**ENSAYO SOBRE LOS PROBLEMAS DE AUDIO**

**(COMPUTACION BLANDA)**

**GERMAN JOSE ANDUQUIA RAMIREZ**

**1225092654**

**UNIVERSIDAD TECNOLOGICA DE PEREIRA**

**2020**

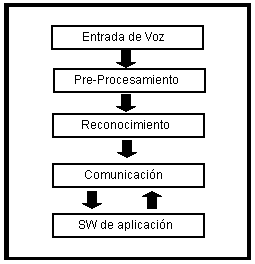
**RECONOCIMIENTO DE VOZ**

Reconocimiento de voz generalmente es utilizado como una interfaz entre humano y computadora para algún software.

Debe cumplir 3 tareas:

* Preprocesamiento: Convierte la entrada de voz a una forma que el reconocedor pueda procesar.
* Reconocimiento: Identifica lo que se dijo (traducción de señal a texto).
* Comunicación: Envía lo reconocido al sistema (Software/Hardware) que lo requiere.

Componentes en una aplicación



Existe una comunicación bilateral en aplicaciones, en las que la interfaz de voz está íntimamente relacionada al resto de la aplicación. Estas pueden guiar al reconocedor especificando las palabras o estructuras que el sistema puede utilizar. Otros sistemas sólo tienen una comunicación unilateral.

Los Procesos de pre-procesamiento, reconocimiento y comunicación deberían ser invisibles al usuario de la interfaz. El usuario lo nota el desempeño del sistema de manera indirecta como: *certeza* en el reconocimiento y *velocidad*. Estas características las utiliza para evaluar una interfaz de reconocimiento de voz.

**Los Datos del Reconocimiento de Voz**

Los sistemas de reconocimiento de voz se enfocan en las palabras y los sonidos que distinguen una palabra de la otra en un idioma. Estas son los fonemas. Por ejemplo, "tapa", "capa", "mapa", "napa", son palabras diferentes puesto que su sonido inicial se reconocen como fonemas diferentes en español.

Existen varias maneras para analizar y describir el habla. Los enfoques más comúnmente usados son:

1. **Articulación:** Análisis de cómo el humano ***produce*** los sonidos del habla.
2. **Acústica**: Análisis de la ***señal*** de voz como una secuencia de sonidos.
3. **Percepción** **Auditiva:** Análisis de cómo el humano ***procesa*** el habla.

Los tres enfoques proveen ideas y herramientas para obtener mejores y más eficientes resultados en el reconocimiento.

**1 Articulación**

La articulación centra su atención en el aparato vocal: Garganta, boca, nariz, en donde se producen los sonidos del habla.

* Garganta: Contiene las cuerdas vocales, cuya vibración produce los fonemas "hablados".
* Boca y Nariz: "Cavidades de resonancia" por que refuerzan ciertas frecuencias sonoras.
  + Cuando el paladar suave baja y deja el aire pasar por la nariz se generan los fonemas nasales (/m/ /n/)
  + La boca consiste de:
    - Puntos de articulación

Dientes

Puente alveolar (puente óseo atrás de los dientes superiores)

Paladar duro

Paladar suave o velum

y de

* + - * articuladores

Labios

Lengua

**2 La Señal (Acústica)**

Un reconocedor no puede analizar los movimientos en la boca!. En su lugar, la fuente de información es la señal de voz misma.

El Habla es una señal **analógica** , es decir, un flujo continuo de ondas sonoras y silencios.

El conocimiento de la ciencia de la acústica se utiliza para identificar y describir los atributos del habla que son necesarios para un reconocimiento de voz efectivo.

Cuatro características importantes del análisis acústico son:

* Frecuencias
* Amplitud
* Estructura Armónica (tono vs ruido)
* Resonancia.

**3 Percepción Auditiva**

La variabilidad del habla producida por coarticulación y otros factores hacen del análisis de la voz extremadamente difícil.

La facilidad del humano en superar estas dificultades sugiere que un sistema basado en la percepción auditiva podría ser un buen enfoque. Desafortunadamente nuestro conocimiento de la percepción humana es incompleto.

Lo que sabemos es que el sistema auditivo está adaptado a la percepción de la de voz.

El oído humano detecta frecuencias de 20Hz a 20,000 Hz pero es más sensible al rango entre 1000 y 6000 Hz. También es más sensible a cambios pequeños en la frecuencia en el ancho de banda crítico para el habla. Además el patrón de sensibilidad a cambios en el tono (pitch) no corresponde a la escala lineal de frecuencias de ciclos por segundo de la acústica.

Para representar mejor al patrón de percepción del oído humano, se desarrollo una escala llamada mel-scale, la cual es una escala logarítmica.

Estudios recientes muestran que el humano no procesa frecuencias individuales independientemente, como lo sugiere el análisis acústico. En su lugar escuchamos grupos de frecuencias y somos capaces de distinguirlas de ruidos alrededor.

**Muestreo (Sampling)**

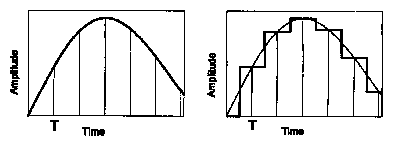
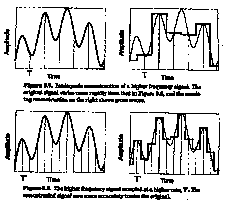
Asigna un valor numérico a la señal en unidades discretas de tiempo (constante) .  
La frecuencia Nyquist: Específica la frecuencia máxima de una señal que puede reproducirse completamente. Esta establece que

*Frecuencia muestreo = 2 veces la frecuencia máxima de la señal*

para poder reproducirla.

Para poder reproducir la señal análoga debe pasar por un filtro ***pasa-bajas*** a la frecuencia de muestreo (quitar ruidos creados por el muestreo).

Al no cumplirse estas condiciones sucede el fenómeno de ***aliasing***

Señal que varia lentamente: muestreada en una frecuencia T  
  
Si la señal varía más rápido se requiere una T más pequeña por lo tanto un menor ancho de banda de frecuencias.  
  
El humano produce señales de Voz desde los 100(hombre)-400(mujer) Hertz hasta los 15000Hz.

Teléfono : 3100Hz por lo tanto se muestrea a 8000 Hz, inteligible pero baja calidad.

Comparado con un CD, se muestrea a 44.1Hz

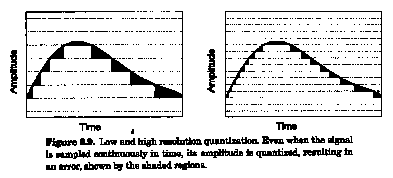
20Khz  
El ancho de banda es mayor para instrumentos que para voz. Pero la diferencia es audible!!  
Por lo tanto se requiere mayor espacio para almacenar y transmitirla.

**Resolución: Cuantización**

Cada muestra se representa con un valor digital limitando el rango de valores discretos correspondiente al original.

Ej.  
Utilizando 4 bits se pueden representar 16 valores diferentes.  
Con 8 bits ya son 256 valores.

Esto se puede ver como la resolución:



El error o diferencia entre la señal original y la reconstruida se percibe como **ruido**.

Por lo tanto, a mayor resolución mayor cuantización y menor ruido como consecuencia.

La resolución del "cuantizer" (# de bits por muestra) se describe generalmente en términos de la relación señal-a-ruido (signal to noise ratio o SNR)

A mayor SNR es mayor la fidelidad de la señal digitalizada.

SNR aprox 2 B (B=bits/muestra)

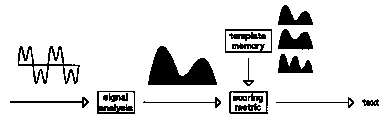
Es independiente de la frecuencia de muestreo.

Teléfono: 8bits/muestra , es decir, si muestreamos a 8kHz tenemos 8000 muestras por segundo y así 8000 x 8 = 64000 bits por segundo.

CD: 16bits/muestra, por lo tanto 44100 muestras por segundo x 16bits = 705600 bits por segundo (mono) para 2 canales (estéreo) se duplica.

**Template-based Approach**

Teoría que ha dado lugar a toda una familia de técnicas de reconocimiento de voz, causando un considerable avance durante las décadas de los 70’s y 80’s.



El principio es simple: Se almacenan patrones de voz típicos (templates) como modelos de referencia en un diccionario de palabras candidato. El reconocimiento se lleva a cabo comparando la expresión desconocida a los patrones almacenados (templates) seleccionando el que más se le parece.

Usualmente se construye patrones para las palabras completas.

Ventaja: Se evitan errores debidos a la segmentación o clasificación de unidades pequeñas que pueden variar mucho acústicamente, como los fonemas.

Desventaja: Cada palabra requiere de su patrón y el tiempo de preparar y compararlos se vuelve demasiado al incrementarse el tamaño del vocabulario.

Inicialmente se pensaba que se restringía sólo a reconocimiento dependiente del locutor. Sin embargo se logro un reconocimiento independiente del loc. Utilizando técnicas de "Clustering" para generar automáticamente grupos de patrones para cada palabra del vocabulario.

También pasó de ser para el reconocimiento de palabras aisladas a habla continua utilizando técnicas de programación dinámica para encontrar la mejor cadena de patrones.

**a) Medición de características:**

Se trata básicamente de una técnica de reducción de datos en la cual el gran número de datos en la señal grabada es transformados en un grupo más pequeño de caracteres que describen las propiedades importantes de la señal.

El tipo de características que se calculan depende de:

1. Tiempo para calcularlas
2. Espacio para almacenarlas
3. Facilidad de Implementación.

Análisis mediante "banco de filtros" (Filter Bank)

Una muestra de voz se puede aproximar como una combinación lineal de muestras anteriores.

**b) Determinación de similitud entere patrón de referencia de voz y serial capturada.**

R(t) T(t)

Se busca una función de alineamiento w(t) que mapee R y las partes correspondientes de T.

El criterio de correspondencia es la minimización de una medida de distancia D(T,R)

Por lo tanto se busca una w(t) tal que mística ***d(T(t), R(W(t))) función de ponderación deriva de w(t)***

D(T,R)= min ***INTEGRAL***

dw(t)

**conjunto** de funciones diferenciables que se incrementan monotónica mente

complicado ***snore***puede ***resalren***en general por lo tanto se debe ***discretizar***.

Se buscan una alineación en tiempo "optima" a través de una curva que relacione el eje de tiempo m de R a el eje de tiempo del patrón T.

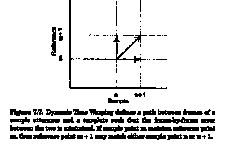
m=w(n)

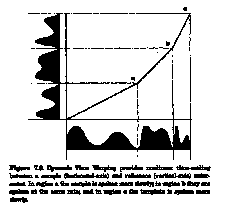
Restricciones w(I)=1

W(NT)=NR

Para determinar el tipo de la alineación se propusieron diversas técnicas:

1. alineamiento lineal m=w(n)=(n-1)[(NR-1)/(NT-1)]+1
2. Mapeo de eventos temporales (significativos) en ambos patrones
3. Máxima Correlación se varía la función para maximizar la correl entre RyT
4. Dynamic Time Warping (-se calcula la distancia entre R y T)





Se debe encontrar una medida que indique qué tan similares con los patrones R y T. Para ello es necesario alinearlos de alguna forma.

C**) Reglas de decisión**

Nearest neighbors el candidato con la menor distancia o una lista ordenada por distancias (de ***menos a mayor***)

Se usa cuando se tienen varios patrones de referencia para cada cantidad R.

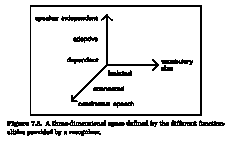
**3. Creación de un Reconocedor de Voz**

Pasos

1) Entrenamiento- adquisición de los conjuntos de características para cada palabra en el vocabulario.

2) Clustering- Creación de los patrones de referencia.

3) Pruebas



A medida que transcurre el tiempo la tecnología y los algoritmos se van actualizando de manera significativa, con la capacidad propia de autoaprendizaje, aplicando así de forma automática todas estas herramientas vistas anteriormente. Al detectar una voz es fácil comparar estos datos con los de otra persona y determinar finalmente a quien pertenece.

**Consultas Web**

<http://ict.udlap.mx/people/ingrid/Clases/IS412/>