UNION INTERNATIONALE DES TÉLÉCOMMUNICATIONS

UIT-T P.851

(11/2003)

SECTEUR DE LA NORMALISATION
DES TÉLÉCOMMUNICATIONS
DE L'UIT

SÉRIE P: QUALITÉ DE TRANSMISSION TÉLÉPHONIQUE, INSTALLATIONS TÉLÉPHONIQUES ET RÉSEAUX LOCAUX

Méthodes d'évaluation objective et subjective de la qualité

Evaluation subjective de la qualité des services téléphoniques basés sur des dialogueurs automatiques

Recommandation UIT-T P.851

RECOMMANDATIONS UIT-T DE LA SÉRIE P

QUALITÉ DE TRANSMISSION TÉLÉPHONIQUE, INSTALLATIONS TÉLÉPHONIQUES ET RÉSEAUX LOCAUX

Vocabulaire et effets des paramètres de transmission sur l'opinion des usagers	Série	P.10
Lignes et postes d'abonnés	Série	P.30
		P.300
Normes de transmission	Série	P.40
Appareils de mesures objectives	Série	P.50
		P.500
Mesures électroacoustiques objectives	Série	P.60
Mesures de la sonie vocale	Série	P.70
Méthodes d'évaluation objective et subjective de la qualité	**Série**	**P.80**
		P.800
Qualité audiovisuelle dans les services multimédias	Série	P.900

Pour plus de détails, voir la Liste des Recommandations de l'UIT-T.

Recommandation UIT-T P.851

Evaluation subjective de la qualité des services téléphoniques basés sur des dialogueurs automatiques

Résumé

La présente Recommandation décrit des méthodes et des procédures permettant d'effectuer des expériences en vue d'une évaluation subjective des services téléphoniques basés sur des dialogueurs automatiques. Les systèmes en question permettent une interaction naturelle par le biais du dialogue parlé et possèdent des capacités de reconnaissance et d'interprétation de la parole, de gestion de dialogue et d'émission de parole. La configuration et la réalisation des essais d'interaction appropriés sont décrits, et des questionnaires quantifiant les aspects pertinents de la qualité sont indiqués.

Source

La Recommandation P.851 de l'UIT-T a été approuvée le 13 novembre 2003 par la Commission d'études 12 (2001-2004) de l'UIT-T selon la procédure définie dans la Recommandation UIT-T A.8.

Mots-clés

Compréhension de la parole, dialogueurs automatiques, évaluation subjective, génération de parole, gestion de dialogue, paramètre d'interaction, reconnaissance vocale.

AVANT-PROPOS

L'UIT (Union internationale des télécommunications) est une institution spécialisée des Nations Unies dans le domaine des télécommunications. L'UIT-T (Secteur de la normalisation des télécommunications) est un organe permanent de l'UIT. Il est chargé de l'étude des questions techniques, d'exploitation et de tarification, et émet à ce sujet des Recommandations en vue de la normalisation des télécommunications à l'échelle mondiale.

L'Assemblée mondiale de normalisation des télécommunications (AMNT), qui se réunit tous les quatre ans, détermine les thèmes d'étude à traiter par les Commissions d'études de l'UIT-T, lesquelles élaborent en retour des Recommandations sur ces thèmes.

L'approbation des Recommandations par les Membres de l'UIT-T s'effectue selon la procédure définie dans la Résolution 1 de l'AMNT.

Dans certains secteurs des technologies de l'information qui correspondent à la sphère de compétence de l'UIT-T, les normes nécessaires se préparent en collaboration avec l'ISO et la CEI.

NOTE

Dans la présente Recommandation, l'expression "Administration" est utilisée pour désigner de façon abrégée aussi bien une administration de télécommunications qu'une exploitation reconnue.

Le respect de cette Recommandation se fait à titre volontaire. Cependant, il se peut que la Recommandation contienne certaines dispositions obligatoires (pour assurer, par exemple, l'interopérabilité et l'applicabilité) et considère que la Recommandation est respectée lorsque toutes ces dispositions sont observées. Le futur d'obligation et les autres moyens d'expression de l'obligation comme le verbe "devoir" ainsi que leurs formes négatives servent à énoncer des prescriptions. L'utilisation de ces formes ne signifie pas qu'il est obligatoire de respecter la Recommandation.

DROITS DE PROPRIÉTÉ INTELLECTUELLE

L'UIT attire l'attention sur la possibilité que l'application ou la mise en œuvre de la présente Recommandation puisse donner lieu à l'utilisation d'un droit de propriété intellectuelle. L'UIT ne prend pas position en ce qui concerne l'existence, la validité ou l'applicabilité des droits de propriété intellectuelle, qu'ils soient revendiqués par un Membre de l'UIT ou par une tierce partie étrangère à la procédure d'élaboration des Recommandations.

A la date d'approbation de la présente Recommandation, l'UIT n'avait pas été avisée de l'existence d'une propriété intellectuelle protégée par des brevets à acquérir pour mettre en œuvre la présente Recommandation. Toutefois, comme il ne s'agit peut-être pas de renseignements les plus récents, il est vivement recommandé aux responsables de la mise en œuvre de consulter la base de données des brevets du TSB.

© UIT 2004

Tous droits réservés. Aucune partie de cette publication ne peut être reproduite, par quelque procédé que ce soit, sans l'accord écrit préalable de l'UIT.

TABLE DES MATIÈRES

Recommandation UIT-T P.851

Evaluation subjective de la qualité des services téléphoniques basés sur des dialogueurs automatiques

1 Domaine d'application

La présente Recommandation décrit des méthodes d'évaluation subjective donnant des informations sur la qualité des services téléphoniques basés sur des dialogueurs automatiques, telle qu'elle est expérimentée par les utilisateurs de ces services. Les dialogueurs automatiques dont il est question dans la Recommandation permettent une interaction en langage parlé effectuée à tour de rôle avec un utilisateur humain par le biais du réseau téléphonique et ils possèdent des capacités de reconnaissance vocale, de compréhension de la parole, de gestion de dialogue, de génération de réponse et d'émission de parole. Ils peuvent donner accès à des renseignements stockés dans une base de données ou permettre la réalisation de différents types de transactions.

Les méthodes d'évaluation décrites ici visent différents aspects de la qualité du point de vue de l'utilisateur, le dialogueur automatique étant considéré comme une boîte noire. Des aspects importants de la qualité sont l'utilisabilité du service, l'efficience de la communication, l'efficience de la tâche et du service, la satisfaction de l'utilisateur, la qualité perçue au niveau de l'émission et de la réception de la parole, la coopérativité du système, la symétrie de l'interaction et le caractère harmonieux de l'interaction, tel qu'il est perçu. Les méthodes sont fondées sur des expériences menées en laboratoire au cours desquelles des sujets interagissent avec le dialogueur automatique afin de réaliser une tâche réaliste prédéfinie. L'opinion des sujets concernant les aspects perceptifs de la qualité peut être évaluée librement ou selon des directives au moyen de questionnaires qui leur sont remis après l'expérience, ou à l'aide d'autres méthodes d'évaluation de l'utilisabilité. La présente Recommandation décrit la configuration et le déroulement des expériences d'interaction, les aspects pertinents de la qualité perçus par l'utilisateur, ainsi que les méthodes qui fourniront des informations sur ces aspects de la qualité. On trouvera dans les Recommandations UIT-T P.800 et P.85, ainsi que dans le Manuel de téléphonométrie, d'autres directives sur les méthodes d'évaluation subjective en général et sur l'évaluation des dispositifs d'émission de parole.

2 Références normatives

La présente Recommandation se réfère à certaines dispositions des Recommandations UIT-T et textes suivants qui, de ce fait, en sont partie intégrante. Les versions indiquées étaient en vigueur au moment de la publication de la présente Recommandation. Toute Recommandation ou tout texte étant sujet à révision, les utilisateurs de la présente Recommandation sont invités à se reporter, si possible, aux versions les plus récentes des références normatives suivantes. La liste des Recommandations de l'UIT-T en vigueur est régulièrement publiée. La référence à un document figurant dans la présente Recommandation ne donne pas à ce document, en tant que tel, le statut d'une Recommandation.

– Recommandation UIT-T E.800 (1994), *Termes et définitions relatifs à la qualité de service et à la qualité de fonctionnement du réseau, y compris la sûreté de fonctionnement*.

– Recommandation UIT-T G.107 (2003), *Le modèle E: modèle de calcul utilisé pour la planification de la transmission*.

– Recommandation UIT-T G.1000 (2001), *Qualité de service des communications: cadre et définitions*.

– Recommandation UIT-T P.85 (1994), *Méthode d'évaluation subjective de la qualité de parole des serveurs vocaux*.

– Recommandation UIT-T P.800 (1996), *Méthodes d'évaluation subjective de la qualité de transmission*.

– UIT-T, *Manuel de téléphonométrie* (1992).

3 Abréviations

La présente Recommandation utilise les abréviations suivantes:

ACR	évaluation par catégories absolues (*absolute category rating*)
ANOVA	analyse de variance (*analysis of variance*)
ASR	reconnaissance automatique de la parole (*automatic speech recognition*)
CCR	évaluation par catégories de comparaison (*comparison category rating*)
DARPA	Defense Advanced Research Projects Agency
DCR	évaluation par catégories de dégradation (*degradation category rating*)
DTMF	multifréquence bitonalité (*dual tone multiple frequency*)
HMM	modèle de Markov caché (*hidden Markov model*)
HSD	différence honnêtement significative (*honestly significant difference*)
MLP	perceptron multicouche (*multi-layer-perceptron*)
MOS	note moyenne d'appréciation (*mean opinion score*)
PARADISE	paradigme d'évaluation de dialogueur (*paradigm for dialogue system evaluation*)
QS	qualité de service
SDS	dialogueur automatique (*spoken dialogue system*)
WoZ	magicien d'Oz (*wizard-of-Oz*)

4 Introduction

Les dialogueurs automatiques (SDS), c'est-à-dire les systèmes informatiques avec lesquels des utilisateurs humains interagissent à tour de rôle au moyen du langage parlé peuvent faire partie des réseaux téléphoniques modernes. Ils donnent accès à des bases de données et à des transactions par l'intermédiaire du téléphone, par exemple pour obtenir des renseignements sur les horaires des trains ou des avions, les cours de la Bourse, ou des renseignements touristiques, ou encore pour effectuer des opérations bancaires, des réservations d'hôtels, etc. A l'opposé de simples systèmes DTMF, les dialogueurs automatiques possèdent des fonctions de reconnaissance automatique et de compréhension de la parole (c'est-à-dire des fonctions syntaxiques/sémantiques/pragmatiques et donc d'interprétation), ainsi qu'un module de gestion de dialogue qui assure un déroulement harmonieux et naturel de l'interaction parlée entre l'utilisateur et le système. En conséquence, l'interaction prend un caractère plus humain et le service fourni par ces systèmes peut attirer une gamme plus large d'utilisateurs potentiels. Très souvent, les systèmes de type DTMF et de type à dialogue parlé sont intégrés et une partie des aspects de la qualité seront identiques pour les deux types de systèmes. Il arrive parfois que les systèmes fondés sur le dialogue parlé ont recours à des structures et à des protocoles d'interface utilisés dans des environnements d'applications Web et sont constitués d'une façon similaire aux interfaces Web; ainsi, ces interfaces peuvent servir de référence pour obtenir les mêmes fonctions.

4.1 Tâches et modules d'un dialogueur automatique

D'un point de vue technique, la meilleure façon de représenter les modules d'un dialogueur automatique mis en œuvre sur le réseau téléphonique est de le faire selon une structure séquentielle.

On trouvera un exemple de cette structure à la Figure 1. Elle comprend six principaux modules auxquels l'utilisateur accède via une interface de serveur téléphonique. Le signal vocal émis par l'utilisateur est d'abord traité par le dispositif de reconnaissance vocale. Pendant le processus de reconnaissance, le signal est transformé en une chaîne de mots ou en un graphe d'hypothèse qui est ensuite soumis à une analyse sémantique. Le résultat est une trame sémantique représentant ce qui a été "compris" à partir des paroles prononcées par l'utilisateur. Le gestionnaire de dialogue a pour tâche d'interpréter la trame sémantique dans le contexte du dialogue et de la tâche, et de garder une trace de l'historique du dialogue. Lorsque toutes les informations pertinentes ont été recueillies auprès de l'utilisateur, une interrogation peut être lancée auprès de l'application sous-jacente (dans cet exemple une base de données). Les renseignements émanant du programme d'application et des autres objectifs communicatifs du gestionnaire de dialogue doivent être transformés en une réponse destinée à l'utilisateur. C'est la tâche du module de génération de réponse. Celui-ci génère une réponse textuelle, qui est ensuite transformée par le synthétiseur vocal en un signal vocal qui est transmis à l'utilisateur humain. Parfois, la génération de réponse et la synthèse vocale sont mises en œuvre dans un seul module (sans passer à la représentation textuelle) et des messages préenregistrés sont utilisés au lieu de la parole synthétisée.

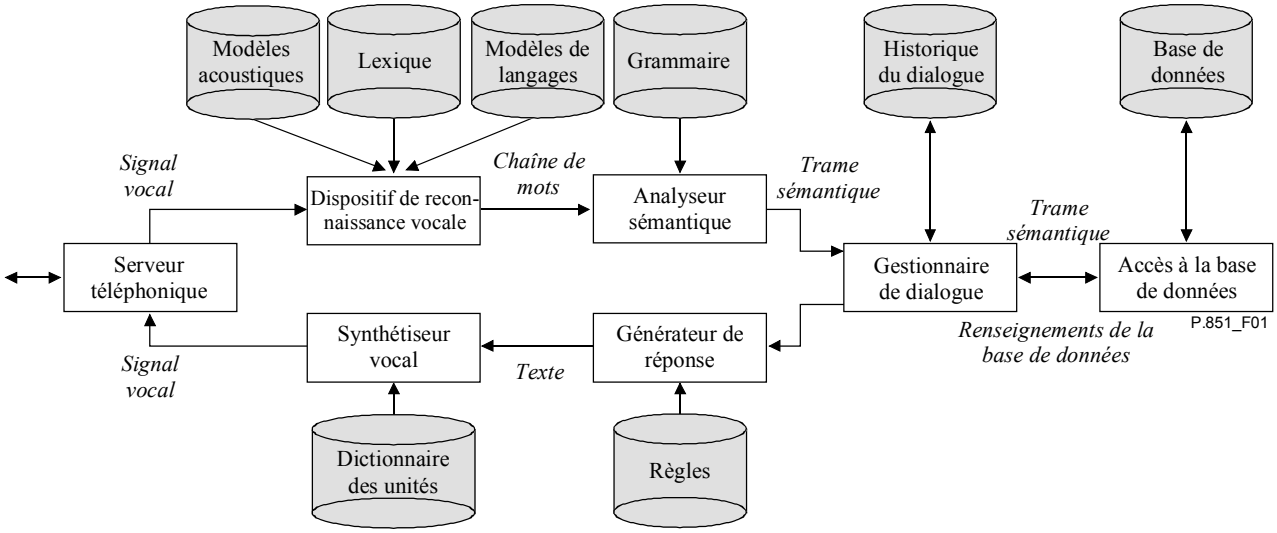


Figure 1/P.851 – Structure séquentielle d'un dialogueur automatique de type téléphonique [27], [33]

La structure de principe précitée peut être implémentée de plusieurs façons différentes. On trouvera des exemples en [4]. Une structure courante est "l'architecture centralisatrice" [37], [45] qui est utilisée dans le projet "Communicator" du DARPA. D'autres structures sont fondées sur des modules à fonctionnement asynchrone destinés à l'interprétation, au comportement (raisonnement et action) et à la génération. Voir [1].

4.2 Interaction téléphonique avec un dialogueur automatique

L'interaction avec le dialogueur automatique passe par un type quelconque de réseau de télécommunication. Ce dernier introduira un certain nombre de dégradations à la transmission qui influeront sur la qualité de la parole transmise et, de ce fait, également sur la qualité de fonctionnement du dispositif de reconnaissance vocale et des modules suivants du dialogueur automatique, qui recourent à la technique vocale et à la technique du langage naturel. Dans le sens de retour vers l'utilisateur humain, la voie de transmission dégradera le signal vocal généré par le système à dialogue. Les réseaux de télécommunication étant confrontés à des scénarios de communication entre humains et à des scénarios d'interaction entre la machine et l'homme, il est important d'examiner les besoins de l'utilisateur humain et du dispositif à technologie vocale. Manifestement, les besoins seront différents car les caractéristiques perceptives qui influent sur le

jugement en matière de qualité ne sont pas identiques aux caractéristiques d'un dispositif à technologie vocale, par exemple un dispositif de reconnaissance automatique de la voix (ASR, *automatic speech recognition*).

L'utilisateur humain effectue l'interaction par l'intermédiaire d'un type quelconque d'interface utilisateur, par exemple un combiné téléphonique, un poste à haut-parleur ou un casque. Les caractéristiques acoustiques des interfaces mentionnées sont d'une grande diversité, de même que leur sensibilité aux phénomènes acoustiques ambiants observés dans l'environnement de dialogue et d'écoute de l'utilisateur. Par exemple, le bruit ambiant peut avoir une incidence considérable sur l'intelligibilité des signaux vocaux transmis par un poste à haut-parleur et il influe également sur le comportement vocal de l'utilisateur. En conséquence, le scénario d'interaction entier, y compris le dialogueur automatique, la voie de transmission et l'interface utilisateur, doit être pris en compte dans la qualité globale de l'interaction.

4.3 Aspects de la qualité et facteurs déterminants

Des êtres humains sont les utilisateurs des services basés sur un dialogueur automatique qui sont offerts par le biais du téléphone. Aussi les facteurs humains doivent-ils être pris en compte lorsque l'on détermine les fonctions d'un système/service et son degré d'accomplissement. La qualité du service découle des perceptions de son utilisateur, relativement à ce qu'il attend ou veut du service. Selon une définition de la qualité élaborée en [24], *la qualité d'un service basé sur un dialogueur automatique* est le résultat de l'évaluation de la composition du service, telle qu'elle est perçue, par rapport à la composition souhaitée. Ainsi, la qualité perçue par l'utilisateur est un compromis entre ce qu'il attend ou veut et les caractéristiques qu'il perçoit pendant l'utilisation du service. Elle est fortement tributaire de la situation dans laquelle la perception et le jugement ont lieu. Ce fait doit être pris en compte lorsque l'on mène des essais d'évaluation subjective de la qualité, notamment en créant un cadre expérimental plus ou moins naturel et une motivation réaliste au niveau de l'utilisateur qui participe à l'essai.

A la différence de la notion de qualité de transmission vocale dans une communication d'homme à homme, l'utilisateur d'un service basé sur un dialogueur automatique participe activement à la production de la parole et au flux de dialogue. Ainsi, les caractéristiques et le comportement de l'utilisateur peuvent être décisifs pour l'accomplissement de la tâche souhaitée. L'utilisateur influera donc fortement sur les caractéristiques du système et du service. Pour simplifier la description du comportement du système et de l'utilisateur, on peut enregistrer les paramètres pendant l'interaction. Ces paramètres d'interaction peuvent être mesurables par des instruments, mais il n'est pas nécessaire qu'ils le soient. On peut citer comme exemples de paramètres le nombre d'énoncés ou la durée d'un dialogue, qui sont mesurables par des instruments, ou le taux d'erreurs dans les mots et la mesure du succès d'une tâche, qui ne peuvent être déterminés qu'avec le concours d'experts humains. On trouvera en [18] et [33] un aperçu des paramètres d'interaction.

En principe, la qualité d'un service basé sur un dialogueur automatique peut être traitée selon deux points de vue différents: celui du fournisseur de services et celui de l'utilisateur[1]. Le fournisseur de services s'intéresse surtout aux effets de chaque élément du service et à la manière dont il est lié au degré de satisfaction ou d'acceptabilité au niveau de l'utilisateur. Les fournisseurs de services appliquent la définition de la qualité de service (QS) donnée dans la Rec. UIT-T E.800. L'utilisateur perçoit les caractéristiques du service, y réfléchit, compare ses perceptions à un type quelconque de référence interne et les juge par rapport au point de savoir si les caractéristiques du service répondent à ses attentes ou à ses désirs. Lorsque l'on étudie la qualité d'un service, il est important de prendre en compte les deux points de vue. Les méthodes d'évaluation subjective telles que celles

[1] La Rec. UIT-T G.1000 définit même quatre points de vue différents: les besoins du client en matière de qualité de service, l'offre de qualité de service de la part du fournisseur de services, la qualité de service assurée ou fournie par le fournisseur et la qualité de service perçue par le client.

qui sont décrites dans la présente Recommandation seront axées sur le point de vue de l'utilisateur. Toutefois, elles seront aussi utiles pour le fournisseur de services car elles donnent des indications sur les caractéristiques du service qui ont besoin d'être améliorées.

Tant l'*efficacité* que l'*efficience* sont liées aux résultats obtenus dans la réalisation de l'objectif pour lequel le service a été conçu. L'efficacité est un indice absolu qui décrit dans quelle mesure l'objectif a été atteint, en termes d'exactitude et de complétude (voir par exemple [14]):

"*Efficacité*: exactitude et complétude des objectifs définis que les utilisateurs peuvent atteindre dans des environnements particuliers."

Les mesures visant l'efficacité dont il est fait mention dans différents ouvrages sont par exemple le succès de la tâche ou les mesures Kappa [33]. Par contre, l'efficience est une mesure relative de la réalisation d'un objectif par rapport aux ressources utilisées [14]:

"*Efficience*: ressources utilisées par rapport à l'exactitude et à la complétude des objectifs atteints."

Les mesures souvent utilisées sont celles par exemple de la durée du dialogue ou du nombre de séquences de paroles prononcées par le système et l'utilisateur.

L'efficience et la demande cognitive sont des critères qui caractérisent un système avec lequel un utilisateur est en mesure d'atteindre son ou ses objectifs pour une tâche donnée. Cela étant, l'*utilisabilité* est généralement définie d'une façon bien plus large et décrit la capacité d'un service à être compris, appris et utilisé par des utilisateurs déterminés dans des conditions bien définies. Elle indique l'aptitude du service à répondre aux besoins de l'utilisateur, comprend l'efficacité et l'efficience du système et conduit à la satisfaction de l'utilisateur [35]. La *satisfaction de l'utilisateur* est un indicateur de l'utilité et de l'utilisabilité du service, telles qu'elles sont perçues, et de l'utilisabilité pour le groupe d'utilisateurs visé. Il s'agit de savoir si l'utilisateur obtient les renseignements qu'il souhaite, est à l'aise avec le service et obtient les renseignements dans un délai acceptable [31].

Les notions de qualité décrites pour un service basé sur un dialogueur automatique peuvent être représentées sous forme de diagramme, ainsi qu'il est proposé en [34]; voir la Figure 2. Mis à part les facteurs utilisateur mentionnés, quatre types de facteurs contribuent à la qualité perçue par l'utilisateur: les facteurs agent (principalement liés au dialogue et au système lui-même), les facteurs tâche (liés à la façon dont le dialogueur automatique appréhende la tâche pour lequel il a été conçu), les facteurs environnementaux (par exemple les facteurs liés à l'environnement acoustique et à la voie de transmission) et les facteurs contextuels tels que les coûts, le type d'accès et la disponibilité. Les aspects de la qualité perçus par l'utilisateur sont décrits dans la partie inférieure du diagramme.

Les facteurs environnementaux, les facteurs agents et les facteurs tâche influent sur la *qualité de l'entrée et de l'émission de parole*, sur la *coopérativité* du comportement du système et sur la *symétrie* de l'interaction dans le dialogue. La qualité de l'entrée et de l'émission de parole comprend des aspects tels que l'intelligibilité, le naturel, l'effort d'écoute nécessaire pour comprendre les messages du système, ou encore la compréhension par le système, telle qu'elle est perçue. La coopérativité est définie ici au sens de la non-violation des principes du comportement coopératif en matière de dialogue, comme le définit Grice [20]. Elle comprend des aspects tels qu'informativité, vérité et faits probants, pertinence, manière, connaissance du domaine et traitement des métacommunications (c'est-à-dire confirmation, clarification, rectification et reprise après des erreurs de communication); voir [6]. L'aspect asymétrie des partenaires (différences dans le comportement d'interaction à attribuer à l'asymétrie des partenaires interactifs) est visé par une catégorie appelée symétrie du dialogue, qui comprend également les effets des capacités d'initiative dans le dialogue et de contrôle de l'interaction.

Les aspects de la qualité mentionnés conduisent à une communication (interaction) plus ou moins efficiente et à une réalisation efficiente de la tâche à effectuer. L'*efficience de la communication* est liée à la rapidité et au rythme de l'interaction, à la concision du dialogue et à son caractère

harmonieux. Par contre, l'*efficience de la tâche* est liée au succès et à la facilité de la tâche. Deux aspects additionnels de la qualité sont importants: la "personnalité" de l'agent machine (politesse, convivialité, naturel du comportement) et l'effort demandé à l'utilisateur humain pour l'interaction (facilité de communication, tension/énervement, etc.). Ces aspects ont été résumés dans le terme *confort*.

L'efficience de la communication, l'efficience de la tâche et le confort contribuent tous à l'utilisabilité du service, pour lequel la satisfaction de l'utilisateur peut être considérée comme un indicateur. En revanche, l'*efficience du service* subit les effets de l'efficience de la tâche et des facteurs contextuels. Elle est importante pour l'adéquation du service (pour l'accomplissement de la tâche souhaitée) et pour la valeur ajoutée attribuée au service (par exemple par rapport à des méthodes similaires permettant d'obtenir les mêmes renseignements, telles qu'une interface Web ou des informations défilantes). L'utilisabilité, l'efficience du service et *l'avantage économique* conduisent à l'*utilité* du service et, enfin, à son *acceptabilité*.

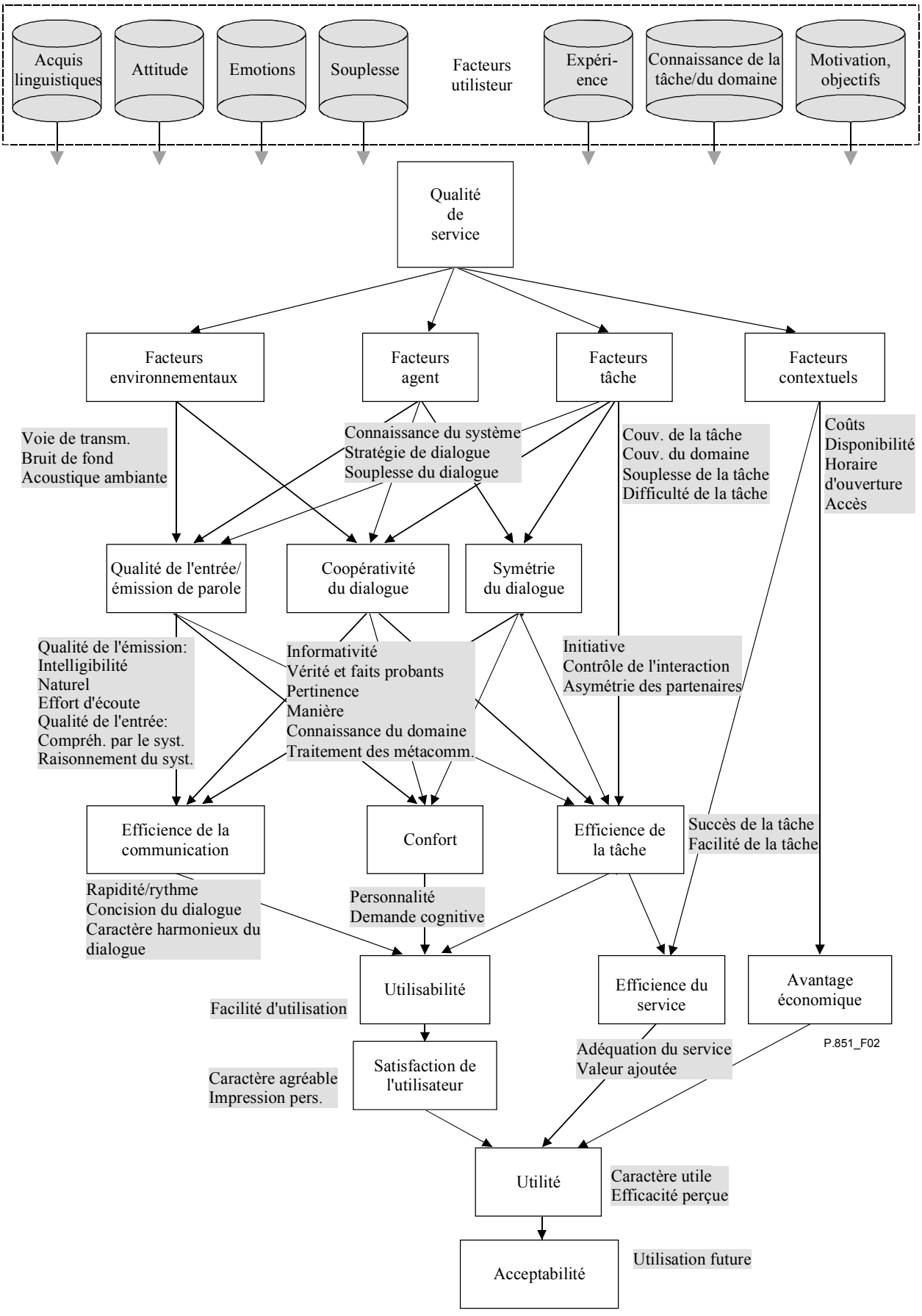


Figure 2/P.851 – Aspects de la qualité et facteurs déterminants; voir [34]

4.4 Méthodes d'évaluation subjective

Les dialogueurs automatiques peuvent être évalués au niveau des modules constitutifs (par exemple le dispositif de reconnaissance vocale, le module de compréhension de la parole ou le module d'émission de parole) ou au niveau de l'ensemble du système (intégré). L'évaluation analytique au niveau des modules est une source précieuse de renseignements pour décrire comment les différentes parties du système réalisent leur tâche. Cependant, elle peut parfois passer à côté des facteurs pertinents qui contribuent à la qualité globale du service perçue par l'utilisateur. Par exemple, des erreurs de reconnaissance ou de compréhension de la parole peuvent être neutralisées par le module de traitement du discours, sans affecter la qualité globale du système. Pour cette raison, afin de déterminer la qualité d'un service basé sur un dialogueur automatique, il est indispensable de mener des essais subjectifs avec des utilisateurs réels ou des utilisateurs agissant à titre expérimental en interaction avec le système entier.

Pour évaluer les différents aspects de la qualité d'un service basé sur un dialogueur automatique, il est nécessaire de mener des expériences subjectives avec des utilisateurs humains. Ces expériences ont deux objectifs principaux:

1) pendant l'interaction, les paramètres système mesurables par des instruments sont recueillis et les paroles prononcées par le système et l'utilisateur sont enregistrées. Les fichiers comptes rendus sont soumis à une évaluation d'expert et le résultat obtenu est un ensemble de paramètres décrivant les aspects spécifiques de l'interaction homme-machine concernant les énoncés, le dialogue et le niveau de tâche du point de vue du concepteur du système;

2) après l'interaction, un questionnaire est remis aux sujets ayant participé à l'expérience, le but étant de collecter des informations sur les caractéristiques perceptives de la qualité qui sont pertinentes pour exprimer l'impression de l'utilisateur humain concernant la qualité globale. De telles expériences peuvent être réalisées avec des systèmes pleinement opérationnels ou avec des systèmes encore au stade de la mise au point et dans lesquels des parties de modules doivent être simulées. On trouvera aux § 6 à 8 des précisions sur la configuration expérimentale, les questionnaires et les méthodes d'évaluation de l'utilisabilité.

En laboratoire, les deux types d'informations peuvent être obtenus en parallèle. Cela étant, dans une situation d'essai sur le terrain avec de vrais utilisateurs, les paramètres d'interaction enregistrés par des instruments sont souvent l'unique source d'information permettant au fournisseur de services de contrôler la qualité du système. Le volume de données qui peut être recueilli auprès d'un service opérationnel peut devenir considérable. En pareil cas, il importe de définir un ensemble principal de mesures qui décrivent la performance du système et de disposer d'outils qui automatisent une grande partie du processus d'analyse des données. La tâche de l'analyste humain est d'analyser et d'interpréter les données, et d'estimer l'effet des mesures de la performance obtenues sur la qualité qui serait perçue par un utilisateur (prototype). Un certain nombre de considérations concernant l'analyse et l'interprétation des résultats d'essais sont données au § 9. Il est possible d'établir une relation entre les paramètres d'interaction et les jugements subjectifs à condition que les deux types de renseignements soient disponibles. Ces modèles de prédiction de la qualité des dialogueurs automatiques de type téléphonique sont encore à l'étude et une brève analyse est donnée au § 9.

Comme il est d'usage pour les expériences d'évaluation subjective, le but visé et les circonstances d'une analyse ou d'une évaluation devraient être explicitement indiqués, et ils devraient être étayés par des documents. Dans le cadre du projet européen DISC, un modèle a été conçu à cette fin [5]. A partir de ce modèle et d'une classification des méthodes donnée en [33], on peut définir les critères suivants:

- motivation de l'analyse/évaluation (par exemple analyse détaillée des mécanismes de reprise du système ou de la satisfaction estimée des futurs utilisateurs);

- objet de l'analyse/évaluation (par exemple dispositif de reconnaissance vocale, gestionnaire de dialogue ou système dans son ensemble);
- environnement de l'analyse/évaluation (par exemple expérience contrôlée en laboratoire ou essai sur le terrain);
- type de méthode de mesure (par exemple mesure instrumentale des paramètres d'interaction ou jugements sur qualité obtenus auprès des utilisateurs, ceux-ci pouvant répondre librement ou devant choisir entre plusieurs réponses);
- symptômes à rechercher (par exemple questions de clarification des utilisateurs ou rejet de la reconnaissance automatique de la parole);
- phase du cycle de vie durant laquelle l'analyse/évaluation a lieu (par exemple simulation, version prototype ou système entièrement opérationnel);
- accessibilité du système et de ses modules (par exemple méthode de la boîte de verre ou de la boîte noire);
- références utilisées pour les mesures (par exemple mesures qualitatives de la performance absolue du système ou valeurs quantitatives par rapport à une référence ou à un repère mesurables);
- outils d'appui disponibles pour l'analyse/évaluation.

Ces critères constituent un ensemble de documents de base qui devrait être fourni avec les essais d'analyse/évaluation. Les documents peuvent être mis en application sous la forme d'une liste de points ou d'une description détaillée des expériences.

5 Caractérisation du dialogueur automatique

Selon le schéma des aspects de la qualité indiqué à la Figure 2, cinq types de facteurs influent sur l'interaction avec un service basé sur un dialogueur automatique: *facteurs agent*, *facteurs tâche*, *facteurs utilisateur*, *facteurs environnementaux* et *facteurs contextuels*. Ces facteurs détermineront la performance du système (modules) et la qualité perçue par l'utilisateur. Ils doivent donc être pris en compte lorsque l'on mène des expériences d'évaluation subjective et que l'on établit la documentation afférente.

5.1 Facteurs agent

En tant que partenaire interactif, le système doit être caractérisé d'une manière technique, c'est-à-dire en définissant les caractéristiques des différents modules du système et leur interconnexion dans la structure séquentielle de la Figure 1, ou en spécifiant les fonctions opérationnelles de l'agent. Les fonctions d'agent les plus importantes à définir sont la capacité de reconnaissance de la parole, la capacité de compréhension du langage naturel, la capacité de gestion de dialogue, la capacité de génération de réponse et la capacité d'émission de parole. Les modules de compréhension du langage naturel et de génération de réponse sont étroitement liés aux modules voisins, à savoir le gestionnaire de dialogue d'un côté et le dispositif de reconnaissance vocale ou le synthétiseur vocal de l'autre. Ainsi, les interfaces de ces modules doivent être décrites avec précision.

5.1.1 Caractérisation du dispositif de reconnaissance vocale

D'un point de vue fonctionnel, les dispositifs de reconnaissance vocale peuvent être classés selon les paramètres suivants [41]:

- richesse du vocabulaire, par exemple dispositifs de reconnaissance dotés d'un vocabulaire limité, moyen ou riche;
- complexité du vocabulaire, par exemple en ce qui concerne la possibilité de confusion au niveau des mots;

- type de discours, par exemple mots isolés, mots enchaînés, discours continu, discours spontané, y compris discontinuités telles que accès de toux, hésitations, interruptions, recommencements, etc;
- langue: dispositifs de reconnaissance vocale unilingue ou multilingue, résultats de la reconnaissance dépendant de la langue, portabilité de la langue;
- dépendance à l'égard du locuteur, par exemple dispositifs de reconnaissance vocale dépendant du locuteur, indépendant du locuteur ou pouvant s'adapter au locuteur;
- type de grammaire et complexité. La complexité de la grammaire peut être déterminée en termes de perplexité, attribut qui permet de mesurer avec quel niveau d'efficacité une séquence de mots peut être prédite par le modèle de langage;
- méthode de formation, par exemple formations multiples concernant des mots isolés explicitement prononcés ou formation intégrée concernant des chaînes de mots dont les points de départ et de fin ne sont pas définis.

Par ailleurs, les modules de reconnaissance vocale peuvent être décrits en termes de caractéristiques techniques générales qui peuvent être implémentées de façon différente dans divers systèmes [27]. Les caractéristiques techniques suivantes ont été partiellement utilisées dans le projet DISC:

- capture du signal: fréquence d'échantillonnage, largeur de bande du signal, quantification, fenêtrage;
- analyse des caractéristiques, par exemple des coefficients mel cepstraux, énergie et dérivées de premier ou de deuxième ordre;
- unités vocales fondamentales, par exemple modèles téléphoniques ou modèles de mots, modélisation du silence ou d'autres sons autres que la parole;
- lexique: nombre d'entrées pour chaque mot, avec une ou plusieurs prononciations; lexique créé à partir de dictionnaires ou de convertisseurs de graphèmes en phonèmes; entrées additionnelles pour les bruits et les mots de remplissage; vocabulaire couvert prévu par rapport au vocabulaire cible;
- modèle acoustique: type de modèle, par exemple réseaux à perceptron multicouche (MLP, *multi-layer-perceptron*) ou modèle de Markov caché (HMM, *hidden Markov model*); données et paramètres de formation; post-traitement du modèle;
- modèle de langage: type de modèle, par exemple modèle de langage statistique N-grammes à repli ou grammaire sans contexte; matériel de formation, notamment une importante base d'apprentissage à usage général ou des données recueillies dans une expérience limitée; modélisation de mots individuels ou de classes de mots pour des catégories spécifiques (par exemple dates ou noms); modèles indépendants ou dépendants de l'état de dialogue;
- type de décodeur, par exemple HMM;
- degré d'utilisation des informations prosodiques.

5.1.2 Caractérisation de la compréhension du langage

Les caractéristiques suivantes sont importantes pour la capacité de compréhension du langage du système:

- description sémantique de la tâche, par exemple par lots (paires de valeurs d'attributs);
- analyse syntaxique-sémantique: capacité d'analyse générale, par exemple analyse complète ou analyse partielle robuste; complexité de la syntaxe autorisée, par exemple nombre de possibilités à un niveau déterminé;
- analyse contextuelle: nombre et complexité des règles;

- interaction avec les modules de reconnaissance vocale et de gestion de dialogue: type et volume d'informations d'entrée et de sortie (hypothèses uniques, listes ordonnées, etc.), dépendance de l'interprétation syntaxique-sémantique et contextuelle à l'égard de l'état de dialogue.

5.1.3 Caractérisation du gestionnaire de dialogue

L'approche adoptée pour la gestion de dialogue peut être définie d'un point de vue technique, par exemple sous la forme d'une grammaire de dialogue, d'une méthode fondée sur un plan ou d'une collaboration [8], [32]. Les caractéristiques les plus importantes du gestionnaire de dialogue sont le type et la quantité de connaissances implémentées dans le gestionnaire, la répartition de l'initiative entre le système et l'utilisateur et les stratégies de métacommunication du système (confirmation, clarification, rectification et reprise):

- connaissances du gestionnaire de dialogue: modèle d'historique de dialogue (renseignements échangés jusque-là dans le dialogue), modèles de tâches et de domaines (scénarios, plans, objectifs et sous-objectifs, objets et leurs caractéristiques), modèle de connaissance mondial, modèle conversationnel et modèle d'utilisateur;
- initiative: initiative du système, initiative mixte ou initiative de l'utilisateur;
- stratégie de confirmation: confirmation explicite, confirmation implicite, confirmation "en écho", confirmation récapitulative;
- stratégies de rectification, de clarification et de reprise;
- adaptativité du gestionnaire de dialogue: gestionnaires constitutifs du système devant apprendre de nouvelles notions en fonctionnement normal ou gestionnaires adaptatifs pouvant inclure un modèle d'utilisateur dynamique et susceptibles de pouvoir apprendre les stratégies de communication de l'utilisateur.

Outre l'interaction avec l'utilisateur, il faut définir l'interaction avec le système d'application, notamment l'interface (langage de programmation) et les mécanismes éventuels de commande pour le traitement de la dynamique du système d'application.

5.1.4 Caractérisation de la génération de parole

La génération de parole comprend l'éventuelle génération d'une réponse textuelle et la traduction en langage parlé. La plupart des systèmes utilisent un des trois types de génération de parole suivants: parole préenregistrée, phrases modèles ou conversion texte-parole. Les caractéristiques suivantes doivent être définies:

- interaction avec le gestionnaire de dialogue: type et quantité d'informations fournies en entrée par le gestionnaire de dialogue, par exemple texte orthographique ou annoté, renseignements ciblés ou prosodiques, etc;
- génération de réponse: stratégie (par exemple grammaire formelle ou modèles simples), souplesse (vocabulaire prédéfini ou non limitatif), type et quantité d'informations à inclure dans chaque énoncé, forme du message (syntaxe, choix des mots);
- voix du système: nombre de voix, sexe, professionnalisme, formation, qualité prosodique, conditions d'enregistrement, adaptativité;
- langue: synthétiseurs unilingue ou multilingue, capacité d'identification de la langue, portabilité de la langue;
- type de génération de parole: messages préenregistrés, phrases modèles, conversion texte-parole, conversion concept-parole;

- caractéristiques de la conversion texte-parole: stratégie (par exemple fondée sur un modèle ou sur un ensemble), capacités de prétraitement du texte, paramètres de modèles, caractéristiques du corpus d'unités (types et longueurs d'unités, couverture du vocabulaire cible, etc.), algorithmes de concaténation et/ou de sélection, stratégies de prosodie (fréquence fondamentale, durée, intensité), etc;
- caractéristiques contextuelles: style d'élocution, vitesse d'élocution, adaptativité contextuelle.

5.2 Facteurs tâche

La tâche qui doit être effectuée par l'utilisateur est un facteur déterminant de l'interaction. Elle peut être caractérisée au niveau du type, du domaine, de la complexité, de la fréquence, des conséquences et de la portabilité:

- type de tâche: selon le document [6] une différenciation est possible entre:
  - les tâches bien structurées, ayant une structure stéréotypée qui indique quel élément d'information doit être échangé et souvent aussi dans quel ordre naturel;
  - les tâches mal structurées, contenant un grand nombre de sous-tâches optionnelles dont la nature et l'ordre sont difficiles à prévoir;

  et entre:
  - les tâches homogènes;
  - les tâches hétérogènes, c'est-à-dire *intrinsèquement* une combinaison de plusieurs tâches qui sont différentes en raison de leur nature même (par exemple ordre, plus informations, plus contrôle de dispositif);
- domaine de tâche: richesse, variabilité, nombre d'utilisateurs qui connaissent bien le domaine, utilité du domaine, généralisabilité, etc;
- complexité de la tâche: nombre de scénarios visés, nombre maximal de sous-objectifs, nombre de sous-tâches pouvant être accomplies en parallèle, nombre minimal d'échanges nécessaires pour résoudre le problème, complexité prévue pour la syntaxe/le vocabulaire, etc;
- fréquence de la tâche, c'est-à-dire fréquence prévue à laquelle les utilisateurs utiliseront le système pour une tâche déterminée. Les systèmes d'acheminement des appels ou donnant des renseignements sur les vols aériens (dits systèmes "walk-up-and-use", c'est-à-dire des systèmes faciles à utiliser sans apprentissage) seront utilisés à une fréquence relativement faible, de sorte que l'on ne peut pas s'attendre à ce que les utilisateurs potentiels connaissent le système, montrent les effets de l'apprentissage (c'est-à-dire qu'ils se souviennent du comportement adopté lors des appels précédents) ou acceptent un apprentissage;
- conséquences de la tâche, par exemple questions de sécurité;
- portabilité de la tâche.

5.3 Facteurs utilisateur

Dans la plupart des cas, la caractérisation de l'utilisateur se limite à une large catégorisation concernant la tâche, le domaine et le contexte mondial, car il est impossible de décrire exactement les facteurs importants pour un utilisateur en particulier (attitude, motivation, émotion, souplesse). Les caractéristiques ci-après sont souvent indiquées dans les protocoles d'évaluation:

- nombre d'utilisateurs;
- âge et sexe: ces facteurs sont censés influer sur la fréquence fondamentale et le spectre vocal, mais aussi sur l'interaction dans le dialogue;

- niveau d'expérience: utilisateur novice par rapport à un utilisateur expérimenté, utilisateur occasionnel par rapport à un utilisateur régulier, utilisateur formé par rapport à un utilisateur non formé;
- niveau de connaissances spécialisées dans le domaine d'application: utilisateurs professionnels par rapport à des utilisateurs privés;
- motivation explicite concernant l'utilisation du service;
- condition physique, effort vocal, vitesse d'élocution, etc;
- langage natif, accent, dialecte, etc.

Pour les applications spécialisées, il pourrait être nécessaire d'être plus explicite dans la spécification de l'expérience et du savoir-faire, par exemple en ce qui concerne la connaissance de l'objet de la tâche, la possibilité d'élaborer des stratégies pour optimiser l'exécution de la tâche et la capacité d'utiliser les dispositifs nécessaires pour exécuter la tâche [29].

5.4 Facteurs environnementaux

L'environnement contient l'ensemble du contexte physique de l'interaction. Une caractérisation complète sera généralement impossible et seuls les facteurs influant directement sur le signal vocal devraient être décrits, à savoir:

- type et propriétés acoustiques de l'interface utilisateur;
- voie de transmission téléphonique: la description peut être donnée à différents niveaux, par exemple en termes d'équipement de transmission, d'équipement de commutation et d'équipement terminal utilisés dans la connexion ou en termes de paramètres d'une configuration de référence pour la planification du réseau voir la Rec. UIT-T G.107;
- situation acoustique ambiante, y compris réverbération, coloration du son, niveaux de bruit ambiant et spectres, présence simultanée d'autres locuteurs, etc.

5.5 Facteurs contextuels

Il s'agit de facteurs non physiques qui caractérisent le contexte d'utilisation du service étudié. Les facteurs types sont les suivants:

- facilité d'accès: disponibilité des numéros de téléphone, liens avec d'autres services, etc;
- disponibilité du service: horaire d'ouverture, restrictions éventuelles de l'accès;
- coûts: coûts fixes et coûts facturés en fonction de la durée de l'interaction, conditions particulières des comptes, etc;
- services ayant une fonctionnalité similaire: ils doivent être comparés à tous les autres facteurs contextuels.

6 Configuration de l'expérience

Les expériences d'interaction subjective avec un dialogueur automatique devraient être configurées selon les règles générales appliquées aux essais d'opinion de conversation figurant dans la Rec. UIT-T P.800. On trouvera une description plus détaillée des questions d'ordre pratique dans le Manuel de téléphonométrie de l'UIT-T. Ce principe s'applique aux conditions physiques des cabines d'essai, aux caractéristiques du bruit ambiant, au modèle expérimental et aux règles générales relatives à l'analyse de données. Dans les paragraphes ci-après, seuls seront décrits les éléments qui sont propres à l'interaction avec le dialogueur automatique, à savoir la configuration du système, les scénarios d'essai et les sujets participant aux essais.

Les expériences subjectives peuvent être menées avec des systèmes pleinement opérationnels ou avec l'aide d'un expérimentateur humain qui simule les parties manquantes du système, ou avec l'ensemble du système (simulation de type "magicien d'Oz"). Pour obtenir des résultats valables et

fiables, le système (simulé), les utilisateurs participant aux essais et la tâche expérimentale doivent satisfaire à plusieurs conditions; voir les § 6.1 à 6.3. Les interactions sont généralement enregistrées et annotées par un expert humain, de manière à pouvoir calculer les paramètres d'interaction. Après chaque interaction et après la séance d'essai, des questionnaires doivent être remplis par les sujets participant à l'expérience. Ces questionnaires permettent de quantifier différents aspects de la qualité d'un service basé sur un dialogueur automatique. La conception de ces questionnaires est examinée au § 7. Au § 8, on trouvera un bref aperçu des méthodes d'évaluation de l'utilisabilité des services.

6.1 Configuration du système et simulation de type "magicien d'Oz"

Pour réaliser des expériences d'interaction avec des utilisateurs humains, il est nécessaire d'implémenter une configuration qui assure la pleine fonctionnalité du système. La nature exacte de la configuration dépendra de la disponibilité des modules du système et donc de la phase de mise au point du système. Si les modules du système n'ont pas encore été implémentés ou si une implémentation n'est pas réalisable (par exemple à cause de données insuffisantes) ou n'est pas économique, une simulation des modules respectifs ou du système dans son ensemble est nécessaire.

La simulation du système interactif par un être humain (le "magicien"), c'est-à-dire la simulation de type "magicien d'Oz" (WoZ, *wizard-of-Oz*) est une technique bien acceptée dans la phase de mise au point du système. En même temps, elle permet d'évaluer le système dans la boucle ou le système bionique (intervention pour moitié du système et pour moitié du "magicien"). L'idée est de simuler le système en utilisant le langage parlé en entrée, de traiter celui-ci selon certains *principes* et de générer des réponses en langage parlé destinées à l'utilisateur. Pour obtenir une situation téléphonique réaliste, l'entrée et l'émission de parole devraient être assurées aux utilisateurs par le biais d'une liaison téléphonique réelle ou simulée, au moyen d'une interface utilisateur normalisée. On trouvera en [16], [6], [3] et [9] une description détaillée de la configuration des expériences WoZ.

Il est intéressant d'utiliser des simulations WoZ lorsque les capacités humaines sont supérieures à celles des ordinateurs comme c'est généralement le cas pour la compréhension de la parole ou l'émission de parole. Le système pouvant être évalué avant d'être entièrement configuré, le fonctionnement de certains modules du système peut être simulé à un degré qui dépasse le point actuel des connaissances. Ainsi, une extrapolation vers les technologies qui seront disponibles à l'avenir devient possible [23]. La simulation WoZ permet de tester d'une manière relativement économique la faisabilité, la couverture et l'adéquation d'un système avant son implémentation. Pour des degrés élevés d'innovation et des modèles d'interaction complexes, il peut être plus facile de procéder à des simulations WoZ que d'adopter une approche "implémentation-essai-révision". Toutefois, cette dernière approche gagnera probablement du terrain avec l'apparition de logiciels normalisés et d'outils de prototypage, et dans des configurations industrielles où des plates-formes sont largement disponibles. La méthode WoZ est néanmoins utile si l'application présente des risques élevés et que le coût de reconstruction du système est suffisamment important [6].

L'interaction entre l'utilisateur humain et le système ou le "magicien" dépend considérablement des cinq types de facteurs décrits au § 5. Du point de vue des expérimentateurs, ces facteurs constituent les variables de la configuration expérimentale. Les variables sont contrôlées par l'expérimentateur (variables de contrôle), accessibles et mesurables par l'expérimentateur (variables de réponse), ou bien il s'agit de facteurs de confusion qui ne sont pas intéressants pour l'expérimentateur ou qui sont hors de son contrôle. Les facteurs de confusion peuvent être traités grâce à des procédures de conception expérimentales minutieuses, à savoir une conception à l'intérieur du sujet, complète ou partielle.

Une des principales caractéristiques de la simulation WoZ est le fait que les sujets testés ne savent pas que le système avec lequel ils sont en interaction est simulé. Les éléments de preuve indiqués en

[16] et [9] montrent que cet objectif est atteint dans près de 100% des cas si la simulation est conçue avec soin. Pour donner cette illusion au sujet, le plus important est la capacité d'entrée et d'émission de parole du système. Plusieurs auteurs mettent l'accent sur le fait que l'illusion d'un dialogue avec un ordinateur devrait être assurée par une distorsion vocale, par exemple [17] et [2]. Toutefois, d'autres paramètres du système sont susceptibles de causer le même effet, par exemple la directionnalité du système.

La méthode WoZ devrait assurer une simulation réaliste de la fonctionnalité du système. En conséquence, une description exacte de la fonctionnalité et du comportement du système est nécessaire avant de configurer une simulation WoZ. Il est important que le "magicien" respecte cette description et fasse abstraction de toutes les connaissances et compétences supérieures qu'il a par rapport au système à tester. Cela exige un niveau élevé de formation et d'aide pour le "magicien". Vu qu'un être humain utiliserait intuitivement ses compétences supérieures, le travail du "magicien" devrait être automatisé au maximum. Un certain nombre d'outils ont été mis au point à cette fin. Ils consistent généralement en une représentation du modèle d'interaction, par exemple graphique visuel ou outil logiciel de prototypage rapide, filtres pour la voie d'entrée et de sortie du système (notamment simulateurs structurés pour la reproduction acoustique, le déguisement de la voix et la reconnaissance vocale), et autres outils tels que les outils d'enregistrement de l'interaction (audio, texte, vidéo) et prise en charge des domaines (par exemple calendriers). Des exemples types sont décrits en [23], [15], [9], [6] et [33].

6.2 Scénarios d'essai

En raison de l'absence d'une motivation réelle, les essais en laboratoire utilisent souvent les tâches expérimentales que les sujets doivent effectuer. La tâche expérimentale fournit un objectif explicite, mais celui-ci ne devrait pas être confondu avec un objectif qu'un utilisateur souhaiterait atteindre dans une situation réelle. A cause de cette divergence, il n'est pas aisé d'obtenir dans le cadre d'un essai en laboratoire un jugement valable de l'utilisateur sur l'utilité et l'acceptabilité du système.

En laboratoire, la tâche expérimentale est définie dans la description d'un scénario. Un scénario décrit une tâche particulière que le sujet doit effectuer en interaction avec le système, par exemple pour obtenir des renseignements sur une correspondance ferroviaire précise ou pour chercher un restaurant en particulier [6]. Des exemples de ces scénarios pour un service d'information sur les restaurants sont donnés à l'Appendice I. L'utilisation d'un scénario prédéfini assure un contrôle maximal de la tâche réalisée par les sujets, tout en couvrant un large éventail de situations possibles (et de problèmes possibles) lors de l'interaction. Des scénarios peuvent être spécialement conçus pour tester des fonctions spécifiques du système (scénarios de mise au point) ou pour couvrir un large éventail de situations d'interaction potentielles qu'il est souhaitable d'évaluer. Ainsi, les scénarios de mise au point sont généralement différents des scénarios d'évaluation.

Les scénarios aident à trouver les diverses lacunes d'un dialogue et donc à accroître l'utilisabilité et l'acceptabilité du système final. Ils définissent les objectifs des utilisateurs en ce qui concerne la tâche et le sous-domaine traités dans un dialogue, et constituent une condition préalable à satisfaire pour déterminer si l'utilisateur a atteint son but. Sans un scénario prédéfini, il serait extrêmement difficile de comparer les résultats obtenus dans différents dialogues, car les demandes de l'utilisateur pourraient différer et sortir du cadre des connaissances de domaine du système. Si l'influence de la tâche est un facteur qui doit être étudié lors de l'expérience, l'expérimentateur doit veiller à ce que tous les utilisateurs exécutent les mêmes tâches. Cela n'est possible qu'avec des scénarios prédéfinis.

Malheureusement, les scénarios prédéfinis peuvent avoir des effets négatifs sur le comportement de l'utilisateur. Bien que n'offrant pas un objectif réel pour les sujets participant à l'expérience, les scénarios mettent les utilisateurs au fait de la façon dont ils doivent interagir avec le système. Les scénarios écrits peuvent inviter les sujets à imiter le langage donné dans le scénario, ce qui conduit à une lecture à haute voix au lieu d'un énoncé spontané. Il a été démontré que le choix des scénarios

pouvait également influer sur les stratégies de solution les plus efficaces pour mener à bien la tâche [43]. Les sujets appliquant des scénarios prédéfinis ne sont généralement pas particulièrement intéressés par la réponse du système, vu qu'ils n'ont pas réellement besoin du renseignement. En conséquence, le succès de la tâche peut ne pas se traduire par un effet important sur le jugement des sujets concernant l'utilisabilité du système. Par ailleurs, il a été signalé que les sujets participant aux essais ne lisaient pas toujours attentivement les instructions et pouvaient ne pas tenir compte des restrictions essentielles indiquées dans les scénarios ou mal les interpréter.

L'effet sur le langage de l'utilisateur peut être réduit à l'aide de descriptions graphiques de scénarios; voir les exemples indiqués à l'Appendice I. Une comparaison entre des scénarios écrits et des scénarios graphiques [6], [13] a montré que l'effet massif des scénarios écrits pouvait être presque totalement évité au moyen d'une représentation graphique, mais que la diversité des éléments linguistiques (nombre total de mots, nombre de mots ne figurant pas dans le vocabulaire) était similaire dans les deux cas. Ainsi, la diversité des langages doit encore être prise en charge en collectant les énoncés d'un nombre suffisamment élevé d'utilisateurs différents, par exemple dans une situation d'essai sur le terrain. Une autre possibilité consiste à présenter aux sujets une description orale préenregistrée des tâches et à leur conseiller de prendre des notes [42]. On espère ainsi que les processus de compréhension et de mémorisation impliqués communiqueront aux sujets le codage du sens de la description de la tâche, mais non la représentation de la forme superficielle. Toutefois, une preuve empirique de cette hypothèse n'a pas encore été donnée.

6.3 Sujets participant aux expériences

Selon la règle générale d'évaluation des expériences, le choix des sujets participant aux essais devrait être guidé par l'objet de l'essai. Par exemple, l'évaluation analytique des caractéristiques spécifiques du système ne sera possible que pour des sujets formés spécialistes du système étudié. Cela étant, ce groupe ne pourra pas juger les aspects globaux de la qualité du système d'une façon indépendante de leur connaissance du système. On ne peut attendre des jugements valables concernant la qualité globale que des sujets qui se rapprochent le plus possible du groupe des futurs utilisateurs du service. De même, les recommandations générales relatives aux conditions à remplir par les sujets données dans la Rec. UIT-T P.800 devraient être respectées pour les expériences d'interaction subjective avec des services basés sur un dialogueur automatique.

Un aperçu des facteurs utilisateur est donné au § 5.3. Certains de ces facteurs influent sur les caractéristiques acoustiques et linguistiques des paroles prononcées par l'utilisateur, à savoir l'âge, le sexe, la condition physique, la vitesse d'élocution, l'effort vocal, le langage natif, le dialecte ou l'accent. Comme ces facteurs peuvent être très critiques pour la reconnaissance vocale et la qualité de la compréhension, les jugements sur la qualité obtenus d'un groupe d'utilisateurs dont les caractéristiques acoustiques et linguistiques peuvent différer pourraient ne pas correspondre à la qualité à laquelle on peut s'attendre pour un groupe d'utilisateurs cible. Les groupes d'utilisateurs sont cependant variables et mal définis. Un service qui est accessible au grand public sera tôt ou tard confronté à un grand nombre d'utilisateurs différents. Les essais effectués avec des utilisateurs précis hors du groupe d'utilisateurs cible donneront donc une mesure de la robustesse du système par rapport aux caractéristiques des utilisateurs.

Un deuxième groupe de facteurs liés aux utilisateurs se rapporte à l'expérience et à la bonne connaissance du système, de la tâche et du domaine. D'après plusieurs études, l'expérience des utilisateurs affecte un large éventail de caractéristiques de la parole et du dialogue. Par exemple, il a été rapporté qu'en général, les utilisateurs résolvaient davantage de problèmes par appel lorsqu'ils s'étaient habitués au système et que l'interaction était plus courte [10]. D'autres études ont montré que le nombre de mots prononcés faisant partie du vocabulaire augmentait lorsque les utilisateurs s'étaient familiarisés avec le système. En même temps, le taux d'accomplissement des tâches s'accroissait [25]. La connaissance du système peut également réduire le nombre d'entrées de la part de l'utilisateur et le nombre de messages d'aide, et écourter la durée de la transaction [26], [28].

Les utilisateurs semblent élaborer des schémas d'interaction spécifiques lorsqu'ils se sont familiarisés avec un système. On est parti du principe qu'un tel schéma représentait un équilibre optimal perçu entre l'effort que chaque utilisateur devait fournir pendant l'interaction et l'efficience de l'interaction [39]. La quasi-totalité des utilisateurs semblent élaborer des schémas stables avec le système, mais les schémas ne sont pas identiques pour tous. Le schéma d'interaction élaboré par un utilisateur peut aussi traduire ce qu'il pense de l'agent machine, en ce sens que l'utilisateur peut avoir un "modèle cognitif" du système qui traduit ce qui est considéré comme la conviction du système [38]. Un tel modèle est déterminé en partie par les énoncés communiqués au système et en partie par les énoncés provenant du système. En général, l'utilisateur suppose que ses énoncés sont bien compris du système. En cas de mauvaise interprétation, l'utilisateur est confus et des problèmes se produiront probablement au niveau du flux de dialogue. Une autre source de divergence entre le modèle cognitif de l'utilisateur et la conviction du système tient au fait que le système a accès à des sources d'information secondaires telles qu'une base de données d'application. L'utilisateur peut être surpris s'il est confronté à des renseignements qu'il n'a pas fournis.

7 Questionnaires

Pour obtenir des renseignements sur les aspects de la qualité perçus par l'utilisateur, des jugements subjectifs doivent être recueillis. Deux principes différents peuvent être appliqués lors de la collecte, soit pour identifier d'une manière plus ou moins libre les aspects pertinents de la qualité, soit pour quantifier des aspects prédéterminés de la qualité en réponse à des questions à sélection de réponse ou des tâches d'évaluation. Les deux façons ont leurs avantages et leurs inconvénients: les enquêtes où les réponses sont librement données aident à trouver des aspects de la qualité qui autrement ne seraient pas détectées et à identifier les aspects qui sont les plus pertinents du point de vue de l'utilisateur. On peut ainsi faciliter l'interprétation des jugements quantitatifs correspondant aux questions à sélection de réponse. Les questions à sélection de réponse ou les tâches d'évaluation facilitent les comparaisons entre les sujets et les expériences, et elles donnent une signification exacte permettant de quantifier les perceptions de l'utilisateur. Elles peuvent être mises en œuvre assez facilement et les sujets non formés préfèrent souvent cette méthode.

Les tâches d'évaluation produiront des résultats valables et fiables lorsque deux conditions essentielles sont remplies: les éléments à juger doivent être choisis de manière adéquate et significative, et la mesure d'évaluation doit obéir à des règles bien établies. Des méthodes d'évaluation sont décrites en détail dans les ouvrages traitant de la psychométrie, par exemple en [21], [12] ou [7]. *Pour évaluer la qualité de la transmission*, l'UIT-T recommande la méthode de l'évaluation par catégories absolues (ACR, *absolute category rating*), la méthode de l'évaluation par catégories de dégradation (DCR, *degradation category rating*) et la méthode de l'évaluation par catégories de comparaison (CCR, *comparison category rating*); voir la Rec. UIT-T P.800. *Pour évaluer la qualité des services basés sur un dialogueur automatique*, on sollicite généralement des sujets des jugements sur des échelles d'évaluation continues ou sur des échelles d'évaluation par catégories absolues différentes. Une échelle ACR comprend un certain nombre de catégories discrètes dont l'une doit être choisie par les sujets. Les catégories sont affichées visuellement et peuvent être étiquetées au moyen d'attributs pour chaque catégorie ou uniquement pour les catégories extrêmes (le plus à gauche et le plus à droite). On trouvera ci-après des exemples d'échelles d'évaluation continues. Bien que l'échelle "impression globale" soit similaire à l'échelle ACR respective pour l'évaluation de la qualité de la transmission; voir la Rec. UIT-T P.800, il n'existe pas de relation directe entre les évaluations et donc pas de loi de la transformation qui lie les notes moyennes obtenues sur l'une des échelles continues aux notes moyennes d'appréciation utilisées pour décrire les jugements sur la qualité globale de la transmission.

La tâche d'évaluation sur l'échelle continue et sur l'échelle par catégories est souvent appelée "énoncé" (par exemple "Le système était facile à comprendre") et les sujets doivent exprimer leur accord en cochant la case ou catégorie correspondante de l'échelle. Cette méthode est inspirée des premières propositions faites par Likert [30] et une échelle est indiquée en exemple à la Figure 3.

Des chiffres sont attribués aux catégories ou aux positions de l'échelle selon que l'énoncé est positif (de 1 pour "fortement en désaccord" à 5 pour "fortement d'accord") ou négatif (de 5 pour "fortement en désaccord" à 1 pour "fortement d'accord"), les différentes évaluations étant additionnées pour tous les sujets. Une autre possibilité consiste à définir des étiquettes explicatives pour chaque catégorie, ainsi qu'il est proposé par l'UIT-T pour les expériences portant sur la qualité de la transmission vocale (voir la Rec. UIT-T P.800).

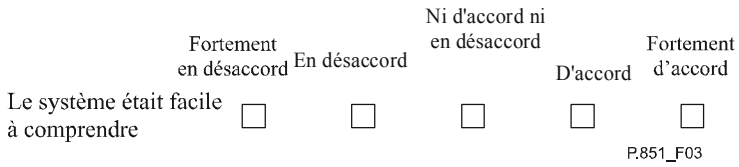


Figure 3/P.851 – Jugement d'un énoncé selon la méthode proposée par Likert [30]

Des échelles bien conçues ne donneront pas de renseignements valables lorsque la caractéristique de qualité à évaluer est mal définie ou lorsqu'elle n'est pas adéquatement choisie pour le service considéré. Dans les paragraphes ci-après, on trouvera à titre d'exemple un choix d'aspects de la qualité à examiner, chacun d'entre eux correspondant à une question précise qui doit être évaluée par les sujets. Des exemples de questions ou d'énoncés formulés sont également énumérés. Le choix des questions à poser pour un service donné dépendra du type de service, des tâches qui peuvent être effectuées, du comportement d'interaction spécifique du service, du groupe de sujets participant aux essais, ainsi que de l'objet précis de l'expérience d'évaluation. En général, le nombre d'éléments à évaluer dans un seul questionnaire devrait être limité à un nombre compris entre 15 et 20, de sorte que les sujets puissent différencier les divers éléments et en rendre compte.

Dans une configuration d'essai en laboratoire, on peut remettre un questionnaire aux sujets directement après une interaction (reflétant éventuellement l'impression dégagée après cette interaction), et/ou après un certain nombre d'interactions (avec intégration de certaines données provenant des expériences précédentes). Dans un essai sur le terrain, l'établissement des questionnaires est impossible à contrôler rigoureusement et le jugement porte généralement sur un certain nombre d'interactions menées pendant une période de temps largement définie. Il peut arriver que les expériences négatives prédominent et influent davantage sur le processus d'intégration temporelle que les expériences positives [11].

7.1 Questions relatives au profil de l'utilisateur

Au début de la séance d'essai, il devrait être répondu à un certain nombre de questions pour décrire l'utilisateur et les caractéristiques de son profil qui sont pertinentes pour l'expérience. Ces questions portent sur les points suivants:

- renseignements personnels: âge, sexe, profession, lieu de naissance, domicile, connaissance des langues;
- renseignements liés à la tâche: fréquence de la tâche, approche habituelle pour la réalisation de la tâche (autres interfaces possibles), motivation, autres aspects importants liés à la tâche et au domaine;
- renseignements liés au système: expérience des services basés sur un dialogueur automatique ou basés sur un système DTMF, expérience des dispositifs à technologie vocale (reconnaissance vocale, synthèse vocale, etc.).

La liste ci-après donne des exemples de questions qui peuvent être posées aux sujets. Ces questions se rapportent à un service d'information sur les restaurants mais peuvent être aisément adaptées à d'autres tâches et services.

Questions relatives au profil de l'utilisateur

Renseignements personnels

Sexe: ☐ féminin ☐ masculin

Age: _____ ans

Profession/instruction: _____

Région/ville de naissance: _____

Domicile actuel: _____

1 A quelle fréquence déjeunez-vous ou dînez-vous à l'extérieur en moyenne?

___ fois par semaine ___ fois par mois ___ fois par an

2 Comment chercheriez-vous un restaurant lorsque vous vous trouvez en un lieu étranger (choix multiples possibles)?

2.1 Magazines ☐
2.2 Dépliants publicitaires ☐
2.3 Guide de la ville ☐
2.4 Annuaire téléphonique ☐
2.5 Internet ☐
2.6 Conseils d'amis ☐
2.7 Appel d'un système automatique basé sur le dialogue ☐
2.8 Autre: _____ ☐

3 Qu'est-ce qui est important pour vous lorsque vous choisissez un restaurant (choix multiples possibles)?

3.1 Prix ☐
3.2 Type de nourriture ☐
3.3 Qualité de la nourriture ☐
3.4 Variété de la nourriture offerte ☐
3.5 Emplacement ☐
3.6 Ambiance ☐
3.7 Horaire d'ouverture ☐
3.8 Rapidité du service ☐
3.9 Convivialité du service ☐
3.10 Autre: _____

4 Avez-vous déjà utilisé un système d'information automatique basé sur le dialogue?
☐ oui ☐ non

4.1 Si oui, à quelle occasion?

_____

4.1.1 Comment qualifieriez-vous votre expérience à cet égard?

Extrêmement Mauvaise | Mauvaise | Médiocre | Passable | Bonne | Excellente | Parfaite

5 Avez-vous une expérience des systèmes de compréhension de la parole?
☐ oui ☐ non

5.1 Si oui, quel type de système?

_____

6 Avez-vous une expérience de la parole synthétisée?
☐ oui ☐ non

6.1 Si oui, à quelle occasion?

_____

7 Quels renseignements concernant un restaurant voulez-vous obtenir d'un système d'information?

_____

_____

7.2 Questions relatives à l'interaction individuelle

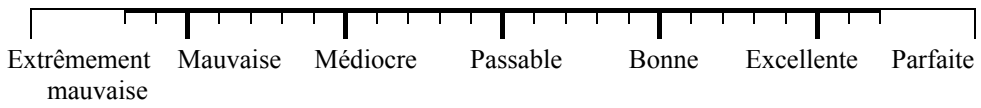
Après chaque interaction avec le service (simulé), les sujets doivent compléter un questionnaire comportant un certain nombre d'éléments liés à l'expérience individuelle en matière d'interaction. Ces éléments peuvent porter sur les aspects suivants:

- renseignements obtenus du système: disponibilité, exactitude, complétude, cohérence, fiabilité, clarté et véracité des renseignements obtenus, etc;
- capacité d'entrée/d'émission de parole: compréhension par le système, telle qu'elle est perçue; fréquence des erreurs du système; raisonnement du système, tel qu'il est perçu; effort d'écoute nécessaire pour comprendre les messages du système; intelligibilité perçue; utilisabilité perçue, etc;
- comportement interactif du système: transparence de l'interaction; conformité aux attentes de l'utilisateur; souplesse de l'interaction; fiabilité du traitement du système, telle qu'elle est perçue; répartition de l'initiative; capacité de contrôle de l'interaction; capacité de confirmation et de correction; reprise après des problèmes d'interaction; naturel de l'interaction; durée du dialogue; rapidité du système, telle qu'elle est perçue; caractère harmonieux du dialogue, etc;
- personnalité du système, telle qu'elle est perçue: convivialité, politesse, etc;
- impression sur l'utilisateur: naturel, tel qu'il est perçu, du comportement de l'utilisateur; caractère agréable; demande cognitive imposée à l'utilisateur; tension; énervement; etc;
- accomplissement de la tâche, tel qu'il est perçu: succès de l'opération, fiabilité des résultats de la tâche.

Des exemples de questions traitant de ces aspects sont indiqués ci-après. L'expérimentateur peut choisir les questions les plus adéquates pour le service étudié.

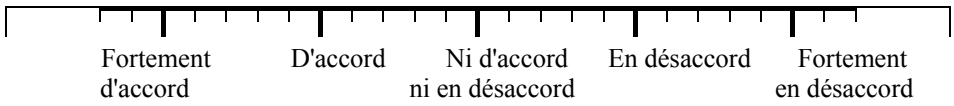
Questions relatives à l'interaction individuelle

Impression globale:

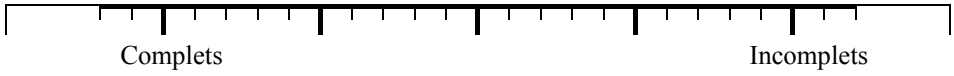


Renseignements obtenus du système

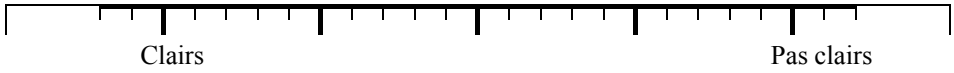
1 Le système a fourni les renseignements désirés:



2 Les renseignements fournis étaient …



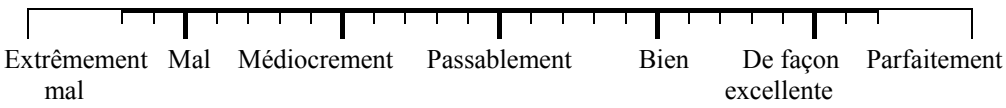
3 Les renseignements étaient …



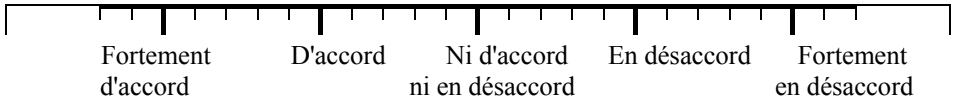
4 Vous évalueriez les renseignements comme étant …

Communication avec le système

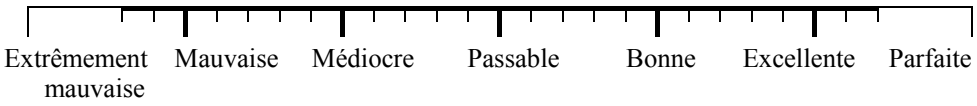
5 Comment pensez-vous avoir été compris par le système?



6 Vous avez dû vous concentrer pour comprendre ce que le système attendait de vous:

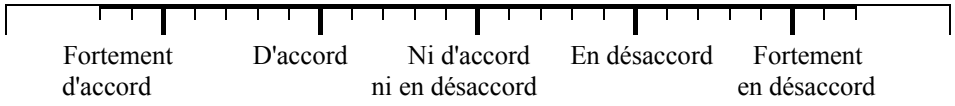


7 Quelle était l'intelligibilité du système du point de vue acoustique?

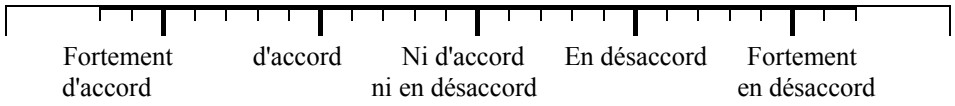


Comportement du système

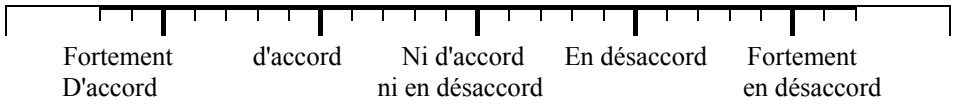
8 Vous saviez à chaque stade du dialogue ce que le système attendait de vous.



9 A votre avis, le système a correctement traité vos indications.

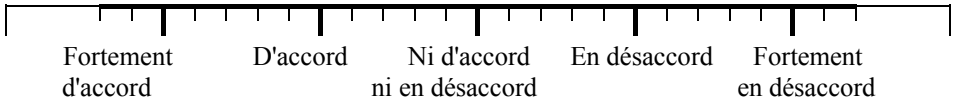


10 Le système s'est toujours comporté comme prévu.

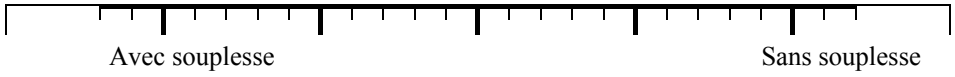


11 A quelle fréquence le système a-t-il fait des erreurs?

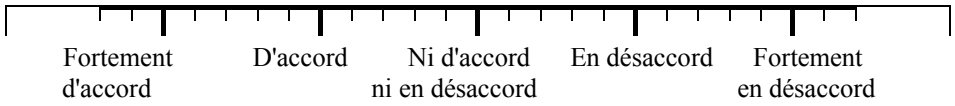
12 Le système a réagi de la même manière qu'un être humain.



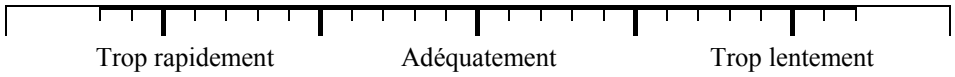
13 Le système a réagi ...



14 Vous avez pu contrôler le dialogue de la manière désirée.



15 Le système a réagi ...



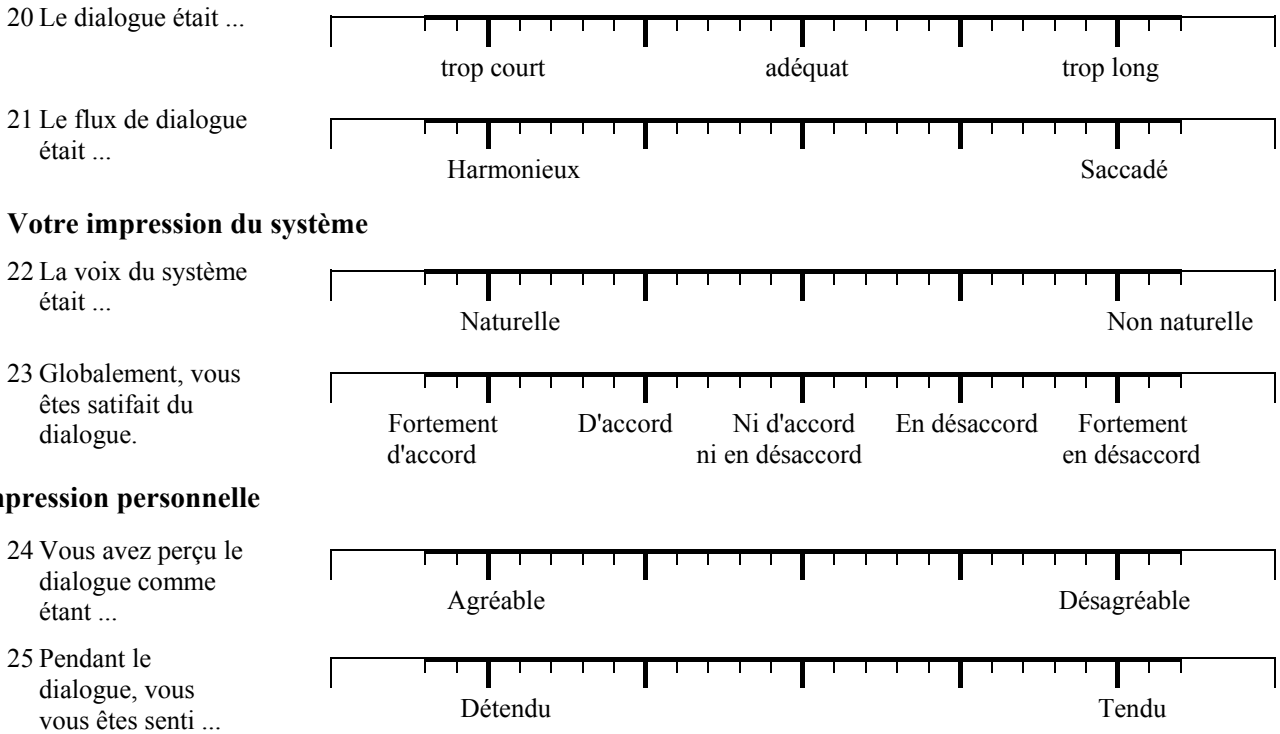
16 Le système a réagi de manière ...

Dialogue

17 Les énoncés du système étaient ...

18 Vous avez perçu le dialogue comme étant ...

19 Le flux de dialogue était ...



7.3 Questions relatives à l'impression globale de l'utilisateur concernant le système

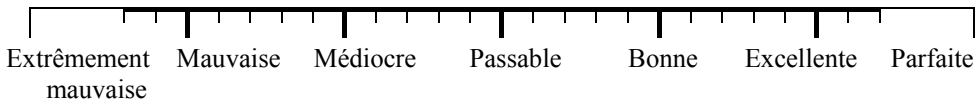
A la fin de toutes les interactions avec le service, les sujets devraient répondre à une série supplémentaire de questions, se rapportant cette fois-ci à l'expérience globale du système qu'ils ont acquise jusque-là. Les points suivants peuvent être inclus dans un tel questionnaire:

- impression globale de l'utilisateur concernant le système/service;
- manière ou expression du système;
- personnalité du système, telle qu'elle est perçue: convivialité, politesse, etc;
- capacités de correction, de reprise et d'aide du système;
- contrôle et initiative en matière d'interaction, tels qu'ils sont perçus;
- confort perçu pendant l'utilisation du système;
- accomplissement de la tâche, tel qu'il est perçu: succès de la tâche, fiabilité des résultats de la tâche;
- utilisabilité perçue: facilité d'utilisation, facilité pour apprendre à utiliser le système, habitabilité du système;
- degré auquel l'utilisateur a aimé utiliser le système, appréciabilité du système;
- adéquation et utilité du système pour l'accomplissement de la tâche;
- valeur ajoutée du système par rapport à d'autres interfaces ou à un opérateur humain;
- amélioration à apporter avant de pouvoir mettre le système en service;
- future utilisation prévue pour le service.

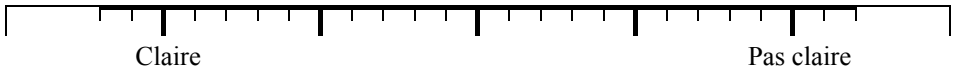
On trouvera ci-après un exemple de ce type de questionnaire pour un service d'information. Il peut être adapté et enrichi selon le service étudié, la tâche qu'il accomplit et le but de l'expérience.

Questions relatives à l'impression globale de l'utilisateur concernant le système

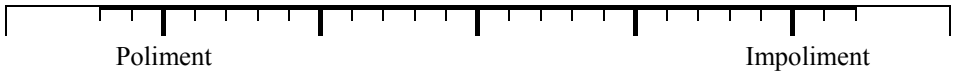
1 Impression globale.



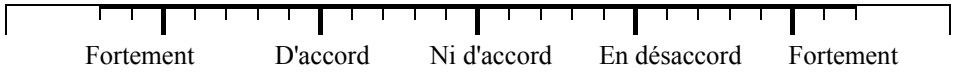
2 La façon de s'exprimer du système était ...



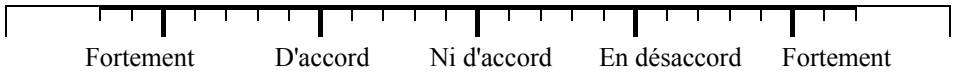
3 Le système a réagi ...



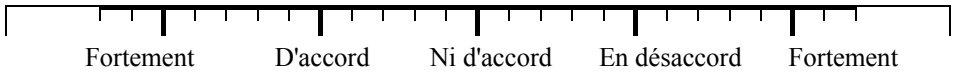
4 Vous auriez souhaité davantage d'aide de la part du système.



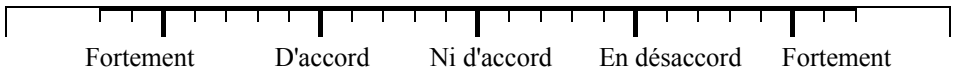
5 Le système a pu répondre à toutes vos questions.



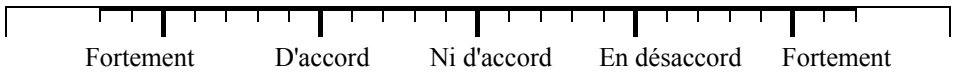
6 Les malentendus ont pu aisément être résolus.



7 Le système a contrôlé le flux de dialogue.

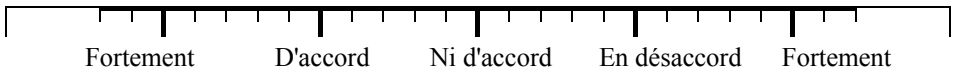


8 Vous avez pu manipuler le système sans problèmes.

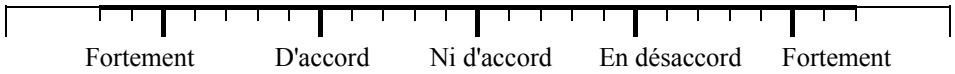


9 Concernant les dialogues, vous êtes ...

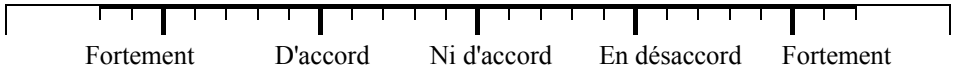
10 Vous avez trouvé les dialogues agréables.



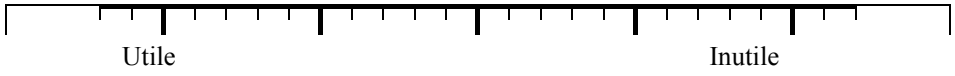
11 Vous estimez être adéquatement informé des possibilités du système.

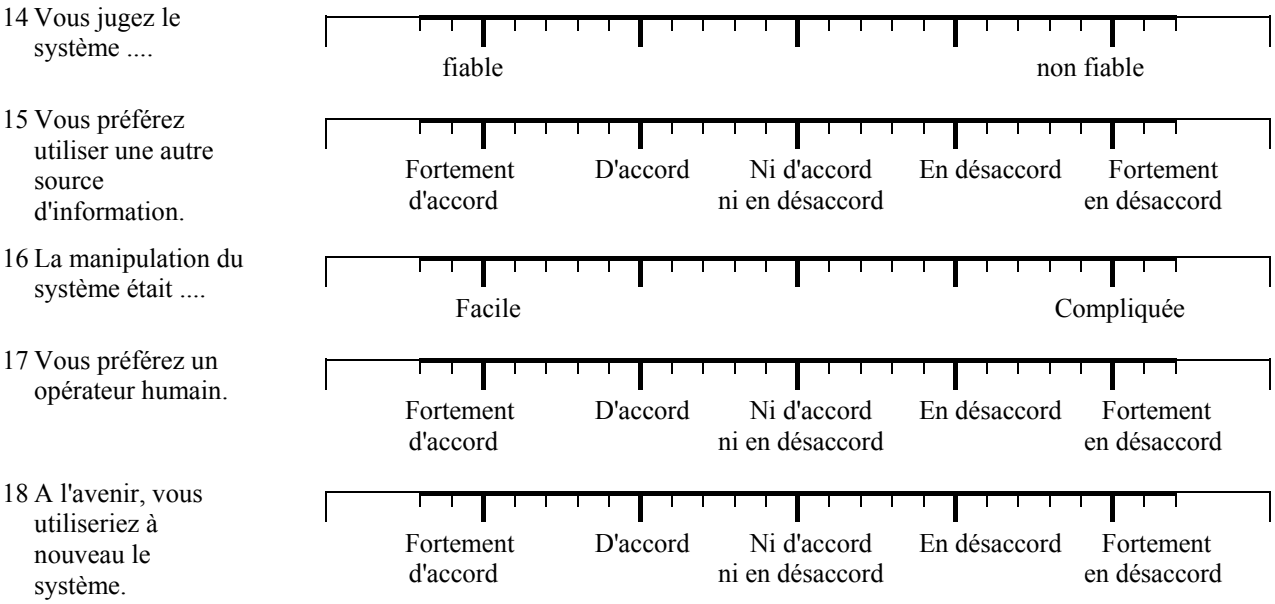


12 Les communications téléphoniques avec le système étaient utiles.



13 Vous avez perçu cette possibilité d'obtenir des renseignements comme étant ...





19 Quelles caractéristiques du système avez-vous le plus appréciées?

_____

_____

20 Quelles caractéristiques du système vous ont le plus perturbé?

_____

_____

21 Avez-vous des suggestions pour améliorer le système?

_____

_____

8 Evaluation de l'utilisabilité

Mis à part les questionnaires décrits qui permettent de traiter les différents aspects de l'utilisabilité, il est également possible de recourir à des méthodes spéciales pour procéder à cette évaluation. L'utilisabilité peut être évaluée avec des utilisateurs réels qui effectuent des essais spécifiques ou au moyen de méthodes d'inspection de l'utilisabilité et avec l'aide d'experts en évaluation. Les deux méthodes sont complémentaires en ce sens que les méthodes d'inspection de l'utilisabilité sont susceptibles de détecter les problèmes d'utilisation non décelés au cours des essais effectués avec des utilisateurs, et inversement [36]. En fait, on a observé un degré élevé de non-chevauchement entre les deux méthodes. Ainsi, l'évaluation devrait combiner des essais empiriques et des inspections de l'utilisabilité.

Les méthodes d'inspection de l'utilisabilité visent à déceler les problèmes dans un modèle d'interface utilisateur existant, éventuellement en évaluant la gravité des problèmes, en formulant des recommandations sur la façon de les corriger et donc en améliorant l'utilisabilité du système. Ces méthodes permettent aisément d'appliquer les connaissances et l'expérience des concepteurs d'interfaces utilisateur pour optimiser les nouveaux systèmes. Une partie importante de l'inspection de l'utilisabilité consiste à compter et à classer les problèmes d'utilisation qui sont observés dans l'interaction homme-machine. Cela étant, l'inspection devrait non seulement se révéler efficace pour détecter les problèmes, mais aussi pour les évaluer selon leur gravité (il n'est pas utile de résoudre les problèmes peu importants) et en particulier pour proposer des modifications et des améliorations au niveau de la conception. Comme de nombreuses méthodes d'inspection sont fondées sur la spécification et non sur l'implémentation du modèle, elles peuvent être appliquées à un stade relativement précoce du processus de conception du système.

On peut différencier les huit types suivants de méthodes d'inspection de l'utilisabilité [36]:

- *évaluation heuristique*: cette méthode informelle fait intervenir des spécialistes de l'utilisabilité qui jugent si un élément de dialogue est conforme aux principes établis de l'utilisabilité; c'est ce que l'on appelle l'heuristique;
- *contrôle du respect des lignes directrices*: inspections au cours desquelles on contrôle le service basé sur un dialogueur automatique pour vérifier s'il est conforme à une liste complète de directives sur l'utilisabilité. Le nombre total de directives pouvant être élevé, cette approche exige un degré élevé de connaissances spécialisées;
- *revues générales pluralistes*: réunions au cours desquelles les utilisateurs, les développeurs et des experts en facteurs humains appliquent ensemble un scénario et discutent des questions relatives à l'utilisabilité liées à des éléments de dialogue intervenant à chaque étape du scénario;
- *inspections de la compatibilité*: une interface est inspectée par plusieurs concepteurs représentant de multiples aspects de la conception, puis évaluée quant au point de savoir si elle est compatible avec toutes les questions de conception;
- *contrôle du respect des normes*: un expert examine une interface spécifique pour vérifier si elle est conforme à une norme définie;
- *revues cognitives*: simulation d'un processus de résolution de problèmes rencontrés par l'utilisateur à chaque étape de l'interaction et vérification du point de savoir si l'on peut supposer que les objectifs de l'utilisateur et sa mémorisation des opérations peuvent le conduire à effectuer correctement l'opération suivante. Ces revues revêtent généralement la forme de questions sur la relation entre les objectifs attribués à l'utilisateur et les opérations du système nécessaires pour réaliser ces objectifs;
- *inspections formelles de l'utilisabilité*: méthode formalisée faisant intervenir une équipe d'inspection. Chaque membre de l'équipe est chargé d'une tâche particulière, par exemple en tant que modérateur, que propriétaire du modèle ou qu'inspecteur. Des réunions sont organisées pour préparer et mener l'inspection, et pour analyser les résultats;
- *inspections des caractéristiques*: elles portent principalement sur les fonctions opérationnelles de l'interface utilisateur et sur la question de savoir si les fonctions offertes répondent aux besoins des utilisateurs finals ciblés.

La plupart de ces méthodes sont examinées en détail dans les ouvrages traitant de l'utilisabilité [36]. Le choix de la bonne méthode dépend des objectifs de l'évaluation, de l'existence de lignes directrices, de l'expérience de la personne chargée de l'évaluation ainsi que des contraintes temporelles et financières.

Une deuxième méthode d'évaluation de l'utilisabilité consiste à recourir à des expériences contrôlées faisant intervenir des utilisateurs. Ces essais peuvent être menées d'une manière "objective non intrusive" ou d'une manière "subjective intrusive" [19]. Les méthodes non intrusives tentent de saisir le comportement de l'utilisateur humain d'une manière naturelle et sans perturbations, par exemple en l'observant avec un équipement audiovisuel ou en l'enregistrant avec un dispositif d'enregistrement. Les méthodes intrusives exigent une participation active des utilisateurs, sous la forme de réponses à des questionnaires ou à des enquêtes (voir le dernier paragraphe), de discussions en groupe ou d'une manière autodescriptive, c'est-à-dire qu'un protocole verbal reflétant les pensées ou opinions des utilisateurs pendant ou après l'interaction est nécessaire. Ces méthodes sont décrites de façon plus détaillée dans les ouvrages traitant de l'évaluation de l'utilisabilité [14].

9 Analyse et interprétation des renseignements collectés

Il est possible d'analyser les jugements obtenus sur des échelles d'évaluation limitatives en recourant à des graphiques en colonnes ou à des distributions cumulatives. Bien que les distributions ne soient pas forcément de type gaussien, il est courant de calculer les moyennes arithmétiques (et non les valeurs médianes) dans toutes les évaluations obtenues avec une configuration de système déterminée; voir la Rec. UIT-T P.800 et le Manuel de téléphonométrie de l'UIT-T. Pour les valeurs moyennes, on évalue les limites de précision et on effectue des tests d'hypothèses à l'aide de l'analyse de variance (ANOVA) classique. Les hypothèses qui sous-tendent une analyse de variance (distribution de Gauss et homogénéité des variances) ne sont pas toujours satisfaites, mais cette méthode semble être assez robuste pour donner des résultats raisonnables également en cas d'écart par rapport à des conditions statistiquement idéales. Dans le cas d'un effet statistiquement significatif de l'une des variables aléatoires (configuration et/ou voix du système, sujet participant à l'expérience, ordre des conditions de l'expérience, séance d'essai, etc.), on peut recourir à un essai *a posteriori* pour effectuer des comparaisons par paires entre les moyennes et déterminer les sources de divergence. A cette fin, il est recommandé d'effectuer l'essai des différences honnêtement significatives (HSD, *honestly significant difference*) de Tukey [40]. Lorsque les hypothèses qui sous-tendent des statistiques paramétriques ne sont pas satisfaites, il est utile de résumer également les résultats sous la forme d'une médiane ou d'un mode, et d'utiliser des tests non paramétriques tels que celui proposé par Kruskal et Wallis à des fins de comparaison.

Lorsque des essais d'interaction sont réalisés dans des conditions contrôlées (en laboratoire), il est possible de collecter les paramètres d'interaction et les évaluations subjectives des utilisateurs pour la même interaction. En pareil cas, il est possible de quantifier la relation entre les deux types de mesures. Dans une deuxième étape, on peut tenter de prédire les jugements des utilisateurs concernant certains aspects de la qualité ou concernant la qualité globale, ou leur satisfaction à partir des paramètres d'interaction mesurables. Cela étant, il n'existe pas encore de méthode universelle (indépendante de la tâche) pour la modélisation de la qualité, qui permettrait de couvrir dans ses prédictions la majeure partie de la variance des jugements des utilisateurs. Un cadre général pour la prédiction de la qualité des systèmes à dialogue parlé a été proposé par Walker et divers collaborateurs dans 1997 ("PARAdigm for DIalogue System Evaluation, PARADISE"), sur la base d'une analyse de régression linéaire multivariable [44]. Le cadre est en principe indépendant de la tâche, mais les paramètres de la fonction de prédiction de la qualité doivent être à nouveau déterminés pour chaque système, à partir des expériences d'interaction subjectives. Par ailleurs, l'efficacité prédictive de cette méthode est encore très limitée (généralement entre 40 et 50% de la variance couverte). En conséquence, les paramètres d'interaction (tant ceux qui sont mesurables par des instruments que ceux qui sont mesurés par des experts) et les jugements des utilisateurs concernant la qualité restent les principales sources d'information qui décrivent la qualité d'une interaction avec un dialogueur automatique.

Appendice I

Exemples de scénarios

Scénario n° 1

Vous souhaiteriez savoir où vous pourriez consommer du canard. Veuillez le demander au système.

Nom du ou des restaurants: _____

Scénario n° 2

Vous projetez d'aller dîner dans un restaurant grec le mardi soir à XXX.

Prix: – ├─x─┼─────┼─────┼─────┤ +

Nom du ou des restaurants: _____

Si le système ne peut pas indiquer un restaurant, veuillez modifier l'indication comme suit:

Vous voulez dîner à YYY.

Nom du ou des restaurants: _____

Scénario n° 3

Vous projetez de déjeuner dans un restaurant chinois en ville.

Prix: – ├─────┼─────┼─────┼─x─┤ +

Nom du ou des restaurants: _____

Scénario n° 4

Vous projetez d'aller au restaurant à XXX. Votre restaurant favori étant fermé à cause des vacances, demandez au système de vous indiquer un restaurant.

Veuillez noter d'abord quelles indications vous voulez donner au système.

Si le système ne peut pas trouver un restaurant correspondant à votre demande, veuillez chercher une autre possibilité jusqu'à ce que le système indique au moins un restaurant.

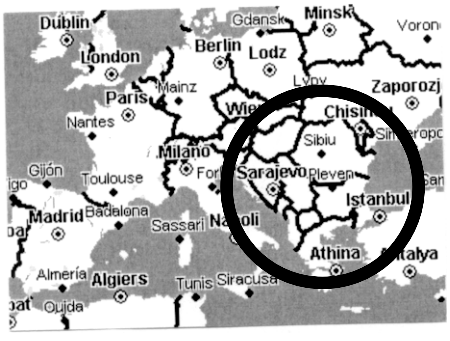
Nom du ou des restaurants: _____

Scénario n° 5

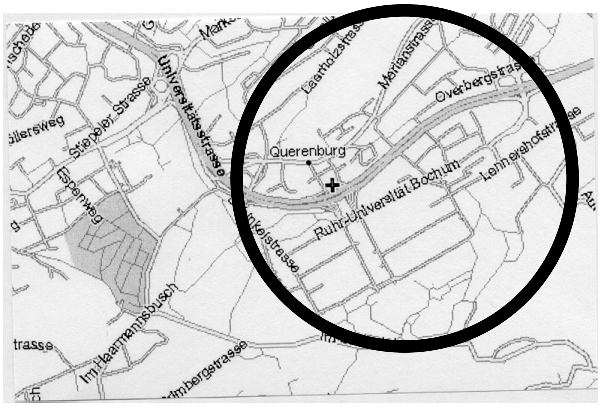
Veuillez collecter vos renseignements d'après les suggestions suivantes:

Prix: – ├─────┼──x──┼─────┤ +

Type de nourriture:



Emplacement:



Nom du ou des restaurants: _____

BIBLIOGRAPHIE

[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, Londres (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, Athènes (2000).

[6] BERNSEN (N.O.), DYBKJÆR (H.), DYBKJÆR (L.): Designing Interactive Speech Systems: From First Ideas to User Testing, *Springer*, Berne (1998).

[7] BORG (I.), STAUFENBIEL (T.): Theorien und Methoden der Skalierung: Eine Einführung, *Verlag Hans Huber*, Berne (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 Quaterly*, 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, Institut européen des normes de télécommunication, 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, Londres (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, Gênes (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*, Thèse d'habilitation (non publiée), 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, Edimbourg (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, Thèse d'habilitation, *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).

SÉRIES DES RECOMMANDATIONS UIT-T

Série A	Organisation du travail de l'UIT-T
Série B	Moyens d'expression: définitions, symboles, classification
Série C	Statistiques générales des télécommunications
Série D	Principes généraux de tarification
Série E	Exploitation générale du réseau, service téléphonique, exploitation des services et facteurs humains
Série F	Services de télécommunication non téléphoniques
Série G	Systèmes et supports de transmission, systèmes et réseaux numériques
Série H	Systèmes audiovisuels et multimédias
Série I	Réseau numérique à intégration de services
Série J	Réseaux câblés et transmission des signaux radiophoniques, télévisuels et autres signaux multimédias
Série K	Protection contre les perturbations
Série L	Construction, installation et protection des câbles et autres éléments des installations extérieures
Série M	RGT et maintenance des réseaux: systèmes de transmission, circuits téléphoniques, télégraphie, télécopie et circuits loués internationaux
Série N	Maintenance: circuits internationaux de transmission radiophonique et télévisuelle
Série O	Spécifications des appareils de mesure
Série P	**Qualité de transmission téléphonique, installations téléphoniques et réseaux locaux**
Série Q	Commutation et signalisation
Série R	Transmission télégraphique
Série S	Equipements terminaux de télégraphie
Série T	Terminaux des services télématiques
Série U	Commutation télégraphique
Série V	Communications de données sur le réseau téléphonique
Série X	Réseaux de données et communication entre systèmes ouverts
Série Y	Infrastructure mondiale de l'information, protocole Internet et réseaux de nouvelle génération
Série Z	Langages et aspects généraux logiciels des systèmes de télécommunication

Imprimé en Suisse
Genève, 2004