Evaluation de la qualité vocale dans les télécommunications

M. Guéguin[1,2,3]*, V. Barriac*[1]*, V. Gautier-Turbin*[1]*, R. Le Bouquin-Jeannès*[2,3]*, G. Faucon*[2,3]

[1]France Télécom R&D, TECH/SSTP/MOV, 22307 Lannion Cedex, France
[2]INSERM, U642, Laboratoire Traitement du Signal et de l'Image, Rennes, France
[3]Université de Rennes 1, LTSI, Campus de Beaulieu, 35042 Rennes Cedex, France
marie.gueguin@francetelecom.com

ABSTRACT

This paper is a review of the methods for speech quality assessment. Subjective methods involve human subjects testing systems in various network conditions and voting on an opinion scale. The scores obtained for each condition are averaged to get a mean opinion score (MOS). These subjective tests are the only way to assess perceived speech quality, but they are complex, cost- and time-consuming. Consequently objective methods have been introduced to predict the speech quality as perceived by users. Here, objective methods are classified depending on the context they deal with. This review of objective methods shows a lack of model in the conversational context. Then we propose an objective model of the conversational speech quality, built on a combination of objective models of the listening and talking speech qualities and the delay.

1. INTRODUCTION

Les systèmes de télécommunications sont en constante évolution depuis plusieurs années et nous avons assisté à l'émergence de nouveaux types de transmission, tels que les réseaux mobiles (GSM, bientôt UMTS) et les réseaux de type paquet (*Internet Protocol*, IP). Ces nouvelles technologies sont en pleine expansion du fait de la valeur ajoutée qu'elles apportent aux utilisateurs par rapport à la téléphonie classique, telle que par exemple : la mobilité, ou la possibilité de transmettre non seulement la voix, mais aussi des données et du contenu multimédia, ou encore le coût réduit des appels longue distance. Cependant, contrairement à la qualité de la voix transmise sur le réseau téléphonique commuté (RTC) relativement stable et prévisible, la qualité de service (*Quality of Service*, QoS) de ces nouvelles technologies est généralement non garantie. En effet, elles sont non seulement sujettes à la plupart des dégradations rencontrées avec le RTC (écho, délai, distorsion de l'effet local, bruits, etc.), mais encore introduisent de nouvelles dégradations (distorsion de la parole due au codage, délais augmentés par le traitement numérique), dont certaines sont non linéaires (délai variable appelé « gigue » et pertes de paquets dans les réseaux IP, bruits de fond non stationnaires dans les réseaux mobiles).

Afin de satisfaire leurs clients et de leur offrir la meilleure QoS possible, les opérateurs de télécommunications se doivent de contrôler la qualité perçue par les utilisateurs de leurs services, et doivent pour cela évaluer cette qualité. Les méthodes subjectives, faisant appel à des participants humains qui testent un système dans différentes conditions réelles d'utilisation définies par l'expérimentateur, restent la solution la plus fiable pour évaluer la qualité perçue par les utilisateurs. Bien que ces méthodes subjectives soient le seul moyen d'atteindre le jugement des utilisateurs, les opérateurs de télécommunications cherchent à éviter le recours à de telles méthodes, du fait du coût et du temps qu'elles demandent. Ces méthodes sont décrites dans la section 2.

Ainsi, des méthodes objectives plus poussées que les mesures objectives simples telles que le rapport signal-à-bruit (RSB) et l'erreur quadratique moyenne (EQM) ont été développées. Elles sont construites afin d'être corrélées avec les résultats de tests subjectifs et ainsi constituent un moyen de substitution aux méthodes subjectives. Ces méthodes objectives sont présentées dans la section 3.

Enfin, nous présentons dans la section 4 notre modèle objectif d'évaluation de la qualité vocale en contexte de conversation.

2. TESTS SUBJECTIFS

Lors d'un test subjectif, on demande à des participants de tester un système de télécommunications dans différentes conditions et de noter sur une échelle de qualité la qualité vocale de ce système. D'une manière générale, la qualité dépend de la personne qui la juge. Sa perception met en jeu l'expérience passée, les attentes et l'humeur de chacun. La qualité vocale, dans le cadre des systèmes de télécommunications, est elle aussi dépendante de celui qui l'évalue. Ainsi, les notes des participants pour une condition de test donnée sont moyennées pour obtenir la note moyenne d'opinion (*Mean Opinion Score*, MOS), qui permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale. De plus, la perception de la qualité vocale dépend du contexte et de l'environnement dans lesquels est placée la personne qui juge. En effet, si elle est simplement en train d'écouter un message vocal (contexte d'écoute) ou si elle est impliquée dans une conversation avec un interlocuteur (contexte de conversation), les processus d'attention mis en jeu ne sont pas les mêmes et le jugement de la qualité en est impacté. De même, l'environnement (bruit, informations visuelles ou sonores supplémentaires, etc.) influence le jugement de la qualité. Ainsi, les conditions à tester sont définies en fonction de l'objectif visé, le participant étant amené à évoluer dans un ou plusieurs contextes (écoute, locution et conversation).

2.1. Tests d'écoute

Le principe des tests d'écoute consiste à placer les participants en situation d'écoute et à leur diffuser des séquences audios correspondant la plupart du temps à dif-

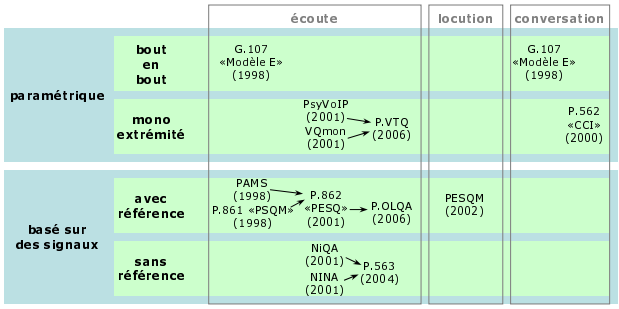

FIG. 1: Les modèles objectifs existants de la qualité vocale

férentes conditions de dégradation. Les conditions testées concernent les dégradations affectant la qualité d'écoute, comme la distorsion de la parole due au codage, le bruit pour l'auditeur et les pertes de paquets. La notation s'effectue selon l'une des méthodes définies par l'Union Internationale des Télécommunications (UIT) dans la Recommandation P.800 [19]. La plus utilisée est la méthode d'évaluation par catégories absolues (*Absolute Category Rating*, ACR) avec les catégories : 5 = Excellente, 4 = Bonne, 3 = Passable, 2 = Médiocre, 1 = Mauvaise. On peut également citer la méthode d'évaluation par catégories de dégradation (*Degradation Category Rating*, DCR) avec les catégories 5 = Dégradation inaudible, 4 = Dégradation audible mais pas gênante, 3 = Dégradation un peu gênante, 2 = Dégradation gênante, 1 = Dégradation très gênante. Plusieurs questions peuvent être posées aux participants permettant ainsi d'évaluer différentes dimensions de la qualité vocale, telles que la qualité globale, le naturel de la voix du locuteur et la dégradation due au bruit.

2.2. Tests de parole et d'écoute

Dans un test de parole et d'écoute, les participants sont placés dans le contexte de locution. Ils doivent donc parler dans le microphone du système à tester et écouter simultanément ce qui arrive du haut-parleur. Cela permet de tester des dégradations affectant la qualité de locution, telles que l'écho, la distorsion de l'effet local et le bruit pour le locuteur. De même que pour les tests d'écoute, les participants notent les conditions testées selon l'une des méthodes définies par les Recommandations P.800 [19] et P.831 [20]. Les questions posées concernent en général la qualité globale, la dégradation due à l'écho et la dégradation due au bruit.

2.3. Tests de conversation

Les tests de conversation sont conçus pour évaluer la qualité dans la situation la plus réaliste. Deux participants sont installés chacun dans une salle et dialoguent via un système de télécommunications. Les conditions testées dans ces tests concernent les dégradations des deux contextes précédents (écoute et locution) ainsi que les dégradations affectant spécifiquement l'interaction de la conversation, comme le délai, la gigue et la double parole. Les conditions testées peuvent être les mêmes pour les deux participants (test symétrique) ou différentes (test asymétrique). Le but étant de reproduire une communication téléphonique réaliste, des prétextes de conversation (sous la forme de dessin à décrire ou de jeu de rôle) sont généralement fournis aux participants. Ainsi, des scénarios de conversation ont été créés [9] ayant pour thèmes par exemple une commande de pizza ou un achat de billet d'avion. Chacun des participants note ensuite la qualité de la conversation qu'il vient d'expérimenter selon l'une des méthodes définies par les Recommandations P.800 [19] et P.831 [20]. Généralement lors de tels tests, il est demandé aux participants d'évaluer la qualité globale, la dégradation due à l'écho, la dégradation due au bruit et l'effort d'interruption. Les tests de conversation sont les plus coûteux et ne permettent pas d'étudier autant de conditions que les tests d'écoute, ils sont donc plus rares.

Quel que soit le type de test subjectif, de nombreuses précautions doivent être prises afin de contrôler les différentes sources de variabilité, telles que le choix des participants, le choix des conditions testées ou l'ordre de présentation des conditions, et afin d'obtenir des résultats fiables et exploitables. Ces tests sont donc fastidieux et coûteux à mettre en place. Les méthodes objectives se présentent comme une alternative aux méthodes subjectives et permettent d'automatiser l'évaluation de la qualité vocale. Cependant, elles doivent avoir une forte corrélation avec les résultats des tests subjectifs, qui représentent le jugement des utilisateurs. Qui dit modélisation objective, dit donc données subjectives pour « alimenter » le modèle.

3. MODÈLES OBJECTIFS

Les modèles objectifs peuvent être classés selon :
- le fait qu'ils se basent sur des mesures physiques du système (paramétriques) ou sur les signaux,
- le besoin qu'ils ont d'avoir accès aux informations des deux côtés du système (bout en bout ou avec référence) ou d'un seul côté seulement (mono-extrémité ou sans référence),
- le contexte dans lequel ils fonctionnent (écoute, locution ou conversation).

La Figure 1 classe les différents modèles existants en fonction de ces trois critères.

3.1. Modèles paramétriques

Les modèles paramétriques utilisent des mesures physiques du système à évaluer pour donner une note de qualité vocale.

Parmi les modèles paramétriques, le modèle E est le plus utilisé. Il a été développé comme un outil bout en bout pour les concepteurs de réseaux et normalisé en 1998 à l'UIT-T dans la Recommandation G.107 [15]. Il a été à la source de nombreux tests subjectifs, qui ont permis de l'optimiser. Le modèle E produit un facteur d'évaluation R compris entre 1 et 100, calculé à partir de mesures physiques des deux côtés du système à évaluer telles que le délai, l'écho, l'atténuation, le bruit de salle, etc. Il peut être utilisé pour estimer la qualité de conversation, et la qualité d'écoute sous réserve de fixer certains de ces paramètres à des valeurs par défaut. Cependant, ce modèle est connu pour donner des résultats faux dans certains cas.

L'équivalent du modèle E en mono-extrémité est le modèle appelé CCI (*Call Clarity Index*), décrit dans la Recommandation P.562 de l'UIT-T [17]. Il permet d'évaluer la qualité de conversation à partir de mesures du système (*e.g.* niveau de parole, niveau de bruit, atténuation de l'écho) effectuées par des sondes non-intrusives appelées

les INMDs (*In-service Non-intrusive Measurement Devices*), décrites dans la Recommandation P.561 de l'UIT-T [16]. Ce modèle permet d'interpréter les mesures faites par les INMDs pour prédire la qualité de conversation, telle que perçue par chaque utilisateur de la communication, en faisant des hypothèses sur le réseau et sur les utilisateurs de chaque extrémité.

Un autre modèle mono-extrémité de la qualité d'écoute, appelé provisoirement P.VTQ [2], est en cours de normalisation à l'UIT-T. Il fixe les objectifs de performances qui doivent être atteintes par des modèles tels que PsyVoIP de Psytechnics [12] et VQmon de Telchemy [5]. Le but de ce type de modèle est de se baser sur les informations des paquets IP sans utiliser les données vocales contenues dans le flot IP (longues à désencapsuler des paquets), afin d'être utilisé dans la surveillance en temps réel de la qualité des réseaux IP. Le modèle estime des paramètres de qualité intermédiaires (taux de perte de paquets, type de perte de paquets et gigue) à partir des informations contenues dans l'en-tête du Real-Time Protocol (RTP). La note de qualité d'écoute est ensuite estimée à partir de ces paramètres.

L'avantage des modèles paramétriques est leur rapidité, ils peuvent donc être facilement embarqués dans des éléments du réseau et les terminaux. Cependant, ils n'atteignent pas les mêmes performances que les modèles basés sur des signaux.

3.2. Modèles basés sur des signaux

Ces modèles, comme leur nom l'indique, utilisent les signaux de référence et dégradé (bout en bout ou avec référence) ou le signal dégradé seul (mono-extrémité ou sans référence) pour prédire la note de qualité vocale du système évalué.

Les modèles avec référence envoient un signal connu (référence) à travers le système à tester, capturent le signal après traversée du système (signal généralement appelé « signal dégradé »), et comparent ces deux signaux afin d'en déduire une note de qualité, qui doit être bien corrélée avec la note MOS.

Parmi les modèles avec référence, les plus utilisés sont les modèles basés sur une comparaison des transformations internes propres à l'oreille humaine, appelés modèles perceptuels. Cette méthode consiste à transformer la représentation physique d'un signal (mesurée en décibels, secondes, hertz) en une représentation psychoacoustique (mesurée en sones, secondes, barks) et est basée sur les travaux de Zwicker et Feldtkeller sur la psychoacoustique [23].

Les modèles perceptuels constituent dorénavant l'approche dominante depuis le développement de méthodes pour évaluer la qualité des signaux audio, qui a abouti à une norme de l'UIT-R [14] : PEAQ (*Perceptual Evaluation of Audio Quality*). Partant de cette norme dans le domaine audio, KPN a développé un outil similaire pour le domaine de la parole appelé PSQM (*Perceptual Speech Quality Measure*) [3], normalisé par l'UIT-T sous le nom P.861 [21]. Cependant, ces deux modèles ont été développés pour évaluer la qualité des codecs audio et vocaux, mais ne sont pas suffisants pour évaluer la qualité d'un système de télécommunications complet, impliquant de nombreuses autres dégradations. En parallèle, British Telecom a développé le modèle perceptuel appelé PAMS (*Perceptual Analysis Measurement System*) [7]. L'avantage de ce dernier est qu'il est plus robuste que PSQM aux délais variables rencontrés en VoIP. Les connaissances de ces deux modèles ont donc été mutualisées et ont permis de créer le modèle PESQ (*Perceptual Evaluation of Speech Quality*), normalisé à l'UIT-T sous le nom P.862 [22]. PESQ permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (perte de paquets, distorsion due au codage et bruit ambiant du côté émission), aboutissant à une corrélation proche de 0,935 avec les données subjectives. Une extension de PESQ au domaine acoustique (avec prise en compte des terminaux) et en bande élargie (de 50 à 7000 Hz, au lieu de 300 à 3400 Hz en bande étroite) est en cours d'étude à l'UIT-T sous le nom provisoire P.OLQA (*Objective Listening Quality Assessment*) [4].

Les méthodes mono-extrémité permettent l'analyse des signaux sans référence connue. L'équivalent de PESQ en mono-extrémité a été normalisé par l'UIT-T sous le nom P.563 [18] à partir des modèles NiQA (*Non-intrusive speech Quality Assessment*) de Psytechnics [13] et NINA (*Non Intrusive Network Assessment*) de SwissQual [8]. Le modèle P.563 permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (distorsion dur aux annuleurs d'écho ou aux systèmes de réduction de bruit, perte de paquets, distorsion due au codage, et bruit ambiant du côté émission), aboutissant à une corrélation proche de 0,89 avec les données subjectives. Le principe de ce modèle est de détecter les trames de parole dans le signal dégradé et d'en extraire un ensemble de paramètres permettant de faire une analyse du conduit vocal et du caractère non naturel de la voix, une analyse des bruits additionnels intenses, une analyse des interruptions, silences et écrêtage temporel. La note de qualité vocale finale est calculée en faisant une combinaison linéaire des différents résultats de l'évaluation de la qualité intermédiaire avec certaines caractéristiques additionnelles du signal.

Tous ces modèles fonctionnent dans le contexte d'écoute. Le modèle perceptuel PESQM (*Perceptual Echo and Sidetone Quality Measure*) [1] évalue la qualité dans le contexte de locution d'un système de communications potentiellement affecté par de l'écho et/ou une distorsion de l'effet local. Ce modèle fonctionne sur le même principe que le modèle PESQ en comparant un signal dégradé avec le signal de référence correspondant. Dans le contexte de locution, le signal de référence est le signal prononcé par le participant dans le microphone et le signal dégradé est le signal retourné par le système dans le haut-parleur du même participant, pouvant donc contenir de l'écho et/ou un effet local distordu. Ce modèle est à la fois un modèle avec référence puisqu'il compare le signal dégradé au signal de référence et un modèle mono-extrémité puisqu'il ne nécessite d'avoir accès aux informations que d'un seul côté du système.

4. MODÈLE OBJECTIF DANS LE CONTEXTE DE CONVERSATION

Comme le montre cet état de l'art des modèles objectifs de la qualité vocale, résumé dans la Figure 1, il n'existe pas encore de modèle non paramétrique de la qualité vocale dans le contexte de conversation. L'UIT-T s'intéresse à la modélisation de la qualité de conversation dans la

Question 20 du Study Group 12, en vue de la normalisation d'un modèle objectif nommé provisoirement P.CQO [10]. L'intérêt existant pour un tel modèle nous a donc conduit naturellement à nous intéresser à sa conception [6]. Lors d'une conversation, chaque utilisateur alterne entre les rôles d'auditeur et de locuteur [11]. La qualité vocale dans ce contexte est donc détériorée par les dégradations rencontrées dans le contexte d'écoute (codage, bruit de fond pour l'auditeur et pertes de paquets) et dans le contexte de locution (écho et distorsion de l'effet local), mais aussi par les dégradations inhérentes à la bidirectionnalité du contexte de conversation, telles que le délai ou la dégradation due au fait que les deux interlocuteurs parlent simultanément (double parole).

Partant de ce constat, notre modèle objectif va donc combiner ces trois composantes de la qualité vocale (écoute, locution et interaction) pour prédire la qualité vocale de conversation correspondante. Pour cela, la relation entre les trois composantes de qualité vocale et la qualité de conversation est déterminée sur le plan subjectif grâce aux notes subjectives correspondantes collectées lors de tests subjectifs. La qualité d'interaction est difficilement évaluable lors de tests subjectifs (il n'existe pas de méthodologie de test normalisée pour ce contexte). Elle est principalement dégradée par le délai. Notre modèle va donc considérer la valeur du délai comme un indicateur de la qualité vocale d'interaction, plutôt que la note subjective d'interaction.

Cette combinaison de trois composantes (qualité d'écoute, qualité de locution et délai) ne consiste pas en une simple juxtaposition (somme, moyenne, etc.). En effet, la qualité en contexte de conversation est plus ou moins influencée par chacune des trois composantes, en fonction des dégradations présentes dans la communication. Ainsi, quand seule une dégradation de la qualité d'écoute est présente, par exemple des pertes de paquets, la note de qualité de conversation sera essentiellement corrélée avec la note de qualité d'écoute, et ne dépendra pas (ou peu) de la note de qualité de locution et du délai. Notre modèle tient donc compte de cette influence du type de dégradation sur la combinaison des trois composantes en introduisant un système de décision, qui pondère l'influence des trois composantes sur la note de qualité de conversation. Ainsi, des tests subjectifs sont nécessaires pour déterminer, en fonction des dégradations, quelle est la relation entre la note de qualité de conversation et les notes de qualités d'écoute et de locution, et le délai.

Une fois déterminée sur le plan subjectif, la relation est transposée sur le plan objectif, en remplaçant les notes subjectives de qualité vocale d'écoute et de locution par des notes objectives fournies par des modèles objectifs, tels que PESQ et PESQM respectivement, et la valeur du délai par sa mesure objective. Le système de décision, déterminé grâce à des tests subjectifs, est alors piloté par les dégradations détectées sur le système de télécommunications évalué.

Appliqué à un test subjectif sur l'écho et le délai, notre modèle aboutit à des corrélations d'environ 0,94 entre les notes objectives et les notes subjectives correspondantes [6].

5. CONCLUSION

Ce papier présente un état de l'art des techniques subjectives et objectives d'évaluation de la qualité vocale. Les méthodes objectives sont classées en fonction de plusieurs critères, notamment celui du contexte dans lequel elles fonctionnent. Ce classement met en évidence le manque de modèles objectifs dans le contexte de conversation, qui est pourtant le contexte le plus courant pour les utilisateurs. Nous présentons donc notre modèle objectif de la qualité de conversation, bâti à partir d'une combinaison des modèles objectifs de la qualité d'écoute et de la qualité de locution et du délai.

RÉFÉRENCES

[1] R. Appel and J. G. Beerends. On the quality of hearing one's own voice. *J Audio Eng Soc*, 50(4) :237–248, 2002.

[2] V. Barriac. Recent standardisation work on non-intrusive evaluation of voice quality in IP environments. In *Proc. CFA/DAGA'04*, pages 53–54, 2004.

[3] J. G. Beerends and J. A. Stemerdink. A perceptual speech-quality measure based on a psychoacoustic sound representation. *J Audio Eng Soc*, 42(3) :115–123, 1994.

[4] J. Berger. Requirements for a new model for objective speech quality assessment P.OLQA. UIT-T COM 12-D.75, 2005.

[5] A. D. Clark. Modeling the effects of burst packet loss and recency on subjective voice quality. In *Proc. IPTEL'01 Workshop*, 2001.

[6] M. Guéguin, R. Le Bouquin-Jeannès, G. Faucon, and V. Barriac. Towards an objective model of the conversational speech quality. In *Proc. ICASSP'06*, 2006.

[7] M. Hollier, A. Rimell, and P. Gray. Verification and use of an auditory perceptual model for subjective analysis of telephone systems. UIT-T COM 12-D.035, 1998.

[8] P. Juric. Non-intrusive speech quality measurement. UIT-T COM 12-27, 2001.

[9] S. Möller. Development of scenarios for a short conversation test. UIT-T COM 12-35, 1997.

[10] J. Pomy. Proposed scope for P.CQO. UIT-T TD 27, 2005.

[11] D. L. Richards. *Telecommunication by Speech : The Transmission Performance of Telephone Networks*. Butterworths, London, 1973.

[12] A. Rix, S. Broom, and R. Reynolds. Non-intrusive monitoring of speech quality in voice over IP networks. UIT-T COM 12-D.49, 2001.

[13] A. Rix and P. Gray. NiQA - Non-intrusive speech Quality Assessment. UIT-T COM 12-D.48, 2001.

[14] Recommandation UIT-R BS.1387. Méthode de mesure objective de la qualité du son perçu, 1998.

[15] Recommandation UIT-T G.107. Le modèle E : modèle de calcul utilisé pour la planification de la transmission, 2003.

[16] Recommandation UIT-T P.561. Dispositif de mesure en service et sans intrusion - Mesures pour les services vocaux, 2002.

[17] Recommandation UIT-T P.562. Analyse et interprétation des mesures en service sans intrusion dans les services vocaux, 2004.

[18] Recommandation UIT-T P.563. Méthode mono-extrémité pour l'évaluation objective de la qualité vocale dans les applications de la téléphonie à bande étroite, 2004.

[19] Recommandation UIT-T P.800. Méthodes d'évaluation subjective de la qualité de transmission, 1996.

[20] Recommandation UIT-T P.831. Evaluation subjective de la qualité de fonctionnement des annuleurs d'écho de réseau, 1998.

[21] Recommandation UIT-T P.861 (supprimée). Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique (300-3400 Hz), 1998.

[22] Recommandation UIT-T P.862. Evaluation de la qualité vocale perçue : méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite, 2001.

[23] E. Zwicker and R. Feldtkeller. *Psychoacoustique : l'oreille récepteur d'information*. Masson, Paris, France, 1981.