ÍNDICE

Prefacio

El reconocimiento de voz, computacionalmente hablando, es un tema en continuo auge gracias al avance de los dispositivos que incorporan sistemas automáticos con diferentes propósitos. Robots y sistemas inteligentes, telefonía móvil, traductores e intérpretes automáticos o el más reciente paradigma conocido como Internet de las Cosas (*Internet of Things*, *IOT*) son buenos ejemplos de ello.

La base en los reconocedores automáticos de voz son las señales acústicas, que representan unidades del habla (palabras, sílabas, fonemas). Dichas señales son en realidad secuencias de valores discretos a lo largo del tiempo. Bajo esta perspectiva, el libro aborda dos aspectos clave. En primer lugar, establece las bases sobre el análisis de señales acústicas, introduciendo los conceptos y técnicas necesarias para su tratamiento. En segundo lugar, aplica tales conceptos y técnicas para concretar los fundamentos y llegar al desarrollo de las técnicas de reconocimiento propuestas. Se obtienen así las propiedades que caracterizan una unidad de voz con fines de su reconocimiento. Estas propiedades son a veces suficientes. No obstante, en cualquier caso constituyen la entrada para los reconocedores más avanzados basados en técnicas de aprendizaje, que quedan fuera del ámbito del presente texto, dado que este aspecto es ampliamente abordado en la literatura especializada. No así, la conjunción de conceptos y técnicas sobre el tratamiento de señales y la extracción de características. Es justamente aquí donde radica el valor añadido de este libro, que proporciona, con la claridad y ejemplos ilustrativos suficientes, los conceptos esenciales en sendos ámbitos. Sin duda, esta es la razón fundamental por la que se publica la obra, siendo conscientes de la necesidad de un texto con las características reunidas por este.

Esa unión del tratamiento de señales y la extracción de características hacen que este texto sea autosuficiente, de suerte que el lector, incluso sin ser experto en los conceptos relativos al tratamiento de señales, puede abordar sin dificultad los contenidos del libro, para llegar al desarrollo de sus propios reconocedores. Sin duda, este hecho otorga un valor añadido de interés a la obra.

En definitiva, desarrolladores, ingenieros, investigadores o estudiantes universitarios encuentran en el libro una referencia de base de suma utilidad para abordar los aspectos conceptuales y de implementación en el desarrollo de reconocedores automáticos de voz, particularmente para quienes se inicien en la materia por su carácter didáctico y autocontenido.

Sobre el autor

Gonzalo Pajares Martinsanz es profesor en la Facultad de Informática en la Universidad Complutense de Madrid en el Departamento de Ingeniería del Software e Inteligencia Artificial. Ha desarrollado una extensa actividad profesional durante más de una década en la industria con aplicación de tecnologías software y de Inteligencia Artificial. Además, está ampliamente involucrado en tareas de investigación en el ámbito de la Inteligencia Artificial donde ha sido y es director de proyectos nacionales e internacionales de investigación, durante más de dos décadas, con transferencia tecnológica a la industria. Es autor y editor de varios libros sobre visión por computador, inteligencia artificial, tecnologías sensoriales, así como autor de numerosas publicaciones en dichas áreas, incluyendo técnicas de reconocimiento de patrones y estructuras en diversos ámbitos industriales, con participación en empresas situadas en la vanguardia tecnológica. Ha publicado numerosos artículos en revistas especializadas de prestigio internacional, a la vez que es editor invitado y asociado en varias revistas con alto índice de impacto, así como editor jefe de la revista *Journal of Imaging* recientemente creada.

Agradecimientos

Es de agradecer a todas las personas que han tenido que soportar la sustracción de tiempo debido a la dedicación al libro. Gracias, Alicia y Belén, por vuestra comprensión. Gracias también al resto de la familia por lo mismo, aunque haya sido a distancia. Por supuesto, en el recuerdo a quienes siempre disfrutaron con el esfuerzo de sus allegados.

Gracias también a las industrias y al departamento al que pertenezco en la facultad, por haberme brindado la oportunidad de poder aplicar tecnologías informáticas de vanguardia en diversos ámbitos.

Finalmente, un especial agradecimiento a la editorial RC Libros por su soporte para la publicación y difusión de la obra. Gracias, José Luis, por tu apoyo continuo en este sentido a lo largo de los muchos años en los que hemos colaborado y especialmente ahora ante el nuevo reto que tienes por delante en la editorial.

1 RECONOCIMIENTO COMPUTACIONAL DE VOZ

INTRODUCCIÓN

El reconocimiento de voz, y en su sentido más amplio el reconocimiento de sonidos, se enmarca dentro de la capacidad perceptual relacionada con el sistema del oído como receptor natural de los sonidos, así como de la emisión de los mismos mediante el correspondiente aparato fonador, que incluye las cavidades bucal y nasal, el sistema pulmonar y los conductos de comunicación entre ellos y con los pulmones. En el caso humano se refiere al mecanismo del habla, mientras que en el caso animal está relacionado con la emisión de sonidos propios, a veces sin la sofisticación de los sistemas de aquellos. A lo largo del texto se utilizarán indistintamente los términos reconocimiento de voz y del habla.

Desde el punto de vista computacional, que es el que nos ocupa, los sistemas de reconocimiento de voz se fundamentan en el análisis de las señales generadas o recibidas a través de los sistemas físicos de emisión o recepción con el fin de conseguir su interpretación. Se excluye, por tanto, lo relativo a su transmisión en el medio, que corresponde a otros campos de la ciencia en el ámbito de las telecomunicaciones.

Hoy día los numerosos dispositivos existentes en el mercado y aquellos otros que sin duda aparecerán como consecuencia de la evolución tecnológica, muchos de ellos dentro del nuevo paradigma en expansión actual como es el Internet de las Cosas, están equipados con sistemas de captura y reconocimiento de la voz, con desarrollos basados en interfaces de usuario (Pearl, 2016). Fijémonos en la función de grabación

de uno de tales dispositivos. El usuario activa el micrófono de forma que comienza la grabación. Si el hablante no emite sonido, aunque haya ruido en el ambiente, el procedimiento de tratamiento de la señal considera que no se ha emitido un mensaje razonable como para ser interpretado. Si por el contrario, el hablante expresa un sonido, independientemente de que luego sea inteligible o no, el método de tratamiento de la señal del sonido traduce, en el idioma seleccionado, la interpretación de dicho sonido, que deja escrito en pantalla para su posterior uso, bien como mensaje de envío, bien para su almacenamiento. Cuando se deja de hablar sobre el micrófono, el procedimiento continúa la grabación para determinar que se ha entrado en una fase de captación de ruido ambiente, procediendo en este caso a eliminar la parte de la señal considerada como sobrante.

Otro ámbito donde el tratamiento de las señales de sonido cobra especial relevancia es en los futuros y no tan futuros vehículos inteligentes, los cuales están y estarán equipados con diversos sistemas que requerirán en determinados momentos la intervención del conductor, por muy sofisticados que aquellos sean. La conducción requiere alta concentración del conductor, por lo que la manipulación física de los dispositivos es ciertamente no recomendable, cuando no prohibida y sancionada, por motivos obvios. Es aquí donde los sistemas de reconocimiento de voz están llamados a jugar un importante papel de cara al uso razonable de los dispositivos que necesiten intervención humana, minimizando así el riesgo de potenciales accidentes con el consiguiente incremento de la seguridad vial.

Por tanto, volviendo al tema sobre el origen y manifestación de los sonidos desde el punto de vista computacional, se trata de señales unidimensionales que evolucionan en el tiempo. Esto significa que la mayoría, por no decir todos los procesos y técnicas contenidos en el presente libro, son aplicables de forma general a cualquier tipo de señal temporal unidimensional, dentro de las cuales se incluyen procedimientos propios de lo que se conoce como tratamiento de señales. Esta circunstancia justifica el hecho de que varios capítulos del libro se dediquen de forma genérica al tratamiento de señales en su sentido más amplio, dentro de las cuales se encuentra precisamente el sonido.

Ciñéndonos al reconocimiento de voz, comencemos por el análisis de los elementos y estructuras que la conforman. Es evidente que las frases están constituidas por palabras, estas por sílabas hasta llegar a los fonemas. Durante el proceso del habla, el hablante emite los sonidos correspondientes a los diferentes fonemas y estos se encadenan produciendo la señal apropiada. En la figura 1.1 se muestra un ejemplo de la señal unidimensional obtenida correspondiente a la expresión “percepción computacional”. En este caso se ha grabado a una frecuencia Fs de 11025 Hz, habiendo obtenido $6 \times Fs$ (66150) muestras de la señal durante un

intervalo de tiempo de 6 s. En ciertos dispositivos computacionales son habituales las siguientes frecuencias de muestreo: 8000, 11050, 22050 y 44100 Hz. En el ejemplo mostrado en dicha figura, aparecen dos secuencias que se corresponden con las dos palabras que componen la expresión representada. En la primera de ellas se distinguen claramente las tres sílabas que conforman la palabra “per-cep-ción”, en esta grabación se ha puesto especial énfasis en el tono de voz para las tres sílabas y más específicamente en relación con la primera. En la segunda secuencia aparecen cuatro tramos correspondientes a las subsecuencias “com-pu-ta-cio-nal”. Tal y como se puede apreciar fácilmente, unas sílabas aparecen más enfatizadas que otras. Este fenómeno depende exclusivamente del hablante, por lo que puede deducirse que según sean las características de este así será la representación de la señal y sus propiedades asociadas. Debido a esta circunstancia, es fácil deducir la dificultad que puede entrañar el proceso de reconocimiento del habla, ya que un factor importante depende de la naturaleza de la fuente que ha generado la secuencia de sonidos. Así, la misma expresión pronunciada por distintos hablantes puede originar diferentes señales, que si bien en esencia pueden contener las mismas unidades para su análisis, su forma, constitución y encadenamiento pueden llegar a diferenciarse considerablemente. Además de lo anterior, conviene señalar que en reconocimiento de voz resulta de gran interés conocer en todo momento detalles tales como dónde comienza y finaliza con exactitud el contenido del mensaje dentro de la señal de voz, identificar zonas de silencio, presencia de ruido y otras características relevantes que afectan al proceso de reconocimiento.

Es el momento de pensar sobre el hecho de que cuando uno se enfrenta a reconocedores automáticos de voz, a veces se llega a la desesperación más absoluta, debido a que el sistema no es capaz de reconocer una palabra o simple expresión que el propio sistema está demandando y esperando dentro de un limitado conjunto de expresiones y posibilidades que él mismo a veces indica. En este sentido, los métodos computacionales destinados al reconocimiento de voz desempeñan un papel esencial dentro del proceso de percepción. Bien es cierto que los sistemas de reconocimiento cada día están más perfeccionados, como es el caso de los dispositivos de telefonía móvil o intérpretes y traductores automáticos, como lo demuestra el hecho del alto grado de aciertos conseguidos.

Los diferentes procedimientos de análisis se centran en la comparación de las características de la señal con una serie de patrones previamente almacenados en el computador y convenientemente identificados. Supongamos que antes de comenzar el análisis de la señal mostrada en la figura 1-1 se han almacenado los patrones correspondientes, que en este caso representan cada uno de los sonidos asociados con las sílabas que se pretende identificar. En la figura 1-2 se muestran dichos patrones separados, que corresponden a las sílabas encadenadas de la figura 1-1.

El objetivo para el análisis computacional consiste en extraer de los patrones determinadas características con el fin de que las mismas sean posteriormente comparables con las que aparecen en la señal a analizar. En las siguientes secciones se aborda esta cuestión con el detalle que corresponde.

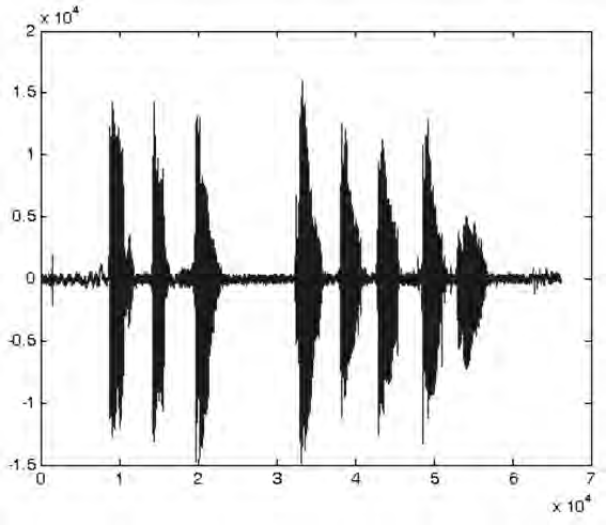

Fig. 1-1 Señal de voz correspondiente a la expresión: percepción computacional

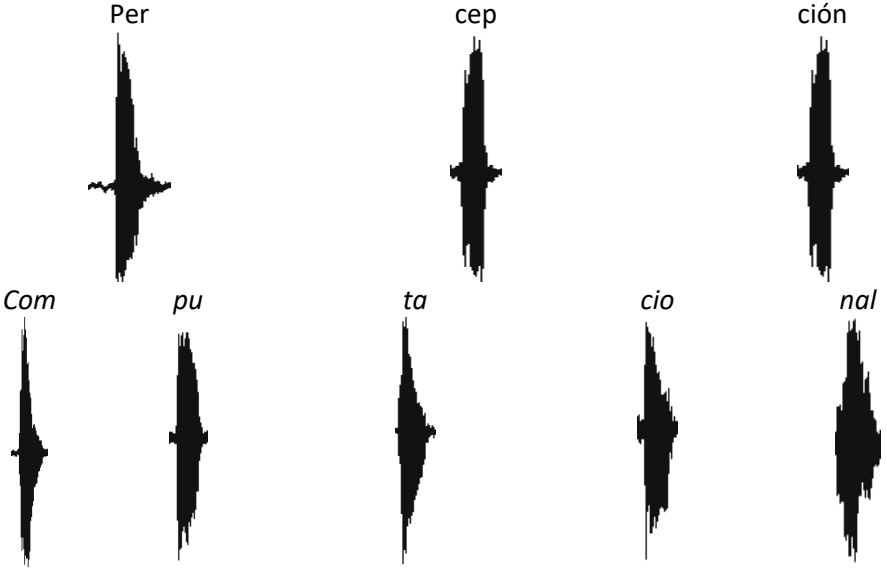

Fig. 1-2 Patrones de análisis que conforman las sílabas correspondientes a la expresión: per-cep-ción com-pu-ta-cio-nal

ORGANIZACIÓN DEL LIBRO

Como se ha expresado previamente, desde el punto de vista computacional el fundamento del análisis de voz, y por tanto su reconocimiento, está basado en una primera etapa en la generación y tratamiento de señales temporales digitales. El objetivo principal es extraer la suficiente información a partir de dichas señales para identificar su contenido. Esta información se estructura en forma de características interpretables desde el punto de vista del análisis computacional, que permiten determinar el contenido de la señal desde el punto de vista de su reconocimiento. Dichas características a veces se utilizan directamente como discriminantes y en ocasiones sirven como entrada para los sistemas de reconocimiento basados en técnicas propias de lo que se conoce como reconocimiento de patrones y de forma más amplia, aprendizaje automático, donde se encuadran técnicas de naturaleza estadística o redes neuronales por citar solo dos ejemplos. Este último aspecto queda excluido del ámbito de este libro, centrándose por tanto, en el procesamiento de la señal para su preparación con fines de una posterior extracción de características para su reconocimiento. Algunas de estas técnicas, como las basadas en programación dinámica descritas en el capítulo ocho, resultan ser computacionalmente costosas, si bien este efecto adverso puede solventarse mediante implementaciones hardware hoy día suficientemente desarrolladas.

Respecto a las técnicas mencionadas sobre reconocimiento de voz mediante técnicas de aprendizaje automático destacan aquellas basadas en clasificadores, tales como *Bayes*, *K-vecinos más cercanos*, *perceptrón*, *árboles de decisión* o *máquinas vectores soporte* o las más recientes que incluyen aprendizaje profundo, por mencionar solo algunas. Existen otras técnicas de naturaleza estocástica tales como los modelos ocultos de Markov (*Hidden Markov Models*, *HMM*). Un *HMM* se caracteriza por presentar una arquitectura basada en nodos con sus correspondientes estados y probabilidades de transición entre estados. Se comienza por definir las unidades de voz a reconocer (palabra, silaba, fonema, etc.), de suerte que a cada unidad se le asocia un *HMM* definido por una matriz de probabilidades de transición entre estados que lo forman, junto con las correspondientes probabilidades de salida. Es habitual modelar las funciones de densidad de probabilidad mediante distribuciones Gaussianas mixtas (*Gaussian mixture model*, *GMM*). Se trata de los correspondientes parámetros que caracterizan tales funciones para determinar la probabilidad de salida. En Solera-Ureña y col. (2012), Li y col. (2016) o Giannakopoulos y Pikrakis (2014) se pueden encontrar detalles de algunas de estas técnicas con sus múltiples referencias asociadas. En cualquier caso, tal y como se ha mencionado previamente, estas técnicas necesitan como entradas vectores de características que definen las señales acústicas y cuya extracción constituye uno de los objetivos del presente libro.

Por otra parte, existen infinidad de programas y aplicaciones que abordan los temas relacionados con el reconocimiento de voz, destacando el programa Praat de Boersma y Weenink (2016). Son también numerosos los programas y aplicaciones al respecto desarrollados bajo el lenguaje Matlab (2016), que además es con el que se han generado las gráficas y los resultados que aparecen en este libro.

El libro consta de ocho capítulos, de suerte que al tratamiento de señales se dedican los capítulos dos a cinco, mientras que los capítulos seis a ocho se centran en la extracción de características.

Así y de forma más concreta y en lo que respecta al tratamiento de las señales, el capítulo dos aborda el tema de los sistemas discretos, dado que a pesar de que el tiempo es continuo en esencia, en el mundo digital solo es posible el tratamiento de muestras obtenidas en determinados instantes de tiempo. El capítulo tres trata el dominio de la frecuencia, como complemento al dominio temporal, de suerte que las señales se transforman a dicho dominio para su tratamiento. El capítulo cuatro describe brevemente el tema relacionado con el filtrado de las señales con un enfoque predominante en el dominio de la frecuencia, cuyo objetivo es la preparación de la señal para un mejor análisis posterior. Existen otros dominios transformados de gran utilidad, tales como wavelets, transformada del coseno o transformada *z*, que se analizan en el capítulo cinco.

Desde el punto de vista del reconocimiento e identificación del contenido, el capítulo seis aborda la segmentación de la señal, es decir, su tratamiento basado en el análisis temporal. Mientras que el capítulo siete segmenta la señal, si bien desde su procesamiento y consideración en el dominio de la frecuencia. Finalmente, también en el ámbito de la segmentación, en el capítulo ocho se aplican otras técnicas específicas dentro de los dominios señalados previamente y descritos básicamente en el capítulo cinco.