Laboratorio de Procesamiento Digital de Voz

Practica 6

Cuantización Vectorial

Objetivos:

Manejar los conceptos del Cuantización Vectorial y los tipos de Distorsión utilizados en el Procesamiento Digital de Señales.

1. Introducción.

La cuantización vectorial (VQ), resulta una generalización de la cuantización escalar, pero ahora aplicada a todo un vector. El cambio de una a varias dimensiones, trae consigo un gran número de ideas, conceptos, técnicas y aplicaciones nuevas. Mientras la cuantización escalar se utiliza, principalmente en la conversión analógico/digital, la cuantización vectorial se enfrenta a las sofisticadas técnicas del procesamiento digital de señales. En la mayoría de los casos, las características más relevantes de las señales de entrada tiene representación digital, por eso, la cuantización vectorial se utiliza generalmente en la compresión de datos. Sin embargo, existen ciertos paralelismos entre ambas cuantizaciones, lo que permite la utilización de varios métodos, en la cuantización vectorial, como una generalización [Gersho, 1997].

Un vector se puede utilizar para describir prácticamente cualquier tipo de patrón, como puede ser un segmento de una señal de voz o de una imagen, simplemente al formar un vector con las muestras de la señal de voz o de la imagen. La cuantización vectorial puede aplicarse al reconocimiento de patrones, ya que un patrón de entrada es comparado y aproximado a alguno de los patrones de referencia almacenados. El reconocimiento permite encontrar el patrón de referencia que más se acopla al patrón de entrada. Por lo tanto, la cuantización vectorial es más que una generalización de la cuantización escalar. En fechas recientes, se ha convertido en la principal herramienta del reconocimiento de voz, además de que se sigue utilizándose en la compresión de señales de voz e imágenes [Gersho, 1997].

2. Distancias y medidas de distorsión

Un componente clave en la mayoría de los algoritmos de comparación de patrones, es formular una medida de distorsión entre dos vectores característicos. Esta medida de distorsión, puede ser manejada con rigor matemático si los patrones son visualizados en un espacio vectorial.

Suponer que se tienen dos vectores característicos, $\boldsymbol{x}$ e $\boldsymbol{y}$, definidos en un espacio vectorial χ. Se define una *métrica* o *función de distancia, d,* en el espacio vectorial χ, como una función de valor real, sobre el producto cartesiano $\chi \times \chi$, que cumpla las siguientes condiciones:

1. $0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < \infty$, para $\boldsymbol{x}, \boldsymbol{y} \in \chi$ y $d(\boldsymbol{x},\boldsymbol{y}) = 0$ sí y solo sí $\boldsymbol{x} = \boldsymbol{y}$;
2. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ para $\boldsymbol{x}, \boldsymbol{y} \in \chi$;
3. $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$ para $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \chi$.

Además, una función de distancia se denomina invariante sí

4. $d(\boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} + \boldsymbol{z}) = d(\boldsymbol{x},\boldsymbol{y})$

Las primeras tres propiedades comúnmente son conocidas como positividad (no negatividad), simetría y desigualdad del triángulo, respectivamente. Una métrica que contenga estas propiedades, permite un alto grado de manejo matemático. Sí una medida de distancia*, d,* satisface solo la propiedad de positividad, se le denomina medida de distorsión, particularmente cuando los vectores son representaciones del espectro de la señal.

Para el procesamiento de voz es importante considerar que la definición (o elección), de la medida de distancia, es significativamente subjetiva. Una medida matemática de la distancia, para ser utilizada en el procesamiento de voz, debe tener una alta correlación entre su valor numérico y su distancia subjetiva aproximada, para evaluar una señal real de voz. Para el reconocimiento de voz, la consistencia psicofísica (los diferentes matices que se le pueden imprimir a una misma palabra o frase), que se desea medir con la distancia, obliga a que se encuentre una medida matemática ajustada por necesidad, a las características lingüísticas conocidas. Estos requisitos tan subjetivos no pueden ser satisfechos con medidas de distancia que proporcionen manejo matemático. Un ejemplo es la tradicional medida del Error Cuadrático, $d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^2$.

Dado que existe una enorme dificultad al querer cumplir simultáneamente ambos objetivos (subjetividad y manejo matemático), algún compromiso es inevitable. Por consiguiente, y dado que se necesita manipular matemáticamente las propiedades de esa medida de distancia, se necesita probar que estas propiedades subjetivas son lo suficientemente buenas como para lograr el reconocimiento de voz. Por otra parte se hablará de "medidas de distorsión" en vez de "métricas" debido a que se relajan las condiciones de simetría y desigualdad del triángulo. No se debe utilizar el término *distancia* en sentido estricto, acorde a la definición de arriba; por otro lado, se debe mantener la costumbre de la literatura de voz, donde el término *distancia* es análogo, a las medidas de distorsión [Rabiner, 1993].

Existen varios tipos de medidas de distorsión, cada una con sus características especiales. Entre ellas tenemos: *Distancia Euclidiana Cuadrática, Distorsión del Error Cuadrático Medio, Distorsión del Error Cuadrático Ponderado, Distancia de Itakura*, etc..

- ***Distancia Euclidiana Cuadrática***. La medida más conveniente y ampliamente usada para calcular distancias, es el Error Cuadrático o Distancia Euclidiana Cuadrática, entre dos vectores, definida como:

$$d(X_{1,}, X_2) = \left\| X_1 - X_2 \right\|^2 = \sum_{j=1}^{N} \left(X_{1j} - X_{2j} \right)^2 \qquad (1)$$

- ***Distorsión del Error Cuadrático Medio***. La distorsión del Error Cuadrático Medio (MSE) es otra de las medidas mas utilizadas y se define como:

$$\mathrm{d}(\mathrm{X}_1, \mathrm{X}_2) = \frac{1}{\mathrm{N}} (\mathrm{X}_1 - \mathrm{X}_2)^{\mathrm{T}} (\mathrm{X}_1 - \mathrm{X}_2) = \frac{1}{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{X}_{1\mathrm{j}} - \mathrm{X}_{2\mathrm{j}})^2 \qquad (2)$$

  en la cual la distorsión está definida por cada dimensión. La popularidad del MSE se basa en su simplicidad y seguimiento matemático.

- ***Distorsión del Error Cuadrático Medio Ponderado***. Otra medida de distorsión es el Error Cuadrático Medio Ponderado. En el MSE la medida asume que las distorsiones contribuyen cuantizando los diferentes parámetros $\{X_{1j}\}$ de igual forma. Y de manera general, se pueden introducir pesos diferentes con el fin de aportar ciertas contribuciones a la distorsión, dependiendo del parámetro. El MSE ponderado general se define como:

$$d(X_1, X_2) = (X_1 - X_2)^T W (X_1 - X_2) \tag{3}$$

  Donde *W* es una matriz de ponderación definida, simétrica y positiva, y los vectores X_1 y X_2 son tratados como vectores columna.

  Cada una de las medidas de distorsión mencionadas anteriormente, resultan simétricas en sus argumentos X_1 y X_2 y pueden ser aplicadas a las características derivadas del análisis de producción lineal de la voz; el uso de algunas presenta ciertas desventajas, tal es el caso de la distancia euclidiana que aunque resulte fácil de calcular, no todas sus características tienen el mismo significado perceptible.

  Por lo anterior, en ciertos casos resulta conveniente y efectivo escoger una matriz de ponderación $W(X_1)$ que dependa explícitamente del vector X_1, para así obtener una medida de distorsión perceptiblemente motivada. En este caso, la distorsión:

$$d(X_1, X_2) = (X_1 - X_2)^T W(X_1)(X_1 - X_2) \tag{4}$$

  es asimétrica.

- ***Distancia de Itakura***. En muchos casos del procesamiento de voz, es necesario tener otra medida de la distancia que existe entre dos vectores LPC. La distancia Euclidiana no es apropiada para medir los parámetros de dos LPC's individuales, en vectores que estén relacionados. Esto es debido a que los vectores LPC dependen del peso de la matriz de autocorrelación correspondiente a cada LPC.

  La medida de distancia más comúnmente utilizada para este propósito es la propuesta por *Itakura*. Esta distancia se deriva utilizando una interpretación intuitiva del rango de predicción en el error de la energía. Fue obtenida originalmente, utilizando la máxima probabilidad existente entre argumentos similares. La distancia de *Itakura* es, probablemente, la medida de distorsión más empleada para encontrar la similitud entre dos vectores LPC [Deller, 1987].

  Esta distancia se define de la siguiente forma: sean $R_y x$ y $R_y y$ las matrices de autocorrelación multiplicadas por las señales de voz de entrada y de comparación, respectivamente. Sean así mismo, $xR_y x^T$ la energía de salida del filtro inverso, tomado como referencia con la entrada, y $yR_y y^T$ la energía mínima posible de salida, del filtro LPC, con respecto a la entrada de la voz. Entonces tenemos que la distancia de *Itakura* se obtiene mediante la ecuación 5:

$$d(x, y) = \log\left(\frac{xR_y x^T}{yR_y y^T}\right) \tag{5}$$

donde

$$yR_y y^T = \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix} \begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p) \\ r(1) & r(0) & r(1) & \cdots & r(p-1) \\ r(2) & r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \vdots & & \vdots \\ r(p) & r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} -1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad (7)$$

o lo que es lo mismo:

$$yR_y y^T = \sum_{i=0}^{P} a_i \sum_{n=0}^{P} r(|i-n|) a_n \quad (8)$$

donde $y = \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}$

De esta forma, si ***y*** es el vector aumentado de coeficientes LPC's de "referencia" o de "plantilla" $\begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}$, y sea ***x*** el vector aumentado de coeficientes LPC's "desconocido" u "observado" $\begin{bmatrix} -1 & a'_1 & a'_2 & \cdots & a'_p \end{bmatrix}$ entonces:

xR_yx^T = Es la energía del filtro inverso formado con la señal de entrada de voz
yR_yy^T = La energía de salida mínima posible para el filtro de predicción lineal con la entrada de voz

Por tanto (5) también se puede escribir como:

$$d(x,y) = \log\left(\frac{E_x}{E_y}\right)$$

3. Cuantización Vectorial (VQ).

Partimos de un conjunto de vectores que pertenecen a un espacio *K*-dimensional, asumiendo que ***x*** es un vector perteneciente a ese conjunto, cuyos componentes [$\mathbf{x}_i$, $1 \le i \le K$] son variables aleatorias reales y de amplitud continua. Un cuantizador vectorial Q, de dimensión *K* y tamaño *N*, es una transformación de un vector **x,** del espacio euclidiano de dimensión R^K, en un conjunto finito *C* que contiene *N* salidas o puntos de reproducción, llamados *code vectors* (vectores de código):

$$\mathrm{Q} : \mathrm{R}^{\mathrm{K}} \rightarrow \mathrm{C}, \quad \mathrm{C} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{\mathrm{N}}\} \quad \mathbf{y}_{\mathrm{i}} \in \mathrm{R}^{\mathrm{K}} \quad \forall \quad \mathrm{i} \in \mathrm{I} \equiv \{0,1,\ldots,\mathrm{N}\} \quad (9)$$

El conjunto *C* es llamado *code book* (libro de códigos) y tiene un tamaño *N*, esto significa que tiene *N* distintos elementos, cada uno de ellos dentro del espacio R^k.

Asociado a cada cuantizador vectorial de *N* puntos, existe una *partición* de R^k en *N* regiones o *celdas*, R_i para $i \in I$. La *i*-ésima celda esta definida por:

$$R_i = \{ \mathbf{x} \in R^K : Q(\mathbf{x}) = \mathbf{y}_i \} \tag{10}$$

algunas veces llamada *imagen inversa* o *pre-imagen* de $\mathbf{y}_i$ dentro del mapeo *Q* y denotada de forma más consistente por $R_i = Q^{-1}(\mathbf{y}_i)$.

De la definición de celdas, tenemos que:

$$\bigcup_i R_i = R^K \quad \text{y} \quad R_i \bigcap R_j = \varnothing \quad \text{para} \quad i \neq j \tag{11}$$

donde las celdas forman una partición de R^k.

La tarea de codificación de un cuantizador vectorial es, examinar cada vector de entrada **x** e identificar a que celda, *k*-dimensional del espacio R^k, pertenece. El codificador vectorial simplemente identifica el índice *i* de la región y el decodificador vectorial genera el vector del código $\mathbf{y}_i$ que representa a esta región [Gersho, 1997].

El conjunto **y** es conocido como diccionario de reconstrucción o simplemente diccionario, donde *N* es el tamaño del diccionario y $\{\mathbf{y}_i\}$ es el conjunto de vectores del código. Los vectores $\mathbf{y}_i$ son conocidos también en la literatura de reconocimiento de patrones, como los patrones de referencia o plantillas. El tamaño *N* del diccionario, también se conoce como número de niveles, término proveniente de la cuantización escalar. De esta forma, se puede hablar de un diccionario de *N* niveles. Al proceso de creación del diccionario, también se le conoce como entrenamiento o población del diccionario

El modelo de operación de este codificador, se define de forma similar al caso escalar, la *función de selección,* $S_i(\mathbf{x})$, como *indicador* o *función miembro* $1_{Ri}(\mathbf{x})$ para la celda R_i de la partición, esto es:

$$S_i(\mathbf{x}) = \begin{cases} 1 & \text{si} \quad \mathbf{x} \in R_i \\ 0 & \text{c.o.c.} \end{cases} \tag{12}$$

La operación de un cuantizador vectorial puede ser representada como:

$$Q(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{y}_i S_i(\mathbf{x}) \tag{13}$$

Una descomposición estructural (como la mostrada anteriormente), es particularmente evaluable para encontrar un algoritmo efectivo, durante la implementación de la cuantización vectorial [Gersho, 1997].

3.1. Agrupamiento

El agrupamiento es la forma en que se realiza la cuantización vectorial; consiste en lo siguiente: a partir de un conjunto de *N* muestras $\chi = \{X_1, X_2, X_3, ..., X_N\}$, se intentan separar en *K* subconjuntos disjuntos $\chi_1, \chi_2, \chi_3, ..., \chi_k$. En donde cada subconjunto representa a un grupo

(*clúster*) y en el cual, las muestras pertenecientes tienen una mayor similitud entre sí, en comparación a las muestras de cualquier otro grupo.

Existen varios algoritmos de agrupamiento, entre los que tenemos: *simple, distancia máxima, K-Medias, ISODATA y LBG*; estos dos últimos son variantes del agrupamiento *K-Medias*, pero con mayor complejidad.

3.2. Definición del algoritmo de K-medias

Se describe a continuación el algoritmo de agrupamiento *K-Medias.* Su criterio de función es:

$$J_e = \sum_{j=1}^{K} \sum_{x\ en\ \chi_j} d(\mathbf{x}, \mathbf{z}_j) \qquad (14)$$

donde:

K = numero de grupos
$\mathbf{z_j}$ = centro del grupo (centroide) para el grupo j
χ_j = subconjunto de muestras asignadas al grupo j
$d(x, z_j)$ = Es la distancia de *Itakura* entre el ventor *x* y el centroide z_j

Algoritmo de K-medias:

1) Escoger *K* centroides iniciales $\mathbf{z_1}(1)$, $\mathbf{z_2}(1)$, ..., $\mathbf{z_k}(1)$.

2) En la iteración *l*, asignar las muestras a los grupos:
Asignar:

$$\mathbf{x} \quad a \quad \chi_i(l) \quad si \quad d(\mathbf{x}, \mathbf{z}_i(l)) \quad \leq \quad d(\mathbf{x}, \mathbf{z}_j(l)) \quad j = 1,2,...,K \quad j \neq i$$

donde $d(x, z_j(l))$ es la distancia (ó distorsión) de *Itakura*

3) Calcular los nuevos centros de grupo:

$$\mathbf{z}_i(1+1) = \frac{1}{N_i} \sum_{x \in \chi_i(l)} \mathbf{x} \qquad i = 1,2,...,K$$

donde N_i es el número de muestras asignadas a $\chi_j(l)$.

4) Si $\mathbf{z_i}(l+1) = \mathbf{z_i}(l)$ para *i = 1, 2, ..., K,* el algoritmo ha convergido y debe terminarse. En caso contrario, regresar al paso 2.

Una característica de este algoritmo es que los centroides o cuantizadores obtenidos, no son los óptimos globales; es decir, se obtienen cuantizadores óptimos locales. Estos dependen de varios factores, como son: asignación inicial de centroides (principalmente), orden de la toma de muestras, propiedades geométricas de los datos, tipo de distorsión empleada, etc.. Una forma de poder alcanzar los óptimos globales, es probar con una gran variedad de centroides iniciales y seleccionar los cuantizadores finales que tengan la menor distorsión, con respecto a los vectores a los cuales representan. Solución poco factible porque existen un gran número de combinaciones para los cuantizadores iniciales. Otra forma es, elegir los centroides iniciales de forma aleatoria, para buscar una distribución homogénea [Deller, 1987].

4. Desarrollo.

1) Escribir una función en *Matlab* que calcule la distancia de *Itakura* (ecuación 5). La función tendrá como entradas el vector de referencia $y = \begin{bmatrix} -1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}$, el vector de autocorrelación (con respecto a el vector *y*) $R_y = \begin{bmatrix} r(0) & r(1) & \cdots & r(p) \end{bmatrix}$ y el vector desconocido u observado $x = \begin{bmatrix} -1 & a'_1 & a'_2 & \cdots & a'_p \end{bmatrix}$. Tendrá como salida, el valor de la distancia resultante *d*.
2) Con base en los coeficietes LPC obtenidos en la práctica anterior, elaborar una función que obtenga el conjunto de coeficientes LPCC.

3) Escribir el algoritmo de *k-medias*, utilizando la función de distancia del punto 1). Utilice *K* = 16 (16 *centroides* o *codebooks*).

4) Obtener los *codebooks* (*centroides*) de cada uno de los fonemas que forman una palabra dada.

    a. Grabe siete archivos de audio con formato **.wav*, de una misma palabra. Utilice la grabadora de sonidos de Windows o el software equivalente con un formato de conversión PCM, 16 bits, mono y 8000 Hz.

    b. Segmente los archivos de audio en sus respectivos fonemas de forma lineal.
    c. Para los segmentos que representan los mismos fonemas se calcula la autocorrelación y los coeficientes LPC's, con tramas de 128 muestras (utilice el algoritmo de la práctica pasada).

    d. Obtenga 16 *centroides* para cada uno de los fonemas utilizando el algoritmo de *K-medias* para k = 16 y la distancia de *Itakura*.

Bibliografía

John R. Deller Jr, John G. Proakis, John H. L. Hansen, "Discrete–Time Processing of Speech Signals", Ed. Prentice Hall 1987

Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Ed. Prentice Hall, U.S.A, 1993.

Gersho, A. & Gray, R. M. "*Vector Quantization and Signal Compression"*. Sexta Edición. Editorial Kluwer Academic Publishers. Norwell, Massachusetts, U.S.A., 1997.

"Apuntes de clases de Procesamiento Digital de Voz", Abel Herrera Camacho

"The Itakura distance measure", Speech Vision Robotics group, Tony Robinson, 1998, http://mi.eng.cam.ac.uk/~ajr/SA95/node47.html#SECTION00079000000000000000

Apéndice

Para utilizar la grabadora de sonido realice lo siguiente:

- Verifique que el micrófono con que cuente este correctamente conectado a la computadora.

- De “clic” en INICIO/Programas/Accesorios/Entretenimiento/**Grabadora de Sonidos.**
- De “clic” en Archivo/Propiedades/Convertir Ahora, para seleccionar el formato deseado.