

Interacción persona-robot

Juan Antonio Corrales Ramón, Francisco Javier García Polo

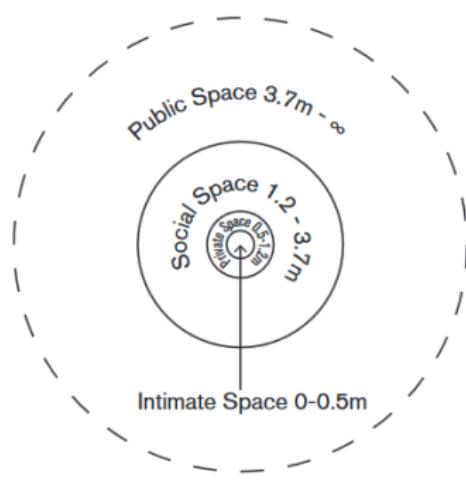
Departamento de Electrónica y Computación
Universidad de Santiago de Compostela

- “*Los interfaces multimodales combinan muchas formas simultáneas de modalidades de entrada/salida*”
- Más de una entrada y salida para la misma tarea
- Los humanos somos multimodales: vista, oído, olfato, tacto...
- La multimodalidad es deseable y necesaria para HRI con robots de servicio:
 - Interacción espacial
 - Interacción no-verbal: gestos, emociones, ...
 - Interacción verbal

- ¿Dónde posicionar a los robots durante la interacción?
- Solo evitar obstáculos: comportamiento incómodo, rudo, o inapropiado
 - Por ejemplo, cuando una aspiradora Roomba trata a las personas como “obstáculos” y sigue chocando con ellos mientras trata de evitarlos
- Hay que tener en cuenta las preferencias de las personas y las **normas sociales** que existen con respecto a dicha ubicación en relación con otros
 - Los otros esperan que se respeten las **distancias sociales** cuando hay espacio disponible

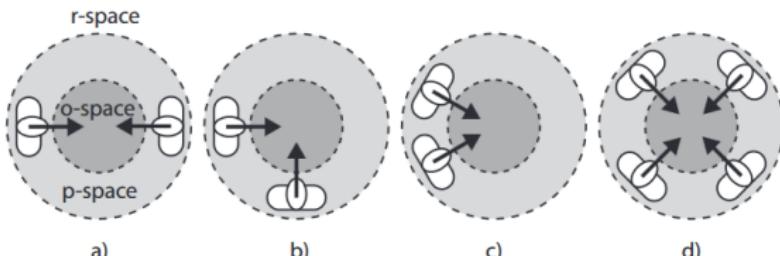
Interacción espacial

- Proxémica ([DOI: 10.1086/200975](https://doi.org/10.1086/200975))
 - Describe cómo las personas ocupan el espacio en relación con otras y cómo el posicionamiento espacial influye en las actitudes, los comportamientos y la interacción interpersonal
 - Cuatro distancias: distancia íntima, distancia personal, distancia social y distancia pública
 - Dimensión de la experiencia cultural: Personas de diferentes culturas tienen diferentes preferencias y expectativas proxémicas



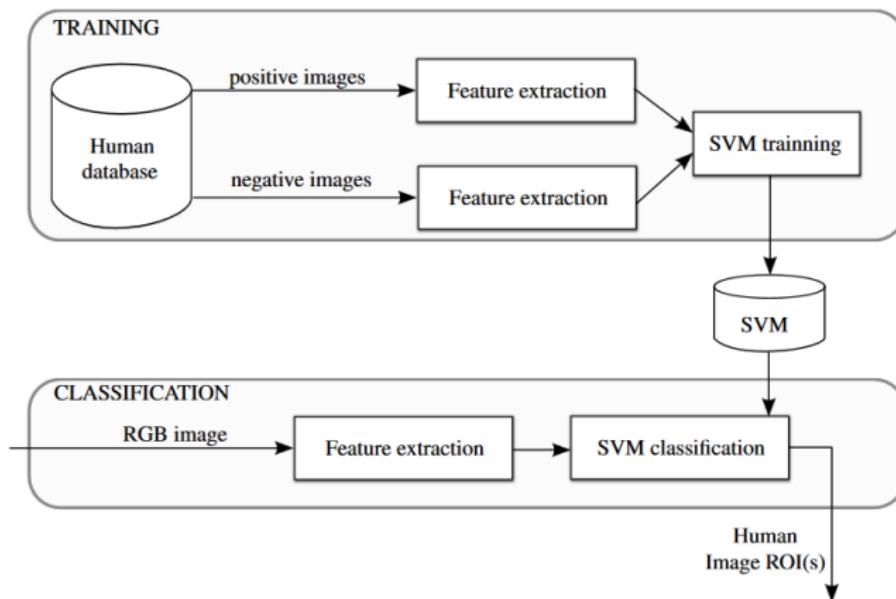
Interacción espacial

- No sólo es importante la distancia, también cómo nos posicionamos con respecto a los otros...
- *F-formation (Facing formation; DOI:10.3390/mti6030018)*:
 - *o-space*: Espacio compartido al que tienen acceso igual, directo y exclusivo los participantes. Si otra persona accede sin permiso, se considera una intrusión.
 - *p-space*: Los participantes lo ocupan alrededor del o-space con varias configuraciones posibles (frontal, en L, lateral o circular).
 - *r-space*: El resto del espacio fuera de los o/p-spaces.
- Un robot necesita estar al tanto de los movimientos de las personas involucradas en una conversación y no caminar entre ellas, incluso si hay suficiente espacio ([Ejemplo](#))



Interacción espacial

- Detección de personas
 - Esencialmente un problema de clasificación...



Interacción espacial

- Detección de personas
 - Esencialmente un problema de clasificación...



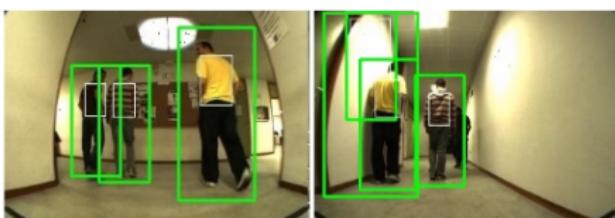
(a) Challenging illumination

(b) Pose variations



(c) Background interference

(d) Human-like objects

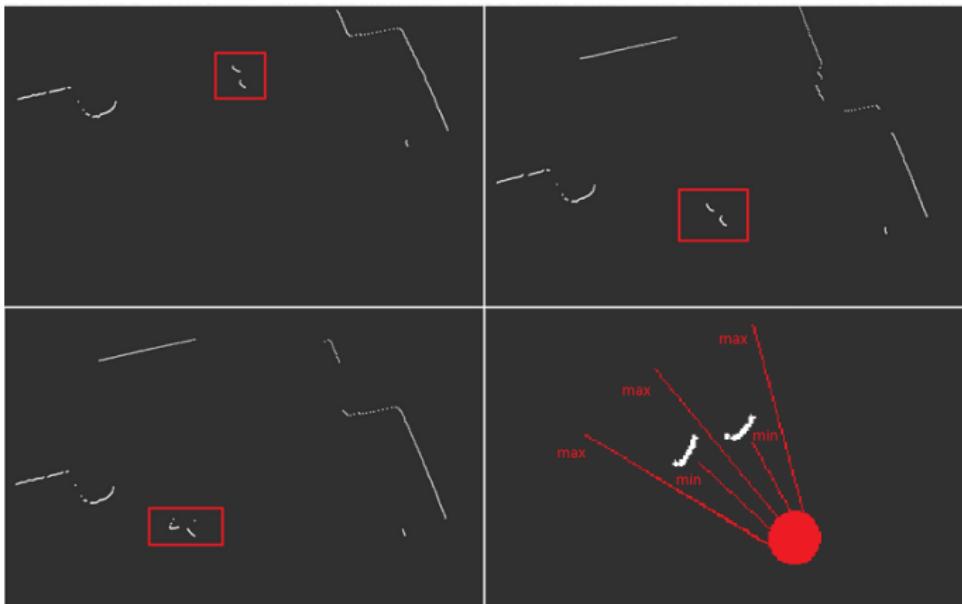


(e) People being correctly detected

(f) Positive detections and false positives

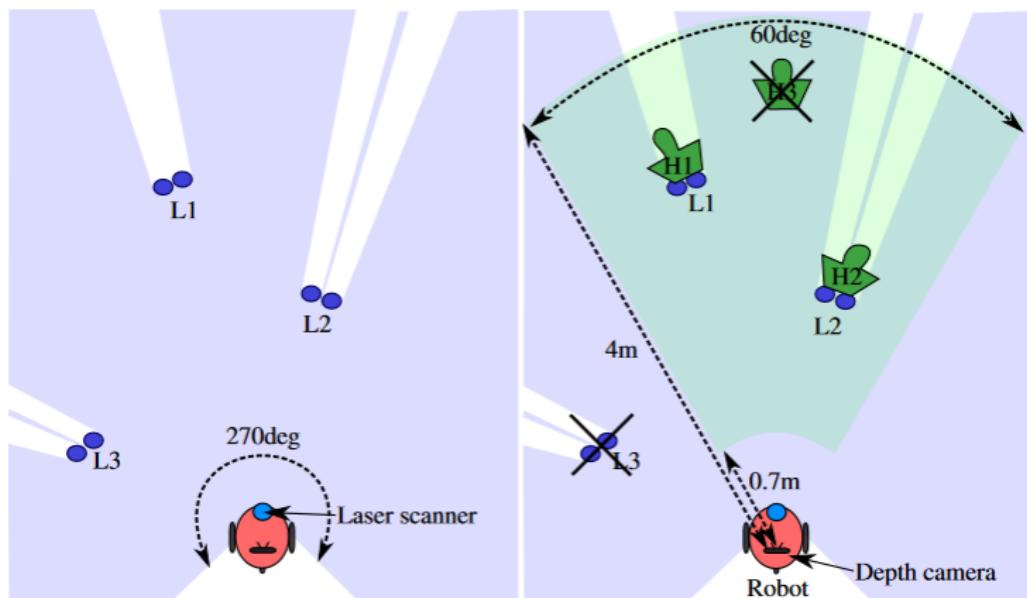
Interacción espacial

- Detección de personas
 - Detección de piernas en información 2D lidar también como un problema de clasificación implementado con Random Forest en OpenCV([leg detector](#) en ROS o la alternativa [leg tracker](#)) o como un problema [geométrico](#) según la forma de la pierna...



Interacción espacial

- Detección de personas
 - Mejora: Fusión con información visual (face detector). Sólo se reconoce como persona aquella que tenga una cara y dos piernas asociadas. Ejemplo: [DOI:10.3390/s20030722](https://doi.org/10.3390/s20030722)



- Localización

- Antes de comenzar la interacción el robot debe saber cuál es su situación en el espacio (odometría, mapeado, localización,...)
- Pérdida de precisión con el tiempo: se fusiona información sensorial de diferentes fuentes de información
- La localización permite al robot saber el tipo de espacio en el que se encuentra (e.g., en el salón, cocina, ...) generando un plano del entorno al mismo tiempo ([SLAM](#), [video](#))
- ... pero no dice nada de las personas dentro de ese espacio
- Localización de las personas (e.g., cámaras, LIDAR, motion capture, ...) e inclusión de información útil para ellas:
 - Diálogo para preguntar a las personas sobre nuevos lugares: [DOI:10.1109/IROS.2006.281770](#)
 - Inclusión de etiquetas visuales para añadir información semántica al plano: [DOI: 10.1145/1957656.1957815](#)
 - Reconocimiento de relaciones espaciales en conversaciones para identificar caminos respecto a puntos de interés: [DOI: 10.1109/HRI.2010.5453186](#)
 - Cognite mapping and navigation

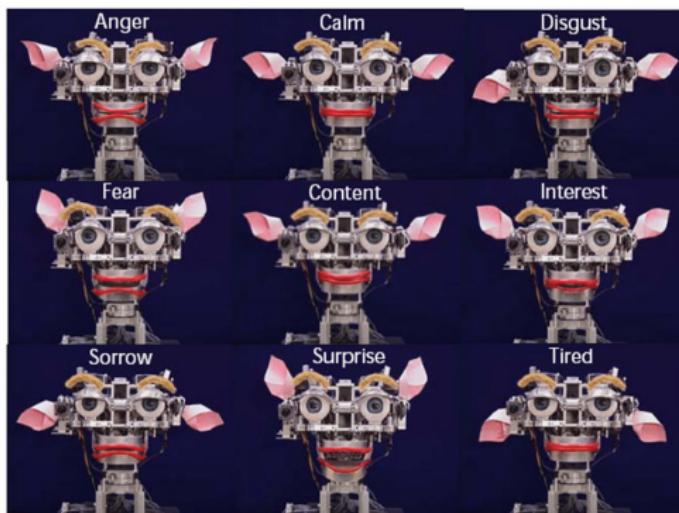
- Navegación
 - Path planning clásico → personas como obstáculos
 - Comportamiento socialmente inapropiado en HRI: los robots esperan hasta el último momento para evitar obstáculos pero esto puede ser considerado agresivo o intimidante.
 - (Re)Planificación considerando aspectos sociales:
 - Los humanos tienen en cuenta la dirección de avance de los otros humanos, intercambian información no verbal sobre el lado que ocuparán al cruzarse y evitan ocupar el espacio personal del otro.
 - El robot debe **anticipar el futuro** en la interacción: teniendo en cuenta las distancias respetadas por los humanos según la proxémica ([DOI: 10.1109/IROS.2006.282535](#)); desarrollando un modelo de los humanos para evitar sorprenderlos durante una acción ([DOI: 10.15607/RSS.2018.XIV.069](#), [Vídeo](#)); (re)planificando su trayectoria para que sea considerada como segura y cómoda aunque implique ineficiencia funcional ([DOI: 10.1109/ROBOT.2004.1308098](#), [Vídeo](#); [DOI: 10.1016/j.rcim.2010.07.005](#), [Vídeo](#))...

- Dinámica espacial para el inicio de la HRI
 - Cómo nos acercamos a las personas y cómo este acercamiento es percibido por los otros tiene implicaciones en la futura interacción
 - El robot debe diseñarse cuidadosamente para iniciar una interacción de manera adecuada al **contexto social**:
 - Cuando un robot se une a una cola, tiene que respetar el espacio personal ([DOI: 10.1023/A:1015273816637](https://doi.org/10.1023/A:1015273816637))
 - Cuando un robot se encuentra con personas, necesita cambiar su modo de navegación a considerar la distancia social, configuración espacial,... ([DOI: 10.1109/ROBOT.2004.1308100](https://doi.org/10.1109/ROBOT.2004.1308100))
 - Dicho inicio también depende del **contexto de la tarea**:
 - Es deseable acercarse desde el frente cuando un robot está tratando de iniciar una conversación,
 - Pero es menos deseable si el robot está entregando un objeto. ([DOI: 10.1145/1121241.1121272](https://doi.org/10.1145/1121241.1121272), Vídeo)
 - Estas estrategias se pueden aprender a partir de **demonstraciones humanas** ([DOI:10.1109/TRO.2016.2588880](https://doi.org/10.1109/TRO.2016.2588880), Vídeo1; [DOI:10.1145/3319502.3374798](https://doi.org/10.1145/3319502.3374798), Vídeo2)

- Informar a los usuarios de las intenciones del robot
 - Los usuarios prefieren que un robot se comporte con "legibilidad", i.e., de una manera que permita a los usuarios saber sus metas e intenciones (E.g. un dron indica su dirección de avance **con LEDs a su alrededor**)
 - El robot puede dar un objeto a una persona de muchas maneras, pero probablemente la más eficiente energéticamente puede ser incomprendible para la persona, por eso es preferible un movimiento fácil de interpretar. (E.g **un brazo robot** se acercará al objeto a agarrar siguiendo **una trayectoria circular** para interpretar mejor su movimiento aunque sea más lento que una trayectoria lineal).
 - El robot necesita saber cómo la persona percibe el mundo que le rodea (e.g., ¿Me das ese objeto?). La arquitectura del robot puede contener un **modelo geométrico del entorno** que contiene la localización de las personas, los robots, los objetos y sus relaciones de visibilidad y alcance. Se pueden utilizar **ontologías** para definir conceptos de más alto nivel (i.e. categorías de los objetos, estados mentales de los agentes...) con los que se implementa **razonamiento simbólico**.

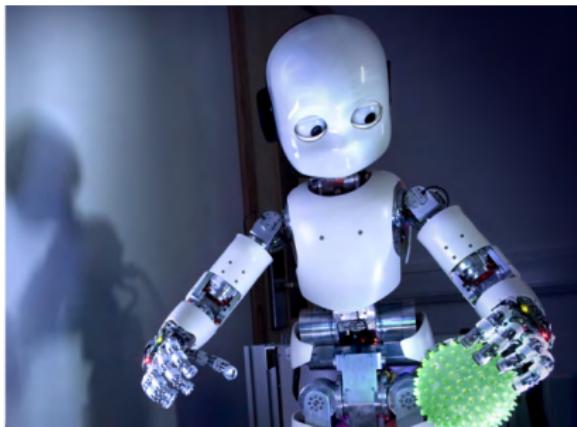
Interacción no-verbal

- Las personas captan constantemente (y automáticamente) una variedad de señales no verbales mientras interactúan
 - Se utilizan para interpretar los matices del significado, la emoción, las intenciones,...
 - Algunas más genuinas que el habla al ser involuntarias
- En robótica, se combina con el habla para proporcionar información complementaria sobre el **estado interno/intenciones** del robot



Interacción no-verbal

- Movimiento de la mirada (gaze) y los ojos
 - La mirada indica interés, comprensión, **atención** y la capacidad y voluntad de las personas para seguir la conversación
 - Facilitan la colaboración, como entregar un objeto a alguien o llamar la atención de alguien sobre la siguiente herramienta necesaria en una tarea
 - La mirada del robot puede ser una forma poderosa de gestionar las interacciones con una (e.g. si mira a la persona, **recuerda mejor** la conversación) o más personas (e.g. el robot **implica a diferentes personas** en una conversación)



- Gestos
 - Después del habla, gesticular es la forma más importante de proporcionar información durante una interacción ([DOI:10.1007/s12369-013-0196-9](https://doi.org/10.1007/s12369-013-0196-9)):
 - Los gestos **deícticos** se refieren a señalar cosas específicas en el entorno (atención conjunta)
 - Los gestos **íconicos** respaldan e ilustran aún más lo que se dice (e.g. mostrar los brazos abiertos mientras se habla de agarrar una pelota grande).
 - Los gestos **simbólicos**, como saludar o despedirse, pueden tener su propio significado
 - Los gestos de **ritmo** (beat gestures) se utilizan para acompañar el ritmo del habla y parecen mover los brazos mientras se habla como si dirigiera una **orquesta invisible**



- Mimetismo e imitación
 - Replicación **inconsciente (mimetismo)** o **consciente (imitación)** de otra persona
 - En la infancia, proporcionan una forma común de aprender nuevos comportamientos y normas sociales culturalmente relevantes
 - **Función social:** si dos personas usan los mismos gestos o adoptan la misma postura durante una conversación, generalmente es porque han establecido una relación positiva. Se llama "**efecto camaleón**" y su uso consciente pretende persuadir a otra persona imitándola.
 - Aprendizaje por demostración es uno de los paradigmas de aprendizaje más utilizados en robótica. El mimetismo aprendido se puede usar para incrementar la aceptación del robot en tareas interactivas [DOI:10.1007/s12193-009-0028-2](https://doi.org/10.1007/s12193-009-0028-2)

Interacción no-verbal

- Tacto
 - **Interacciones cercanas** entre personas, como entre los amigos, o entre cuidadores y pacientes (calmar, consolar, ayudar, . . .)
 - **Efecto Midas**: los camareros obtienen una propina más alta si tocan accidentalmente a los clientes antes de que paguen
 - Parte integral de la HRI tanto en **tareas funcionales** (i.e. la entrega y manipulación de objetos) como en **tareas sociales** (i.e. apretón de manos para saludar)
 - Considerar las implicaciones psicológicas del contacto:
 - La interacción táctil con robots diseñados para el contacto (como **Telenoid**) y en aplicaciones de **cuidado** es aceptada
 - Aquellas personas que tienen una **actitud negativa** hacia el robot pueden incrementarla si son obligados a tocarlo



Interacción no-verbal

- Postura y movimientos
 - Las posturas se pueden utilizar para interpretar el **estado emocional** de una persona
 - Los movimientos lentos, los hombros caídos y los gestos letárgicos sugieren un estado mental abatido, mientras que los movimientos rápidos y el porte erguido son signos positivos
 - Cuando el robot **carece de rasgos faciales expresivos** (**Nao** en el proyecto [Aliz-E](#)), el cuerpo será el medio de comunicación principal de las emociones

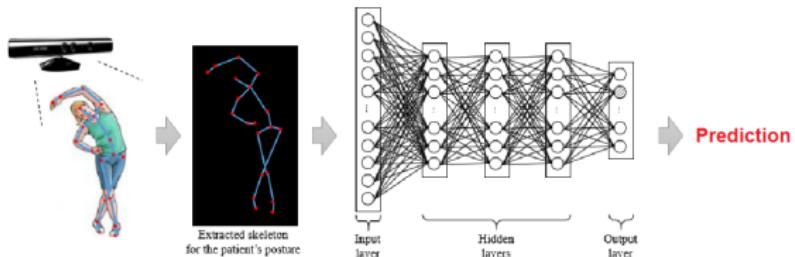


Figure 2a-c: The Thrifty Faucet in different postures: Seeking, curious, and rejecting.

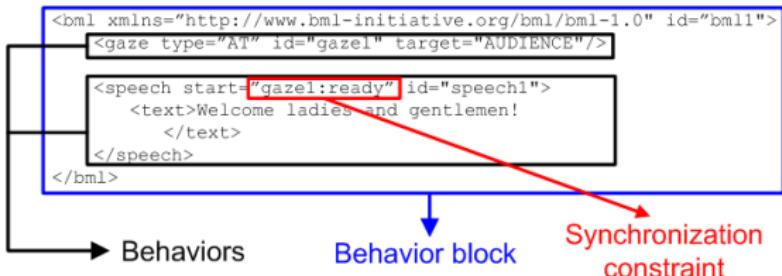
Robot-grifo para concienciar sobre el uso del agua

Interacción no-verbal

- En robótica:
 - **Percepción:** reconocimiento de patrones, reconocimiento de posturas y gestos con cámaras 2D, 3D, sensores iniciales → Aprendizaje automático



- **Generación:** frameworks de animación como [Choregraphe](#), o Behaviour Markup Language ([BML](#))



- La forma más obvia de comunicación: audible y explícita
- Hacer hablar a un robot es mucho más simple que comprender el habla humana, lo que crea un desequilibrio entre las expectativas de las personas y las capacidades reales del robot
- El habla es increíblemente compleja y abierta a multitud de interpretaciones



Thank You for Sharing that Interesting Fact!

Robot: Hello my name is Waiterbot. I will be helping you today.

Participant: Great, can I order something?

Robot: Please tell me your order.

Participant: Um, can I have one water?

Robot: Yes, that is permissible.

Participant: Great.

Robot: Please tell me your order.

Participant: Can I have one water?

Robot: Yes, that is permissible.

Participant: Great, I'll take one water.

Robot: Thank you for sharing that interesting prediction. Please tell me your order.

Participant: May you please bring me one water?

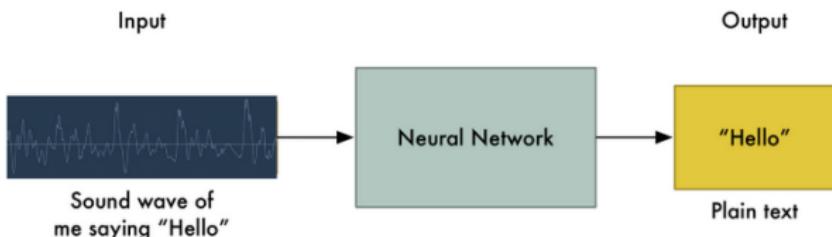
Robot: Yes, I am able to do that.

Participant: My order is water.

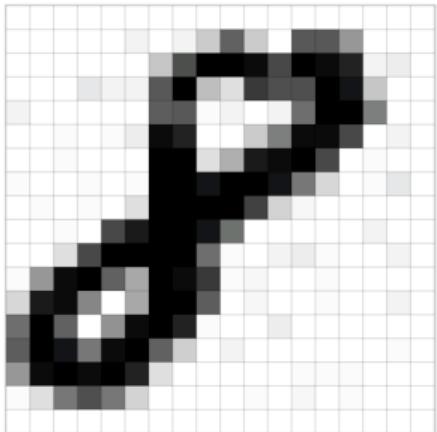
Robot: Okay. (Robot retrieves meal card for participant)

- Componentes:
 - **Enunciado:** una palabra o conjunto de palabras que comunica una idea con sentido completo en una situación determinada
 - **Palabras:** las unidades más pequeñas que podemos pronunciar para transmitir significado
 - **Fonemas:** pequeñas unidades de sonido que forman la palabra
 - Mientras que las personas esperan una adherencia bastante estricta a las reglas gramaticales y la sintaxis en el texto escrito, se vuelven mucho más liberales cuando hablan → la tarea es mucho más compleja
 - El robot debe transcribir el habla en palabras, comprender las palabras y dar una respuesta adecuada

- Speech recognition
 - Es el proceso que toma una grabación digital de voz y la transcribe
 - Por sí solo no comprende ni interpreta lo que se ha dicho
 - Posicionar el micrófono de forma adecuada en el robot, posición del usuario, ...
 - Aprendizaje automático

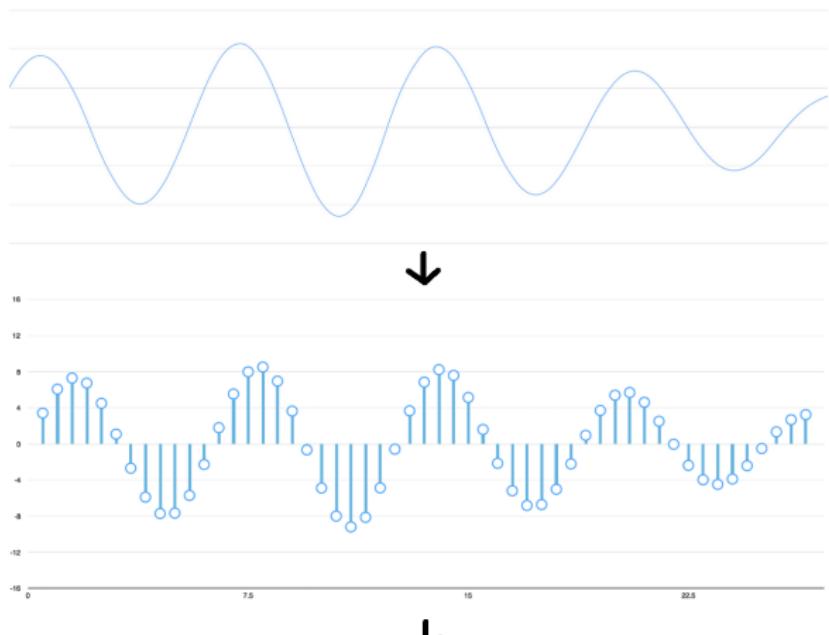


- Convertir la señal de entrada (imagen/sonido) en números...
 - Imagen:



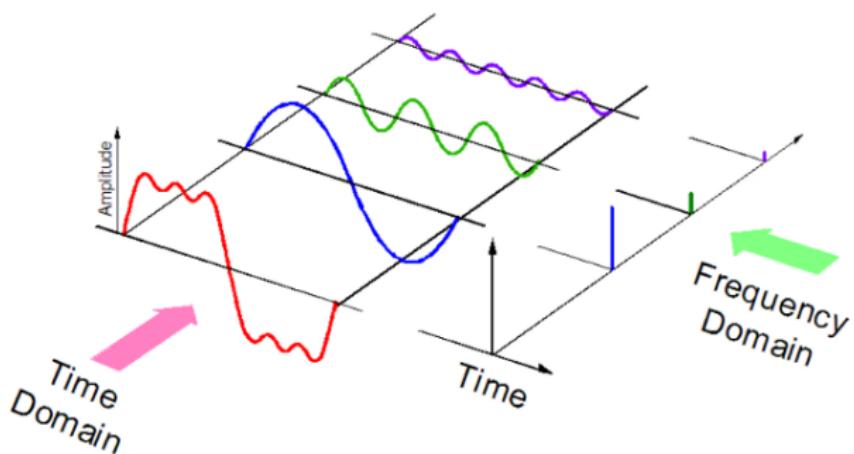
Speech recognition

- Sonido:



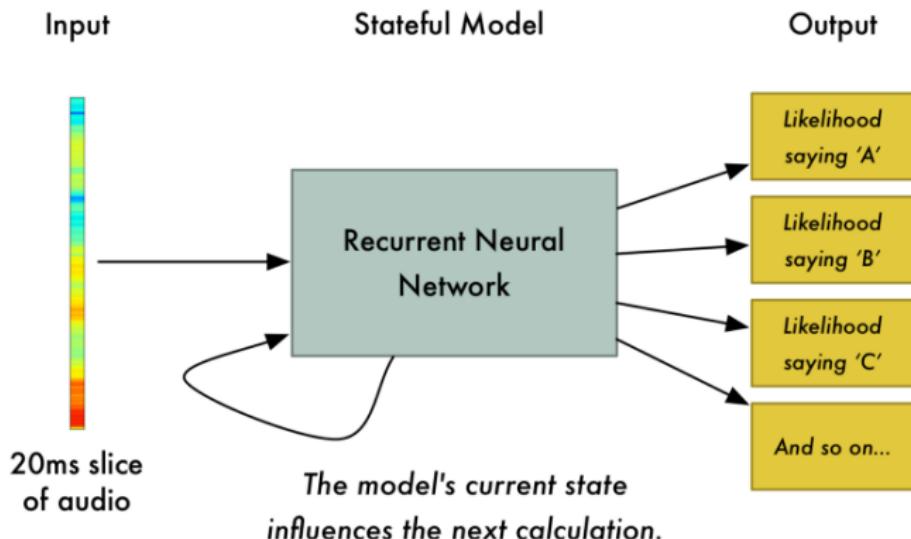
Speech recognition

- Ahora tenemos un array de números para suministrársela a la red neuronal...
- ...pero aún requiere mucho espacio porque el ratio de muestreo puede ser bastante elevado
- Transformada de Fourier: Los fonemas se identifican mejor en el **dominio de la frecuencia**.



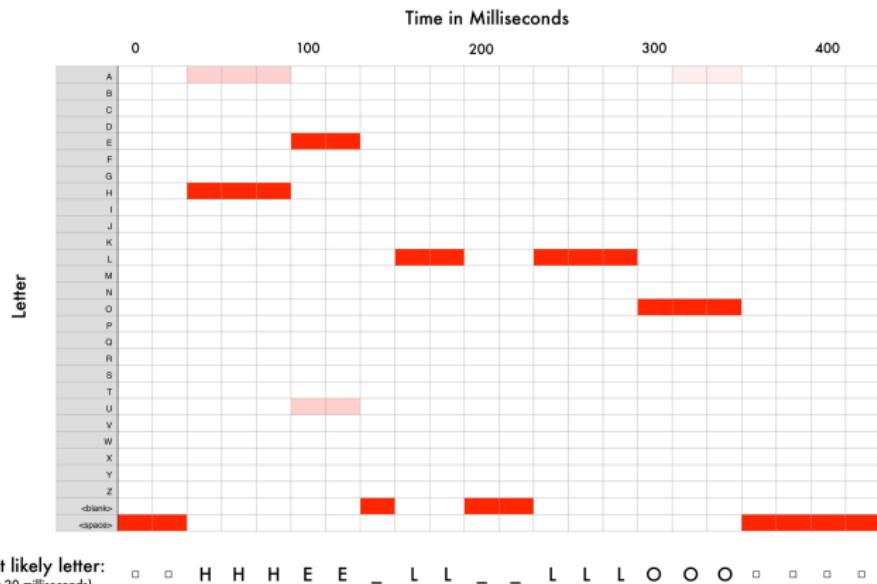
Speech recognition

- Se sumisten porciones de audio en el dominio de la frecuencia a la red neuronal que decide qué letra es la más probable



Speech recognition

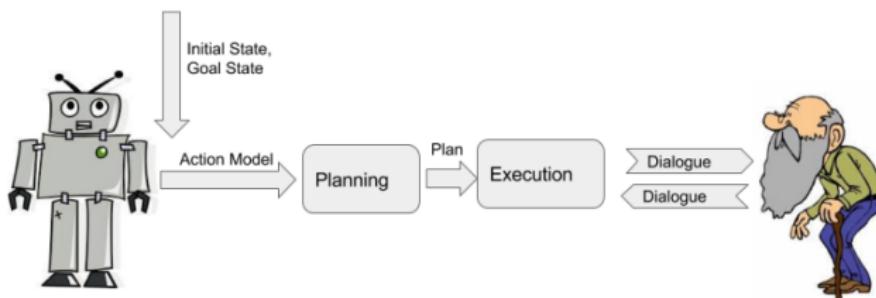
- Se sumisten porciones de audio en el dominio de la frecuencia a la red neuronal que decide qué letra es la más probable



- Limitaciones
 - Formas de hablar atípicas
 - Modelos insuficientemente entrenados
 - Dialectos locales
 - Ruido, multitudes,...

- Comprensión:
 - *Speech recognition* no significa **comprensión**
 - *Sentiment analysis* en las redes sociales o en sitios web de e-commerce permiten conocer las opiniones de los usuarios para mostrar elementos que les interesen ([vídeo en HRI](#)):
 - Afinidad léxica: Clasifica en categorías de afecto al asociarlas con ciertas palabras clave (e.g. triste, feliz...)
 - Análisis gramatical: Encuentra relaciones de contexto entre opinión y objeto
 - Enfoques a nivel de concepto: Utilizan representaciones del conocimiento (e.g. ontologías y redes semánticas) para detectar relaciones semánticas sutiles mediante conceptos ligados entre sí.
 - Comprensión del lenguaje natural (NLU): trata de representar la información percibida utilizando lógica de predicados, redes semánticas, en definitiva en alguna representación que le permita después “razonar”
 - **Aún no se ha logrado una comprensión genuina del lenguaje natural**

- Gestión del diálogo:
 - *Finite-state machines* para diálogos sencillos: requieren escribir un script con la secuencia del diálogo (e.g., QiChat en SoftBank Robotics)
 - Planificación: mantienen un estado interno del mundo y son capaces de anticipar el futuro, por lo tanto no requieren definir explícitamente un script



- Mediante planificación automática generación de textos policíacos, mitológicos, ...
- Acciones:
 - Move-to-kill: Un personaje se mueve al lugar donde está otro al que tiene intención de matar
 - Move-to-kidnap: Un personaje se mueve al lugar donde está otro al que intención de secuestrar
 - Pick-up-object-to-kill: Un personaje desarmado con intención de matar a otro coge un objeto que se encuentra en su misma posición
 - Pay-debt: Un personaje endeudado con otro paga al segundo su deuda
 - Find-clue-object: Un policía descubre una prueba delictiva que incrimina a otro personaje y pretende capturarlo
 - ...

"Lucabradi goes from his house to Barzini's house and picks up rifle to rob. Lucabradi betrays Vito and quits Corleone family. Vito command Michael to kill Lucabradi for treason. Michael goes from coleone's house to lucabradi's house and picks up revolver to kill Lucabradi by Vito's command. Michael goes from Lucabradi's house to Barzini's house to kill Lucabradi by Vito's command. Michael kills Lucabradi with revolver at Barzini's house by Vito's command."

"Lucabradi betrays Vito and quits Corleone-family. Barzini decide to kidnap Kay due to an unpaid debt. Barzini goes from House-Barzini to House-Corleone to kidnap Kay. Barzini attacks Kay at House-Corleone. Kay gets kidnapped by Barzini. Lucabradi goes from House-Lucabradi to House-Corleone and picks up Machine-Gun to rob. Michael decide to kill Sollozzo for vengaenze. Lucabradi goes from House-Corleone to bank to commit a bank robbery. Vito goes from House-Corleone to Jewelry-store-nino to kiss Marie. Lucabradi commits a bank robbery at bank. Lucabradi goes from bank to Restauran-Louis, Louis' business, to commit a robbery. Michael goes from House-Corleone to Restaurant-Louis and picks up gun to kill Sollozzo. Vito command Michael to kill Lucabradi for treason. Michael tries to kill Lucabradi at Restauran-louis by Vito's command, but fails. Lucabradi gets hurt. Lucabradi goes from Restauran-Louis to Jewelry-store-nino... "

Interacción verbal

- Producción del habla (Agentes conversacionales):
 - Convertir una respuesta de texto a sonido: *text-to-speech* (TTS)
 - Actualmente casi indistinguible del habla humana
 - Producción de habla concatenativa: se graba la voz de un actor y se corta en fonemas, que luego se “pegan” y se suavizan para proporcionar un discurso que suene natural
 - DNN: relacionan el texto en la entrada con sus sonidos en la salida
 - Basada en una arquitectura multi-componente

