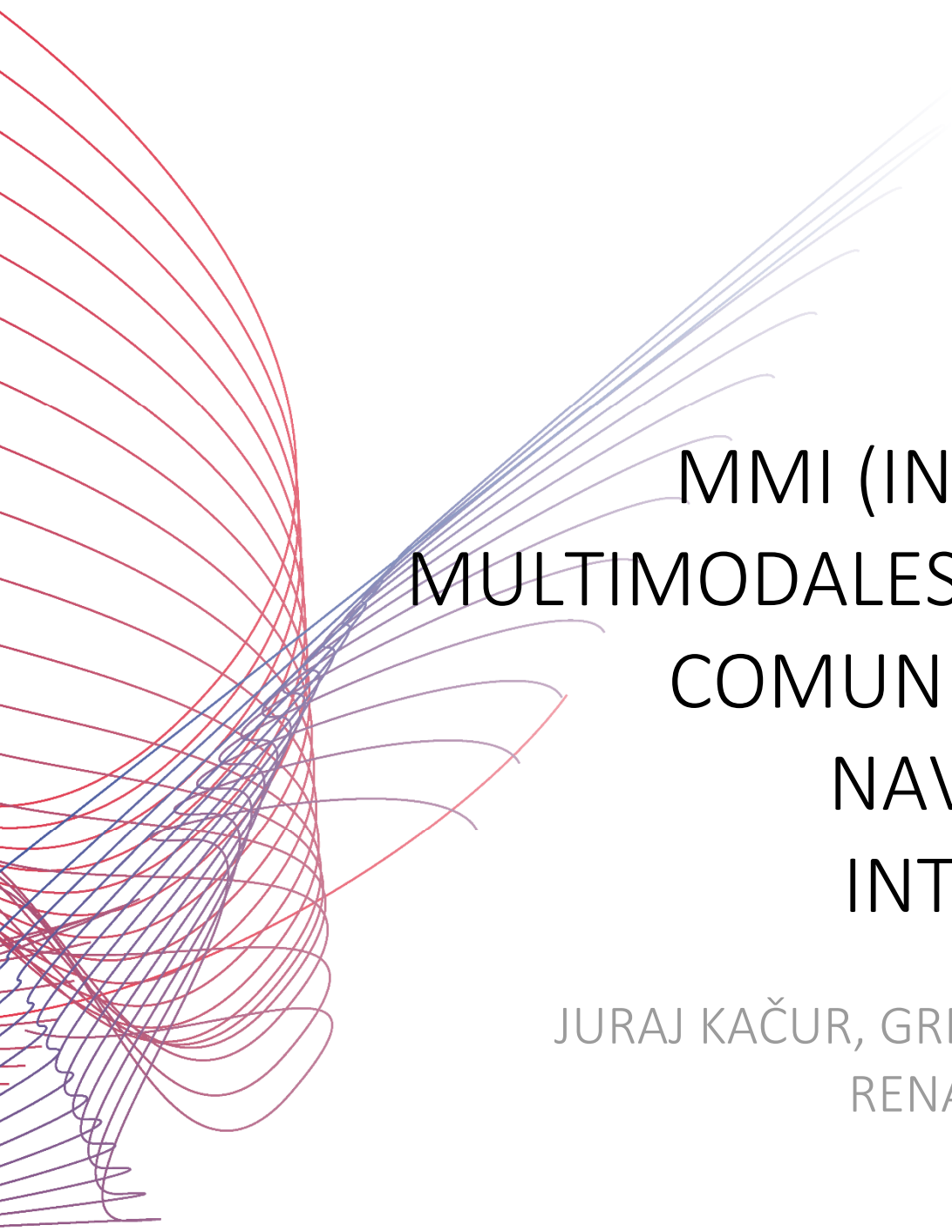
en cs de sk es

TECH pedia



MMI (INTERFACES MULTIMODALES) PARA LA COMUNICACIÓN Y NAVEGACIÓN INTERACTIVA

JURAJ KAČUR, GREGOR ROZINAJ, RENATA RYBÁROVÁ

Título: MMI (interfaces multimodales) para la comunicación y navegación interactiva

Autor: Juraj Kačur, Gregor Rozinaj, Renata Rybárová

Traducido por: Miquel Soriano

Publicado por: České vysoké učení technické v Praze

Fakulta elektrotechnická

Dirección de contacto: Technická 2, Praha 6, Czech Republic

Número de teléfono: +420 224352084

Print: (only electronic form)

Número de páginas: 41

Edición: Primera edición, 2017

ISBN 978-80-01-06244-9

TechPedia
European Virtual Learning Platform for Electrical and Information Engineering

http://www.techpedia.eu

Erasmus+

El presente proyecto ha sido financiado con el apoyo de la Comisión Europea.
Esta publicación (comunicación) es responsabilidad exclusiva de su autor. La Comisión no es responsable del uso que pueda hacerse de la información aquí difundida.

NOTAS EXPLICATIVAS

	Definición
	Interesante
	Nota
	Ejemplo
	Resumen
	Ventajas
	Desventajas

ANOTACIÓN

La interfaz multimodal (MMI) consta de varios temas y módulos que sirven para una comunicación natural y fácil de usar con el sistema. En conjunto, estos módulos representan la funcionalidad de MMI. La lista de módulos que pueden ser parte de la interfaz multimodal es la siguiente: el reconocimiento del hablante, el reconocimiento de comandos y del habla, el reconocimiento multi facial, la navegación por gestos, la navegación con los ojos y la síntesis de voz. Este módulo introducirá algunos de ellos para ayudar a entender los principios básicos y los algoritmos que se nos presentan en nuestra vida cotidiana.

OBJETIVOS

El objetivo principal del módulo es introducir al estudiante en lo fundamental de la interfaz multimodal y su aplicación. El estudiante se familiarizará con los principios básicos del reconocimiento facial, la identificación del hablante y la navegación por comandos de voz y por gesto. Al final, se describen las posibles aplicaciones multimodales de todas las modalidades mencionadas.

LITERATURA

[1] Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. EVALUATION: Final Multimodal Interface for User/Group-Aware Personalisation, Deliverable 5.5.1, available at http://www.hbb-next.eu/index.php/documents, Decmber 2013.

[2] Vančo, Marek; Minárik, Ivan; Rybárová, Renata. Evolution of static gesture recognition. In: Redžúr 2014 proceedings; 8th International Workshop on Multimedia and Signal Processing; 13 May 2014, Dubrovnik, Croatia. Bratislava: Nakladateľstvo STU, 2014, p. 41-44. ISBN 978-80-227-4162-0.

[3] Rautaray, Siddhart S., Agrawal, Anupam. Vision based hand gesture recognition for human computer interaction: a survey, Springer, 2012, pp 106-113. ISBN 978-3-642-27386-5

[4] Kristensson, Per Ola; Denby, L.C. Continuous recognition and visualization of pen strokes and touch-screen gestures. In Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling, 2011, pp. 95-102.

[5] Gülbahar, Mark; Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. Design and Protocol: Intermediate Multimodal Interface and Context Aware Recommendation Engine, Deliverable 5.3.1, available at http://www.hbb-next.eu/index.php/documents, October 2012.

[6] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech communication, Vol. 52, pp. 12-40, Elsevier, 2010

[7] Benesty J., Sondhi M. M., Huang Y., Springer Handbook of Speech Processing, Springer-Verlag, ISBN 978-3-540-49125-5, 2008

Indice

1 Arquitectura MMI

En la actualidad, las interfaces más utilizadas para la comunicación hombre-máquina son el teclado, el ratón o tableta táctil. Estos dispositivos representan la adaptación de los seres humanos a las limitaciones de los ordenadores en lugar de una comunicación natural con el ordenador. En los últimos años, ha aparecido un nuevo requisito y es que los seres humanos necesitan comunicarse con las máquinas de la misma manera como lo hacen entre sí: por el habla, mímica y gestos, ya que éstos conciben mucha más información que los anteriores dispositivos periféricos. Esto nos lleva al término *interfaz multimodal* (**MMI -** *multimodal interface*).

La interfaz multimodal se compone de varios temas y módulos que sirven para la comunicación natural y fácil de usar con el sistema. En conjunto, estos módulos representan la funcionalidad de MMI. Estos módulos pueden formar parte de la interfaz multimodal:

- Identificación de voz múltiple
- Reconocimiento de comandos y habla
- Reconocimiento de comandos de voz y habla
- Reconocimiento facial múltiple
- Navegación por gestos
- Navegación con los ojos
- Síntesis del habla
- Motor de recomendación

La arquitectura general de la interfaz multimodal se compone de varias capas. La capa física representa los dispositivos hardware de entrada y salida que permiten la interacción con el mundo real. Los datos multimodales proporcionados por los dispositivos de entrada (cámaras, sensores, micrófonos, etc.) son procesados en paralelo por cada módulo por separado. El controlador MMI recoge los datos de salida de todos los módulos, los evalúa y los combina en un solo flujo de datos de salida. El flujo contiene información sobre los usuarios reconocidos y sus acciones solicitadas.

2 Identificación de usuario

Hoy en día las tecnologías de la información están cada vez más en un primer plano. Nos encontramos con ellas en nuestra vida cotidiana, desde los teléfonos inteligentes a los dispositivos más sofisticados.

Todos los dispositivos que manejan datos personales deben ser protegidos mediante algún tipo de sistema de seguridad. El tipo de autenticación más utilizado es la contraseña, pero con el desarrollo de las tecnologías de la información y los algoritmos de seguridad, los sistemas y las aplicaciones comienzan a utilizar la autenticación basada en factores biométricos. Existen varios beneficios por el uso de señales biométricas: la biometría es exclusiva de una persona, por lo que es suficiente la presencia de una persona frente al sensor y no existe la necesidad de recordar varias contraseñas o PINs. Conocemos sistemas de identificación basados en diferentes modalidades biométricas como el iris, huellas digitales, voz, cara, etc. En esta parte nos centraremos en los sistemas de reconocimiento facial y en los sistemas de identificación del hablante en general.

2.1 Reconocimiento facial 2D y 3D

El reconocimiento facial es un subconjunto de gran campo de la investigación y la tecnología del reconocimiento de patrones. El reconocimiento de rostro humano se ha convertido en uno de los métodos de autenticación biométrica más importantes de las últimas décadas, debido a su potencial para una amplia variedad de aplicaciones y áreas (vigilancia, seguridad en el hogar, control de fronteras, etc.). Los sistemas biométricos de identificación personal, desarrollados por varios proveedores, consiguen una muy alta precisión en el reconocimiento de la cara. La mayoría de estas aplicaciones requieren [1]:

- Sistemas de reconocimiento que puedan reconocer varias caras a partir de una imagen de vídeo o de fotografía
- Alta tasa de reconocimiento
- Invarianza a la iluminación
- Estabilidad ante el cambio de expresiones faciales y poses
- Reconocimiento en tiempo real, etc.

La principal ventaja en comparación con otros enfoques es que el reconocimiento facial no requiere la colaboración por parte del usuario puesto que las imágenes de la cara se pueden adquirir desde la distancia por una cámara. Otra ventaja es que los dispositivos de adquisición son baratos y se están convirtiendo en elementos básicos.

El principal inconveniente del reconocimiento facial es la actual relativa facilidad con que puede ser burlado.

Sin embargo, en comparación con el reconocimiento del hablante, el reconocimiento facial alcanza resultados mucho mejores. En general, hay tres enfoques principales basados en el tipo de datos que se utilizan en el proceso de reconocimiento: los métodos basados en la imagen de intensidad en 2D, los que se basan en métodos de datos faciales en 3D y la técnica que utiliza ambos tipos de datos. El proceso completo de reconocimiento consta principalmente de 3 etapas. La primera de ellas es la de adquisición y el procesado previo, la segunda es el registro de datos y la tercera etapa es el reconocimiento. Una descripción detallada sobre el reconocimiento facial 2D y 3D se puede encontrar en el módulo de *Identificación de Usuario*.

En los sistemas modernos se puede implementar el reconocimiento facial múltiple. En ese caso se detectan todos los rostros en las imágenes. Las posiciones de las caras detectadas están asociadas con la imagen. A continuación, la imagen se divide en varias muestras basadas en las posiciones de las caras detectadas. Estas muestras generadas representan todas las caras en la escena y permiten realizar el seguimiento de cada cara por separado. La división de la imagen es un paso principal hacia el reconocimiento facial múltiple dentro de los sistemas.

2.2 Identificación del hablante

En términos generales, la tarea de identificación de hablantes consiste en decidir de forma automática a quien pertenece la muestra de voz desconocida. La decisión se basa en un conjunto de usuarios almacenados en una base de datos durante una fase de entrenamiento. Sin embargo, si la fiabilidad en la decisión observada es demasiado baja, el sistema puede no reconocer a nadie.

Como hay un conjunto fijo de usuarios que puede ser reconocido a la vez, esta tarea se llama a menudo un problema de grupo cerrado. La identificación del hablante ha estado bajo seria investigación científica desde hace 40 años y todavía la hay e incluso se continúa con un esfuerzo científico creciente. Con la llegada de nuevas y ampliamente extendidas tecnologías, se están encontrando aplicaciones en muchas áreas, sólo por mencionar algunas de ellas:

- **Ciencia forense**
- **Indexación automática de voz y grabaciones de audio almacenadas en bases de datos**
- **Método natural y no invasivo para el acceso y la protección de datos y de servicios,**
- **Aplicaciones para la industria del juego**
- **Ayuda para personas con discapacidad**

La identificación del hablante es un problema bastante complicado debido a muchas razones. Sólo mencionar algunas de ellas:

- **Variabilidad acústica del hablante**
- **Modificaciones causadas por el estado actual de salud, físico y mental**
- **Presencia de ruidos de fondo**
- **Influencia del dispositivo de grabación y del lugar de grabación (ecos)**
- **Fácil de replicar por los dispositivos de grabación**

Sin embargo, utilizar el habla como señal biométrica tiene las siguientes ventajas:

- **No hay necesidad de utilizar contraseñas**
- **Método no invasivo**
- **Fácil método de adquisición de datos para la obtención de una señal biométrica**

Existen muchas aplicaciones para la identificación del hablante que presentan diferentes niveles de complejidad, requisitos, confidencialidad, tiempo de respuesta, etc. Por lo tanto, se distinguen varias clases de sistemas de identificación:

- **Sistemas de texto independiente** - trabajan independientemente de lo que se dijo, por lo que no hay requisitos en las frases, oraciones o incluso en las palabras que se digan, tal vez con la excepción de la longitud de un discurso, es decir, el habla no corrompida y sin pausas. El rango de precisión de tales sistemas se encuentra entre el 80 y el 95%. Actualmente son los sistemas más frecuentes, ya que implican ninguna o muy pocas restricciones. Sin embargo, debido a su menor precisión no se pueden utilizar en aplicaciones críticas de seguridad.

- **Sistemas de texto dependiente** - requieren frases precisas pronunciadas por un desconocido. Tales sistemas pueden utilizar las características especiales de los individuos pronunciando ciertas frases (denominado efecto de co-articulación). Estos sistemas exhiben las tasas de precisión más altas, que van desde el 95% al 99%. Además, estos sistemas pueden ser incluso más precisos y robustos si el texto pronunciado es una contraseña específica para cada usuario. En determinados casos estos sistemas pueden alterar el texto que se diga haciendo diferentes preguntas.

- **Sistemas basados en información acústica**- utilizan sólo información acústica de un intervalo de tiempo corto derivado del habla. Es sobre todo una especie de espectro modificado y filtrado que captura el número, posición y ancho de las principales frecuencias que forman el habla. Estos sistemas forman la base de todos los sistemas, ya que contienen información sobre las propiedades físicas de los órganos vocales (tamaños, formas, peso, dureza, etc.) de cada individuo.

- **Sistemas de información prosódica** - evalúan, menos la información acústica, también la información prosódica que refleja los hábitos individuales del habla como el ritmo, la dinámica, la velocidad del habla, el estado de ánimo, la modulación de la voz, etc. Estas características son observadas y derivadas a partir de períodos de tiempo más largos.

- **Sistemas basados en características de alto nivel** - incorporan también características como el dialecto, un vocabulario a partir del cual se puede inferir una situación social, etc. Estas características sin embargo no proporcionan suficiente discriminación por lo tanto no se pueden utilizar con éxito sin otras características (acústicas y prosódicas).

- **Sistemas de tiempo real -** proporcionan información instantánea sobre la identidad del hablante; sin embargo se necesita un cierto retraso de varios segundos para la fiabilidad del sistema.

- **Sistemas *Offline* -** pueden utilizar algoritmos más complejos para proporcionar decisiones tan precisas como sean posibles mientras se utiliza una grabación de voz entera.

Las definiciones y conceptos arriba mencionados pueden ser extendidos a muchas más áreas y a varias ramas de la ciencia. Así, para una introducción más completa al problema de la identificación del hablante, por favor refiérase al ejemplo [6].

2.3 Identificación multinivel

Algunas aplicaciones (por ejemplo *AppStore*) pueden ser accedidas por múltiples usuarios de uno o de diferente tipo de dispositivos. Para este propósito se introduce una identificación multinivel.

La identificación multinivel se ocupa de diversos requisitos de confidencialidad (como diferentes permisos de acceso o de control) a través de múltiples modalidades biométricas en combinación con métodos de contraseña y PIN. Del grupo de modalidades biométricas, el reconocimiento facial y del habla se utiliza para identificar al usuario. Dependiendo del resultado de este proceso, podría permitir el acceso al sistema o a la aplicación o bien solicitar a la persona la entrada de otras credenciales - un nombre de usuario, contraseña o PIN. En el futuro, los sistemas pueden incorporar el reconocimiento de iris y reemplazar el método PIN o contraseña y, así, mejorar la seguridad.

3 Comunicación por comandos de voz

La comunicación con las máquinas a través de comandos de voz cae en un amplio campo de investigación denominado de reconocimiento automático de voz (**ASR -** *automatic speech recognition*).

Este término significa generalmente la transcripción automática (por una máquina) de un lenguaje hablado. La entrada es una señal de voz, generalmente en forma digital como una secuencia de números. La salida es un texto en forma de cadenas de palabras que existen en un vocabulario (como puede ser bastante extenso por lo general sólo se utiliza un vocabulario restringido en un dominio particular). Además, la cadena de salida sigue las reglas gramaticales regulares o bien patrones frecuentes que pueden ser observados en el lenguaje hablado (modelo estadístico de lenguaje). Esta tarea es a menudo denominada como el problema de voz al texto.

ASR debe ser diferenciada de la tarea de comprensión de lo que se dijo la cual opera en un nivel más alto (la entrada es un texto) y la rama de la ciencia que se ocupa de este problema se denomina inteligencia artificial.

Desde hace varias décadas ha habido un gran esfuerzo continuado en construir un sistema ASR que pudiera ser ampliamente utilizado, especialmente en las siguientes áreas: sistemas de recuperación de información, sistemas de diálogo, ayuda a las personas con discapacidad, etc. Pero no ha sido hasta hace poco que las aplicaciones orientadas a dominio han salido de los laboratorios. Actualmente, la tecnología y el conocimiento han hecho pasos cruciales y están surgiendo otras aplicaciones más sofisticadas, como los sistemas de dictado o incluso la transcripción automática del habla natural. Debido a que la tarea es tan compleja afrontando muchísimos obstáculos que éstos deben ser resueltos por los diferentes dominios de la ciencia. Los sistemas prácticos para el público en general deben operar en condiciones reales y en la mayoría de los casos en condiciones muy adversas (gran variabilidad de ruidos, dispositivos de grabación, lugares de uso, etc.), deben aceptar una gran variabilidad del lenguaje hablado (las reglas desaparecen), la variabilidad de los hablantes (en términos de una forma acústica) y grandes vocabularios, sólo por mencionar algunas de ellas. Además, un usuario en general requiere una respuesta inmediata del sistema (que trabaja en tiempo real), no está dispuesto a cambiar su hábito de hablar o de restringir su vocabulario y pierde rápidamente su paciencia si el sistema no funciona con una alta precisión.

Como la gama de posibles aplicaciones ASR es bastante amplia y también lo es la complejidad de los sistemas necesarios para la solución de tareas particulares (en realidad está creciendo aún más rápidamente con el aumento de requerimientos) existen varias clases de sistemas que de acuerdo a ellos se clasifican los sistemas ASR. El principal criterio relacionado con la clasificación ASR está basado en el tamaño del vocabulario, y diferenciamos:

- **Pequeño-** varias docenas o incluso centenares de palabras
- **Mediano-** varios centenares o incluso miles de palabras

- **Grande-** varias decenas o centenas de miles de palabras

Sin embargo, estos números cambian al tiempo que progresa la tecnología.

Los sistemas pueden ser independientes o dependientes del hablante lo que significa que si pueden o no trabajar por igual, independientemente de quién está hablando. Además, es importante si el sistema proporciona una respuesta inmediata o trabaja fuera de línea (*Offline*), por lo tanto tenemos sistemas en tiempo real o fuera de línea. A continuación es muy importante saber en qué forma estos sistemas esperan las muestras de voz para ser procesados. Así distinguimos:

- Reconocimiento de palabras aisladas – donde se espera una sola palabra de un diccionario
- Sistemas de dictado - requiere de una cadena de palabras que estén bien separadas para que puedan ser procesadas en bases individuales
- Habla contínua - que sin embargo sigue más o menos limitaciones graves dadas por algún tipo de gramática
- Habla natural- es la situación más difícil que abarca tanto el habla continua sin limitaciones o muy sueltas.

Finalmente, podemos clasificar los sistemas en base a lo que modelan las unidades del habla (fonemas, palabras, sílabas, frases, ...) y a qué modelos que utilizan, por ejemplo, modelos estadísticos como los *modelos ocultos de Markov* (**HMM**).

Una señal de voz es producida por los órganos vocales humanos y se observa a través de las vibraciones del aire. A diferencia de muchas otras, lleva la información de léxico (lo que realmente se dice) que se representa como una secuencia de diferentes sonidos acústicos. Un conjunto de sonidos básicos, llamados fonemas, se utilizan para construir palabras de un idioma en particular. Su número puede variar entre idiomas (por lo general de 40 a 60). Sin embargo, sus formas acústicas difieren de un hablante a otro y los fonemas vecinos se influyen entre sí (fenómeno de co-articulación).

Toda la información que es irrelevante o incluso que puede dificultar el reconocimiento correcto del habla debe ser suprimida antes del proceso de reconocimiento. Básicamente, sólo estamos interesados en la información léxica por lo que la restante, como la identidad del interlocutor, el estado de ánimo, el estado de salud actual, dialecto, deficiencias y hábitos del habla, debería ser suprimida. La tarea de un método de extracción del habla es la de recoger sólo la información necesaria que se ha de procesar por los siguientes bloques. Aproximadamente, existe un flujo de información léxica de 10b/s, mientras que la tasa de bits para la señal de voz es de aproximadamente 100 kb/s. Según lo anterior, el método de extracción se puede considerar como un compresor de velocidad de bits.

El objetivo es simular el sistema auditivo de los seres humanos, describirlo matemáticamente, simplificarlo para que su manipulación sea práctica y opcionalmente adaptarlo para un uso correcto y sencillo con los tipos de métodos de reconocimiento y clasificación seleccionados.

Existen varios métodos de extracción de características adecuados que, o bien simulan el proceso de producción del habla o que imitan el sistema auditivo humano (bandas críticas). Esto es así porque el sistema auditivo ha evolucionado para centrarse sólo en la información relevante, suprimiendo los ruidos y distorsiones ubícuos.

A lo largo de años de investigación se descubrió que excepto en el dominio del tiempo la información discriminativa más significativa para clasificar los fonemas entre sí se establece en el dominio de la frecuencia. Más precisamente, se encuentra en la posición y forma de las componentes de frecuencia dominantes. Para demostrarlo, en la Fig. 3.1 se presenta un espectro de frecuencias de la vocal "e", la envolvente de su magnitud con las frecuencias contenidas en ella (principales picos espectrales). La Fig. 3.2 es una representación en el dominio del tiempo de la vocal correspondiente 'e'.

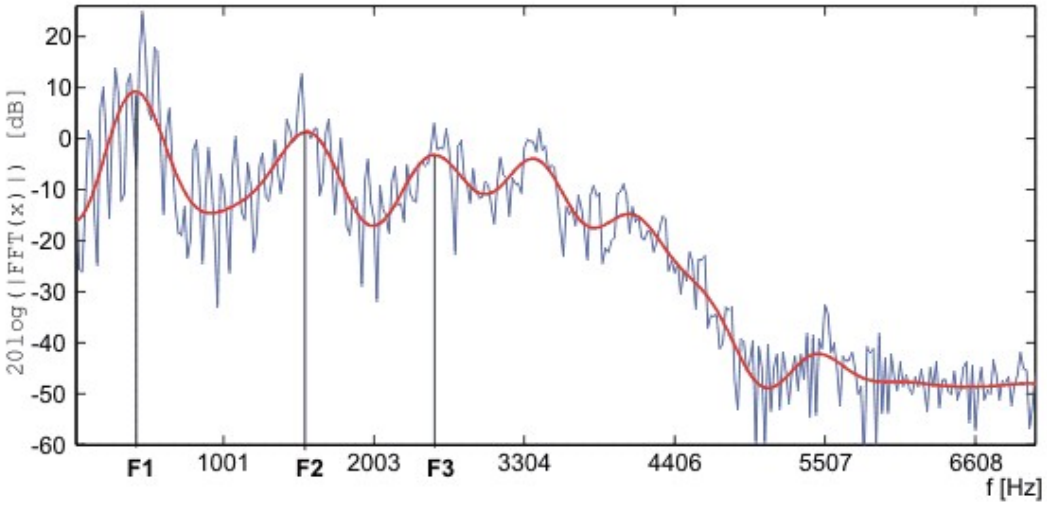


Figure 3.1. Espectro, componentes frecuenciales y envolvente de la magnitud de la vocal “e”.

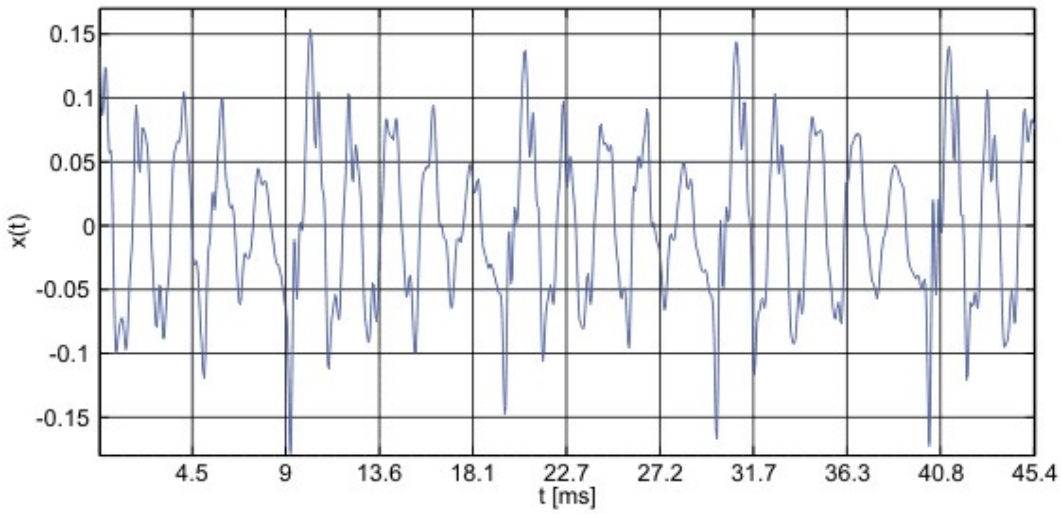


Figure 3.2. Señal temporal de la vocal “e”.

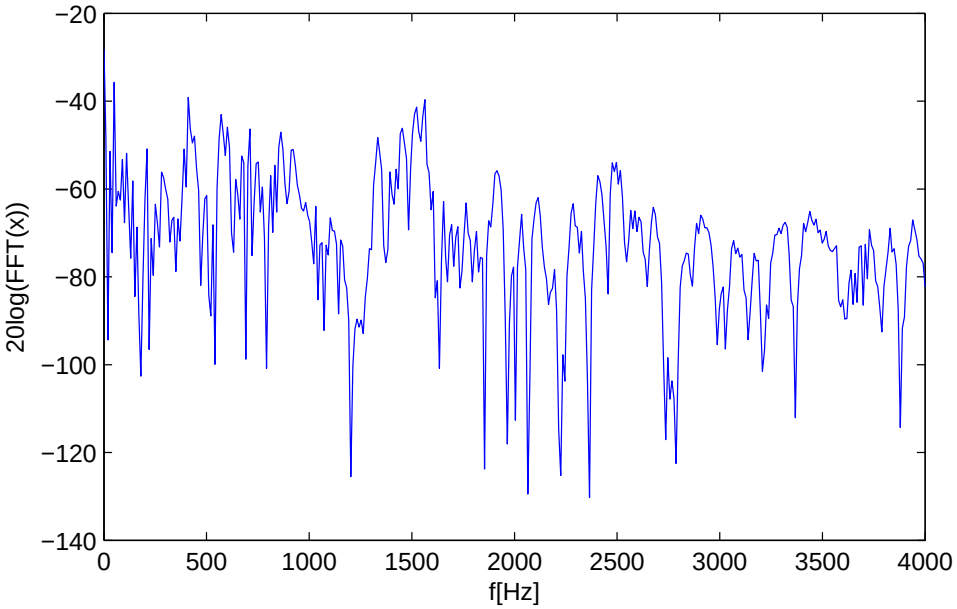


Figure 3.3. Espectro del fonema “t”.

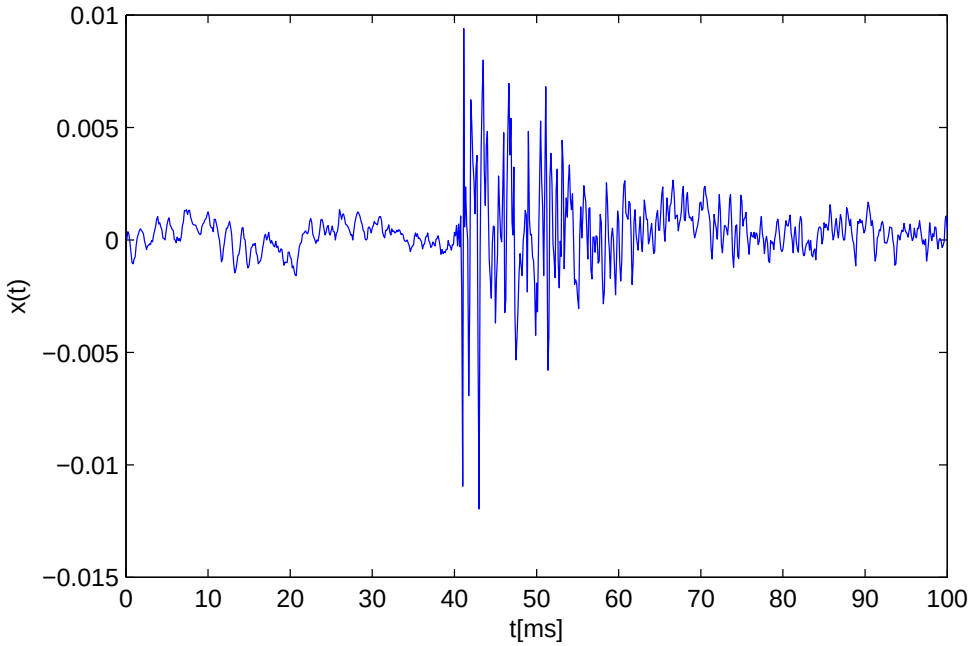


Figure 3.4. Señal temporal del fonema “t”.

Para ver las diferencias temporales y frecuenciales entre diferentes fonemas, en la Fig. 3.3 se representa el espectro y en la figura 3.4 la respuesta temporal de un fonema "t". En la siguiente tabla se enumeran por separado y de forma resumida las dos primeras componentes frecuenciales de las vocales en hombres y mujeres (cifras medias obtenidas de la población). Estas posiciones proporcionan una manera muy basta pero sencilla de cómo clasificar los fonemas en base a sus representaciones espectrales.

Table 3.1. Primeras dos componentes frecuenciales en vocales observadas en hombres y mujeres

vocal	Hombres		Mujeres	
	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [Hz]
a	730	1100	850	1200
e	530	1850	600	2350
i	400	2000	430	2500
o	570	850	590	900
u	440	1000	470	1150

Las características extraídas deberían ser capaces de estimar y discriminar las diferencias en las componentes frecuenciales. Por otro lado, deberían desestimar las variaciones que son naturales e inaudibles. En la siguiente tabla se enumeran las modificaciones espectrales audibles e inaudibles más relevantes.

Table 3.2. Modificaciones espectrales audibles e inaudibles

Tipo de modificación	
Audible	**Inaudible**
Número componentes frecuenciales	Inclinación general del espectro
Posición de las componentes frecuenciales	Frecuencias bajo la primera componente frecuencial
Ancho de banda de las componentes frecuenciales	Frecuencias por encima del tercer armónico
-	Filtrado de rechazo de banda estrecha

Además, la intensidad de las señales se percibe de forma no lineal que puede ser aproximada por una función logarítmica.

Desde el punto de vista del ruido, las características deberían ser insensibles a los ruidos aditivos y convolucionales. Por último, pero no menos importante, una buena característica debe ser fácil de implementar, matemáticamente tratable, y poseer una representación compacta. Por lo general, es beneficioso si las características son linealmente independientes entre sí ya que facilitan su posterior procesado.

3.1 Características acústicas para el reconocimiento de voz

En primer lugar hay que señalar que no hay ninguna característica que cumple completamente con todas las propiedades mencionadas anteriormente. Por tanto, la investigación está todavía activa y muchas características acústicas del habla han sido definidas, pero las más utilizadas son *Mel frequency cepstral coefficients* (**MFCC**) y *Perceptual Linear Prediction* (**PLP**). PLP y MFCC tratan de simular el sistema auditivo humano proporcionando un buen rendimiento en las tareas de reconocimiento de voz. Ambas características son capaces de capturar las posiciones y anchuras de formantes que son más perceptibles. A pesar de las similitudes obvias, difieren en los fenómenos psicoacústicos que abarcan.

MFCC aplica filtro de paso alto (supresión de la radiación de los labios), segmentación de voz por medio de una ventana de Hamming seguida por una conversión a espectro por DFT. A continuación, el espectro es deformado no linealmente en la escala Mel (escala psicoacústica que refleja la percepción humana) sobre la que se colocan ventanas triangulares, espaciadas por igual con un 50% de superposición para simular un banco de filtros (véase la Fig. 3.5). En la etapa final de cálculo, se aplican transformaciones logarítmicas y la **transformada discreta del coseno** (*DCT*). Por otra parte DCT suprime la dependencia entre las características.

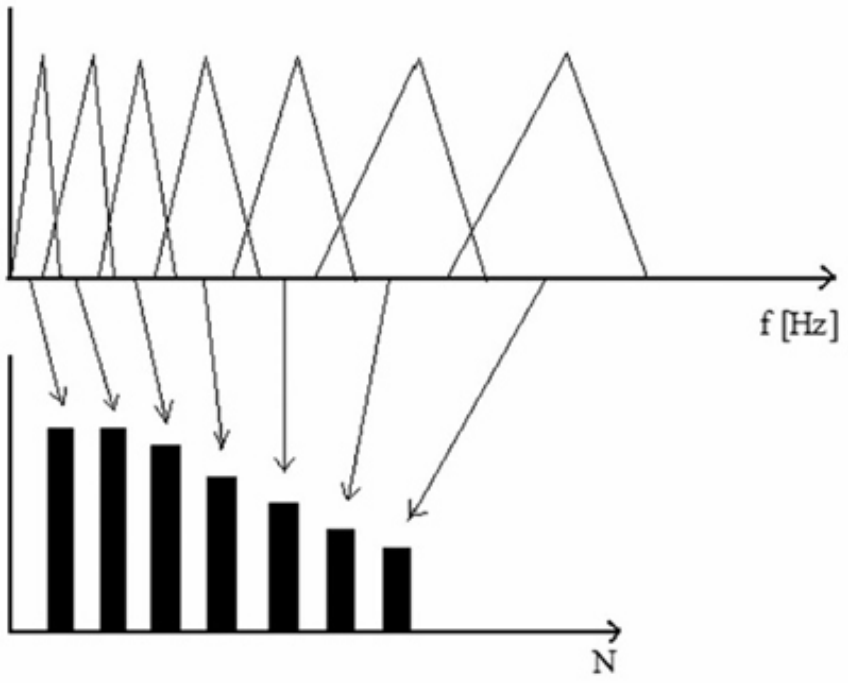


Figure 3.5. División espectral de banco de filtros utilizando la escala Mel y ventanas triangulares

Las características PLP difieren de las de MFCC en varios aspectos: el uso de la escala Bark, el suavizado y el muestreo del espectro bark en intervalos de 1 bark, ponderación de igual sonoridad, transformación de energías en sonoridad, cálculo de un modelo lineal de producción del habla, y su transformación en un cepstrum.

PLP aplica un procesado psicoacústico más complejo que MFCC. Sin embargo, por lo general ambos producen resultados similares en el reconocimiento de voz y en condiciones de laboratorio.

Características dinámicas y energía

El habla es básicamente una secuencia particular de sonidos diferentes por lo que tiene sentido medir y evaluar las transiciones adecuadas entre ellos. El método más común de hacerlo es a través de coeficientes delta y de aceleración a partir de las características acústicas en el tiempo. Éstos se calculan como diferencias entre dos tramas consecutivas o en el caso más general como una combinación lineal de las diferencias que cubren un período de tiempo más amplio. Además, se ha demostrado que la envolvente de la energía puede localizar la posición de las vocales de alta energía y las consonantes sordas de baja energía que aumenta la información general de discriminación. Así, la característica de energía (normalizada) se añade también a menudo a las características acústicas.

3.2 Métodos de reconocimiento del habla

Una vez que se han extraído las características adecuadas del habla, el habla se encuentra en forma de una secuencia de vectores de características, por ejemplo MFCC. A continuación, el proceso de reconocimiento, dicho de forma somera, toma muestras o modelos de unidades conocidas del habla (fonemas, palabras, etc., desde una base de datos de entrenamiento) y los compara con la muestra de voz desconocida, es decir, con sus vectores de características. A continuación, la muestra o modelo con la mayor coincidencia (puntuación) resulta ser la palabra reconocida. Las propiedades especiales de las señales de voz hacen que cada par de señales únicas de una misma palabra se diferencien por su longitud (alguien que hable más rápido o lento, etc.). Por otra parte, esta variabilidad en longitud no se distribuye uniformemente a lo largo del tiempo, por lo que algunas partes pueden durar más tiempo, mientras que otras pueden ser pronunciadas más rápido. Por tanto, el enfoque básico para resolver las diferencias de longitud mediante interpolación lineal o decimación no se puede aplicar con éxito aquí. Además, basándose en los modelos que utiliza el sistema, por lo general es necesario concatenar una secuencia de muestras o modelos para representar una determinada palabra o incluso una frase entera. Estos dos fenómenos (variabilidad no uniforme en la longitud y la concatenación de modelos) específicos del habla dan lugar al desarrollo de modelos de clasificación específicos. Actualmente, los métodos más comunes son la Deformación Dinámica de Tiempo (**DTW -** *Dynamic Time Warping*) y el Modelo Oculto de Markov (**HMM -** *Hidden Markov Model*). Sin embargo existen más modificaciones o incluso combinaciones que pueden ser elegibles en aplicaciones particulares. A continuación se proporciona una breve introducción a estos dos métodos.

DTW

DTW es la abreviatura de deformación dinámica de tiempo que es un método que acústicamente compara secuencias de dos expresiones características del habla (la de referencia y la prueba de uno). Se basa en la deformación no lineal de tiempo durante el proceso de comparación de modo que estas dos secuencias sean lo más parecidas posible (evaluadas por una medida acústica adecuada). Por lo tanto, se compensan las variaciones no lineales en longitud dentro de las palabras.

Para ello los vectores primero y último de las dos secuencias deben alinearse. Por lo tanto, este proceder requiere del conocimiento previo de las palabras de los límites, lo que puede ser una tarea difícil por sí misma mientras que se debe hacer de forma automática. Sin embargo existen modificaciones a DTW que relajan este requisito tan estricto.

El método intenta encontrar una correspondencia entre las dos secuencias de vectores de diferentes longitudes de modo que cada vector tiene un vector asociado de la otra secuencia para ser comparado. Esto significa que algunos vectores en momentos determinados pueden omitirse o que un vector puede corresponder a más de un vector de la otra secuencia. Por supuesto, este proceso no puede realizarse de manera arbitraria por lo que este mapeo debe seguir ciertas limitaciones lógicas es

decir los vectores inicial y final de una secuencia deben asignarse a sus homólogos en la segunda secuencia, las funciones de deformación deben ser no decrecientes, existe una discrepancia máxima permitida que este mapeo no lineal puede superar (por lo general los vectores cuyos índices son más del doble de la otra no pueden ser comparados, etc.). En el proceso de cálculo DTW se utilizan dos matrices (de distancia local y de distancia global). En la matriz local, se almacenan distancias acústicas entre los vectores de referencia y de características desconocidas. La matriz global se utiliza para calcular la trayectoria (*mapping*) y para acumular la distancia mínima a lo largo del camino óptimo. Por lo tanto, hay una distancia mínima y un camino óptimo en relación con cualquier elemento de la matriz global que conecta su posición al punto de inicio que está en la esquina inferior izquierda. Esta situación se ilustra en la Fig. 3.6. Por supuesto, hay limitaciones naturales en las direcciones, cómo desplazarse de un punto a los siguientes (no decreciente en las direcciones horizontales y verticales). Una vez que el proceso de búsqueda alcanza el punto final de la matriz global (esquina superior derecha) la comparación finaliza y se ha encontrado la distancia. Este proceso se repite para cada palabra de un diccionario y la palabra con la menor distancia global resulta ser la palabra desconocida. Como se puede ver, este método es elegible para el reconocimiento de palabras aisladas, o comandos.

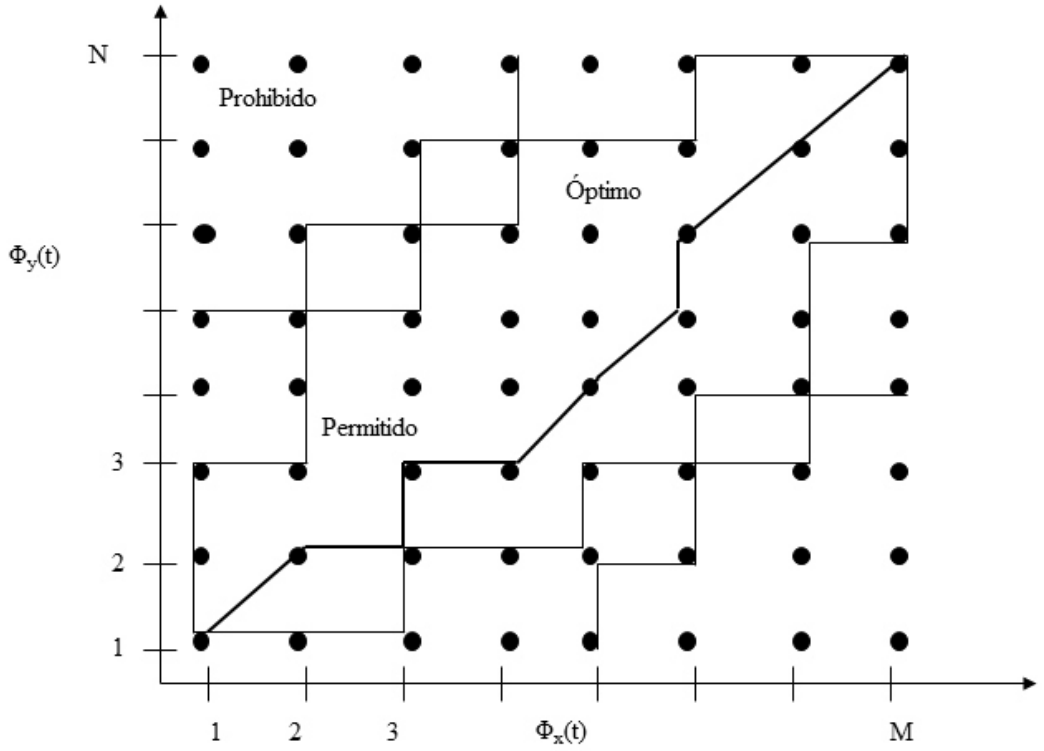


Figure 3.6. Matriz global con ruta óptima y limitaciones globales de ruta.

DTW tenía una posición importante en el reconocimiento de voz, especialmente para el problema del reconocimiento de palabras aisladas que depende del hablante. Sin embargo, como los requisitos iban creciendo, por ejemplo, reconocimiento de habla continua e independiente del hablante, fue perdiendo gradualmente su posición frente al método HMM.

Modelo Oculto de Markov (HMM)

El modelo oculto de Markov es una técnica estadística de modelado utilizada principalmente para el reconocimiento de voz que resuelve tanto la independencia del hablante y la concatenación de modelos básicos (para formar palabras, frases, oraciones e incluso cubre habla continua) de una manera matemáticamente elegante

Para cada unidad de voz seleccionada (fonemas, sílabas, palabras, ...) se crea un modelo HMM de una cierta estructura. Por lo general, todos los modelos comparten la misma estructura y difieren solamente por parámetros libres del modelo. En el proceso de entrenamiento sólo los parámetros libres se establecen utilizando la base de datos de entrenamiento. La base de datos de entrenamiento consta de locuciones de voz que están etiquetados de forma que se conoce exactamente qué se dijo. Los parámetros de los modelos de HMM se ajustan de tal manera que los modelos describen los datos de entrenamiento con la probabilidad más alta, también denominado criterio de máxima verosimilitud. Sin embargo, algunos sistemas que utilizan estrategias diferentes basadas en la separación máxima entre modelos o minimizando las tasas de error (criterios discriminativos) pueden proporcionar resultados más precisos.

Cada modelo se compone de varios estados que están conectados el uno con el otro. A cada conexión se le asocia una probabilidad de transición (p). Además, existe una probabilidad inicial (π) para cada estado que es la probabilidad con la que el modelo comienza en dicho estado.

Con dicho modelo, la probabilidad de la secuencia de estado $S_1, S_2, S_3, .. S_N$ vendrá dada por:

$$P(S_1, S_2, ..., S_{N-1}, S_N) = \pi(s_1) p_{12} ... p_{(N-1)N}$$

Por otra parte existe una probabilidad adicional asociada con cada estado y que es la probabilidad de observar un vector de características X dado un estado S, es decir $P(X/S)$. Entonces, la probabilidad de observación de secuencias de vectores de características $X_1, ..., X_M$, y estados $S_1, S_2, .. S_N$ es la siguiente:

$$P(S_1, X_1, S_2, X_2, ..., S_N, X_N) =$$
$$= \pi_1 P(X_1 / S_1) p_{12} P(X_2 / S_2) p_{23} ... P(X_{N-1} / S_{N-1}) p_{N-1,N} P(X_N / S_N)$$

La Fig. 3.7 muestra un ejemplo de un modelo HMM de 4 estados.

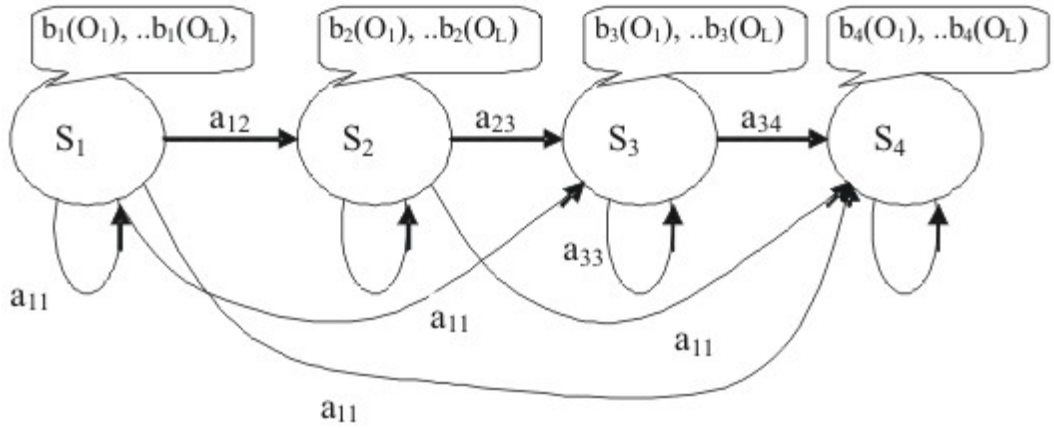


Figure 3.7. Modelo HMM de 4 estados.

A continuación, el proceso de reconocimiento calcula la probabilidad de una secuencia desconocida de todos los modelos HMM en un diccionario y elige aquella con la más alta probabilidad. La Fig.3.8 representa esquemáticamente el proceso.

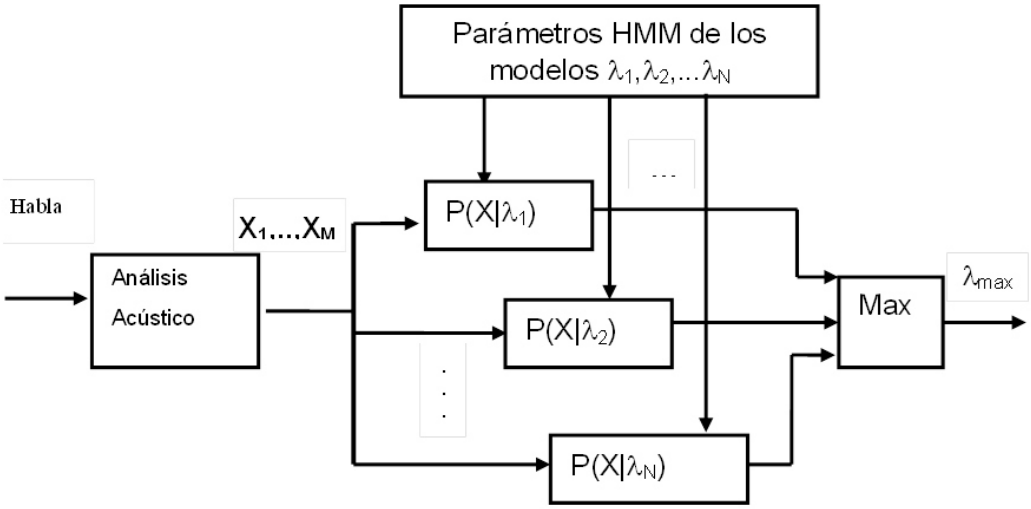


Figure 3.8. Proceso de reconocimiento de voz basado en HMM.

En la actualidad, los sistemas HMM más avanzados que consiguen las tasas de error de palabra (**WER -** *word error rates*) más bajos utilizan diferentes estrategias como la información mutua máxima (**MMI -** *Maximal Mutual Information*), error de clasificación mínimo (**MCE -** *Minimal classification error*), etc. Tanto MMI como MCE pertenecen al entrenamiento discriminativo. Por último, hay conexiones híbridas de gran éxito con otros métodos de clasificación como máquinas de vectores soporte y redes neuronales que proporcionan una WER aún más baja en los datos que no se conocen.

Una descripción más detallada de la tecnología de reconocimiento de voz y los métodos utilizados se pueden encontrar, por ejemplo, en [7].

4 Navegación por gestos

Hoy en día los gestos son una forma muy popular para el control de aplicaciones, siendo utilizados por muchas personas todos los días. En realidad, podemos utilizar gestos en nuestros dispositivos móviles, aplicaciones de ordenadores, consolas de juegos, etc. Se espera que la interacción por gestos esté disponible en casi todos los dispositivos en tan sólo unos pocos años. La gran popularidad de la navegación por gestos fuerza a los investigadores a mejorar estas tecnologías. Esta es una tendencia clara ya que el rendimiento del equipo ya no representa un cuello de botella para una navegación y control más naturales mediante el uso de gestos [2].

Hay dos áreas principales en el campo de reconocimiento de gestos en función del hardware de entrada. La primera área se centra en los gestos diseñados para dispositivos táctiles, tales como tabletas, teléfonos inteligentes o paneles táctiles. La otra área se ocupa de los gestos realizados frente a una cámara de 2 dimensiones o estereoscópica. Con algunos ajustes, se pueden aplicar los mismos algoritmos para reconocer gestos en ambas áreas. Además, los algoritmos para cámaras estereoscópicas pueden ser modificados para usar coordenadas 3D y realizar un seguimiento en las tres dimensiones [2].

4.1 Clasificación de gestos basada en varios criterios

Las aplicaciones controladas por gestos se vuelven más y más populares. Los gestos se pueden dividir en dos categorías básicas según la experiencia de usuario. Los gestos innatos se basan en la experiencia general de todos los usuarios como para mover un objeto a la derecha moviendo la mano hacia la derecha, coger un objeto con los dedos cerrados, etc. Naturalmente, el gesto innato puede verse afectado por los hábitos o la cultura. Mediante el uso de estos gestos no tenemos ninguna necesidad de enseñar al usuario a obtener experiencia para el control por gestos. Suponemos que el usuario puede controlar la aplicación de forma natural. El segundo grupo de gestos son gestos aprendidos. En este caso es necesario enseñar a los usuarios. Los gestos utilizados para la navegación en los sistemas y las aplicaciones tienen que ser fáciles, naturales y tienen que gastar un mínimo de energía humana.

Gestos estáticos

Los gestos estáticos no dependen de movimiento. Los gestos estáticos representan las formas de las extremidades que gesticulan, que llevan información significativa (Fig. 4.1).

Son tipos de gestos "sin movimiento" en los que no se necesita información sobre el movimiento. Por lo tanto no hay necesidad de investigar una secuencia de tramas, más que la trama de imagen real. Una secuencia de varias tramas que contienen diferentes gestos sólo es importante en el nivel más alto de comprensión.

Gestos contínuos

Los gestos contínuos sirven como base para la interacción con la aplicación. El gesto contínuo es un seguimiento prolongado del movimiento en el que no se reconoce ninguna pose específica pero el movimiento se utiliza para interactuar con la aplicación o el entorno virtual (Fig. 4.1).

El ejemplo típico de gestos continuos es el control de juegos usando una tecnología sin contacto tal como Microsoft Kinect, etc., donde el sistema traduce los cambios en las posturas a cambios en el vídeojuego, pero no hay movimiento específico para provocar una secuencia de cambios.

Gestos dinámicos

Los gestos dinámicos representan el movimiento que permite a los usuarios manipular directamente una aplicaciónde un objeto o de control (Fig. 4.1).

Los gestos dinámicos pueden definirse de dos maneras. La primera consiste en establecer algorítmicamente reglas y condiciones que el gesto realizado tiene que pasar para ser evaluado y reconocido éxito. La segunda forma es usar plantillas. Las plantillas son conjuntos de puntos que identifican la forma del gesto. Se utilizan

sofisticados algoritmos para evaluar la tasa de éxito cuando se comparan los movimientos del usuario con plantillas.

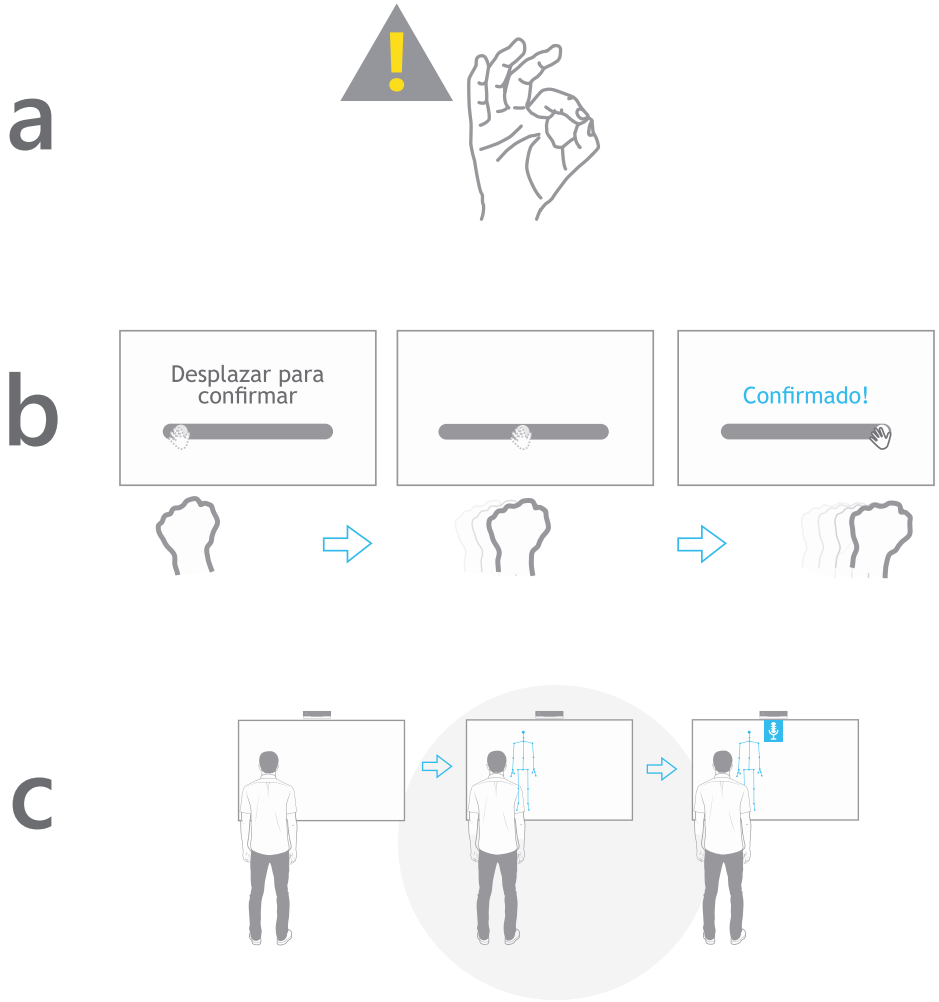


Fig. 4.1 – Tipos de gestos: a) gesto estático, b) gesto dinámico, c) gesto contínuo

4.2 Adquisición de imágenes

En las primeras investigaciones sobre el reconocimiento de gestos, el diseño de algoritmos se basó en la entrada **RGB** (*Red Green Blue*) de una cámara. Este enfoque aporta algunas ventajas, las cuales, sin embargo, se ven compensadas por desventajas. La imagen RGB se compone de canales de color existiendo muchos factores que influyen negativamente en la calidad de la imagen RGB.

El cambio de iluminación puede influir radicalmente en la calidad de la imagen. Un cambio en las condiciones de iluminación en una habitación cambia absolutamente el brillo y el contraste de la imagen. Esto hace que el reconocimiento mediante una imagen RGB sea incierto [2].

En últimos años, los investigadores empezaron a utilizar imágenes de profundidad para su procesado. Se trata de una imagen bidimensional que contiene información adicional - profundidad. La mayoría de los sensores de profundidad se basan en la emisión *infrarroja* (**IR**) (ejemplo del sensor Kinect en la Fig. 4.2). El sensor devuelve datos que representan la distancia de cada píxel en la imagen al sensor. Las distancias obtenidas se pueden transformar simplemente en una representación en escala de grises, de este modo, los datos de profundidad se representan como secuencia de vídeo en escala de grises. La obtención de datos de profundidad no requiere de condiciones especiales.

Las cámaras de profundidad ofrecen más ventajas frente a las RGB. La principal ventaja es que la cámara de profundidad hace uso de luz infrarroja y en la banda de infrarrojos no hay tanto ruido. Por tanto, el sensor de infrarrojos no es tan sensible a perturbaciones como el sensor RGB [2].

Fig. 4.2 – Imagen de una cámara de profundidad

La distancia de cada píxel al sensor se define en milímetros.

Si queremos convertir una imágen de profundidad en una imagen en escala de grises lo que necesitamos saber son las distancias mínima y máxima posibles.

$$intensity = 255 - \frac{255 - (d - d_{\min})}{d_{\max}}$$

donde, d es la distance real del pixel, d_{min} es la mínima distancia posible al sensor y d_{max} es la máxima distancia posible al sensor [2].

Para ayudar con la adquisición de datos de imagen, se procesan algunos elementos adicionales, como una pulsera luminosa, anillo o bola pequeña especial en la mano.

4.3 Metodología de diseño de gestos

La cuestión del reconocimiento de gestos no es trivial. Todas las limitaciones e inexactitudes que se producen al grabar imágenes utilizando diversos dispositivos tienen que ser tenidas en cuenta durante el proceso de diseño de gestos para el control de una aplicación o sistema. La metodología para el diseño de gestos tiene que cumplir varios requisitos: naturalidad, fácil de recordar y comodidad. La comodidad es el criterio más importante para la aplicación controlada por gestos.

Como en cualquier campo nuevo, también en el control por gestos, uno de los principales problemas en el proceso de adaptación es la estandarización existente. Cuando a diez personas al azar se les pide que hagan un gesto para iniciar la música, es muy probable que se propongan al menos 5 gestos diferentes. Por lo tanto, no es posible establecer un conjunto de gestos sólido que se adapte a todo el mundo. Esta es la razón por la cual todo fabricante que ofrezca el control por gestos utiliza un conjunto diferente de los gestos. Es difícil para un uso práctico, ya que el usuario tiene que aprender nuevos gestos para cualquier sistema nuevo. La existencia de al menos un conjunto estándar de gestos para controlar funciones básicas podría acelerar significativamente que los usuarios comenzaran a utilizar gestos de control así como su puesta en práctica. Hoy en día, se ofrecen algunos gestos de control para aparatos de televisión que se basan en la detección de una imagen vía una cámara 2D. Esto permite una fácil navegación a través del menú del televisor. A pesar de sus imperfecciones, el efecto educativo sobre las personas que poco a poco se están acostumbrando a este tipo de control es beneficioso.

Los conjuntos actuales de gestos se basan en dispositivos periféricos de entrada utilizados con los ordenadores. Básicamente tratan de quitar el dispositivo, pero mantienen los mismos patrones de uso. Con el fin de acercarse a una operación basada en gestos naturales (sin contacto) el concepto tiene que cambiar para que los conjuntos de gestos sean diseñados desde cero, como si no hubiera periféricos a excepción de sensores.

4.4 Técnicas de reconocimiento de gestos básicos de mano

El objetivo general del reconocimiento de gestos de mano es la interpretación de la posición, postura o movimiento de la mano.

El reconocimiento de gestos puede llevarse a cabo de dos maneras. Mediante el uso de un guante de datos que transforma las flexiones del cuerpo en información de movimiento, o mediante visión donde una cámara sirve de ojo humano para grabar las posiciones del cuerpo que serán extraídas por medio de procesado de imágenes [5].

Está claro que el primer método puede dar resultados precisos.

El enfoque basado en la visión, por el contrario, no posee requerimientos de equipo para el usuario final (a excepción de la cámara), por lo que es adecuado para aplicaciones generales [5].

El guante de datos es bastante incómodo en términos de comodidad del usuario. Además, el equipo necesario para emplear el método sería inaceptablemente costoso para la mayoría de los clientes estándar por lo que sólo es adecuado en usos especiales.

El inconveniente del método basado en la visión, sin embargo, radica en la complejidad algorítmica que requiere de una gran cantidad de tiempo y potencia de cálculo para extraer los movimientos del cuerpo [5].

Existen varios algoritmos disponibles que se centran en diferentes aspectos de una persona haciendo gestos (bajo diferentes supuestos). En general, se pueden dividir en dos categorías, los algoritmos basados en la apariencia y los basados en modelos 3D. El enfoque basado en modelos 3D compara los parámetros de entrada de una extremidad con la proyección 2D de un modelo 3D de la extremidad. El enfoque basado en la apariencia utiliza características de la imagen para modelar la apariencia visual de una extremidad y la compara con características de la imagen extraída de la entrada de vídeo [5].

En la sección 4.1 se definieron tres tipos de gestos. Se utiliza un clasificador general para detectar gestos estáticos (es decir, posturas). El clasificador nos dice qué gesto o postura fue reconocido. Sin embargo, los gestos dinámicos de mano tienen un aspecto temporal y requieren de técnicas que manejen esta dimensión, es decir, los modelos ocultos de Markov (HMM). La otra forma es utilizar modelo basado en movimiento.

Algunas de las técnicas utilizadas para el reconocimiento del gesto estático (y dinámico) de la mano son el algoritmo K-Means, KNN, SVM, el ya mencionado algoritmo HMM, el algoritmo DTW o las redes neuronales [3].

Técnicas utilizadas para el reconocimiento de gestos estáticos

Para un procesado más rápido de los datos en el proceso de reconocimiento de gestos, no toda el área de la imagen puede ser procesada.

Una vez es detectada automáticamente la mano, se asigna sólo el área alrededor de la mano que es luego procesada reduciendo así la carga de procesado. Las distancias obtenidas se convierten en una imagen en escala de grises para obtener el contorno de la mano.

A continuación, se describen dos métodos para el proceso de reconocimiento de gestos.

Defectos de convexidad

El primer paso consiste en separar la mano del fondo. La separación se puede obtener utilizando la información de profundidad, determinando qué píxeles de imagen pertenecen a mano. El segundo paso es detectar el contorno de la mano.

Las formas de muchos objetos complejos están bien caracterizadas por defectos de convexidad. La Figura. 4.3 ilustra el concepto de defecto de convexidad utilizando una imagen de una mano humana. El casco convexo se representa como una línea oscura alrededor de la mano, y las regiones etiquetadas de la A a la H son cada uno de los "defectos" en relación a ese casco. Como se puede ver, estos defectos de convexidad caracterizan también el estado de la mano. El algoritmo retorna las coordenadas de tres puntos, el *punto de inicio*, el *punto más profundo* y el *punto final* (Fig. 4.4) del defecto, en el que el punto más profundo es entendido como la distancia máxima entre el casco y el contorno de la mano.

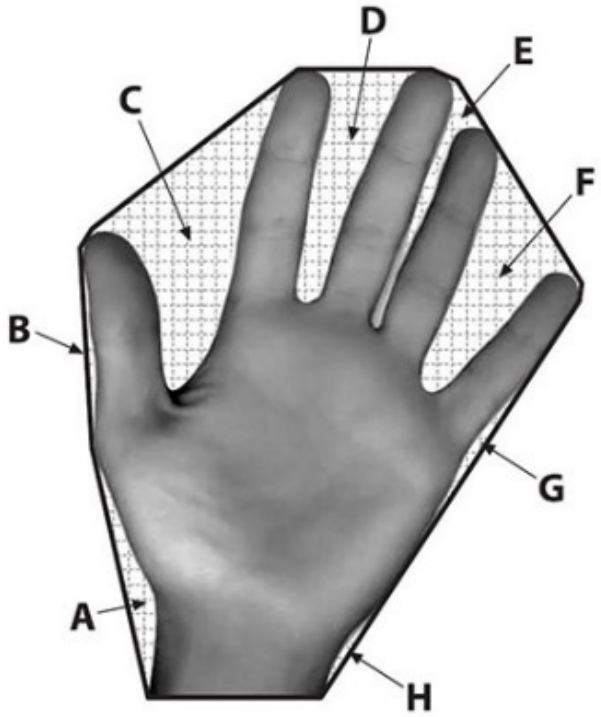


Fig. 4.3 – Casco convexo representado como una línea oscura alrededor de la mano

Fig. 4.4 – Las áreas marcadas son defectos de convexidad (arriba) y los tres puntos: punto de inicio (izquierda), punto más profundo (abajo) y punto final (superior)

El objetivo del algoritmo es encontrar el punto en el dedo que se encuentre más alejado del centro de la mano. El primer paso es eliminar todos los defectos cuya altura sea inferior a un valor especificado. Después se eliminan los defectos que tengan una distancia entre el punto de inicio y el punto máximo por encima de un valor especificado, definido dinámicamente (si dos puntos están demasiado lejos, significa que no pueden representar un dedo). A continuación también se eliminan los defectos que tienen una distancia entre el punto de inicio y el punto de profundidad menor a un valor especificado. Este valor cambia dinámicamente de acuerdo con el tamaño de la región en la que se detecta la mano (Fig. 4.5). El último paso consiste en eliminar todos los defectos que se producen por debajo de la muñeca.

Fig. 4.5 – Para ilustrar la lógica detrás del uso de triángulos para eliminar defectos innecesarios

Reconocimiento de gestos de la mano en base a partes

En el primer paso de este algoritmo, se obtienen y se convierten a imagen binaria los segmentos de datos que contienen las manos. El centro de la palma (Fig. 4.6) se calcula a partir del círculo interior y la suma de un punto sobre la envolvente del casco que pertenece a la mano y que tiene una distancia máxima desde el defecto encontrado.

Fig. 4.6 – Encontrando el centro de la palma

Todos los puntos del contorno de la mano se mapean en el eje X. El eje Y describe la distancia relativa de cada punto desde el centro de la palma. Los puntos mapeados crean una curva (Fig. 4.7).

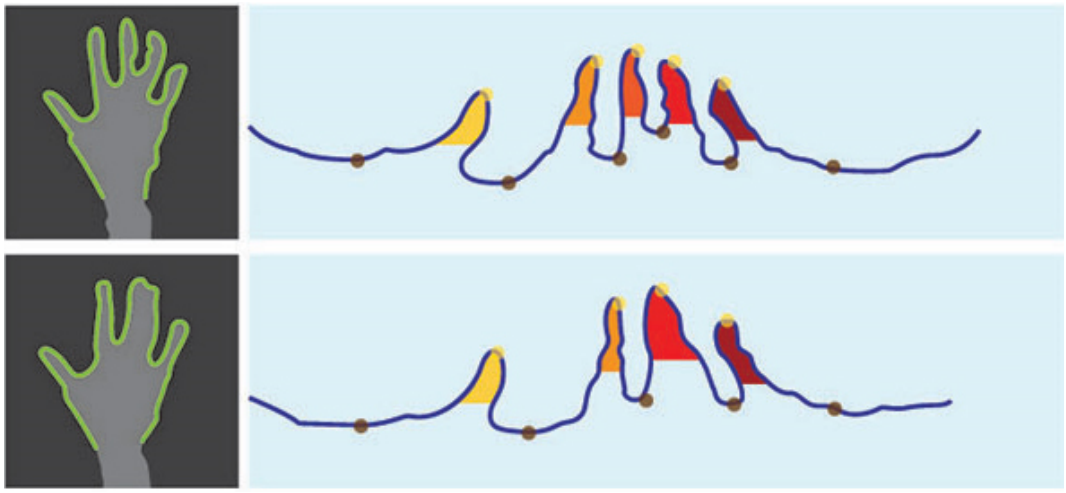


Fig. 4.7 – Representación de la curva

El siguiente paso del algoritmo es el análisis de la curva con el objetivo de encontrar un máximo local. Una vez finalizado el análisis de la curva, se inicia la extracción del dedo. Se compara la distancia de cada máximo para establecer un umbral. Cada dedo tiene un peso específico. Los dedos se detectan a partir de unas relaciones definidas (la relación se encuentra comparando el peso con unos valores dados y en base a ello se conoce el número de dedos, así, si el peso <1,5x peso medio - El segmento contiene un dedo, etc.). Este algoritmo funciona también para dedos unidos (Fig. 4.8).

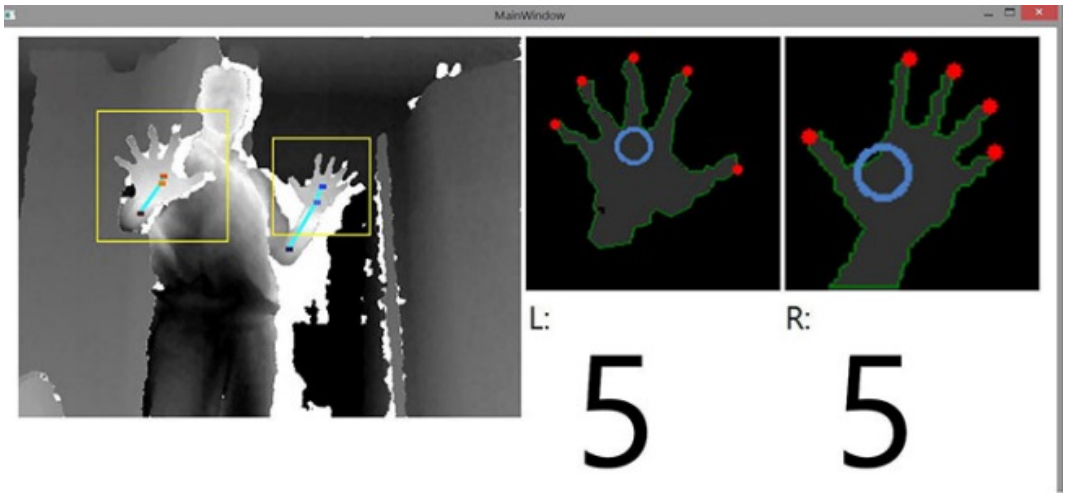


Fig. 4.8 –Dedos detectados

Técnicas utilizadas para el reconocimiento de gestos dinámicos

Los gestos dinámicos aseguran el acceso a contenidos privados de usuarios o al control de un sistema o una aplicación. Su uso se puede utilizar también como una contraseña. Las redes neuronales y los algoritmos genéticos se utilizaron sobre todo en los comienzos del reconocimiento de gestos.

Estos métodos tenían una tasa de reconocimiento aceptable, pero el mayor inconveniente era la cantidad de potencia de cálculo necesaria y el tiempo necesario para el entrenamiento de las redes neuronales, los cuales eran inaceptablemente elevados para aplicaciones prácticas.

Hoy en día, se utilizan nuevas técnicas para reconocer los gestos. Se han inventado algoritmos que no requieren de redes neuronales, por ejemplo, la búsqueda de la sección áurea, el algoritmo incremental de reconocimiento y modelos probabilísticos como el modelo oculto de Markov. Para aumentar la tasa de éxito de los algoritmos mencionados se puede utilizar *Machine learning*. Existen muchos enfoques de cómo hacer el reconocimiento de gestos. Los métodos HMM se han convertido en muy populares en los últimos años. La razón principal es que el método HMM es bien conocido y utilizado en muchas áreas.

El algoritmo propuesto por Kristensson y Denby [4], originalmente para los trazos del lápiz digital y dispositivos de pantalla táctil, se puede extender también a los gestos dinámicos.

Para esta propuesta, se define una plantilla como un conjunto de segmentos que describen un gesto. Es evidente que, dado un conjunto de gestos suficientemente distinguibles el uno del otro, el reconocimiento puede tener éxito después de tan sólo realizar una parte del gesto (Fig. 4.9).

Fig. 4.9 – Vista previa de la aplicación de prueba. La línea roja representa el gesto realizado. Las plantillas debajo de la imagen representan los gestos realizados más probables (de izquierda a derecha).

El uso de gestos se amplia con los gestos de deslizamiento. Este tipo de gesto resulta muy natural y cómodo. Los gestos de deslizamiento están diseñados para una navegación rápida y rutinaria en menú, programas y la galería. Poseen 4 direcciones para cada mano y un par de combinaciones de la mano izquierda-derecha. El método CDGR (*Circle Dynamic Gesture Recognition*) publicado en [1] se basa en la detección de la mano, la velocidad del movimiento y la distancia. Si la mano del hombre ejecuta un movimiento más rápido y el círculo ínterior sale del círculo exterior, el sistema procesa este movimiento y determina el gesto. El gesto viene dado por el ángulo del movimiento ejecutado desde el círculo central hacia el círculo exterior. Los gestos posibles son: deslizamiento a la izquierda, derecha, arriba y abajo. Los mismos gestos también se pueden hacer con ambas manos. El usuario también puede realizar gestos de *zoom in* y *zoom out* (Fig. 4.10).

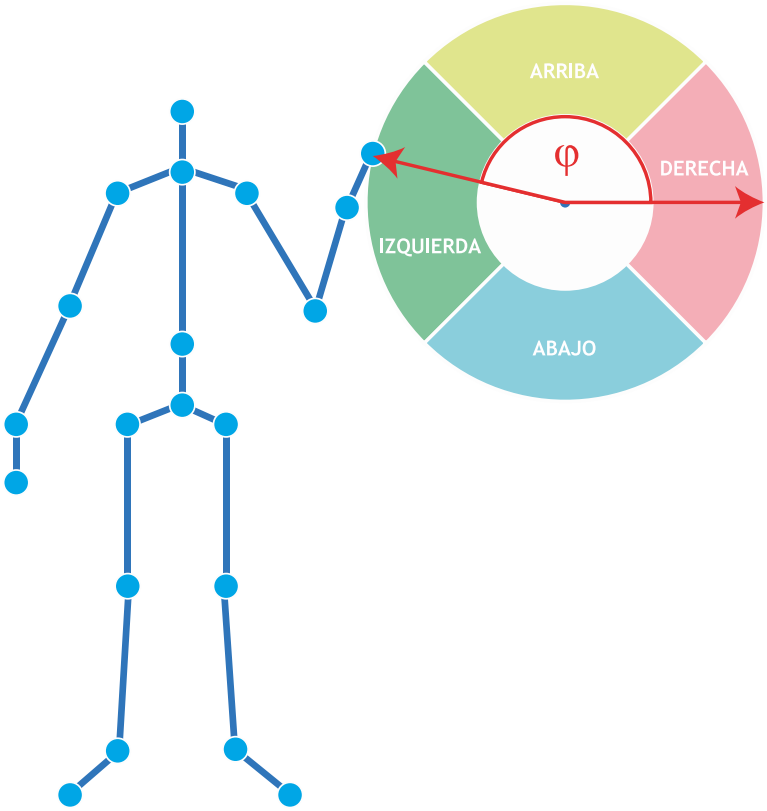


Fig. 4.10 – CDGR (*circle dynamic gesture recognition*)

4.5 Los gestos en la práctica

El reconocimiento de gestos se puede interpretar como una de las primeras formas posibles de cómo los ordenadores pueden entender a los seres humanos o al lenguaje corporal humano. Ello puede resultar en un gran logro cuando dispositivos de entrada como el teclado y el ratón se conviertan en minoría. Utilizando el reconocimiento de gestos, es posible señalar con el dedo a la pantalla del ordenador para que el cursor se mueva en consecuencia.

La navegación por gestos se vuelve cada vez más popular principalmente en los teléfonos inteligentes, tabletas y televisores modernos. Todo el mundo conoce el gesto de arrastrar el dedo para desplazarse hacia arriba/abajo o moverse hacia la izquierda/derecha. Ya hay aplicaciones que ofrecen un conjunto de gestos para la navegación en dispositivos multitáctiles, como el control de volumen, el bloqueo de pantalla, el control del reproductor de música, el tomar una captura de pantalla, el ir al menú de inicio, hacia atrás, las últimas aplicaciones y menú.

El área para el uso de gestos no está limitada, como ejemplo, se puede mencionar una aplicación de entrenamiento de tenis utilizando el reconocimiento de gestos 3D publicado por Cristian García Bauza y su equipo.

A modo de ejemplo, Samsung presentó el reconocimiento por gestos en sus TV insignia en el 2013. Consistía en un algoritmo de seguimiento de la mano con reconocimiento del gesto "click" realizado por el cierre de la palma de la mano. Para que el usuario pueda cambiar de canal o ajustar el volumen, primero él/ella tenía que levantar una mano frente la TV, a continuación, realizar el seguimiento del cursor que aparecía (un remanente de la era de la computadora) a cualquiera de los dos lados de la pantalla en la que aparecían los botones. A continuación, para aumentar el volumen, el usuario tenía que hacer "click" con su mano varias veces hasta alcanzar el volumen deseado mientras lo mantiene en su posición. Este proceder podría considerarse intuitivo, al menos por los usuarios con conocimientos informáticos, pero difícilmente podría servir como un ejemplo de gesto natural.

5 Diseño de GUI

La *interfaz gráfica de usuario* (**GUI**) representa una forma de interacción entre humanos y máquinas. La interfaz de usuario incluye hardware (físico) y los componentes de software (lógicos).

El usuario envía comandos o tareas a la máquina/sistema y la máquina responde mediante el cumplimiento de la orden o la tarea requerida. En el pasado, la GUI aceptaba la entrada a través de dispositivos tales como un teclado de ordenador y un ratón. Con la aparición de los dispositivos de pantalla táctil, la interfaz gráfica de usuario se adaptó y también se acepta la introducción por tacto de los dedos o de un lápiz óptico.

La tarea de crear una GUI adecuada para una aplicación multimodal (o cualquier otra) requiere de buenas prácticas, experiencia y de la realimentación por parte del usuario (hoy en día esta área de investigación se denomina experiencia de usuario). La interfaz gráfica de usuario para las aplicaciones multimodales debe permitir operar con una máquina de manera fácil, eficiente y fácil de usar. Se desea que el usuario necesite proporcionar un esfuerzo mínimo para configurar la entrada y también que la máquina requiera un tiempo de operación mínimo en proporcionar una salida. El tópico conectado a la *experiencia de usuario* (**UX**) es el diseño de la experiencia de usuario (**UXD** - *User experience design*).

6 Aplicaciones MMI

La investigación reciente de aplicaciones se centra en la interfaz multimodal natural y su integración en el sistema multimedia que utiliza diariamente. El uso de la interfaz multimodal no se limita sólo al sistema de televisión, sino que tiene muchas aplicaciones diferentes. La idea de futuras aplicaciones y sistemas no es sólo para controlar el contenido multimedia de la televisión y su acceso, sino para construir un sistema que sea consciente de los hábitos, preferencias e intereses de su usuario. Tal sistema puede convertirse en un conserje inteligente del hogar y puede proporcionar interconexiones entre los diversos servicios avanzados (de compras mientras se ve la TV, de voto o de portero digital).

Con el fin de hacer el mejor uso de la interfaz multimodal, es importante saber qué modalidad se debe utilizar para llevar a cabo cada una de las acciones. Algunas acciones siempre serán mejor ejecutadas mediante el uso de una modalidad diferente. Es decir. La entrada de texto sería difícil, consumiría tiempo y sería incómoda si se utilizan gestos, pero se podría realizar fácilmente y más rápidamente mediante reconocimiento de voz. Con esto en mente, se hace necesario introducir una plataforma de integración que proporcionará aplicaciones multimodales a las entradas solicitadas. La aplicación no necesita conocer la modalidad de origen, si no se requiere explícitamente.

Un simple ejemplo de un posible escenario: La aplicación MMI para el control de TV muestra una lista de usuarios identificados por los módulos de voz o de reconocimiento facial. Sólo a los usuarios de esta lista se les permite controlar el televisor mediante un conjunto predefinido de gestos, comandos de voz u otra modalidad. Cuando el usuario sale de la habitación, se le retira automáticamente de la lista. La otra parte de la aplicación muestra los canales recomendados. Dependiendo de las preferencias de visualización de usuario, el sistema muestra las recomendaciones que mejor se adapten a todos los usuarios en frente de la televisión. Utilizando gestos de desplazamiento, un usuario es capaz de navegar en esta lista, reproducir o detener el vídeo. Algunos de los canales recomendados están bloqueados. Esto significa, que a los usuarios sin el permiso no se les permite ver dicho contenido hasta que entren el código secreto. Para introducir el código secreto aplicamos los gestos dinámicos.