Caractérisation Perceptuelle des Interactions Audiovisuelles: revue

R. Pastrana-Vidal[1] C. Colomes[1] J. Gicquel[1] H. Cherifi[2]

[1] France Telecom R&D, DIH/EQS/M@I

4, rue de Clos Courtel
35512 Cesson Sévigné -France

{ricardo.pastrana, catherine.colomes, jeancharles.gicquel}@francetelecom.fr

[2] LIRSIA

Université de Bourgogne, Faculté de Science Mirande
BP 400, 21011 Dijon cedex-France

cherifi@crid-u.bourgogne.fr

Résumé

La qualité d'un service de communication audiovisuel est souvent estimée en fonction de l'ensemble des paramètres techniques intervenant dans la chaîne de communication ou à partir des caractéristiques inhérentes au signal numérique (qualité objective). Néanmoins cette approche n'est pas suffisamment adaptée aux services de diffusion pour lesquels la perception de l'utilisateur est le centre d'intérêt. Ce document présente les études réalisées sur les interactions existantes entre l'audio et la vidéo ainsi que leurs contributions à la qualité globale audiovisuelle du point de vue de la perception humaine (qualité perçue). L'ensemble des auteurs arrive à une conclusion commune : la qualité globale perçue d'un signal audiovisuel est principalement déterminée par la qualité du signal vidéo. Cependant, dans le cas d'une transmission sur Internet à très bas débit, ce résultat pourrait être remis en cause.

Mots clefs

Multimédia, Audiovisuel, Qualité, Modélisation.

1 Introduction

Le terme multimédia évoque la combinaison de différentes formes d'information : texte, son, images, vidéo et graphiques. Dans les services de communication multimédia, chaque type d'information est souvent traité de façon indépendante sans considérer leurs interdépendances. Dans le cas des applications audiovisuelles, comme la vidéoconférence ou la TV numérique, les interactions entre l'information sonore et l'information visuelle sont très importantes. Il a été constaté, par exemple, que la perception de la parole est bimodale : une personne écoute la voix en observant les organes articulatoires et les expressions du visage; nous nous appuyons sur la lecture des lèvres lorsque l'environnement est fortement bruité.

La qualité d'un service audiovisuel est souvent estimée en fonction de l'ensemble des paramètres techniques intervenant dans la chaîne de communication ou à partir des caractéristiques inhérentes au signal numérique (qualité objective). Néanmoins cette approche n'est pas suffisamment adaptée aux services de diffusion pour lesquels la perception de l'utilisateur est le centre d'intérêt. La plupart du temps cet élément est négligé. Dans cet esprit, il est inutile de quantifier les défauts que l'utilisateur final ne percevra pas. Il est donc indispensable d'étudier les interactions existantes entre l'audio et la vidéo ainsi que leurs contributions à la qualité globale audiovisuelle du point de vue de la perception humaine (qualité perçue).

La qualité perçue par l'utilisateur est estimée grâce à la note moyenne de jugement (MOS[1]) provenant de tests subjectifs. Un ensemble d'observateurs évalue la qualité en utilisant une échelle d'appréciation subjective (par exemple: excellente, bonne, assez bonne, mauvaise, médiocre) liée à une échelle de valeurs (0 à 5, 0 à 9 ou 0 à 100). L'évaluation peut se réaliser de façon instantanée [1] pendant l'observation et l'écoute de la séquence (évaluation en continu) ou de façon globale une fois la séquence finie (jugement en absolu) [2]. Les tests sont faits en respectant des protocoles expérimentaux établis par des groupes de normalisation comme l'Union Internationale des Télécommunications[2]. Dans les

[1] Mean Opinion Score
[2] www.itu.int

recommandations sont décrites les méthodes d'évaluation subjective: les échelles d'appréciation et de valeur, le type de jugement et la structure de la séance. Nous trouvons aussi les caractéristiques des séquences à utiliser: durée, type de contenu, nombre de séquences et la liste des conditions à examiner. Un autre élément à respecter est l'environnement normalisé de test qui spécifie les conditions d'observation (l'éclairement de la salle, les distances entre l'observateur et les dispositifs reproducteurs, etc.). La procédure pour le traitement des données (le rejet d'opinion, l'analyse statistique et la présentation de résultats) est aussi spécifiée.

Nous présentons une étude comparative des travaux existants sur la caractérisation perceptuelle des interactions audiovisuelles. Tout d'abord, la section 2 traite de l'influence de l'audio sur la vidéo et vice versa. Ensuite, nous discutons de la prépondérance de chaque média par rapport à la qualité globale audiovisuelle. Puis, un ensemble de modèles de prédiction de la qualité audiovisuelle est présenté. Enfin, nous présentons le résultat d'un test de préférence audiovisuelle dans le cas d'une transmission par Internet à bas débit.

Le tableau 1 montre le cadre d'expérimentation des six travaux représentatifs trouvés dans la littérature.

Auteur	Contenu			Format (vidéo)	Méthode subjective
	Vidéo	Audio	Durée		
[3,4]	Bandes annonces	Musique, voix	25 secs	720x576	Absolu
[5]	Vidéoconférence	Voix	10 secs	352x288, 176x144	Absolu
[6]	Journaux, sports, reportage	Voix, audio mono	14 min	720x576	Continu
[7]	Réalité virtuelle	Voix	10 secs	720x576	Absolu
[8]	Gens, graphics	Voix	10 secs	352x288, 176x144	Absolu

Tableau 1 – Etudes sur les interactions audiovisuelles.

2 Interactions audio - vidéo

Dans son article, [3] montre qu'il existe une influence mutuelle significative entre l'audio et la vidéo dans le cadre de la TV numérique. Les auteurs ont constaté que la qualité objective vidéo (QOV) a une contribution significative sur la qualité perçue de l'audio (MOSA), selon une proportion d'environ 13%. En revanche, l'influence de l'audio sur la qualité perçue vidéo (MOSV) est pratiquement négligeable (environ 2%).

Pour sa part [5] a mené son étude dans deux contextes de vidéoconférence : passif (la personne ne fait que regarder et écouter) et interactif (la personne participe à la conversation). Il a constaté que la qualité objective de la vidéo a un effet sur la MOSA. Dans un contexte passif l'influence est faible (mais non négligeable) tandis que dans un contexte interactif elle est très importante. Ce dernier résultat coïncide avec le travail de [9] qui a démontré que la perception de la parole peut être modifiée par l'information visuelle provenant du mouvement des lèvres[3]. En revanche, [5] affirme aussi que la perception de la vidéo est indépendante de l'audio dans les deux contextes. Il faut souligner que dans son cas, le signal audio employé est un signal de parole, possédant une qualité supérieure au seuil d'intelligibilité. Si la qualité de la parole n'atteint pas ce seuil, les résultats obtenus pourraient changer.

L'étude faite par [7] discute de l'influence de la vidéo sur la perception de l'audio. En utilisant comme contenu une séquence de réalité virtuelle (survol de bâtiments), en format TV numérique, et des commentaires avec des voix d'hommes et de femmes, il a été trouvé que la qualité perçue de l'audio diminue lorsque la vidéo se dégrade. Cette influence est similaire à celle qu'on trouve dans [3] et [5].

Dans un contexte de TV numérique, les auteurs de [6] ont mis en évidence des relations entre l'audio et la vidéo différentes de celles trouvées par [5]. Ils affirment que la QOV n'a aucune influence sur la qualité perçue de l'audio. En revanche, il semble que la qualité de l'audio a une influence sur la perception des dégradations de la vidéo.

Auteur	Interaction entre médias	
	A->V	V->A
[3,4]	faible	Forte
[5]	Négligeable	Forte
[6]	Forte	Négligeable
[7]	Non traité	Forte
[8]	faible	Forte

Tableau 2 – *Influence entre médias*

Le tableau 2 résume les figures d'interactions entre médias: l'audio sur la vidéo (A->V) et la vidéo sur l'audio (V->A).

[3] Pour percevoir la parole, un observateur écoute le signal acoustique en regardant les organes articulatoires et les expressions du visage.

3 Prépondérance des médias

En ce qui concerne l'influence de la qualité d'un media sur la qualité globale perçue audiovisuelle (MOSAV), le travail de [3] signale que la composante vidéo joue un rôle prépondérant tandis que l'audio prend une place secondaire. Cette constatation est partagée par [5] qui conclut que la perception Audiovisuelle est fortement dépendante de la qualité objective de la vidéo. Les variations de qualité objective de l'audio ont un effet faible sur la MOSAV, effet toutefois non négligeable. Par la suite [6] arrive aux même conclusions bien que son étude ait été faite avec des séquences de longue durée.

En utilisant comme contenu une séquence de réalité virtuelle [7] arrive à la même conclusion, quant à l'influence de la QOV sur la note MOSA, que [3,5,6]. L'effet contraire n'a pas été étudié.

Auteur	Influence de A et V sur AV	
	A -> AV	V -> AV
[3,4]	faible	Forte
[5]	faible	Forte
[6]	faible	Forte
[7]	Non traité	Forte
[8]	faible	Forte

Tableau 3 – *Influences sur la perception audiovisuelle*

Le tableau 3 résume les figures d'interactions de l'audio sur l'audiovisuel (A->AV) et de la vidéo sur l'audiovisuel (V->AV).

4 Modèles de prédiction

Dans la littérature, chacun s'accorde sur le fait que la qualité globale audiovisuelle peut être prédite à partir de la note de qualité perçue audio (test audio uniquement) et de la qualité perçue vidéo (test vidéo uniquement). Les auteurs de [3] proposent un modèle non linéaire pour estimer la note MOS audiovisuelle,

$$MOSAV = 1.12 + 0.007\,MOSA + 0.24\,MOSV + 0{,}88 * MOSV * MOSA\,.$$

Le dernier terme de l'équation correspond à l'interaction audio/vidéo. La corrélation entre la note prédite avec ce modèle et la note observée est de 0.98. Il faut noter que le poids attribué à la note MOS Audio est extrêmement faible. Les auteurs proposent un autre modèle qui atteint 0.97 de corrélation entre la note MOSAV subjective et celle prédite, modèle utilisant uniquement le terme représentant l'interaction audiovisuelle,

$$MOSAV = 1.45 + 0.11 * MOSA * MOSV\,. \qquad 4.1$$

Cette relation est très proche du modèle obtenu par [8] où les séquences ont été dégradées par différents types de codages à bas débit,

$$MOSAV = 1.51 + 0.12 * MOSA * MOSV\,. \qquad 4.2$$

Dans [5] on trouve deux modèles de prédiction de la qualité perçue d'un signal audiovisuel dans un contexte passif. Le modèle non linéaire a une corrélation de 0.96 avec les résultats subjectifs,

$$MOSAV = -0{,}10 + 0{,}21 * MOSV + 0{,}12 * MOSA.$$

Un autre modèle non linéaire a été testé. Il réalise le produit pondéré des notes MOSA et MOSV,

$$MOSAV = 1{,}76 + 0{,}10 * MOSA * MOSV.$$

Dans ce cas, la corrélation entre les notes estimées et les résultats subjectifs est inférieure au premier modèle mais il est intéressant de noter que ce modèle ressemble aux modèles trouvés dans la littérature bien que les types de dégradations utilisées dans les tests soient très différents :

$MOSAV = 1.30 + 0.11 * MOSA * MOSV$ [10],

$MOSAV = 1.07 + 0.11 * MOSA * MOSV$ [11].

De plus, la relation est aussi conforme avec [4] et [8] donnés par les équations (4.1) et (4.2) respectivement. Malgré des conditions expérimentales très différentes, ces modèles mettent en évidence la nécessité de prendre en compte les interactions audio/vidéo dans l'évaluation globale de la qualité audiovisuelle.

5 Qualité Audiovisuelle sur Internet

Dans la littérature nous n'avons pas trouvé pour l'instant de tests d'évaluation subjective de la qualité au sens des interactions audiovisuelles prenant en compte les effets d'une transmission Internet. Ce type de communication par paquets introduit de nouveaux éléments de dégradation sur le signal audiovisuel et sur la perception de celui-ci. Les relations audiovisuelles auparavant citées peuvent donc changer lors d'une transmission par paquets à service non garanti. Par exemple, il est bien connu que dans IP[4] on peut avoir des perte de paquets en rafale, du délais et des variations de délais considérables. Les effets sur la qualité perçue peuvent être très importants. A ce sujet, un test de préférence audiovisuelle, mené au sein de

[4] Internet Protocol

notre équipe de recherche [12], nous a montré que l'audio est privilégiée par rapport à la vidéo dans le cas d'une transmission sur IP à très bas débit. Ce résultat remet en cause la prépondérance de la vidéo sur l'audio dans l'évaluation de la qualité globale audiovisuelle, généralement trouvée dans la littérature. Cette problématique constitue l'un de nos axes de recherche actuel.

6 Conclusions

En ce qui concerne les interactions entre l'audio et la vidéo, les conclusions sont hétérogènes et quelques-unes contradictoires. Ceci est dû à la diversité des conditions d'expérimentation et des méthodes de validation. Cependant, on peut dire que les interactions varient en fonction du contexte d'application et dépendent du type de contenus audio et vidéo.

Malgré la diversité des types de contenus audiovisuels, des durées des séquences, des débits audio et vidéo, des types de dégradations introduits, et des formats d'image, l'ensemble des auteurs arrive à une conclusion commune : La qualité perçue d'un signal audiovisuel est principalement déterminée par la qualité objective de la vidéo.

D'autre part, on note que la qualité globale audiovisuelle peut être prédite à partir de la note de qualité de chaque média. Ainsi, dans les différentes études analysées, un modèle commun combinant ces deux notes est proposé. Le modèle calcule avec une certaine précision la note de qualité globale en tenant compte des interactions audiovisuelles (produit des notes audio et vidéo).

Enfin, dans le cas d'une transmission par Internet à bas débit, les interactions audiovisuelles doivent être étudiées relativement au type de contenu et au contexte d'application.

Références

[1] ITU-R Recommendation BT.500-10. Methodology for the Subjective Assessment of the Quality of Television Pictures, 2000.

[2] ITU-T Recommendation P.910. Subjective Video Quality Assessment Methods for Multimedia Applications, Novembre 1999.

[3] J. Beerends et F. de Caluwe. The Influence of Video Quality on Perceived Audio Quality and Vice Versa, *Journal-Audio Engineering Society*, 5(47): 355-362, Mai 1999.

[4] ITU-T SG12 COM12-19-E (KPN). Relations between Audio, Video and Audiovisual Quality, Décembre 1997.

[5] N. Château. Study Of The Influence Of Experimental Context On The Relationships Between Audio, Video, and Audiovisual Subjective Qualities, *ITU-T SG12 COM 12 (CNET/France Telecom)*, Novembre 1998.

[6] A. Joly et N. Montard et Marcel Buttin. Audio-Visual Quality and Interactions between Television Audio et Video. *International Symposium on Signal Processing and its Applications*, pages: 438-441, Août 2001.

[7] M. Hollier et R. Voelcker. Objective Performance Assessment: Video Quality as an influence on Audio Perception, Dans *the 103rd Audio Engineering Society (AES) Convention*, preprint: 4590 (L-10), New York, September 1997.

[8] ITU-T SG12 COM12 D.038 (NTIA/ITS). Results of an Audiovisual Desktop Video Teleconferencing Subjective Experiment, Février 1998.

[9] H. McGurk et J. MacDonald. Hearing Lips and Seeing Voices, *Nature*, (264): 746-748, Décembre 1976.

[10] ITU-T SG12 COM12-20 (BELLCORE). Experimental Combined Audio/Video Subjectif Tests Methode, Décembre 1993.

[11] ITU-T SG12 COM12-37 (BELLCORE). Extension of Combined Audio/Video Quality Model, Septembre 1994.

[12] J. Blin. *UMTS Optimisation et évaluation de la qualité audiovisuelle perçue pour l'accès à des contenus vidéo sur mobile*, rapport interne France Telecom R&D, Décembre 2001.