Departamento de Teoría de la señal y comunicaciones y Sistemas telemáticos y Computación
Universidad Rey Juan Carlos

Grado en Ingeniería en Sistemas de Telecomunicación

Procesamiento de Señales

Curso 2016/2017

PRÁCTICA: CODIFICACIÓN DE VOZ

1. Introducción

En esta práctica analizaremos el funcionamiento de un codificador-decodificador de voz básico basado en predicción lineal, un Vocodificador LPC. En el codificador se calculan los parámetros del modelo y en el decodificador se reconstruye la señal de voz a partir de los parámetros recibidos.

Con este tipo de codificación se consiguen tasas binarias muy bajas. La voz sintética reconstruida es inteligible pero poco natural.

Introducción Teórica

Existe una fuerte correlación entre muestras consecutivas de la señal de voz, una muestra de voz puede modelarse como una combinación lineal de las p muestras anteriores:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \tag{1}$$

El error que se comete en la estimación (error de predicción) será:

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \tag{2}$$

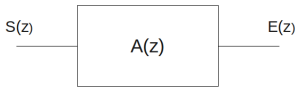
Tomando transformadas Z:

$$E(z) = TZ\{s[n] - \sum_{k=1}^{p} a_k s[n-k]\} = S(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) \tag{3}$$

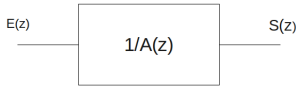
Si denominamos $A(z)$ a la expresión:

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

$A(z)$ es la función de transferencia del sistema:



La función de transferencia de un sistema inverso al anterior será:



Excitando con el error de predicción un sistema cuya función de transferencia sea $1/A(z)$, a la salida se obtiene la señal deseada de voz, $s[n]$.

$H(z)$ **será el filtro de predicción lineal mediante el cual se puede modelar el tracto vocal**. El número de polos del modelo será igual al orden de predicción lineal, p.

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5}$$

Debemos determinar el conjunto de coeficientes de predicción a_k que minimizan la energía del error de predicción (criterio de mínimos cuadrados) en cada trama de análisis:

$$E = \sum_{n} e^2[n] \tag{6}$$

Obtendremos los a_k que cumplan:

$$\frac{\partial E}{\partial a_k} = 0 \quad para \quad k = 1, ..., p \tag{7}$$

$$\frac{\partial E}{\partial a_1} = 0 = \sum_{n=-\infty}^{\infty} s[n]s[n-1] - \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s[n-k]s[n-1]$$

$$\frac{\partial E}{\partial a_2} = 0 = \sum_{n=-\infty}^{\infty} s[n]s[n-2] - \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s[n-k]s[n-2]$$

$$\vdots$$

$$\frac{\partial E}{\partial a_p} = 0 = \sum_{n=-\infty}^{\infty} s[n]s[n-p] - \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s[n-k]s[n-p]$$

Dado que la autocorrelación de una señal se define como:

$$R(k) = \sum_{n=-\infty}^{\infty} s[n]s[n-k] \tag{8}$$

El sistema de ecuaciones queda:

$$\begin{aligned} 0 &= R[1] - \sum_{k=1}^{p} a_k R[k-1] \\ 0 &= R[2] - \sum_{k=1}^{p} a_k R[k-2] \\ &\vdots \\ 0 &= R[p] - \sum_{k=1}^{p} a_k R[k-p] \end{aligned}$$

Expresado de forma matricial:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & & & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}$$

De forma compacta:

$$\mathbf{Ra = r} \tag{9}$$

Existen diferentes métodos resolver este sistema y hallar los a_k. En esta práctica resolveremos el sistema mediante inversión de la matriz, ya que consideramos un valor de p pequeño.

2. Actividades

- Cargue la señal *confront.mat* y reprodúzcala utilizando el comando *soundsc(y,fs)*, donde el primer parámetro de entrada representa la señal que se va a reproducir y el segundo parámetro determina la frecuencia de muestreo de la señal.
- Represente la señal de voz en el dominio temporal.

Cálculo de la respuesta del tracto vocal

Obtenga una trama (s) de la señal (muestras de la 4181 a 4400) y realice los siguientes pasos:

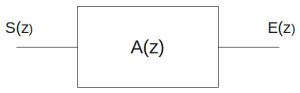
1. Calcule la autocorrelación de la trama con el comando *xcorr*.

2. Genere el vetor **r** tomando las p primeras muestras de la autocorrelación ($R(1)...R(p)$). Tome el valor $p = 12$, suficiente para modelar correctamente la respuesta del tracto vocal.

3. Genere la matriz **R** utilizando en comando *toeplitz*.

4. Obtenga los coeficientes de predicción lineal a_k resolviendo el sistema $\mathbf{Ra} = \mathbf{r}$.

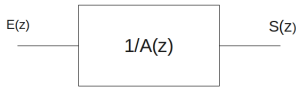
5. Obtenga el error de predicción:



   e=filter([1 -ak'], 1, s)

   donde, *ak* son los coeficientes de predicción lineal obtenidos y *s* es la trama de señal de voz.

6. Para comprobar que ha obtenido los coeficientes *ak* correctamente utilize el error de predicción para recuperar la señal original:



   s=filter(1,[1 -ak'],e)

7. Represente la trama de señal de voz original y la señal recuperada. ¿Cómo son?

8. Implemente una función para obtener los coeficientes de predicción lineal *ak* de cualquier trama de forma genérica. La función *[ak] = tractoVocal(s,p)* recibirá como parámetros una trama de la señal de voz (s) y el orden de predicción lineal (p); devolverá como salida los coeficientes de predicción lineal (ak). Esta función debe contener las lineas de código generadas en los puntos 1–4.

Voz sintética

A continuación vamos a obtener voz sintética utilizando señales sencillas que realicen el papel del error de predicción como entrada o excitación del sistema:

- Un tren de deltas de periodo igual al de la voz original para las tramas sonoros.
- Ruido blanco para las tramas sordos

1. obtenga dos tramas de voz de la señal original, trama 1: muestras 14200-14419 y trama2: muestras 15500-15719.

2. Obtenga la energía y la tasa de cruces por cero de ambas tramas con la función *[E,TCC]=energy_TCC(trama)*.

3. Obtenga la autocorrelación de las tramas con la función *xcorr(trama)* y representelas.

4. Obtenga el espectro de las tramas con la función *myspectra(s,fs)*.

5. Determine si se trata de tramos sonoros o sordos. En el caso de ser sonoros, estime el *pitch* o frecuencia fundamental.

Ejecute el script *vc_lpc*, este script hace uso de la función *tractoVocal* que ha implementado, también hace uso de la función *sonoro_sordo* que determina si cada trama de señal es sonora o sorda y construye la entrada apropiada al sistema en cada caso.

6. Enventanado: el tamaño de las ventanas es de 20 msg, pruebe a utilizar ventanas de 200 msg. ¿Qué ocurre? ¿por qué?.