ETIQUETADO NO FONEMATICO PARA UN SISTEMA AUTOMATICO MONOLOCUTOR DE RECONOCIMIENTO DE PALABRAS AISLADAS

Por Antonio J. Rubio Ayuso (*) y M. Carmen Carrion Perez (**)
(*) Departamento de Electronica
(**) Departamento de Electricidad
Universidad de Granada

I. Introduccion

Presentamos en estas lineas una de las partes de nuestro trabajo que pueden considerarse terminadas desde el punto de vista logico. Trabajamos en otros temas dentro del mismo campo, pero eso lo presentaremos en otros numeros.

Aqui presentamos un Sistema Automatico de Reconocimiento Monolocutor con Etiquetado No Fonematico (SARMENF) (1). Funciona a base de dividir la señal sonora en segmentos de la misma duracion e intentando clasificar cada segmento entre unos prototipos obtenidos durante la Fase de Entrenamiento, mediante un criterio de distancia euclidea. La etiqueta que se asigna a cada segmento no guarda ninguna relacion con los fonemas del lenguaje. Este tipo de etiquetado puede ser ventajoso debido a la dificil localizacion del fonema como unidad de analisis.

II. Esquema general

Hay dos fases en el funcionamiento del SARMENF: fase de entremiento y fase de reconocimiento.

2.1 Fase de Reconocimiento

Durante la Fase de Reconocimiento se realizan los siguientes pasos:

1) Una vez que el locutor ha pronunciado una palabra del vocabulario permitido, el primer paso consiste en muestrear la señal, in-

troduciendo las muestras en un ordenador.

La frecuencia de muestreo es de 10 kHz. La conversion A/D se hace a traves de un conversor A/D de 12 bits. Para cada palabra, se toman 15000 muestras, durante 1'5 s.

2) Con objeto de reducir el volumen de calculo, los extremos de la señal sonora son localizados. En general, ninguna palabra ocupa completamente las 15000 muestras.

3) Una vez que los extremos han sido localizados, se establece una segmentacion regular de la señal sonora. Los segmentos son de 10 ms (100 muestras) y se solapan a la mitad.

Sobre cada segmento se calcula un conjunto de parametros. En primer lugar se hace una clasificacion previa entre segmentos sonoros y no sonoros, utilizando la Energia y el Numero de pasos por cero del segmento (ZCR). Se clasifican como segmentos sonoros aquellos que cumplen la inecuacion

Energia) A+B*ZCR

donde las constantes A = .1605E+08 y B = 1.075E+06 se han determinado experimentalmente y son funcion de la ganancia global del sistema de audio.

Se clasifican como no sonoros aquellos segmentos que cumplen la desigualdad

Energia) C+D*ZCR

donde las constantes C = .2012E+08 y D = 8.047E+05 tambien se han determinado experimentalmente.

Aquellos segmentos que no cumplen ninguna de las dos condiciones anteriores son clasificados como dudosos y son ignorados en el tratamiento posterior.

Si un segmento se clasifica como sonoro, se calculan sobre el otros cuatro parametros adicionales: los tres primeros formantes y el logaritmo de la razon de las areas de la primera y segunda seccion del tubo acustico equivalente al tracto vocal. Esta determinacion se hace mediante Prediccion Lineal (2).

Por otra parte, si el segmento se clasifica como no sonoro, los cuatro parametros adicionales son los valores medios del espectro de la señal dividido en cuatro zonas o bandas (0 - 1.25 kHz, 1.25 - 2.5 kHz, 2.5 - 3.75 kHz y 3.75 - 5 kHz).

4) Los parametros correspondientes a un segmento se comparan con los de los Segmentos de Referencia obtenidos en la Fase de Entrenamiento. De esta comparacion se obtiene una etiqueta para el segmento en cuestion. Asi se consigue una cadena de etiquetas (una etiqueta por cada 5 ms) que representa a la palabra emitida.

5) El ultimo paso en la Fase de Reconocimiento consiste en comparar la cadena de etiquetas con las Cadenas de Referencia, obtenidas tambien en la Fase de Entrenamiento (una Cadena de Referencia por cada palabra del vocabulario). Esta comparacion se lleva a cabo por medio de un algoritmo de Programacion Dinamica (3). Este algoritmo permite comparar cadenas de etiquetas de diferente longitud, pro-

porcionando la "Similitud" entre ellas en funcion de la distancia euclidea entre los diferentes simbolos de cada una de las cadenas comparadas.

Un esquema de bloques para la Fase de Reconocimiento puede verse en la parte derecha de la Figura 1.

2.1 Fase de Entrenamiento

Esta Fase esta representada en la parte izquierda de la Figura 1.

El Entrenamiento consiste en la pronunciacion de todas las palabras del vocabulario, cada una de ellas repetida cuatro veces.

Despues del muestreo de cada palabra de entrenamiento, se calculan sus extremos y se obtienen los parametros correspondientes a los segmentos de la señal, del mismo modo que en la Fase de Reconocimiento.

Cada segmento se considera como un punto de un hiper-espacio euclidiano de seis dimensiones. En realidad se manejan dos hiper-espacios: uno para los segmentos sonoros y otro para las no sonoros. De este modo, la voz del locutor se caracteriza por una cierta distribucion de puntos a lo largo de ambos hiper-espacios.

Los pasos que siguen en la Fase de Entrenamiento son:

1) Analisis de las nubes formadas por los segmentos de las palabras pronunciadas, en ambos hiper-espacios.

Aquellas nubes claramente separadas del resto de puntos se consideran como correspondientes a segmentos de señal del mismo tipo. Su centro geometrico representara al Segmento de Referencia, con una etiqueta asociada.

La obtencion de las nubes y Segmentos de Referencia se lleva a cabo mediante el algoritmo "k-medias" modificado (4).

La modificacion puede describirse brevemente. La principal fuente de fallos en el algoritmo "k-medias" original esta en la seleccion del conjunto inicial de centros. Evitamos este problema de la siguiente forma:

Se hace una normalizacion de todos los parametros de modo que la media de cada parametro sea 1. Asi, el punto (1,1,1,1,1,1) es el centro de la distribucion completa. Entonces, se colocan los puntos iniciales de modo que esten uniformemente repartidos alrededor del punto (1,1,1,1,1,1).

El numero de Segmentos de Referencia (k) se ha establecido experimentalmente. Para los segmentos sonoros $k = 21$ y para los segmentos no sonoros $k = 7$.

2) De acuerdo con los Segmentos de Referencia, cada palabra del vocabulario se analiza con el fin de elaborar una cadena de etiquetas (Cadena de Referencia) para esa palabra. Estas Cadenas de Referencia son las que se usan en el algoritmo de Programacion Dinamica (Fase

de Reconocimiento) para comparar una palabra emitida con todas las del vocabulario.

III. Conclusion

En este trabajo se ha presentado un sistema automatico de reconocimiento monolocutor con etiquetado no fonematico. Las principales caracteristicas son:

1) Es un sistema de reconocimiento monolocutor para palabras aisladas.

2) Se lleva a cabo un etiquetado no fonematico. Hay 21 diferentes tipos de segmentos sonoros y 7 tipos de segmentos no sonoros.

3) El vocabulario es un conjunto de 19 palabras castellanas relacionadas con el calculo aritmetico: "Cero", "Uno", "Dos", "Tres", "Cuatro", "Cinco", "Seis", "Siete", "Ocho", "Nueve", "Parentesis", "Cierra", "Mas", "Menos", "Por", "Partido", "Raiz", "Igual?" y "Punto". Este vocabulario permitira el manejo oral del ordenador en calculos aritmeticos simples.

4) Los resultados experimentales estan hasta el momento alrededor del 65% de reconocimiento correcto. Merece la pena destacar que no se ha puesto especial atencion en la seleccion de las Cadenas de Referencia.

5) El sistema ha sido implementado sobre un ordenador ECLIPSE-250 de DATA GENERAL, en lenguaje de programacion FORTRAN V.

IV. Referencias

(1) R. Reddy
"Speech recongnition by machine: a review"
Proc. IEEE, vol. 64, pp. 505-531

(2) J. D. Markel, A. H. Gray
"Linear Prediction of Speech"
Springer-Verlag, 1976.

(3) H. Sakoe, S. Chiba
"Dynamic programming algorithm optimization for spoken word recognition"
IEEE Trans. on ASSP. Vol. ASSP-26, n. 1, pp. 43-49, 1978.

(4) J. T. Tou, R. C. Gonzalez
"Pattern recognition principles"
Addison-Wesley, 1974.

Nota: Este trabajo ha sido presentado al "1983 SPAIN WORKSHOP ON SIGNAL PROCESSING AND ITS APPLICATIONS" que se celebrara en Sitges (Barcelona) en Septiembre de 1983.

ESQUEMA - RESUMEN

Martin S. Ruiperez, Catedrático de Filologia Griega
Universidad Complutense, Madrid

Aparte de temas de Filología Clásica y Lingüística Comparada Indoeuropea ha trabajado en:

-Lingüística estructural y funcional (griego y español), concretamente morfosintaxis y fonología.

-Con equipo de alumnos y colaboradores, sin ordenador y sólo con calculadoras, ha establecido (1968) un vocabulario de frecuencias de griego ático de 400-350 a. C. Igualmente un índice morfológico verbal de frecuencias. Todo ello para la elaboración de un método de enseñanza del griego antiguo, que está en fase de experimentación en varios Institutos y Facultades.

-Ha dirigido una tesis doctoral de lingüística cuantitativa aplicada al problema de los arcaísmos en Homero.

-Desde 1973 se ha familiarizado con cálculo programado.

-Desde 1980 se ha familiarizado con ordenadores, concretamente con un HEWLETT-PACKARD HP250, provisto de base de datos.

Además de programas comerciales, ha realizado programas lingüísticos:

1. Construcción de cadenas a partir de segmentos conmutables. Concretamente, conversión de un número introducido en dígitos en texto español, con perfecta coherencia sintáctica.

2. Corte de palabras conforme a las reglas de la Academia y al buen uso tipográfico para programas de textos cuando al final de una línea no cabe la palabra entera. Versiones para Castellano, Catalán, francés e italiano. Disponible en el mercado.

   Tiene un programa de calendario (que incluye la reforma del calendario gregoriano) para determinar en qué día de la semana cae cualquier fecha entre el 1 de enero del año 1 (!) y el 28 de febrero del 2100; número de días entre dos fechas y fecha a partir de otra y de un número de días de intervalo.

Todo en lenguaje BASIC y "Extended BASIC" o "Business BASIC" (en realidad BASIC enriquecido con PASCAL)

—Desde diciembre de 1982 dispone en su cátedra de un modesto equipo compuesto por:

-Ordenador HEWLETT-PACKARD HP87 con 128 Kb de memoria RAM, más un ROM de Input/Output y otro de Assembler.
-Periférico de una unidad dual de diskettes floppies con capacidad total de 540 Kb
-Periférico de impresora de agujas EPSON III/100

Actualmente con este equipo está desarrollando "software" para tratamiento científico de textos: introducción, almacenamiento, modificación, impresión, búsqueda y confección de índices alfabetizados (con alfabetización distinta para el griego, en el que, por ejemplo, la z no va al final sino después de la e).

Se prevé la aplicación inmediata al análisis de textos métricos griegos codificados y de un repertorio de inscripciones.

— En 1982-83 curso monográfico de doctorado sobre los ordenadores en la investigación lingüística y filológica

Sevilla, 15 de abril de 1983.