

Instituto tecnológico de Massachussetts
Departamento de ingeniería eléctrica e informática

6.345 Reconocimiento automático del habla
Primavera de 2003

Publicado: 21/02/03
Entregar: 05/03/03

Trabajo 3
Representación de la señal

El objetivo de este trabajo es familiarizarle con los principios del análisis de Fourier de corto tiempo (1ª parte) y el análisis cepstral (2ª parte) aplicado al habla. Sería aconsejable que las siguientes tareas, señaladas con **T**, se acabaran durante cada sesión de prácticas. Las respuestas a las preguntas (marcadas con **P**) se entregarán dentro del plazo estipulado.

Para comenzar la práctica, introduzca el siguiente comando en la línea de comandos de UNIX:

```
% start_lab3.cmd
```

La práctica estará fundamentalmente controlada por la ventana *Lab 3* (práctica 3). Esta ventana presenta un conjunto de paneles que contienen opciones de distribución (se conocen como opciones del *Programa* en la ventana *Lab 3*) y un grupo de botones para los enunciados. Para mostrar un enunciado empleando una distribución en concreto, seleccione primero la opción de distribución y haga clic en el botón del enunciado que desee mostrar. En la práctica se utilizará también la ventana de control del *Analizador del espectro*. Desde esta ventana, usted podrá modificar muchos de los rasgos del análisis espectral tales como la longitud de la ventana, el tipo de ventana, etc.

Parte 1: Análisis de Fourier de corto tiempo

Ejercicio de prácticas en el laboratorio

T1: En esta parte de la práctica examinará las características espectrales de dos ventanas distintas para llevar a cabo el análisis de Fourier de corto tiempo. Un modo sencillo de hacer esto es aplicar la ventana a una forma de onda de amplitud constante $x[n] = K$, modificando luego las características de la ventana para examinar los resultados.

Para hacer esto, seleccione la distribución *Forma de onda y espectros* y el enunciado *constante*, que contiene una amplitud de onda de amplitud constante, desde la ventana *Lab3*.

Para una longitud de ventana determinada, compare tanto el ancho del lóbulo principal, como la amplitud del pico de los lóbulos laterales (definida como la diferencia de amplitud entre el lóbulo principal y el lóbulo lateral mayor), para el tipo de ventana *rectangular* frente a *hamming*. Puede especificar la ventana, utilizando el control del tipo de ventana en el *Analizador del espectro*. Es posible que desee trazar varias líneas en la ventana de visualización del espectro para facilitar las comparaciones. Si enfoca la región de baja frecuencia de los espectros, le resultará más fácil medir el peso y ancho de los lóbulos laterales y principal.

Para un tipo de ventana determinada, examine también el efecto de las longitudes de distintas ventanas sobre al ancho del lóbulo principal, y la amplitud del pico de los lóbulos laterales. Utilice longitudes de ventana de 100, 300 y 500 puntos. Las puede especificar en el *Analizador del espectro*, si utiliza el control *tamaño (seg.)*. Recuerde que la señal es muestreada a 16kHz, y que será necesario que calcule lo que la longitud de la ventana significa (en segundos), para las longitudes del punto superior. No utilice el teclado numérico para introducir el tamaño de la ventana. Si teclea sobre el teclado numérico, y *Num Lock* (bloqueo numérico) está activado, el *Analizador del espectro* morirá. Después de introducir en nuevo tamaño en segundos, pulse *Intro* o *Retorno de carro* para comprobar los cambios en la ventana del espectro.

T2: En esta parte de la práctica investigará los efectos de la aplicación de distintas ventanas a una forma de onda totalmente periódica. Por los resultados de la sección anterior, debería ser capaz de predecir (o explicar) las distorsiones que esas ventanas pueden crear en la magnitud del espectro.

Mientras esté en la misma distribución, selecciones el enunciado sintético-ah, que es una vocal generada sintéticamente con una función de excitación totalmente periódica. Compare la magnitud de los espectros para los mismos casos de antes (ej., ventanas *Rectangular* y de *Hamming* de longitud de 100, 300 y 500 puntos cada una).

P1: Determine el periodo fundamental de la vocal desde la forma de onda. ¿Cómo y bajo qué condiciones puede determinarse este valor a partir de la magnitud de los espectros?

P2: Para los espectros computados con ventanas de *Hamming*, ¿cuál es el resultado general del aumento del tamaño de la ventana?

P3: ¿Por qué se computan los espectros con ventanas rectangulares de aspecto irregular?

T3: Repetir la tarea 2 (T2) con un enunciado emitido de forma natural.

Seleccione la *Forma de onda*, el *Espectrograma* y la distribución de los *espectros*, que es igual que la distribución anterior, excepto que en este caso se ha añadido una línea de espectrograma. El espectrograma muestra el tiempo en el eje horizontal, y la frecuencia en el eje vertical. El grado de oscuridad del visualizado representa la energía. La variación de amplitud con frecuencia para un punto concreto en el tiempo, corresponde al STFT de la señal de voz centrada en este punto. (Las STFT para este espectrograma se calcularon con una ventana de *Hamming* de 67 ms).

Seleccione *Voz femenina natural*, que es un enunciado emitido por un hablante femenino, o *Voz masculina natural*, que es un enunciado emitido por un hablante masculino.

Sítue el cursor en las distintas porciones de fricativas y vocales del enunciado, y observe los diferentes espectros, manteniendo pulsado el botón derecho del ratón mientras el cursor se encuentra dentro de la ventana de la forma de onda, y seleccionando la opción *xspectrum*.

P4: ¿Por qué es mayor el nivel de las porciones de alta frecuencia de los espectros computados con ventanas rectangulares, que el de los espectros correspondientes a ventanas de *Hamming*?

Trabajo para casa

P5: Una señal de voz se muestrea a una velocidad de 16.000 muestras por segundo. Una ventana de 10 ms se aplica a la señal de voz para un análisis espectral de corto tiempo, y la ventana avanza a 40 muestras cada vez. Suponga que para computar la transformada de Fourier discreta (DFT), sólo se encuentran disponibles los algoritmos de raíz 2 FFT.

- (a) ¿Cuántas muestras existen en el segmento de voz seleccionado?
- (b) ¿Cuál es la velocidad de tramo del análisis espectral de corto tiempo, ej., cuál es la duración (en milisegundos) entre cada computación de la DFT?
- (c) ¿Cuál es el tamaño *mínimo* de la DFT para que no se de solapamiento temporal?
- (d) ¿Cuál es el espaciado en Hz entre las muestras de la DFT para el tamaño de la DFT como se determinó en la parte (c)?

Ahora es necesario pasar la señal digital a través de un sistema de reconocimiento de voz telefónico que supone una velocidad de muestreo de 8 KHz.

- (e) ¿Qué preprocesado requiere la señal para satisfacer esta nueva velocidad de muestreo?
- (f) Asumiendo el mismo tamaño de ventana de 10 ms, ¿cuántas muestras existen en el segmento de voz seleccionado?
- (g) Para mantener la misma velocidad de tramo que en la parte (b), ¿en cuántas muestras debería avanzar la ventana DFT?
- (h) ¿Cómo se verían afectadas las respuestas a las partes (c) y (d)?

Parte II: Análisis cepstral del discurso

Ejercicio de prácticas en el laboratorio

En esta parte de la práctica, primero examinará algunas de las propiedades del cepstrum complejo, y mostrará luego cómo éste se puede manipular para obtener o bien la información del tracto vocal, o la información de excitación. Estudiará primero estas cuestiones utilizando enunciados sintéticos cuyas propiedades son totalmente conocidas. Más tarde, podrá examinar el discurso natural, a partir de los enunciados proporcionados.

El cepstrum (complejo)

Estudiará primero algunas de las propiedades del cepstrum complejo utilizando dos enunciados sintéticos. El primero consiste en un impulso en el origen, y un impulso escalado 20 muestras más adelante. El segundo es una /a/ sintética, el mismo enunciado utilizado en la 1ª parte.

T4: En la ventana *Lab3*, seleccione la distribución *WaveForm Only* (sólo forma de onda) y elija luego el enunciado *impulse-pair* (impulso-par). Compute el cepstrum complejo (fase mínima), manteniendo pulsado el botón derecho del ratón en la ventana de la forma de onda, y seleccionando del menú que aparece, la opción *complex cepstrum* (cepstrum complejo). En un cuadro de diálogo le aparecerá una línea de comando que le inducirá al orden de la DFT N (ej., una DFT de $2n$ puntos). Introduzca un número razonable entre 5 y 12 y pulse retorno de carro. Aparecerá una nueva ventana con el cepstrum complejo.

Asegúrese de que comprende cómo se computa el cepstrum complejo. Mida y verifique que los impulsos *válidos* del cepstrum complejo poseen la amplitud correcta. Es posible que desee reescalar el eje vertical para ver mejor el cepstrum. Haga esto manteniendo pulsado el botón derecho del ratón en la ventana del cepstrum y seleccionando *fixed vertical zoom* (enfoco fijo vertical) del menú. Puede reestablecer la ventana seleccionando *vertical auto-zoom* (auto enfoque vertical).

Varíe el tamaño de la DFT y observe el efecto que posee sobre la extensión del solapamiento de espectros.

P6: ¿Puede explicar por qué el valor cero del cepstrum complejo es un valor no cero? Realice algunas mediciones para verificar su hipótesis.

T5: Seleccione la distribución *Waveform only* (sólo forma de onda) y el enunciado sintético -ah. Señale una región de la forma de onda que encierre varios periodos tonales.

Compute el cepstrum (cepstrum complejo de fase cero) manteniendo pulsado el botón derecho del ratón en la ventana de la forma de onda y seleccionando *cepstrum* del menú. Utilice una DFT de décimo orden. Es posible que quiera intentar también esto sobre alguna porción de discurso real.

P7: Mida la frecuencia fundamental de sonoridad (F_0) de la vocal sintética del cepstrum. Verifique su medición desde la forma de onda temporal.

Recuperación de la información del tracto vocal

T6: Utilizando la distribución *Waveform and Spectra* (forma de onda y espectros), seleccione el enunciado sintético ah-sintético.

En el *Spectrum Analyzer* (anализador del espectro), ajuste el *tipo de análisis* a DFT, el *tipo de ventana* a *Hamming* y el *tamaño en seg.* para incluir varios periodos tonales (ej., 0'025 segs.). En la ventana del espectro, guarde el espectro como espectro de referencia.

Compute el cepstrum manteniendo pulsado el botón derecho del ratón en la ventana de la forma de onda y seleccionando *cepstrum* del menú. Utilice una DFT de 10º orden. Observe el cepstrum y familiarícese con los lugares que los picos ocupan en él.

En el *Spectrum Analyzer* (anализador del espectro), modifique *el tipo de análisis* a CEPST. Esta opción *recomputará* los espectros después de realizar operaciones de filtrado cepstral sobre la señal. Las operaciones de filtrado están controladas por los siguientes parámetros:

- *Liftering* (filtrado): selecciona filtrado paso alto, paso bajo, o ninguno.
- *Cep. cut (sec)* (corte cepstral (seg.)): facilita la frecuencia de corte nominal.
- *Cep. trans.*(transición cepstral): proporciona la duración de la región de transición de frecuencia entre la potencias cero y absoluta.

Practique con los distintos valores de los parámetros de arriba. Con la operación adecuada de filtrado de la señal, debería ser capaz de recuperar sólo la magnitud logarítmica de la respuesta de frecuencia del tracto vocal sin ningún tipo de información de fuente / de excitación.

P8: En este ejemplo, ¿cuál es la operación de filtrado correcta (paso alto, paso bajo o ninguna) y el corte correcto para recuperar sólo la información del tracto vocal?

Recuperación de la información de excitación

T7: Como en la T6, practique con los distintos valores de los parámetros de filtrado, pero esta vez deberá recuperar sólo la información de fuente / excitación.

P9: En este ejemplo, ¿cuál es la operación de filtrado correcta, y el corte adecuado, para recuperar sólo la información de fuente / excitación?

T8: Intente este procedimiento de análisis sobre el lenguaje natural. Observe la diferencia en las características del cepstrum complejo para el discurso sonoro y sordo.

Es posible que desee comparar el espectro de magnitud logarítmica cepstralmente suavizado, con el espectro original, para ver cuál es el preferido para extraer las frecuencias del formante y/o la frecuencia fundamental de sonoridad.

P10: Según sus observaciones sobre el lenguaje *natural*, ¿cuál es el filtrado adecuado para recuperar la información del tracto vocal? ¿y para la información de excitación?

Coefficientes cepstrales

Los coeficientes cepstrales y sus derivadas temporales se utilizan en la actualidad en muchos sistemas de reconocimiento de voz. En esta parte de la práctica, usted examinará algunos de esas características, y seguramente obtendrá conocimientos que le permitirán comprender por qué resultan útiles.

T9: Seleccione las distribuciones *Waveform* (forma de onda), *Spectrogram* (espectrograma) y *Cepstral coefficients* (coeficientes cepstrales) y elija el enunciado *natural-female* (femenino natural).

En la ventana de los coeficientes cepstrales, se trazan los coeficientes cepstrales $c[0]$ y $c[1]$ (arriba) y sus derivadas temporales correspondientes (abajo). Las derivadas temporales en este caso se computan tomando la diferencia de los cuatro tramos de

coeficientes cepstrales (ej., 20 ms) antes y después del tramo actual. Estudie el comportamiento de estos cuatro parámetros para los distintos sonidos del habla.

P11: Basándose en lo que ve y sabe acerca del comportamiento de estos parámetros, sugiera cómo estos parámetros pueden correlacionarse con las propiedades acústicas de los sonidos del habla. Por ejemplo, ¿qué puede decir sobre el valor de $c[1]$ para vocales y fricativas?

Coeficientes cepstrales de frecuencia Mel

Ahora utilizará MATLAB para volver a sintetizar algunos enunciados, utilizando sólo los coeficientes cepstrales de frecuencia Mel (MFCC) derivados de estos enunciados. Esto debería darle alguna idea sobre la información contenida en los MFCC, y si esta información es o no adecuada para propósitos de reconocimiento de voz.

Para arrancar MATLAB, escriba lo siguiente en la línea de comandos de Linux:

```
% start_lab3_matlab.cmd
```

Luego, escriba lo siguiente en la línea de comandos de MATLAB:

```
>> init_lab3
```

Esto cargará tres enunciados, cada uno de ellos muestreados a 8 kHz en las variables *unusual*, *pathological* y *cupcakes*.

T10: Escuche los tres enunciados utilizando la función reproducir. Para escuchar el tipo *unusual*, escriba:

```
>> play (unusual, 8000);
```

Para volver a sintetizar estos enunciados utilizando los MFCC, utilizará la función *resyn_from-mfccs*. Esta función opera invirtiendo primero los MFCC para obtener un espectro de magnitud en cada tramo, y utilizando un algoritmo de iteración para calcular la fase de cada componente de frecuencia en cada tramo. La estimación de la fase es necesaria, dado que la fase se ignora cuando se calculan los MFCC.

La primera iteración de la parte iterativa del algoritmo comienza con las funciones de la fase inicial de cero por cada tramo. Estas funciones de fase se combinan con los espectros de magnitud obtenidos de los MFCC para producir STFT completos. Dado que los tramos en los que el STFT es computado se solapan, no todos los grupos de STFT corresponden a señales válidas. Para explicar esto, se aplica un algoritmo para obtener la señal válida cuyos STFT se corresponden más estrechamente, en un sentido menos restrictivo, con los STFT creados. La señal resultante se considera el resultado de la primera iteración. La segunda iteración comienza con la combinación de la función de fase de esta señal y de los espectros de magnitud obtenidos de los MFCC. En cada iteración, la señal resultante se parece más a la señal original en que los espectros de magnitud están más cercanos en un sentido menos restrictivo. No obstante, el algoritmo no garantiza la convergencia de la señal original. Para más información, consulte la siguiente referencia:

D. W. Griffin y J. S. Lim; "Signal Estimation from Modified Short-Time Fourier Transform", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 32, n° 2, abril, 1984.

Para volver a sintetizar el tipo *unusual* con este algoritmo, utilizando los primeros 14 MFCC, escriba lo siguiente en la línea de comandos de MATLAB:

```
>> resyn = resyn_from_mfccs (unusual, 14);
```

Esto almacenará la forma de onda en *resyn*, que puede entonces reproducirse utilizando la función *reproducir*.

Vuelva a sintetizar los enunciados utilizando varios números de coeficientes y escuche los resultados. Concretamente, vuelva a sintetizar las formas de onda utilizando los coeficientes MFCC 2, 14, 25 y 128.

P12: Para cada enunciado, compare el resultado resintetizado utilizando distintos números de coeficientes MFCC. ¿Para cuántos MFCC es bien capturada la forma espectral total para hacer que el enunciado sea inteligible? (Es posible que desee pedirle a un compañero que no conozca cuáles son los resultados, que le ayude con esto). ¿Cuántos coeficientes MFCC son necesarios para capturar la información tonal?

P13: Muchos reconocedores de voz utilizan los primeros 14 coeficientes MFCC. Utilizando su respuesta a la pregunta anterior, explique por qué esto podría ser una buena elección.

P14: Escuche los enunciados resintetizados para *unusual* y *cupcake* utilizando sólo los dos primeros coeficientes MFCC. ¿Cuál es más inteligible? ¿Por qué es esto así? (Pista: Tenga en cuenta qué información es capturada en los dos primeros coeficientes cepstrales y relacione esto a las características espectrales de los fonos en el enunciado *unusual*. Preste atención también al número de sílabas de cada palabra, y cómo afecta éste al espacio de búsqueda léxica.

Ejercicio para casa

P15: El cepstrum complejo $\hat{x}[n]$, se relaciona con la señal $x[n]$ por :

$$\hat{X}(z) = \log X(z)$$

Como hemos visto en clase, el cepstrum complejo de una señal causal de fase mínima, puede generarse con el uso de las DFT. Sin embargo, existe también una relación

recursiva para determinar $\hat{x}[n]$, directamente a partir de $x[n]$.

- (a) Diferenciando las transformadas Z, muestre que $\hat{x}[n]$ y $x[n]$ se relacionan por:

$$n\hat{x}[n] * x[n] = nx[n]$$

(b) Suponiendo que $x[n]$ es una fase mínima, utilice esta expresión para derivar la siguiente fórmula recursiva para la generación de $\hat{x}[n]$:

$$\hat{x}[n] = \begin{cases} 0 & n < 0 \\ \log x[0] & n = 0 \\ \frac{x[n]}{x[0]} - \frac{1}{nx[0]} \sum_{k=0}^{n-1} k\hat{x}[k]x[n-k] & n > 0 \end{cases}$$

(Utilice el teorema del valor inicial para $n=0$).

P16: Este problema trata algunas cuestiones del procesamiento de la señal relativos a la computación de los coeficientes cepstrales de frecuencia Mel (MFCC) a partir de los coeficientes espectrales de frecuencia Mel (MFSC). Asumiremos que la longitud del segmento de $(s[n])$ es N , el tamaño de la transformada de Fourier discreta es $(S[k])$ es M , y el número de filtros de frecuencia Mel es L . Supondremos también que el tamaño de la DFT es lo bastante grande como para que el efecto del solapamiento de espectros sea insignificante.

(a) Muestre que el cepstrum $c[n]$ puede computarse como una transformada de coseno discreta, ej.,

$$c[n] = \sum_{k=0}^{M-1} \log |S[k]| \cos \frac{2\pi}{M} kn \quad 0 \leq n \leq M-1$$

(Un factor de escala de $1/M$ es ignorado en la expresión mostrada arriba).

(b) El siguiente procedimiento se utiliza normalmente para computar los coeficientes MFCC:

- Los coeficientes de Fourier son $S[k]$ son cuadrados.
- El espectro resultante de magnitud cuadrada es pasado a través de los bancos de filtro triangulares de frecuencia Mel mostrados en las fotocopias de clase.
- Las salidas de energía logarítmica (en decibelios) de los filtros $X_k, k = 1, 2, \dots, L$, forman colectivamente el vector MFSC L -dimensional. Observe que no existe coeficiente MFSC para $k=0$. Es posible que esta información le sea útil para la parte (c).
- Los coeficientes MFCC se computan por tanto mediante una transformada de coseno discreta.

Para llevar a cabo el último paso, debemos tratar al coeficiente MFSC como una transformada de Fourier discreta de una señal real. ¿Qué propiedades de simetría deben imponerse? ¿Cuál es el tamaño mínimo de la transformada de Fourier discreta?

(c) Muestre que los coeficientes MFCC $Y_i, i = 1, 2, \dots, L$, están determinados por:

$$Y_i = \sum_{k=1}^L X_k \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{L}\right].$$

PISTA: La transformada DFT inversa puede computarse a partir de cualquier grupo de puntos igualmente espaciados sobre el círculo unidad.