

y altavoces). Lo incluimos aquí pues por su interés como técnica avanzada y como ejemplo de utilización de estos periféricos que se estudian en el apartado 2.4.1.

2.2.2.3.1. Reconocimiento de voz

2.2.2.3.1.1. Concepto y características

En la comunicación hablada hay que diferenciar el proceso de generar lenguaje hablado a través del ordenador, denominado síntesis de habla, del proceso de comprender el lenguaje natural mediante ordenador o reconocimiento del habla. Aunque ambos procesos requieren de complejas técnicas, son totalmente diferentes. Reconocer el lenguaje natural es mucho más difícil que generarlo a partir de un texto, requiriendo el proceso de reconocimiento de mayor potencia de cálculo.

El reconocimiento de voz es el proceso por el cual un ordenador es capaz de interpretar palabras o frases que le son transmitidas por un ser humano de forma hablada. A partir de esta interpretación el ordenador puede transcribir la información en un fichero de texto o utilizarla para el control de procesos (por ejemplo, de su propio funcionamiento).

El proceso por el cual el ordenador interpreta la voz humana no es un proceso simple. En el habla pueden influir muy diversos parámetros como las características morfológicas del hablante (género, edad), posibles trastornos orgánicos (o simples diferencias físicas no clasificables como trastorno), características culturales del hablante (dialectos, acentos, modismos, giros), características del habla (precisión, velocidad de articulación, apertura vocal), etc. Además, el hecho del habla no se produce normalmente en ambientes silenciosos, sino todo lo contrario, con lo que el ruido ambiente dificulta aún más el proceso de reconocimiento.

Si difícil es el reconocimiento de palabras aisladas, más lo será cuando éstas se articulan en frases completas. Al hacerlo así se concatenan formando conjuntos de sonidos que suele ser difícil separar. Otros problemas adicionales se dan por factores inherentes al lenguaje como la existencia de palabras homófonas (aya, haya y halla, vaca y baca³, herrar y errar, etc.), pero todos ellos van siendo resueltos por los sistemas actuales de reconocimiento de voz que ya han dejado de ser modelos experimentales para entrar en una fase de aprovechamiento comercial.

Un sistema de reconocimiento ideal debería contar con las características siguientes:

- Reconocer el conjunto más amplio posible del idioma del hablante: es preciso que el sistema disponga de diccionarios extensos del idioma utilizado, con el fin de que no se produzcan fallos debidos al desconocimiento de vocablos. En este sentido es muy útil que los diccionarios puedan ser ampliados por el propio usuario final.
- Reconocer habla continua, también conocida como habla natural: debe adaptarse a las situaciones que se dan en la vida real, y no limitarse a casos "de laboratorio" en los que se ha pedido al orador que siga unas determinadas pautas en su discurso para adaptarse a las carencias del sistema (como hablar pausadamente – reconocimiento discreto– o marcar los fonemas). En el habla continua la generación de cada fonema se ve afectada por la generación de los fonemas adyacentes y, de modo parecido, el comienzo y final de las palabras se ven afectados por los vocablos que las preceden y suceden.
- Ser independiente del sujeto que habla: cualquier hablante medio (otro caso sería, por ejemplo, un sistema específico para discapacitados) debe poder dirigirse al sistema y que este transcriba fielmente la información contenida en su discurso hablado.

Estas tres características implican una gran capacidad de cómputo del sistema. De todos modos, aunque para alcanzar la extraordinaria capacidad de los seres humanos en el procesamiento del lenguaje queda aún un camino que recorrer, los sistemas de reconocimiento de voz actuales son ya perfectamente capaces de interpretar correctamente la mayor parte del discurso de un hablante.

El reconocimiento de voz se compone, al menos, de los procesos siguientes:

- Digitalización de la señal acústica del habla: para ello el ordenador precisará una tarjeta de sonido (ver 2.4.1.1) que permita convertir la señal analógica del discurso en señales digitales interpretables. Este tipo de tarjetas son muy comunes en el sector de ocio de la informática y su uso se ha extendido, siendo ya un periférico usual en cualquier PC.

³ Baca y vaca no son exactamente homófonas pero muchos hablantes las pronuncian igual.

- Transformación de la señal: tras el proceso de digitalización, el sistema deberá traducir la información obtenida a un formato que le permita distinguir palabras. Para ello se utilizan diversas técnicas como que tienen como base el análisis espectral mediante transformadas de Fourier. De este modo se obtienen una serie de vectores que constituyen la información que se utilizará en el proceso de reconocimiento.
- Reconocimiento de fonemas, grupos de fonemas y palabras. Este paso puede ejecutarse de varias formas, aunque las técnicas más utilizadas son los modelos ocultos de Markov⁴ (HMM) y las redes neuronales (NN, Neural Networks) en combinación con sistemas expertos y otras técnicas de apoyo. Están en fase de desarrollo sistemas que mezclan estas dos técnicas.

Símbolo fonético	Ejemplo	Pronunciación
/b/	belfo	<i>b</i> elfo
/dZ/	lleno	<i>d</i> zeno
/i/	pise	<i>p</i> ise
/p/	paso	<i>p</i> aso

Figura 16 Algunos fonemas del español.

2.2.2.3.1.2. Proceso de reconocimiento de voz

Los elementos lógicos necesarios para que un sistema pueda desarrollar procesos de reconocimiento de voz son:

- Dominio: es el diccionario de palabras que el sistema puede reconocer. Los sistemas comerciales actuales superan, para idiomas como el castellano, las cien mil palabras.
- Inventario de pronunciación (fichero de referencia del usuario): base de datos en la que se almacenarán los elementos característicos de la pronunciación de los usuarios para los elementos del vocabulario. Su construcción requiere de una fase de aprendizaje (por ejemplo, en algunos productos comerciales se "enseña" al sistema como habla el usuario leyéndole pasajes de El Quijote de Miguel de Cervantes)
- Corpus: es el modelo lingüístico, esto es, la base de datos estadística y estocástica que ayuda al motor de reconocimiento a determinar las palabras que deben reconocerse. El modelo lingüístico de un idioma se diseña paralelamente al vocabulario, y se basa en textos leídos por una población de usuarios (locutores). Contiene pues información sobre la utilización de palabras y la estructura de las frases, lo que será muy útil para eliminar algunas de las dificultades del reconocimiento de voz.

Los procesos de reconocimiento de voz han ido evolucionando para incluir algoritmos cada vez más sofisticados que intentan simular el comportamiento del reconocimiento humano de códigos de información. El sistema más utilizado actualmente parte del reconocimiento de fonemas. Para ello se disgregan los vocablos en las unidades de información oral más pequeñas, los fonemas, con las que se construye la base de datos de reconocimiento. A partir de aquí la construcción de una palabra requerirá de un modelo lingüístico en el que se señalen las formas en que se relacionan los fonemas y las palabras básicas para el idioma que se interprete.

2.2.2.3.2. Identificación de locutores

Por identificación de hablantes se conoce al conjunto de técnicas biométricas encaminadas a la identificación/autenticación de una persona que habla. Usualmente, el reconocimiento de locutores obligaba al usuario que debía ser identificado a leer un texto fijo. En la actualidad existen ya productos que son independientes del texto que el usuario emplee, e incluso pueden trabajar con diferentes lenguajes.

La principal utilidad de esta técnica es la autenticación de personas empleando algo tan simple como su propia voz. Otras posibilidades pueden ser permitir que ciertos sistemas se autoconfiguren para un usuario determinado a partir de que lo reconocen por su voz. Por ejemplo, un automóvil podría reconocer a partir de su voz cuál de sus dos usuarios habituales está entrando en el habitáculo y proceder a colocar los reglajes de volante, asiento, climatizador, sintonía musical, etc. según las preferencias del mismo.

⁴ Actualmente los sistemas basados en HMM son los más frecuentemente utilizados dado su mayor porcentaje de éxito y su contrastada fiabilidad.