Introducción al habla

Helenca Duxans Barrobés
Marta Ruiz Costa-jussà

PID_00188069

UOC
Universitat Oberta de Catalunya

Los textos e imágenes publicados en esta obra están sujetos –excepto que se indique lo contrario– a una licencia de Reconocimiento-NoComercial-SinObraDerivada (BY-NC-ND) v.3.0 España de Creative Commons. Podéis copiarlos, distribuirlos y transmitirlos públicamente siempre que citéis el autor y la fuente (FUOC. Fundación para la Universitat Oberta de Catalunya), no hagáis de ellos un uso comercial y ni obra derivada. La licencia completa se puede consultar en http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.es

Índice

Introducción

El habla es el medio de comunicación más natural entre las personas; por lo tanto, la presencia que tiene en el mundo digital (ya sea en transmisiones o en interfaces hombre-máquina, en servicios de atención al cliente, en aplicaciones de ocio, etc.) es muy elevada y se prevé que todavía lo será más en el futuro.

Las características propias de la voz, diferentes de las del audio en general o de las de la música, permiten el diseño de algoritmos pensados específicamente para este dominio. Además, el habla no es solo audio, sino que también incluye aspectos de fonética, sintaxis y semántica que amplían la información que se puede utilizar en las tecnologías que están implicadas en ella. Por lo tanto, consideramos que es interesante dedicar un bloque específico al habla dentro de la asignatura *Procesamiento de audio*.

Objetivos

Este módulo pretende dar a conocer los conceptos básicos sobre la voz y el habla que son necesarios para entender los módulos de reconocimiento automático y de síntesis del habla de este tercer bloque. Concretamente, al acabar este módulo deberéis ser capaces de lo siguiente:

1. Explicar cuál es el proceso y los factores que intervienen en la producción de la voz.

2. Conocer las propiedades acústicas más relevantes de la voz, es decir, el tono, el timbre y la prosodia, y saber cómo se miden por medio de la frecuencia fundamental (o f_0) los formantes y la envolvente de la señal, y la evolución de la f_0, de la energía y de la duración de los sonidos.

3. Clasificar los sonidos según el tipo (consonante o vocálico), la sonoridad, el punto y el modo de articulación, la altura y la profundidad.

4. Distinguir los conceptos de trifonema y difonema y cómo les afecta el fenómeno de la coarticulación.

1. Introducción a las tecnologías del habla

Las tecnologías del habla incluyen todas las tecnologías cuyo elemento principal es el habla. Los dos ejemplos de tecnologías más habituales son el **reconocimiento automático del habla**, que consiste en transcribir el contenido de una señal de voz sin intervención humana, y la **síntesis del habla**, que se puede definir como todo proceso de creación de habla de manera artificial. Dedicaremos los dos módulos siguientes a explicar estas dos tecnologías con más detalle.

Otras tecnologías del habla clásicas que podemos destacar son las siguientes:

- **Reconocimiento automático del idioma hablado en una locución o discurso**. Esta tecnología analiza la voz de manera independiente del locutor y del mensaje transmitido para detectar el idioma utilizado. También se puede utilizar para identificar diferentes dialectos de una lengua. Las aplicaciones de reconocimiento del idioma se encuentran sobre todo en aplicaciones para sistemas de atención al usuario, con el fin de mejorar el servicio a partir de la detección del idioma utilizado por el usuario.

- **Reconocimiento de la edad y el sexo del locutor**. Los reconocedores de edad clasifican al locutor en tres o cuatro rangos de edad: niños, adultos jóvenes, adultos y gente mayor. Los reconocedores de sexo clasifican al locutor entre el sexo masculino y el femenino. En ambos casos, la identidad del locutor no es conocida ni se analiza. Estas tecnologías se pueden aplicar, además de a la mejora de servicios de atención al cliente, como paso previo del reconocimiento automático del habla para utilizar reconocedores adaptados específicamente a locutores masculinos o femeninos o de un rango de edad concreto.

- **Reconocimiento de locutor**. Esta tecnología consiste en detectar automáticamente la identidad de un locutor a partir de la voz. Dentro del reconocimiento del locutor se puede distinguir entre verificación e identificación.
  - La **verificación de locutor** consiste en dilucidar si una persona es quien dice ser a partir de la voz. Por lo tanto, el resultado de los verificadores es una decisión del tipo sí/no, que puede ir acompañada de un nivel de confianza del resultado. La aplicación más habitual de los verificadores se encuentra en sistemas de seguridad para accesos, tanto a medios físicos como digitales, normalmente en combinación con claves de paso u otros tipos de verificaciones biométricas (huella dactilar, iris, etc.).

  - La **identificación de locutor** consiste en determinar la identidad del locutor entre un conjunto de identidades conocidas por el sistema.

Por lo tanto, en este caso no se ha de comprobar si el locutor es quien dice ser, sino que se trata de encontrar cuál de las identidades que se conocen es más probable que le corresponda. Entre las aplicaciones de la identificación de locutor se encuentran los sistemas que se adaptan a los usuarios sin necesidad de que se identifiquen previamente. Otra aplicación es el análisis de grabaciones en las que se requiere encontrar automáticamente la identidad del locutor.

- **Diarización de locutor**. Esta tecnología consiste en segmentar, de manera automática, una grabación de una conversación en turnos de locutor y detectar todos los turnos que pertenecen al mismo locutor. Esta tecnología constituye el paso previo para aplicar las tecnologías anteriores cuando nos encontramos con grabaciones que contienen múltiples locutores. Una vez se ha diarizado la grabación, se puede reconocer la identidad de cada locutor o sus características de edad y sexo o detectar el idioma de cada turno.

Últimamente las tecnologías del habla se están uniendo con las tecnologías de la imagen, el vídeo y el texto para crear tecnologías multimodales.

2. La producción de la voz en tres pasos

El sistema humano de producción de sonidos se basa en tres acciones:

- La producción de un flujo de aire.
- La conversión del flujo de aire en sonido.
- La modificación del sonido anterior para producir todo el abanico posible de sonidos.

Físicamente, estas tres acciones tienen lugar en los pulmones, la laringe y el tracto vocal (podéis ver la figura 1):

Sistema humano de producción de sonidos

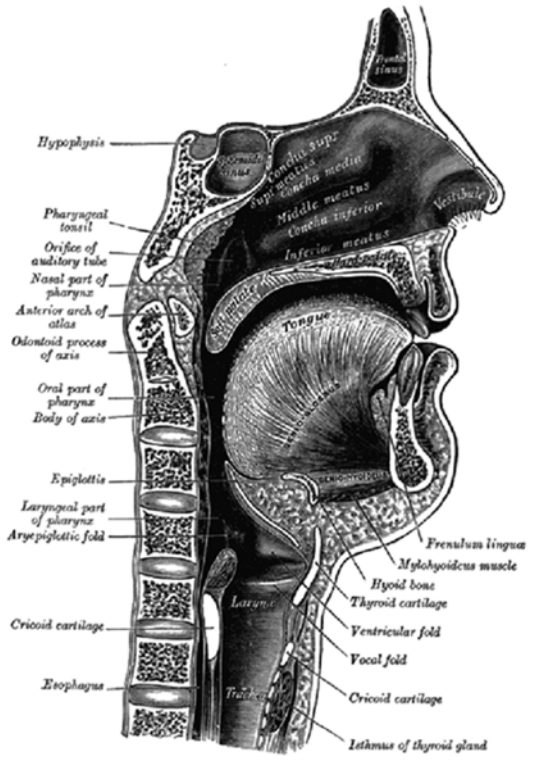

Figura 1. Sección sagital del sistema humano de producción de sonidos (*Gray's Anatomy*, edición de 1918)

La producción de un sonido empieza cuando los pulmones impulsan aire por la tráquea hacia la laringe. En la laringe se encuentran las cuerdas vocales, dos membranas que se interponen horizontalmente al paso del flujo de aire. Según cómo se sitúen estas membranas se producen dos tipos diferentes de sonoridad:

- Sonoridad sonora: las membranas se encuentran cerradas para obstruir el flujo de aire. La presión del aire hace vibrar las cuerdas vocales cuando atraviesa estas membranas.

- Sonoridad sorda: las membranas se encuentran separadas, de modo que dejan pasar el flujo de aire.

En la figura 2 se muestran representaciones temporales de señales de voz, correspondientes a un sonido sonoro y un sonido sordo. En la primera representación se ve cómo la vibración de las cuerdas vocales produce una periodicidad en la señal de voz. En cambio, esta periodicidad no está presente en las señales correspondientes a sonidos sordos:

Sonido sonoro y sonido sordo

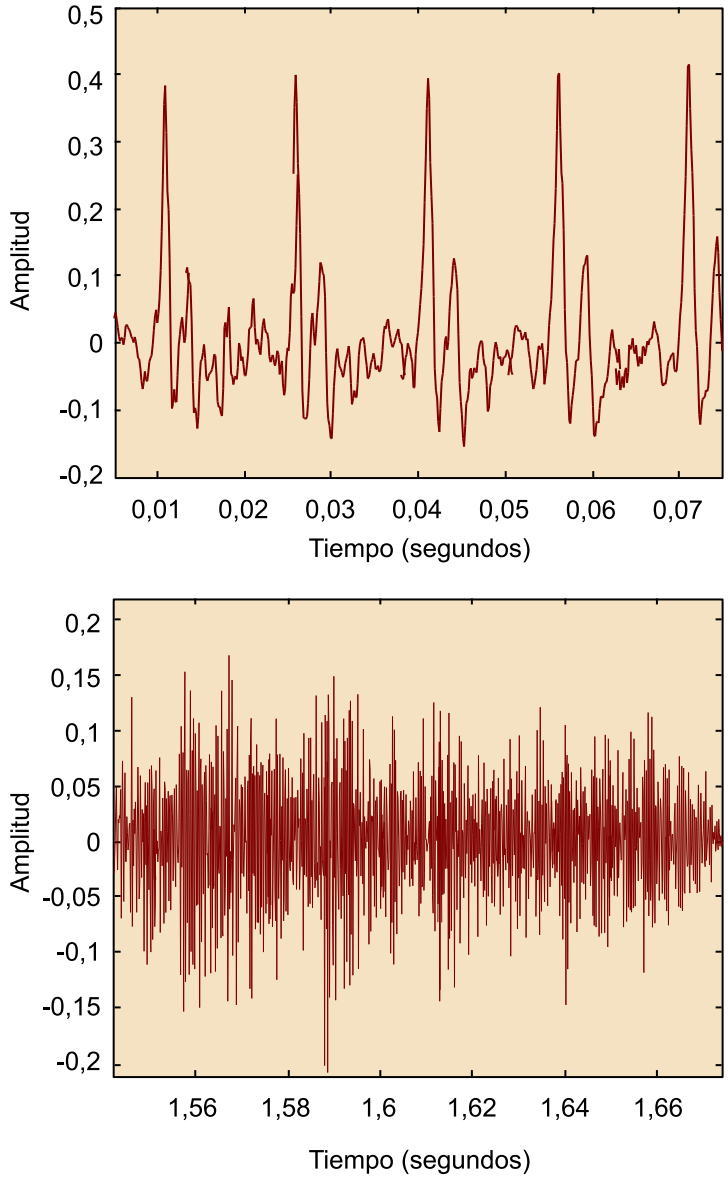

Figura 2. Representación temporal de dos segmentos de voz: un sonido sonoro (arriba) y un sonido sordo (abajo)

Finalmente, el flujo de aire llega al **tracto vocal**. El tracto vocal es el área que se encuentra entre las cuerdas vocales y los labios, incluyendo tanto la cavidad bucal como las cavidades nasales. La forma del tracto vocal, definida tanto por aspectos anatómicos como por la posición de los labios, las encías, los dientes y la lengua en cada momento, modifica el sonido producido por las cuerdas vocales y crea así todo el abanico de sonidos posibles.

3. Propiedades acústicas de la señal de voz

La voz está formada por un sonido o una secuencia de sonidos. Las propiedades de la señal de voz se dividen en dos grupos, según si se refieren a propiedades de la voz en un instante concreto (propiedades segmentales) o a la evolución de la voz en un período de más duración (propiedades suprasegmentales).

3.1. Propiedades segmentales

Altura tonal o *pitch*: es el nivel perceptivo de grave o agudo con el que habla una persona. Esta característica perceptiva está relacionada con la velocidad de vibración de las cuerdas vocales: cuanto más rápidamente vibren estas, más agudo se percibe el sonido.

La frecuencia fundamental, denominada también f_0, es la medida en hercios de la altura tonal. Suele presentar valores más elevados en el caso de las mujeres (sonidos más agudos) que en el de los hombres (sonidos más graves, es decir, bajas frecuencias).

Timbre: es la "personalidad" del sonido, el "color" que tiene. La percepción del timbre de un sonido se encuentra muy relacionada con la forma del tracto vocal en el momento de producción del sonido.

La figura 3 muestra una representación frecuencial de un sonido sonoro. Todos los sonidos sonoros tienen un espectro con tres propiedades:

Frecuencia de resonancia

Frecuencia a la que la función de transferencia de un sistema (por ejemplo, el tracto vocal) toma el valor máximo.

- La envolvente decae con la frecuencia. La envolvente espectral habitualmente se denomina tracto vocal, puesto que la forma que tiene se encuentra relacionada con la forma del tracto vocal durante la producción del sonido.

- Existen unas frecuencias de resonancia del tracto vocal, llamadas formantes, que son diferentes según el sonido que se pronuncia y la persona que lo produce.

- La envolvente está muestreada a una frecuencia fija, denominada también f_0, que es la medida de la característica perceptiva de la altura tonal del sonido.

Propiedades segmentales de las señales de voz

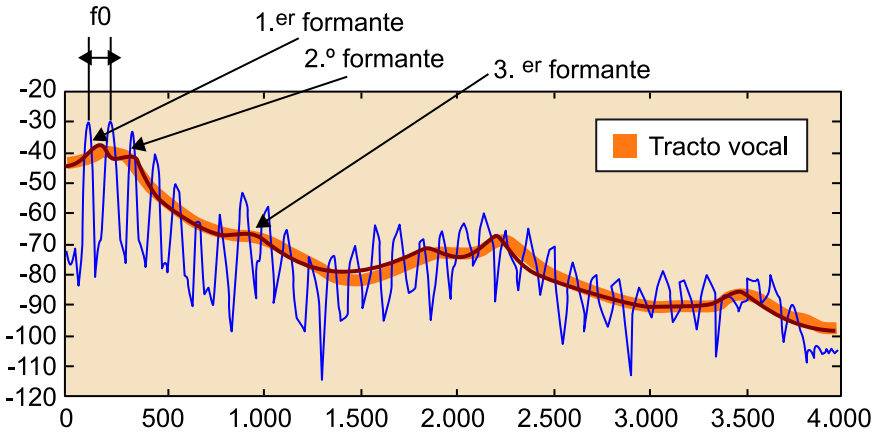

Figura 3. Espectro de un segmento sonoro de voz donde se aprecian el tracto vocal, los formantes y la f_0.

3.2. Propiedades suprasegmentales

La **prosodia** es la característica de la voz que engloba cómo se perciben la melodía, el ritmo y el volumen de la voz. Los parámetros que normalmente se utilizan para representar la prosodia son los siguientes:

- El contorno melódico, es decir, la curva de evolución de f_0 con el tiempo.
- La duración de cada sonido y también la posición y duración de las pausas.
- La evolución de la energía de la señal.

4. Clasificación fonética de los sonidos

Un **fonema** es la unidad sonora más pequeña con entidad propia. Algunos ejemplos de fonemas, en nomenclatura de la IPA, son [a] y [u]. Los fonemas son abstracciones teóricas; las diferentes pronunciaciones de un fonema se denominan **alófonos**.

Alfabeto fonético internacional

El alfabeto fonético internacional, o *international phonetic alphabet* (IPA), es un alfabeto fonético que representa de manera unívoca todo el inventario de fonemas que puede producir el aparato vocal humano.

Los fonemas se describen según las propiedades siguientes:

- **Tipo de sonido**: vocálico, consonántico o semiconsonántico. En los sonidos vocálicos, la lengua se sitúa de modo que no impide la salida del flujo de aire; en cambio, en la articulación de los sonidos consonánticos, la lengua obstruye parte del flujo. Los sonidos semiconsonánticos son similares a los vocálicos, pero tienen una duración más corta y no pueden estar acentuados.

- **Sonoridad**: según si las cuerdas vocales vibran o no en la producción, el sonido es sonoro o sordo. Todas las vocales son sonoras, pero existen consonantes sonoras y consonantes sordas.

- **Punto de articulación** en los sonidos consonánticos. El punto de articulación es el lugar donde se produce la obstrucción del flujo de aire. Según el punto de articulación, las consonantes se pueden clasificar de la manera siguiente:
  - Bilabial: obstrucción entre los labios.
  - Labiodental: obstrucción entre el labio inferior y los dientes superiores.
  - Dental: obstrucción entre la punta de la lengua y los dientes superiores.
  - Velar: obstrucción entre el fondo de la lengua y el paladar blando.
  - Alveolar: obstrucción entre la punta de la lengua y los alveolos.
  - Palatal: obstrucción entre el medio de la lengua y el paladar duro.

- **Modo de articulación** en los sonidos consonánticos. El modo de articulación describe cómo se produce el contacto entre los articuladores (lengua, labios, dientes, etc.) en el punto de articulación.
  - Oclusivo: oclusión completa de las cavidades bucales y nasales.
  - Nasal: oclusión completa de la cavidad bucal.

Ejemplo

- El fonema [a] de *c**a**sa* es un sonido vocálico, sonoro, abierto y central.
- El fonema [p] de ***p**az* es un sonido consonántico, sordo, oclusivo y bilabial.

  - Fricativo: fricción continua del flujo de aire en el punto de articulación.

  - Africado: comienzo oclusivo y final fricativo.

  - Aproximante: obstrucción muy ligera.

  - Lateral: aproximante pero utilizando el lado de la lengua.

  - Vibrante: el articulador, normalmente la lengua, vibra.

- **Altura y profundidad** en los sonidos vocálicos: las vocales se clasifican según la posición de la lengua durante la producción; en concreto, según la altura de la lengua respecto a las mandíbulas (abierta, medio abierta, cerrada, medio cerrada) y la profundidad en la cavidad bucal (frontal, central, posterior).

5. Unidades acústicas utilizadas en las tecnologías del habla

Acústicamente, en la realización de un fonema se distinguen tres partes: comienzo, centro y final. El centro de la realización del fonema es la parte más estable, donde se mantienen todas las características acústicas del sonido. En cambio, al comienzo y al final se da una transición entre los parámetros acústicos del sonido anterior y el posterior y el sonido que realmente se quiere producir. Por lo tanto, si comparamos diferentes pronunciaciones de un fonema, veremos que la parte central es similar, pero el comienzo y el final pueden ser muy diferentes. Este fenómeno de adaptación del comienzo y final del sonido al contexto en el que se produce se denomina **coarticulación**.

En las tecnologías del habla se utilizan unidades acústicas derivadas de los fonemas para tener en cuenta el fenómeno de coarticulación. En reconocimiento automático del habla se utilizan habitualmente los trifonemas y en síntesis del habla, los difonemas.

Un **trifonema**, denominado también *fonema con contexto*, es un fonema en el que se indica el fonema anterior y posterior. El trifonema $p^{+}au^{-}$, por ejemplo, se corresponde con el fonema [a], pero solo cuando se encuentra detrás del fonema [p] y ante el fonema [u], como sucede en ***pau*** (*pau*, en castellano, es *paz*). Si una lengua tiene N fonemas diferentes, habrá N^3 posibles trifonemas.

Nomenclatura fonética

La nomenclatura p^{+} indica la parte del fonema [p] que va desde la parte estable hasta el final del sonido, mientras que p^{-} hace referencia a la parte que va desde el comienzo del fonema hasta el medio de la parte estable.

Un **difonema** es una unidad que empieza en medio de la zona estable de un fonema y acaba en medio de la zona estable del fonema siguiente. Por ejemplo, la palabra *pau* construida a partir de difonemas es $\_p^{-}p^{+}a^{-}a^{+}u^{-}u^{+}\_$ (donde "_" indica silencio).

Segmentación de una señal de voz

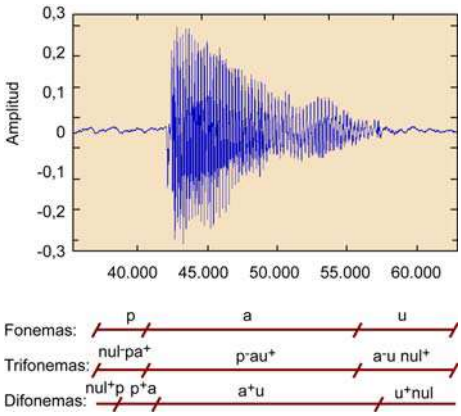

Figura 4. Segmentación según fonemas, trifonemas y difonemas de la locución *pau*

Glosario

altura tonal *f* Nivel perceptivo de grave o agudo con el que habla una persona. La frecuencia fundamental (f_0) es la medida de la característica perceptiva de la altura tonal del sonido. En inglés, *pitch*.

coarticulación *f* Fenómeno que provoca que la pronunciación de un sonido dependa del contexto, es decir, pronunciamos diferente un mismo fonema dependiendo del fonema anterior y posterior y del nivel de acentuación que tiene. El fonema se adapta al entorno para que la voz suene fluida.

difonema *m* Unidad acústica que empieza en medio de la zona estable de un fonema y acaba en medio de la zona estable del fonema siguiente.

pitch *m* Altura tonal.

fonema *m* Unidad sonora más pequeña con entidad propia. Acústicamente se puede dividir en tres zonas: el comienzo, el medio y el final. En el centro, las características acústicas son estables, mientras que el comienzo y el final son zonas de transición.

formante *m* Frecuencia de resonancia del tracto vocal.

prosodia *f* Característica de la voz que engloba cómo se percibe la melodía, el ritmo y el volumen de la voz. Se describe por medio de la curva de evolución de f_0, la duración de los fonemas, la duración y posición de las pausas, y la curva de energía.

tracto vocal *m* Físicamente, es el área que hay entre las cuerdas vocales y los labios, incluyendo tanto la cavidad bucal como las cavidades nasales. También se denomina *tracto vocal* a la envolvente del espectro de la señal de voz. La forma que tiene determina qué sonido se pronuncia.

trifonema *m* Unidad acústica que consiste en un fonema dependiente del contexto. Es decir, es un fonema en el que se indica el fonema anterior y posterior.