan en otros dos arrays bidimensionales $PX(f)_n$ y $PY(f)_n$, donde una dimensión contiene las tramas temporales y la otra dimensión divide cada trama en sus bandas bark.

3.4.4. Compensación parcial de la densidad de potencia de la altura del sonido de la señal de referencia para la ecualización de la función de transferencia

Para cada trama de la señal de referencia, se evalúa si es una trama activa o silenciosa. Una trama se considera activa si la suma total de potencias de todas sus bandas supera cierto umbral. Para el cálculo de esa potencia total sólo se contabilizan aquellas bandas cuya potencia supera un umbral mínimo de audición. El umbral mínimo de audición de una persona depende de la frecuencia e indica la mínima potencia que debe tener un sonido para ser percibido. Para frecuencias bajas (20-1000 Hz) el umbral es muy alto. Después hay un margen donde el umbral es bajo y aproximadamente constante (1000-4000 Hz). A partir de 4 kHz el umbral aumenta poco a poco. En PESQ se realiza una interpolación del umbral de audición para obtener su valor en las frecuencias bark utilizadas. La potencia de cada banda bark modificada debe estar 20 dB por encima de su umbral de audición correspondiente para contribuir a la potencia total. Si la potencia total de una trama supera los 70 dB, la trama se considera activa.

Para tratar el efecto del filtrado, se obtiene el valor promedio de potencia para cada banda bark de ambas señales. A este promedio sólo contribuyen las tramas activas cuya banda supere 20 dB el umbral de audición. Para obtener una estimación del filtrado del sistema sometido a prueba, se realiza el cociente entre el valor de potencia promedio de la señal degradada y la de referencia, obteniendo una respuesta frecuencial en función de cada banda bark. Esta función es acotada entre -20 y +20 dB. Se multiplica la densidad de potencia de la altura del sonido de cada trama de la señal de referencia por esta respuesta frecuencial para ecualizar la señal inicial a la degradada, obteniéndose $PX'(f)_n$. Esta compensación parcial se utiliza porque un filtrado intenso puede ser molesto para el oyente.

3.4.5. Compensación parcial de la densidad de potencia de la altura del sonido degradada para tener en cuenta variaciones de la ganancia en función del tiempo entre la señal de referencia y la degradada

Las variaciones de ganancia de corta duración son compensadas parcialmente procesando las densidades de potencia de la altura del sonido trama por trama. Para ambas densidades, se calcula para cada trama la potencia total audible, siendo esta la suma de las potencias de las bandas que rebasan el umbral de audición. Se calcula el cociente entre la potencia total audible de la señal de referencia y la potencia total audible de la degradada, limitando el resultado entre los valores $[3 * 10^{-4}, 5]$. A este cociente se aplica un filtro paso bajo de orden inferior (a lo largo del eje del tiempo). La densidad de potencia de la altura del sonido de la señal degradada es multiplicada por este cociente, trama a trama, obteniéndose la densidad de potencia de la altura del sonido degradada con compensación parcial de la ganancia $PY'(f)_n$.

3.4.5. Cálculo de las densidades de sonoridad

La sonoridad es la medida subjetiva de la intensidad con la que un sonido es percibido por el oído humano. Es dependiente de la frecuencia, por lo que un sonido de cierto nivel de potencia, será percibido con mayor o menor sonoridad dependiendo de la frecuencia. La sonoridad se mide en sones y su relación con la frecuencia se puede encontrar en las curvas establecidas por Munson y Fletcher [9].

Como la sonoridad es la verdadera magnitud de percepción del oído humano, las densidades de potencia de la altura del sonido son transformadas de una escala de potencia a una escala de sonoridad, utilizando la ley de Zwicker.

$$LX(f)_n = S_l * \left(\frac{P_0(f)}{0.5}\right)^{\gamma} * \left[\left(0.5 + 0.5 * \frac{PX'(f)_n}{P_0(f)}\right)^{\gamma} - 1\right] \qquad (3.1)$$

donde $P_0(f)$ es el umbral absoluto de audición, γ es la potencia Zwicker, $PX(f)_n$ la densidad de potencia de la altura del sonido y S_l es el factor de escala de sonoridad. Este factor es obtenido con el mismo tono de referencia del apartado 3.4.3. Después de obtener su densidad de potencia en función de la frecuencia (en la escala bark) se utiliza la ley de Zwicker para obtener su densidad de sonoridad. Su integral a lo largo del eje de frecuencias, multiplicada por el factor S_l, debe ser igual a 1. Para ambas frecuencias de muestreo, el valor de S_l resulta $1.866055 * 10^{-1}$.

Por encima de 4 barks, la potencia Zwicker vale 0.23 (valor dado en la literatura). Por debajo de 4 barks, se aumenta algo para tener en cuenta el efecto denominado de reclutamiento, pero acotando a 0.2552 el valor máximo.

$$\gamma \begin{cases} 0.23 & \text{f} \geq 4 \\ 0.23 * \left(\frac{6}{f(barks)+2}\right)^{0.15} & \text{f} < 4 \end{cases}$$

Figura 3.9: Valor de la potencia Zwicker en función de la frecuencia

Los arrays bidimensionales resultantes $LX(f)_n$ y $LY(f)_n$ son las densidades de sonoridad.

3.4.6. Cálculo de la densidad de perturbación

Para cada trama se calcula la diferencia con signo entre la densidad de sonoridad distorsionada y la inicial (array denominado densidad de perturbación), y el mínimo de ambas (array denominado zona muerta). La zona muerta se multiplica por 0.25. Después, en cada trama se procede así:

-A cada valor de la densidad de perturbación que es superior al de la zona muerta se le resta el valor de la zona muerta.

-A cada valor de la densidad de perturbación que es inferior a menos el valor de la zona muerta, se le suma el valor de la zona muerta.

-Cuando el valor de la densidad de perturbación es menor o igual que el valor absoluto de la zona muerta, la densidad de perturbación es puesta a 0.

El efecto es que las densidades de perturbación tienden a desplazarse a cero. Con esto se modela el tratamiento de las pequeñas diferencias que no son audibles en presencia de señales de gran sonoridad (enmascaramiento). El resultado de este paso es un array bidimensional de la densidad de perturbación $D(f)_n$.

3.4.7. Multiplicación por un factor de asimetría

Cuando un códec deforma la señal de entrada, es muy difícil que vuelva a reconstruirla sin producir una distorsión claramente audible, pudiéndose descomponer la señal como suma de señal más distorsión. Cuando lo que se produce es una pequeña pérdida de una componente de tiempo-frecuencia, la distorsión se nota menos. Este efecto se modela calculando una densidad de perturbación asimétrica $DA(f)_n$ por trama, mediante el producto de la densidad de perturbación por un factor de asimetría. Este factor es igual al cociente de las densidades de potencia de la altura del sonido degradada e inicial, elevado a 1.2. Si es mayor que 12 se fija a 12, mientras que si es inferior a 3 se fija a 0

3.4.8. Obtención de las perturbaciones de trama

Las densidades de perturbación se integran a lo largo del eje de la frecuencia mediante dos normas diferentes:

$$D_n = M_n * W * \sqrt{\frac{\sum_{f=1}^{Nbandasbark}\left(|D(f)_n| * W_f\right)^2}{W}} \tag{3.2}$$

$$DA_n = M_n * W * \frac{\sum_{f=1}^{Nbandasbark}\left(|DA(f)_n| * W_f\right)}{W} \tag{3.3}$$

siendo $M_n = \left(\frac{1*10^7}{pot.totalaudible.ref+1*10^5}\right)^{0.04}$ un factor que produce una acentuación

de las perturbaciones producidas durante períodos de silencio. W_f es el ancho de banda de cada banda bark y W el ancho de banda total. Los valores de perturbación de trama son limitados a un máximo de 45.

3.4.9. Puesta a cero de las perturbaciones de trama en las tramas durante las cuales el retardo aumenta apreciablemente y realineación de intervalos malos

Si la variación del retardo de la señal degradada respecto de la referente de una trama a otra es superior a media ventana (16 ms), los valores de perturbación de esa trama y de las siguientes que estén dentro de ese tiempo de variación son puestos a cero. Esto se debe a la observación de que a efectos de determinar la calidad del habla por medios objetivos es mejor no tener en cuenta las perturbaciones durante tales efectos.

Si durante varias tramas consecutivas el valor de la perturbación de trama es superior a un umbral, ese intervalo es considerado como malo. El valor de perturbación puede ser alto como consecuencia de un mal alineamiento. Por ello, se procede al realineamiento de dichos intervalos realizando la correlación cruzada entre el valor absoluto del segmento de señal perteneciente al intervalo malo de referencia y el valor absoluto de su intervalo homólogo de la señal degradada ya ajustado de acuerdo a los retardos obtenidos previamente. Si la confianza de la correlación es baja, se decide que se está concordando ruido con ruido y se deja de calificar al intervalo como malo. En caso contrario, se vuelven a realinear los intervalos malos con los retardos obtenidos y se vuelve a recalcular las perturbaciones de trama para las tramas pertenecientes a dichos intervalos. Si son menores, sustituyen a las calculadas anteriormente. Las nuevas densidades de perturbación de trama se guardan en dos nuevos arrays D_n' y DA_n'.

3.4.10. Obtención de los indicadores de perturbación finales y de las notas PESQ y MOS-LQO

Para el cálculo final, se consideran sólo las tramas pertenecientes al intervalo de tiempo de habla activo. Se realiza el mismo algoritmo para las dos perturbaciones de trama que, por motivos de sencillez, se presenta a modo de esquema en la figura 3.10, donde tw_n es un factor de peso que depende de la trama. Si la señal completa no es muy larga, este factor vale 1 independientemente de la trama. Si tiene más tramas que un determinado umbral, el peso de cada trama va creciendo poco a poco.

Se obtienen así dos indicadores finales de perturbación indd1 e indd2 según se utilice D_n ó DA_n en la figura 3.10, que mediante una ponderación nos proporcionan la nota PESQ.

$$NotaPESQ = 4.5 - 0.1 \times indd1 - 0.0309 \times indd2 \quad (3.4)$$

cuyo valor está comprendido entre -0.5 y 4.5.

Como ya se ha comentado, esta nota se pasa por una función monotónica para obtener la nota MOS-LQO, que es la predicción de la nota de un experimento subjetivo obtenida de forma objetiva. La función de correspondencia es la siguiente:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945x + 4.6607}} \quad (3.5)$$

que proporciona un rango de valores de salida entre 1 y 4.5 (no alcanza a 5 porque en un experimento subjetivo dar un 5 es decir que ambas señales son iguales y en la práctica casi nunca se da).

$rt, ttw = 0$

for n = trama de comienzo; paso de 10 tramas; hasta la trama de parada

$$rs = \sqrt[6]{\frac{\sum_n^{n+19} (D_n{}')^6}{20}}$$

$$rt = rt + (tw_n * rs)^2$$

$$ttw = ttw + (tw_n)^2$$

$$indicador\ final = \sqrt{\frac{rt}{ttw}}$$

Figura 3.10: Algoritmo de obtención de los indicadores finales de perturbación

4. Metodología de evaluación de los vocoders

Como ya se ha comentado, se ha analizado la calidad perceptual de cada vocoder en resíntesis, ya que el objetivo es comparar el nivel de calidad de cada vocoder de forma individual y aislada del proceso de síntesis con HMM. La única excepción es la de STRAIGHT, comentada anteriormente, en la que hay una transformación de parámetros entre el análisis y la síntesis que hace más eficiente el proceso de entrenamiento con HMM y que se ha tenido en cuenta a la hora de evaluar su calidad.

Para la evaluación, se ha utilizado el audio de la base de datos de voz en español Albayzin. Su corpus fonético está dividido en dos grupos: subcorpus de aprendizaje y subcorpus de prueba. El subcorpus de aprendizaje está formado por 200 frases diferentes. 4 locutores pronuncian las 200 frases y 160 locutores 25 frases cada uno. Un total de 4800 frases, las 200 frases repetidas 24 veces por diferentes locutores. El subcorpus de prueba está constituido por 500 frases distintas, en las que 40 locutores distintos al subcorpus de aprendizaje pronuncian 50 frases cada uno. Un total de 2000 frases.

El número total de ficheros que contiene la base de datos es por lo tanto de 6800, que contienen sonidos del castellano hablado de forma estadísticamente equilibrada. Estos ficheros son ficheros binarios cuya extensión es ".ses" y cuyas muestras están cuantificadas con enteros de 16 bits con signo a una frecuencia de muestreo de 16 kHz.

El trabajo realizado en lo que respecta a este punto de evaluación de los vocoders ha sido la realización del software en matlab necesario para convertir la base de datos de archivos binarios en una base de datos con archivos ".wav" reproducibles. Esta base de datos ha sido resintetizada con los tres vocoders en estudio, generando otras tres nuevas bases de datos con los archivos ".wav" de las frases resintetizadas. Finalmente, se ha utilizado PESQ para calcular la nota MOS-LQO de cada frase resintetizada de cada vocoder. Como resultado final se obtiene un fichero de texto con todas las notas MOS-LQO de cada señal y una variable matricial que puede ser cargada en matlab con los valores de todas las notas.

Este software se ha ejecutado en un sistema linux con matlab instalado, debido a que sólo se disponía de un ejecutable para linux de AHO-coder. Sin embargo, para los otros dos vocoders puede realizarse la resíntesis también en Windows (con matlab). Para ejecutar parte de los vocoders desde matlab se ha utilizado el signo "!" de exclamación delante de algunas sentencias. Este signo permite ejecutar la sentencia que le sigue como una llamada de comandos del sistema operativo.

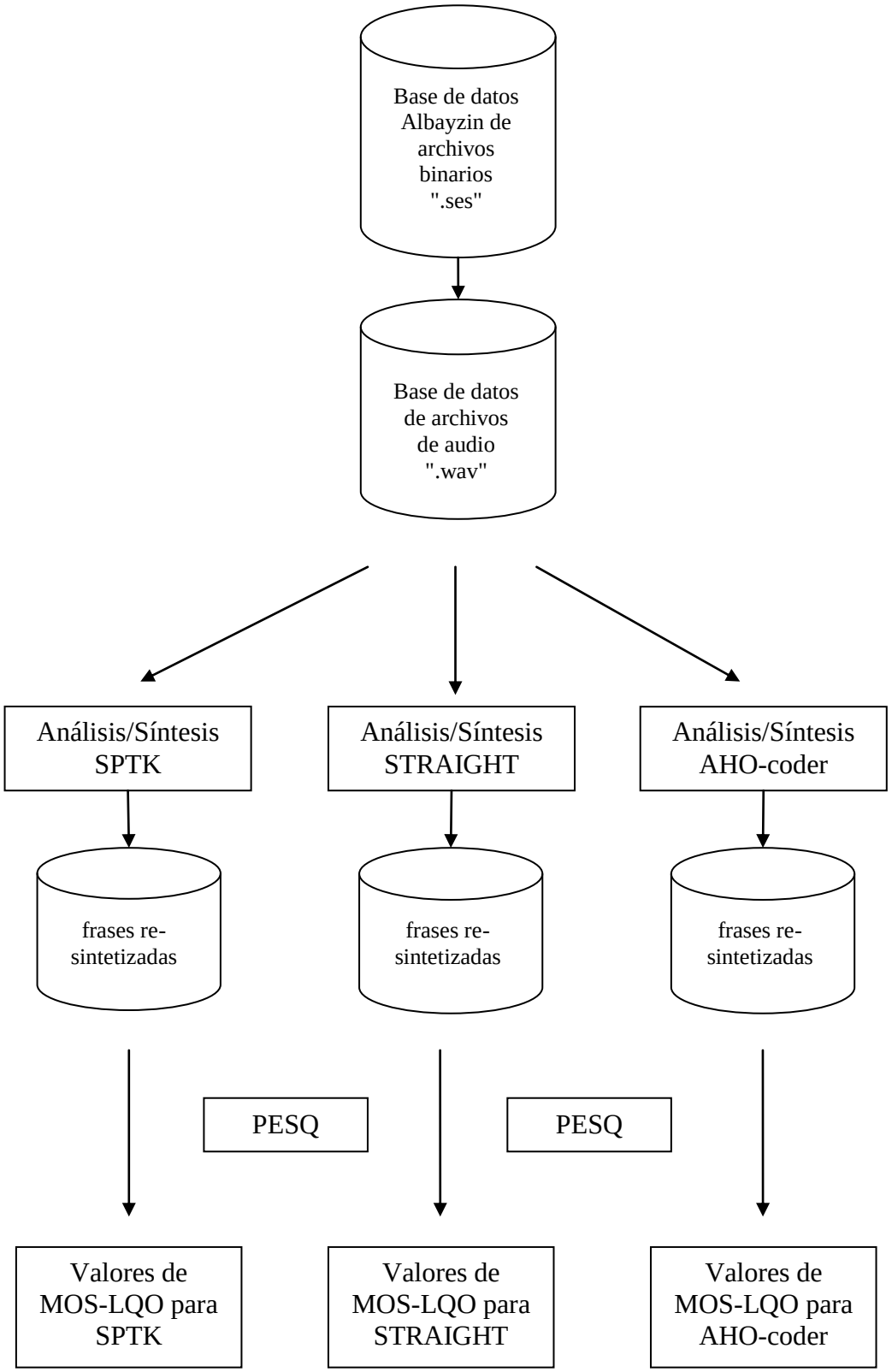


Figura 4.1: Esquema de evaluación de los vocoders

Se ha ejecutado el análisis de los tres vocoders con los siguientes parámetros: para el rango de búsqueda de la frecuencia de pitch se ha elegido en todos los casos el rango 60-500 Hz para obtener buenos resultados tanto en mujeres como en hombres. El desplazamiento temporal de trama se ha escogido de 5 milisegundos para SPTK y AHO-coder, mientras que para STRAIGHT ha sido necesario bajarlo a 1 ms para poder realizar un alisado temporal de la señal correcto en el proceso de obtención de la envolvente. Por último, el orden de los coeficientes mel-cepstrum obtenidos en el análisis es de 39 para todos los casos.

En realidad, STRAIGHT proporciona 513 coeficientes de aperiodicidad y otros 513 de la envolvente del espectro de la señal (para 16 kHz), pero son demasiados coeficientes para entrenar con ellos. Por lo tanto, se obtienen 39 coeficientes mel-cepstrum a partir del espectro y 5 coeficientes de aperiodicidad agrupando todos y promediando en 5 subbandas. Esto se ha realizado utilizando funciones de SPTK. Para realizar la síntesis se vuelven a obtener los coeficientes del espectro y de aperiodicidad (513 coeficientes) a partir de los parámetros usados para entrenar.

A modo de resumen, a continuación se muestra una tabla con los parámetros extraídos de cada vocoder para cada trama de señal.

SPTK
39 coeficientes mel-cepstrum
1 coeficiente del pitch
STRAIGHT
39 coeficientes mel-cepstrum
1 coeficiente del pitch
5 coeficientes de aperiodicidad
AHO-coder
39 coeficientes mel-cepstrum
1 coeficiente del pitch
1 coeficiente de la MVF

Figura 4.2: Parámetros extraídos de cada vocoder

5. Resultados

Realizando el proceso de resíntesis a la base de datos con los 3 vocoders y los parámetros definidos en el apartado anterior, se obtienen los resultados visibles en la figura 5.1.

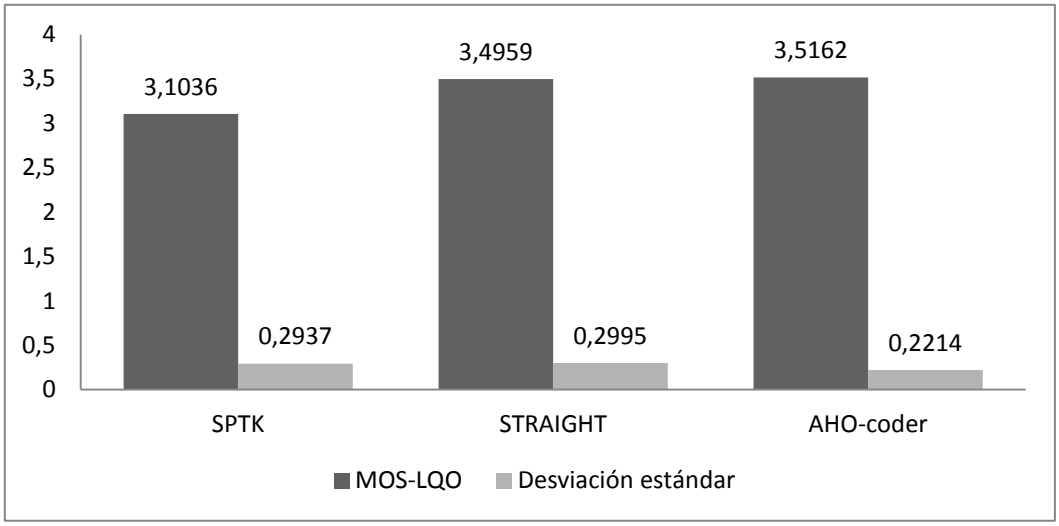


Figura 5.1: Valores de calidad MOS-LQO obtenidos para los 3 vocoders

Se puede ver que el vocoder que proporciona una mayor calidad de resíntesis es AHO-coder, con una nota de 3.52. Sin embargo, no se puede decir que sea mejor que STRAIGHT, ya que la diferencia entre ambos es de apenas 0.02. El que queda demostrado ser claramente inferior es SPTK, con una puntuación de 3.10. En cuanto a las desviaciones estándar, para SPTK y STRAIGHT está entre 0.29 y 0.3 mientras que AHO-coder la presenta ligeramente más baja, de 0.22.

La base de datos Albayzin tiene un sistema de etiquetado de sus frases que permite dividir a los hablantes en función de su sexo y edad. Esto permite realizar un análisis más detallado de la calidad de los vocoders. En la tabla 5.1 puede observarse el valor de la nota MOS-LQO y su desviación estándar para cada agrupación que permite hacer la base de datos Albayzin.

En ella puede verse como SPTK es el único vocoder que funciona mejor para hombres que para mujeres. STRAIGHT es el que mayor diferencia de calidad presenta entre hombres y mujeres, siendo AHO-coder el que más equilibrio presenta en este aspecto. Esta gran diferencia entre hombres y mujeres para STRAIGHT hace que incluso tenga un nivel de calidad superior que AHO-coder para el caso de las mujeres.

Grupo	SPTK MOS-LQO/desv.std	STRAIGHT MOS-LQO/desv.std	AHO-coder MOS-LQO/desv.std
Hombres	*3.2186/ 0.2993*	*3.3623/ 0.3168*	*3.4659/ 0.2450*
Mujeres	*2.9887/ 0.2378*	*3.6294/ 0.2083*	*3.5666/ 0.1816*
Hombres entre 18-30 años	*3.1741/ 0.3081*	*3.3299/ 0.3339*	*3.4459/ 0.2548*
Hombres entre 31-40 años	*3.3032/ 0.2609*	*3.4198/ 0.3098*	*3.5144/ 0.2279*
Hombres entre 41-55 años	*3.2464/ 0.2912*	*3.3867/ 0.2652*	*3.4681/ 0.2288*
Mujeres entre 18-30 años	*3.0107/ 0.2331*	*3.6373/ 0.2024*	*3.5685/ 0.1762*
Mujeres entre 31-40 años	*2.9972/ 0.2300*	*3.6481/ 0.1987*	*3.5838/ 0.1894*
Mujeres entre 41-55 años	*2.9242/ 0.2458*	*3.5907/ 0.2270*	*3.5445/ 0.1850*

Tabla 5.1: Valores de calidad clasificados en grupos según la edad y el sexo

Si nos fijamos en la relación que hay entre STRAIGHT y AHO-coder, podemos ver que tenemos la misma tendencia si ordenamos los grupos de mayor a menor calidad. La principal diferencia entre ambos es el rango en que se mueve la nota MOS-LQO. Para ambos vocoders, el grupo con el que mejor resultado se obtiene es el formado por las mujeres de entre 31 y 40 años, siendo el peor el formado por hombres de entre 18 y 30 años. Para SPTK el grupo que obtiene una mayor calidad es el formado por hombres de entre 31 y 40 años, siendo el peor el formado por mujeres de entre 41 y 55 años.

En base a estos resultados, se puede afirmar que AHO-coder y STRAIGHT ofrecen una calidad muy similar. El pico máximo de calidad se obtiene con STRAIGHT para mujeres de entre 31 y 40 años. Pero el tiempo de análisis de STRAIGHT es mayor y el tiempo de entrenamiento que requiere también es mayor al tener más parámetros, lo que hace que sea más pesado de utilizar en una síntesis con HMM. El más rápido computacionalmente es SPTK, pero la calidad que ofrece es notablemente menor. Una ventaja adicional de AHO-coder sobre STRAIGHT es que es más sencillo de utilizar, está escrito en C y se ejecuta fácilmente con dos comandos en un sistema Linux. Para ejecutar STRAIGHT es necesario matlab. Es por todo ello que podemos considerar a AHO-coder el vocoder óptimo entre los 3 estudiados.

También puede resultar llamativo que un códec que utiliza una excitación mezclada tan simple como en la que para frecuencias bajas se generan impulsos periódicos y para frecuencias altas se genera ruido gausiano proporcione tanta calidad como una excitación mezclada más completa como es el caso de STRAIGHT.

Al escuchar diferentes frases resintetizadas con los 3 vocoders se aprecia claramente que las de peor calidad son las de SPTK, mientras que es casi imposible apreciar la diferencia entre AHO-coder y STRAIGHT.

Para terminar con el trabajo, se han tomado una serie de medidas cuyo único fin es cuantificar el efecto que tiene el realizar un par de cambios sobre los vocoders. Estas pruebas solo han sido efectuadas sobre SPTK y STRAIGHT y básicamente son las siguientes:

Para SPTK se ha realizado la resíntesis reduciendo el orden de los coeficientes mel-cepstrum a 24. También se ha cuantificado la mejora que tiene el hecho de utilizar excitación mezclada en frecuencia respecto a no utilizarla. Por último, y por mera curiosidad, se ha realizado la resíntesis de STRAIGHT sin hacer la transformación de parámetros explicada anteriormente, necesaria para entrenar. El objetivo de esta última medida es averiguar qué nivel de calidad máximo es capaz de ofrecer STRAIGHT, sin importar el coste computacional.

Los valores globales y desglosados por grupos pueden verse en la figura 5.2 y en la tabla 5.2 respectivamente.

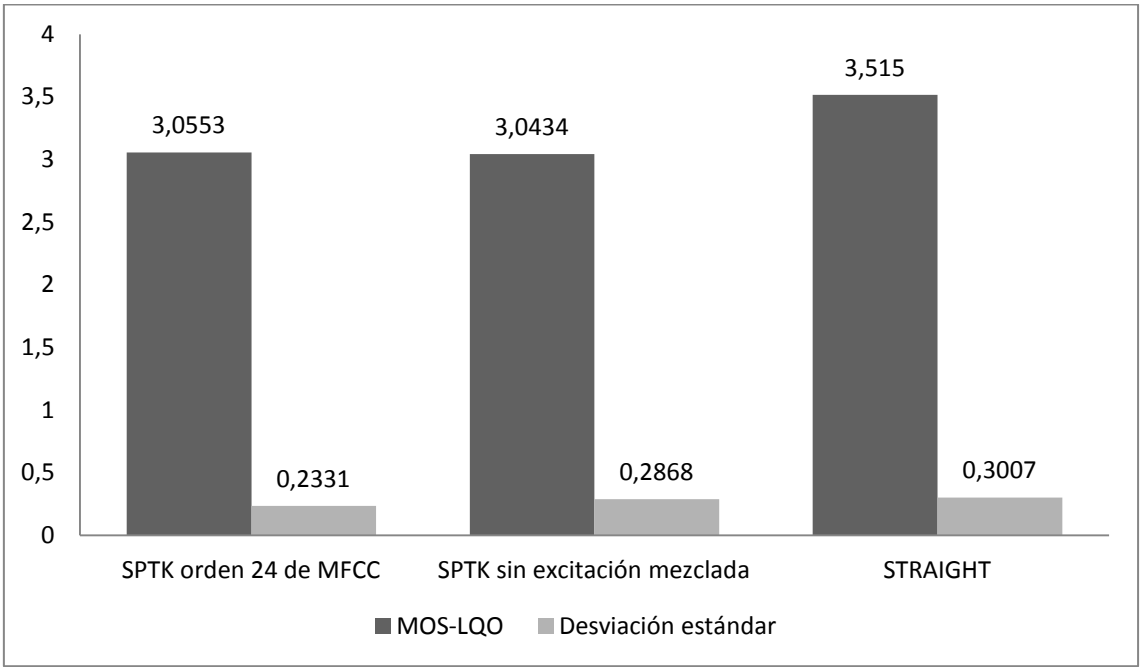


Figura 5.2: Medidas de calidad adicionales

De estos resultados se puede concluir que el empeoramiento de STRAIGHT debido a la transformación de coeficientes es apenas imperceptible, de apenas 0.02. La mejora obtenida por extraer 39 coeficientes en lugar de 24 es de aproximadamente 0.05 y la obtenida por utilizar una excitación mezclada es de 0.06. La tendencia obtenida en la clasificación por grupos es la misma que para las medidas iniciales.

Grupo	**SPTK orden 24 de los mel-cepstrum** MOS-LQO/desv.std	**SPTK sin excitación mezclada** MOS-LQO/desv.std	**STRAIGHT** MOS-LQO/desv.std
Hombres	*3.1225/ 0.2474*	*3.1400/ 0.2991*	*3.3769/ 0.3134*
Mujeres	*2.9881/ 0.1960*	*2.9468/ 0.2375*	*3.6530/ 0.2109*
Hombres entre 18-30 años	*3.1017/ 0.2483*	*3.1014/ 0.3044*	*3.3494/ 0.3288*
Hombres entre 31-40 años	*3.1655/ 0.2365*	*3.2155/ 0.2713*	*3.4306/ 0.3081*
Hombres entre 41-55 años	*3.1322/ 0.2500*	*3.1621/ 0.2957*	*3.3926/ 0.2676*
Mujeres entre 18-30 años	*2.9778/ 0.1872*	*2.9722/ 0.2327*	*3.6633/ 0.2010*
Mujeres entre 31-40 años	*3.0170/ 0.2076*	*2.9490/ 0.2313*	*3.6672/ 0.2052*
Mujeres entre 41-55 años	*2.9852/ 0.2032*	*2.8804/ 0.2431*	*3.6126/ 0.2352*

Tabla 5.2: Medidas de calidad adicionales clasificadas en grupos según la edad y el sexo

Por último, y a modo de comentario, indicar que para un desplazamiento de trama de 5 ms STRAIGHT obtiene un valor MOS-LQO de 3.3308, casi 2 décimas por debajo del máximo que puede alcanzar. Esto es debido a que, por su propia implementación, STRAIGHT necesita tener un desplazamiento de trama inferior a 1 ms para poder realizar un alisado correcto del espectro de potencia y así obtener una buena envolvente.

6. Conclusiones y líneas futuras

6.1. Conclusiones del trabajo

Como resultado de este trabajo, se puede concluir que el mejor vocoder de los 3 estudiados, si valoramos la calidad que ofrece, el coste computacional y su sencillez de utilización, es AHO-coder. Este vocoder obtiene una nota MOS-LQO de casi 3.52 de un rango que va desde 1 hasta 4.5. En otras palabras, alcanza una calidad bastante alta para tratarse de un codificador paramétrico.

AHO-coder está además bien enfocado a ser utilizado en sistemas de síntesis de voz basados en HMM, como HTS. Ofrece unos resultados muy buenos tanto para hombres como mujeres y es el que menor desviación típica tiene, lo que lo hace menos sensible a variaciones de calidad. Pero es necesario recordar que STRAIGHT tiene la misma calidad prácticamente, con lo que cualquiera de los dos sería una buena elección.

En este trabajo se ha observado además que una excitación mezclada más sencilla con un buen algoritmo para detectar correctamente la máxima frecuencia de voz (MVF) es tanto o más efectivo que hacer una excitación mezclada tan completa como la que hace STRAIGHT. Por ello, en una excitación mezclada puede ser suficiente con calcular correctamente hasta que frecuencia llega la parte de voz periódica y realizar únicamente una composición o suma de señal sonora a bajas frecuencias y señal sorda a altas frecuencias.

Por último, insistir en que los resultados de calidad de este trabajo han sido obtenidos en el proceso de resíntesis. Este es el modo de evaluar la calidad de un vocoder de forma objetiva, ya que se dispone de una señal original y otra decodificada. Es decir, a priori debería considerarse que AHO-coder funcionaría con buenas prestaciones en un sistema basado en HMM, pero no quiere decir que sea necesariamente así. Sería necesario usarlo en un sistema como HTS por ejemplo, y comprobar si realmente proporciona una alta calidad.

6.2. Líneas futuras

Como líneas futuras de trabajo o estudio a este proyecto podrían citarse 2 principalmente:

La primera sería, una vez que sabemos que AHO-coder y STRAIGHT son los que mejor calidad proporcionan, realizar un estudio más exhaustivo tanto teórico

como práctico a nivel de código (ambos códigos son propietarios y habría que solicitárselo a sus autores para investigación académica) para conseguir las mejores propiedades de cada uno de ellos y obtener un nuevo vocoder que mejore las prestaciones de ambos. No obstante, la solución más sencilla sería partir de SPTK e investigar nuevas mejoras en la señal de excitación para conseguir un vocoder mejor que pueda incluso superar a los dos anteriores.

La otra línea futura estaría más relacionada con los estándares que permiten evaluar objetivamente la calidad vocal de códecs. En este trabajo se ha implementado PESQ pero hay un estándar que le sigue y es más completo. Este estándar es la recomendación P.863 de la ITU, también conocido como POLQA (Perceptual objective listening quality assessment). Además de señales con frecuencias de muestreo de 8 y 16 kHz, también puede evaluar la calidad con señales de banda ancha muestreadas a 48 kHz. Debido a la dificultad y a la gran extensión de esta recomendación, además de que ya se disponía en este trabajo de una base de datos de voz con señales de 16 kHz, AHO-coder sólo funciona a 16 kHz y con PESQ era suficiente, no se llegó a implementar.

Bibliografía

[1] Ivan Folgueira Bande. "*Síntesis de voz mediante Modelos Ocultos de Markov*". Proyecto Final de Carrera de la UPC. Noviembre de 2008.

[2] Takashi Masuko. "*HMM-Based Speech Synthesis and Its Applications*". pp. 5-14, November 2002.

[3] David Talkin. "*A Robust Algorithm for Pitch Tracking (RAPT)*" in "*Speech Coding and Synthesis*". Elsevier, 1995.

[4] Hideki Kawahara. *"STRAIGHT, Exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds"*. Acoustic Science and Technology, Vol. 27, No. 6, pp. 349-353, 2006.

[5] Hideki Banno, Hiroaki Hata, Masanori Morise, Toru Takahashi, Toshio Irino and Hideki Kawahara. *"Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation"*. Acoustic Science and Technology, Vol. 28, No. 3, pp. 140-146, 2007.

[6] D. Erro, I. Sainz, E. Navas, I. Hernaez. "*Improved HNM-based Vocoder for Statistical Synthesizers*". InterSpeech, pp. 1809-1812, Florence, August 2011.

[7] D. Erro, I. Sainz, E. Navas, I. Hernaez. "*HNM-based MFCC+F0 Extractor applied to Statistical Speech Synthesis*". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4728-4731, Prague, May 2011.

[8] D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernaez. *"MFCC+F0 extraction and waveform reconstruction using HNM: preliminary results in an HMM-based synthesizer"*. VI Jornadas en Tecnología del Habla & II Iberian SLTech (FALA), pp. 29-32, Vigo, November 2010.

[9] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.