Union internationale des télécommunications

UIT-T

SECTEUR DE LA NORMALISATION
DES TÉLÉCOMMUNICATIONS
DE L'UIT

Série P

Supplément 24

(10/2005)

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

Paramètres décrivant l'interaction avec les dialogueurs automatiques

Recommandations UIT-T de la série P – Supplément 24

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	Series	P.10
Lignes et postes d'abonnés	Series	P.30
		P.300
Normes de transmission	Series	P.40
Appareils de mesures objectives	Series	P.50
		P.500
Mesures électroacoustiques objectives	Series	P.60
Mesures de la sonie vocale	Series	P.70
Méthodes d'évaluation objective et subjective de la qualité	Series	P.80
		P.800
Qualité audiovisuelle dans les services multimédias	Series	P.900
Aspects relatifs à la qualité de transmission et à la qualité de service aux points de terminaison des réseaux à protocole Internet	Series	P.1000

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

Supplément 24 aux Recommandations UIT-T de la série P

Paramètres décrivant l'interaction avec les dialogueurs automatiques

Résumé

Le présent Supplément fournit les définitions d'un ensemble de paramètres pouvant être extraits de services employant des dialogueurs automatiques. Ces paramètres peuvent être obtenus à partir des interactions entre un utilisateur enregistré (participant à un essai) et le service considéré. Ils permettent de quantifier le flux de l'interaction, le comportement de l'utilisateur et du système, ainsi que la performance des dispositifs à technologie vocale qui participent à l'interaction. Ils donnent par ailleurs des informations utiles relatives au développement, à l'optimisation et à la maintenance des systèmes et complètent les jugements subjectifs sur la qualité, recueillis conformément à la Rec. UIT-T P.851.

Source

Le Supplément 24 aux Recommandations UIT-T de la série P a été agréé le 21 octobre 2005 par la Commission d'études 12 (2005-2008) de l'UIT-T.

Mots clés

Compréhension automatique de la parole, dialogueur automatique, évaluation, génération de parole, gestion de dialogue, paramètre d'interaction, reconnaissance automatique de la parole, technologie 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 publication, l'expression "Administration" est utilisée pour désigner de façon abrégée aussi bien une administration de télécommunications qu'une exploitation reconnue.

Le respect de cette publication se fait à titre volontaire. Cependant, il se peut que la publication contienne certaines dispositions obligatoires (pour assurer, par exemple, l'interopérabilité et l'applicabilité) et considère que la publication 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 publication.

DROITS DE PROPRIÉTÉ INTELLECTUELLE

L'UIT attire l'attention sur la possibilité que l'application ou la mise en œuvre de la présente publication 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 publications.

A la date d'approbation de la présente publication, 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 publication. 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.

TABLE DES MATIÈRES

Supplément 24 aux Recommandations UIT-T de la série P

Paramètres décrivant l'interaction avec les dialogueurs automatiques

1 Domaine d'application

Le présent Supplément décrit les paramètres donnant des informations sur l'interaction avec les services qui emploient des dialogueurs, telle qu'elle est perçue par le développeur de systèmes ou par l'opérateur de services. Les dialogueurs automatiques abordés dans le présent Supplément permettent une interaction en langage parlé, à tour de rôle avec une personne humaine, par le biais du réseau téléphonique, et possèdent des capacités de reconnaissance vocale automatique, de compréhension de la parole, de gestion de dialogue, de génération de réponse et d'émission de parole. Ils permettent d'accéder à des informations stockées dans une base de données ou d'exécuter différents types de transactions.

Les paramètres définis dans le présent Supplément permettent de quantifier le flux de l'interaction, le comportement de l'utilisateur et du système, ainsi que la performance des dispositifs à technologie vocale qui participent à l'interaction. Afin d'extraire tous les paramètres, le dialogueur automatique doit être accessible comme une boîte en verre; toutefois, certains paramètres peuvent également être extraits selon une méthode de type boîte noire, c'est-à-dire sans accéder aux différents composants du système. L'extraction peut être effectuée en partie automatiquement et en partie manuellement, un expert transcrivant et annotant les fichiers de consignation des interactions. Les paramètres concernent la performance d'un système du point de vue du développeur; ils fournissent donc des informations venant compléter les expériences d'évaluation subjective effectuées avec des dialogueurs, pour lesquelles des orientations sont données dans la Rec. UIT-T P.851. On trouvera dans les Recommandations UIT-T P.800 et P.85, ainsi que dans le Manuel de téléphonométrie, d'autres lignes directrices concernant les méthodes d'évaluation subjective en général et l'évaluation des dispositifs d'émission de parole. Les paramètres énumérés dans le présent Supplément ne concernent pas expressément les éventuelles dégradations dues au canal de transmission. Ces effets doivent être étudiés plus avant par la CE 12 de l'UIT-T.

2 Références normatives

– 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*.

– Recommandation UIT-T P.851 (2003), *Evaluation subjective de la qualité des services téléphoniques basés sur des dialogueurs automatiques*.

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

3 Définitions

En ce qui concerne les définitions qui ne sont pas énumérées ci-après, se reporter à la Rec. UIT-T P.10.

3.1 intervention: capacité d'une personne à parler après une invite ou une sortie du système [10].

3.2 dialogue: conversation ou échange d'informations. En tant qu'unité d'évaluation: un des différents trajets possibles de la structure de dialogue.

3.3 **efficacité**: mesures de la précision et de la complétude des tâches système relatives aux ressources (temps, effort humain, ...) utilisées pour exécuter les différentes tâches du système.

3.4 **échange**: paire de tours de parole successifs en rapport l'un avec l'autre pris par chaque participant au dialogue [8].

3.5 **fonctionnalité**: capacité du système à assurer des fonctions répondant à des besoins exprimés ou implicites lorsque le système est utilisé dans des conditions particulières.

3.6 **métacommunication**: communication sur la communication; par exemple, dans le but de dissiper des malentendus ("Vous ai-je bien compris?") ou de parvenir à un accord sur la langue utilisée.

3.7 **performance**: capacité d'une unité à assurer la fonction pour laquelle elle a été conçue.

3.8 **technologie de la parole**: discipline se rapportant à la recherche et au développement de systèmes d'entrée et de sortie en langage parlé, utilisant les acquis des disciplines voisines telles que l'acoustique, l'ingénierie électrique, la statistique, la phonétique, le traitement du langage naturel, et qui concerne la spécification, la conception, l'implémentation et l'évaluation des systèmes, le traitement des corpus et des ressources linguistiques et l'évaluation du produit orienté consommateur [10].

3.9 **dialogueur automatique**: système informatique avec lequel des personnes interagissent en langage parlé par tours de parole.

3.10 **tâche**: ensemble des activités qu'un utilisateur doit développer pour atteindre un objectif dans un domaine donné.

3.11 **dialogue orienté tâche, dialogue finalisé**: dialogue portant sur un sujet particulier, qui vise un objectif précis (par exemple, la résolution d'un problème ou l'obtention d'informations particulières) [8].

3.12 **transaction**: partie d'un dialogue consacrée à une seule tâche de haut niveau (par exemple, la réservation d'un voyage ou la vérification du solde de son compte bancaire). Une transaction peut constituer le dialogue ou le dialogue comporter plusieurs transactions [8].

3.13 **tour de parole, énoncé**: séquence de parole prononcée par un participant à un dialogue, entre le moment où ce participant commence à parler et le moment où un autre participant prend la parole [1].

3.14 **énoncé**: voir tour de parole.

4 Abréviations

ASR	reconnaissance automatique de la parole (*automatic speech recognition*)
AVM	matrice attribut-valeur (*attribute-value matrix*)
AVP	paire attribut-valeur (*attribute-value pair*)
DARPA	Defense Advanced Research Projects Agency
DP	programmation dynamique (*dynamic programming*)
DTMF	multifréquence bitonalité (*dual tone multiple frequency*)
IVR	réponse vocale interactive (*interactive voice response*)
MOS	note moyenne d'opinion (*mean opinion score*)
SDS	dialogueur automatique (*spoken dialogue system*)
WoZ	magicien d'Oz (*wizard-of-Oz*)

5 Introduction

Les dialogueurs automatiques (SDS, *spoken dialogue system*), c'est-à-dire les systèmes informatiques avec lesquels des personnes interagissent à tour de rôle par l'intermédiaire du langage parlé, peuvent faire partie des réseaux téléphoniques modernes. Ils permettent d'accéder à des bases de données et à des transactions par le biais du téléphone, par exemple pour obtenir des renseignements sur les horaires des trains ou des avions, sur les cours de la Bourse, ou des renseignements touristiques, ou encore pour effectuer des opérations bancaires, des réservations d'hôtel, etc. A l'opposé des simples systèmes à réponse vocale interactive (IVR, *interactive voice response*) avec entrée multifréquence bitonalité (DTMF, *dual tone multiple frequency*), les dialogueurs automatiques possèdent toutes les fonctions d'interaction au moyen de la parole, y compris la reconnaissance de la parole de l'utilisateur, l'attribution d'un sens aux mots reconnus, la décision concernant la manière de poursuivre le dialogue, la formulation d'une réponse linguistique et la génération d'un résultat parlé à l'attention de l'utilisateur. Ainsi, une interaction parlée plus ou moins "naturelle" est possible entre l'utilisateur et le système.

Pour évaluer la qualité des services employant des dialogueurs automatiques du point de vue de l'utilisateur, la Commission d'études 12 de l'UIT-T a élaboré en 2003 la Rec. UIT-T P.851. Cette Recommandation décrit des méthodes permettant de réaliser des expériences d'évaluation subjective afin de déterminer la qualité *du point de vue de l'utilisateur*, le dialogueur automatique étant considéré comme une boîte noire. Grâce aux expériences réalisées conformément à la Rec. UIT-T P.851, il est possible d'obtenir des informations utiles relatives à la qualité, telle qu'elle est perçue par l'utilisateur. Toutefois, il peut être difficile de déterminer comment chacun des composants du système contribue à la qualité globale perçue par l'utilisateur, par exemple de déterminer quel élément il faut améliorer en cas de problème d'interaction. Par conséquent, l'évaluation devrait être complétée par des informations concernant la performance du système *du point de vue du concepteur du système et du point de vue de l'opérateur de services*.

Les renseignements concernant le système peuvent être décrits en termes de *paramètres d'interaction*. Ces paramètres permettent de quantifier le flux de l'interaction, le comportement de l'utilisateur et du système, ainsi que la performance des dispositifs à technologie vocale qui participent à l'interaction. Ils portent sur la performance du système du point de vue du concepteur du système et du point de vue de l'opérateur de services, et par conséquent, fournissent des informations venant compléter les données d'évaluation subjective. Pour pouvoir extraire certains des paramètres, le dialogueur automatique doit être accessible comme une boîte en verre; il est en outre possible d'extraire d'autres paramètres selon la méthode de type boîte noire, c'est-à-dire sans accéder à chacun des composants du système.

On trouvera dans le présent Supplément un ensemble de paramètres d'interaction qui ont été utilisés pour évaluer les dialogueurs automatiques au cours des 15 dernières années. Les paramètres énumérés concernent la communication globale de l'information entre un utilisateur et un système, la métacommunication dans le cas de malentendus, la coopérativité du système, la tâche qui peut être accomplie avec l'aide du système, et les fonctions d'entrée de parole du système. On ne dispose pas encore d'une description paramétrique de la qualité de sortie de parole (par exemple, par rapport à la qualité de la parole synthétisée). Cet ensemble de paramètres repose sur les travaux théoriques décrits dans la référence [17].

Tous les paramètres d'interaction ne se sont pas révélés être en relation directe avec la qualité perçue des services employant un dialogueur automatique. En fait, la corrélation entre les différents paramètres et les jugements des utilisateurs sur la qualité est généralement faible. Cela étant, il est utile de disposer d'un vaste ensemble de paramètres qui décrit l'interaction entre l'utilisateur et le système, de façon à recueillir le plus d'informations potentiellement pertinentes sur la qualité perçue du point de vue du concepteur du système. Ces paramètres donnent des informations utiles au développement, à l'optimisation et à la maintenance du système.

Les paramètres ayant été définis et utilisés dans des essais d'évaluation en des emplacements différents, il est possible d'estimer leur incidence sur la qualité perçue pour un grand nombre de systèmes et de services. On peut ainsi mettre au point des algorithmes permettant de prévoir la qualité sur la base des paramètres d'interaction. Des travaux dans ce domaine sont toujours en cours au sein de la Commission d'études 12 de l'UIT-T et d'autres organismes.

6 Caractéristiques des paramètres d'interaction

L'extraction des paramètres d'interaction peut se faire lorsque des utilisateurs réels ou des utilisateurs agissant à titre expérimental sont en interaction avec le service. Elle peut être effectuée en partie au moyen d'instruments et en partie avec l'aide de fichiers de consignation qui doivent être transcrits et annotés par un expert. Les paramètres simples, comme la durée de l'interaction ou des énoncés isolés, peuvent généralement être mesurés entièrement à l'aide d'instruments en employant les algorithmes appropriés. Une transcription et une annotation humaines sont toutefois nécessaires lorsque sont examinés non seulement la forme superficielle (signaux de parole), mais aussi le contenu et la signification des énoncés du système ou de l'utilisateur (par exemple, pour déterminer l'exactitude d'un mot ou d'un concept).

Les dialogueurs automatiques sont d'une telle complexité qu'il est nécessaire de décrire le comportement du système et de comparer les différents systèmes ou les différentes versions d'un même système sur la base d'une multitude de paramètres différents [24]. En conséquence, pour obtenir le plus d'informations possible, il convient d'appliquer les deux méthodes (celle faisant appel à des instruments et celle faisant appel à des experts) de collecte des paramètres d'interaction. Sur la base des informations recueillies, les services de dialogue automatique peuvent être optimisés et gérés avec une grande efficacité.

Puisqu'ils sont fondés sur des données qui ont été recueillies dans une interaction entre un utilisateur et un système, les paramètres d'interaction sont influencés par les caractéristiques du système, de l'utilisateur et de l'interaction entre eux. Ces effets ne peuvent généralement pas être pris isolément car le comportement de l'utilisateur est grandement influencé par celui du système (par exemple, les questions posées par le système), et vice versa (par exemple, le vocabulaire et le style d'élocution employés par l'utilisateur influent sur la capacité du système à reconnaître et à comprendre exactement la parole). Par conséquent, les paramètres d'interaction reflètent grandement les caractéristiques du groupe d'utilisateurs avec lesquelles ils ont été collectés.

Les paramètres d'interaction sont déterminés soit au moyen d'un montage expérimental en laboratoire dans des conditions bien définies, soit au moyen d'un essai sur le terrain. Dans ce dernier cas, il arrive qu'on ne puisse pas extraire tous les paramètres, toutes les informations nécessaires ne pouvant pas être rassemblées. Par exemple, si l'on doit établir qu'une interaction orientée tâche (par exemple, l'obtention d'horaires de train est réussie), il faut connaître les intentions exactes de l'utilisateur. Ces informations ne peuvent être recueillies que dans le cadre d'un montage en laboratoire (par exemple, selon la méthode décrite dans la Rec. UIT-T P.851). Au cas où le système entièrement intégré ne serait pas encore disponible, il est possible de recueillir les paramètres au moyen d'une simulation de type "magicien d'Oz" (WoZ, *wizard-of-Oz*), dans laquelle un expérimentateur remplace les parties manquantes du système à l'essai. Les caractéristiques d'une telle simulation doivent être prises en considération lors de l'interprétation des paramètres obtenus.

Les paramètres d'interaction peuvent être calculés au niveau des mots, au niveau des phrases ou des énoncés, ou au niveau d'une interaction complète ou d'un dialogue complet. Dans le cas de paramètres au niveau des mots ou des énoncés, on calcule souvent des valeurs moyennes pour chaque dialogue. Les paramètres recueillis à partir d'un groupe particulier d'utilisateurs peuvent être analysés en fonction de l'effet du système (ou de sa version), du groupe d'utilisateurs et du montage expérimental (scénarios, environnement d'essai, etc.), au moyen de méthodes statistiques normalisées. Les particularités de ces effets sont indiquées dans la Rec. UIT-T P.851.

7 Aperçu des paramètres d'interaction

Sur la base d'une vaste étude des travaux publiés, on a recensé les paramètres qui avaient été utilisés dans différentes évaluations et expériences au cours des 15 dernières années. Ces différents travaux font l'objet des références dans les documents [2][3][4][6][7][8][9][11][12][14][16][21][22][23][24][25][26][27][28][30][31][32], les paramètres étant résumés dans la référence [17]. Ces paramètres peuvent grosso modo être classés comme suit:

- paramètres liés au dialogue et à la communication;
- paramètres liés à la métacommunication;
- paramètres liés à la coopérativité;
- paramètres liés aux tâches;
- paramètres liés à l'entrée vocale.

Ces catégories seront brièvement examinées dans les paragraphes qui suivent. Pour chaque catégorie, on donnera la liste des paramètres correspondants, ainsi qu'une définition, le niveau d'interaction concerné par le paramètre (mot, énoncé ou dialogue), et la méthode de mesure employée (fondée sur des instruments ou des annotations d'experts).

7.1 Paramètres liés au dialogue et à la communication

Les paramètres qui se rapportent au dialogue dans son ensemble et à la communication des informations donnent une indication très grossière de la manière dont se déroule l'interaction. Ils ne spécifient pas de façon détaillée la fonction communicative de chaque énoncé. Les paramètres appartenant à cette catégorie sont énumérés dans le Tableau 1; ils comprennent les paramètres liés à la durée (durée du dialogue dans son ensemble, durée des tours de parole du système et de l'utilisateur, délai de réponse du système et de l'utilisateur), et les paramètres liés aux mots et aux tours de parole (nombre moyen de tours de parole du système et de l'utilisateur, nombre moyen de mots par tour du système et de l'utilisateur, nombre de questions posées par le système et par l'utilisateur).

Deux paramètres, proposés dans la référence [11], méritent d'être mentionnés: la *densité des interrogations (query density)*, qui donne une indication sur la manière dont un utilisateur peut fournir de façon efficace de nouvelles informations à un système, et l'*efficience des concepts (concept efficiency)*, qui décrit la manière dont le système peut absorber de façon efficace les informations fournies par l'utilisateur. Même si ces paramètres se rapportent également à la capacité de compréhension du langage du système, ils ont été inclus dans le présent paragraphe car ils résultent des capacités d'interaction du système dans leur ensemble et non pas simplement des capacités de compréhension du langage.

Tous les paramètres dans cette catégorie revêtent un caractère général et se rapportent au dialogue dans son ensemble, même s'ils sont calculés en partie au niveau des énoncés. Les paramètres globaux sont parfois problématiques dans la mesure où les différences individuelles en termes de faculté cognitive peuvent être importantes par rapport aux différences produites par le système, et dans la mesure où les sujets peuvent apprendre des stratégies de résolution de tâches qui ont une grande incidence sur les paramètres globaux.

Tableau 1 – Paramètres d'interaction liés au dialogue et à la communication

Abréviation	Nom	Définition	Niveau d'inter-action	Méthode d'évaluation
DD	Durée du dialogue	Durée totale d'un dialogue en [ms], voir par exemple [8][6][12][21].	Dialogue	Instruments
STD	Durée du tour de parole du système	Durée moyenne d'un tour de parole du système, entre le moment où le système commence à parler et le moment où il s'arrête de parler, en [ms]. Un tour de parole est un énoncé, c'est-à-dire une étendue de parole prononcée par une entité participant au dialogue. [8]	Enoncé	Instruments
UTD	Durée du tour de parole de l'utilisateur	Durée moyenne d'un tour de parole de l'utilisateur, entre le moment où l'utilisateur commence à parler et le moment où il s'arrête de parler, en [ms]. [8]	Enoncé	Instruments
SRD	Délai de réponse du système	Délai moyen d'une réponse du système, entre le moment où l'utilisateur s'arrête de parler et le moment où le système commence à parler, en [ms]. [22]	Enoncé	Instruments
URD	Délai de réponse de l'utilisateur	Délai moyen d'une réponse de l'utilisateur, entre le moment où le système s'arrête de parler et le moment où l'utilisateur commence à parler en [ms]. [22]	Enoncé	Instruments
# turns	Nombre de tours de parole	Nombre total de tours de parole pris dans un dialogue. [30]	Dialogue	Instruments/ experts
# system turns	Nombre de tours de parole du système	Nombre total de tours de parole de système pris dans un dialogue. [30]	Dialogue	Instruments/ experts
# user turns	Nombre de tours de parole de l'utilisateur	Nombre total de tours de parole pris par le système dans un dialogue. [30]	Dialogue	Instruments/ experts
WPST	Mots par tour de parole du système	Nombre moyen de mots par tour de parole du système dans un dialogue. [6]	Enoncé	Instruments/ experts
WPUT	Mots par tour de parole de l'utilisateur	Nombre moyen de mots par tour de parole de l'utilisateur dans un dialogue. [6]	Enoncé	Instruments/ experts
# system questions	Nombre de questions posées par le système	Nombre total de questions posées par le système par dialogue.	Dialogue	Experts
# user questions	Nombre de questions posées par l'utilisateur	Nombre total de questions posées par l'utilisateur par dialogue. [12][21]	Dialogue	Experts

Tableau 1 – Paramètres d'interaction liés au dialogue et à la communication

Abréviation	Nom	Définition	Niveau d'inter-action	Méthode d'évaluation
QD	Densité des interrogations	Nombre moyen de nouveaux concepts (créneaux, voir § 7.4) introduits par interrogation de l'utilisateur. Soit n_d le nombre de dialogues, $n_q(i)$ le nombre total d'interrogations de l'utilisateur dans le *ième* dialogue, et $n_u(i)$ le nombre de concepts uniques correctement "compris" par le système dans le *ième* dialogue, alors $QD = \frac{1}{n_d}\sum_{i=1}^{n_d}\frac{n_u(i)}{n_q(i)}$ Un concept n'est pas inclus dans le nombre $n_u(i)$ si le système l'a déjà compris dans un des énoncés précédents. [11]	Ensemble de dialogues	Experts
CE	Efficience des concepts	Nombre moyen de tours de parole nécessaires pour que chaque concept soit "compris" par le système. Soit n_d le nombre de dialogues, $n_u(i)$ le nombre de concepts uniques correctement "compris" par le système dans le *ième* dialogue, et $n_c(i)$ le nombre total de concepts dans le *ième* dialogue, alors $CE = \frac{1}{n_d}\sum_{i=1}^{N_d}\frac{n_u(i)}{n_c(i)}$ Il est tenu compte d'un concept à chaque fois qu'il a été énoncé par l'utilisateur et qu'il n'a pas déjà été compris par le système. [11]	Ensemble de dialogues	Experts

7.2 Paramètres liés à la métacommunication

La métacommunication, c'est-à-dire la communication sur la communication, est particulièrement importante pour l'interaction parlée avec les systèmes qui ont des capacités de reconnaissance, de compréhension et de raisonnement limitées. Dans ce cas, des énoncés de correction et de clarification ou même des sous-dialogues sont nécessaires pour dissiper les malentendus.

Les paramètres appartenant à ce groupe quantifient le nombre d'énoncés du système ou de l'utilisateur, qui font partie d'une métacommunication. On calcule la plupart des paramètres comme étant le nombre absolu d'énoncés dans un dialogue, qui se rapportent à un problème d'interaction particulier, puis on établit leur moyenne sur un ensemble de dialogues. Sont inclus le nombre de demandes d'aide émanant de l'utilisateur, le nombre d'annonces d'expiration de temporisation émises par le système, le nombre d'énoncés d'utilisateur rejetés par le système dans le cas où aucun contenu sémantique n'a pu être extrait (rejets dans le cadre de la reconnaissance automatique de la parole (ASR, *automatic speech recognition*), le nombre de messages d'erreur de diagnostic émis par le système, ainsi que le nombre de tentatives d'intervention de l'utilisateur et de tentatives d'annulation d'une action précédente par l'utilisateur.

La capacité du système (et de l'utilisateur) à récupérer lors de problèmes d'interaction peut être décrite de deux manières: soit explicitement au moyen du taux de correction, c'est-à-dire le pourcentage des tours de parole (du système ou de l'utilisateur) qui concernent en premier lieu la résolution d'un problème d'interaction, soit implicitement au moyen du paramètre *récupération implicite*, qui quantifie la capacité du système à récupérer les énoncés qui n'ont en partie pas pu être reconnus ou compris.

Contrairement aux grandeurs globales, la plupart des paramètres liés à la métacommunication décrivent la fonction des énoncés du système et de l'utilisateur dans le processus de communication. Ainsi, la plupart des paramètres doivent être déterminés avec l'aide d'un expert chargé de l'annotation. Ces paramètres sont énumérés dans le Tableau 2.

Tableau 2 – Paramètres d'interaction relatifs à la métacommunication

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'évaluation
# help request	Nombre de demandes d'aide émanant de l'utilisateur	Nombre total de demandes d'aide émanant de l'utilisateur dans un dialogue. Une demande d'aide est étiquetée par l'expert chargé de l'annotation si l'utilisateur demande explicitement de l'aide. Cette demande peut être formulée sous la forme d'une question (par exemple "Quelles sont les options disponibles?") ou sous la forme d'une assertion ("Donnez-moi les options disponibles!"). [30]	Enoncé	Experts
# system help	Nombre de messages d'aide concernant le diagnostic émis par le système	Nombre total de messages d'aide émis par le système dans un dialogue. Un message d'aide est un énoncé du système qui informe l'utilisateur des options disponibles à un certain point du dialogue.	Enoncé	Instruments/ experts
# time-out	Nombre d'annonces d'expiration de tempori-sation	Nombre total d'annonces d'expiration de temporisation, dues à l'absence de réponse de la part de l'utilisateur, dans un dialogue. [30]	Enoncé	Instruments
# ASR rejection	Nombre de rejets ASR	Nombre total de rejets ASR dans un dialogue. On entend par "rejet ASR" une annonce du système indiquant que celui-ci n'a pas été en mesure "d'entendre" ou de "comprendre" l'utilisateur (le système n'a pu extraire aucune signification de l'énoncé de l'utilisateur). [30]	Enoncé	Instruments
# system error	Nombre de messages d'erreur de diagnostic émis par le système	Nombre total de messages d'erreur de diagnostic émis par le système dans un dialogue. On entend par "message d'erreur de diagnostic" un énoncé du système dans lequel celui-ci indique qu'il n'est pas en mesure d'accomplir une certaine tâche ou de fournir une certaine information. [22]	Enoncé	Instruments/ experts

Tableau 2 – Paramètres d'interaction relatifs à la métacommunication

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'évaluation
# barge-in	Nombre de tentatives d'intervention faites par l'utilisateur	Nombre total de tentatives d'intervention faites par l'utilisateur dans un dialogue. Un utilisateur fait une tentative d'intervention lorsqu'il s'adresse volontairement au système alors que celui-ci est toujours en train de parler. Selon cette définition, les énoncés d'utilisateur qui ne sont pas destinés à influencer le cours du dialogue (rires, expressions de colère ou marques de politesse) ne sont pas considérés comme des interventions. [30]	Enoncé	Experts
# cancel	Nombre de tentatives d'annulation faites par l'utilisateur	Nombre total de tentatives d'annulation faites par l'utilisateur dans un dialogue. Un tour de parole d'un utilisateur est considéré comme une tentative d'annulation si cet utilisateur essaie de reprendre le dialogue depuis le début ou s'il souhaite explicitement revenir sur ses pas dans la hiérarchie du dialogue. [16][23]	Enoncé	Experts
SCT, SCR	Nombre de tours de parole du système concernant une correction, taux de correction par le système	Nombre (SCT) ou pourcentage (SCR) total des tours de parole du système dans un dialogue, qui concernent en premier lieu la résolution d'un "problème" et qui interrompent le flux du dialogue sans y ajouter de nouveaux contenus propositionnels. Un "problème" peut être causé par des erreurs de reconnaissance ou de compréhension vocale ou par des énoncés d'utilisateur illogiques, contradictoires ou non définis. Dans le cas où l'utilisateur ne donne pas de réponse à une question posée par le système, la réponse correspondante donnée par le système est étiquetée comme un tour de parole de type correction par le système, sauf lorsque l'utilisateur demande une information ou une action qui n'est pas prise en charge par la fonction courante du système. [8][24][9][7]	Enoncé	Experts
UCT, UCR	Nombre de tours de parole de l'utilisateur concernant une correction, taux de correction par l'utilisateur	Nombre (UCT) ou pourcentage (UCR) total des tours de parole de l'utilisateur dans un dialogue, qui concernent en premier lieu la résolution d'un "problème" et qui interrompent le flux du dialogue sans y ajouter de nouveaux contenus propositionnels (voir SCT, SCR). [8][24][9][7]	Enoncé	Experts

Tableau 2 – Paramètres d'interaction relatifs à la métacommunication

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'évaluation
IR	Récupé-ration implicite	Capacité du système à récupérer lors d'énoncés de l'utilisateur pour lesquels le processus de reconnaissance ou de compréhension de la parole a en partie échoué. Elle est déterminée par l'étiquetage des énoncés partiellement analysés (voir la définition de PA:PA au § 7.5), qui indique si la réponse du système était "pertinente" ou pas: $IR = \frac{nombre\ d'énoncés\ avec\ réponse\ pertinente\ du\ système}{PA:PA}$ Pour la définition de "la pertinence" (*appropriateness*), se reporter au § 7.3. [7]	Enoncé	Experts

7.3 Paramètres liés à la coopérativité

La coopérativité est considérée comme un facteur essentiel à la réussite d'une interaction avec un dialogueur automatique [1]. Malheureusement, il est difficile de quantifier si un système se comporte de façon coopérative ou non. Plusieurs des paramètres liés au dialogue et à la métacommunication se rapportent d'une façon ou d'une autre à la coopérativité du système, mais ils ne sont pas censés quantifier ce facteur.

Des mesures directes de la coopérativité sont obtenues au moyen des paramètres de pertinence contextuelle introduits par Simpson et Fraser [24]. Chaque énoncé de système doit être évalué par un certain nombre d'experts pour savoir s'il viole une ou plusieurs des maximes de Grice relatives à la coopérativité (voir la référence [13]):

- *quantité* d'informations: rendez votre contribution aussi informative que nécessaire (pour l'objectif ordinaire de l'échange); ne rendez pas votre contribution plus informative que nécessaire;
- *qualité*: rendez votre contribution véridique; n'affirmez pas ce que vous croyez être faux; n'affirmez pas ce pour quoi vous manquez de preuves;
- *relation*: soyez pertinent;
- *modalité*: soyez clair; évitez de vous exprimer peu clairement; évitez d'être ambigu; soyez bref (ne soyez pas plus prolixe qu'il n'est nécessaire); soyez méthodique.

Ces principes ont été énoncés avec plus de précision par Bernsen et Dybkjær [1] dans le cas des dialogueurs automatiques.

Les énoncés sont classés en énoncés pertinents (ne violant pas les maximes de Grice), non pertinents (violant une ou plusieurs maximes), pertinents ou non pertinents (les experts ne s'accordent pas quant à leur classement), incompréhensibles (le contenu de l'énoncé n'est pas perceptible dans le contexte du dialogue), ou échec total (aucune réponse linguistique de la part du système). Il convient de noter que cette classification n'est pas toujours évidente et que des principes d'interprétation peuvent être nécessaires.

Tableau 3 – Paramètres d'interaction liés à la coopérativité

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'évalua-tion
CA:AP, CA:IA, CA:TF, CA:IC, *%CA:AP, %CA:IA, %CA:TF, %CA:IC*	Pertinence contextuelle	Nombre ou pourcentage total des énoncés du système qui sont considérés comme étant pertinents dans le contexte immédiat du dialogue. Ce paramètre est déterminé par l'étiquetage des énoncés selon qu'ils violent ou pas une ou plusieurs des maximes de Grice relatives à la coopérativité: • *CA:AP*: énoncé pertinent, ne violant pas les maximes de Grice qui n'est pas, contre toute attente, marqué ou caractérisé d'une certaine façon; • *CA:IA*: énoncé non pertinent, violant une ou plusieurs des maximes de Grice; • *CA:TF*: échec total, aucune réponse linguistique; • *CA:IC*: énoncé incompréhensible; le contenu ne peut être compris par l'expert chargé de l'annotation. Pour plus de détails, voir les références [24][8][9]; la classification est analogue à celle adoptée dans la référence [14].	Enoncé	Experts

7.4 Paramètres liés aux tâches

Les services courants les plus modernes permettent des interactions orientées tâche entre le système et l'utilisateur, la réussite d'une tâche étant un facteur essentiel pour qu'un service soit utile. Cette réussite peut être le mieux évaluée en laboratoire où les sujets participant à l'expérience se voient attribuer des tâches explicites (voir la Rec. UIT-T P.851). Toutefois, les mesures réalistes de la réussite des tâches doivent tenir compte des écarts par rapport au scénario que l'utilisateur a pu utiliser, soit parce qu'il n'a pas fait attention aux consignes données dans le scénario en raison de son inattention aux énoncés du système, soit parce que la tâche était insoluble et qu'elle a dû être modifiée au cours du dialogue.

Il est tenu compte de la modification de la tâche expérimentale dans la plupart des définitions de la réussite d'une tâche, figurant dans les publications sur le sujet. On réussit après avoir simplement fourni la réponse correcte aux injonctions énoncées dans les consignes; après que le système ou l'utilisateur (ou les deux) a assoupli les règles; ou après avoir constaté qu'aucune solution n'existe pour la tâche donnée. L'échec d'une tâche peut provisoirement être attribué au comportement du système ou de l'utilisateur, le comportement de l'utilisateur étant toutefois influencé par celui du système.

Le coefficient κ est un autre moyen pour déterminer la réussite d'une tâche. Cette méthode repose sur une stratégie de compréhension de la parole qui fait appel à des attributs (concepts, lots) auxquels des valeurs autorisées doivent être attribuées au cours du dialogue entre le système et l'utilisateur. Les paires d'attributs et de valeurs attribuées sont appelées "paires attribut-valeur" (AVP, *attribute-value pairs*). L'ensemble de tous les attributs disponibles ainsi que des valeurs attribuées par la tâche matrice attribut-valeur (AVM, *attribute-value matrix*) décrit complètement une tâche qui peut être exécutée avec l'aide du système. Pour déterminer le coefficient κ, on établit une matrice de confusion *M(i,j)* pour les attributs dans la clé (définition du scénario) et dans la solution indiquée (fichier de consignation du dialogue). Ensuite, on peut calculer l'accord entre la clé et la solution *P(A)* et l'accord de chance *P(E)* à partir de cette matrice (voir le Tableau 4). La

matrice *M(i,j)* peut être calculée pour chacun des dialogues ou pour un ensemble de dialogues appartenant à un système particulier ou à une configuration de système particulière.

Le coefficient κ dépend de la disponibilité d'un simple système de codage de tâche, c'est-à-dire d'une matrice AVM. Toutefois, certaines tâches ne peuvent pas être caractérisées aussi facilement. Il est alors nécessaire de mettre au point des méthodes plus élaborées d'évaluation de la réussite des tâches qui dépendent généralement du type de tâche considéré.

Tableau 4 – Paramètres d'interaction liés aux tâches

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'éva-luation
TS	Réussite de la tâche	Etiquette de réussite de la tâche indiquant si l'utilisateur a atteint son objectif avant la fin d'un dialogue, à condition que cet objectif ait pu être atteint avec l'aide du système. Les étiquettes indiquent si l'objectif a été atteint ou non et précisent l'origine supposée des problèmes: • *TS:S*: réussite (tâche pour laquelle il existe des solutions) • *TS:SCs*: réussite avec assouplissement des règles par le système • *TS:SCu*: réussite avec assouplissement des règles par l'utilisateur • *TS:SCsCu*: réussite avec assouplissement des règles par le système et par l'utilisateur • *TS:SN*: réussite après constat qu'aucune solution n'existe • *TS:Fs*: échec à cause du comportement du système (adéquation) • *TS:Fu*: échec à cause du comportement de l'utilisateur (comportement non coopératif) Voir également les références [8][7][24].	Dialogue	Experts

Tableau 4 – Paramètres d'interaction liés aux tâches

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'éva-luation
κ	Coefficient kappa	Pourcentage des tâches accomplies conformément aux statistiques kappa. Ce pourcentage est déterminé en fonction de l'exactitude de la matrice AVM résultante, obtenue à la fin d'un dialogue, par rapport à la matrice AVM du scénario (clé). Une matrice de confusion *M(i,j)* est établie pour les attributs dans le résultat et dans la clé, *T* étant le nombre de comptages dans *M* et t_i la somme des comptages dans la *ième* colonne de *M*. Alors $\kappa = \frac{P(A) - P(E)}{1 - P(E)}$ *P(A)* étant le nombre de fois que les matrices AVM du dialogue en cours et de la clé correspondent, $P(A) = \sum_{i=1}^{n} \frac{M(i,i)}{T}$. *P(E)* peut être estimé à partir du nombre de fois qu'elles correspondent par hasard, $P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2$. [31][4]	Dialogue ou ensemble de dialogues	Experts

7.5 Paramètres liés à l'entrée vocale

La capacité d'entrée vocale d'un dialogueur automatique est déterminée par sa capacité à reconnaître les mots et les énoncés et à extraire la signification de la chaîne reconnue (la "compréhension de la parole"). On distingue deux méthodes de reconnaissance automatique de la parole: celle qui fait appel aux reconnaisseurs de mots pouvant extraire des mots isolés du discours de l'utilisateur, lorsque ceux-ci sont prononcés séparément (reconnaissance de mots isolés) ou continûment (détection de mots clés) et celle qui fait appel aux reconnaisseurs de parole continue pouvant reconnaître des phrases entières ou des énoncés entiers. La compréhension de la parole se fait souvent sur la base de paires attribut-valeur (voir le § 7.4). Les paramètres décrits ci-après concernent tant la reconnaissance de la parole que la compréhension de la parole.

Les reconnaisseurs de parole continue produisent généralement une chaîne de mots hypothétique. Afin de déterminer si la chaîne représente correctement ce qui a été dit, l'expert chargé de la transcription doit fournir une transcription de référence. Pour chaque énoncé, la chaîne hypothétique et la chaîne de référence sont d'abord alignées au niveau des mots au moyen d'une programmation dynamique (DP, *dynamic programming*) fondée sur un algorithme de concordance [19][20]. Sur la base de cet alignement, on compte le nombre de mots c_w déterminés correctement, le nombre de substitutions s_w, le nombre d'insertions i_w et le nombre de suppressions d_w. Ces comptages peuvent être reliés au nombre total de mots dans la référence n_w, ce qui donne deux autres mesures de la qualité de reconnaissance, à savoir le taux d'erreurs sur les mots (WER, *word error rate*) et l'exactitude des mots (WA, *word accuracy*) (voir le Tableau 5).

Des mesures de qualité complémentaires peuvent être définies au niveau de la phrase: l'exactitude de la phrase (SA, *sentence accuracy*) ou le taux d'erreurs sur la phrase (SER, *sentence error rate*)

(voir le Tableau 5). En règle générale, le taux *SA* est inférieur au taux *WA*, parce qu'un seul mot mal reconnu dans une phrase influe sur le paramètre *SA*. Il arrive toutefois qu'il soit supérieur au taux *WA*, en particulier lorsque de nombreuses phrases à un mot sont correctement reconnues. Strik et al. ont attiré l'attention sur le fait que les taux *SER* et *SA* pénalisent un énoncé entier lorsqu'un seul mot est mal reconnu [26] [27]; ce problème peut être évité au moyen des paramètres *NES* et *WES* (voir le Tableau 5). Lorsque les énoncés ne sont pas subdivisés en phrases, tous les paramètres liés à la phrase peuvent aussi être calculés au niveau d'un énoncé plutôt qu'au niveau d'une phrase.

Les reconnaisseurs de mots isolés produisent un mot ou énoncé hypothétique à la sortie pour chaque mot ou énoncé entré. Il est possible de comparer directement les mots entrés et sortis et de définir des mesures de qualité analogues à celles définies dans le cas de la reconnaissance continue, en omettant les insertions. Au lieu des insertions, on peut compter le nombre de fausses alarmes au cours d'une période de temps donnée (voir van Leeuwen et Steeneken [28]). Les paramètres *WA* et *WER* peuvent également être déterminés pour les mots clés seulement, lorsque le reconnaisseur fonctionne en mode détection des mots clés.

En ce qui concerne l'évaluation de la compréhension de la parole, il convient de distinguer deux méthodes courantes. La première méthode consiste à classer les réponses du système aux questions de l'utilisateur dans les catégories suivantes: réponses correctes, réponses partiellement correctes, réponses incorrectes ou réponses défaillantes. Les différentes catégories de réponses regroupées permettent d'obtenir les grandeurs qui ont été utilisées dans le programme américain DARPA (voir le Tableau 5). La seconde méthode consiste à classer les capacités d'analyse du système en fonction soit des énoncés correctement analysés, soit des paires AVP correctement identifiées. Sur la base des paires AVP identifiées, on peut calculer des grandeurs globales telles que la précision des concepts (CA, *concept accuracy*), le taux d'erreur de concept (CER, *concept error rate*), ou l'exactitude de la compréhension (UA, *understanding accuracy*). Tous ces paramètres sont énumérés dans le Tableau 5.

Tableau 5 – Paramètres d'interaction liés à l'entrée vocale

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'éva-luation
WER, WA	Taux d'erreur sur les mots, exactitude des mots	Pourcentage des mots qui ont été reconnus correctement, sur la base de la forme orthographique de l'énoncé hypothétique et de l'énoncé de référence (transcrit), et d'un alignement effectué à l'aide de l'algorithme "sclite" (voir la référence [18]). Soit n_w le nombre total de mots contenus dans tous les énoncés d'utilisateur d'un dialogue, et s_w, d_w et i_w le nombre de mots respectivement substitués, supprimés et insérés, alors le taux d'erreur sur les mots et l'exactitude des mots sont obtenus de la façon suivante: $WER = \frac{s_w + i_w + d_w}{n_w}$ $WA = 1 - \frac{s_w + i_w + d_w}{n_w} = 1 - WER$ Voir la référence [24]; on trouvera dans la référence [28] des détails sur la manière dont ces paramètres peuvent être calculés dans le cas d'une reconnaissance de mots isolés.	Mot	Instruments/ experts

Tableau 5 – Paramètres d'interaction liés à l'entrée vocale

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'éva-luation
SER, SA	Taux d'erreur sur les phrases, exactitude des phrases	Pourcentage des phrases entières qui ont été correctement identifiées. Soit n_s le nombre total de phrases et s_s, i_s et d_s le nombre de phrases respectivement remplacées, insérées et supprimées, alors: $SER = \frac{s_s + i_s + d_s}{n_s}$ $SA = 1 - \frac{s_s + i_s + d_s}{n_s} = 1 - SER$ [24]	Enoncé	Instruments/ experts
NES	Nombre d'erreurs par phrase	Nombre moyen d'erreurs de reconnaissance dans une phrase. Soit $s_w(k)$, $i_w(k)$ et $d_w(k)$ le nombre de mots respectivement remplacés, insérés et supprimés dans une phrase k, alors $NES(k) = s_w(k) + i_w(k) + d_w(k)$ Le nombre *NES* moyen peut être calculé comme suit: $NES = \frac{\sum_{k=1}^{\text{nbre de tours de parole de l'utilisateur}} NES(k)}{\text{nbre de tours de parole de l'utilisateur}} = \frac{WER \cdot \text{nbre de mots de l'utilisateur}}{\text{nbre de tours de parole de l'utilisateur}}$ [26]	Enoncé	Instruments/ experts
WES	Erreur sur les mots par phrase	En rapport avec le paramètre *NES*, mais normalisé au nombre de mots dans une phrase k, $w(k)$: $WES(k) = \frac{NES(k)}{w(k)}$ Le nombre WES moyen peut être calculé comme suit: $WES = \frac{\sum_{k=1}^{\text{nombre de tours de parole de l'utilisateur}} WES(k)}{\text{nombre de tours de parole de l'utilisateur}}$ [26]	Mot	Instruments/ experts
AN:CO, AN:IN, AN:PA, AN:FA, %AN:CO, %AN:IN, %AN:PA, %AN:FA	Nombre ou pourcentage de réponses correctes/ incorrectes/ partiel-lement correctes/ défaillantes du système	Nombre ou pourcentage total des questions posées par l'utilisateur, auxquelles les réponses données par le système, par dialogue sont: • correctes (*AN:CO*); • incorrectes (*AN:IC*); • partiellement correctes (*AN:PA*); • défaillantes (*AN:FA*). voir les références [21][12][14].	Enoncé	Experts

Tableau 5 – Paramètres d'interaction liés à l'entrée vocale

Abré-viation	Nom	Définition	Niveau d'inter-action	Méthode d'éva-luation
$DARPA_s$, $DARPA_{me}$	Note DARPA, erreur modifiée DARPA	Mesures selon l'initiative DARPA de compréhension de la parole, modifiées par Skowronek [25][17] afin de comptabiliser les réponses partiellement correctes: $DARPA_S = \frac{AN:CO - AN:IC}{nombre\ de\ questions\ de\ l'utilisateur}$ $DARPA_{me} = \frac{AN:FA + 2\cdot(AN:IC + AN:PA)}{nombre\ de\ questions\ de\ l'utilisateur}$ [21][12][25]	Enoncé	Experts
PA:CO, PA:PA, PA:IC, %PA:CO, %PA:PA, %PA:IC	Nombre d'énoncés de l'utilisateur analysés correc-tement/ en partie correc-tement/ incor-rectement	Evaluation du nombre des concepts (paires attribut-valeur, AVP) dans un énoncé, qui ont été extraits par le système: • *PA:CO*: tous les concepts d'un énoncé d'utilisateur ont été correctement compris par le système; • *PA:PA*: pas la totalité mais au moins un concept d'un énoncé d'utilisateur a été correctement compris par le système; • *PA:IC*: aucun concept d'un énoncé d'utilisateur n'a été correctement compris par le système. Cette évaluation s'exprime sous la forme du nombre ou du pourcentage total des énoncés d'utilisateur dans un dialogue, qui ont été analysés correctement/en partie correctement/incorrectement. [7]	Enoncé	Experts
CA, CER	Exactitude des concepts, taux d'erreur de concept	Pourcentage, par dialogue, d'unités sémantiques correctement comprises. Les concepts sont définis comme des paires attribut-valeur (AVP), n_{AVP} étant le nombre total de paires AVP et s_{AVP}, i_{AVP} et d_{AVP} le nombre de paires AVP respectivement remplacées, insérées et supprimées. On peut alors déterminer l'exactitude des concepts et le taux d'erreur sur les concepts de la façon suivante: $CA = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ $CER = \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ [9][24][3][2]	Enoncé	Experts
UA	Exactitude de compré-hension	Pourcentage des énoncés d'utilisateur dans lesquels toutes les unités sémantiques (AVP) ont été correctement extraites: $UA = \frac{PA:CO}{nombre\ de\ tours\ de\ parole\ de\ l'utilisateur}$ [32]	Enoncé	Experts

7.6 Autres paramètres

La majorité des paramètres d'interaction énumérés dans les tableaux décrivent le comportement du système, chose évidente car il s'agit de décrire la qualité du système et du service en question. Outre ces paramètres, il est possible de définir des paramètres liés à l'utilisateur qui sont propres à un groupe d'utilisateurs participant à une expérience mais qui peuvent être étroitement liés aux aspects qualité perçus par l'utilisateur.

Lorsque l'on distingue les différents aspects de la qualité d'un service employant un dialogueur automatique comme décrit dans le § 5.3/P.851, on constate que plusieurs aspects qualitatifs ne sont pas pris en compte par les paramètres d'interaction. En effet, aucun paramètre ne concerne directement la facilité d'emploi, la satisfaction de l'utilisateur, l'acceptabilité ou la qualité de sortie de la parole. Pour l'instant, très peu de méthodes abordent la qualité de sortie de la parole (qu'elle soit concaténée ou synthétisée) d'une manière paramétrique. Des mesures instrumentales liées à l'intelligibilité de la parole sont définies par exemple dans la norme CEI 60268-16 [15], mais elles ne s'appliquent pas à un environnement téléphonique. Des mesures du coût de concaténation, pouvant être calculées à partir du texte d'entrée et de la base de données vocale d'un système de concaténation, ont été proposées [5]. Même si elles peuvent présenter de fortes corrélations avec les notes moyennes d'opinion (MOS, *mean opinion score*) obtenues dans des expériences auditives, ces mesures sont propres au synthétiseur de parole et à ses corpus de concaténation.

8 Interprétation des valeurs des paramètres d'interaction

Même si les paramètres d'interaction, comme ceux qui sont définis dans le présent Supplément, sont importants pour la conception, l'optimisation et la maintenance d'un système, ils ne sont pas directement associés à la qualité perçue par l'utilisateur. En conséquence, l'ensemble des paramètres d'interaction devrait être complété par une série de jugements portés par l'utilisateur sur différents aspects relatifs à la qualité, tels qu'ils sont décrits dans la Rec. UIT-T P.851. C'est la seule manière d'obtenir des informations valables sur la qualité des services faisant appel aux dialogueurs automatiques.

Les valeurs des paramètres d'interaction peuvent être interprétées sur la base des résultats d'expériences, mais ces résultats sont souvent propres au système ou au service considéré. A titre d'exemple, une augmentation du nombre d'annonces d'expiration de temporisation peut indiquer que l'utilisateur ne sait pas quoi dire à certains points d'un dialogue ou qu'il ne comprend pas certaines actions du système [29]. Par ailleurs, une augmentation des tentatives d'intervention peut simplement indiquer que l'utilisateur a appris qu'il était possible d'interrompre le système. A l'inverse, une réduction du nombre de tentatives peut également indiquer que l'utilisateur ne sait pas quoi dire au système. De longs énoncés d'utilisateur peuvent être la conséquence de nombreuses interventions faites par l'utilisateur. Une diminution des valeurs des paramètres liés à la métacommunication (en particulier, celles liées à la métacommunication établie par l'utilisateur) devrait avoir pour effet d'augmenter la robustesse du système, la fluidité du dialogue et l'efficience de la communication [1].

BIBLIOGRAPHIE

[1] BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L.: *Designing interactive speech systems: From first ideas to user testing,* Springer, Berlin (Allemagne), 1998.

[2] BILLI, R., CASTAGNERI, G., DANIELI, M.: Field trial evaluations of two different information inquiry systems, *Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96)*, Basking Ridge NJ (Etat-Unis d'Amérique), pp. 129-134, 1996.

[3] BOROS, M., ECKERT, W., GALLWITZ, F., GORZ, G., HANRIEDER, G., NIEMANN, H.: Towards understanding spontaneous speech: Word accuracy vs. concept accuracy, *Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96)*, IEEE, Piscataway NJ (Etat-Unis d'Amérique), pp. 2, 1009-1012, 1996.

[4] CARLETTA, J.: Assessing agreement of classification tasks: The kappa statistics, *Computational Linguistics*, Vol. 22(2), pp. 249-254, 1996.

[5] CHU, M., PENG, H.: An objective measure for estimating MOS of synthesized speech, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavie*), Aalborg (*Danemark*), 3, pp. 2087-2090, 2001.

[6] COOKSON, S.: Final evaluation of VODIS – Voice operated data inquiry system, *Proc. of Speech'88, 7th FASE Symposium*, Edimbourg (Royaume-Uni), 4, pp. 1311-1320, 1988.

[7] DANIELI, M., GERBINO, E.: Metrics for evaluating dialogue strategies in a spoken language system, *Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAAI Symposium*, Stanford CA (Etat-Unis d'Amérique), AAAI Press, Menlo Park CA (Etat-Unis d'Amérique), pp. 34-39, 1995.

[8] FRASER, N.: Assessment of interactive systems, *Handbook on Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, Berlin (Allemagne), pp. 564-615, 1997.

[9] GERBINO, E., BAGGIA, P., CIARAMELLA, A., RULLENT, C.: Test and evaluation of a spoken dialogue system, *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP'93)*, IEEE, Piscataway NJ (Etat-Unis d'Amérique), 2, pp. 135-138, 1993.

[10] GIBBON, D., MOORE, R., WINSKI, R., Eds.: *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin (Allemagne), 2000.

[11] GLASS, J., POLIFRONI, J., SENEFF, S., ZUE, V.: Data collection and performance evaluation of spoken dialogue systems: The MIT experience, *Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000)*, Beijing (Chine), 4, pp. 1-4., 2000.

[12] GOODINE, D., HIRSCHMAN, L., POLIFRONI, J., SENEFF, S., ZUE, V.: Evaluating interactive spoken language systems, *Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'92)*, CA-Banff, 1, pp. 201-204, 1992.

[13] GRICE, H.P.: Logic and conversation, *Syntax and Semantics, Vol. 3: Speech Acts* (P. Cole and J.L. Morgan, eds.), Academic Press, New York NY (Etat-Unis d'Amérique), pp. 41-58, 1975.

[14] HIRSCHMAN, L., PAO, C.: The cost of errors in a spoken language system, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93)*, Berlin (Allemagne), 2, pp. 1419-1422, 1993.

[15] CEI 60268-16 (2003), *Equipements pour systèmes électroacoustiques – Partie 16: Evaluation objective de l'intelligibilité de la parole au moyen de l'indice de transmission de la parole*. Commission électrotechnique internationale, Genève (Suisse).

[16] KAMM, C.A., LITMAN, D.J., WALKER, M.A.: From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, Sydney (Australie), 4, pp. 1211-1214, 1998.

[17] MÖLLER, S.: *Quality of telephone-based spoken dialogue systems*. Springer, New York NY (Etat-Unis d'Amérique), 2005.

[18] NIST Speech Recognition Scoring Toolkit, *Speech recognition scoring toolkit,* National Institute of Standards and technology, http://www.nist.gov/speech/tools, Gaithersburg MD, (Etat-Unis d'Amérique) 2001.

[19] PICONE, J., DODDINGTON, G.R., PALLETT, D.S.: Phone-mediated word alignment for speech recognition evaluation, *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 38(3), pp. 559-562, 1990.

[20] PICONE, J., GOUDIE-MARSHALL, K.M., DODDINGTON, G.R., FISHER, W.: Automatic text alignment for speech system evaluation, *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 34(4), pp. 780-784, 1986.

[21] POLIFRONI, J., HIRSCHMAN, L., SENEFF, S., ZUE, V.: Experiments in evaluating interactive spoken language systems, *Proc. DARPA Speech and Natural Language Workshop*, Harriman CA (Etat-Unis d'Amérique), pp. 28-33, 1992.

[22] PRICE, P.J., HIRSCHMAN, L., SHRIBERG, E., WADE, E.: Subject-based evaluation measures for interactive spoken language systems, *Proc. DARPA Speech and Natural Language Workshop*, Harriman CA (Etat-Unis d'Amérique), pp. 34-39, 1992.

[23] SAN-SEGUNDO, R., MONTERO, J.M., COLÁS, J., GUTIÉRREZ, J., RAMOS, J.M., PARDO, J.M.; Methodology for dialogue design in telephone-based spoken dialogue systems: A Spanish train information system, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavie)*, Aalborg (Danemark), 3, pp. 2165-2168, 2001.

[24] SIMPSON, A., FRASER, N.M.: Black box and glass box evaluation of the SUNDIAL system, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93)*, Berlin (Allemagne), 2, pp. 1423-1426, 1993.

[25] SKOWRONEK, J.: *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität, Bochum (Allemagne), 2002.

[26] STRIK, H., CUCCHIARINI, C., KESSENS, J.M.: Comparing the performance of two CSRs: How to determine the significance level of the differences, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavie)*, Aalborg (Danemark), 3, pp. 2091-2094, 2001.

[27] STRIK, H., CUCCHIARINI, C., KESSENS, J.M.: Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test, *Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000)*, Beijing (Chine), 4, pp. 740-743, 2000.

[28] VAN LEEUWEN, D., STEENEKEN, H.: Assessment of recognition systems, *Handbook on Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, Berlin (Allemagne), pp. 381-407, 1997.

[29] 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 Conference Proc.*, Los Angeles CA (Etat-Unis d'Amérique), ACM, New York NY (Etat-Unis d'Amérique), pp. 582-589, 1998.

[30] WALKER, M.A., LITMAN, D.J., KAMM, C.A., ABELLA, A.: Evaluating spoken dialogue agents with PARADISE: Two case studies, *Computer Speech and Language*, Vol. 12(3), pp. 317-347, 1998.

[31] WALKER, M.A., LITMAN, D.J., KAMM, C.A., ABELLA, A.: PARADISE: A framework for evaluating spoken dialogue agents, *Proc. of the 35th Ann. Meeting of the Assoc. for Computational Linguistics*, Madrid (Espagne), pp. 271-280, 1997.

[32] ZUE, V., SENEFF, S., GLASS, J.R., POLIFRONI, J., PAO, C., HAZEN, T.J., HETHERINGTON, L.: JUPITER: A telephone-based conversational interface for weather information, *IEEE Trans. Speech and Audio Processing*, Vol. 8(1), pp. 85-96, 2000.

SÉRIES DES RECOMMANDATIONS UIT-T

Série A	Organisation du travail de l'UIT-T
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	Gestion des télécommunications y compris le RGT et maintenance des réseaux
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, communication entre systèmes ouverts et sécurité
Série Y	Infrastructure mondiale de l'information, protocole Internet et réseaux de prochaine génération
Série Z	Langages et aspects généraux logiciels des systèmes de télécommunication

Imprimé en Suisse
Genève, 2006