REVISIONES

Síntesis visual del habla

Visual synthesis of speech

Y. Blanco, A. Villanueva, R. Cabeza

RESUMEN

En pacientes cuya discapacidad motora ha alcanzado un grado elevado, los ojos pueden llegar a ser el único modo de comunicación.

Con tecnología adecuada se pueden interpretar los movimientos oculares y aumentar sus posibilidades de comunicación con su entorno permitiéndole el uso de sintetizadores de habla.

Un sistema de estas características deberá disponer de un sintetizador de habla propiamente, de un programa de interacción con el usuario para la construcción de texto y un sistema de interpretación de la mirada. De esta forma se consigue que el usuario maneje el sistema únicamente con sus ojos.

En esta revisión se presenta el estado del arte de los tres módulos que componen un sistema de este tipo y finalmente de forma breve se presenta el sistema de síntesis visual del habla (SiVHa) que se está desarrollando en la Universidad Pública de Navarra.

Palabras clave: Síntesis del habla. Seguimiento ocular. Discapacitado.

ABSTRACT

The eyes can come to be the sole tool of communication for highly disabled patients.

With the appropriate technology it is possible to successfully interpret eye movements, increasing the possibilities of patient communication with the use of speech synthesisers.

A system of these characteristics will have to include a speech synthesiser, an interface for the user to construct the text and a method of gaze interpretation. In this way a situation will be achieved in which the user will manage the system solely with his eyes.

This review sets out the state of the art of the three modules that make up a system of this type, and finally it introduces the speech synthesis system (Síntesis Visual del Habla [SiVHa]), which is being developed in the Public University of Navarra.

Key words: Speech Synthesis. Eye tracking. Disability.

ANALES Sis San Navarra 2000; 23 (1): 41-66.

Departamento de Ingeniería Eléctrica y Electrónica. Universidad Pública de Navarra.

Aceptado para su publicación el 26 de julio de 1999.

Correspondencia
Rafael Cabeza
Dpto. Ingeniería Eléctrica y Electrónica
Universidad Pública de Navarra
Campus de Arrosadía
31006 Pamplona

INTRODUCCIÓN

Desde la introducción de los ordenadores en la sociedad a mediados del siglo XX, éstos han sufrido una evolución considerable. Por una parte, los avances tecnológicos han dado lugar a máquinas más potentes y, por otro lado, se han desarrollado herramientas especializadas de software que se han extendido a todos los sectores de la sociedad: medicina, industria, ocio, servicios, enseñanza, etc. En el diseño de estas herramientas el usuario tiene cada vez un papel más importante y activo. El objetivo que se persigue en el desarrollo de cualquier herramienta es conseguir llegar a toda la sociedad sin necesidad de una preparación específica y tratando de facilitar al máximo la utilización por parte del usuario.

Sin embargo, el uso del ordenador requiere un control motor mínimamente preciso y cierta habilidad para la utilización del teclado y el ratón, de modo que un sector importante de la sociedad queda excluido. Las personas con el sistema motor dañado quizá sean las que más necesidad puedan tener de utilizar un ordenador debido a que podría mejorar en gran medida su calidad de vida pero, sin embargo, tienen muy limitado el acceso a su uso. Distonías, tetraplejias, tumores medulares superiores, esclerosis múltiples, esclerosis laterales amiotróficas, atáxias de Friedrich, son algunas de las patologías que dificultan o impiden la utilización del ordenador.

Para todas estas personas el ordenador podría ser un elemento de conexión con el entorno de vital importancia. El ordenador puede facilitar el control de tareas diarias en el hogar (control de luces, puertas, persianas, teléfono, televisión...); puede ser un elemento de comunicación para aquellas personas que además han perdido el habla; puede ser un elemento de ocio (acceso a lectura, Internet, juegos...) e incluso puede ser una herramienta de trabajo.

Para interactuar con el ordenador se han desarrollado herramientas alternativas que son muy dependientes del tipo de enfermedad y del usuario para el que han sido desarrolladas. Existen ratones controlados por la boca, mediante movimientos de cabeza, con soplidos, etc. Todos estos métodos exigen cierta movilidad y pueden ser engorrosos. Una línea interesante de investigación es la utilización de los ojos como medio de comunicación con el ordenador, por ser la mirada una característica que incluso en fases muy avanzadas de estas enfermedades permanece intacta.

Nuestra investigación está orientada a la integración de personas con este tipo de discapacidades en la sociedad, haciendo uso del ordenador controlado por la vista. Nuestro primer objetivo ha sido facilitar la comunicación a aquellas personas que han perdido el habla y para ello se está trabajando en un sistema de síntesis de habla agradable, rápido, de fácil manejo, que sea capaz de aprender las preferencias de usuario y adaptarse a sus habilidades y que pueda ser manejado fácilmente haciendo uso únicamente de la mirada.

Un sistema de este tipo contempla tres partes diferenciadas: el control del ordenador con la vista (eyetrack), una aplicación informática para generar texto (interfaz), y un sintetizador de habla de alta calidad (conversor texto voz).

En este artículo se presenta el estado del arte de los tres módulos por separado, ya que no se ha encontrado ningún sistema completo con la misma funcionalidad. Finalmente se presenta la solución aportada en nuestra investigación, su estado actual y las líneas futuras.

EYETRACK

Para controlar el ordenador con la mirada es necesario determinar dónde está mirando el usuario en cada instante. Es decir, es necesario seguir el movimiento de los ojos en todo momento y saber relacionar la posición de los ojos con un punto en la pantalla. Además de esto, una vez que se determina dónde está mirando el usuario, es necesario adivinar sus intenciones: si quiere seleccionar algo, si está buscando algún objeto determinado, o simplemente si está pensando con la mirada perdida en la pantalla.

El primer módulo de *eyetrack* se encarga únicamente de la no fácil tarea de determinar a qué punto de la pantalla está mirando el usuario. Los procedimientos para determinar las intenciones del usua-

rio o definir unas reglas de selección, descanso, etc. los controlará el segundo módulo, la denominada interfaz de usuario.

Determinar a qué punto de la pantalla está mirando el usuario una vez que sabemos la posición de los ojos es relativamente sencillo mediante un calibrado. El usuario, al inicio de cualquier sesión, deberá mirar a los puntos en pantalla que se exija, para poder asociar posiciones de pantalla y posiciones oculares. No es necesario que se realice esta operación para todos los puntos del monitor ya que, con el uso de puntos estratégicos y funciones polinómicas de superficie, pueden interpolarse el resto de puntos de la pantalla, limitando el proceso de calibrado a cuestión de segundos. La tarea más difícil es determinar la posición de los ojos en todo momento. En cuanto a los principios básicos, las técnicas de seguimiento de la mirada son las mismas hoy que hace 20 años. Young y Sheena, en un artículo publicado en 1975, recogieron las diversas técnicas conocidas para la medida de los movimientos oculares[1]. Hemos extraído de este artículo lo que consideramos más relevante, añadiendo los sistemas comerciales actuales que hacen uso de las distintas técnicas y sus especificaciones.

Tipos de movimientos oculares

Los movimientos oculares son rotaciones en torno a un eje horizontal, rotaciones en torno a un eje vertical y torsiones en torno al eje de la mirada. Las distintas combinaciones de estos movimientos básicos dan lugar a los movimientos característicos de los ojos.

Las sacudidas son movimientos rápidos, conjugados, para cambiar la fijación de un punto a otro de forma voluntaria. Cada sacudida se caracteriza por una aceleración inicial y deceleración final que puede alcanzar los 40000 grados/segundo^2 y una velocidad de pico durante el movimiento que varía con la amplitud de la sacudida y puede llegar a ser de 400 ó 600 grados/seg. La duración de la sacudida varía asimismo con su magnitud entre 30 y 120 mseg[2]. Las sacudidas observadas generalmente en una búsqueda son del rango de 1 a 40 grados y el movimiento de cabeza acompaña a la sacudida cuando el cambio del punto de mira excede los 30 grados. En respuesta a un estímulo visual las sacudidas presentan una latencia de entre 100 y 300 mseg. Las sacudidas verticales u oblicuas pueden llevar asociado un componente de torsión debido a la superposición de la acción de los 6 músculos extraoculares. El propósito del sistema de sacudidas parece ser fijar la imagen del objeto en la fóvea, una región de alta precisión en la retina que abarca entre 0,6 y 1 grado del ángulo visual. Hay un retraso o periodo refractario entre dos sacudidas consecutivas de duración entre 100 y 200 mseg. El umbral de visión es significativamente elevado durante la sacudida así como en instantes previos a la misma[3-6].

Los seguimientos son movimientos conjugados de los ojos para seguir un objeto visual en movimiento en un rango de 1 a 30 grados por segundo. Los movimientos de seguimiento son suaves y parecen estabilizar parcialmente la imagen del objeto en movimiento dentro de la retina. Este tipo de movimiento está limitado tanto en velocidad como en aceleración[7,8].

Las compensaciones son movimientos suaves similares a los seguimientos, que compensan los movimientos activos o pasivos tanto de la cabeza como del tronco. Estos movimientos estabilizan la imagen de los objetos estáticos en la retina durante el movimiento de la cabeza o el cuerpo.

Las divergencias son movimientos de los dos ojos en direcciones opuestas para difuminar la imagen de los objetos cercanos o lejanos. Son movimientos más lentos y más suaves que los movimientos de ojos conjugados, son no predecibles y alcanzan una velocidad máxima de 10 grados por segundo sobre un rango de unos 15 grados. Estos movimientos son estimulados por errores de enfoque y disparidad binocular[9,10].

Las fijaciones o miniaturas incluyen una variedad de movimientos de amplitud generalmente inferior a un grado que ocurren cuando se pretende fijar la mirada sobre un objeto. A continuación se detallan las más conocidas. Los *Drifts* son movimientos lentos y aleatorios del ojo, fuera del punto de fijación, a velocidades de tan sólo unos pocos minutos de arco por segundo. Los *Flicks* o *microsacudidas*, son movimientos

rápidos, pequeños, de la misma naturaleza que las sacudidas, de magnitudes máximas de un grado y con intervalos mínimos de unos 30 ms, que generalmente redireccionan el ojo hacia la posición necesaria para la fijación del objeto[11]. Adicionalmente individuos normales presentan un *tremor* de alta frecuencia entre 30 y 150 Hz, con amplitudes aproximadas de 10 segundos de arco en los 70 Hz. Debido a la presencia de estos movimientos de fijación, una precisión de 0,5-1 grado es normalmente suficiente en tareas de seguimiento de ojos diseñadas para mostrar qué parte del campo visual está siendo observada.

El nystagmus optocinético o de tren es un patrón de movimiento ocular en forma de diente de sierra que se produce ante un campo visual en movimiento con patrones repetitivos. Consiste en una fase lenta en la cual el ojo se fija en una porción del campo en movimiento y lo sigue con un movimiento lento y una fase rápida, o sacudida de retorno, en la que el ojo se fija en una nueva porción del campo. El tiempo mínimo entre las fases rápidas es de aproximadamente 0,2 segundos, resultando en una frecuencia máxima de 5 Hz. La amplitud es variable, generalmente entre 1 y 10 grados. Si existe un punto de fijación estático en el campo visual la respuesta de nystagmus puede verse reducida a una fracción de grado y no ser observable directamente.

El nystagmus vestibular es un movimiento oscilatorio del ojo, similar en apariencia al nystagmus optocinético, que contiene una fase lenta y una sacudida rápida de retorno. Se atribuye a la estimulación de los canales semicirculares durante la rotación de la cabeza respecto al espacio inercial. Una rotación de la cabeza en el sentido contrario al de las agujas del reloj sobre un eje vertical induce una estabilización de la imagen mediante un movimiento lento de los ojos en el sentido de las agujas del reloj. Si el movimiento de la cabeza persiste, los ojos saltan hacia atrás rápidamente para recoger otro punto y repetir el patrón de diente de sierra.

El nystagmus espontáneo o de mirada es una anomalía de nystagmus asociada a desórdenes neurológicos. Puede ser lo suficientemente amplio para poder ser apreciado directamente o inferior a 1 grado, requiriendo entonces grabaciones para su percepción. Este tipo de nystagmus se observa únicamente cuando el paciente mira en cierta dirección. El nystagmus puede ser bien asimétrico, conteniendo una fase lenta y una rápida, o bien pendular, mostrando oscilaciones de alta frecuencia de entre 4 y 10 Hz. Algunos de estos movimientos anormales son muy pequeños para ser observados en un examen clínico y son muy importantes desde el punto de vista del diagnóstico.

Los movimientos de torsión son movimientos de rotación del ojo en torno al eje de la mirada y están generalmente limitados a ángulos inferiores a 10 grados. Este tipo de movimientos puede ser estimulado por nystagmus optocinético de rotación o bien por respuestas vestibulares.

CARACTERÍSTICAS FÍSICAS DEL OJO UTILIZADAS EN LA MEDIDA DE SU MOVIMIENTO

Los ojos presentan una serie de características geométricas, ópticas y eléctricas que han sido utilizadas por diversos métodos de seguimiento de su movimiento (Fig. 1). He aquí las más importantes a tener en cuenta.

Potencial córnea-retina

Existe una diferencia de potencial de hasta 1 mV entre la córnea y la retina del ojo. Este potencial tiene importantes variaciones diurnas y con el nivel de adaptación a la luz [12]. Por ello, antes de su utilización para posibles medidas de movimiento, es necesario un periodo de entre 30 y 60 minutos para la adaptación del usuario a las condiciones lumínicas.

Debido a que el campo eléctrico no está alineado con el eje óptico, una torsión del ojo produce un cambio de potencial que puede inducir a error.

Impedancia eléctrica

La impedancia medida entre electrodos colocados en los cantos exteriores de los dos ojos varía con la posición de los ojos. La variación de la resistencia se asocia tanto a la naturaleza anisotrópica de las características eléctricas de los tejidos oculares como a la forma no esférica del

ojo, que supone un cambio en el camino resistivo entre los dos ojos cuando se produce un cambio de posición [13,14].

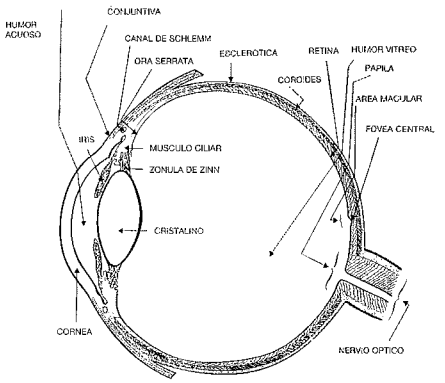


Figura 1. Sección esquemática del ojo.

Curvatura de la córnea

La córnea, unida a la esclerótica en la parte frontal del ojo y centrada cercana al eje óptico, tiene un radio de curvatura menor que el del propio ojo: la curvatura nominal de la córnea para un adulto es aproximadamente de 8 mm para un radio de ojo de 13,3 mm.

Estas diferencias son la base de un número importante de métodos de medidas de movimiento de ojos. Se han utilizado desde transductores de presión sobre los párpados hasta lentillas ajustadas a la córnea que hacen uso de otro tipo de medida para determinar su posición. Pero debido a que la propia córnea se puede desplazar sobre la esclerótica, para obtener una medida fiable este tipo de lentes deben abarcar toda la esclerótica y la propia córnea.

Reflexiones en la córnea

La superficie frontal de la córnea, a pesar de no ser una superficie óptica perfecta, se aproxima a una sección de esfera de 25 grados. Como en el caso de un espejo convexo, las reflexiones de un objeto brillante en esta superficie forman una imagen virtual detrás de la superficie que puede ser grabada. La posición de la reflexión, comúnmente vista como un brillo en el ojo, es función de la posición del ojo. La rotación del ojo sobre su centro produce una translación relativa y una rotación de la córnea, formando la base de una importante clase de instrumentos de medida de los movimientos oculares, conocidos como sistemas de reflexión en la córnea.

Reflexiones de otras curvaturas ópticas en las imágenes de Purkinje del ojo

A pesar de que la reflexión más brillante de una luz incidente es la anteriormente citada, procedente de la superficie de la córnea, la luz también se refleja en cada capa del ojo donde hay un cambio del índice de refracción. Las reflexiones se producen por lo tanto también en la superficie trasera de la córnea y en las superficies frontal y trasera del cristalino. Estas cuatro se conocen como las imágenes de Purkinje. La reflexiones más visibles son la primera y la cuarta, y la medida de su posición relativa representa una técnica de medida activa de la orientación espacial del ojo, independiente de su relación con la posición de la cabeza.

Limbus

El iris, normalmente visible y claramente distinguible respecto a la esclerótica, es la base para algunos métodos de medida del ángulo de la mirada. La posición del límite entre el iris y la esclerótica (limbus) puede ser medida respecto a la cabeza. La relación entre el iris oscuro y la esclerótica brillante en ambos lados izquierdo y derecho del ojo puede ser medida directamente, mediante sensores, o indirectamente, en una imagen del ojo. Esta relación está inequívocamente asociada a la posición horizontal del ojo.

Pupila

La pupila puede distinguirse del iris por la diferencia en sus índices de reflexión. Se puede conseguir que la pupila aparezca mucho más oscura que el iris bajo la mayor parte de condiciones lumínicas cuando la mayoría de la luz incidente no viene del mismo eje de medida. Por otro lado, se puede conseguir que la pupila aparezca

muy brillante cuando la mayoría de la luz incidente entra a lo largo del eje óptico y se refleja atrás en la retina. En ambos casos la pupila puede separarse de su alrededor ópticamente. Esto puede ser especialmente acentuado con el uso de luz infrarroja, que será prácticamente absorbida en su totalidad cuando entra en el ojo, apareciendo de esta manera la pupila mucho más oscura. La pupila normalmente varía entre 2 y 8 mm de diámetro en adultos. A pesar de que su forma real es algo elíptica, puede aproximarse por un círculo y es fácil encontrar su centro. La pupila aparece elíptica si se mira desde cualquier eje diferente al eje óptico y su excentricidad podría servir de base para medida del ángulo del ojo.

Otras características ópticas y no ópticas

Además de la pupila y el iris, pueden utilizarse otras características del ojo. Las propias venas y arterias de la esclerótica pueden identificarse y utilizarse para la medida de movimientos oculares. Asimismo los vasos sanguíneos de la retina pueden ser identificados y seguidos y constituyen una medida precisa de la posición sobre la retina en que se forma la imagen, esto es, donde el objeto se está viendo, lo que implica el punto exacto de mira del ojo.

También se han colocado algunas marcas artificiales en el ojo para poder hacer un seguimiento de sus movimientos. Entre los materiales utilizados para este fin destacan mercurio, carbón, membrana de huevo e incluso una pieza de metal incrustada en la esclerótica que permite un seguimiento magnético de la posición del ojo.

TÉCNICAS DE MEDIDA DE MOVIMIENTOS OCULARES

A continuación se describen las técnicas más extendidas de medida de movimientos oculares, los principios básicos, su desarrollo práctico, la evaluación y los sistemas comerciales con sus especificaciones.

Electro-oculografía

Principio

A principios del siglo XX, se descubrió que la posición del ojo podía medirse colocando electrodos superficiales alrededor del mismo y midiendo las diferencias de potencial[15,16] (Fig. 2). La fuente de la energía eléctrica es el potencial córneo-retinal: la córnea se mantiene entre 0,4 y 1 mV más positiva que la retina; esto es atribuible a la velocidad metabólica más alta de la retina. Al girar el ojo, el dipolo electrostático gira solidario a él, y consecuentemente la diferencia de potencial en un plano normal al eje principal varía teóricamente como el seno del ángulo de desviación. Por supuesto la naturaleza no homogénea del medio conductor ocasiona grandes desviaciones de los valores teóricos. Los potenciales medidos son pequeños, del orden de 15 a 200 mV con sensibilidades nominales del orden de 20 mV/grado de movimiento ocular[17].

Desarrollo

Las señales asociadas al potencial córneo-retinal son difíciles de detectar debido a la presencia de potenciales de acción musculares de mayor amplitud recogidos por los mismos electrodos[18]. La presencia de interferencias electrónicas externas puede llegar a ser también un problema si no se toman las precauciones adecuadas.

Normalmente la adquisición de la señal continua que es necesaria para conocer la posición del ojo se conoce con el nombre de electro-oculografía (EOG) y la adquisición de la señal alterna, que es asimismo útil para la medición de movimientos oculares, incluyendo las fases rápida y lenta del nystagmus, se conoce con el nombre de electronystagmografía (ENG).

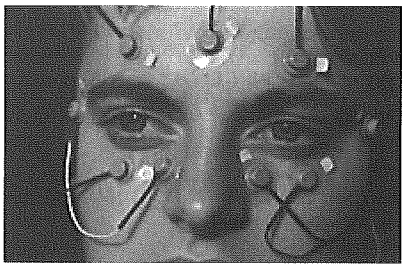


Figura 2. Electro-oculografía (EOG).

Los movimientos conjugados horizontales se graban mediante electrodos super-

ficiales colocados en los cantos exteriores de los ojos. Cuanto más atrás se coloquen los electrodos más disminuye el ruido debido a movimientos musculares. Para estudiar los movimientos de cada ojo independientemente se utiliza un tercer electrodo común colocado sobre el puente nasal. Cuando se realizan adquisiciones simultáneas de movimientos horizontales y verticales los errores introducidos por acoplo entre los ejes y por la no linealidad de las mediciones pueden ser considerables. En el caso de la medición vertical se añade además el ruido debido a los parpadeos. La relación señal ruido mejora considerablemente haciendo uso de electrodos implantados, si bien este método conlleva todos los inconvenientes de las medidas intrusivas.

Evaluación

La EOG presenta el mayor rango de medida debido a que no necesita visualización del ojo. El método se puede utilizar en rango de ±70 grados, si bien la linealidad empeora progresivamente a partir de 30 grados, especialmente en el eje vertical. La resolución típica con electrodos superficiales es de ±1,5-2 grados.

Las mayores fuentes de error son los movimientos musculares, interferencias debidas al parpadeo, no linealidad en la técnica y variaciones del potencial córneo-retinal atribuibles a la adaptación a la luz, variaciones diurnas y estado de alerta.

Se podrían conseguir mejoras incorporando amplificadores integrados directamente unidos a la piel para eliminar la susceptibilidad al ruido y minimizar los requerimientos de apantallamiento.

Reflexión en la córnea

Principio

El abultamiento de la córnea produce una imagen virtual de las luces en el campo visual. Debido a que el radio de curvatura de la córnea es menor que el del ojo, el reflejo en la córnea se mueve en la dirección del movimiento del ojo relativo a la cabeza. La luz incidente se refleja en la superficie cóncava de la córnea en un patrón de luz divergente y mediante una lente cóncava puede plasmarse en una película mediante cámara de vídeo o fotodiodos [19,20]. Existe una relación entre el ángulo de reflexión del rayo incidente y los centros de la curvatura de la córnea y del ojo. Las variaciones del ángulo de reflexión permiten calcular la posición relativa de los dos centros y con ello la dirección de la mirada. El método es sensible a los movimientos laterales de cabeza, debido a que el ángulo de reflexión de la luz incidente varía con los desplazamientos. El sistema debe diseñarse con mecanismos que eliminen o compensen ese error.

Desarrollo

Hay dos tipos básicos de métodos dependiendo de la localización de la fuente de luz. En el primero la fuente de luz se fija respecto a la cabeza del sujeto. Para determinar la posición de la mirada sobre el campo visual se requiere un sistema totalmente fijado a la cabeza incluyendo el campo visual, un método de medición de los movimientos de la cabeza, o bien un sistema de adquisición del campo de visión relativo a la cabeza en todo instante. En el segundo grupo de técnicas la luz se coloca fija al campo de visión en lugar de a la cabeza. Los movimientos del reflejo de la córnea relativos a la pupila indican el punto de mira en el campo y no es relativo a la posición de la cabeza. Estos métodos son mucho menos sensibles a los movimientos de cabeza.

Evaluación

El rango lineal sin corrección de todos los sistemas de reflejo de córnea que utilizan una única luz para el reflejo se limita a ±12-15 grados horizontal o vertical. Excursiones más largas localizan el reflejo en la porción periférica no esférica de la córnea, que requiere una calibración y linealización compleja. El rango del reflejo es finalmente limitado por el tamaño de la córnea y su desaparición parcial detrás de los párpados. Además de los movimientos de la cabeza, otro factor que puede limitar la precisión de los métodos de reflexión en la córnea a 0,5-1 grado son las variaciones en la forma de la córnea, la densidad del fluido lacrimoso, astigmatismo de la córnea y la producción de otros reflejos por cristales de gafas, lentillas, etc[21].

Seguimiento de pupila, limbus y párpados

Técnicas básicas

La frontera entre el iris y la esclerótica (limbus) es un límite fácilmente identificable que puede ser detectado ópticamente y seguido por diversidad de medios. Si el iris completo fuera siempre visible y no estuviera parcialmente cubierto por los párpados, sería cuestión de trazar su circunferencia y determinar su centro. Sin embargo, debido a que normalmente sólo una parte del iris es visible, son necesarios otros métodos ópticos, como el seguimiento de la pupila para encontrar su centro.

Cuando únicamente interesan los desplazamientos horizontales, se puede realizar un seguimiento de los extremos derecho e izquierdo del iris por diferencias de la luz reflejada o por un sistema de vídeo. Cuando se necesitan mediciones verticales se puede seguir los niveles de los párpados, la posición de la pupila o el movimiento vertical de una zona visible del limbus. Casi todos los sistemas de seguimiento de limbus usan iluminación invisible, usualmente infrarroja. Todos ellos miden la posición del limbus relativa a fotodetectores. Para el caso de fotodetectores e iluminación fija a la cabeza, los movimientos libres de cabeza son posibles y las medidas de la posición del ojo son relativas a la cabeza.

La pupila ofrece varias ventajas frente al limbus. Primero, es más pequeña y por ello queda oculta por el párpado en menos ocasiones: para movimientos del ojo amplios presenta al instrumento de observación una porción mayor de círculo o de la forma ligeramente elíptica. El centro de la pupila coincide virtualmente con el eje óptico de la fóvea del ojo. Existe una desviación de 5 ó 6 grados pero puede corregirse mediante el calibrado. El borde de la pupila es más limpio y definido que el existente entre iris y esclerótica. Todo ello lleva a una medida de mayor resolución.

Por otro lado, la pupila cuando se observa bajo condiciones lumínicas normales, aparece negra y por ello presenta menor contraste con el iris que el iris con respecto a la esclerótica. Esto hace más complicado discriminar la pupila de forma automática. Sin embargo, si se hace uso de luz colimada, la luz se refleja desde el interior del ojo y para un observador colocado en el eje de la iluminación la pupila aparece brillante. Este efecto se observa en las cámaras de fotos cuando se utiliza el flash cercano al objetivo de la cámara y los ojos aparecen rojos en las fotografías.

Otra característica que presenta la pupila y puede resultar en ocasiones una ventaja y en otras una desventaja es el hecho de que su diámetro varía por influencias fisiológicas y psicológicas. Esto dificulta la medida del centro de la pupila pero en ocasiones, en aquellos métodos que indican el diámetro de la pupila, puede usarse para estudiar el interés en la escena observada en cualquier instante.

Desarrollo

Existen dos grupos de sistemas que tratan de realizar el seguimiento del limbus. El primer grupo utiliza la señal de vídeo adquirida mediante una cámara de vídeo normal o algún dispositivo de escaneado de vídeo y a partir de esa señal trata de encontrar el iris y su centro. Inicialmente tratan de encontrar la línea horizontal que cruza el centro del iris, basándose en la distancia entre los dos límites con la esclerótica y una vez hallada esa línea calcula su centro. Se consiguen rangos de movimiento de ±15 grados y resoluciones de 0,1 grados. El segundo grupo utiliza dos o más fotocélulas que observan zonas específicas del ojo, bien directamente o bien a partir de su imagen. A estas técnicas se les conoce con el nombre de Oculografía de Infrarrojo *IROG, IR-Oculography*. Se puede realizar con una fuente de iluminación amplia y campos muy limitados de fotodetección o bien fuentes de iluminación limitadas y zonas de detección amplias. La técnica consiste en determinar la posición del iris a partir de la diferencia de iluminación medida en las zonas seleccionadas, que se eligen apropiadamente para que pequeños movimientos puedan ser detectados. Se consiguen de esta manera rangos de movimientos oculares de 15 grados y con una precisión de 15 minutos de arco, y pueden conseguirse precisiones de 10 segundos de arco en un rango de unos pocos grados [11]. La interferencia por cambios de iluminación se elimina haciendo uso de una modulación

"*chopper*" de la luz y demodulando a la misma frecuencia.

Evaluación y métodos existentes actualmente en el mercado

Muchas implementaciones de estas técnicas dan un buen resultado con precisión en un rango aceptable. Los movimientos verticales son, sin embargo, un problema. La técnica también requiere dispositivos fijos a la cabeza.

ASL ofrece un sistema basado en IROG que proporciona un rango de ±15 grados tanto en horizontal como en vertical y una precisión de 0,25 grados en horizontal y algo menor en vertical y la velocidad del sistema es de 1000 Hz. El sistema se monta sobre gafas o sobre la cabeza. Express Eye ofrece un sistema también montado en la cabeza y basado en el seguimiento del limbus, que ofrece el mismo rango y velocidad con una precisión mejor, de aproximadamente 0,1 grados. Microguide ofrece un rango mayor de ±30 grados en horizontal y ±20 grados en vertical con una precisión de 0,1 grado, pero su velocidad es menor: de 250 Hz. Skalar ofrece un sistema basado en la reflexión de IR diferencial con modulación de amplitud y detección síncrona. La resolución que proporciona es de 2 minutos de arco en un rango de ±30 grados en horizontal y ±20 grados en vertical y una velocidad de 500 Hz.

Lentes de contacto

Técnicas básicas

Las medidas más precisas de movimientos oculares se consiguen con técnicas que mediante lentes de contacto unen algún dispositivo de forma ajustada al ojo[22-24]. Las lentes convencionales de córnea son demasiado móviles para poder ser utilizadas y todos estos sistemas de medida utilizan lentes especiales consistentes en dos superficies individuales esféricas que ajustan perfectamente sobre la córnea y la esclerótica. Para obtener medidas precisas es indispensable que la lente se mueva solidaria al ojo, tanto en desplazamientos a velocidad constante como en aceleraciones. El ajuste se consigue mediante presiones de –20 mmHg o más negativas entre la lentilla y el ojo. Para obtener estas presiones, se propone rellenar la cavidad con bicarbonato sódico al 2%[25] o retirar mediante una válvula el fluido o aire existente entre la lente y el ojo [24]; si bien retirar aire supone un tiempo limitado de uso debido a la falta de irrigación de la córnea. Ningún sistema de lentes resulta confortable, y las más ajustadas requieren el uso de anestesia local. Se estima que incluso con el mejor de los ajustes existe un desplazamiento de la lente de 1 minuto de arco en sacudidas de 1 grado y de 6 minutos de arco en sacudidas de 9 grados[25]. La lente de contacto y el material asociado a la misma no deben ser de tamaño o volumen que interfiera con los movimientos oculares.

Desarrollo

El sistema más común que utiliza lentes hace uso de una o más superficies de espejo planas fijadas en la lente que reflejan la luz proveniente de una fuente. El reflejo de un espejo, que es adquirido por una placa fotográfica o un cuadrante de fotodetectores, presenta ventajas en cuanto a precisión respecto a las técnicas de reflejo en la córnea. En primer lugar, el cambio en el ángulo de reflexión es el doble del ángulo de rotación del ojo, frente al factor de 1,3 que ofrece el reflejo de la córnea. En segundo lugar, se eliminan las imperfecciones de la córnea. Y lo más importante es que el ángulo de reflexión depende únicamente de la rotación del ojo y es independiente de desplazamientos puramente lineales siempre que el haz incidente siga iluminando el espejo. Esto implica que el sistema se mantiene inalterable frente a movimientos pequeños inevitables de la cabeza. La luz colimada es reflejada por uno de los espejos planos y se enfoca con una lente sobre el plano de la imagen. Cuando el ojo se mueve, el movimiento lateral del plano del espejo en un haz colimado no causa un desplazamiento de la posición de la imagen. Únicamente la rotación del ojo y del espejo causa una desviación de la imagen proyectada [26]. En cualquier caso, debido a la alta precisión inherente al sistema, se estabiliza cuidadosamente la cabeza con respecto a los dispositivos de medida. Se pueden obtener medidas de movimientos de 5 segundos de arco sobre un rango de ±5 grados. Cuando el plano del espejo no es normal al eje óptico los movimientos de torsión del ojo también provocan desplazamientos en la imagen reflejada. Haciendo uso de dos

espejos cuyos reflejos se mueven en diferente dirección durante la torsión, se detectan movimientos de rotación en los tres ejes con precisiones de 2 segundos de arco[27]. En algunos de estos sistemas, los espejos no se colocan directamente sobre la lentilla porque sus propiedades de reflexión pueden cambiar con el fluido lacrimoso, sino que la lentilla lleva una pequeña patilla sobre la que se coloca el espejo[25] . Basados en el mismo principio existen también sistemas que en lugar de colocar un espejo colocan una pequeña lámpara en un extremo de la lente con dirección paralela al eje óptico [28] . Estos últimos sistemas no permiten el cierre total del ojo y resultan incómodos durante el parpadeo. En lugar de lámparas pueden también utilizarse sustancias radiactivas (como tritio) colocadas en la lente que, bajo iluminación ultravioleta, aparecen brillantes[29] . De esta forma se consiguen precisiones de 20 minutos de arco, que resultan un tanto pobres para sistemas de lentes pero que permiten un amplio rango: de 0,3 a 30 grados.

También existen medidas no ópticas que hacen uso de las lentes. La principal coloca dos hilos espirales metálicos sobre la lente, orientados perpendicularmente el uno al otro, donde se induce un voltaje debido a dos espiras electromagnéticas perpendiculares próximas al ojo[30] (Fig.3). La tensión inducida en cada espira sobre la lentilla varía únicamente con el seno del ángulo relativo al campo magnético y es independiente de la posición de la cabeza dentro de la zona uniforme del campo. Se consiguen localizaciones angulares en tres dimensiones del orden de 1 minuto de arco sobre excursiones angulares muy amplias. Los voltajes se miden mediante extensiones de cable ligeras y flexibles de la lente.

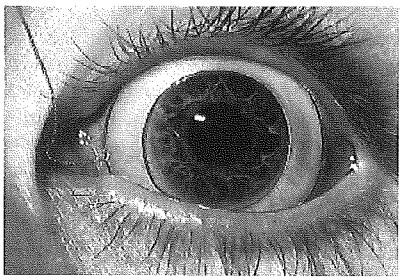


Figura 3. Lente electromagnética

Evaluación y métodos existentes actualmente en el mercado

A pesar de que las lentes ofrecen las mejores resoluciones de hasta 5-10 segundos de arco, lo hacen en general sacrificando el rango angular. Normalmente son aplicables para el estudio de movimientos oculares pequeños, y son inapropiadas para movimientos superiores a 5 grados, a excepción de los métodos electromagnéticos. El coste y la incomodidad de la técnica la hace más apropiada para casos muy específicos de estudios fisiológicos de fijaciones que para su uso generalizado. La presión negativa necesaria supone también riesgos. Existe la posibilidad de deformar la córnea o dañar los músculos de acomodación como resultado de la presión a la que son expuestos.

Skalar ofrece un sistema basado en lentes con espiras metálicas en las que se induce un voltaje mediante un campo electromagnético externo que varía con el movimiento de la lente. Incluyen el kit de colocación de las lentes, que exige anestesiar brevemente los ojos, para poder evacuar el fluido y aire existente entre la lente y el ojo. Se aconseja un tiempo de utilización máximo de media hora, analiza movimientos horizontales, verticales y torsionales, y ofrece un error inferior al minuto de arco pico a pico en un rango aproximado de 30 grados.

Seguimiento del centro del reflejo de la córnea con respecto al centro de la pupila

Principio

Los sistemas anteriores estudiaban los movimientos oculares respecto a la cabeza, más que el lugar donde dirigía el usuario la mirada. Cuando la cabeza se mantiene quieta las dos medidas son equivalentes. Y la posición de la mirada siempre se puede obtener si se conoce la posición del eje óptico respecto a la cabeza y mediante otras técnicas se mide la posición de la cabeza respecto al entorno.

Para poder determinar el punto de mira independientemente de los movimientos de desplazamiento que se realicen, es necesario disponer de dos puntos

del ojo que se muevan de diferente manera en función de los movimientos de cabeza y de la rotación del ojo, de modo que se pueda deducir la dirección de la mirada independientemente de los movimientos de la cabeza. Dos características que cumplen estos requisitos son la reflexión en la córnea y el centro de la pupila[31] (Fig. 4).

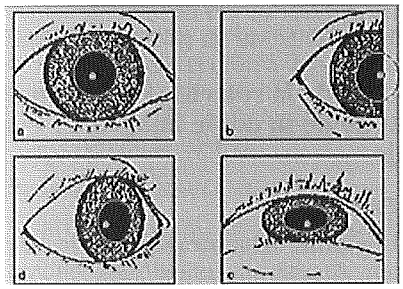


Figura 4. Evolución del reflejo de la córnea en torsiones y rotaciones.

Desarrollo

Si se mira directamente a una fuente de luz, un observador cercano a dicha fuente observará el reflejo de la misma en el centro de la pupila. Esto tiene dos consecuencias. La primera es que la imagen del punto donde el sujeto está mirando aparece en el centro de la pupila. Y la segunda consecuencia es que el ángulo de la mirada con respecto a la fuente de luz es proporcional a la distancia entre la imagen de la fuente de luz en la córnea y el centro de la pupila. Estas dos propiedades son equivalentes. Haciendo uso de la primera se utiliza un conjunto de fuentes de luz, y el punto donde se está mirando se determina por la fuente de luz que aparece en el centro de la pupila. Haciendo uso de la segunda, se utiliza una única fuente de luz y se mide la distancia existente entre su imagen y el centro de la pupila.

En una realización real del primer tipo se utiliza la escena iluminada como fuente de luz, y a partir de una imagen ampliada del ojo se determina qué parte de la escena aparece en el centro de la pupila. El método no ofrece una gran resolución debido a que toda la escena aparece reflejada en la córnea. Además, para poder obtener la escena reflejada, ésta debe poseer varios puntos muy luminosos frente a un fondo oscuro. Sin embargo, el método no presenta restricciones con respecto al movimiento de cabeza. En otra realización práctica se ilumina con una matriz de infrarrojos y se determina el punto de mira. Se han conseguido así rangos de 40 grados con resoluciones de ±2,5 grados.

En el segundo tipo se debe medir la distancia entre el reflejo producido por una única fuente de luz y el centro de la pupila. En cuanto a la fuente de luz, se ha utilizado luz infrarroja mayormente por ser invisible y por lo tanto no molesta, pero también existen realizaciones con luz visible. En una de ellas, toda la luz ambiente se polariza excepto la de una pequeña zona. La luz de retorno se pasa a través de un filtro de polarización y se obtiene una imagen nítida del reflejo.

Una variación del segundo tipo de medida utiliza dos puntos reflejados en la córnea para eliminar los efectos de variación de distancia a la óptica[32] (Fig. 5). La separación entre los dos puntos se convierte en la longitud básica respecto a la que se normalizan el resto de medidas. Esto elimina la necesidad de un calibrado absoluto y la posición de la pupila se mide con respecto al punto medio entre los dos puntos reflejados en la córnea.

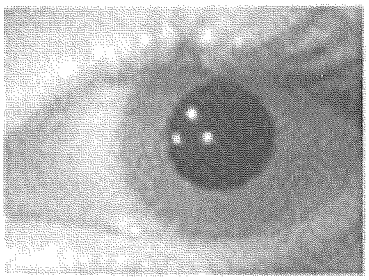


Figura 5. Imagen del ojo con reflejos sobre la pupila oscurecida.

En general los métodos que hacen uso de reflexiones en la córnea están limitados por la propia curvatura de la córnea a un rango de ±15 grados. A partir de ese punto la córnea se aplana y la medida se convierte en no lineal aunque todavía monótona. También pueden existir imperfecciones en la córnea que impliquen cierta no linealidad en la medida. Además de todo esto, la película

lacrimosa y las dilataciones y contracciones de la pupila que desplazan el centro de la misma respecto al globo ocular constituyen un problema.

Evaluación y métodos existentes actualmente en el mercado

Mantener la cabeza decididamente fija o disponer de dispositivos unidos a la misma resulta difícil, no apropiado para muchas tareas, nada confortable y puede requerir bastante tiempo de preparación. Por ello, cualquier tipo de métodos que puedan medir el punto de mira sin necesidad de determinar la posición de la cabeza o estabilizarla ofrecen muchas ventajas: son confortables; ofrecen los datos de una manera cómoda; hacen uso de luces invisibles y por lo tanto no molestas, y permiten un rango amplio de movimientos tanto oculares como de cabeza. Los sistemas resultan más caros, de momento algo limitados en velocidad debido a los sistemas de adquisición y ofrecen menor resolución que las lentes fijas a los ojos o el propio seguimiento del limbus o del reflejo de la córnea.

Este quizá sea el sistema más utilizado comercialmente. Alphabio ofrece un rango de medida de ±30 grados en horizontal ±20 grados en vertical y ±45 en torsiones. La precisión y resolución depende de la velocidad y está limitada mayormente por la cámara de adquisición. De manera que a 60 imágenes por segundo ofrece una precisión de ±0,1 grado y una resolución de ±0,016 grados, a 240 imágenes por segundo la precisión es de ±0,15 grados y la resolución de ±0,15 grados y a 480 imágenes por segundo la precisión es de ±0,3 grados y la resolución de ±0,15 grados. ASL ofrece un sistema a 50/60 Hz con una precisión de 0,5 grados en un rango de 50 grados en horizontal y 40 grados en vertical. DBA ofrece un sistema similar, con la misma velocidad, 50/60 Hz, un rango de ±40 grados y una precisión mejor de 0,5 grados, típicamente 0,2 grados. SMI presenta varios modelos, con dispositivos colocables sobre la cabeza y sistemas de grabación remota. Para medidas únicamente horizontales y verticales, emplea rangos de ±30 en horizontal y ±25 en vertical y precisiones inferiores a 0,1 grados. Para medidas que incluyan movimientos torsionales utiliza tres fuentes de iluminación y consigue rangos de ±20 grados en vertical, ±25 grados en horizontal y ± 20 grados en torsiones, con resoluciones espaciales respectivas de 0,03 grados verticales, 0,02 grados horizontales y 0,1 grado en torsiones. Estos sistemas exigen la colocación del dispositivo en la cabeza. En ambos casos se ofrecen versiones con velocidad de 50 Hz (PAL), 60 Hz (NTSC) y 250 Hz.

Medida de la rotación ocular por el método de la doble imagen de Purkinje

Principio

Conforme la luz atraviesa el ojo, se producen sucesivas reflexiones en varias capas (Fig. 6). En la superficie de la córnea aparece la conocida reflexión de la córnea o primera imagen de Purkinje; una segunda reflexión ocurre en la parte posterior de la córnea, la tercera en la parte frontal de la lente o cristalino y la cuarta en la parte posterior de la lente, donde se encuentra en contacto con el humor vítreo. La segunda imagen de Purkinje es relativamente débil y la tercera se forma en un plano lejano a las demás; de modo que estas dos no se utilizan en el método de medida.

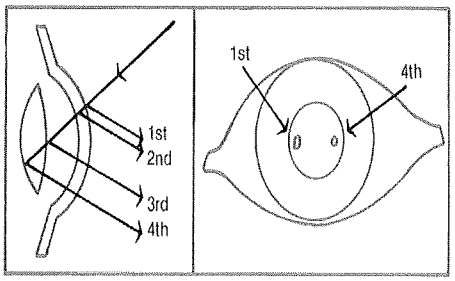


Figura 6. Reflexiones de Purkinje

Igual que en el caso del centro de la pupila y la primera imagen de Purkinje, la primera y cuarta imágenes de Purkinje son dos características del ojo que se mueven conjuntamente frente a translaciones del ojo pero de manera diferente frente a rotaciones[33].

Por cuestiones de simplicidad se supone que las dos superficies (córnea y cara

posterior del cristalino) tienen el mismo radio de curvatura y están separadas entre sí por una distancia igual a este radio. Si las dos superficies se suponen esféricas y el ojo está mirando en la dirección de la luz incidente, la luz colimada incidente creará dos imágenes (primera y cuarta) que se superponen en un punto medio equidistante de ambas superficies. Frente a una rotación del ojo, las imágenes de Purkinje dejarán de coincidir en este punto medio, y aparecerán a una distancia relativa que es proporcional al seno del ángulo de rotación e independiente de translaciones de la cabeza.

Desarrollo

A partir de una fuente de luz, una abertura circular forma las dos imágenes de Purkinje en el ojo. La óptica de adquisición se encarga de trasladar las dos reflexiones a dos fotodetectores, cada uno de los cuales genera una señal eléctrica proporcional a la posición de la imagen respecto al centro. La salida del sistema es la diferencia entre las dos señales eléctricas generadas.

Se permiten movimientos de cabeza de ± 0,5 cm en cualquiera de los tres ejes y se utiliza enfoque automático para permitir movimientos de cabeza en el eje óptico. El rango de movimiento ocular alcanza ±15 grados en los ejes vertical y horizontal con una resolución de hasta 2 minutos de arco. El ancho de banda del sistema puede alcanzar los 300 Hz.

Evaluación

Esta técnica es la única que permite movimientos de cabeza (no muy amplios) obteniendo un seguimiento de mirada con muy alta resolución. Tiene una respuesta frecuencial más amplia que los sistemas de vídeo y por eso es superior al sistema del vector diferencia centro de pupila - centro de reflejo. Como contrapartida, el rango está limitado a ±15 grados debido al diámetro de la pupila y necesita una iluminación mayor para conseguir que la cuarta imagen de Purkinje sea distinguible por encima de los valores de ruido

INTERFAZ

La interfaz de usuario es la parte "visible" del sistema con la que el usuario va a interactuar. Se puede entender una interfaz como la frontera entre dos sistemas de distinta naturaleza. En este caso se trataría pues de la parte existente entre el usuario y el ordenador.

Es posible definir toda una ciencia alrededor del diseño de las interfaces de usuario; sin embargo, no existen teoremas ni reglas exactas que permitan llegar a diseñar la interfaz perfecta[34,35].

A la hora de diseñar una interfaz de usuario los principios utilizados abarcan conocimientos de psicología, pensamiento cognitivo, diseño... Al fin y al cabo se trata de desarrollar algo que permita al usuario interactuar con el ordenador de la manera más sencilla posible.

Evolución

Igual que hemos hablado al principio de este trabajo de la evolución de los ordenadores, se podría hablar también de una evolución en el diseño de las interfaces de usuario. A lo largo de la historia de los ordenadores, el usuario ha podido interactuar con distintos tipos de "pantallas". Una de las fuerzas que impulsa el desarrollo de nuevos diseños es la necesidad existente de facilitar su manejo y de adaptar el sistema al usuario[34,36]. De ahí que el enfoque que guía el diseño de nuevos sistemas haya ido variando desde sus comienzos y, hoy por hoy, el usuario cobre una importancia mucho mayor a la hora de crear una interfaz de usuario.

Paralelamente a la evolución de las interfaces de usuario se puede hablar de una expansión de los ordenadores y sistemas informáticos. No está muy claro cuál de las dos evoluciones ha sido consecuencia de la otra. Es decir, se podría entender una mayor utilización del ordenador por parte del ser humano gracias a que su manejo es cada vez más sencillo o, por el contrario, se podría plantear la demanda por parte de nuevos colectivos como razón para la realización de interfaces de usuario más sencillas.

En definitiva, se puede hablar de una adaptación progresiva del ordenador hacia el usuario. Así en principio se contaba con interfaces orientadas a comando (tipo MS-DOS) en las que el usuario debía

conocer el grupo de comandos y la sintaxis correcta para llegar a transmitir a la máquina sus deseos. Más tarde aparecen las interfaces orientadas a objeto (tipo Windows) en las que se cuenta con entornos mucho más visuales, y en las que cobra más importancia el objeto sobre el que se quiere actuar. Se puede decir que hoy en día éste sería el tipo de interfaz de usuario más común. Pero, dando un paso adelante, se podría llegar a hablar de una nueva generación de interfaces de usuario[34,36]. La característica más sobresaliente de esta nueva clase de interfaces es que se trata de interfaces orientadas a usuario. En este tipo de entornos cobra especial importancia el tipo de usuario; es decir, quién es, qué espera del sistema y cómo va a interactuar con éste. En este caso la comunicación entre el usuario y el ordenador puede estar basada en información sonora e incluso en gestos que pueda realizar el usuario.

Aunque el aspecto externo de las interfaces de usuario haya variado, la comunicación entre ordenador y usuario sigue basándose en comandos simples que el ordenador y el usuario entienden. La diferencia radica en el tipo de comandos intercambiados que han evolucionado de ser una serie de palabras ordenadas según una sintaxis concreta a estar constituidos por simples gestos o información sonora procedente del usuario.

Este grado de adaptación que posee el ordenador hacia el usuario ha dotado a los sistemas informáticos de una mayor inteligencia e independencia de actuación. A medida que el tipo de interfaz ha ido evolucionando, la máquina ha adquirido una mayor capacidad para desarrollar acciones por sí misma. A esta capacidad de desarrollar acciones de forma transparente al usuario (sin que el usuario tenga conciencia de que están ocurriendo) se le ha venido a llamar desarrollo de agentes de interfaz. Se trata de procesos autónomos que se ejecutan sin necesidad de ser lanzados por el usuario. Anteriormente al desarrollo de estos agentes de interfaz el ordenador era capaz de desarrollar únicamente lo que el usuario le ordenaba. Un ejemplo de agente de interfaz sería el sistema de ayuda activa[37] de algunos programas que detectan si el usuario tiene dificultades para desarrollar algún tipo de acción y ofrecen un asistente sin que el usuario lo haya solicitado.

Esta evolución viene a romper con el paradigma de interfaz única y unificada[36] para todos los usuarios y surge la necesidad de crear interfaces dedicadas y adaptadas a cada perfil de usuario. Es decir, se trata de diseñar interfaces de usuario a partir de un usuario, una situación y unos objetivos concretos. No es el usuario el que debe adaptarse a la máquina, sino que es ésta última la que debe adaptarse y aprender del usuario.

Por esto resulta muy complicado definir un estado del arte para la interfaz de usuario. Son muchos los ejemplos de interfaces dedicadas que se encuentran, pero se trata en todo caso de interfaces adaptadas a un tipo de usuario específico y una situación concreta.

Un grupo de usuarios muy a tener en cuenta en este contexto sería el de aquellos individuos con algún tipo de discapacidad. Es en estos casos cuando más se desarrolla el concepto de interfaz dedicada, intentando suplir de alguna forma la dificultad añadida que posee el usuario. Se trata de conseguir fuentes de información alternativas que constituyan los comandos de interacción con el sistema.

Se han desarrollado varios modelos de interfaces controladas a partir de información visual, información sonora o gestual. Es interesante resaltar que antes de esta década no era posible este tipo de diseño, debido a la limitación en la capacidad de procesado de los ordenadores. El manejo de información visual, así como información sonora o gestual dota a la interfaz de una mayor efectividad. En el caso concreto de utilizar información visual, el aumento de efectividad se ha estimado en un 30%.

Midas Touch

Un problema común a todas las interfaces de usuario controladas por información visual sería lo que se ha denominado *Midas Touch*, que recoge la dificultad que supone el distinguir entre movimientos naturales de la vista y de aquéllos con los que realmente se desea realizar una acción. Se han propuesto diversas soluciones para solventar este problema.

Muchos de los movimientos oculares que realiza el ser humano son movimientos reflejos, condicionados por el entorno en que se encuentra. Es decir, el individuo no es consciente de todos los movimientos que realiza con los ojos. A la hora de interactuar con una interfaz visual, se debería distinguir entre los movimientos que realiza cuando está buscando algo, cuando quiere seleccionar algo o cuando simplemente no desea hacer nada. Uno de los factores que diferencia los distintos tipos de movimientos sería el tiempo. Esto es, se podría considerar el tiempo que el usuario permanece con la mirada fija en un punto como parámetro para determinar el tipo de movimiento que se encuentra realizando[38]. Se establece una frontera de 500 ms de latencia sobre un objeto para determinar si el usuario realmente desea seleccionar o activar dicho objeto. Así, por ejemplo, tiempos de fijación inferiores a 100 ms significarían que el usuario está localizando algún objeto concreto o que simplemente la vista realiza movimientos reflejos.

Otra de las alternativas propuestas sería la de fijar una zona de descanso, en la que el usuario debería fijar la vista cuando no desea seleccionar ni activar ningún elemento[39]. Esto no resultaría muy operativo ya que se está forzando al usuario a mantener la mirada en cierta zona lo que, al fin y al cabo, se traduce en prohibir al usuario la realización de movimientos naturales de la vista.

Una de las soluciones quizá más eficiente sería la utilización de una fuente de información adicional que constituya el comando de activación o selección de un objeto que previamente se ha seleccionado con la mirada. Es decir, en primer lugar se seleccionaría el objeto con la mirada, y a continuación éste se activaría al realizar el usuario una acción adicional, como emitir voz o realizar un gesto concreto. Existen sistemas que basan la selección en guiños, soplidos o movimientos de cabeza. Este tipo de interfaces soluciona sin duda el problema del *Midas Touch*, ya que el usuario puede realizar cualquier movimiento ocular sin que esto suponga la activación de los objetos de la interfaz. Sin embargo, como ya se ha dicho, este tipo de interfaces requieren que el usuario sea capaz de proporcionar esta información adicional.

En caso de no contemplar esta posibilidad, la única información en la que puede basarse la comunicación usuario-ordenador sería la de la posición de la mirada. En tales condiciones la alternativa para solucionar el *Midas Touch*, como se ha comentado antes, sería tener en cuenta el factor tiempo como parámetro de distinción entre los distintos movimientos oculares.

Teclados de pantalla

Existen algunos ejemplos de interfaces cuyo control se basa únicamente en información visual. Uno de estos sería la realización de teclados de pantalla (*on-screen keyboard*) o teclados visuales (*visual keyboard*). Se trata de teclados virtuales, impresionados en pantalla, que en principio simularían o sustituirían al teclado típico de manejo manual.

Se puede distinguir entre tres tipos de teclados de pantalla:

- Los dedicados exclusivamente a la entrada de texto.
- Los dedicados a alguna aplicación específica.
- Los de propósito general.

Teclados de texto

Algunos tipos de teclados de pantalla están únicamente destinados al procesamiento de textos, y no permiten el manejo de otros comandos como menús u otras herramientas del sistema. Cosas tan simples como un doble-click o un scroll no podrían ser realizadas[40].

Una de las técnicas más empleadas en los teclados de pantalla para texto sería la de la predicción de palabras. Existen varias modalidades de predicción[41]:

- Algunas se basan en predecir la palabra basándose en la primera letra seleccionada, a partir de cómputos de frecuencias de uso para cada palabra. Es decir, una vez seleccionada una letra, se ofrece en la pantalla una lista de las posibles palabras que se podrían escribir.
- Existen también técnicas que se basan en predecir no la palabra sino solamente la letra que podría seguir a la anteriormente seleccionada[42].

– Por último, se dispone de sistemas que permiten predecir la siguiente palabra (sin necesidad de seleccionar ninguna letra) a partir de la última palabra escrita y basándose en patrones de pares de palabras. Una vez escrita una palabra aparece un conjunto de "palabras siguientes" de las cuales el usuario podría seleccionar una[43].

Este tipo de sistemas de predicción tiene la ventaja notable del ahorro de tiempo y el aumento de la velocidad de escritura. El ahorro de tecleado se ha estimado en un 40-50%. Sin embargo existen argumentos que se oponen o por lo menos ponen de manifiesto las desventajas de este tipo de sistemas de predicción[41].

En primer lugar está el tiempo que el usuario emplea en revisar la lista ofrecida por el sistema: ello supone desviar la vista de su posición actual y consultar la lista de opciones; tiempo que, dependiendo de la palabra que se desea escribir, podría ser mayor que el necesario para teclearla.

En segundo lugar: se pierde automaticidad. La predicción de palabras impide el desarrollo de movimientos automáticos de la mirada sobre la pantalla. Es importante practicar los movimientos hasta llegar a adquirir la capacidad de realizarlos de una manera no consciente (similar a cuando un experto teclea manualmente) con el consiguiente aumento del rendimiento y la velocidad. Sin embargo, el sistema de predicción de palabras se opone al desarrollo de esta automaticidad ya que obliga constantemente al usuario a desviar la mirada hacia las sugerencias ofrecidas y revisarlas, impidiéndole así desarrollar patrones de movimiento automáticos.

Realizar la revisión de las palabras ofrecidas es más lento que la selección directa. En un sistema el que las palabras aparezcan colocadas en filas el usuario debe hacer al menos dos movimientos, uno para colocarse en la fila correspondiente y otro para seleccionarla. En este tipo de sistemas se estima que el ahorro de "tecleo" es de un 45%, sin embargo ofrecen sólo un 28 de aumento en la eficiencia respecto a los sistemas de selección directa[44].

En este tipo de estudios realizados hay que considerar sin embargo algunos factores que podrían suponer variaciones en los resultados obtenidos. Algunos de estos factores serían: el grado de entrenamiento de los usuarios, la extensión del vocabulario y la forma de ofrecer las sugerencias por parte del sistema (tamaño, disposición en la pantalla).

En definitiva los teclados dedicados a texto con sistema de predicción de palabras suponen un ahorro de tiempo aunque su eficiencia pueda ser cuestionada.

Teclados específicos de cada aplicación

Se han desarrollado teclados para trabajar con las distintas aplicaciones informáticas en el ordenador.

De hecho existen numerosas aplicaciones para las que la entrada de texto no es tan importante y sí lo es en cambio el poder activar las diferentes opciones del programa. Un ejemplo serían los navegadores de Internet. En este tipo de programas, aunque la entrada de información de texto por parte del usuario siga siendo importante, resulta más interesante el poder activar ciertas opciones del programa: acciones de búsqueda, almacenar ciertas direcciones...

Otro ejemplo serían aquellos teclados para control de dispositivos, como el teléfono (Fig. 7).

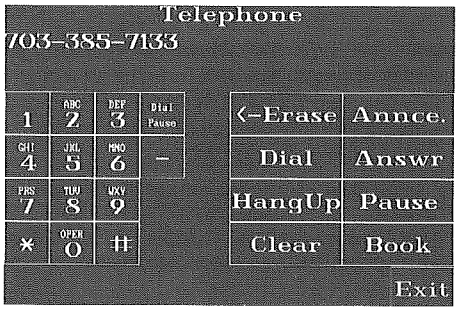


Figura 7. Teclado visual del teléfono.

Teclados de propósito general

Por tanto parece que la solución ideal sería el tener un sistema que permitiese trabajar con texto y a la vez permitiese trabajar con el resto de aplicaciones del ordenador (Fig. 8).

Existen algunos ejemplos de teclados de

pantalla para propósito general[40]. Uno de los problemas a solucionar por este tipo de teclados sería el permitir o simular combinaciones de teclas que se realizan en un teclado manual, *<ctrl+alt+supr>, <ctrl+c>...*

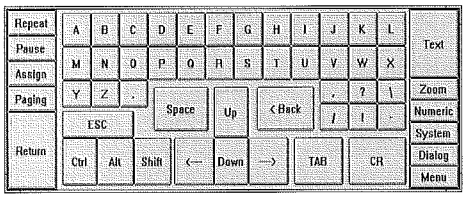


Figura 8. Teclado de propósito general.

Además este tipo de teclados debe permitir de alguna manera la simulación del ratón, así como la conmutación entre aplicaciones. En definitiva, un teclado de pantalla de propósito general debe simular todas las acciones que se pueden realizar combinando un teclado y un ratón de manejo manual.

Algunos sistemas se basan en disponer de distintos tipos de teclados, cada uno con una función determinada[45] (Fig. 9). Es decir, se distingue entre teclados para texto, los cuales están únicamente dedicados al procesamiento de textos; teclados de diálogo, que poseen las herramientas necesarias para gestionar las ventanas de diálogo (posibilidad de movimiento en la ventana, *<esc>*, *<tab>*...); teclados para la gestión de menús; etc.

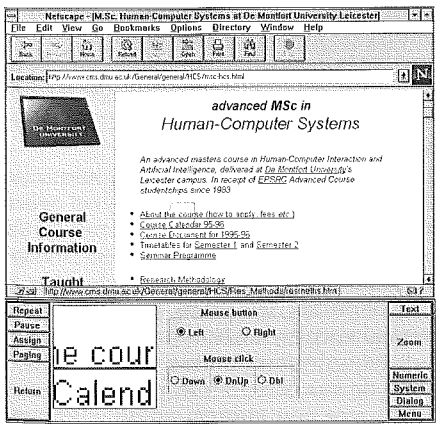


Figura 9. Teclado de propósito general aplicado a Netscape.

Estos teclados son comunes a todas las aplicaciones del ordenador. La activación de las diferentes teclas se realiza por tiempo, estableciendo diferentes tiempos para diferentes teclas y solucionando así el problema del *Midas Touch*[45].

Además del *Midas Touch*, se podría plantear el problema de la resolución del ojo en la pantalla. El grado de resolución da idea de la cantidad de puntos en la pantalla que podría llegar a distinguir el usuario. Para algunas aplicaciones podría ocurrir que esta resolución no fuese suficiente (los ítems no están suficientemente separados). Una de las soluciones planteadas sería la de realizar un zoom de la zona que el usuario se encuentra observando, de forma que esta parte de la interfaz aparece mayor y ocupa un mayor espacio en la pantalla. De esta forma los objetos aparecen más separados y permiten una mayor exactitud.

Se ha demostrado que este tipo de teclado resulta totalmente válido y que podría llegar a proporcionar a personas con algún tipo de discapacidad la posibilidad de llegar a manejar un ordenador y todas sus aplicaciones[45].

CONVERSOR TEXTO-VOZ

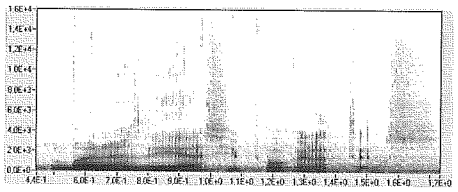
Características del habla

Características físicas

El habla físicamente es una onda acústica de presión cuasiestacionaria, con un ancho de banda aproximado de 7 kHz. Sus características van cambiando a lo largo del tiempo, pero en periodos pequeños de la señal (~30 ms) mantiene sus características y puede considerarse estacionaria. Las características espectrales de la onda dependen del fonema que se está pronunciando en cada momento. Así, en el caso de fonemas sonoros, la onda es periódica y rica en armónicos, con un periodo fundamental que se conoce con el nombre de *Pitch*. El pitch medio es característico de cada hablante y es función de la frecuencia a la que vibran sus cuerdas vocales. En el caso de los hombres la frecuencia fundamental oscila entre 80 y 200 Hz frente a la de las mujeres que se encuentra entre los 150 y los 400 Hz.

Efectivamente, a lo largo del discurso,

el pitch sufre pequeñas variaciones con lo que conseguimos darle entonación al habla. En el caso de sonidos sordos, la señal del habla no es periódica sino ruidosa y la mayor parte de su energía se concentra en las frecuencias más altas de la banda característica del habla. En ambos casos, sonidos sonoros y sordos, existen unas bandas de frecuencia denominadas *formantes* donde se concentra la energía de la señal y que caracterizan el tipo de fonema que se está pronunciando. En cuanto a las características temporales de la onda, también dependen del tipo de fonema: en el caso de sonidos sonoros la onda presenta mayor amplitud y, como ya se ha comentado, su forma es periódica, frente a los sonidos sordos, que se caracterizan por tener una amplitud o energía muy baja y presentar un aspecto ruidoso. Los sonidos oclusivos se caracterizan por un silencio de pocos milisegundos y una explosión de energía de corta duración. En la figura 10 puede observarse la evolución del espectro, o del reparto de la energía en frecuencia a lo largo del tiempo en la frase *"Buenos días Hipócrates"*.



/b/u/e/n/o/s/d/i/a/s/ i/ p/o/c/r/a/ t/e/s/

Figura 10. Espectrograma.

Las manchas más oscuras son las formantes, es decir las frecuencias de mayor energía cuya disposición caracteriza a los distintos fonemas. La evolución de estas bandas de energía es continua de un fonema a otro y es difícil determinar cuándo termina uno y comienza el otro.

Generación del habla

En la generación del habla intervienen los pulmones, la tráquea y el tracto vocal, compuesto por la laringe, la faringe, y las cavidades nasal y oral. El aire emitido por los pulmones atraviesa la tráquea y alcanza la laringe donde, en el caso de la generación de sonidos sonoros, se producirá la sonorización de la onda. En el centro de la laringe se encuentran las cuerdas vocales, dos repliegues musculares en forma de V, rodeadas de una estructura de cartílago y músculo que forma una hendidura: la glotis. Mediante estos músculos la glotis, que durante la respiración permanece abierta, se puede cerrar y mediante la presión del aire proveniente de los pulmones se abre periódicamente generando un tren de pulsos de aire que dan lugar a los sonidos sonoros. La apertura periódica de la glotis se rige por el principio de Bernouilli. El aire emitido por los pulmones y el cierre de la glotis, produce un aumento de la presión subglotal, sobrepasando los 40-60 Pa, y pudiendo alcanzar los 200 Pa. Debido a la presión, las cuerdas vocales se separan, el aire fluye, su velocidad aumenta y la presión subglotal disminuye. Por efecto de la tensión de las cuerdas, de nuevo superior a la presión subglotal, la glotis vuelve a cerrarse y el ciclo comienza de nuevo. La frecuencia de vibración de las cuerdas marca la frecuencia fundamental del sonido producido y es función tanto de características físicas de las cuerdas, como su longitud y grosor, como de la tensión que nosotros apliquemos a los músculos adyacentes. En la generación de los sonidos sordos, sin embargo, las cuerdas vocales permanecen abiertas dejando fluir el aire libremente. Tanto en el caso de la producción de sonidos sordos, como en la producción de sonidos sonoros, el aire atraviesa el resto del tracto vocal, que actúa como cavidad resonante atenuando unas frecuencias y permitiendo pasar otras. Mediante movimientos de la boca y la lengua se varía la forma de la cavidad resonante y con ello las bandas de paso y atenuación de energía, dando lugar a los diferentes sonidos que producimos al hablar.

En la generación de vocales, interviene la vibración de las cuerdas, y el tracto vocal se mantiene durante su producción, destacándose generalmente tres frecuencias de resonancia que las caracterizan. En el caso de las fricativas y silbantes las cuerdas vocales no intervienen, pero se produce un estrechamiento del tracto en alguno de sus puntos que genera turbulencias de aire con energía en altas frecuencias. En las oclusivas, las cuerdas vocales pueden intervenir o no, dependiendo de si

se trata de oclusivas sonoras o sordas, y se caracterizan por una evolución del tracto vocal, que inicialmente permanece cerrado provocando un silencio y a continuación se abre súbitamente, produciendo una explosión de energía repartida en frecuencia, dependiendo de la posición del tracto previo a la apertura, del lugar del cierre del tracto y de su posición final tras su apertura. Por último, en el caso de las nasales, la cavidad nasal se acopla a la cavidad oral, produciendo resonancias y antiresonancias adicionales.

Prosodia y coarticulación

Debido a la continuidad del habla el tracto vocal está continuamente cambiando de forma y mientras pronunciamos un fonema se va adaptando a la pronunciación del siguiente. A este fenómeno se le conoce con el nombre de *coarticualación* y convierte al habla en una señal con cambios suaves y continuos. Otro fenómeno importante para la comprensión del habla es la *prosodia*. La prosodia engloba los aspectos melódicos del habla: el ritmo, el tono y el volumen. Mediante el control de estos tres parámetros dotamos al mensaje de sentido y somos capaces de transmitir emociones, hacer preguntas, realizar exclamaciones, enfatizar tramos del discurso etc.

Sintetizadores del habla

Un sintetizador de habla es un dispositivo capaz de producir habla humana artificialmente. Hoy en día existen dos tipos de sintetizadores de habla. Los sintetizadores por reglas y los sintetizadores por concatenación. Los sintetizadores por reglas tratan de alguna manera de imitar la producción de habla natural, y se les conoce con el nombre de sintetizadores por reglas porque para la producción de cada sonido en cada contexto es necesaria la aplicación de un número elevado de reglas. Los sintetizadores por concatenación almacenan unidades de voz natural y las concatenan para sintetizar el nuevo mensaje, procesándola si es preciso para conseguir los parámetros de prosodia necesarios en cada momento.

Sintetizadores por reglas

Generan una señal similar a la que generan los pulmones junto con la laringe y, mediante filtros digitales, tratan de simular el comportamiento del tracto vocal. Para ello es necesario conocer los parámetros necesarios para la producción de cada sonido como las frecuencias de resonancia del tracto, la periodicidad del sonido, el ancho de banda, la energía, la evolución en el tiempo y cómo modificarlos para pasar de un fonema a otro. Este método analiza los datos acústicos de la voz y los utiliza para sintetizarla. En una primera fase se crea un corpus fonético representativo de las transiciones y coarticulaciones a estudiar. Se realiza una grabación y mediante un analizador de voz se parametriza la señal separando la contribución de las cuerdas y del tracto y se presenta en forma más compacta y adecuada. De este análisis se obtienen una serie de parámetros y reglas que nos describen el habla. La calidad depende de la eficiencia de las reglas, la calidad del corpus, tanto en la elección como en la grabación y el modelo del habla que se utilice en el análisis. Este método describe el habla como evolución dinámica de hasta 60 parámetros [46], la mayoría relacionados con las frecuencias centrales de formantes (máximos de la envolvente espectral) y antiformantes (mínimos de la envolvente espectral) y sus anchos de banda y forma [47]. Como resultado, dicha técnica está casi libre de errores internos de modelización. En oposición a esto el amplio número de parámetros complica el análisis. Además, las frecuencias de las formantes son inherentemente difíciles de estimar a partir de los datos de voz.

La síntesis se consigue mediante un banco de filtros digitales conectados bien en serie o en paralelo, cuyos parámetros y señal de excitación (ruidosa o periódica) se ajustan con los datos obtenidos del analizador de voz[48] (Fig. 11).

Actualmente la calidad de síntesis conseguida muestra problemas típicos de silbido que las propias reglas generan. "Introducir un grado alto de naturalidad es posible teóricamente, pero las reglas para hacerlo están todavía por descubrir".

Los sintetizadores por reglas potencialmente son los más potentes en cuanto al acercamiento al problema de la sín-

tesis del habla. No es de extrañar que hayan sido integrados en sistemas texto-habla: MITtalk[47], JSRU[49] para inglés,[50] para castellano, INTOVOX sistema multilingüe[51], INRS[52] o[53] para francés.

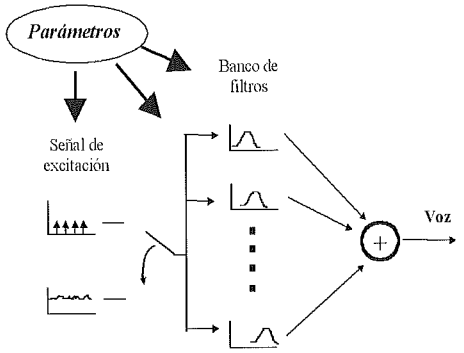


Figura 11. Diagrama de bloques de un sintetizador por reglas.

Síntesis por concatenación

En oposición a los basados en reglas, los sintetizadores por concatenación poseen un conocimiento muy limitado de los datos con los que tratan. Se limitan a concatenar segmentos de voz previamente grabados.

Antes de que el sintetizador pueda producir ninguna pronunciación se debe crear una base de datos con segmentos y acondicionarlos para la posterior concatenación. En primer lugar hay que determinar la longitud de los segmentos, pudiendo ser frases, palabras, sílabas o unidades más pequeñas. Se utilizan normalmente *difonemas* (unidad que comienza en el centro de un fonema y termina en el centro del fonema siguiente) y *trifonemas* (comienza en el centro de un fonema, abarca la totalidad del siguiente y termina en el centro del tercer fonema), dado que incluyen la mayoría de transiciones y coarticulaciones, requiriendo una memoria permisible. Cuando se tiene una lista completa de segmentos se selecciona una lista de palabras donde el segmento aparezca como mínimo una vez. Se excluyen posiciones desfavorables, como dentro de sílabas acentuadas o en contextos coarticulados. El corpus se graba y almacena digitalmente y se extraen los segmentos seleccionados, bien manualmente con herramientas de visualización de voz o semiautomáticamente, con algoritmos de segmentación, cuyas decisiones se comprueban y corrigen iterativamente. El resultado es una base de datos que contiene la lista con los nombres de todos los segmentos y sus duraciones. Hasta aquí el proceso es fuertemente dependiente del idioma.

Los segmentos se almacenan a menudo de forma paramétrica, obteniéndolos a la salida de un analizador de voz. Esta operación recuerda en muchos aspectos al análisis realizado en sintetizadores por reglas, pero su objetivo es diferente. Esta parametrización se realiza por dos motivos. En primer lugar permite reducir el número de datos, algo no desechable en este tipo de sintetizadores dada la gran cantidad de los mismos. Consecuentemente, el analizador es seguido de un codificador paramétrico. En segundo lugar algunos modelos paramétricos separan la contribución respectiva de la fuente y el tracto vocal, una operación que resulta beneficiosa para operaciones de presíntesis, como ajuste de prosodia y concatenación de segmentos.

La tarea del sintetizador es producir en tiempo real una secuencia adecuada de segmentos concatenados extraídos de la base de datos y ajustar la entonación y duración. Esta parte resulta más fácil si los segmentos han sido parametrizados adecuadamente. Debido a que los segmentos a concatenar se extraen de diferentes contextos fonéticos presentan desajustes de amplitud y timbre que derivan en discontinuidades audibles en la concatenación. El efecto puede atenuarse en la constitución de la base de segmentos mediante una ecualización que impone un espectro en amplitud similar en los extremos de los segmentos, siendo distribuida la diferencia en la vecindad. Esta operación está restringida a parámetros de amplitud. Los conflictos de timbre se tratan mejor en el momento de ejecución, suavizando la transición entre los segmentos individualmente cuando es necesario.

El cambio de entonación y duración de los segmentos se realiza por algoritmos de procesado de señal. Existen diferentes propuestas, PSOLA, MRB-PSOLA, MREMBROLA... basadas en desplazamientos y solapamientos temporales de la señal[54,55].

La concatenación se realiza por interpolación de la parte final del segmento previo y la inicial del segmento posterior. Debido a que las unidades usadas como segmentos son difonemas, es decir comienzan en el centro de un fonema y terminan en el centro del siguiente fonema, las transiciones están en el centro de los difonemas, y la concatenación se da en la parte estacionaria (central) del fonema, resultando una unión limpia y relativamente sencilla de realizar.

Sistemas texto-habla

Los sintetizadores de voz, en algunas de sus aplicaciones, requieren tomar como entrada un texto. A un sistema que toma como entrada un texto y lo reproduce en forma de voz sintética se le conoce como sistema texto-habla o más popularmente en ingles TtS (*Text to Speech*). La utilización de texto como entrada a un sintetizador de voz es cómoda de cara al usuario pero no es óptima para la máquina. El texto es ambiguo y no lo suficientemente específico en términos de información lingüística, por ello requiere un tratamiento de coste computacional elevado antes de ser sintetizado. Por lo tanto, un sistema texto-habla constará de dos etapas: un analizador de texto y un sintetizador de voz (Fig. 12).

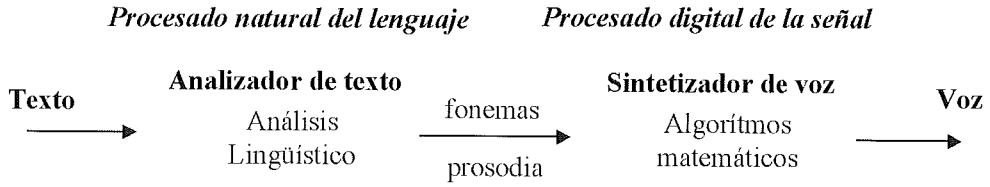


Figura 12. Diagrama de bloques de un conversor texto-voz.

Analizador de texto

Esta etapa se encarga del procesado del lenguaje y es capaz de transcribir el texto a fonemas, entonación y ritmo. Para ello hay que tratar distintos puntos:

– En primer lugar, es posible que el texto conlleve algún tipo de cabecera, como puede ocurrir con cartas, documentos, artículos, correo electrónico, etc. Lo primero que se debe hacer en este caso es separar la parte a sintetizar de la parte que no nos interesa, o incluso extraer la información interesante de la cabecera como "informe enviado por....el día..." y prepararla para su síntesis.

– Otro de los aspectos que habrá que considerar es la expansión de abreviaturas, si las hubiera, y la pronunciación de siglas, fechas y números. Así, si en el texto aparece 12-5-92 debería extenderse en una primera fase a "doce de mayo de mil novecientos noventa y dos".

– En el caso de idiomas como el castellano la limitación del contorno de las frases y de las palabras es relativamente sencillo debido a que las palabras en los textos aparecen separadas por espacios en blanco o símbolos de puntuación. En contraste, en muchos idiomas asiáticos la situación no es tan simple debido a que nunca utilizan espacios en blanco para separar las palabras.

– Pronunciación de cada palabra. Se trata de transcribir grafemas (o letras) a fonemas. Esta fase es fuertemente dependiente del idioma y no hay generalizaciones posibles: es necesario un módulo de transcripción para cada idioma. En castellano de nuevo el proceso es relativamente sencillo si lo comparamos con otros idiomas como el inglés o danés, debido a que la escritura se acerca bastante a la pronunciación y la regla ideal 1carácter = 1fonema se respeta en la mayoría de los casos. Existen algunas excepciones (c, g,

...), pero siguen unas reglas ortográficas muy sencillas que nos permiten extraer fácilmente a partir del texto la secuencia de fonemas. En otros idiomas el problema de obtener la pronunciación de las palabras se resuelve mediante la utilización de un diccionario de pronunciación, donde aparecen todas las palabras junto a su pronunciación.

– Coarticulación. Este punto es más complejo que los precedentes. El problema de coarticulación consiste en que un mismo fonema se pronuncia de diferente manera dependiendo de los fonemas que le precedan o sigan. Esto es debido a que mientras pronunciamos un fonema, nos preparamos para pronunciar el siguiente (adecuando la forma de la boca, lengua, etc.). El cambio de un fonema a otro lo hacemos de forma continua. Un sintetizador debe prestar atención a la transición entre fonemas.

– Prosodia. La impresión de naturalidad en un sistema texto-habla depende mucho de la riqueza de entonación y calidad de los patrones rítmicos. A estos dos aspectos, ritmo y entonación, se les conoce con el nombre de prosodia. La prosodia puede describirse numéricamente en forma de secuencia de valores de la frecuencia fundamental, conocida como pitch, y duración de los fonemas con ayuda de algoritmos de análisis de pitch y segmentación. La prosodia aparece a distintos niveles, por un lado la acentuación de las distintas sílabas dentro de una misma palabra, por otro la acentuación de las palabras dentro de una frase, y por último el fraseado o entonación global de la frase.

– En cuanto a qué sílaba debe ser acentuada dentro de una palabra, también en el caso del castellano podemos obtenerlo directamente del texto gracias a las reglas ortográficas de acentuación. En idiomas donde estas reglas no existen, al igual que en el caso anterior, el problema se resuelve con los diccionarios de pronunciación que deben indicar la sílaba tónica dentro de la palabra.

– La acentuación de las palabras dentro de una frase, asociada con el contenido semántico y su importancia dentro de la frase y la entonación de toda la frase marcando grupos sintácticos es más difícil de determinar, siendo necesaria tanto información léxica como sintáctica. El cambio de entonación para preguntas o exclamaciones en castellano también es algo más fácil de realizar debido a que tenemos indicación de ello antes de comenzar la frase y al finalizarla, marcando adecuadamente los límites de la misma.

– La prosodia puede asimismo transmitir el estado de ánimo del hablante. Aspectos como la velocidad de discurso, el tono medio y su varianza, así como el volumen y su varianza permiten expresar el estado de ánimo del hablante y puede simularse.

SiVHa

El objetivo de SiVHa es diseñar un sistema de comunicación automático, rápido, cómodo, fácil y práctico, controlado únicamente con la vista. Los tres módulos que componen el sistema han sido diseñados atendiendo a las necesidades del usuario y las soluciones adoptadas han sido determinadas por los objetivos impuestos.

Eyetrack

Los objetivos que debe cumplir este módulo son:

– Determinación inequívoca del punto de pantalla al que está mirando el usuario con precisión y resolución

– Comodidad para el usuario: se han desechado aquellos métodos que exigen contacto físico con el usuario

– Posibilitar cierta movilidad: se necesita un método insensible al movimiento

– Un método independiente de las condiciones lumínicas

Por todo ello se ha optado por un método VOG (Vídeo Oculography) que no exige contacto directo con el usuario y la estimación del punto de mira basada en la posición relativa del reflejo de la córnea y el centro de la pupila para permitir movilidad. La iluminación elegida para provocar el reflejo ha sido un led de potencia controlada a 8 mW en 880 nm. La longitud de onda en la banda del infrarrojo es invisible para el ojo humano y no resulta molesta, y la potencia viene determinada por el mínimo necesario para que la cámara la detecte y el máximo permisible para evitar

dañar el ojo, incluso en exposiciones prolongadas. La cámara CCD debe tener una respuesta espectral hasta 900 nm y la velocidad escogida ha sido de 50/60 imágenes por segundo. La independencia de las condiciones lumínicas se consigue mediante la utilización de un filtro que elimine el visible.

La determinación de los dos puntos de interés se consigue procesando cada una de las imágenes adquiridas por la cámara una vez digitalizadas por la tarjeta de adquisición. Para facilitar la discriminación de la pupila del resto de la imagen, el led se coloca en el eje de la cámara, apareciendo de esta forma la pupila mucho más brillante. Con el uso de histogramas se determinan umbrales para encontrar el reflejo y la pupila y a partir de ahí, matemáticamente, se calcula la posición relativa del centro de la pupila y el reflejo.

La asociación de la posición calculada y los puntos en pantalla se consigue mediante un calibrado al inicio de la sesión.

Las líneas de trabajo actuales y futuras se centran en conseguir un sistema lo más robusto posible (frente a desenfoques, movimientos en cualquier dirección, etc.), en aumentar la precisión mediante el uso de método mixtos de seguimiento de ojo y técnicas de predicción y recalibrados automáticos durante la sesión, transparentes al usuario.

Interfaz

Los objetivos en este módulo son
- Construir texto con la mirada
- Facilidad de manejo
- Rapidez
- Adaptación a las preferencias del usuario

El primer objetivo no es nada trivial. A partir de una posición de pantalla, la interfaz debe adivinar las intenciones del usuario. Debe discernir la voluntad de selección, la búsqueda o la mirada pensativa. Para ello se puede optar por un análisis de movimientos oculares y tratar de predecir las intenciones del usuario o bien determinar un protocolo de comunicación. La primera solución se basa en utilizar el tiempo de las fijaciones y la evolución de las sacudidas para determinar qué tipo de tarea está realizando el usuario: búsqueda, selección o descanso. En el segundo caso se establecen unas zonas de selección o determinados movimientos oculares para indicar la voluntad de selección. En nuestro sistema se han diseñado varias interfaces con diferentes soluciones para que el propio usuario pueda elegir cuál le resulta más fácil de utilizar. Para solucionar el problema de Midas Touch se opta por poner un botón de activación y desactivación de la interfaz.

En cuanto a la facilidad de manejo, se ha diseñado una interfaz visual de ventanas, limpia, sencilla, con los elementos imprescindibles en pantalla. Para facilitar la construcción de texto se ha introducido un diccionario con las palabras más usuales y un abecedario como complemento al diccionario. De esta manera el usuario no debe escribir letra a letra todo el texto, puede seleccionar las palabras del diccionario y, en caso de que no aparezca alguna, deletrearla mediante el abecedario.

Para obtener mayor rapidez, se han desarrollado herramientas de búsqueda, para posicionar de forma más rápida el diccionario en las palabras de interés, y se están desarrollando herramientas de predicción de palabras para acelerar aún más la síntesis del texto.

Para adaptarse a las preferencias del usuario, el sistema aprende el vocabulario más usado por el usuario y lo presenta en primer lugar. Además ofrece la posibilidad de generar un diccionario personalizado de frases más usuales. El sistema guarda las preferencias del usuario de una sesión a otra, de manera que con su uso la interfaz cada vez es más personal.

Actualmente se están desarrollando herramientas para medir la eficacia del sistema en cuanto a velocidad y para poder analizar el comportamiento de la mirada del usuario y optimizar el sistema.

Conversor texto-voz

El sintetizador debería cumplir
- Reproducir oralmente el texto de forma inteligible
- Naturalidad del discurso
- Conseguir una voz agradable para el usuario

– Permitir transmitir emociones con el discurso

Para conseguir una síntesis inteligible y con un sonido natural se ha optado por un método de síntesis de voz basado en la concatenación de segmentos reales de voz pregrabados. Como segmentos de voz se han elegido los difonemas, porque permiten la síntesis de cualquier texto con una necesidad de memoria relativamente pequeña y la concatenación de los segmentos es relativamente sencilla debido a que se da en la zona estacionaria de los fonemas. Las técnicas elegidas para ajustar la cadena de difonemas a los valores de entonación y duración deseados han sido MBR-PSOLA y PSOLA, por ser dos técnicas de bajo cómputo y buena calidad de síntesis.

La naturalidad del discurso se consigue mediante una buena estimación de prosodia. Éste es un campo de investigación muy abierto todavía, en el que queda mucho trabajo por hacer. Nuestra investigación se basa en el análisis sintáctico del texto y la introducción de cierta variabilidad en los parámetros para evitar la monotonía y la repetición de patrones melódicos.

Para obtener una voz agradable para el usuario en primer lugar se ha considerado interesante poder ofrecer varias voces al usuario y que pueda elegir entre ellas. Para conseguir varias voces, es necesario realizar grabaciones de difonemas de diferentes personas o bien disponer de herramientas que transformen un tipo de voz en otra. Se han preparado voces femeninas y masculinas mediante grabaciones y ahora se están desarrollando herramientas de cambio de voz para ampliar el abanico de voces disponibles.

Por último, para que el usuario pueda añadir emociones al discurso oral, se está investigando en los aspectos del habla que conllevan connotaciones emotivas, normalmente expresadas mediante el ritmo, el volumen y la entonación del discurso. Actualmente el sistema incluye 6 posibles estados de ánimo.

La investigación en este módulo está muy abierta y se está trabajando en todos los puntos comentados (naturalidad, transformación de voz y emociones) para obtener el discurso más acorde con los deseos del usuario.

BIBLIOGRAFÍA

1. Young L, Sheena D. Survey of eye movement recording methods. Behav Res Methods Instrumentation 1975; 7: 397-429.
2. Mackensen G. Die Geschwindigkeit horizontaler Blickbewegungen Unteresuchungen mit Hilfe der Electrooculographic. Arch Ophthalmol 1958; 169: 47-64.
3. Cook G. Control sytem study of the saccadic eye movement system. ScD Thesis, Massachussetts Institute of Technology 1965.
4. Robinson DA. The mechanics of human saccadic eye movements. J Physiol 1964; 174: 245-264.
5. Vossius G. Das System der Augenbewegung. Zeitschrift für Biologie 1960; 112: 27.
6. Young LR, Zuber BL, Stark L. Visual and control aspects of saccadic eye movements. NASA CR-564. 1966
7. Robinson DA. The mechanics of human smooth pursuit eye movement. J Physiol 1965; 180: 569-591.
8. Westheimer G. Eye movement responses to a horizontally moving visual stimulus. AMA Arch Ophtalmol 1954; 52: 932.
9. Rashbass C, Westheimer G. Disjunctive eye movements. J Physiol 1961; 159: 361.
10. Zuber BL, Stark L. Dynamical characteristics of fusional vergence eye-movement system. IEEE Transactions Systems. Sci Cybernetics 1968; 72-74.
11. Zuber BL. Physiological control of eye movements in humans. PhD Thesis, Massachussetts Institute of Technology 1965.
12. Gonshoor A, Malcom R. Effect of changes in illumination level on electro-oculography. Aerospace Medicine 1971; 41: 138-140.
13. Geddes LA, McCrady JD, Hoff HE. The impedance nystagmogram-A record of the level of anesthesia in the horse. Southwest Veterinarian 1965; 19.
14. Sullivan G, Weltmen G. The impedance oculogram-A new technique. J Appl Physiol 1963; 18: 215-216.
15. Mowrer OH, Ruth RC, Miller NE. The corneoretinal potential difference as the basis of the galvanometric method of recording eye movements. Am J Physiol 1936; 114: 423.
16. Schott E. Über die Registrierung des Nystagmus und anderer Augenbewegung vermittels des Seitengalvanometers.

Deutsches Archiv für Klinische Medizin 1922: 140: 79-90.

17. SCHACKEL B. Eye movement recording by electro-oculography. In P. H. Venables & I. Martion (Eds.), An manual of psychophysiological methods. Amsterdam: North-Holland Publishing Co. 1967; 300-334.
18. SCHACKEL B. Review of the past and present in oculography. Medical Electronics Proceedings of the Second International Conference, London: Hiffe 1960; 57.
19. CARMICHEL L, DEARBORN WF. Reading and visual fatigue. Boston: Houghton Mifflin 1947.
20. TAYLOR SE. The dynamic cativity of reading: A model of the process. EDL Research and Information Bulletin. New York: McGraw-Hill 1971; 9.
21. HALL RJ, CUSACK BL. The measurement of the eye behavior: Critical and selected reviews of voluntary eye movement and blinking. U.S. Army Human Engineering Laboratory, Technical Memorandum 1972; 18-72.
22. DITCHBURN RW, GINSBORG BL. Involuntary eye movements during fixation. J Physiol 1953; 119: 1.
23. RIGGS LA, RATLIFF R, CORNSWEET JC, CORNSWEET TN. The disappearance of steadily fixated test objects. J Optical Society America 1953; 43: 495.
24. YARBUS AL. Eye movements and vision. New York: Plenum Press 1967.
25. FENDER DH. Contact lens stability. Biomedical Science and Instrumentation 1964; 2: 43-52.
26. RATLIFF F, RIGGS A. Involuntary motions of the eye during monocular fixation. Journal of Experimental Psychology 1953; 40: 687-701.
27. MATIN L, PEARCE DG. Three dimensional recording of rotational eye movements by a new contact lens technique. Biomed Sci Instrum 1964; 2: 79-95.
28. BYFORD GH. A sensitive contact lens photoelectric eye movement recorder. IRE Transactions on Bio-Medical Electronics 1962; 9: 236-243.
30. NAYRAC P, MILBLED G, PARQUET PHJ, LECLERCQ M, DHEDIN G. Un nouveau procédé d'enregistrement des mouvements oculaires. Application aux tests de tracking. Lille Medical 1969; 14: 685-687.
31. ROBINSON DA. A method of measuring eye movement using a scleral search coil in a magnetic field. IEEE Trans Biomed Electronics 1963; 10: 137-145.
32. MERCHANT J, MORRISETTE R. Remote measurement of eye direction allowing subject motion over one cubic foot of space. IEEE Trans Biomed Engineering 1974; 21: 309-317.
33. HOCHBERG J. Personal communication to D. Sheena 1974.
34. CORNSWEET TN, CRANE HD. Accurate two-dimensional eye tracker using first and fourth Purkinje images. J Optical Soc America 1973; 63: 921.
35. MANDEL T. Elements of user design Wiley USA 1997.
36. SCHNEIDER-HUFSCHMIDT M, KÜHME T, MALINOWSKI U. Adaptative user interfaces, principles and practice Human Factors in Information Technology 10 N·H Holanda 1993.
37. NIELSEN J. Non command user interfaces Communications of the ACM 1993; 83-99.
38. FISHER G, LEMKE A. Knowledge-based help systems ACM CHI´85 Conferencia de Factores Humanos en Sistemas, San Francisco CA 1985.
40. VELICHKOVSKY B, SPRENGER A, UNEMA P. Towards gaze-mediated interaction: Collecting solutions of the Midas Touch Problem 1997 http://www.phy.tudresden.de/psycho/sydney.htm.
41. DOWNING A.R. Eye controlled and other fast communicators for speech impaired and phisically handicapped persons Australasians. Phys Engin Sci Medicine 1985; 8, No 1: 17-21.
42. FRIEDMAN MB, KILIANY G, DZMURA M. The eyetracker communication system Johns Hopkins. APL Technical Digest 1992; 3: 250-252.
43. STEWARD H. Just how useful is word prediction? Ability Research Centre: Newsletter Nov.1996.
44. MACKAY D. Arithmetic coding data entry device 1998. http://wol.ra.phy.cam.ac.uk/mackay/dasher/
45. KLUND J, NOVAL M. If word prediction can help, which program do you choose? 1995
46. KOESTER HH, LEVINE SP. Modeling the speed of text entry with a word prediction interface IEEE. Trans Rehab Engin 1994; 2.
47. HOWELL O. Providing motor impaired users Interface software via eye based interaction Proc. 1st Euro. Conf. Disability, Virtual Reality & Assoc. Tech. Maidenhead UK 1996.
48. STEVENS KN. Control parameters for synthesis by rule. Proceedings of the ESCA tutorial day on speech synthesis, Autrans 1990; 27-37.

49. ALLEN J, HUNNICUT S, KLATT D. From text to speech: the MITalk System. Cambridge University Press 1987; 213.

50. KLATT D. Software for a cascade/parallel formant synthesizer. J Acoust Soc AM 1980; 67: 971-995.

51. HOLMES J. Formant Synthesizers: cascade or parallel?. Speech Communication 1983; 2: 251-273.

52. SANTOS JM, NOMBELA JR. Text-To-Speech conversion in Spanish: a complete rule-based system. ICASSP 82, Paris 1982; 1593-1596.

53. CARLSON R, GRANSTRÖM B, HUNNICUT S. A multi-language Text-To-Speech module. ICASSP 82, Paris 1982; 3: 1604-1607.

54. O'SHAUGHNESSY D. Design of real -time French Text-To-Speech system. Speech Communication Vol 3: 233-243.

55. BAILLY G, MURILLO G, AL DAKKAK O, GUERIN B. A text-to-speech system for French using formant synthesis. SPEECH '88, 7th FASE Symposium, Edimburgh, U.K 1988, 255-260.

56. DUTOIT T, LEICH H, MBR-PSOLA. Text-To-Speech synthesis based on a MBE re-synthesis of the segments database. Speech Communication 1993; 13: 435-440.

57. MOULINES E, CHARPENTIER F. Pitch synchronous waveform processing techniques for Text-To-Speech synthesis using diphones. Speech Communication 1990; 9: 453-467.