



UNIÓN INTERNACIONAL DE TELECOMUNICACIONES

UIT-T

SECTOR DE NORMALIZACIÓN
DE LAS TELECOMUNICACIONES
DE LA UIT

P.851

(11/2003)

SERIE P: CALIDAD DE TRANSMISIÓN TELEFÓNICA,
INSTALACIONES TELEFÓNICAS Y REDES LOCALES

Métodos de evaluación objetiva y subjetiva de la calidad

**Evaluación de la calidad subjetiva de los
servicios telefónicos basados en sistemas
conversacionales**

Recomendación UIT-T P.851

RECOMENDACIONES UIT-T DE LA SERIE P

CALIDAD DE TRANSMISIÓN TELEFÓNICA, INSTALACIONES TELEFÓNICAS Y REDES LOCALES

Vocabulario y efectos de los parámetros de transmisión sobre la opinión de los clientes	Serie	P.10
Líneas y aparatos de abonado	Serie	P.30
		P.300
Patrones de transmisión	Serie	P.40
Aparatos para mediciones objetivas	Serie	P.50
		P.500
Medidas electroacústicas objetivas	Serie	P.60
Medidas relativas a la sonoridad vocal	Serie	P.70
Métodos de evaluación objetiva y subjetiva de la calidad	Serie	P.80
		P.800
Calidad audiovisual en servicios multimedios	Serie	P.900

Para más información, véase la Lista de Recomendaciones del UIT-T.

Recomendación UIT-T P.851

Evaluación de la calidad subjetiva de los servicios telefónicos basados en sistemas conversacionales

Resumen

En esta Recomendación se describen métodos y procedimientos para llevar a cabo experimentos de evaluación subjetiva de servicios telefónicos que emplean sistemas conversacionales. Los sistemas correspondientes permiten una interacción natural empleando lenguaje oral y disponen de capacidades de reconocimiento e interpretación vocal, gestión del diálogo y emisión vocal. Se describe la configuración y la realización de los experimentos de interacción adecuados, y se proponen cuestionarios para cuantificar los parámetros de calidad pertinentes que percibe el usuario.

Orígenes

La Recomendación UIT-T P.851 fue aprobada el 13 de noviembre de 2003 por la Comisión de Estudio 12 (2001-2004) del UIT-T por el procedimiento de la Recomendación UIT-T A.8.

Palabras clave

Comprensión vocal, evaluación subjetiva, generación vocal, gestión de diálogo, parámetro de interacción, reconocimiento vocal, sistema conversacional.

PREFACIO

La UIT (Unión Internacional de Telecomunicaciones) es el organismo especializado de las Naciones Unidas en el campo de las telecomunicaciones. El UIT-T (Sector de Normalización de las Telecomunicaciones de la UIT) es un órgano permanente de la UIT. Este órgano estudia los aspectos técnicos, de explotación y tarifarios y publica Recomendaciones sobre los mismos, con miras a la normalización de las telecomunicaciones en el plano mundial.

La Asamblea Mundial de Normalización de las Telecomunicaciones (AMNT), que se celebra cada cuatro años, establece los temas que han de estudiar las Comisiones de Estudio del UIT-T, que a su vez producen Recomendaciones sobre dichos temas.

La aprobación de Recomendaciones por los Miembros del UIT-T es el objeto del procedimiento establecido en la Resolución 1 de la AMNT.

En ciertos sectores de la tecnología de la información que corresponden a la esfera de competencia del UIT-T, se preparan las normas necesarias en colaboración con la ISO y la CEI.

NOTA

En esta Recomendación, la expresión "Administración" se utiliza para designar, en forma abreviada, tanto una administración de telecomunicaciones como una empresa de explotación reconocida de telecomunicaciones.

La observancia de esta Recomendación es voluntaria. Ahora bien, la Recomendación puede contener ciertas disposiciones obligatorias (para asegurar, por ejemplo, la aplicabilidad o la interoperabilidad), por lo que la observancia se consigue con el cumplimiento exacto y puntual de todas las disposiciones obligatorias. La obligatoriedad de un elemento preceptivo o requisito se expresa mediante las frases "tener que, haber de, hay que + infinitivo" o el verbo principal en tiempo futuro simple de mandato, en modo afirmativo o negativo. El hecho de que se utilice esta formulación no entraña que la observancia se imponga a ninguna de las partes.

PROPIEDAD INTELECTUAL

La UIT señala a la atención la posibilidad de que la utilización o aplicación de la presente Recomendación suponga el empleo de un derecho de propiedad intelectual reivindicado. La UIT no adopta ninguna posición en cuanto a la demostración, validez o aplicabilidad de los derechos de propiedad intelectual reivindicados, ya sea por los miembros de la UIT o por terceros ajenos al proceso de elaboración de Recomendaciones.

En la fecha de aprobación de la presente Recomendación, la UIT no ha recibido notificación de propiedad intelectual, protegida por patente, que puede ser necesaria para aplicar esta Recomendación. Sin embargo, debe señalarse a los usuarios que puede que esta información no se encuentre totalmente actualizada al respecto, por lo que se les insta encarecidamente a consultar la base de datos sobre patentes de la TSB.

© UIT 2004

Reservados todos los derechos. Ninguna parte de esta publicación puede reproducirse por ningún procedimiento sin previa autorización escrita por parte de la UIT.

ÍNDICE

	Página
1 Alcance	1
2 Referencias	1
3 Abreviaturas.....	2
4 Introducción.....	2
4.1 Funciones y componentes de un sistema conversacional.....	3
4.2 Interacción telefónica con un sistema conversacional.....	3
4.3 Aspectos de calidad y factores que pueden influir	4
4.4 Métodos de evaluación subjetiva.....	8
5 Caracterización del sistema conversacional	9
5.1 Factores de agente	9
5.2 Factores de tarea	12
5.3 Factores de usuario	12
5.4 Factores de entorno.....	13
5.5 Factores de contexto	13
6 Organización del experimento	13
6.1 Organización del sistema y simulación de Mago de Oz.....	14
6.2 Casos de prueba	15
6.3 Agentes de prueba	16
7 Cuestionarios	17
7.1 Preguntas relativas a los antecedentes del usuario	19
7.2 Preguntas relativas a la interacción individual	20
7.3 Preguntas relativas a la impresión general del usuario con relación al sistema	23
8 Evaluación de la capacidad de uso	25
9 Análisis e interpretación de la información recopilada	26
Apéndice I – Ejemplos de hipótesis.....	27
BIBLIOGRAFÍA	29

Recomendación UIT-T P.851

Evaluación de la calidad subjetiva de los servicios telefónicos basados en sistemas conversacionales

1 Alcance

En la presente Recomendación se describen métodos de evaluación subjetiva que proporcionan información relativa a la calidad de los servicios telefónicos basados en sistemas conversacionales, según las experiencias obtenidas por los usuarios de esos servicios. Los sistemas conversacionales que se abordan en esta Recomendación permiten la interacción mediante el lenguaje oral con un usuario físico por la red telefónica basándose en el modo de turno por turno, y disponen de capacidades de reconocimiento y comprensión vocal, gestión del diálogo, generación de respuesta y emisión vocal. Estos sistemas pueden permitir el acceso a información almacenada en una base de datos, o facilitar la realización de distintos tipos de transacciones.

Los métodos de evaluación que se describen en esta Recomendación abordan distintos aspectos de calidad desde la perspectiva del usuario, y consideran el sistema conversacional como una caja negra. Los aspectos importantes de calidad son las posibilidades de uso del servicio, la eficacia de la comunicación, la eficacia de la tarea y del servicio, la satisfacción del usuario, la calidad de la entrada y la salida vocal que se percibe, la capacidad de cooperación del sistema, la simetría de la interacción y la regularidad de la interacción que se percibe. Los métodos se fundamentan en experimentos de laboratorio mediante agentes que interactúan con el sistema conversacional para llevar a cabo una tarea predefinida y realista. La opinión de los agentes con relación a los parámetros de calidad que se perciben puede evaluarse de una manera dirigida o no, mediante cuestionarios que se les proporcionan después del experimento, o con la ayuda de otros métodos de evaluación de las posibilidades de uso. En esta Recomendación se describen la configuración y realización de los experimentos de interacción, los parámetros de calidad pertinentes percibidos por el usuario y los métodos que permitirán proporcionar la información relativa a estos parámetros de calidad. En las Recomendaciones UIT-T P.800 y P.85, así como en el Manual sobre telefonometría, pueden obtenerse directrices adicionales relativas a los métodos de evaluación subjetiva en general y con relación a la evaluación de los dispositivos con emisión vocal.

2 Referencias

Las siguientes Recomendaciones del UIT-T y otras referencias contienen disposiciones que, mediante su referencia en este texto, constituyen disposiciones de la presente Recomendación. Al efectuar esta publicación, estaban en vigor las ediciones indicadas. Todas las Recomendaciones y otras referencias son objeto de revisiones por lo que se preconiza que los usuarios de esta Recomendación investiguen la posibilidad de aplicar las ediciones más recientes de las Recomendaciones y otras referencias citadas a continuación. Se publica periódicamente una lista de las Recomendaciones UIT-T actualmente vigentes. En esta Recomendación, la referencia a un documento, en tanto que autónomo, no le otorga el rango de una Recomendación.

- Recomendación UIT-T E.800 (1994), *Términos y definiciones relativos a la calidad de servicio y a la calidad de funcionamiento de la red, incluida la seguridad de funcionamiento.*
- Recomendación UIT-T G.107 (2003), *El modelo E, un modelo informático para utilización en planificación de la transmisión.*
- Recomendación UIT-T G.1000 (2001), *Calidad de servicio de las comunicaciones: Marco y definiciones.*

- Recomendación UIT-T P.85 (1994), *Método para la evaluación subjetiva de la calidad vocal de los dispositivos generadores de voz.*
- Recomendación UIT-T P.800 (1996), *Métodos de determinación subjetiva de la calidad de transmisión.*
- *Manual sobre telefonometría del UIT-T* (1992).

3 Abreviaturas

En esta Recomendación se utilizan las siguientes siglas.

ACR	Evaluación por categorías absolutas (<i>absolute category rating</i>)
ANOVA	Análisis de la varianza (<i>analysis of variance</i>)
ASR	Reconocimiento automático de la voz (<i>automatic speech recognition</i>)
CCR	Evaluación por categorías de comparación (<i>comparison category rating</i>)
DARPA	Agencia de proyectos de investigación avanzada para defensa (<i>defense advanced research projects agency</i>)
DCR	Evaluación por categorías de degradación (<i>degradation category rating</i>)
DTMF	Multifrecuencia bitono (<i>dual tone multiple frequency</i>)
HMM	Modelo oculto de Markov (<i>hidden Markov model</i>)
HSD	Diferencia enteramente significativa (<i>honestly significant difference</i>)
MLP	Perceptron multicapas (<i>multi-layer perceptron</i>)
MOS	Nota media de opinión (<i>mean opinion score</i>)
PARADISE	Paradigma de evaluación del sistema de diálogo (<i>paradigm for dialogue system evaluation</i>)
QoS	Calidad de servicio (<i>quality of service</i>)
SDS	Sistema conversacional (<i>spoken dialogue system</i>)
WoZ	Mago de Oz (<i>wizard-of-Oz</i>)

4 Introducción

Es posible que los sistemas conversacionales (SDS), es decir los sistemas informáticos mediante los cuales los usuarios físicos interactúan con lenguaje oral, turno por turno, lleguen a ser parte integral de las redes telefónicas modernas. Estos sistemas permiten el acceso a bases de datos y transacciones a través del aparato telefónico, por ejemplo, para obtener información de horarios de trenes o aviones, movimientos de bolsa, información turística, realización de operaciones de cuenta bancaria, reservaciones de hotel, y otras. En comparación con los sistemas DTMF simples, los sistemas conversacionales disponen de capacidades de reconocimiento vocal automático y de comprensión vocal (es decir, sintáctica/semántica/pragmática y por consiguiente interpretativa), y de un módulo de gestión de diálogo que garantiza el funcionamiento regular y natural de la interacción oral entre el usuario y el sistema. Como resultado, la interacción se vuelve más parecida a la humana, y el servicio que proporcionan esos sistemas podrá interesar a una amplia gama de usuarios potenciales. Frecuentemente, los tipos de sistemas basados en DTMF y en la conversación se realizan de un modo integrado, y una parte de los aspectos de calidad correspondientes será idéntica para ambos tipos. Algunas veces, los tipos de sistemas basados en conversación utilizan estructuras y protocolos de interfaz que se emplean en entornos de aplicación web, y se construyen

de una manera similar a las interfaces web; por consiguiente, estas interfaces pueden constituir una referencia para obtener la misma funcionalidad.

4.1 Funciones y componentes de un sistema conversacional

Desde un punto de vista técnico, se pueden visualizar mejor los componentes de un sistema conversacional que funciona por la red telefónica, a partir de una estructura secuencial. En la figura 1 se ilustra un ejemplo de ese tipo de estructura que consta de seis componentes principales a los que puede acceder el usuario a través de la interfaz del servidor telefónico. En primer lugar, la unidad de reconocimiento vocal procesa la señal vocal del usuario. Durante el proceso de reconocimiento, la señal se transforma en una cadena de palabras o en una gráfica de posibilidades de palabras que se analiza a continuación desde un punto de vista semántico. La salida da por resultado una trama semántica que representa lo que se ha "comprendido" a partir de las unidades de habla del usuario. El gestor del diálogo debe interpretar la trama semántica en el contexto del diálogo y de la tarea, y mantener el seguimiento de la historia del diálogo. Cuando se ha recopilado toda la información pertinente del usuario, podrá iniciarse una consulta a la aplicación subyacente (que en este ejemplo es una base de datos). La información que se origina a partir del programa de aplicación, así como otros objetivos de comunicación del gestor del diálogo, se transforman en una respuesta al usuario. Ésta es la tarea del módulo de generación de respuesta el cual da una respuesta en forma de texto, que a su vez el sintetizador vocal la transforma en una señal vocal que se transmite al usuario físico. En algunas ocasiones, la generación de respuesta y la síntesis vocal se realizan en un módulo simple (sin pasar por la representación textual), y se utilizan mensajes grabados en lugar de voz sintetizada.

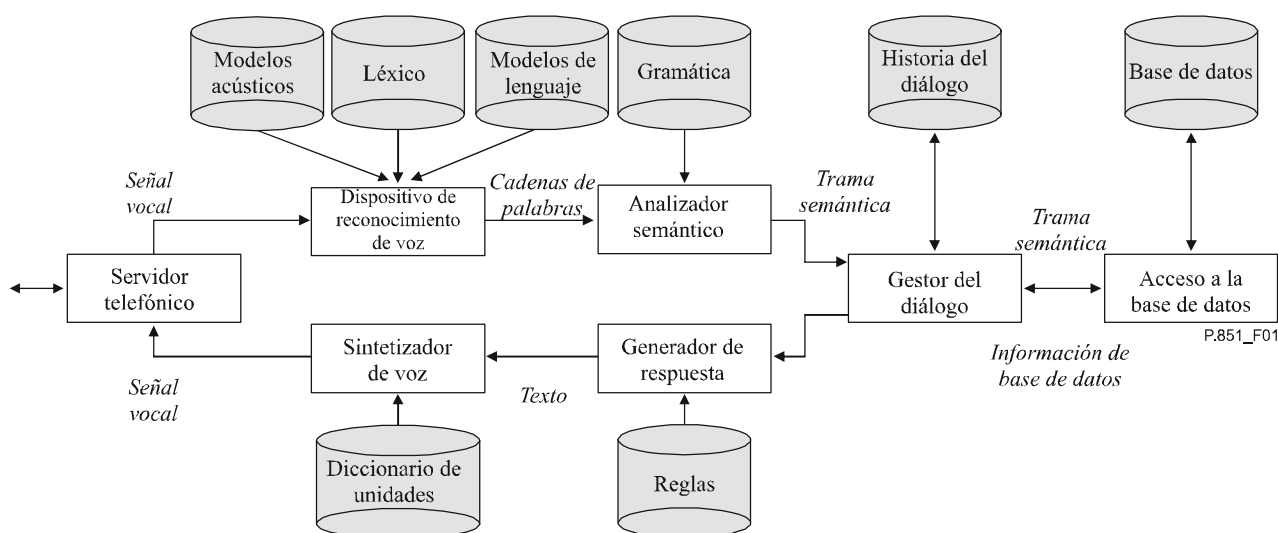


Figura 1/P.851 – Estructura secuencial de un sistema conversacional basado en la telefonía [27], [33]

Esta estructura básica podrá realizarse de distintas maneras. En [4] podrán encontrarse algunos ejemplos. Una manera común es la denominada "arquitectura centralizada" [37], [45] que se utilizó en el proyecto Communicator de DARPA. Otras estructuras se apoyan en módulos de funcionamiento asíncrono para la interpretación, comportamiento (razonamiento y actuación) y generación, véase [1].

4.2 Interacción telefónica con un sistema conversacional

La interacción con el sistema conversacional se lleva a cabo a través de algún tipo de red de telecomunicaciones. Esta red introducirá probablemente una determinada cantidad de degradaciones

de transmisión que repercutirán en la calidad de la voz que se transmite, y por consiguiente también en la calidad de funcionamiento de la unidad de reconocimiento vocal, y de otros componentes subsiguientes de tecnología de lenguaje vocal y natural en el sistema conversacional. En la ruta de retorno al usuario físico, el canal de transmisión degradará la señal vocal generada por el sistema conversacional. Ya que las redes de telecomunicaciones se verán confrontadas con las comunicaciones persona a persona así como con los casos de interacción entre las personas y las máquinas, es importante tener en cuenta los requisitos tanto del usuario físico como del dispositivo de tecnología vocal. Evidentemente, estos requisitos serán diferentes, ya que las características de percepción que tienen influencia sobre la opinión del usuario con relación a la calidad no son idénticas a las características del dispositivo de tecnología vocal, por ejemplo, de una unidad de reconocimiento automático de la voz (ASR, *automatic speech recognition*).

El usuario físico realiza la interacción a través de algún tipo de interfaz de usuario, por ejemplo, un auricular telefónico, un terminal manos libres o un casco telefónico. Las características acústicas de las interfaces indicadas son muy diversas, así como su sensibilidad a los fenómenos acústicos del recinto que se producen en el entorno de habla y escucha del usuario. Por ejemplo, el ruido ambiental puede repercutir significativamente en la inteligibilidad de las señales vocales que se transmiten por un terminal manos libres, y además se añade la influencia del comportamiento de habla del usuario. Como resultado, es necesario tener en cuenta todos los elementos de interacción incluidos el sistema conversacional, el canal de transmisión y la interfaz del usuario para determinar la calidad total de la interacción.

4.3 Aspectos de calidad y factores que pueden influir

Las personas son los usuarios de los servicios basados en SDS que se ofrecen por la línea telefónica. Por consiguiente, deben tenerse en cuenta los factores humanos para determinar las funciones de un sistema/servicio y su grado de cumplimiento. La calidad de servicio resulta realmente de las percepciones de su usuario, con relación a lo que espera o desea de ese servicio. De acuerdo a la definición de calidad de [24], la *calidad de un servicio basado en un sistema conversacional* es el resultado de la evaluación de la composición percibida del servicio con relación a la composición deseada. Por consiguiente, la calidad que percibe el usuario es un compromiso entre lo que espera o desea, y las características que percibe mientras lo utiliza. La calidad depende, en gran medida, de la situación en la que se realiza la percepción y se emite la opinión. Este hecho habrá de tenerse en cuenta cuando se realicen experimentos de evaluación subjetiva de la calidad, concretamente creando una situación de prueba más o menos natural y una motivación del usuario con una prueba realista.

A diferencia de la noción de calidad de transmisión vocal en los casos de comunicación persona a persona, el usuario de un servicio basado en un sistema conversacional adquiere una parte activa en la producción vocal y en el flujo del diálogo. Por consiguiente, las características y el comportamiento del usuario serán decisivos para el cumplimiento de la tarea deseada. Por lo tanto, las características del sistema y del servicio se verán considerablemente influenciadas por el usuario. A fin de poder describir el funcionamiento del sistema y del usuario de una manera simplificada, se podrán registrar algunos parámetros durante la interacción. Esos parámetros de interacción pueden medirse con instrumentos, aunque no es indispensable. Algunos ejemplos incluyen el número de palabras o la duración de un diálogo que puede medirse con instrumentos, o una tasa de errores de palabras y una medición de éxito de la tarea que sólo podrán determinarse con ayuda de expertos. En [18] y [33] se presenta una síntesis de los parámetros de interacción.

En principio, la calidad de un servicio basado en un sistema conversacional puede abordarse desde dos puntos de vista distintos: el del proveedor de servicio y el del usuario¹. Al proveedor de servicio le interesan principalmente los efectos de cada elemento del servicio, y cómo se relacionan con el grado de satisfacción o aceptabilidad del usuario. Los proveedores de servicio utilizan la definición de calidad de servicio (QoS) de la Rec. UIT-T E.800. El usuario percibe y refleja las características que experimenta (prestaciones) del servicio, compara las percepciones con algún tipo de referencia interna y las examina con relación al cumplimiento de sus expectativas o deseos. Cuando se examina la calidad de un servicio, es importante tener en cuenta ambos puntos de vista. Los métodos de evaluación subjetiva como los que se describen en esta Recomendación se concentrarán en el punto de vista del usuario. No obstante, también serán útiles para el proveedor de servicios, ya que proporcionan indicaciones con relación a las características del servicio que deben mejorarse.

Tanto la *eficacia* como la *eficiencia* se relacionan a la calidad de funcionamiento para lograr el objetivo para el que se despliega el servicio. La eficacia es un índice absoluto que describe hasta qué grado se alcanza el objetivo, con relación a la precisión y perfección de los objetivos, véase por ejemplo [14]:

"*Eficacia*: precisión y perfección con la que algunos usuarios especificados pueden lograr objetivos específicos en entornos particulares."

Las mediciones de la eficacia que pueden encontrarse en los libros son por ejemplo, el éxito de la tarea o los criterios de medición kappa [33]. Por otro lado, la eficiencia, es una medición relativa del logro del objetivo con relación a los recursos utilizados [14]:

"*Eficiencia*: relación entre los recursos dedicados y la precisión y perfección de los objetivos logrados."

Los criterios de medición que se utilizan comúnmente son, por ejemplo, la duración del diálogo o la cantidad de cambios producidos por el sistema o por el usuario.

La eficiencia y la demanda cognoscitiva son criterios que caracterizan un sistema con el que un usuario puede lograr sus objetivos. No obstante, *la capacidad de uso*, por lo general se define en un sentido mucho más amplio, y describe la capacidad en la que un servicio puede ser comprendido, aprendido y utilizado por usuarios especificados bajo condiciones específicas. Indica la idoneidad del servicio para cumplir con los requisitos del usuario, incluye la eficacia y la eficiencia del sistema y resulta en la satisfacción del usuario [35]. La *satisfacción del usuario* es un indicador de la utilidad y la capacidad de uso que se percibe por el grupo destinado de usuarios del servicio. Incluye si el usuario obtiene la información que desea, si está satisfecho con el servicio y si obtiene la información dentro de un periodo de tiempo aceptable [31].

Las nociones descritas de la calidad de un servicio conversacional basado en la telefonía pueden ilustrarse en términos de un diagrama, como se ha propuesto en [34], y que se muestra en la figura 2. Aparte de los factores del usuario ya mencionados, otros cuatro tipos de factores contribuyen a la calidad percibida por el usuario: factores de agente (relacionados principalmente con el diálogo y el propio sistema), factores de tarea (relacionados con la forma cómo el sistema conversacional registra la tarea para la que ha sido desarrollado), factores de entorno (por ejemplo, factores relacionados con el entorno acústico y el canal de transmisión), y factores de contexto tales como los costos, tipo de acceso o la disponibilidad. Los aspectos de calidad percibidos por el usuario se ilustran en la parte inferior del diagrama.

¹ En la Rec. UIT-T G.1000 se describen incluso cuatro puntos de vista distintos. Los requisitos del cliente con relación a la QoS, la oferta de QoS del proveedor de servicio, la QoS alcanzada u ofrecida por el proveedor, y la QoS percibida por el cliente.

Los factores de entorno, de agente y de tarea tienen influencia sobre la *calidad de entrada y salida vocal*, sobre la *cooperatividad* del comportamiento del sistema, y sobre la *simetría* de la interacción del diálogo. La calidad de entrada y salida vocal incluye aspectos como inteligibilidad, naturalidad, esfuerzo de escucha necesario para comprender los mensajes del sistema, o la comprensión percibida del sistema. En este documento se define la cooperatividad en el sentido de no violación de los principios del comportamiento del diálogo cooperativo, como se define en [20] por Grice. Incluye los aspectos de capacidad de información, verdad y evidencia, significación, forma, conocimiento previo, y tratamiento de metacomunicaciones (es decir, confirmación, aclaración, reparación y recuperación de errores de comunicaciones), véase [6]. El aspecto de la asimetría de la contraparte (diferencias en el comportamiento de interacción que se atribuyen a la asimetría de las partes que interactúan) se trata mediante una categoría denominada simetría de diálogo. Esta categoría también incluye los efectos de las capacidades de iniciativa de diálogo y de control de interacción.

Los aspectos de calidad mencionados dan por resultado una comunicación (interacción) (más o menos) eficiente, y una solución eficiente de la tarea que se debe realizar. La *eficiencia de comunicación* se relaciona con la velocidad o ritmo de la interacción, la concisión y la homogeneidad del diálogo. Por otro lado, *la eficiencia de la tarea*, está vinculada al éxito y a la facilidad de ésta. Hay otros dos aspectos de calidad adicionales que son importantes: la "personalidad" del agente de la máquina (cortesía, simpatía, naturalidad de comportamiento) y el esfuerzo necesario del usuario físico para la interacción (facilidad de comunicación, estrés, nervios, etc.). Estos aspectos se incluyen en el término *comodidad*.

La eficiencia de la comunicación, la eficiencia de la tarea y la comodidad contribuyen a la capacidad de uso del servicio, por lo cual la satisfacción del usuario puede considerarse como un indicador. La *eficiencia del servicio*, por otro lado, se ve influenciada por la eficiencia de la tarea y por los factores de contexto. Es importante para la idoneidad del servicio (para cumplir con la tarea deseada), y para el valor añadido que se atribuye al servicio [por ejemplo, en comparación con métodos similares para obtener la misma información, como una interfaz web o un rastreador de noticias (*new sticker*)]. La capacidad de uso, la eficiencia del servicio y la *ventaja económica* dan por resultado la *utilidad* del servicio, y finalmente su *aceptabilidad*.

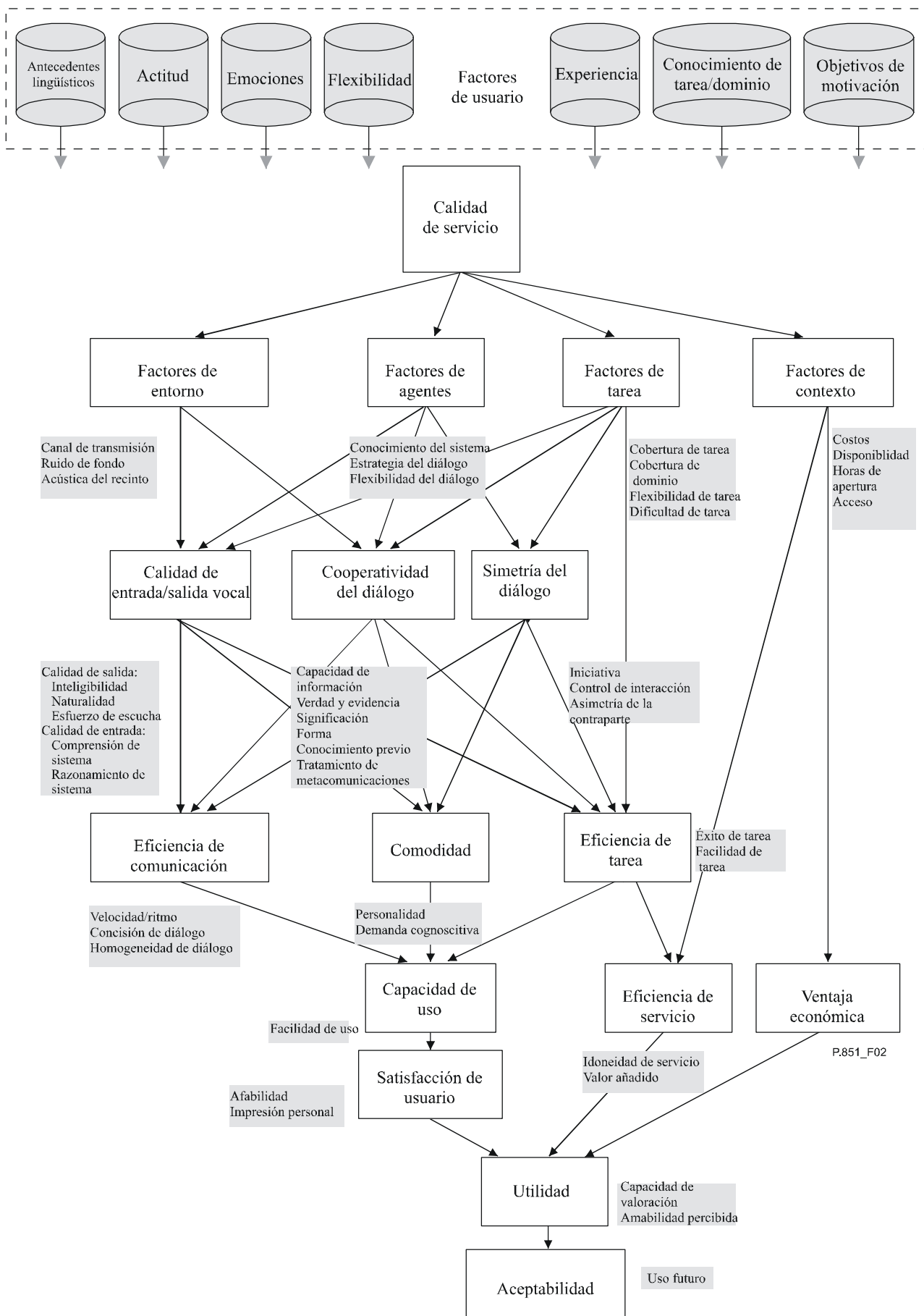


Figura 2/P.851 – Aspectos de calidad y factores que influyen; véase [34]

4.4 Métodos de evaluación subjetiva

Los sistemas conversacionales pueden evaluarse a nivel de componente (por ejemplo, la unidad de reconocimiento vocal, al componente de comprensión vocal, o al componente de salida vocal), o con relación al sistema total (integrado). La evaluación analítica a nivel de componente es una fuente de información valiosa para describir cómo cumplen su tarea las partes individuales del sistema. No obstante, es posible que con ella se escapen algunas veces contribuciones importantes a la calidad total del servicio, conforme a la percepción del usuario. Por ejemplo, el reconocimiento o la comprensión vocal erróneos pueden compensarse mediante el componente de tratamiento de la conversación, sin afectar la calidad del sistema total. Por esta razón, resultan indispensables los experimentos subjetivos con usuarios reales o de prueba que interactúen con todo el sistema cuando se va a determinar la calidad de un servicio basado en un sistema conversacional.

Para evaluar los distintos aspectos de la calidad de un servicio basado en un sistema conversacional, se tienen que realizar experimentos subjetivos con usuarios físicos. Estos experimentos sirven para dos finalidades principales:

- 1) Durante la interacción, se recopilan parámetros del sistema que pueden medirse con instrumentos, y se registran las palabras del sistema y del usuario. Los ficheros de registro se envían a la evaluación de un experto, cuyo resultado es un conjunto de parámetros que describen aspectos particulares de la interacción persona-máquina con relación al nivel de palabra, diálogo y tarea, desde el punto de vista de un desarrollador de sistema.
- 2) Después de la interacción, los agentes de prueba reciben un cuestionario que persigue recopilar información relativa a las características de calidad percibidas que son importantes para integrar la impresión de calidad total del usuario físico. Esos experimentos podrán realizarse con sistemas funcionales completos, o con sistemas que estén aún en la fase de desarrollo y en los que se tengan que simular partes de los módulos del sistema. En las cláusulas 6 a 8 se presentan los detalles relativos a la organización experimental, a los cuestionarios y a los métodos de evaluación de la capacidad de uso.

En experimentos de laboratorio, se pueden obtener ambos tipos de información en paralelo. No obstante, en una situación de prueba de campo con usuarios reales, los parámetros de interacción registrados mediante instrumentos son a menudo la única fuente de información para el proveedor de servicio a fin de que pueda supervisar la calidad del sistema. La cantidad de datos que pueden recopilarse de un servicio en funcionamiento puede resultar muy grande. En este caso, es importante definir un conjunto central de criterios de medición que describan la calidad de funcionamiento del sistema, y disponer de herramientas que permitan automatizar una gran parte del proceso de análisis de los datos. Por consiguiente, la tarea del evaluador físico es analizar e interpretar esos datos, y estimar el efecto de las mediciones de calidad de funcionamiento recopiladas sobre la calidad que sería percibida por un usuario (prototipo). En la cláusula 9 se presentan algunas consideraciones generales relativas al análisis e interpretación de los resultados de prueba. Siempre que estén disponibles ambos tipos de información, pueden establecerse relaciones entre los parámetros de interacción y las opiniones subjetivas. Esos modelos de predicción de calidad para los sistemas conversacionales basados en telefonía se encuentran aún en estudio, y en la cláusula 9 se presenta una breve discusión.

Como es práctica común en los experimentos de evaluación subjetiva, el objetivo y las circunstancias de un experimento de valoración o evaluación deberían manifestarse explícitamente, y documentarse. En el proyecto DISC europeo, se ha desarrollado una plantilla para esta finalidad [5]. Basándose en dicha plantilla y en métodos de clasificación dados en [33], pueden definirse los siguientes criterios:

- Motivación de la valoración/evaluación (por ejemplo, un análisis detallado de los mecanismos de recuperación del sistema, o la estimación de la satisfacción de usuarios futuros).

- Objeto de la valoración/evaluación (por ejemplo, la unidad de reconocimiento vocal, el gestor de diálogo o el sistema completo).
- Entorno de la valoración/evaluación (por ejemplo, en un experimento de un laboratorio controlado o en un campo de prueba).
- Tipo de métodos de medición (por ejemplo, a través de una medición de los parámetros de interacción mediante instrumentos, o a través de opiniones de calidad abiertas o cerradas que se obtienen de los usuarios).
- Síntomas que se pueden esperar (por ejemplo, preguntas de aclaración de los usuarios o rechazos de ASR).
- Fase de vida útil durante la cual se realiza la valoración/evaluación (por ejemplo, de una simulación, una versión prototipo o de un sistema en funcionamiento completo).
- Accesibilidad del sistema y de sus componentes (por ejemplo, con un método de caja de cristal o de caja negra).
- Referencia utilizada para las mediciones (por ejemplo, mediciones cualitativas de la calidad de funcionamiento del sistema absoluto, o valores cuantitativos con relación a una referencia que se puede medir o a un banco de pruebas).
- Herramientas de soporte disponibles para la valoración/evaluación.

Estos criterios constituyen un conjunto básico de la documentación que se deberá proporcionar con los experimentos de valoración o evaluación. La documentación podrá implementarse en términos de una relación de puntos como se da aquí, o mediante una descripción experimental detallada.

5 Caracterización del sistema conversacional

Si se sigue el diagrama de aspectos de calidad de la figura 2, hay cinco tipos de factores que pueden tener influencia sobre la interacción con un servicio basado en un sistema conversacional: *factores de agente*, *factores de tarea*, *factores de usuario*, *factores de entorno* y *factores de contexto*. Estos factores determinarán la calidad de funcionamiento del sistema (componentes) y la calidad que percibe el usuario. Por consiguiente, tendrán que tenerse en cuenta cuando se realicen y documenten experimentos de evaluación subjetiva.

5.1 Factores de agente

El sistema, considerado como un par de interacción, puede caracterizarse de una manera técnica, concretamente definiendo las características de cada uno de sus componentes y de su interconexión en la estructura secuencial de la figura 1, o especificando las funciones operacionales del agente. Las funciones más importantes del agente que se han de caracterizar son la capacidad de reconocimiento vocal, la capacidad de comprensión del lenguaje natural, la capacidad de gestión del diálogo, la capacidad de generación de respuesta y la capacidad de salida vocal. Los componentes correspondientes a la comprensión del lenguaje natural y a la generación de respuesta están vinculados estrechamente a los componentes colindantes, a saber el gestor de diálogo en un lado y la unidad de reconocimiento vocal o el sintetizador vocal en el otro. Por consiguiente, han de describirse con precisión las interfaces para esos componentes.

5.1.1 Caracterización de la unidad de reconocimiento vocal

Desde un punto de vista funcional, las unidades de reconocimiento vocal pueden clasificarse conforme a los siguientes parámetros [41]:

- Tamaño de vocabulario, por ejemplo, unidades de reconocimiento vocal con vocabulario reducido, mediano o grande.

- Complejidad de vocabulario, por ejemplo, con respecto a la posibilidad de confusión de las palabras.
- Tipo de habla, por ejemplo, palabras aisladas, palabras conectadas, voz continua, voz espontánea incluyendo discontinuidades como tos, titubeos, interrupciones, reinicios, y otros.
- Lenguaje: unidades de reconocimiento monolingües o multilingües, dependencia del lenguaje de los resultados de reconocimiento, transferibilidad del lenguaje.
- Dependencia del hablante; por ejemplo unidades de reconocimiento que dependan del hablante, independientes de éste o que se adapten a él.
- Tipo y complejidad de la gramática. La complejidad de una gramática puede establecerse en términos de su perplejidad, que es una medida de lo bien que puede predecir el modelo de lenguaje una secuencia de palabras.
- Método de aprendizaje, por ejemplo, aprendizaje múltiple de palabras aisladas pronunciadas explícitamente, o entrenamiento integrado en cadenas de palabras de las cuales no se definen los puntos de inicio y de terminación.

Por otro lado, los componentes de la unidad de reconocimiento vocal podrán describirse en términos de características técnicas generales que podrán realizarse de modo diferente en sistemas particulares [27]. Las siguientes características técnicas han sido utilizadas parcialmente en el proyecto DISC:

- Registro de señal: frecuencia de muestreo, ancho de banda de señal, cuantificación, visualización de información en recuadros.
- Análisis de características, por ejemplo, coeficientes cepstral en escala mel, energía y derivadas de primero y segundo orden.
- Unidades vocales fundamentales, por ejemplo, modelos telefónicos o modelos de palabras, modelización de silencio o de otros sonidos no vocales.
- Léxico: número de asientos por cada palabra, con una o varias pronunciaciones; generadas ya sea a partir de diccionarios o de convertidores de grafema a fonema; asientos adicionales para palabras y ruidos de relleno; cobertura prevista del vocabulario con relación al vocabulario objetivo.
- Modelo acústico: tipo de modelo, por ejemplo, redes tipo Perceptron multiniveles (MLP, *multi-layer-Perceptron*) o modelos ocultos de Markov (HMM, *hidden Markov models*); datos y parámetros de aprendizaje; postratamiento del modelo.
- Modelo de lenguaje: tipo de modelo, por ejemplo, un modelo de lenguaje estadístico tipo N-gram back-off (retirada), o una gramática sin contexto; material de aprendizaje, por ejemplo, un gran cuerpo de capacitación de propósito general o datos recopilados de un experimento limitado; modelización de palabras o clases de categorías específicas (por ejemplo, fechas o nombres); modelos independientes del estado del diálogo o dependientes del estado del diálogo.
- Tipo de decodificador, por ejemplo, basado en HMM.
- Grado de uso de información prosódica.

5.1.2 Caracterización de la comprensión del lenguaje

Las siguientes características son importantes para la capacidad de comprensión del lenguaje del sistema:

- Descripción semántica de la tarea, por ejemplo, mediante intervalos de tiempo (pares atributo-valor).

- Análisis sintáctico-semántico: capacidad de análisis sintáctico general, por ejemplo, análisis sintáctico completo o parcial robusto; número y complejidad de la sintaxis permitida, por ejemplo, número de alternativas disponibles a un determinado nivel.
- Análisis de contexto: número y complejidad de las reglas.
- Interacción con los módulos de reconocimiento vocal y de gestión del diálogo: tipo y cantidad de información de entrada y de salida (hipótesis simples, relaciones clasificadas, etc.), dependencia de la interpretación sintáctica-semántica y de contexto relativa al estado del diálogo.

5.1.3 Caracterización del gestor del diálogo

El método elegido para la gestión del diálogo puede definirse desde un punto de vista técnico, por ejemplo, como una gramática de diálogo, un método basado en planificación o un método colaborativo [8], [32]. Las características más importantes del gestor de diálogo son el tipo y la cantidad de conocimiento asignado al gestor, la distribución de iniciativa entre el sistema y el usuario, y las estrategias de metacomunicación del sistema (confirmación, aclaración, reparación y recuperación):

- Conocimiento del gestor de diálogo: modelo de historia del diálogo (información que se ha intercambiado en el diálogo hasta el momento), modelos de tarea y dominio (caso, planificaciones, objetivos y subobjetivos, objetos y sus características), modelo de conocimiento global, modelo conversacional y modelo de usuario.
- Iniciativa: iniciativa de sistema, iniciativa mixta o iniciativa de usuario.
- Estrategia de confirmación: confirmación explícita, confirmación implícita, confirmación "eco", confirmación resumida.
- Estrategias relativas a la reparación, aclaración y recuperación.
- Capacidad de adaptación del gestor de diálogo: gestores esenciales que tienen que aprender nuevas nociones durante su funcionamiento normal, o gestores adaptativos que podrían incluir un modelo de usuario dinámico y podrían ser capaces de aprender las estrategias de comunicación del usuario.

Además de la interacción con el usuario, se ha de definir la interacción con el sistema de aplicación, incluida la interfaz (lenguaje de programación), y los posibles mecanismos de control para manejar la dinámica del sistema de aplicación.

5.1.4 Caracterización de la generación vocal

La generación vocal incluye la posible generación de una respuesta de texto, y la traducción a lenguaje oral. La mayor parte de los sistemas utilizan uno de los tres tipos siguientes de generación vocal: voz pregrabada, oraciones modelo o texto a voz. Es necesario definir las siguientes características:

- Interacción con el gestor de diálogo: tipo y cantidad de información de entrada proporcionada por el gestor de diálogo, por ejemplo, texto ortográfico o comentado, información focal o prosódica, etc.
- Generación de respuesta: estrategia (por ejemplo, gramática formal o plantillas simples), flexibilidad (vocabulario predefinido o vocabulario abierto), tipo y cantidad de información que debe incluirse en cada unidad de habla, forma del mensaje (sintaxis, selección de palabras).
- Voz del sistema: número de voces, género, profesionalismo, aprendizaje, calidad prosódica, condiciones de registro, capacidad de adaptación.
- Lenguaje: sintetizadores monolingües o multilingües, capacidad de identificación del lenguaje, transferibilidad de lenguaje.

- Tipo de generación vocal: mensajes pregrabados, oraciones modelo, texto-voz, concepto-voz.
- Características de la generación texto a voz: estrategia (por ejemplo basado en modelo o basado en cuerpo), capacidades de preprocesamiento de texto, parámetros de modelo, características del cuerpo unitario (tipos y longitud de las unidades, cobertura del vocabulario objetivo, etc.), algoritmos de concatenación y/o selección, estrategias de generación de prosodia (frecuencia fundamental, duración, intensidad), etc.
- Características de contexto: estilo de habla, velocidad de habla, capacidad de adaptación de contexto.

5.2 Factores de tarea

La tarea que va a realizar el usuario es un factor determinante de la interacción. Puede caracterizarse con relación al tipo de tarea, dominio de tarea, complejidad de tarea, frecuencia de tarea, consecuencias de tarea y transferibilidad:

- Tipo de tarea: puede diferenciarse conforme a [6] entre:
 - tareas bien estructuradas, que tienen una estructura estereotípica que recomienda qué parte de la información debe intercambiarse, y a menudo también en qué orden natural; y
 - tareas mal estructuradas, que contienen un gran número de subtareas facultativas cuya naturaleza y orden son difíciles de predecir;
 así como entre:
 - homogéneo; y
 - heterogéneo, que significa *inherentemente* una combinación de varias tareas distintas que difieren con relación a su naturaleza real (por ejemplo, orden más información más control de dispositivo).
- Dominio de tarea: riqueza (sonoridad), escalabilidad, número de usuarios familiarizados con el dominio, utilidad del dominio, generalizabilidad, etc.
- Complejidad de tarea: número de casos cubiertos, número máximo de subobjetivos, número de subtareas que pueden completarse en paralelo, número mínimo de intercambios necesarios para resolver el problema, complejidad prevista de sintaxis/vocabulario, etc.
- Frecuencia de tarea, es decir, la frecuencia con la que los usuarios prevén utilizar el sistema para una tarea determinada. Los sistemas de encaminamiento de llamadas o de información de vuelos (denominados "sistemas para acercarse y utilizarlos") se emplearán con una frecuencia relativamente reducida, de modo que no puede preverse que los posibles usuarios tengan conocimiento del sistema, ni muestren efectos de aprendizaje (como recordar su comportamiento durante llamadas anteriores) o que acepten la capacitación correspondiente.
- Consecuencias de tarea, por ejemplo, cuestiones de seguridad.
- Transferibilidad de tarea.

5.3 Factores de usuario

En la mayoría de los casos, la caracterización del usuario se limita a una categorización amplia con relación a su tarea, contexto de dominio y global, ya que no podrá lograrse una descripción exacta de los factores importantes de un usuario particular (actitud, motivación, emociones, flexibilidad). Las siguientes características se incluyen a menudo en los protocolos de evaluación:

- Número de usuarios.

- Edad y género: se prevé que estas características influirán en la frecuencia fundamental y en el espectro vocal, aunque también en la interacción del diálogo.
- Nivel de experiencia: principiante en oposición a experimentado, usuario ocasional en oposición a usuario regular, usuario capacitado en oposición a usuario no capacitado.
- Nivel de pericia en el dominio de aplicación: usuarios profesionales en oposición a usuarios privados.
- Motivación explícita para utilización del servicio.
- Estado físico, esfuerzo vocal, velocidad de habla, etc.
- Lenguaje nativo, acento, dialecto, etc.

En el caso de aplicaciones especializadas, podría resultar necesario un mayor detalle al especificar la experiencia y la pericia, por ejemplo, con relación al conocimiento de los objetivos de la tarea, la capacidad para desarrollar estrategias tendientes a optimizar el rendimiento de la tarea, y la capacidad para utilizar los dispositivos necesarios para realizar la tarea [29].

5.4 Factores de entorno

El entorno incluye todo el contexto físico de la interacción. Por lo general, será imposible una caracterización total, y sólo deben describirse los factores que repercuten directamente en la señal vocal, concretamente:

- Tipo y propiedades acústicas de la interfaz de usuario.
- Canal de transmisión telefónica: la descripción puede realizarse en distintos niveles, por ejemplo, en términos del equipo de transmisión, conmutación y terminal que se emplea en la conexión, o en términos de los parámetros de una configuración de referencia de planificación de la red, véase la Rec. UIT-T G.107.
- Situación acústica del recinto: incluye reverberación, coloración de sonido, niveles y espectros de ruido ambiental, oradores simultáneos, etc.

5.5 Factores de contexto

Se trata de factores no físicos que caracterizan el contexto de utilización del servicio que se está considerando. Los factores comunes incluyen:

- Facilidad de acceso: disponibilidad de números telefónicos, enlaces a otros servicios y desde éstos, etc.
- Disponibilidad de servicio: horarios de apertura, posibles restricciones de acceso.
- Costos: costos de interacción fijos y los que varían con el tiempo, condiciones de cuenta particulares, etc.
- Servicios con funcionalidad similar: los que tienen que compararse con relación al resto de los factores de contexto.

6 Organización del experimento

Los experimentos de interacción subjetiva con un sistema conversacional deben organizarse y configurarse conforme a las reglas generales de las pruebas de opinión sobre la conversación correspondientes a la Rec. UIT-T P.800. En el Manual sobre telefonometría del UIT-T puede encontrarse una descripción más detallada de las cuestiones prácticas sobre este tema. Este principio se aplica a las condiciones físicas de los gabinetes de prueba, a las características de ruido en el entorno, al diseño experimental y a las reglas generales de análisis de datos. En las siguientes cláusulas, sólo se describen las cuestiones que son particulares a la interacción del sistema conversacional, a saber, la organización del sistema, los casos de prueba y los agentes de prueba.

Los experimentos subjetivos pueden realizarse con sistemas en funcionamiento total, o con la ayuda de un experimentador humano que simula las partes faltantes del sistema, o el sistema completo (denominada "simulación de Mago de Oz"). A fin de obtener resultados válidos y fiables, el sistema (simulado), los usuarios de prueba y la tarea experimental tienen que cumplir con diversos requisitos, véanse las cláusulas 6.1 a 6.3. Por lo general, un experto registra y anota las interacciones de modo que puedan calcularse los parámetros correspondientes. Después de cada interacción y después de la sesión de prueba completa, los agentes de prueba tienen que rellenar algunos cuestionarios que permiten la cuantificación de distintos aspectos de la calidad de un servicio basado en un sistema conversacional. El diseño de esos cuestionarios se examina en la cláusula 7. En la cláusula 8 se presenta una síntesis de los métodos de evaluación que abordan la capacidad de utilización de los servicios.

6.1 Organización del sistema y simulación de Mago de Oz

Para realizar experimentos de interacción con usuarios físicos, ha de implementarse una disposición que ofrezca una funcionalidad completa. La naturaleza exacta de la configuración dependerá de la disponibilidad de los componentes del sistema, y por consiguiente, de la fase de desarrollo del sistema. Si los componentes del sistema aún no han sido implementados, o si una determinada implementación no puede ser factible (por ejemplo, debido a la falta de datos) o no resulta económica, será indispensable la simulación de los componentes respectivos o del sistema completo.

La simulación del sistema interactivo mediante un ser humano (denominado "mago"), es decir, la simulación Mago de Oz (WoZ, *Wizard of-Oz*), es una técnica ampliamente aceptada en la fase de desarrollo del sistema. Al mismo tiempo, puede utilizarse como una herramienta para la evaluación del sistema en el bucle, o del sistema biónico (mitad sistema, mitad mago). La idea es simular el sistema utilizando lenguaje oral como entrada, procesándola de una manera *racional* y generando respuestas de lenguaje oral hacia el usuario. A fin de ofrecer una situación de servicio telefónico realista, la entrada y salida vocal debe proporcionarse a los usuarios a través de una conexión telefónica simulada o real, utilizando una interfaz de usuario normalizada. En [16], [6], [3] y [9] pueden encontrarse descripciones detalladas de la organización de los experimentos WoZ.

Las simulaciones WoZ pueden utilizarse ventajosamente en los casos en los que las capacidades de las personas son superiores a las de los ordenadores, como es normalmente el caso de la comprensión o de la salida vocal. Como el sistema puede evaluarse antes de haberse dispuesto completamente, podrá simularse la calidad de funcionamiento de algunos componentes del sistema a un grado más allá de la tecnología moderna actual. Por consiguiente, resulta posible la extrapolación a las tecnologías que estarán disponibles en el futuro [23]. La simulación WoZ permite probar la factibilidad, cobertura e idoneidad antes de la implementación, de un modo relativamente económico. Con el método WoZ resultará más fácil simular modelos de interacción con alto grado de novedad y complejidad que con la implementación de un método del tipo implementar-probar-revisar. No obstante, es probable que este último gane popularidad conforme surjan herramientas de soporte lógico y prototípicas normalizadas, y en configuraciones industriales donde las plataformas están disponibles con amplitud. No obstante, el método WoZ merece la pena si la aplicación tiene un riesgo considerable, y los costos para reconstruir el sistema son lo suficientemente elevados [6].

La interacción entre el usuario físico y el sistema o el mago está influenciada considerablemente por los cinco tipos de factores descritos en la cláusula 5. Desde la perspectiva de los experimentadores, estos factores constituyen variables de la configuración experimental. Las variables se encuentran bajo control del experimentador (variables de control), el experimentador puede tener acceso a ellas para su medición (variables de respuesta) o pueden ser factores de confusión cuando el experimentador no tiene interés o control sobre ellas. Los factores de confusión pueden ser tratados mediante cuidadosos procedimientos de diseño experimental, concretamente a través de un diseño completo o parcial relativo al agente.

Una característica principal de un método de simulación WoZ es que los agentes de prueba no se dan cuenta de que el sistema con el que están interactuando es simulado. Las evidencias proporcionadas en [16] y [9] demuestran que se puede alcanzar este objetivo en prácticamente el 100% de los casos si la simulación se diseña cuidadosamente. El aspecto más importante para crear la ilusión del agente es la capacidad de entrada y salida vocal del sistema. Varios autores enfatizan que la ilusión de un diálogo con un ordenador debería soportarse mediante distorsión vocal, por ejemplo [17] y [2]. No obstante, otros parámetros del sistema podrán provocar el mismo efecto, por ejemplo, la capacidad de franqueza del sistema.

Las simulaciones WoZ deben ofrecer una simulación realista de la funcionalidad del sistema. Por consiguiente, se necesita una descripción exacta de la funcionalidad y el comportamiento del sistema antes de que se pueda organizar la simulación WoZ. Es importante que el mago se apegue en la medida posible a esta descripción, e ignore cualquier conocimiento y aptitudes superiores que pueda tener al hacer comparaciones con el sistema que se encuentra en prueba. Esto requiere una cantidad significativa de capacitación y soporte para el mago. Ya que un humano tiende a utilizar intuitivamente sus habilidades de mayor nivel, el trabajo del mago debe automatizarse en la medida posible. Para este fin se han desarrollado una diversidad de herramientas que por lo general consisten en una representación del modelo de interacción, por ejemplo, en términos de una gráfica visual o de una herramienta de software para realizar prototipos rápidamente, filtros para el canal de entrada y salida del sistema (por ejemplo, reproducción estructurada de audio, modificación de la voz y simuladores de reconocimiento), y otras herramientas de soporte como el registro de las interacciones (audio, texto, vídeo) y soporte de dominio (por ejemplo, horarios). En [23], [15], [9], [6] y [33] se describen algunos ejemplos típicos.

6.2 Casos de prueba

Debido a la falta de motivación real, durante las pruebas de laboratorio a menudo se utilizan tareas experimentales que deben efectuar los agentes. Este tipo de tareas proporciona un objetivo explícito, aunque no se debe confundir el mismo con un objetivo que un usuario desearía alcanzar en una situación de la vida real. Dada esta discrepancia, no podrán obtenerse fácilmente opiniones válidas del usuario con relación a la utilidad y aceptabilidad del sistema con una configuración de prueba de laboratorio.

En este último caso, la tarea experimental se define mediante una descripción de una situación hipotética relativa a una tarea particular en la cual el agente se desempeña interactuando con el sistema, por ejemplo, para recopilar información relativa a una conexión de un determinado tren, o para buscar un restaurante específico [6]. En el apéndice I se presentan algunos ejemplos de hipótesis de servicio de información de restaurante. La utilización de una hipótesis predeterminada facilita el máximo control sobre la tarea que deben realizar los agentes de prueba, y al mismo tiempo abarca una amplia gama de posibles situaciones (y de posibles problemas) durante la interacción. Las hipótesis pueden diseñarse específicamente para probar funcionalidades particulares de sistema (denominados casos de desarrollo), o para cubrir una amplia gama de posibles situaciones de interacción que son deseables para efectos de la evaluación. Por consiguiente, las hipótesis de desarrollo por lo general, difieren de los casos de evaluación.

Las hipótesis pueden ayudar a encontrar distintos puntos débiles en un diálogo, y por consiguiente permiten aumentar la capacidad de uso y la aceptabilidad del sistema final. Éstas definen objetivos de usuario en términos de la tarea y del subdominio considerados en un diálogo, y son un prerrequisito para determinar si el usuario alcanzó su objetivo. Sin una hipótesis predeterminada resultaría extremadamente difícil comparar los resultados obtenidos de distintos diálogos, ya que las peticiones de usuario pueden diferir y caer fuera del dominio del sistema. Si la influencia de la tarea es un factor que tiene que investigarse durante el experimento, el experimentador deberá garantizar que todos los usuarios ejecutan las mismas tareas. Éstos sólo puede alcanzarse a través de hipótesis predeterminadas.

Desafortunadamente, las hipótesis predeterminadas pueden tener algunos efectos negativos del comportamiento del usuario. Aunque no presentan un objetivo de la vida real a los agentes de prueba, sí preparan a los usuarios para interactuar con el sistema. Las hipótesis descritas pueden instar a los agentes de prueba a imitar el lenguaje propuesto, lo que conduce a leer a viva voz en lugar del habla espontánea. Se ha demostrado que la selección de casos también puede tener influencia en las estrategias de solución que son más eficaces para resolver la tarea [43]. Por lo general, los agentes de prueba que realizan hipótesis predeterminadas no se preocupan particularmente por las respuestas del sistema, ya que en realidad no necesitan la información. Por lo tanto, es posible que el resultado de la tarea no muestre un efecto importante con relación a las opiniones del grado de utilización de los agentes de prueba. Además, se sabe que los agentes de prueba no siempre leen las instrucciones cuidadosamente, y que pueden ignorar o malinterpretar las restricciones clave de las hipótesis.

El efecto de preparación relativo al lenguaje del usuario puede reducirse con la ayuda de descripciones gráficas de la hipótesis, véanse los ejemplos en el apéndice I. Una comparación entre hipótesis escritas y gráficas [6], [13] demuestra que el efecto de preparación masiva de las hipótesis escritas puede evitarse casi completamente mediante una representación gráfica, pero que la diversidad de cuestiones lingüísticas (número total de palabras, número de palabras fuera de vocabulario) es similar en ambos casos. Por consiguiente, aún tiene que garantizarse la diversidad del lenguaje recopilando unidades de habla de un número suficientemente grande de distintos usuarios, por ejemplo, en una situación de prueba de campo. Otra posibilidad es presentar descripciones de voz grabadas de las tareas de los agentes de prueba y solicitarles que tomen notas [42]. De esta manera, se espera que los procesos involucrados de comprensión y memoria aportarán a los agentes una codificación del significado de la descripción de la tarea, pero no con una representación de la forma superficial. Sin embargo, aún no se dispone de una prueba empírica de esta suposición.

6.3 Agentes de prueba

La regla general para los experimentos de evaluación es que la elección de los agentes debe guiarse por la finalidad de la prueba. Por ejemplo, la valoración analítica de las características de un sistema particular sólo será posible con agentes de prueba capacitados que sean expertos en el sistema que se considera. No obstante, este grupo no podrá juzgar los aspectos generales de la calidad del sistema de una forma que no sea influenciada por su conocimiento del sistema. Los juicios de calidad general válidos, sólo pueden esperarse de los agentes de prueba que se ajusten tanto como sea posible al grupo de usuarios de servicios futuros. Para los experimentos de interacción subjetiva con servicios basados en un sistema conversacional también deberán observarse las recomendaciones generales relativas a la elegibilidad de los agentes de prueba que se proporcionan en la Rec. UIT-T P.800.

En la cláusula 5.3 se presenta una síntesis de los factores de usuario. Algunos de estos factores tienen la responsabilidad de las características acústicas y lingüísticas de la voz producida por el usuario, concretamente edad y género, estado físico, velocidad de habla, esfuerzo vocal, lenguaje nativo, dialecto o acento. Ya que estos factores pueden ser muy críticos para el rendimiento del reconocimiento y la comprensión vocal, podría ser que las opiniones de calidad obtenidas de un grupo de usuarios que difieran en las características acústicas y de lenguaje no reflejen la calidad que puede esperarse del grupo de usuarios objetivo. Sin embargo, los grupos de usuarios son variables y mal definidos. Un servicio abierto a todo el público tarde o temprano tendrá que enfrentarse a una amplia gama de distintos usuarios. Las pruebas con usuarios especificados fuera del grupo de usuarios objetivo proporcionará por consiguiente una medida de la robustez del sistema con relación a las características del usuario.

Un segundo grupo de factores de usuario se relaciona a la experiencia y la pericia con el sistema, la tarea y el dominio. Distintas investigaciones muestran que la experiencia del usuario afecta a una amplia gama de características de la voz y del diálogo. Por ejemplo, se sabe que los usuarios tienen tendencia a resolver más problemas por llamada cuando se familiarizan con el sistema, y que la interacción es más corta [10]. Otras investigaciones muestran que el número de palabras dentro del vocabulario aumenta cuando los usuarios se familiarizan con el sistema. Al mismo tiempo, aumenta la tasa de terminación de tareas [25]. La confianza en el sistema también puede conducir a un número reducido de entradas de usuario y de mensajes de apoyo, y a disminuir el tiempo de transacción [26], [28].

Da la impresión de que los usuarios desarrollan patrones de interacción particulares cuando se familiarizan con un sistema. Se ha dado por supuesto que ese tipo de patrón es un equilibrio óptimo percibido entre el esfuerzo que cada usuario individual tiene que hacer para la interacción, y la eficacia con la que se realice ésta [39]. Tal parece que casi todos los usuarios desarrollan patrones estables con el sistema, pero éstos no son idénticos para todos ellos. El patrón de interacción que desarrolla un usuario también puede reflejar su confianza en el agente máquina, en el sentido de que el usuario puede tener un "modelo cognoscitivo" del sistema que refleja lo que se puede considerar como la opinión del sistema actual [38]. Ese modelo se determina en parte mediante las palabras entregadas al sistema, y en parte mediante las que salen del sistema. Generalmente, el usuario supone que sus palabras son bien comprendidas por el sistema. En caso de comprensiones erróneas, el usuario se confunde, y es probable que se produzcan problemas con el flujo del diálogo. Otra fuente de divergencia entre el modelo cognoscitivo del usuario y la confianza en el sistema es que este último tiene acceso a fuentes de información secundaria como una base de datos de aplicación. El usuario podría sorprenderse si se enfrenta a información que él no proporcionó.

7 Cuestionarios

Para obtener información relativa a las características de calidad percibidas por el usuario, tienen que recopilarse opiniones subjetivas. Para la recopilación pueden aplicarse dos principios distintos: ya sea para identificar las características de calidad pertinentes de una manera más o menos no orientada, o bien para cuantificar aspectos predeterminados de calidad como respuestas a preguntas cerradas o tareas de graduación de opiniones. Ambos modos tienen ventajas e inconvenientes: las consultas abiertas ayudan a encontrar valores de calidad que de otra manera permanecerían sin detectarse y a identificar los aspectos de calidad que son más importantes desde el punto de vista del usuario. De esta manera, puede facilitarse la interpretación de opiniones cuantitativas cerradas. Las preguntas cerradas o las tareas de graduación facilitan la comparación entre agentes y experimentos, y dan los medios exactos para cuantificar las percepciones del usuario. Éstas pueden efectuarse con relativa facilidad, y a menudo los agentes sin capacitación prefieren este método de juicio.

Las tareas de graduación producirán resultados válidos y fiables cuando se satisfagan dos requisitos principales: los elementos que se han de juzgar deben seleccionarse adecuada y significativamente, y la medición de escala ha de seguir reglas bien establecidas. Los métodos de escala se describen con detalle en los libros relativos a la psicometría, por ejemplo en [21], [12] o [7]. *Para la evaluación de la calidad de transmisión*, el UIT-T recomienda los métodos de determinación de evaluación por categorías absolutas (ACR), la determinación de evaluación por categorías de degradación (DCR) y la determinación de evaluación por categorías de comparación (CCR), véase la Rec. UIT-T P.800. *Para la evaluación de la calidad de los servicios basados en un sistema conversacional*, por lo general se solicita a los agentes de prueba sus opiniones relativas a las escalas de determinación de evaluaciones continuas o a las distintas escalas de determinación de evaluaciones por categorías absolutas. Una escala ACR consiste en un número de categorías discretas una de las cuales será seleccionada por el agente de prueba. Las categorías podrán visualizarse y etiquetarse con atributos para cada categoría, o para las categorías de extremo únicamente (la más a la izquierda y la más a la derecha). En las siguientes cláusulas se presentan

ejemplos de escalas de determinación de evaluaciones continuas. Aunque la escala de "impresión general" es similar a la escala ACR correspondiente a la determinación de evaluación de calidad de transmisión (véase la Rec. UIT-T P.800), no existe una relación directa entre esas evaluaciones de evaluaciones, y por consiguiente tampoco existe una ley de transformación que vincule los resultados medios obtenidos de una de las escalas continuas a los resultados de MOS que se utilizan para describir las opiniones de calidad de la transmisión general.

La tarea de evaluación relativa a las escalas continua o por categorías a menudo se describe en términos de una declaración (por ejemplo, "El sistema resultó fácil de comprender"), y los agentes de prueba han de expresar su acuerdo con respecto a la declaración al introducir la marca respectiva o seleccionar la categoría correspondiente de la escala. Este método se basa en propuestas previas de Likert [30], y en la figura 3 se ilustra el ejemplo de una escala. Se asignan números a las categorías o a las posiciones de la escala, en función de si la declaración es positiva (desde 1 para "desacuerdo total" hasta 5 para "acuerdo total") o negativo (desde 5 para "desacuerdo total" hasta 1 para "acuerdo total"), y se suman las evaluaciones individuales de todos los agentes. Otra posibilidad es definir etiquetas muy fáciles de entender para cada categoría, como propone por ejemplo el UIT-T para los experimentos de calidad de la transmisión vocal en su Rec. UIT-T P.800.

	Total desacuerdo	Desa- cuerdo	Ni acuerdo ni desacuerdo	Acuerdo	Acuerdo total
El sistema resultó fácil de comprender	☐	☐	☐	☐	☐

P.851_F03

Figura 3/P.851 – Opinión relativa a una declaración en una forma propuesta por Likert [30]

Las escalas bien construidas no proporcionarán información válida cuando la característica de calidad sobre la que se debe opinar está mal definida, o cuando no se selecciona adecuadamente para el servicio bajo consideración. En las siguientes cláusulas, se presenta un ejemplo de aspectos de selección de calidad, cada uno de los cuales puede abordarse mediante una pregunta específica que será evaluada por los agentes de prueba. Además, se relacionan ejemplos de preguntas o declaraciones formuladas. La elección de las preguntas que han de incluirse para un servicio particular dependerá del tipo de servicio, las tareas que se van a realizar, el comportamiento de interacción específico del servicio, el grupo de agentes de prueba, así como de la finalidad específica para la que se realiza el experimento de evaluación. Generalmente, el número de puntos sobre los que se va a emitir un juicio en un solo cuestionario debe limitarse a 15 ó 20, de modo que los agentes de prueba puedan distinguir y considerar los puntos individuales.

En una configuración de laboratorio, el cuestionario se entregará a los agentes de prueba inmediatamente después de realizar una interacción (reflejando potencialmente su impresión después de ella), y/o después de un determinado número de interacciones (lo que permite la integración de experiencias anteriores). En una prueba de campo, no podrá controlarse de modo estricto la recopilación de los cuestionarios y, por lo general, la opinión se refiere a un número de interacciones realizadas en un periodo de tiempo definido con amplitud. Podría suceder que las experiencias negativas se destaquen más y tengan una mayor influencia en el proceso de integración de tiempo que las experiencias positivas [11].

7.1 Preguntas relativas a los antecedentes del usuario

Al inicio de la sesión de prueba se deberán contestar diversas preguntas a fin de describir al usuario y sus antecedentes, lo cual es muy importante para el experimento. Estas preguntas se refieren a las siguientes cuestiones:

- Información personal: edad, género, profesión, lugar de nacimiento, dirección actual, competencia de lenguaje.
- Información relativa a la tarea: frecuencia de la tarea, método usual para resolver la tarea (interfaces facultativas), motivación, otros aspectos importantes relacionados con la tarea y el dominio.
- Información relativa al sistema: experiencia con servicios basados en DTMF o en un sistema conversacional, experiencia con equipos de tecnología vocal (reconocimiento vocal, voz sintetizada y otros).

En la siguiente relación se dan ejemplos de preguntas que pueden plantearse a los agentes de prueba. Las preguntas son relativas a un servicio de información de restaurante pero podrán adaptarse fácilmente a otras tareas y servicios.

Preguntas relativas a los antecedentes del usuario

Datos personales

Sexo: ☐ Femenino ☐ Masculino

Edad: _____ años

Profesión/educación: _____

País/ciudad de nacimiento: _____

Dirección actual: _____

1 ¿Con qué frecuencia come usted fuera de su domicilio, en promedio?

_____ veces por semana _____ veces por mes _____ veces por año

2 ¿De qué manera busca usted un restaurante cuando se encuentra en alguna ciudad del extranjero (existe la posibilidad de múltiples opciones)?

- | | | | |
|---------------------------|--------------------------|--------------------------------------|--------------------------|
| 2.1 Revistas | ☐ | 2.6 Recomendaciones de amigos | ☐ |
| 2.2 Folletos comerciales | ☐ | 2.7 Consulta telefónica a un sistema | |
| 2.3 Guía de la ciudad | ☐ | automático basado en la voz | ☐ |
| 2.4 Directorio telefónico | ☐ | 2.8 Otros: _____ | ☐ |
| 2.5 Internet | ☐ | | |

3 ¿Qué es importante para usted cuando elige un restaurante (existe la posibilidad de múltiples opciones)?

- | | | | |
|---|--------------------------|-------------------------------|--------------------------|
| 3.1 Precio | ☐ | 3.6 Ambiente | ☐ |
| 3.2 Tipo de alimento | ☐ | 3.7 Horario de funcionamiento | ☐ |
| 3.3 Calidad de los alimentos | ☐ | 3.8 Rapidez de servicio | ☐ |
| 3.4 Variedad de los alimentos ofrecidos | ☐ | 3.9 Amabilidad de servicio | ☐ |
| 3.5 Ubicación | ☐ | 3.10 Otros: _____ | ☐ |

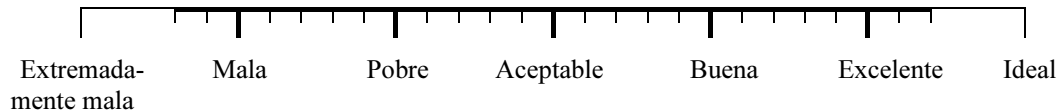
4 ¿Alguna vez ha utilizado un sistema de información automático basado en la voz?

☐ Sí

☐ No

4.1 ¿Si la respuesta es sí, cuándo?

4.1.1 ¿Cómo podría usted caracterizar su experiencia con este tipo de sistema?



5 ¿Tiene usted experiencia con un sistema de comprensión vocal?

☐ Sí

☐ No

5.1 ¿Si la respuesta es sí, con qué clase de sistema?

6 ¿Tiene usted experiencia con la voz sintetizada?

☐ Sí

☐ No

6.1. ¿Si la respuesta es sí, en qué ocasión?

7 ¿Qué información relativa a un restaurante desearía obtener de un sistema de información?

7.2 Preguntas relativas a la interacción individual

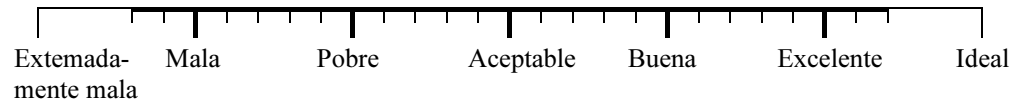
Después de cada interacción con el servicio (simulado), los agentes de prueba tienen que rellenar un cuestionario con varios puntos relativos a la experiencia de interacción particular. Estos puntos pueden abordar los siguientes aspectos:

- Información obtenida del sistema: disponibilidad, precisión, perfección, coherencia, fiabilidad, claridad y veracidad de la información obtenida, etc.
- Capacidad de entrada/salida vocal: comprensión percibida del sistema, frecuencia de errores de sistema, razonamiento percibido del sistema, esfuerzo de escucha necesario para comprender los mensajes del sistema, inteligibilidad percibida, comprensibilidad percibida, etc.
- Comportamiento de la interacción del sistema: transparencia de la interacción, congruencia con las expectativas del usuario, flexibilidad de la interacción, fiabilidad percibida del tratamiento del sistema, distribución de iniciativa, capacidad de control de la interacción, capacidades de confirmación y corrección, recuperación de problemas de interacción, naturalidad de la interacción, longitud del diálogo, velocidad percibida del sistema, homogeneidad del diálogo, etc.
- Personalidad percibida del sistema: cordialidad, cortesía, etc.
- Impresión en el usuario: naturalidad percibida del propio comportamiento del usuario, afabilidad, demanda cognoscitiva impuesta al usuario, estrés, nerviosismo, etc.
- Cumplimiento percibido de la tarea: éxito de la tarea, fiabilidad de los resultados de la tarea.

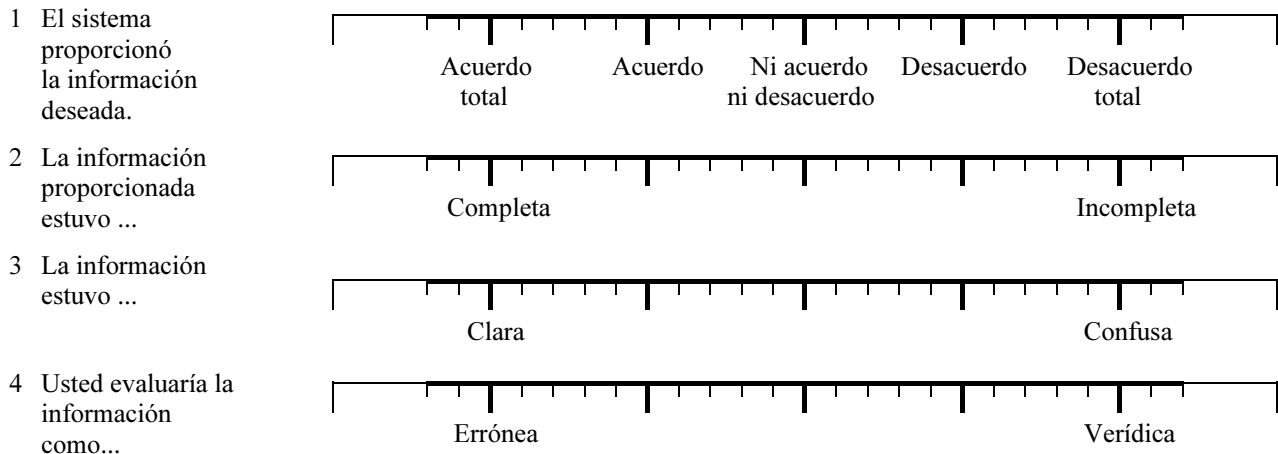
Más adelante se dan ejemplos de preguntas que abordan estos aspectos. El experimentador puede elegir las más apropiadas para el servicio que se desea examinar.

Preguntas relativas a la interacción individual

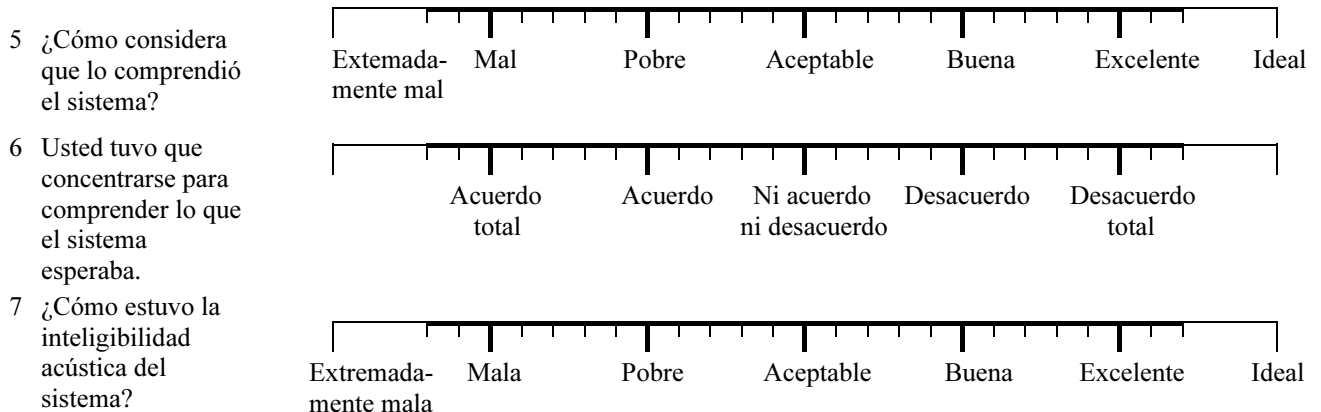
Impresión general:



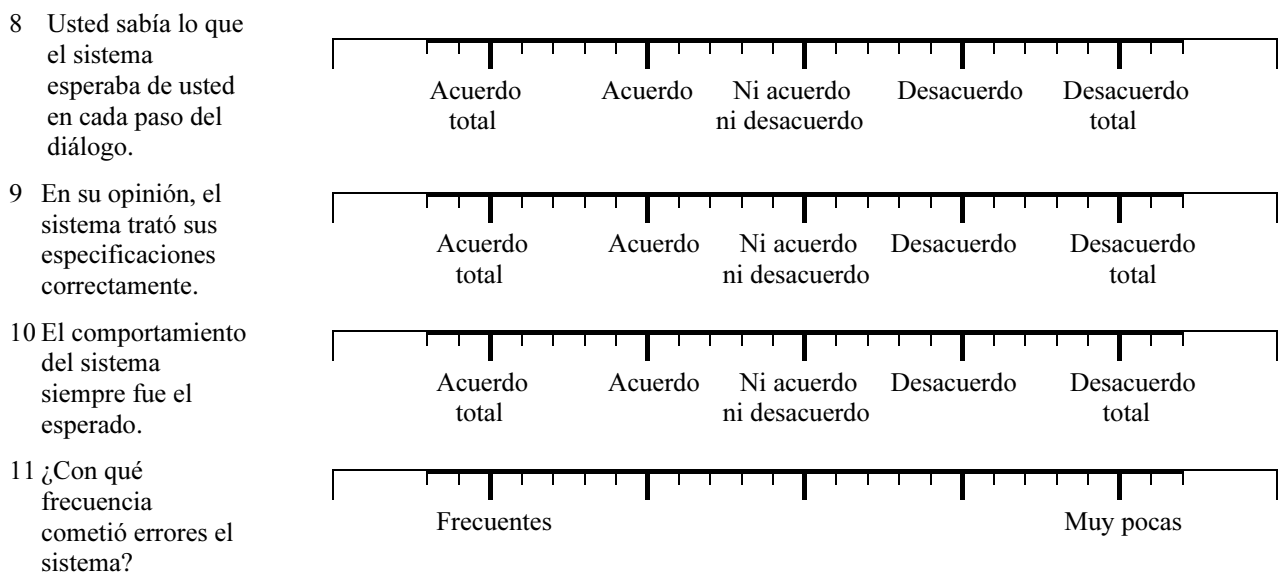
Información obtenida del sistema

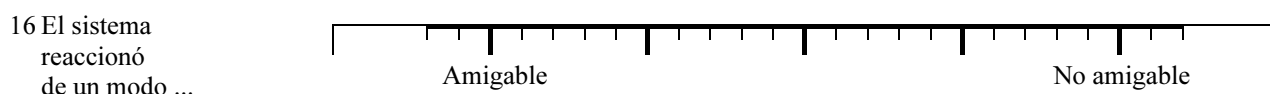
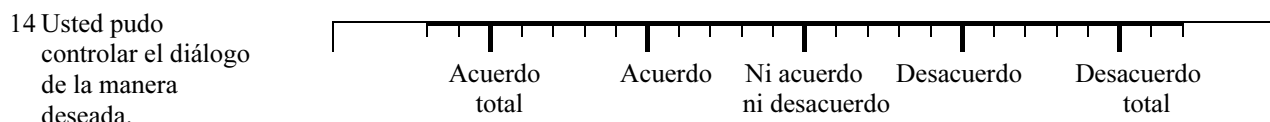
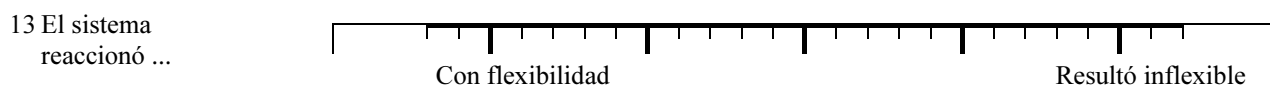
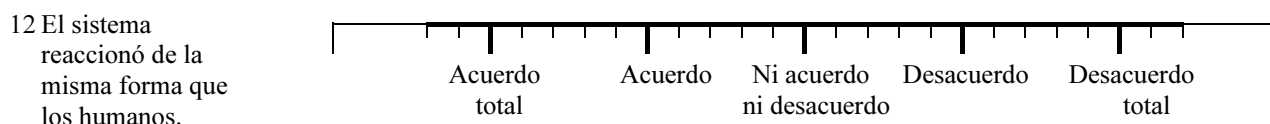


Comunicación con el sistema

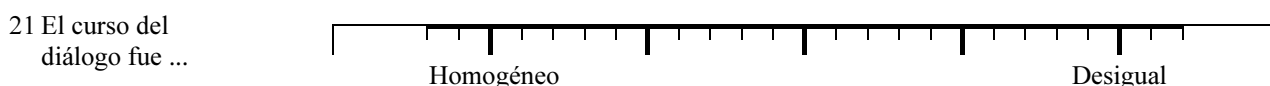
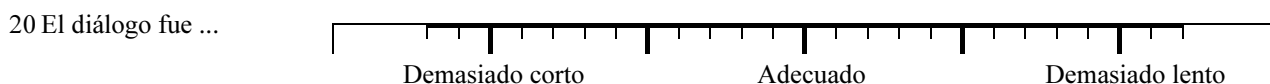
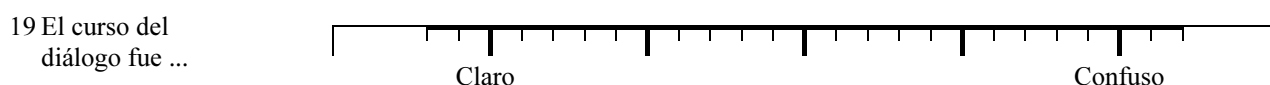
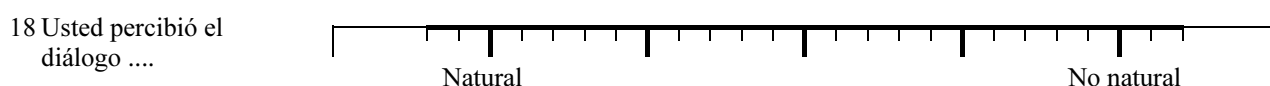
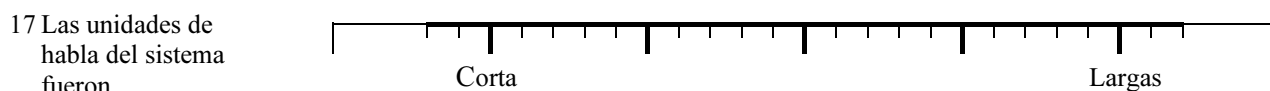


Comportamiento del sistema

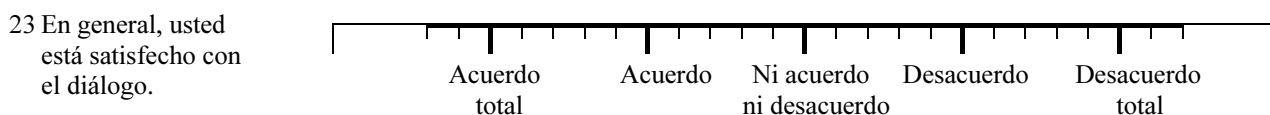
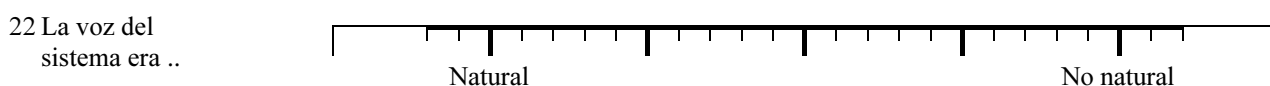




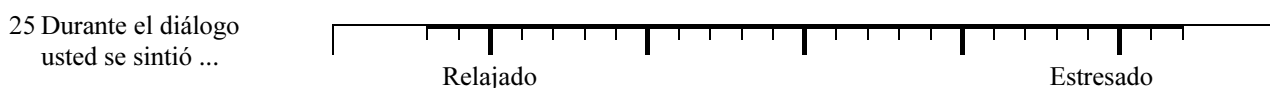
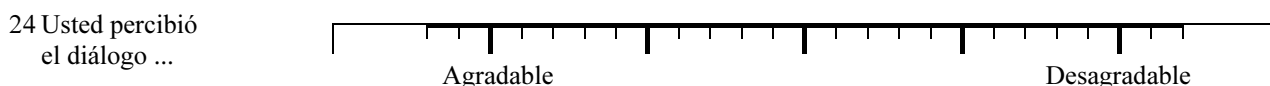
Diálogo



Su impresión del sistema



Impresión personal



7.3 Preguntas relativas a la impresión general del usuario con relación al sistema

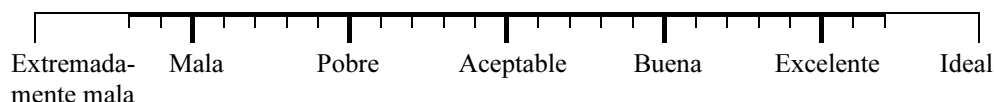
Una vez completadas todas las interacciones con el servicio, los agentes de prueba deben contestar un conjunto adicional de preguntas, esta vez con referencia a la experiencia global ganada hasta este momento con el sistema. Los siguientes puntos pueden incluirse en ese cuestionario:

- Impresión general del usuario con relación al sistema/servicio.
- Trato o expresión del sistema.
- Personalidad percibida del sistema: cordialidad, cortesía, etc.
- Capacidades de corrección, recuperación y ayuda del sistema.
- Control de interacción e iniciativa percibidos.
- Comodidad percibida cuando se utiliza el sistema.
- Percepción del cumplimiento de la tarea: éxito de la tarea, fiabilidad de los resultados de la tarea.
- Percepción de las posibilidades de uso: facilidad de uso, facilidad para aprender a utilizar el sistema, habitabilidad del sistema.
- Grado de placer del usuario, amabilidad del sistema.
- Aplicabilidad y utilidad del sistema para dar cumplimiento a la tarea.
- Valor añadido del sistema, en comparación al resto de las interfaces o al operador físico.
- Mejoras necesarias antes de poner en servicio el sistema.
- Utilización futura prevista para el servicio.

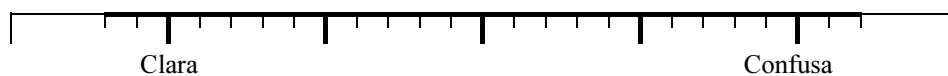
A continuación se da un ejemplo de un cuestionario correspondiente a un servicio de información. El mismo puede adaptarse y ampliarse conforme al servicio bajo consideración, la tarea que cumple, así como los objetivos del experimento.

Preguntas relacionadas con la impresión general del usuario sobre el sistema

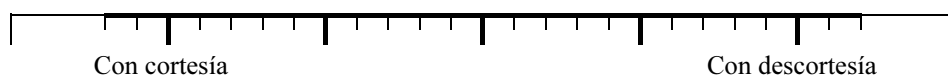
1 Impresión general.



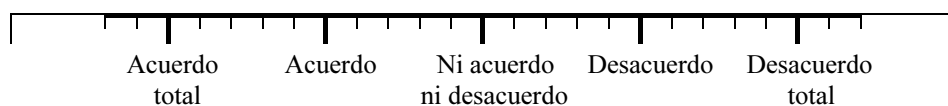
2 La forma de expresión del sistema fue ...



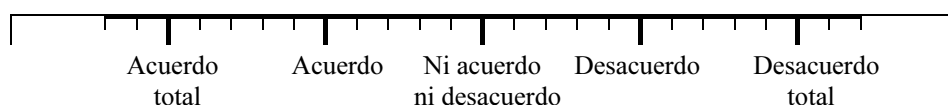
3 El sistema reaccionó ...



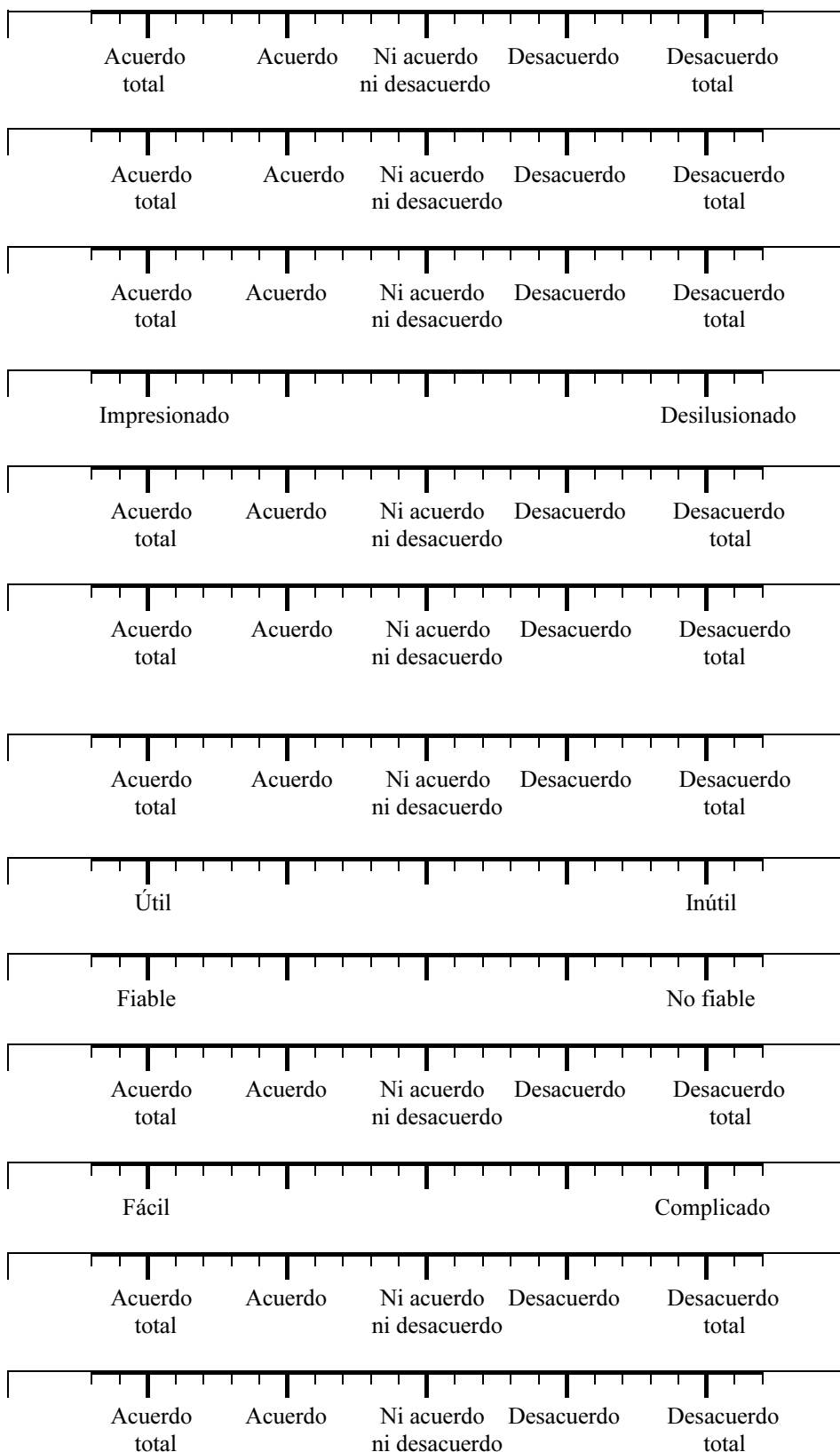
4 Usted habría esperado mayor ayuda del sistema.



5 El sistema pudo contestar todas sus preguntas.



- 6 Los malentendidos pudieron aclararse fácilmente.
- 7 El sistema pudo controlar el curso del diálogo.
- 8 Usted fue capaz de manejar el sistema sin ningún problema.
- 9 Con relación a los diálogos, usted está ...
- 10 Usted disfrutó los diálogos.
- 11 Usted se considera bien informado respecto a las posibilidades del sistema.
- 12 Las llamadas telefónicas al sistema merecieron la pena.
- 13 Usted percibió que este método para obtener información es ...
- 14 Usted evalúa el sistema como ...
- 15 Usted prefiere utilizar otra fuente de información.
- 16 El manejo del sistema fue ...
- 17 Usted prefiere un operador físico.
- 18 En el futuro, usted utilizará el sistema nuevamente.
- 19 ¿Qué características del sistema son las que más le gustaron?



8 Evaluación de la capacidad de uso

Además de abordar los aspectos individuales de las posibilidades de uso mediante los cuestionarios ya descritos, se dispone de métodos de evaluación de las posibilidades de uso dedicadas. Éstas pueden evaluarse bien sea con usuarios reales que realizan pruebas específicas, o mediante métodos de inspección de esas posibilidades con ayuda de expertos en evaluación. Ambos métodos se complementan entre ellos, en el sentido de que los métodos de inspección de las posibilidades de uso pueden detectar problemas relativos a esas posibilidades que fueron pasados por alto durante la prueba del usuario, y viceversa [36]. De hecho, se ha observado un alto grado de no coincidencia entre los dos. Por consiguiente, la evaluación de las posibilidades de uso debe combinar pruebas empíricas e inspecciones de esas posibilidades.

Los métodos de inspección de las posibilidades de uso pretenden encontrar problemas en las mismas en un diseño de interfaz de usuario existente, evaluando potencialmente la severidad de los problemas, dando recomendaciones sobre cómo resolver los problemas, y por consiguiente mejorando las posibilidades de uso del sistema. Esos métodos permiten aplicar fácilmente el conocimiento y la experiencia de los diseñadores de interfaces de usuario a la optimización de nuevos sistemas. Una parte importante de la inspección de las posibilidades de uso consiste en contar y clasificar los problemas de las mismas que se observan durante la interacción persona-máquina. La inspección de las posibilidades de uso debería ser eficaz no solamente para detectar problemas, sino también para ponderarlos conforme a su gravedad (no tiene sentido resolver problemas que no son importantes), y especialmente para sugerir modificaciones y mejoras de su diseño. Ya que muchos métodos de inspección se apoyan en la especificación del diseño en lugar de su implementación, éstos podrán aplicarse en una etapa relativamente prematura del proceso de diseño del sistema.

Pueden distinguirse los siguientes ocho tipos de métodos de inspección de las posibilidades de uso [36]:

- *Evaluación heurística*: este método informal incluye especialistas de las posibilidades de uso quienes emiten un juicio cuando un elemento del diálogo se conforma a los principios de las posibilidades de uso establecidos, denominados heurísticos.
- *Verificaciones del cumplimiento de las directrices*: inspecciones en las que se verifica la conformidad del servicio basado en un sistema conversacional con ayuda de una relación completa de directrices de posibilidades de uso. Ya que el número total de directrices puede ser elevado, este método requiere un alto grado de pericia.
- *Reuniones pluridisciplinarias*: reuniones en las que los usuarios, desarrolladores y expertos en factores humanos participan conjuntamente en una hipótesis, y debaten cuestiones de las posibilidades de uso asociadas con los elementos de diálogo involucrados en cada paso de la hipótesis.
- *Inspecciones de congruencia*: varios diseñadores examinan una interfaz que representa múltiples aspectos de diseño, y a continuación evalúan si es congruente con todas las cuestiones de diseño.
- *Inspecciones de normas*: un experto examina la conformidad de una interfaz específica a una norma definida.

- *Ensayos cognoscitivos*: simulan un proceso de resolución de problemas de usuario en cada fase de la interacción, y verifican si puede suponerse que los objetivos del usuario y la memoria de las acciones conducirán a la próxima acción correcta. Por lo general se presentan en forma de preguntas relativas a la relación entre los objetivos de tarea atribuidos al usuario, y las acciones necesarias del sistema para lograrlos.
- *Examen de las posibilidades de uso formal*: método formal integrado por un equipo de examen de las posibilidades de uso. Cada miembro del equipo tiene una tarea particular en el proceso de examen, por ejemplo, como moderador, propietario del diseño o examinador. Se organizan reuniones para preparar y realizar el examen, y para analizar sus resultados.
- *Examen de características*: se centra en las funciones operacionales de la interfaz de usuario, y si las funciones previstas cumplen los requisitos de los usuarios extremo que se pretenden.

La mayor parte de estos métodos se analizan en detalle en la literatura sobre posibilidades de uso [36]. La elección del método adecuado depende de los objetivos de la evaluación, la disponibilidad de directrices, la experiencia del evaluador y las limitaciones de tiempo y dinero.

La segunda alternativa es la evaluación de las posibilidades de uso con experimentos de usuario controlado. Esas pruebas pueden realizarse de un modo "objetivo, no intrusivo" o bien "subjetivo, intrusivo" [19]. Los métodos no intrusivos pretenden registrar el comportamiento del usuario físico de un modo natural y tranquilo, por ejemplo, mediante observaciones con equipo audiovisual o a través de registros con un equipo de grabación. Los métodos intrusivos necesitan una participación activa de los usuarios, por ejemplo, dando respuesta a cuestionarios o entrevistas (véase la última cláusula), mediante debates de grupo, o de un modo autodescriptivo, es decir, introduciendo un protocolo verbal que refleje los pensamientos u opiniones del usuario durante o después de la interacción. Estos métodos se describen con mayor detalle en los libros de evaluación de las posibilidades de uso [14].

9 **Análisis e interpretación de la información recopilada**

Las opiniones que se obtienen con relación a las escalas de determinación de evaluaciones cerradas pueden analizarse mediante gráficos de barras o distribuciones acumulativas. Aunque las distribuciones no son necesariamente gaussianas, es práctica común calcular valores medios aritméticos (y no medianas) de todas las evaluaciones obtenidas con una configuración de sistema especificada, véase la Rec. UIT-T P.800 y el Manual sobre telefonometría del UIT-T. En el caso de los valores medios, se evalúan límites de confianza y se realizan pruebas de relevancia mediante análisis convencional de la varianza (ANOVA, *analysis of variance*). Las suposiciones que soportan un análisis de varianza (distribución gaussiana y homogeneidad de varianzas) no siempre se satisfacen; no obstante, este método parece ser lo suficientemente robusto para proporcionar resultados razonables aun en el caso de alejarse de las condiciones estadísticas ideales. En caso de que una de las variables estadísticas tenga un efecto significativo estadísticamente (configuración de sistema y/o voz, agente de prueba, hipótesis, orden de las condiciones en el experimento, sesión de prueba, etc.), podrá utilizarse una prueba post-hoc para realizar comparaciones por pares entre las medias, y para determinar las fuentes de las diferencias. Para este fin se recomienda la prueba de diferencia enteramente significativa (HSD, *honestly significant difference*) de Tukey [40]. Cuando no se satisfacen las hipótesis subyacentes de las estadísticas paramétricas, resulta útil sumar los resultados adicionalmente en términos de una media o modo, y emplear pruebas no paramétricas como la de Kruskal y Wallis para efectos de comparación.

Cuando los experimentos de interacción se realizan en condiciones controladas (laboratorio), es posible recopilar parámetros de interacción y evaluaciones subjetivas de usuario de la misma interacción. En este caso, podrá cuantificarse la relación entre ambos tipos de criterios de medición. En una segunda fase, es posible tratar de predecir las opiniones del usuario con relación a aspectos de calidad particulares o sobre la calidad o satisfacción general basándose en parámetros de

interacción que pueden medirse. No obstante, no existe aún un método de modelización de calidad universal (independiente de la tarea) que permita cubrir en sus predicciones la mayoría de las varianzas de las opiniones del usuario. Walker et al., propusieron en 1997 un marco general para la predicción de la calidad de sistemas conversacionales (*PARAdigm for Dialogue System Evaluation*, PARADISE), basándose en un análisis de regresión lineal con múltiples variables estadísticas [44]. En principio, el marco es independiente de la tarea, aunque deben determinarse los parámetros de la función de predicción de calidad de nuevo para cada sistema particular, basándose en experimentos de interacción subjetiva. Además, su capacidad predictiva es aún muy limitada (por lo general en la gama de 40 a 50% de la varianza cubierta). Por consiguiente, los parámetros de interacción (tanto los que se pueden medir con instrumentos como los basados en expertos) y las opiniones de calidad de los usuarios siguen siendo las principales fuentes de información que describen la calidad de una interacción con un sistema conversacional.

Apéndice I

Ejemplos de hipótesis

Hipótesis N.º 1

A usted le gustaría saber dónde puede comer pato. Por favor consulte el sistema.

Nombre(s) del restaurante: _____

Hipótesis N.º 2

Usted desea ir a cenar comida griega la noche del martes en XXX.

Precio: $-|x| \text{---} | \text{---} | \text{---} | \text{---} | +$

Nombre(s) del restaurante: _____

Si el sistema no puede indicar un restaurante, por favor modifique la siguiente especificación:
Desea usted cenar en YYY.

Nombre(s) del restaurante: _____

Hipótesis N.º 3

Usted desea comer en un restaurante chino en el centro de la ciudad.

Precio: $-| \text{---} | \text{---} | \text{---} |x| +$

Nombre(s) del restaurante: _____

Hipótesis N.º 4

Usted desea salir a comer el XXX. Como su restaurante favorito está cerrado por vacaciones, solicite al sistema que le recomiende un restaurante.

Por favor escriba primero las especificaciones que desea darle al sistema.

Si el sistema no puede encontrar un restaurante que concuerde con sus especificaciones, por favor busque una opción hasta que el sistema indique al menos un restaurante.

Nombre(s) del restaurante: _____

Hipótesis N.º 5

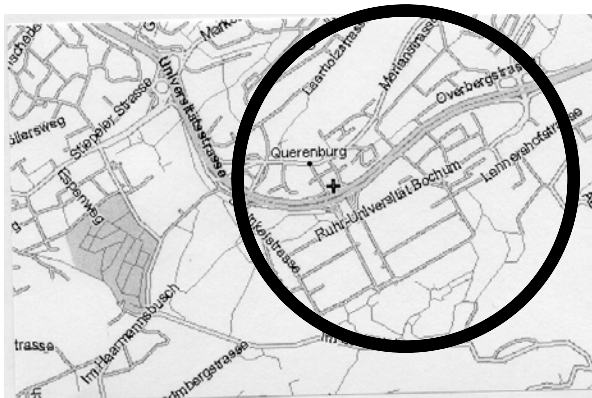
Sírvase recoger la información de los siguientes consejos:

Precio: $- \overline{\text{---} \mid \text{---} \textbf{x} \text{---} \mid \text{---} \mid} +$

Tipo de comida:



Ubicación:



Nombre(s) del restaurante:

BIBLIOGRAFÍA

- [1] ALLEN (J.), FERGUSON (G.), STENT (A.): An Architecture for More Realistic Conversational Systems, *Proc. of Intelligent User Interfaces 2001 (IUI-01)*, 1-8 Santa Fe NM (2001).
- [2] AMALBERTI (R.), CARBONELL (N.), FALZON (P.): User Representations of Computer Systems in Human-Computer Speech Interaction, *Int. Journal on Man-Machine Studies*, 38, 547-566 (1993).
- [3] ANDERNACH (T.), DEVILLE (G.), MORTIER (L.): The Design of a Real World Wizard of Oz Experiment for a Speech Driven Telephone Directory Information Service, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1165-1168, Berlin (1993).
- [4] ANTONIOL (G.), FIUTEM (R.), LAZZARI (G.), DE MORI, (R.): System Architectures and Applications, *Spoken Dialogues with Computers*, R. de Mori, ed., 583-609, Academic Press, London (1998).
- [5] BERNSEN (N.O.), DYBKJÆR (L.): A Methodology for Evaluating Spoken Dialogue Systems and Their Components, *Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC 2000)*, 2, 183-188, Athens (2000).
- [6] BERNSEN (N.O.), DYBKJÆR (H.), DYBKJÆR (L.): Designing Interactive Speech Systems: From First Ideas to User Testing, *Springer*, Berlin (1998).
- [7] BORG (I.), STAUFENBIEL (T.): Theorien und Methoden der Skalierung: Eine Einführung, *Verlag Hans Huber*, Bern (1993).
- [8] CHURCHER (G.E.), ATWELL (E.S.), SOUTER (C.): Dialogue Management Systems: A Survey and Overview, *Report 97.06, School of Computer Studies, University of Leeds*, Leeds (1997).
- [9] DAHLBÄCK (N.), JÖNSSON (A.), AHRENBERG (L.): Wizard of Oz Studies – Why and How? *Knowledge-Based Systems*, 6(4), 258-266 (1993).
- [10] DELOGU (C.), DI CARLO (A.), SEMENTINA (C.), STECCONI (S.): A Methodology for Evaluating Human-Machine Spoken Language Interaction, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1427-1430, Berlin (1993).
- [11] DUNCANSON (J.P.): The Average Telephone Call Is Better Than the Average Telephone Call, *The Public Opinion Quarterly*, 33(1), 112-116 (1969).
- [12] DUNN-RANKIN (P.): Scaling Methods, *Lawrence Erlbaum Assoc.*, Hillsdale NJ (1983).
- [13] DYBKJÆR (L.), BERNSEN (N.O.), DYBKJÆR (H.): Scenario Design for Spoken Language Dialogue Systems Development, *Proc. ESCA Workshop on Spoken Dialogue Systems*, P. Dalsgaard, L.B. Larsen, L. Boves and I. Thomsen, eds., 93-96, Vigsø (1995).
- [14] ETSI Technical Report ETR 095: Human Factors (HF); Guide for Usability Evaluations of Telecommunication Systems and Services, *European Telecommunications Standards Institute*, Sophia Antipolis (1993).
- [15] FOSTER (J.C.), DUTTON (R.), JACK (M.A.), LOVE (S.), NAIRN (I.A.), VERGEYNST (N.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Interactive Speech Technology: Human Factor Issues in the Application of Speech Input/Output to Computers*, C. Baber and J.M. Noyes, eds., 167-175, Taylor and Francis, London (1993).

- [16] FRASER (N.M.), GILBERT (G.N.): Simulating Speech Systems, *Computer Speech and Language*, 5, 81-99 (1991).
- [17] FRASER (N.M.), GILBERT (G.N.): Effects of System Voice Quality on User Utterances in Speech Dialogue Systems, *Proc. 2nd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'91)*, 1, 57-60, Genova (1991).
- [18] GIBBON (D.), MOORE (R.), WINSKY (R.), eds.: Handbook on Standards and Resources for Spoken Language Systems, *Mouton de Gruyter*, Berlin (1997).
- [19] GLEISS (N.): Usability – Concepts and Evaluation, *TELE (English Edition)*, 2/92, 24-30, Swedish Telecommunications Administration, Stockholm (1992).
- [20] GRICE (H.P.): Logic and Conversation, *Syntax and Semantics, Vol. 3: Speech Acts*, P. Cole and J.L. Morgan, eds., 41-58, Academic Press, New York NY (1975).
- [21] GUILFORD (J.P.): Psychometric Methods, *McGraw-Hill Book Company*, New York NY (1954).
- [22] HONE (K.S.), GRAHAM (R.): Towards a Tool for Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering*, 6(3-4), 287-303 (2000).
- [23] JACK (M.A.), FOSTER (J.C.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'91)*, 1, 715-718, Banff (1992).
- [24] JEKOSCH (U.): *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, Habilitation thesis (unpublished), University/GH Essen, Essen (2000).
- [25] KAMM (C.), NARAYANAN (S.), DUTTON (D.), RITENOUR (R.): Evaluating Spoken Dialogue Systems for Telecommunication Services, *Proc. 5th Europ. Conf. on Speech Communication and Technology (EUROSPEECH'97)*, 4, 2203-2206, Rhodes (1997).
- [26] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Speech Communication*, 38, 131-139 (2002).
- [27] LAMEL (L.), MINKER (W.), PAROUBEK (P.): Towards Best Practice in the Development and Evaluation of Speech Recognition Components for a Spoken Language Dialogue System, *Natural Language Engineering*, 6(3-4), 305-322 (2000).
- [28] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 7, 2875-2878, Sydney (1998).
- [29] LIFE (M.A.), LEE (B.P.), LONG (J.B.): Assessing the Usability of Future Speech Technology: Towards a Method, *Proc. of SPEECH'88*, 7th FASE Symposium, 4, 1297-1304, Edinburgh (1988).
- [30] LIKERT (R.): A Technique for the Measurement of Attitudes, *Archives of Psychology*, 140, 1-55 (1932).
- [31] MAIER (E.), MAST (A.), LUPERFOY (S.): Overview. Dialogue Processing in Spoken Language Systems, *Proc. of the ECAI'96 Workshop*, Budapest, E. Maier, M. Mast and S. LuperFoy, eds., Lecture Notes in Artificial Intelligence No. 1236, 1-13, Springer, Berlin (1997).
- [32] McTEAR (M.F.): Spoken Dialogue Technology: Enabling the Conversational Interface, *ACM Computing Surveys*, 34(1), 90-169 (2002).
- [33] MÖLLER (S.): Quality of Telephone-Based Spoken Dialogue Systems, *Habilitation thesis, Institute of Communication Acoustics, Ruhr-University, Bochum (to appear)* (2003).

- [34] MÖLLER (S.): A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems, *Proc. 3rd SIGdial Workshop on Discourse and Dialogue*, 142-153, Philadelphia PA (2002).
- [35] MÖLLER (S.): Assessment and Prediction of Speech Quality in Telecommunications, *Kluwer Academic Publ.*, Boston MA (2000).
- [36] NIELSEN (J.), MACK (R.L.), eds.: Usability Inspection Methods, *John Wiley & Sons*, New York NY (1994).
- [37] SENEFF (S.): Galaxy-II: A Reference Architecture for Conversational System Development, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 3, 931-934, Sydney (1998).
- [38] SOUVIGNIER (B.), KELLNER (A.), RUEBER (B.), SCHRAMM (H.), SEIDE (F.): The Thoughtful Elephant: Strategies for Spoken Dialog Systems, *IEEE Trans. Speech and Audio Processing*, 8(1), 51-62 (2000).
- [39] STURM (J.), BAKX (I.), CRANEN (B.), TERKEN (J.), WANG (F.): The Effect of Prolonged Use of Multimodal Interaction, *Proc. ISCA Workshop on Multi-Modal Dialogue in Mobile Environments*, L. Dybkjær, E. André, W. Minker and P. Heisterkamp, eds., 1-15, Kloster Irsee (2002).
- [40] TUKEY (J.W.): Exploratory Data Analysis, *Addison-Wesley*, Reading MA (1997).
- [41] VAN LEEUWEN (D.), STEENEKEN (H.): Assessment of Recognition Systems, *Handbook on Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore and R. Winsky, eds., 381-407, Mouton de Gruyter, Berlin (1997).
- [42] WALKER (M.A.), RUDNICKY (A.), PRASAD (R.), ABERDEEN (J.), BRATT (E.O.), GAROFOLO (J.), HASTIE (H.), LE (A.), PELLOM (B.), POTAMIANOS (A.), PASSONNEAU (R.), ROUKOS (S.), SANDERS (G.), SENEFF (S.), STALLARD (D.): DARPA Communicator: Cross System Results for the 2001 Evaluation, *Proc. 7th Int. Conf. on Spoken Language Processing (ICSLP 2002)*, 1, 269-272, Denver CO (2002).
- [43] WALKER (M.A.), FROMER (J.), DI FABBRIZIO (G.), MESTEL (C.), HINDLE (D.): What Can I Say? Evaluating a Spoken Language Interface to Email, *Human Factors in Computing Systems. CHI'98 Conf. Proc.*, Los Angeles CA, 582-589, Assoc. for Computing Machinery (ACM), New York NY (1998).
- [44] WALKER (M.A.), LITMAN (D.J.), KAMM (C.A.), ABELLA (A.): PARADISE: A Framework for Evaluating Spoken Dialogue Agents, *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, 271-280 (1997).
- [45] ZUE (V.), SENEFF (S.), GLASS (J.R.), POLIFRONI (J.), PAO (C.), HAZEN (T.J.), HETHERINGTON (L.): JUPITER: A Telephone-Based Conversational Interface to Weather Information, *IEEE Trans. Speech and Audio Processing*, 8(1), 85-96 (2000).

SERIES DE RECOMENDACIONES DEL UIT-T

Serie A	Organización del trabajo del UIT-T
Serie B	Medios de expresión: definiciones, símbolos, clasificación
Serie C	Estadísticas generales de telecomunicaciones
Serie D	Principios generales de tarificación
Serie E	Explotación general de la red, servicio telefónico, explotación del servicio y factores humanos
Serie F	Servicios de telecomunicación no telefónicos
Serie G	Sistemas y medios de transmisión, sistemas y redes digitales
Serie H	Sistemas audiovisuales y multimedios
Serie I	Red digital de servicios integrados
Serie J	Redes de cable y transmisión de programas radiofónicos y televisivos, y de otras señales multimedios
Serie K	Protección contra las interferencias
Serie L	Construcción, instalación y protección de los cables y otros elementos de planta exterior
Serie M	RGT y mantenimiento de redes: sistemas de transmisión, circuitos telefónicos, telegrafía, facsímil y circuitos arrendados internacionales
Serie N	Mantenimiento: circuitos internacionales para transmisiones radiofónicas y de televisión
Serie O	Especificaciones de los aparatos de medida
Serie P	Calidad de transmisión telefónica, instalaciones telefónicas y redes locales
Serie Q	Conmutación y señalización
Serie R	Transmisión telegráfica
Serie S	Equipos terminales para servicios de telegrafía
Serie T	Terminales para servicios de telemática
Serie U	Conmutación telegráfica
Serie V	Comunicación de datos por la red telefónica
Serie X	Redes de datos y comunicación entre sistemas abiertos
Serie Y	Infraestructura mundial de la información, aspectos del protocolo Internet y Redes de la próxima generación
Serie Z	Lenguajes y aspectos generales de soporte lógico para sistemas de telecomunicación