PRÁCTICA 3.- Codificación de voz.

Duración: 6 horas (3 sesiones)

Conceptos desarrollados: muestreo, predicción lineal, filtrado, transformada de Fourier, correlación y autocorrelación, tratamiento de voz, compresión, síntesis de voz.

Material necesario: auriculares.

3.1.- Introducción.

Se va a estudiar una técnica de codificación digital de la voz que emplea un ancho de banda muy reducido mediante la transmisión de tan sólo unos pocos parámetros (admitiendo que se puede clasificar cada trama de voz -20 a 40 ms- bien como cuasi-periódica cuando se trate de sonidos sonoros, bien como ruido cuando se trate de sonidos sordos). Con estas técnicas, el ancho de banda es disminuido hasta alcanzar tasas inferiores a 1 kbps, cuando la codificación de voz habitual origina tasas de 64 Kbps. Como contrapartida, la calidad de la voz es disminuida hasta un punto en el que resulta complicado reconocer al locutor, pero sí se puede comprender su mensaje.

La voz es un fenómeno en que existe correlación entre sus muestras tanto a largo como a corto plazo. Dicha correlación implica la presencia de cierta redundancia en la señal que ocasiona un empleo de recursos mayor que el que a priori puede ser necesario y, en el caso de transmisión digital, un aumento de la velocidad binaria que puede imposibilitar el uso de ciertos canales con ancho de banda limitado.

La correlación a corto plazo se debe a la lenta variación de la envolvente del espectro de amplitud de la señal de voz. En los sonidos sonoros, las muestras se relacionan unas con otras de forma que puede predecirse aproximadamente el valor de una muestra a partir de las anteriores. De esta forma, una técnica de nominada *predicción lineal* aparece como una solución razonable para la eliminación de redundancia en estas circunstancias.

La correlación a largo plazo está estrechamente ligada a la sonoridad de los sonidos, ya que aquellos clasificables como sonoros poseen un alto grado de periodicidad que de nuevo implica una importante redundancia y excesivo uso de recursos. El conocimiento del período fundamental a partir de algoritmos basados en la predicción permite la eliminación de este tipo de redundancias.

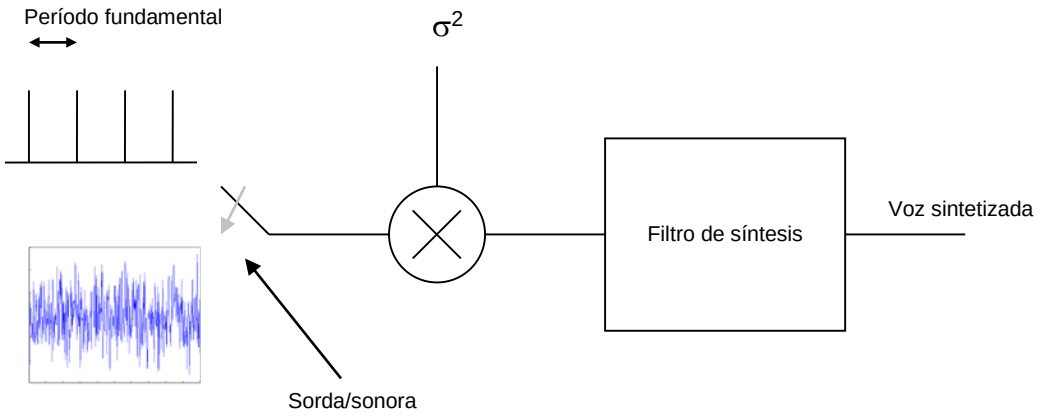
3.2.- Modelo de producción de la voz: síntesis de voz en el receptor.

El codificador de voz que se va a estudiar se basa en el modelado de un sistema físico que es el tracto vocal humano, que actúa sobre el aire que, procedente de los pulmones, atraviesa la glotis para producir voz en los labios. Según este modelo se dividen los sonidos constitutivos de la voz en dos grupos básicos: sonoros y sordos. Para producir un sonido sonoro, el tracto vocal es excitado por una serie de pulsos casi periódicos generados por las cuerdas vocales. Si lo que se obtiene es un sonido sordo, el tracto vocal es excitado por una turbulencia de aire similar al ruido gaussiano.

Con el fin de poder realizar un análisis adecuado de la voz en orden a esta clasificación, es necesario dividirla en segmentos o tramas en los que se puede considerar que las características de la voz permanecen constantes (se modela como un proceso estocástico estacionario). Esta longitud de trama puede ser de unos 20 a 40

milisegundos. Una duración de 30 ms, por ejemplo, con un muestro de 8 KHz se corresponde con una longitud de la trama de 240 muestras.

En un sistema de comunicación por voz, el objetivo final es que el receptor disponga de la información de voz, de forma que se puede comenzar el diseño a partir del funcionamiento que debe tener el sistema de recepción. Imitando el modelo humano de producción de voz, el receptor o sintetizador de voz debe responder al siguiente esquema:



Si el sonido que se debe producir es sordo, la excitación del sistema será un ruido gaussiano como el aire turbulento procedente de los pulmones que pasa por el tracto vocal. Por ello se generará una trama, en este caso de 240 muestras, de ruido gaussiano. Si el sonido es sonoro, la excitación en el aparato fonador humano es un tren de impulsos periódicos procedentes de la apertura y cierre de las cuerdas vocales, por lo que se generará una trama formada igualmente por impulsos (en este caso deltas, empleando un modelo sencillo). El tracto vocal se modela como un filtro (denominado *filtro de síntesis* en el esquema), ya que consiste en aplicar una serie de resonancias a las señales acústicas que recibe; del mismo modo, en el sintetizador aplicaremos un filtro que igualmente se puede considerar como un conjunto de resonancias a diferentes frecuencias. La salida de este filtro será una trama de voz sintética (de nuevo se considerará de 240 muestras).

Como consideración adicional, las señales de excitación que se generen deben tener una amplitud/volumen equivalente al que se produciría por el aparato fonador humano, por lo que es necesario emplear un término de ganancia que las ajuste en amplitud.

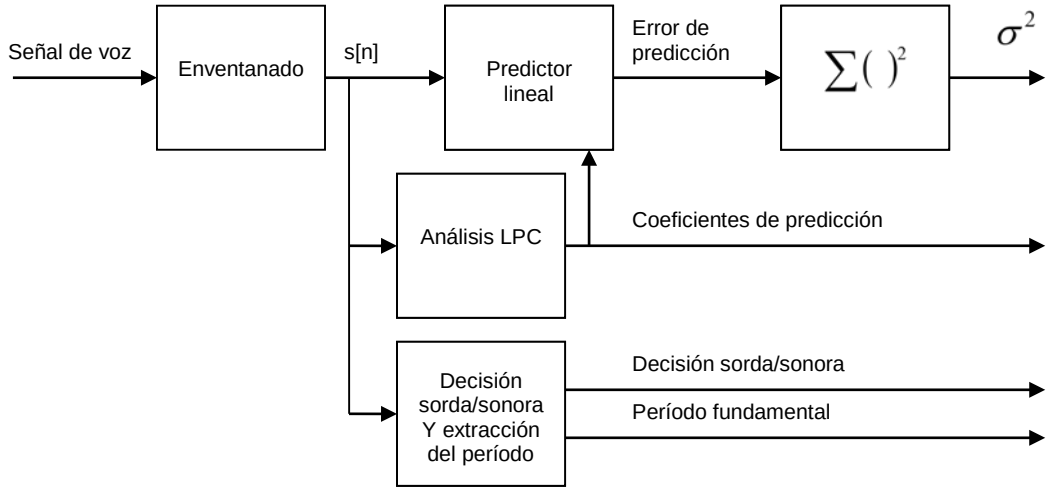
Para implementar correctamente estos sistemas de síntesis de voz se necesitarán una serie de parámetros que son los que deben ser enviados:

- Un valor o bit que indique si la trama analizada corresponde a un sonido sordo o sonoro, y así determinar qué tipo de señal debe generarse.
- El período fundamental (en el caso de ser un sonido sonoro), que identifica la distancia entre las deltas de la señal generada.
- Un factor que denote la energía o volumen de la señal.
- Los coeficientes de predicción.

De esta forma, el esquema del proceso de análisis que corresponde con la codificación de la voz en un transmisor debe estar diseñado para obtener y transmitir esos parámetros para cada trama de voz.

3.3.- Análisis de la voz: diseño del codificador o transmisor.

El esquema general de un codificador de voz, que obtenga y transmita los parámetros que necesita el receptor para sintetizar la voz, es el siguiente:



Los procesos que intervienen es este esquema son:

- Entramado / enventanado. La voz es un proceso no estacionario. No es posible analizar íntegramente un segmento completo de voz. La solución más habitual es dividir la voz en pequeñas tramas en las que las características permanezcan aproximadamente constantes en el tiempo. Diferentes estudios muestran que las características de la voz permanecen constantes durante un tiempo de aproximadamente 20-40 ms. Además, es importante para los algoritmos que se van a emplear eliminar la media de cada trama de forma que siempre se analicen procesos de media cero.
- Análisis LPC y predicción lineal. El análisis LPC (*Linear Predictive Coding*) es el procedimiento por el cual se seleccionan unos coeficientes de predicción que contienen información sobre la relación existente entre las muestras de una señal. El estándar americano LPC-10 para predicción y codificación de voz emplea 10 coeficientes de predicción además de un undécimo coeficiente de valor unidad que se emplea para mantener la notación y la simplicidad en los cálculos recursivos. En la transmisión digital de la voz deberá enviarse para cada trama el conjunto de coeficientes de predicción que la caracteriza. Además, esos mismos coeficientes son empleados como coeficientes (FIR) en un filtro predictor lineal para el análisis de la voz, que proporciona como salida una señal llamada *residuo*, formada por la parte impredecible de la voz.
- El residuo emplea para estimar la energía de la señal en esa trama y a partir de ella determinar la ganancia que debe aplicarse en el receptor.
- Decisión de si el sonido es sordo/sonor, y extracción en su caso del período fundamental. Existen muchos métodos para determinar el período fundamental de una trama sonora. Entre ellos se va a estudiar el conocido como SIFT (*Simple Inverse Filtering Tracking*), basado en la autocorrelación, y descrito más adelante.

Con estos procesos, se extraen los parámetros que caracterizan completamente el modelo de voz empleado y que necesita el receptor para sintetizar la voz. Por supuesto,

se trata de un modelo simplificado, pero suficiente para la comprensión de mensajes de voz en tiempo real en comunicaciones de muy baja velocidad.

3.4.- Análisis LPC y predicción lineal.

Cargue el registro de voz contenido en el fichero `frase1.wav` disponible en el servidor. Divídalo en tramas correspondientes a 30 ms cada una, sabiendo que se trata de un registro de voz muestreado a 8 KHz. Elimine la media de cada trama y obtenga los coeficientes de predicción de orden 10 por medio de la instrucción `lpc.m` de MatLab para cada una de las tramas.

Emplee la instrucción `filter.m` para obtener el residuo sin redundancia a corto plazo correspondiente. Implemente también el filtro inverso y compruebe que es capaz de recuperar perfectamente la señal original concatenando el filtro directo (predictor lineal) y el inverso.

El problema de los coeficientes obtenidos en el análisis LPC es que al ser codificados digitalmente resultan muy difíciles de cuantificar. Sus valores no están acotados en modo alguno, y para pasarlos a bits no es posible establecer unos límites razonables para los cuantificadores. Por ese motivo, es habitual calcular a partir de ellos otros coeficientes, denominador PARCOR, cuyo interés en el codificador reside en que se encuentran acotados en su mayoría entre +1 y -1, facilitando su cuantificación. Además, la conversión de coeficientes LPC a PARCOR es completamente reversible, por lo que se suelen emplear siempre como elemento intermedio para la transmisión, volviendo a calcular los LPC en el receptor. Implemente la conversión de los coeficientes LPC (de predicción) de cada trama a sus equivalente coeficientes PARCOR y viceversa[2], e incorpore dichas conversiones en las rutinas generadas.

3.5.- Decisión sorda/sonora y periodo fundamental.

Implemente un sistema que decida si la trama de voz analizada es sonora o sorda, y que detecte su período fundamental en caso de ser sonora. Para ello tenga en cuenta las siguientes consideraciones.

La mayor parte de los métodos de detección del período fundamental de una señal se basan en la autocorrelación, del mismo modo que implícitamente lo hace la predicción lineal. Aunque la función de autocorrelación de una sección sonora de voz normalmente muestra un pico prominente en el período, también están presentes otros picos debidos a la estructura de los formantes de la señal de voz, originados por las oscilaciones en la respuesta del tracto vocal. Por tanto, un problema es decidir cuál de los picos en la función de autocorrelación corresponde al período fundamental. Hay que acentuar la periodicidad suprimiendo otras características que pueden crear errores. Para eliminar parcialmente los efectos de la estructura formante superior, la trama de la señal de voz de entrada (cada trama) es filtrada con un fir (instrucción `fir1.m`, orden 25) paso bajo con una frecuencia de corte de 900 Hz. Esto mantiene el suficiente número de armónicos para una detección del período bastante afinada, pero elimina los formantes superiores.

Una vez filtrada la trama, emplee el algoritmo SIFT: el algoritmo propone emplear un predictor lineal de orden 4 para analizar la trama ya filtrada. Un filtro predictor lineal de orden 4 es suficiente para modelar el espectro de la señal en el rango de frecuencias de

[2] En el anexo dispone de los algoritmos que permiten la conversión de LPC a PARCOR y viceversa.

0 a 1 KHz aproximadamente. Se realiza entonces la autocorrelación de la señal error de predicción (residuo) resultante (si la señal analizada tiene un período *P*, la autocorrelación de dicha señal también lo tiene).

La señal de voz puede contener un determinado rango de frecuencias, de forma que no es posible que cualquier valor de período fundamental sea válido. Se selecciona por ello el mayor pico en el rango de períodos posibles, de 2.5 a 20 ms. Recuerde que los periodos se corresponden con la distancia al punto central de la autocorrelación.

Esta elección de período fundamental tiene sentido en el caso de señales sonoras, pero si se está analizando una trama sorda existirán innumerables picos debidos al ruido que constituye el sonido. Cuando se estudia una señal sonora, su periodicidad es tan marcada que la amplitud del pico correspondiente al período fundamental es equiparable a la del pico del origen. Sin embargo, en sonidos sordos el pico del origen es mucho mayor que cualquier pico debido al parecido de la voz con el ruido. Por ello, si el pico seleccionado tiene una amplitud menor de un 25% de la del pico en el origen se puede estimar que la trama analizada es sorda; en caso contrario se considerará sonora, y se almacenará o transmitirá el valor del periodo calculado (la distancia al punto central de la autocorrelación).

3.6.- Desarrollo completo del codificador.

Unifique todos los apartados anteriores de forma que disponga de un sistema que, a partir de un registro de voz, proporcione para cada trama un conjunto de parámetros que la caracterice, incluyendo:

- Los coeficientes PARCOR.
- Una decisión sorda/sonora.
- El período fundamental para el caso de tramas sonoras.
- Una estimación de la *energía* de la trama según el esquema del proceso de análisis de la voz.

3.7.- Desarrollo del sintetizador de voz.

Genere un sistema de generación de voz sintética para el receptor, como el indicado en la introducción de la práctica, que reciba los parámetros del apartado anterior y genere una señal de voz según el modelo de codificación planteado. Estudie con detenimiento cómo aplicar correctamente la energía de la señal a cada trama para que la naturalidad de la voz sintetizada permita entender el mensaje correctamente.

Compare objetiva y subjetivamente la señal de voz sintetizada con la original. Para la comparación objetiva compruebe las diferencias en la representación de las señales y calcule el error cuadrático medio cometido. Para la comparación subjetiva escuche ambas señales por medio de la instrucción `sound.m` y analice la comprensibilidad del mensaje, la capacidad de reconocer a un locutor determinado o la naturalidad de la voz sintetizada.