J. M. Pardo

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS
DE TELECOMUNICACIÓN

TESIS DOCTORAL

SISTEMAS DE RECONOCIMIENTO DE HABLA CONTINUA Y AISLADA: COMPARACIÓN Y OPTIMIZACIÓN DE LOS SISTEMAS DE MODELADO Y PARAMETRIZACIÓN

Ricardo de Córdoba Herralde
Ingeniero de Telecomunicación

Director de la tesis
Dr. Ingeniero José Manuel Pardo Muñoz

1995

Resumen

El título de esta tesis puede parecer ambicioso por la amplitud de temas planteados, y puede que realmente lo sea, pues su objetivo es tratar de dominar toda una serie de técnicas con las que mejorar el modelado y parametrización de dos sistemas de reconocimiento de habla, que tienen las siguientes características:
1) Habla continua, vocabulario de 1000 palabras, dependiente del locutor, en inglés.
2) Habla aislada, vocabulario de dígitos, independiente del locutor, en castellano.

Los puntos básicos que se han tratado son:
1) *Parametrización:* se ha trabajado en la manera de caracterizar el espectro de la señal de voz, en cuanto a la forma de obtener la energía en distintos filtros en frecuencia, como paso previo a la obtención de los parámetros. Se han introducido distintas maneras de realizar un filtrado de dichos valores para conseguir que sean robustos frente al ruido telefónico.

2) *Tipo de modelado:* se ha trabajado con tres tipos de modelado: discreto, continuo y semicontinuo. En este último se ha introducido una técnica de preselección con la que reducir el tiempo de cálculo sin perder tasa de reconocimiento. En modelado continuo se ha tratado el equilibrio "número de unidades/número de mezclas por unidad", y la manera de "guiar" los experimentos hasta alcanzar un óptimo.

3) *Unidad de modelado:* se han estudiado distintas posibilidades en nuestros dos sistemas: palabra, fonema, trifonema generalizado y agrupamiento a nivel de estado. Se han introducido nuevas técnicas de agrupamiento de distribuciones con las que se consigue una compartición de parámetros óptima. También se ha

tratado el tema del suavizamiento de las unidades contextuales (mediante su interpolación con unidades no contextuales). Así mismo, se ha introducido una técnica con la que duplicar el número de modelos con el objeto de mejorar el reconocimiento suponiendo dos tipos de locutores.

4) *Detección de principio y fin:* (habla aislada)
Se ha introducido una técnica basada en una red neuronal con la que se consigue una mejora significativa tanto en la precisión de las marcas obtenidas como en la tasa de reconocimiento que se alcanza a continuación cuando se utiliza.
Así mismo, utilizando modelos de ruido inicial y final comunes a todos los modelos, y como algoritmo de reconocimiento un *one-pass,* también se ha conseguido superar el efecto de una detección defectuosa.

Abstract

The title of the thesis may sound too ambitious, because it covers a wide range of topics, and may be it is, as the there are many techniques involved in our purpose of improving the modelling and parametrization of two speech recognition systems, with the following characteristics:
1) Continuous speech, 1000 words vocabulary, speaker dependent, in English.
2) Isolated speech, digits vocabulary, speaker independent, in Spanish.

The basic points covered are:
1) *Parametrization:* we have worked in the way to deal with the speech spectrum, to obtain the energy using different filters, as the previous step to obtain the parameters. We have introduced different ways to filter these values of energy to obtain robustness in the telephonic environment.

2) *Modelling technique:* we have worked with three types of modelling: discrete, continuous and semicontinuous. With semicontinuous, we have introduced a preselection technique to reduce computing time with no decrease in recognition rate. With continuous modelling we have dealt with the balance between the number of units and the number of mixtures in each unit, and the way to "guide" the experiments to reach an optimum.

3) *Modelling unit:* we have used different posibilities in our systems: word, phoneme, generalized triphone and state clustering. We have introduced new ways to cluster the distributions of the models in order to obtain an optimum parameter sharing. We have also dealt with the smoothing of contextual units (interpolating them with context independent units). We have introduced a technique to double the number of models in order to improve the recognition asuming there are two kinds of speakers.

4) *Begin-end detection:* (isolated speech)

We introduced a technique based on a neural network that obtains a significant improvement both in accuracy of the end-pointing and in the recognition rate obtained after it.

By the other hand, using begin and end noise models shared by all the digit models, and one-pass as the recognition algorithm, we have override the effect of the unaccurate end-pointing.

Índice