Representación de la voz en el reconocimiento del habla

Voice representation in speech recognition

Climent Nadeu

Para reconocer el habla de forma automática se requiere una representación paramétrica de la voz que retenga sus características relevantes. En este artículo se exponen las ideas básicas del proceso de extracción de dichas características a partir de la señal de voz recogida por el micrófono, indicando las propiedades de producción y percepción de la voz que están en juego y fijándose particularmente en la consecución de representaciones robustas al entorno.

Automatic speech recognition requires a parametric representation of the speech signal which carries its relevant features. In this paper, the basic ideas underlying the process of feature extraction from the speech signal are described, the involved properties of voice production and perception are summarised, and a particular attention is payed to environmental robust techniques.

Los sistemas informáticos que permiten interacción oral con el usuario van ganando, lentamente, en prestaciones y en naturalidad. Así, en nuestra vida cotidiana se nos irán haciendo cada vez más familiares acciones como el control de un dispositivo mediante órdenes orales o el acceso a un servicio de información *dialogando* con el sistema. Y, a más largo plazo, veremos cómo la voz sustituye en gran medida al teclado, o el asistente personal digital nos traduce lo que está diciendo nuestro interlocutor humano.

Todas estas aplicaciones requieren que el sistema sea capaz de convertir la voz captada por un micrófono en una secuencia de palabras, proceso que se denomina *reconocimiento del habla*. Dicha secuencia puede ser el resultado final que se persigue, como en el caso del dictado automático, o bien la entrada a una etapa de procesamiento posterior que permita comprender el significado de lo que está diciendo el usuario.

Los sistemas actuales de reconocimiento automático representan el habla mediante modelos estadísticos (modelos de Markov ocultos) de las unidades fonéticas elementales (normalmente fonemas contextuales), así como de las relaciones que se establecen entre dichas unidades para componer las palabras (transcripciones fonéticas) y entre las palabras para componer las frases (gramática). Con estos modelos estadísticos, y mediante el cálculo de probabilidades, se estima la secuencia de palabras que ha sido pronunciada. La comprensión del habla utiliza adicionalmente el conocimiento semántico del dominio de la aplicación para captar el significado de la elocución de entrada al sistema a partir de la cadena (o cadenas alternativas) de palabras que suministra el reconocedor.

La inmensa mayoría de sistemas de reconocimiento actuales se basan en modelos estadísticos obtenidos con algoritmos de aprendizaje o entrenamiento que extraen las características del habla implícitamente contenidas en grandes bases de datos orales (y también textuales, para entrenar las gramáticas), convenientemente transcritas y etiquetadas.

Como se comprenderá, para que el sistema ofrezca una buena tasa de reconocimiento, debe existir una correspondencia entre la base de datos de entrenamiento y la tarea que deberá abordar el reconocedor cuando esté funcionando. Por ejemplo, para entrenar adecuadamente modelos de las unidades fonéticas elementales, dichas unidades deben aparecer un número suficiente de veces en la base de aprendizaje y en variedad de contextos. Si se quiere que el sistema sirva para cualquier hablante de un idioma o dialecto (sistema independiente del locutor), la base de datos de aprendizaje deberá contener las voces de un número elevado de locutores (por ejemplo 5000 en las bases SPEECHDAT recogidas para la mayoría de lenguas de la Europa occidental).

Puesto que la voz consiste en una secuencia de sonidos elementales diferenciados, el cálculo de probabilidad que realiza el sistema de reconocimiento se lleva a cabo sobre una secuencia temporal de observaciones, conteniendo cada una de ellas las características de la señal de voz correspondientes a un segmento temporal determinado que son relevantes para el reconocimiento. Idealmente, estas características deberían representar un segmento de voz uniforme, es decir, un evento acústico bien definido. Sin embargo, una simple mirada a la forma de onda de la señal de voz entregada por el micrófono muestra que ésta no puede describirse simplemente como la

concatenación de segmentos uniformes, y que resulta imposible determinar una frontera nítida entre sonidos, como consecuencia de la coarticulación entreexistente.
Esta es la razón que hace preferible que la voz se segmente en tramos de longitud y desplazamiento fijos, y que los intentos de representarla como una sucesión de segmentos variables no hayan dado lugar a una mejora de resultados de reconocimiento.
Así pues, se trata de representar cada tramo de la señal de voz (de longitud entre 20 y 30 milisegundos) mediante un conjunto de características que se corresponden con un modelo paramétrico de dicha señal. Como veremos más adelante, dichas características, a menudo denominadas parámetros, no pretenden captar toda la complejidad de la señal acústica, sino sólo su espectro, es decir, la distribución de la energía de la señal a lo largo de la frecuencia. Y, simplificando más todavía, los parámetros representan únicamente la forma que envuelve el espectro (envolvente espectral o formantes), dejando de lado la estructura armónica que caracteriza los segmentos sonoros, es decir, los que se generan a partir del efecto de vibración de las cuerdas vocales.
Dicha estructura armónica, que es debida a la periodicidad de la señal, determina el tono de la voz y éste se ha usado raramente en reconocimiento hasta el momento presente, excepto en lenguas tonales como el mandarín; no ocurre así en el proceso de compresión y codificación digital de la voz para su transmisión y posterior reproducción, puesto que la entonación es un elemento perceptivo importante.
Cuando un sistema de reconocimiento funciona en situaciones reales se encuentra con condiciones adversas, como ruidos, distorsiones y otros, que degradan la calidad de la voz y, además, hacen que los parámetros representativos ya no estén en sintonía con los modelos estadísticos desarrollados en la fase de aprendizaje del sistema. La capacidad del sistema para hacer frente a estos cambios de las condiciones del entorno se denomina *robustez*. En la actualidad, se trabaja intensamente en el desarrollo de técnicas para el reconocimiento robusto del habla.
A continuación, se van a presentar las técnicas de obtención de los parámetros que caracterizan –cada tramo de– la voz en un sistema de reconocimiento y también, aunque más brevemente, las que se utilizan en la compresión de la voz para su transmisión digital. Son técnicas de tratamiento de señal, muchas de las cuales se utilizan también en otros ámbitos de aplicación, como telecomunicaciones, robótica, bioingeniería, geofísica, etc. Antes de abordarlas, se repasará brevemente el conocimiento sobre la producción y percepción de la voz que se utiliza en ellas.

Modelado de la producción y percepción de la voz

Contemplando el funcionamiento del aparato fonador humano se observa ante todo el movimiento de los órganos articulatorios que dan forma a una cavidad acústica, el tracto vocal. Pero para que se produzca el sonido es necesaria una fuente de ondas de presión del aire, conformada por la vibración de las cuerdas vocales (caso sonoro), o por una fricación o aspiración (caso sordo). Ante un observador habituado a la teoría de sistemas, este mecanismo acústico sugiere enseguida un modelo de entrada-salida, en el que la señal de la voz es la salida de un sistema lineal o filtro en cuya entrada se encuentra la fuente acústica antes mencionada.
Dicho modelo de producción de la voz basa su sencillez en la separación que realiza entre el filtro, que simula el funcionamiento del tracto vocal, el cual a su vez confiere a cada sonido su timbre característico, y la excitación o entrada, que da cuenta del tipo de fuente acústica (sorda o sonora) y, en el caso sonoro, de la frecuencia de vibración de las cuerdas vocales, denominada *frecuencia fundamental* o *tono de la voz*.
En este modelo simplificado se considera que el filtro es puramente recurrente, es decir, que es un sistema lineal sin ceros, sólo con unos 10 polos. Puesto que la envolvente espectral de la señal de voz generada con este modelo la determina el filtro, ello equivale a suponer que sólo interesan los picos de la envolvente espectral, no sus valles.
Gracias a dicho modelo, basado en el análisis de la transmisión de las ondas planas del sonido a través de un tubo sin pérdidas de sección no uniforme, es posible el uso de una técnica eficiente de determinación de los parámetros del filtro que, fundamentada en la idea de predicción lineal óptima, constituye la base de la mayoría de sistemas de compresión de la señal de voz. En reconocimiento del habla, también se ha empleado de modo extenso la predicción lineal, pero la técnica alternativa

basada en subbandas, que veremos más adelante, la ha ido desplazando gracias a su mejor comportamiento frente a las perturbaciones de la señal y a la mayor flexibilidad que le confiere el hecho de trabajar en bandas separadas.
Hasta ahora hemos tratado el modo cómo se produce la voz. Pero al ser la voz un medio de comunicación, interesa conocer también el mecanismo de percepción ya que, lógicamente, las características de la señal estarán en función no sólo del aparato productor sino también del receptor, el oído.
Hay dos propiedades del aparato auditivo humano que son relevantes en el desarrollo de sistemas tanto de reconocimiento del habla como de compresión y codificación de la voz. La primera es el efecto de enmascaramiento, por el que un sonido puede dejar de oírse cuando está situado frecuencialmente (o temporalmente) cerca de otro sonido de intensidad suficientemente alta. Esta propiedad ha resultado de gran valor en el desarrollo de sistemas de compresión de audio y, en particular, de voz. En efecto, a la distorsión introducida por efecto de la reducción del número de bits que representan la voz se le hace jugar el papel del sonido enmascarado.
En segundo lugar, la cóclea del oído funciona como un analizador espectral, trabajando en bandas frecuenciales no uniformes que se hacen sucesivamente más anchas a medida que crece la frecuencia (tenemos una muestra de ello en el piano, donde las teclas de sonidos agudos están más distanciadas en frecuencia que las de sonidos más graves). Pues bien, la técnica basada en subbandas y que, según hemos afirmado, es la más utilizada para representar la envolvente espectral de la voz en el reconocimiento del habla, imita de algún modo el análisis frecuencial realizado por la cóclea. Los parámetros representativos del tramo de voz que esta técnica determina son las fracciones de energía de la señal correspondientes a unas 20 bandas frecuenciales distribuidas según una escala no uniforme denominada *mel* que, determinada con experimentos perceptivos, refleja la resolución frecuencial del oído humano.

Compresión y codificación de la voz

La señal de voz es altamente redundante, como se nota a simple vista observando, por ejemplo, la forma de onda de una vocal. El problema que se plantea en la compresión de voz es la extracción de dicha redundancia para poder almacenar la voz o transmitirla por vía digital (en telefonía móvil, por ejemplo) de forma eficiente. En los sistemas de compresión y codificación más utilizados se codifican, por un lado, los parámetros del filtro, obtenidos mediante el cálculo de un predictor de la señal optimizado y, por otro, la señal que actúa de entrada al filtro en la reconstrucción de la voz en el terminal receptor. Los distintos sistemas de compresión se diferencian en la forma de modelar dicha señal de excitación.
El sonido de la voz, que consiste originalmente en una onda de presión acústica del aire, se convierte en el micrófono en una magnitud eléctrica variante en el tiempo: la señal de voz. Esta señal se digitaliza, tomando una muestra cada cierto intervalo de tiempo y expresándola con una cadena de bits. La velocidad de muestreo suele ser de 8000 muestras por segundo en aplicaciones de telefonía, pues es suficiente para conservar la calidad de la voz telefónica, pero puede ascender hasta las 16 000 muestras por segundo si se quiere preservar toda la calidad acústica de la voz.
Si se quiere una alta calidad del habla a costa de enviar un número elevado de bits por segundo, la señal excitación que se usa para reconstruir la voz consiste en el error de la predicción codificado con un número de bits por muestra suficiente para no perder casi información; por ejemplo, el sistema de compresión-codificación denominado ADPCM (*Adaptive Differential Pulse Code Modulation*), trabaja a 32 000 bits por segundo. Si el objetivo es reducir al máximo la velocidad de transmisión (a menos de un bit por muestra de señal), la excitación se genera en el receptor a partir del código (encontrado en el emisor y transmitido) de una secuencia artificial que, introducida en el filtro, da una secuencia de salida que difiere un mínimo, en términos perceptivos, de la señal de voz original; por ejemplo, el sistema de compresión-codificación denominado CELP (*Code-Excited Linear Prediction*), puede trabajar a velocidades de 5 o 6 bits por segundo. Cuando se desea grabar voz de calidad resulta útil una técnica del primer tipo; en cambio, una técnica del segundo tipo se usa, por ejemplo, para transmisión de voz en telefonía móvil GSM.

Extracción de características

En el reconocimiento del habla, la señal de voz, una vez digitalizada, se procesa para producir una nueva representación de la voz en forma de secuencia de vectores o agrupaciones de unos valores que denominamos *parámetros* y que, como se ha dicho anteriormente, deben representar la información contenida en la envolvente del espectro. El número de parámetros debe ser reducido, puesto que la base de datos de entrenamiento siempre es limitada, por lo que cuantos más parámetros tenga la representación menos fiables son los valores entrenados y, por otro lado, más costoso es el proceso de reconocimiento.

El esquema de la figura 1 muestra las distintas etapas en las que podemos dividir el proceso utilizado de modo habitual en la obtención de los parámetros (parametrización), que se basa en la descomposición en bandas espectrales. El análisis se realiza tramo a tramo, es decir, desplazando de forma regular y a saltos la ventana a través de la cual se observa la señal.

En primer lugar, se aísla el tramo bajo análisis multiplicando la señal por una secuencia en forma de arco (ventana de Hamming) cuya longitud suele ser de 25 o 30 milisegundos y que se desplaza unos 10 milisegundos entre un tramo y el siguiente. A continuación, se calcula (estima) el espectro y se integra su valor en cada una de las bandas en que previamente se ha dividido el margen de frecuencias de acuerdo con la escala *mel*. Cuando se tienen 8000 muestras por segundo, este margen va de 0 a 4000 Hz, siendo suficiente para la voz típica de telefonía, pues ésta se extiende de 300 a 3400 Hz. Con esta integración se elimina prácticamente el efecto de las fluctuaciones armónicas que caen dentro de la banda y el resultado es una secuencia de energías (ordenadas en frecuencia) que sigue la forma de la envolvente espectral.

Para evaluar las probabilidades no se usan directamente estos parámetros de energía sino que se les somete antes a unas transformaciones para adaptarlos a las hipótesis que hacen los modelos estadísticos de las unidades del habla. En primer lugar, se comprime la amplitud de las energías calculando su logaritmo a fin de obtener valores con distribuciones estadísticas más parecidas a las gaussianas que suponen los modelos usados en el reconocimiento. Otra razón de peso para usar logaritmos es el carácter aditivo que pasan a tener las distorsiones que, introducidas por el micrófono o la línea telefónica, son multiplicativas en el dominio lineal. Al presentarse en forma de suma se pueden eliminar de manera mucho más fácil.

Las energías logarítmicas de bandas cercanas están correlacionadas con fuerza, mientras que los modelos de Markov ocultos típicos suponen que los parámetros no guardan correlación estadística entre ellos. Para eliminar la correlación se suele llevar a cabo la transformación coseno discreta que, aplicada al conjunto de energías de un tramo (unas 24) entrega los denominados coeficientes cepstrales, de los que sólo hace falta conservar los primeros (unos 12), lo cual representa una ventaja adicional. Se han propuesto unos parámetros alternativos a dichos coeficientes, que tampoco están correlacionados y, a diferencia de ellos, tienen la ventaja de ubicarse en frecuencia. Estos parámetros, unos 12 de forma típica, se obtienen de una manera muy simple, restando cada dos energías logarítmicas no consecutivas (la 1ª y la 3ª, la 2ª y la 4ª, etc.), por lo que son medidas de pendiente espectral, característica que ha mostrado ser un buen correlato perceptivo.

La última operación indicada en la figura consiste en un filtrado lineal de las secuencias temporales de parámetros espectrales, el resultado del cual es un nuevo vector de parámetros por cada tramo que se suministra al algoritmo de reconocimiento junto con el vector original. El filtro se elige de tal manera que este segundo vector sea una derivada temporal, es decir, una medida de la pendiente de la trayectoria que siguen los parámetros espectrales. A veces, también se usa la segunda derivada, además de la primera, para conseguir así una descripción todavía más completa del cambio que sufren dichos parámetros a lo largo del tiempo. Además, en el conjunto de parámetros finales se suelen incorporar las derivadas temporales de la energía del tramo de señal.

De esta forma, aunque el número de parámetros y, en consecuencia, la complejidad de cálculo, se incrementan de forma sustancial (hasta duplicarse o triplicarse), el porcentaje de acierto del reconocimiento mejora significativamente. La razón de ello estriba de nuevo en una limitación de los modelos de Markov ocultos, pues suponen que los vectores de observaciones (de parámetros) de dos tramos distintos son independientes, algo que no se corresponde con la realidad ya que, con la corta distancia (10 milisegundos) que hay entre tramo y tramo, no se aprecia en general un cambio substancial en la señal de voz al pasar de un tramo al siguiente.

Reconocimiento robusto en condiciones adversas

Las representaciones actuales de la voz, aunque poco eficientes debido a que conllevan mucha redundancia, permiten conseguir unas buenas prestaciones del reconocimiento siempre que la señal de voz se registre en condiciones favorables. Sin embargo, cuando un sistema de reconocimiento se pone a funcionar en situaciones reales se encuentra con condiciones adversas tales como cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, entre otras) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (como ruidos o distorsiones de la señal provocados por el micrófono o el canal de transmisión), que son irrelevantes desde el punto lingüístico pero que pueden degradar en gran medida la tasa de reconocimiento.
Si la variabilidad de las características de la voz aportada por las distintas condiciones ambientales posibles fuera recogida por completo en la base de datos de entrenamiento del sistema podríamos esperar que el resultado del reconocimiento no se degradara mucho. Pero esto sólo resulta útil cuando las condiciones ambientales en las que debe trabajar el sistema son muy específicas, como en el caso de reconocimiento del habla dentro del automóvil (recientemente, se han producido bases de datos en entorno de coche para varias lenguas europeas en el proyecto SPEECHDAT-CAR). Y aún así es imposible recoger todas las alteraciones que se pueden encontrar en un entorno específico, ya sea porque se desconocen, por ser demasiado numerosas o por variar con el tiempo.
Llevar el micrófono colgando, o tener que mantener la posición de la cabeza frente al micrófono de sobremesa, son condiciones incómodas pero necesarias actualmente para que la voz captada por el micrófono sea lo bastante limpia, en especial cuando existe un ruido ambiental molesto. La situación deseable es que el micrófono se encuentre a cierta distancia del o de los parlantes y éstos puedan moverse con libertad (*hands-free*). En realidad, resultaría conveniente que hubiera varios micrófonos para captar señales distintas y luego procesarlas en conjunto y compararlas. De hecho, el sistema auditivo humano dispone de dos entradas de voz y dicha binauralidad le permite separar fuentes de sonido situadas en puntos distintos.
Se pueden distinguir diferentes clases de ruidos. Los más benignos para el reconocimiento del habla son los estacionarios, es decir, los que mantienen sus características estadísticas a lo largo del tiempo. Puesto que en los intervalos sin voz (silencios) aparecen aislados, se pueden determinar en ellos sus parámetros espectrales y así, teniendo los ruidos caracterizados, resulta mucho más fácil eliminarlos de los intervalos donde reside la voz. De ahí la importancia de disponer de detectores fiables de actividad oral que permitan separar los intervalos temporales de voz y de silencio.
Los ruidos más difíciles de captar y eliminar suelen ser los de tipo impulsivo, tales como golpes de puerta, pitidos cortos, tos y, sobre todo, la voz de otra persona que se encuentre cerca; ésta es la situación más perjudicial para el reconocimiento, puesto que un mismo segmento de señal contiene las dos voces y resulta muy difícil separarlas para pasar a reconocer la que interesa.
En cualquier caso, pero en especial cuando la base de aprendizaje no recoge las degradaciones de la voz, hay que recurrir a técnicas de reconocimiento robusto, todavía en fase de desarrollo en los laboratorios, que atacan el problema de varias formas:

1. Obtención de una señal más limpia (*speech enhancement*). En este apartado se hallan las agrupaciones (*arrays*) de varios micrófonos, que actúan como una antena orientable hacia la fuente de sonido deseada gracias el tratamiento de las señales recogidas en toda la agrupación. Si se puede suponer que la señal de voz y la de ruido son *aditivas y no correlativas* (algo que parece realista), el espectro de la señal ruidosa es la suma de los espectros de voz y ruido, y se pueden aplicar varias técnicas de cancelación de ruido. Por ejemplo, la señal se puede procesar con un filtro de Wiener para reducir la presencia de ruido mezclado con la voz; el filtro se entrena durante los silencios para que aprenda la estadística del ruido. Una forma alternativa de eliminar (nunca totalmente) el ruido de la señal es la sustracción espectral, que estima el espectro del ruido en los silencios y luego lo sustrae del espectro de señal de voz ruidosa.
2. Determinación de parámetros más robustos. Es un hecho bien conocido que el sistema auditivo humano es más robusto que cualquier sistema automático no sólo frente al ruido aditivo y las distorsiones en general, sino frente a cualquier factor de variabilidad de la voz,

incluidos los cambios de articulación cuando el hablante está inmerso en un ruido intenso (en una discoteca, por ejemplo). Por tanto, sería de esperar que un sistema de reconocimiento del habla fuera más robusto a todos estos factores si la representación de la señal de voz siguiera de cerca las características perceptivas del sistema auditivo humano. Pero los intentos realizados hasta el presente no se han visto muy favorecidos por el éxito. Por otro lado, cada una de las etapas de parametrización expresadas en la figura 1 es susceptible de modificaciones que mejoren la robustez del sistema. Por ejemplo, la transformación logarítmica no es la más adecuada en los valles del espectro, ya que éstos se contaminan con más facilidad por el ruido aditivo que los picos y, como el logaritmo comprime menos las amplitudes bajas que las altas, una pequeña diferencia en un valle del espectro se convertirá en una desviación relativamente grande en su logaritmo. Existen otras funciones de compresión, como la raíz *n*-ésima, que mitigan de forma apreciable este problema. El filtrado temporal expuesto en la misma figura se utiliza también para mejorar la robustez del sistema. Cuando las señales se distorsionan de forma lineal por el micrófono, el canal telefónico, etc, y dicha distorsión no es idéntica para todas ellas, la tasa de error aumenta de modo sustancial. Puesto que la distorsión únicamente contribuye con un término aditivo en los parámetros logarítmicos, se puede cancelar su influencia eliminando con un filtro la componente continua (frecuencia cero) de las secuencias temporales de parámetros.

3. Compensación de los parámetros distorsionados y adaptación de los modelos a las nuevas condiciones del entorno. En estas técnicas, los parámetros son procesados a fin de que se asemejen estadísticamente a los que se hubieran obtenido con las condiciones ambientales de entrenamiento. Una alternativa es adaptar a las nuevas condiciones los modelos estadísticos de las unidades fonéticas desarrollados para habla limpia; a menudo, estas técnicas usan para ello conocimiento del ruido o la distorsión. Puesto que estas técnicas se basan en optimizaciones estadísticas, se apartan del objetivo de este artículo.

Conclusiones

En experimentos comparativos con palabras sin sentido se observa que la capacidad de percepción auditiva humana es superior a las prestaciones actuales de los sistemas de reconocimiento, especialmente en entornos acústicos adversos. Esta constatación sugiere que la extracción de características que se realiza todavía no está suficientemente conseguida, e impulsa el desarrollo de representaciones de la voz más efectivas. Es posible que un mejor conocimiento del funcionamiento del oído humano pueda ayudar en este sentido, pero es probable que se obtengan progresos más remarcables cuando la representación de la voz pueda ligarse a los movimientos de los órganos articulatorios, pues ellos son su causa y, además, pueden describirse con unos pocos parámetros. Quizás llegue un día en el que las elevadas prestaciones de un sistema de reconocimiento del habla necesiten la extracción de todas las características que transporta la señal de voz, tal como ocurre ahora en la compresión de la voz para su transmisión. De momento, uno de los objetivos principales que se persigue es aumentar la robustez de la representación frente a las condiciones adversas. De todas formas, no hace falta esperar que los problemas de robustez se resuelvan completamente en la parametrización, ya que son abordables también, y de forma complementaria, desde otras partes del sistema de reconocimiento y comprensión del habla.

Bibliografía general

DELLER, J.R.; HANSEN, J.H.L.; PROAKIS, J.G.: *Discrete-time Processing of Speech Signals*, IEEE Press, 1993.
HUANG, X.D.; ACERO, A.; HON, H.W.: *Spoken Language Processing*, Prentice-Hall PTR, 2001.
JUNQUA, J.C.; HATON, J.P.: *Robustness in Automatic Speech Recognition*, Kluwer Acad. Publ., 1996.
NADEU, C.; MACHO, D.; HERNANDO, J: «Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition», *Speech Communication* (Special Issue on Noise Robust ASR) 2001; 34 (April): 93-114.
O'SHAUGHNESSY, D.: *Speech Communication, Human and Machine*, Addison-Wesley Series in Electrical Engineering, Digital Signal Processing, 1990.
PICONE, J.W.: «Signal Modeling Techniques in Speech Recognition», *Proceedings of the IEEE* 1993; 81 (Sept.): 1215-1247.
RABINER, L.R.; JUANG, B.H.: *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
RABINER, R.L.; SCHAFER, R.W.: *Digital Processing of Speech Signals*, Englewood Cliffs, Prentice-Hall, 1978.

Climent Nadeu i Camprubí

Ingeniero de telecomunicación y doctor por la Universitat Politècnica de Catalunya (UPC). Profesor de la UPC desde 1977, actualmente es catedrático del Departament de Teoria del Senyal i Comunicacions, e imparte materias de procesado de señal en la ETSETB. Ha sido investigador visitante en AT&T Bell Labs y en el ICSI de Berkeley. Su investigación se ha llevado a cabo sobre todo en el campo del tratamiento del habla, teniendo publicados más de un centenar de artículos en libros, revistas y actas de congresos. Actualmente es director del Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) de la UPC.
climent@talp.upc.es

Figura 1 Esquema típico de determinación de los parámetros representativos de la envolvente espectral de cada tramo de señal de voz