Mémoire pour le Master Sciences et Technologies de l'UPMC
Mention Sciences de l'Ingénieur
Spécialité MIS, Parcours ATIAM

Virginie DURIN

Evaluation indirecte de la qualité vocale perçue

Maîtres de stage : Laetitia GROS, Nœl CHATEAU
France Télécom R&D TECH/QVP/USE
Technopole Anticipa
2 avenue P. Marzin
22307 Lannion Cedex
laetitia.gros@francetelecom.com
noel.chateau@francetelecom.com

12 Septembre 2005

Remerciements

Le mémoire et le stage tendent à la leur fin et il est tant de remercier tous ceux qui ont rendu ce stage possible, et par la suite agréable à vivre.

Je souhaite alors exprimer toute ma gratitude à Laetitia qui m'a suivie tout au long de ce stage, qui a su m'accorder une part de son temps pour m'encourager quand il le fallait et me guider dans mon travail avec clairvoyance et toujours avec bonne humeur (malgré quelques péripéties !) Un grand merci pour cette disponibilité, ces conseils et remarques, et cette gentillesse ; ces attentions, je les ai profondément appréciées.

Je pense bien sûr à Noël qui m'a accueillie au sein de son équipe et a participé à mon encadrement malgré un emploi du temps chargé. Des réunions, j'ai recueilli des connaissances mais aussi de l'enthousiasme, de l'énergie, une motivation supplémentaire pour avancer. Je le remercie aussi de m'avoir fait découvrir la psychoacoustique il y a deux ans en MST Image et Son.

Cette reconnaissance pour avoir découvert et confirmé mon goût pour la psychoacoustique, je la dois à Laetitia mais aussi et surtout à Daniel PRESSNITZER, Michèle CASTELLENGO, Barbara TILLMANN, mes "professeurs " qui m'ont convaincue par la passion qui les a animés lorsqu'il s'agissait de transmettre leur savoir. Ils m'ont étonnée, appris, fait rire et surtout captivée.

J'ai une pensée particulière pour Gilles, qui a subi mes diverses questions en Matlab et m'a apporté des réponses précieuses, qui m'ont permises de progresser plus facilement et avec davantage de plaisir. J'ai aussi une attention particulière pour Martine qui m'a aidé, elle aussi, dans mon travail.

Un gros merci à toutes les personnes qui ont accepté de prendre un peu sur leur temps pour passer le test (et qui ont cru que j'allais les torturer avec mes capteurs !), sans quoi le travail présent n'aurait plus de raison d'être. J'ai été particulièrement sensible à l'intérêt qu'elles y ont porté.

Un merci chaleureux à Gildas, Virginie, Gaëtan, Mélanie, Véronique, Antoine qui m'ont accompagné quotidiennement tout au long de ses cinq mois de stage, qui ont partagé les pauses café avec moi et su égayer l'atmosphère du bureau quand elle devenait trop studieuse !

Enfin, il a évidemment mes proches, qui ont largement contribué à ce bien être.

Grâce à toutes ces personnes, une douce alchimie s'est installée ; ce stage fut un plaisir. J'espère leur avoir procuré autant de joie de vivre qu'elles m'en ont donnée.

Table des matières

Table des figures

Liste des tableaux

Introduction

La téléphonie mobile et via Internet, la visiophonie, la visioconférence sont autant de témoins de la pluralité des services et moyens de télécommunications qui s'offrent à nous aujourd'hui. Puisque les sons qu'ils véhiculent sont toujours destinés à des auditeurs, il est nécessaire de disposer d'outils et méthodes pour évaluer la qualité vocale et/ou audio perçue par ces auditeurs. Une des missions du laboratoire qui m'a accueillie est de développer des méthodes d'évaluation de la qualité vocale/audio, adaptées aux nouvelles problématiques apparaissant avec les nouveaux services (multi modalité, nouveaux types de dégradations de la qualité, nouveaux contextes d'usages, etc.).

Actuellement, par qualité vocale, nous entendons donc la qualité de la parole transmise par un système de communication, qui comprend si l'on se place du point de vue de l'utilisateur, la sonie (l'intensité perçue), l'intelligibilité du signal restitué (ou degré de clarté avec lequel est perçu le signal de parole) mais aussi l'agrément, le confort d'écoute, la fidélité de la voix, le naturel de la voix, etc. Cependant, on peut considérer que les contraintes d'intensité sont aujourd'hui satisfaites (ne serait-ce parce que tout utilisateur peut régler le niveau de son téléphone) et l'on ne tiendra pas compte de ce facteur. De plus, les contraintes d'intelligibilité sont généralement aussi satisfaites, sauf parfois avec la téléphonie mobile lorsque le nombre de coupures et pertes de paquets d'informations deviennent trop importants. Nous appelons donc qualité vocale la satisfaction de l'utilisateur quant au confort d'écoute, à la fidélité et au naturel de la voix, etc., par rapport à un ensemble de référence propre à chaque individu. Avec les méthodes actuelles de mesure, la simplification suivante est faite : la qualité est considérée comme un phénomène unidimensionnel qui amalgame différents critères, et qui peut être caractérisée par une note sur une échelle linéaire.

Parmi les méthodes d'évaluation de la qualité vocale, on en distingue deux types, les méthodes dites "objectives" basées sur l'utilisation d'instruments de mesure captant le signal dans, ou en sortie de réseau, pour l'analyser et prédire une note de qualité. Cependant, le développement et la validation de ces outils sont fortement basés sur la mesure dite "subjective" faisant intervenir des groupes d'auditeurs qui évaluent la qualité vocale des séquences entendues lors de tests d'écoute ou des communications effectuées lors de tests de conversation.

L'un des enjeux que constitue ce stage est d'explorer une nouvelle voie pour l'évaluation de la qualité vocale, que nous appellerons évaluation indirecte. Notre perspective de recherche consiste en effet à explorer la possibilité de mesurer indirectement la qualité vocale, non plus à travers le jugement conscient des sujets mais à travers des indicateurs comportementaux et des indicateurs électrophysiologiques et cela en plaçant le sujet dans une situation adéquate, comme une situation de double tâche. La situation de double tâche, par exemple la

réalisation d'une tâche de communication associée à une tâche annexe perturbatrice est en effet appropriée puisqu'elle place le sujet dans une situation de surcharge cognitive et est ainsi propice à des réactions émotionnelles plus intenses et à des performances moindres. A terme, nous espérons que cette étude mènera à la découverte de marqueurs comportementaux et / ou physiologiques de la qualité vocale qui pourront compléter les méthodes subjectives actuelles.

Chapitre 1

Les méthodes subjectives d'évaluation de la qualité vocale et leurs limites

Comme l'annonce l'introduction, les outils à partir desquels se détermine la qualité vocale sont essentiellement basés sur la mesure dite "subjective". C'est ce que ce chapitre s'attache donc à décrire. Nous nous restreignons aux méthodes dites "normalisées" dans la mesure où ce sont celles qu'on emploie aujourd'hui le plus fréquemment dans les laboratoires de tests. Elles font intervenir des groupes d'auditeurs qui évaluent la qualité vocale des séquences entendues lors de tests d'écoute ou des communications effectuées lors de tests de conversation. L'UIT, Union Internationale des Télécommunications, présente dans la recommandation P.800 [UIT-T P.800, 1996] les méthodes normalisées d'évaluation subjective de la qualité de transmission.

C'est la nature même des dégradations qui influence principalement le choix d'une méthode. On distingue trois types de dégradations :

- Type I : les dégradations qui entraînent une augmentation de la difficulté de compréhension telles un affaiblissement, un bruit, une distorsion par exemple, quand la communication est unidirectionnelle, *i.e.* en situation d'écoute.
- Type II : les dégradations qui entraînent une difficulté à parler comme l'écho pour celui qui parle.
- Type III : celles qui entraînent une difficulté à converser, par exemple, un écho ou un retard.

Selon la nature des dégradations, on envisagera plutôt des essais d'opinion de conversation pour les types II et III ou plutôt des essais d'opinion d'écoute pour le type I.

1.1 Essais d'opinion de conversations

Les essais de conversation en laboratoire visent à reproduire dans la mesure du possible les conditions réelles de service perçues par les usagers du téléphone.

Deux sujets sont mis en situation de conversation. Pour cela, on fournit des prétextes de conversations qui nourrissent le dialogue, par exemple des scénarios de la vie courante à savoir par exemple, la réservation d'une chambre d'hôtel. Au terme de la conversation, soit environ 5

minutes plus tard pour que le système soit correctement évalué, les sujets notent les différents aspects de la conversations au moyen d' échelle d'appréciation subjective.

1.1.1 Echelle d'appréciation subjective de conversation

Cette première échelle se présente sous la forme de cinq catégories auxquelles sont associées des notes. Elle permet au sujet de noter la qualité de la liaison.

Quelle est votre opinion au sujet de la connexion que vous venez d'utiliser :

Excellente
Bonne
Passable
Médiocre
Mauvaise

L'opérateur affecte les valeurs suivantes aux catégories : Excellente = 5 ; Bonne = 4 ; Passable = 3 ; Médiocre = 2 ; Mauvaise = 1.

La moyenne arithmétique de ces notes d'opinion s'appelle note moyenne d'opinions sur la conversation ; elle est représentée par le symbole MOS_C.

1.1.2 Echelle de difficulté

Comme son nom l'indique, le sujet va juger de la difficulté à parler ou écouter sur une échelle binaire :

Avez-vous, ou votre partenaire, éprouvé des difficultés pour parler ou écouter dans cette connexion ?

Oui
Non

L'opérateur affecte les valeurs suivantes à la réponse :
Oui = 1, Non = 0.

La quantité évaluée (pourcentage de réponses "oui") s'appelle pourcentage de difficulté ; elle est désignée par le symbole %D. La proportion simple correspondante est désignée par le symbole d ; autrement dit, %D = 100d.

1.1.3 Autres échelles

D'autres échelles d'appréciation subjective peuvent convenir. Par exemple l'expérimentateur peut présenter une série numérique de catégories 1, 2, 3, 4, 5 avec les descriptions jointes

seulement à la première et à la dernière pour identifier la dimension subjective. Une autre possibilité envisageable est une échelle graduée de 1 à 10 voire 100 au lieu de 5 ; ou encore il peut proposer une simple ligne droite donnée à partir de laquelle le sujet définit une longueur proportionnelle à une caractéristique, par exemple, la qualité. Pour compléter l'analyse de la qualité vocale, la recommandation [UIT-T P.831, 1998] propose les quatre échelles suivantes :

Comment jugeriez vous la qualité de la communication ?

Inacceptable
Acceptable

Comment jugeriez vous la dégradation due à l'écho de votre propre voix ?

Imperceptible
Perceptible, mais non gênante
Légèrement gênante
Gênante
Très gênante

Comment jugeriez vous les autres dégradations (troncature, bruits divers,...) ?

Imperceptibles
Perceptibles, mais non gênantes
Légèrement gênantes
Gênantes
Très gênantes

Comment trouvez vous la voix de votre interlocuteur ?

Pas naturelle
...
Naturelle

Quelle note d'appréciation donneriez-vous pour la qualité de la communication que vous venez d'utiliser ?

Excellente
Bonne
Passable
Médiocre
Mauvaise

La recommandation [UIT-T P.832, 2000] propose les deux questions suivantes :

Quelle note d'appréciation donneriez-vous pour la capacité de dialoguer ? Ou : ***Quelle note d'appréciation donneriez-vous pour votre aptitude à tenir une conversation avec votre correspondant ?***

Excellente
Bonne
Passable
Médiocre
Mauvaise

Quelle note d'appréciation donneriez-vous pour la qualité sonore de la voix des correspondants ?

Excellente
Bonne
Passable
Médiocre
Mauvaise

1.1.4 Avantages et inconvénients des essais d'opinion de conversation

Ils présentent l'avantage de constituer le seul moyen d'évaluer de manière réaliste l'effet subjectif combiné de tous les paramètres ayant une incidence sur la qualité de conversation. Des effets tels que les variations de niveau, l'écho et la parole simultanée, peuvent avoir une incidence sensible sur la qualité de fonctionnement des téléphones mains-libres.
En revanche, ils sont souvent longs à réaliser. En règle générale, on ne peut évaluer qu'un ensemble limité de paramètres. Les conditions qu'il est possible d'évaluer de manière réaliste au cours d'un essai sont elles aussi limitées en nombre, en raison du temps que nécessitent les conversations types. Sans parler de la complexité de réalisation initiale du montage d'essai. On réalise donc plus fréquemment des essais d'opinions d'écoutes qui sont moins lourds à mettre en place.

1.2 Essais d'opinion d'écoutes

Dans ces tests, le sujet écoute des séries de phrases courtes, simples et claires de 2 à 3 secondes chacune, sans relation évidente de sens entre elles (par exemple, "Nos parents sont nos tuteurs naturels.", "Ce moyeu de roues grinçait continuellement.", etc.) Il doit ensuite les évaluer, soit individuellement, soit par paires sur les échelles décrites ci-après.

Les essais d'opinion d'écoute ne parviennent sans doute pas au même niveau de réalisme que les essais de conversation, mais ils trouvent des applications directes pour l'évaluation de systèmes de transmission physique qui sont essentiellement unidirectionnels, qu'il s'agisse par exemple de circuits de diffusion, de systèmes d'annonce en direct ou d'annonces enregistrées qui peuvent faire l'objet d'affaiblissement, de bruit et de distorsion (soit des dégradations de type I).
La méthode d'essai recommandée pour des essais d'écoute est celle de l'évaluation par catégories absolues (ACR, Absolute Category Rating) décrite ci-dessus. Elle est conforme à la méthode des jugements par catégories recommandée pour les essais de conversation.

1.2.1 Méthode ACR (Absolute Category Rating)

De la même manière que pour les jugements de conversations, on a pour les jugements d'écoutes les échelles suivantes :

- Echelle de qualité d'écoute

Qualité de la parole	***Note***
Excellente	5
Bonne	4
Passable	3
Médiocre	2
Mauvaise	1

La quantité évaluée d'après les notes est la note moyenne d'appréciation ou MOS.

- Echelle des efforts d'écoute

Effort nécessaire pour comprendre le sens des phrases	***Note***
Détente absolue ; aucun effort	5
Attention nécessaire, pas d'effort appréciable	4
Effort modéré	3
Effort considérable	2
Incompréhensible en dépit de tous les efforts possibles	1

La quantité évaluée d'après les notes est la note moyenne d'appréciation d'effort d'écoute ou MOS_{LE}.

- Echelle de niveau sonore préféré

Niveau sonore préféré	***Note***
Beaucoup plus fort que préféré	5
Plus fort que préféré	4
Selon préférence	3
Plus faible que préféré	2
Bien plus faible que préféré	1

La quantité évaluée d'après les notes est la note moyenne d'appréciation de niveau sonore préféré ou MOS_{LP}.

Ces trois échelles de qualité, d'effort et de niveau sonore rendent compte respectivement des trois critères essentiels de la qualité : l'agrément, l'intelligibilité et la sonie. Aujourd'hui, la méthode ACR ne représente plus que le critère "qualité d'écoute" pour les raisons données dans l'introduction.
Contrairement aux essais de conversation qui nécessitent une post annotation, toutes les échelles sont ici pré numérotées, ce qui présente l'avantage d'induire, par le biais de ces chiffres équidistants, une échelle d'intervalles.

Cette méthode est rapide, facile à mettre en œuvre et apporte des informations sur la qualité lorsque aucune référence n'est disponible. Cependant, elle reste peu discriminante de par l'utilisation des cinq catégories couvrant une échelle de grande étendue. C'est donc pour des échantillons présentant des qualités très variées qu'on emploie cette méthode.

1.2.2 Méthode d'évaluation par catégories de dégradation ou DCR (Degradation Category Rating)

Moins efficace pour distinguer des échantillons de bonne qualité, la méthode ACR est alors remplacée par la méthode d'évaluation par catégories de dégradation DCR. Pour cette méthode, les stimuli sont présentés par paires (A-B) ou par paires répétées (A-B-A-B) dans lesquelles A est l'échantillon de référence de haute qualité et B l'échantillon traité. L'échantillon de référence précède toujours l'échantillon traité et permet ainsi d'ancrer chaque jugement de l'auditeur. C'est aussi, en conséquence, une méthode plus longue que la méthode ACR.
Des "paires nulles" (A-A) sont présentées pour contrôler la qualité de l'ancrage. Les sujets notent le niveau relatif de dégradation sur une échelle de 1 à 5 :

5	Dégradation inaudible
4	Dégradation audible mais pas gênante
3	Dégradation un peu gênante
2	Dégradation gênante
1	Dégradation très gênante

La quantité évaluée d'après les notes est la note d'appréciation moyenne de la dégradation ou DMOS.

1.2.3 Méthode d'évaluation par catégories de comparaison ou CCR (Comparison category rating)

Cette méthode est analogue à celle de l'évaluation par catégories de dégradation (DCR). Dans la procédure DCR, on présente d'abord un échantillon de référence (non traité), suivi du même échantillon de parole qui a été traité par une technique quelconque. Dans la méthode DCR, les auditeurs évaluent toujours l'ampleur de la dégradation de l'échantillon traité (deuxième échantillon) par rapport à l'échantillon non traité (premier échantillon). Dans la procédure CCR, l'ordre des échantillons traités et non traités est choisi de manière aléatoire pour chaque essai. Pour la moitié des essais, l'échantillon non traité est suivi de l'échantillon traité. Pour les autres essais, l'ordre est inversé. L'avantage de cette méthode par rapport à la procédure DCR réside dans l'évaluation du traitement de la parole qui dégrade ou améliore sa qualité.
Cette méthode est donc particulièrement employée lorsqu'on ne peut prédire l'impact de la dégradation sur la parole.

Les auditeurs évaluent la qualité du deuxième échantillon par rapport à celle du premier et, de fait, formulent deux jugements avec une même réponse : "Quel est l'échantillon de

meilleure qualité ?" et "Quelle est la différence de qualité entre les deux échantillons ?".

La qualité du deuxième échantillon par rapport à celle du premier est la suivante :

3	Bien meilleure
2	Meilleure
1	Légèrement meilleure
0	A peu près équivalente
-1	Un peu moins bonne
-2	Moins bonne
-3	Beaucoup moins bonne

La quantité évaluée à l'aide des notes est la note moyenne d'appréciation par comparaison ou CMOS.

1.2.4 Méthode de détection de la réponse discontinue alternative.

Au lieu de quantifier une dimension subjective, cette méthode fournit des renseignements sur la possibilité de détection d'un défaut souvent (comme l'écho) en fonction d'un paramètre physique (comme le niveau d'écoute). L'échelle dite d'appréciation subjective de possibilité de détection est la suivante :

A	Gênant
B	Décelable mais pas gênant
C	Indécelable

L'échelle décrite ci-dessus peut être utilisée pour la détection d'écho, de réverbération ou d'un effet local, alors que la diaphonie, voire l'écho dans certains cas, peut être estimée d'après l'échelle Intelligible - Décelable - Indécelable.

L'étude du bruit, de l'évanouissement ou d'autres perturbations s'effectue sur une échelle comprenant beaucoup plus de notes, par exemple :

A	Inaudible - Bruit absolument indécelable
B	Juste audible - Une écoute attentive permet tout juste de déceler le bruit.
C	Léger - Bruit décelable, mais pas gênant.
D	Modéré - Bruit légèrement gênant.
E	Plutôt fort - Le bruit cause une perturbation appréciable.
F	Fort - Le bruit est très gênant, mais la communication continue.
G	Intolérable - Le bruit est si fort que la communication est abandonnée, ou que l'opérateur est prié de changer la ligne.

Bien qu'elles soient celles recommandées aujourd'hui pour l'évaluation de la qualité vocale et largement utilisées dans la plupart des tests de laboratoires, ces méthodes ne font pas l'unanimité. Elles présentent en effet de nombreuses limites que nous allons tenter d'expliciter. Cette réflexion sur les méthodes subjectives constitue les fondements d'une vision beaucoup plus large de l'évaluation de la qualité vocale sur laquelle est basée mon stage.

1.3 Les limites des méthodes actuelles.

1.3.1 Les limites de l'échelle de catégories.

L'une des limites des méthodes décrites ci-dessus est sans doute liée à l'utilisation même de catégories [Gros, 2001]. Ces problèmes sont issus des échelles elles-mêmes. En effet, l'emploi des catégories est problématique : deux stimuli dans une même catégorie seront jugés plus similaires ou plus difficilement discriminables que deux stimuli dans deux catégories différentes. De la même manière, où placer un stimulus au sein d'une catégorie ? Ou plutôt, quel jugement porter sur un stimulus dans une catégorie ? Ne le jugerions-nous pas plus au centre qu'il n'est en réalité ? Autrement dit, ces catégories faussent le jugement. Des améliorations ont bien été proposées, notamment avec des échelles dont le pas est plus précis (l'échelle de 100 intervalles en est un exemple), mais malheureusement, elles demeurent d'une efficacité très relative puisqu'il existe un seuil de cinq ou six catégories au-delà duquel le sujet n'est plus capable de classifier les stimuli sans confusion [Gros, 2001].

De plus, chacune des catégories est associée à un attribut qui est sensé expliciter son contenu. Cette description verbale est là encore sujette à caution. Une expérience de Jones de 1986, rapportée dans [Gros, 2001], dévoile un aspect du problème : il demande aux sujets de placer quinze qualificatifs sur une échelle continue bipolaire allant de "best imaginable" à "worst imaginable". Pour chaque qualificatif présenté séparément et dans un ordre différent pour chaque sujet, les sujets font une marque sur l'échelle de façon à ce que sa position reflète la perception qu'ils ont du mot. Il ressort de cette expérience que les cinq qualificatifs "excellent", "good", "fair", "poor", "bad" ne sont pas placés à intervalles réguliers sur l'échelle. D'autres expérimentations ont donné des résultats similaires, notamment celle rapportée dans l'article de [Mullin et al., 2002]. Dans cette étude, vingt quatre sujets anglais positionnent vingt et un qualificatifs sur une échelle continue de 200 mm. Les cinq qualificatifs cités ci-dessus ne sont à nouveau pas placés à intervalles réguliers, comme l'illustre (cf. figure 1.1). Ce sont pourtant ces mêmes attributs ("excellent", "good", "fair", "poor", "bad") qui apparaissent sur l'échelle MOS utilisée comme une échelle d'intervalles.

En outre, il est intéressant de noter que les résultats de ces expériences varient d'une langue à l'autre et que les intervalles entre les qualificatifs ne sont pas les mêmes pour les italiens et pour les américains. Jones a montré qu'un italien associe "OK" à "Good" alors qu'un américain l'assimile à "Fair ". Dès lors, comment uniformiser les échelles sur le plan international sans se heurter aux questions de traduction, de sémantique et d'interprétation puisque de nombreux termes dans une langue ont un sens bien spécifique et n'ont pas d'équivalent dans une autre langue ?

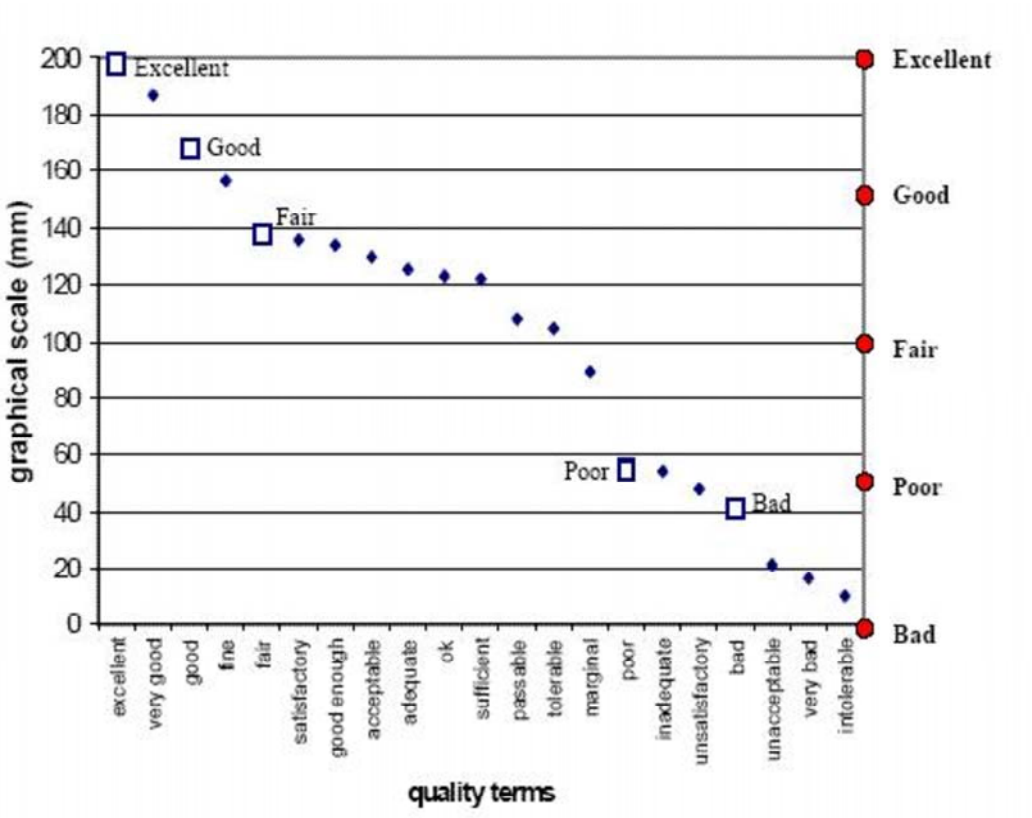

FIG. 1.1 – Positions moyennes des attributs sur une échelle de 200 mm. Les termes de l'UIT sont indiqués par des carrés bleus. L'axe de droite montre les positions théoriques des termes de l'UIT sur une échelle de cinq catégories.

De la même façon, pour une même langue, on se retrouve encore face à l'aspect sémantique de ces qualificatifs : au sein même d'une langue, comment maîtriser les diverses interprétations que chacun peut donner, inconsciemment ou non, à un mot ? Ce problème est particulièrement flagrant sur des échelles bipolaires qui servent parfois à qualifier des sons. Pour exemple : placez sur l'échelle suivante le son que vous venez d'entendre :

clair ——————— **flou**

Qu'entendez-vous par clair ? [Cheminée, 2004] Un son clair, c'est-à-dire défini, net, précis, finalement dont la définition est très bonne ou alors, un son sec, plutôt brillant, percussif, percutant, métallique, clinquant, perçant, dur, aigu, criard, agressif, pauvre en fait en harmonique et qui "perce" l'oreille ? Ou bien un son lumineux, brillant, riche, incisif, cuivré, chaud, rond, ouvert... une certaine transparence ?

Il semble donc que la question du choix et de la définition des attributs verbaux de façon à ce qu'ils soient compris sans ambiguïté par les sujets et en adéquation avec le son testé soit

épineuse.

1.3.2 L'absence de prise en considération du contexte

[Gros et al., 2005] font tout d'abord remarquer que l'acte même de porter un jugement n'est pas du tout représentatif des situations de l'utilisation des services de télécommunications. Par ailleurs, les méthodologies subjectives considèrent la qualité vocale uniquement comme une caractéristique du signal transmis indépendamment du contexte de la communication. Or, aujourd'hui, le contexte, c'est-à-dire l'environnement physique, l'utilité et le but de la communication mais aussi l'expérience de l'utilisateur, la présence ou non de tâches secondaires etc. contribue fortement à la satisfaction de l'utilisateur. Compte tenu de la diversité des services, des nombreuses possibilités d'utilisation, en particulier géographiques, la satisfaction ne se limite pas à l'agrément et au confort de l'écoute de l'utilisateur par rapport à un signal plus ou moins dégradé. Prenons l'exemple d'un alpiniste de haute montagne pris dans d'une avalanche. Aussitôt, il alerte les secours avec son téléphone portable. Bien sûr, dans un tel lieu si peu civilisé, la qualité est exécrable, la communication hachée, mais le message est passé, il est sauvé. Soulagé et comblé ! Imaginons le même dialogue dans test de laboratoire : le jugement d'un sujet est assurément mauvais, 1 sur une échelle de 1 à 5.

Force est de constater que seul le confort d'écoute de l'utilisateur ne détermine donc pas la qualité perçue d'une communication ; l'efficacité de la communication, en termes de résultats et de ressources utilisées, dans un contexte donné pour répondre à un besoin spécifique y concourre fortement.

Chapitre 2

Une autre approche de la qualité

Une approche unidimensionnelle de la qualité dans laquelle seule la satisfaction de l'utilisateur d'un service quant à la qualité du signal de parole transmis, mesurée à travers un jugement subjectif sur une échelle de qualité à cinq catégories est prise en compte, nous semble un peu réductrice pour la multitude de services qui nous sont offerts aujourd'hui.

2.1 Les critères comportementaux et électrophysiologiques

Dans le cadre de l'application à la visio-conférence, [Mullin et al., 2001] et [Wilson et Sasse, 2001] proposent d'élargir la notion de qualité telle qu'elle est conçue aujourd'hui et de la considérer selon une approche tri dimensionnelle (cf. figure 2.1) qui regroupe à la fois la satisfaction de l'utilisateur, son stress et ses performances. Le poids de chaque dimension dépend du contexte ; par exemple, si la visioconférence est utilisée dans le cadre du divertissement, c'est la satisfaction qui devient prioritaire sur les deux autres dimensions.

FIG. 2.1 – L'approche tri dimensionnelle. Le "user cost" est défini comme le stress de l'utilisateur.

En laboratoire, l'efficacité que nous évoquions précédemment pourrait être jugée à travers les performances obtenues pour différentes tâches, plus ou moins complexes, plus ou moins

explicites, en parallèle ou non, et le coût pour l'utilisateur en termes de ressources cognitives et de stress par exemple.

Cette représentation suggère donc de mesurer l'impact de la qualité vocale à travers des indicateurs comportementaux tels que le temps de réaction, les erreurs, etc. pour évaluer les performances des sujets, et à travers des indicateurs électrophysiologiques tels que la conductance électrodermale, le rythme cardiaque, la pression sanguine périphérique pour une mesure du stress.

2.2 Les premiers pas

Les premières expérimentations se sont attachées à mesurer le stress sur des sujets et à observer dans quelles conditions il s'installe. [Wilson et Sasse, 2001] posent les jalons de notre cheminement puisqu'ils explorent la voie des mesures électrophysiologiques pour mesurer la qualité audio et vidéo. La conductance électrodermale (GSR), la pression sanguine périphérique (BVP) et les battements du coeur (HR)[1], indicateurs fiables du stress et de la charge cognitive, sont enregistrés dans les situations suivantes : diverses dégradations audio en écoute passive et en interview audiovisuel préenregistré que le sujet regarde. Les résultats de ces expérimentations établissent que des dégradations audio se lisent sur les signaux BVP et HR dès l'écoute passive. Il faut ajouter le canal visuel pour que le signal GSR fournisse des conclusions significatives. De plus, les signaux physiologiques mesurés répondent différemment pour les dégradations audio et vidéo. Ainsi, il semblerait qu'il y ait différents patterns d'excitation pour différentes dégradations et qu'ils dépendent partiellement de la tâche.

Du coté des critères comportementaux, un travail sur l'analyse des risques de l'utilisation du téléphone portable en situation de conduite effectué par l'INRETS [Pachiaudi, 2002] a retenu notre attention : ils ont observé des sujets en situation de conduite simple et dans un deuxième temps en situation de double tâche, téléphoner en conduisant. Ils ont pu ainsi vérifier que l'addition de la tâche secondaire déclenche une augmentation significative de la charge mentale. Ils ont découvert, en plus de la réduction du champ de vision des conducteurs, une forte élévation du temps de réaction lorsqu'il s'agit de freiner dès qu'une alarme rouge s'allume sur le tableau de bord du simulateur. Cette augmentation peut atteindre jusqu'à 70% du temps de réaction en conduite simple selon l'âge du sujet, la complexité de la tâche de conduite et de la conversation.

[Sonntag et al., 1998] ont mesuré la compréhension de séquences vocales prononcées par six voix de synthèse de qualité diverses, et une voix naturelle, codées en GSM ou non (donc susceptibles d'être dégradées ou non) dans une situation de double tâche : la voix demandait aux sujets d'effectuer un calcul mental pendant qu'ils cliquaient sur des carrés de couleurs à l'écran d'un ordinateur. Cette expérience a permis de mettre en évidence des différences significatives de temps de réaction à la réalisation des deux tâches pour les différents types de voix. De plus, pour deux des sept types de voix, les différences entre les séquences codées et non codées GSM sont significatives.

[1] Pour les explications complètes sur le fonctionnement des mesures physiologiques, se reporter au chapitre 3.8.

Enfin, [Lai et al., 2001] montrent par le biais d'une expérience dans laquelle les sujets doivent à la fois conduire un simulateur et prêter attention à des messages que la compréhension de messages en voix naturelle est plus rapide que celle des messages en voix de synthèse sans pour autant que les performances de conduite ne soient influencées. Ainsi, temps de réaction et performances ne sont pas toujours liés.

Ces quatre études montrent la pertinence à la fois de la mesure physiologique et de la mesure comportementale pour une mesure de la qualité vocale ou audio d'un service, au-delà de la seule satisfaction de l'utilisateur en termes de confort d'écoute.

Dans le cadre d'un stage de trois mois l'an passé, l'équipe qui m'a accueilli a commencé à explorer cette voie en réalisant une première expérimentation utilisant la même double tâche que [Sonntag et al., 1998]. Des réponses comportementales (temps de réponse et performances) ainsi qu'un signal physiologique (la conductance électrodermale) ont été mesurés dans une situation de double tâche, pour différents types de voix (naturelle et de synthèse), et différents types de dégradations (bruit, perte de paquet IP, etc.), cf. [Gros et al., 2005]. En première analyse, il semblerait qu'il y ait un effet du type de qualité sur les temps de réaction de la tâche secondaire, mais qu'il n'y ait pas d'effet sur les temps de réaction de la tâche primaire et la conductance électrodermale.

Cependant, plusieurs éléments peuvent éclairer ces résultats. D'une part, l'article de [Sonntag et al., 1998] est publié en 1998 et depuis, les voix de synthèse ont considérablement évolué et il ne serait pas étonnant que l'amélioration de la qualité ait inhibé les résultats trouvés de 1998. D'autre part, il semblerait que la forte variabilité des temps de réaction de la tâche secondaire provienne de la variabilité de la complexité des calculs et masque ainsi l'influence potentielle de la qualité vocale. De plus, une analyse détaillée du comportement des sujets montre que lors de l'exécution du calcul mental, les sujets ont tendance à suspendre l'activité de la seconde tâche tant qu'ils n'ont pas donné leur réponse. Ce comportement suggère un traitement séquentiel et non parallèle des tâches comme cela était prévu, et visiblement amplifié lorsque la qualité est altérée. Cette observation nous amène à réfléchir sur la situation de double tâche.

2.3 Réflexion sur la double tâche

La configuration de double tâche permet, nous l'avons vu, de mettre le sujet dans un état de surcharge cognitive propice à des réactions émotionnelles plus intenses, mesurables à travers de mesures électrophysiologiques, et à des diminutions de performances, mesurables par les temps de réaction et le taux d'erreurs. Le choix de la double tâche est un problème délicat car il existe une panoplie de tâches réalisables en laboratoire : comment choisir les deux tâches concurrentes ? Pourquoi choisir du calcul mental plutôt que la répétition de phrases par exemple ? Quelle tâche secondaire privilégier et pourquoi ?

Il s'avère que les mécanismes cognitifs que ces tâches vont déclencher lorsqu'elles sont en concurrence restent parfois imprévisibles. Dans les tests de [Landström et al., 2002], les sujets se soumettent à trois tâches : la première tâche consiste à lire mentalement un texte dans

lequel ils doivent repérer des erreurs. La deuxième tâche est un raisonnement dit grammatical, dans lequel ils jugent si la proposition énoncée est une description correcte ou non de l'ordre des lettres A et B par rapport à ce qu'il est inscrit sur leur formulaire. Par exemple, si la voix annonce : "A n'est pas suivi de B" et que le formulaire indique "AB", le sujet doit répondre faux sur le papier. Dans la troisième tâche, les sujets trient des cartes en fonction des adresses de celles-ci. Les sujets sont perturbés dans leur travail par une voix ou un bruit de fond, qui joue le rôle de distracteur. Les expérimentateurs ont choisi des tâches verbales (de lecture et de raisonnement) pour interférer le plus possible avec la parole, du moins le pensaient-ils. Or ce n'est pas ce qu'ils observent : les performances des sujets, *i.e.* le nombre d'erreurs comptabilisées pour chaque épreuve, ne montrent aucune différence significative entre les deux types de distracteurs pour les tâches de lecture et de raisonnement et de plus, elles indiquent que les sujets ont moins bien réussi dans la tâche de tri (et non dans les tâches verbales) en présence de parole.

A ce premier obstacle vient se joindre l'estimation, non moins aisée, de la difficulté des tâches, qu'il ne faut ni trop simple, ni trop compliquée. De plus on ne peut savoir si une absence de résultats est liée à la trop grande facilité de la tâche ou au fait qu'il n'y ait réellement pas d'interférence. C'est à une tâche de conduite trop simpliste que les chercheurs ont imputé l'absence d'effets des messages sur les performances (cf. [Lai et al., 2001]).

Ces problèmes embarrassants découlent des phénomènes d'interférences faisant l'objet de différentes théories. Selon [Pashler, 1994], les différents processus susceptibles de se mettre en œuvre lors d'une situation de double tâche sont les suivants.

- L'approche la plus largement acceptée probablement est celle des ressources partagées **(Capacity Sharing)** qui s'apparente à un traitement parallèle des tâches. Ainsi, tout se passe comme si chacun d'entre nous avait un réservoir de ressources qu'il partageait parmi les différentes tâches à effectuer simultanément. Il en résulterait des capacités moindres allouées à chacune et en conséquence des performances réduites.

- Cependant, une alternative à cette représentation est de dire que ce traitement parallèle peut être impossible pour certaines opérations mentales. Il se peut qu'elles nécessitent un mécanisme qui leur est dédié pour une période de temps définie. Quand deux tâches font appel au même mécanisme, elles provoquent alors un goulot d'étranglement retardant ou altérant ainsi une ou les deux tâches à l'image des grains de sable dans un sablier. Ce modèle s'appelle, comme on peut s'en douter, le modèle du goulot d'étranglement **(Bottleneck or task switching Model)**. Cependant, ce modèle suscite de nombreuses recherches par les questions qu'il soulève : y a-t-il un ou de multiples goulots qui seraient associés aux divers étages du processus cognitif mis en jeu. S'il n'y en a qu'un seul, à quelle étape du traitement se trouve-t-il ? Est-ce au niveau de l'attention, de la mémoire sensorielle, de l'analyse perceptuelle, sémantique, de la production de la réponse... ?

- Enfin, la dernière possibilité envisagée aujourd'hui, la diaphonie (ou **CrossTalk**), repose sur le fait que l'interférence peut dépendre, non pas du processus engendré, mais du contenu de l'information traitée. Ce point de vue rencontre celui qui considère que notre perception des objets est d'abord fonction de leurs attributs physiques élémentaires. Alors des éléments comme les entrées sensorielles, les réponses produites, les pensées de la personne etc. prendraient leur importance et interviendraient dans l'inter-

férence. Selon les théoriciens, il serait plus difficile de réaliser deux tâches qui impliquent les mêmes informations, les mêmes entrées sensorielles par exemple.

Tout l'enjeu consiste à savoir si une théorie prime sur les deux autres ou bien comme le suggère l'auteur, s'il n'y a pas une stratégie qui se met en place pour choisir la voie la plus adaptée et pouvoir exécuter les tâches au mieux. Nous n'avons pas la prétention ni l'ambition d'exposer exhaustivement les tenants et les aboutissants de ces théories ; il s'agit plutôt ici de mettre en évidence la complexité des phénomènes d'interférence associés aux situations de double tâche utilisées dans les études rapportées précédemment.

2.4 La démarche vers l'expérimentation

Plutôt que de "tâtonner" pour trouver une double tâche adéquate nous permettant de mettre en évidence un impact de la qualité vocale sur les performances et/ou l'état émotionnel du sujet, il nous semble plus judicieux d'utiliser une tâche simple faisant intervenir un processus cognitif spécifique. Autrement dit, cette position semble préférable à celle qui consisterait à imaginer différentes situations de double tâche faisant intervenir plusieurs processus cognitifs (attention, mémoire, etc.) et rechercher par tâtonnements, à travers ces situations de double tâche, un effet du niveau de qualité vocale sur différents critères physiologiques et comportementaux.

Nous pensons notamment nous appuyer sur les résultats de l'écoute dichotique, nous inspirer des expériences sur l'attention sélective par le biais de tâches de filature et de prospection de cibles, relatés dans [Fortin et Rousseau, 1993] et les appliquer à notre recherche sur la qualité vocale. L'écoute dichotique présente en effet l'avantage d'être un protocole propre à l'attention et dans lequel l'attention entre en jeu.

Par la suite, cela pourrait nous guider dans le choix des deux tâches concurrentes dans un protocole de doubles tâches plus complexes. Sachant quels sont les processus perceptifs et cognitifs mis en jeu dans le traitement de la qualité vocale, on pourrait alors optimiser une ou plusieurs situations multi-tâches ou non d'ailleurs permettant de mesurer l'impact de la qualité vocale sur l'utilisateur.

Chapitre 3

L'expérimentation

Nous testons l'hypothèse suivante : la qualité d'un signal de parole transmise par un système de communication peut être mesurée à travers son impact sur les performances des sujets dans la réalisation d'une tâche basée sur ce signal de parole. Plus la qualité se dégrade, plus les performances diminuent. Les performances des sujets sont mesurées à travers leurs temps de réaction et les taux d'erreurs, et le stress des sujets, à travers des signaux électrophysiologiques.

3.1 La méthodologie

Au lieu de s'orienter vers une situation de double tâche, nous optons pour une seule tâche, une situation d'écoute dichotique. Le principe de l'écoute dichotique présente l'avantage, comme la situation de double tâche de partager l'attention du sujet. Et nous augmentons ainsi la charge cognitive et espérons ainsi marquer davantage l'effet de la qualité s'il tant est qu'il y en ait un.

3.2 La tâche

Le principe de l'écoute dichotique est simple : le sujet entend sur ses deux oreilles des listes de mots différentes. Les mots sont présentés à cadence irrégulière s'enchaînant par l'intermédiaire de pauses aléatoires variant de 50 à 200 ms, ce qui impose un rythme assez rapide. Ainsi entre un mot de l'oreille droite et un mot de l'oreille gauche, il peut y avoir un chevauchement, amplifiant ainsi la difficulté. Des pré tests sont effectués pour déterminer la vitesse de défilement des mots sur chaque oreille, le nombre de mots cible, ainsi que le temps imparti au sujet pour réagir. Nous fixons le nombre de mots cible à vingt et le temps de réaction maximal à 800 ms. C'est aussi à ce moment que nous déterminons la durée minimale et maximale des pauses (50 à 200 ms).

Sur un écran d'ordinateur s'affichent successivement des mots cibles, que le sujet doit détecter sur l'oreille droite ou sur l'oreille gauche. De cette manière, l'attention du sujet est partagée sur ces deux oreilles. Un fois le mot détecté, le prochain mot à détecter apparaît sur l'écran.

Les instructions données au sujet sont en annexe 1. Il était notamment demandé d'être concentré pour être le plus rapide et plus performant, emphi.e. commettre le moins d'erreurs possible.

3.3 Les stimuli

Un corpus de cent mots est enregistré dans en studio d'enregistrement par deux locuteurs professionnels (un homme et une femme), en 16 bits, à 48 kHz. Il formera les listes destinées aux oreilles droites et gauches. Ces mots sont tous de type dissyllabiques. Ils sont simples et issus du registre courant. (cf. annexe 2 et cd.)

Avant toute manipulation des signaux, les mots sont égalisés à -26 dBov[1].

Deux types de dégradations sont retenus :

- la dégradation dite MNRU, Modulated Noise Reference Unit (décrite dans la recommandation P. 810 de l'UIT [UIT-T P.810, 1996]), qui permet une dégradation référencée et contrôlée des signaux vocaux. Elle ajoute un bruit modulé par le signal de parole selon la formule suivante :

  $$y(i) = x(i)(1 + 10^{\frac{-Q}{20}} \times N(i))$$

  où x(i) est le signal d'entrée à bruiter
  N(i) est le bruit aléatoire
  Q est le rapport de puissance des signaux vocaux à la puissance du bruit modulé
  y(i) est le niveau de puissance de sortie de parole + bruit modulé.

  L'indice Q varie de 30 pour la dégradation la plus faible à 5 pour la plus forte, par palier de 5. On choisit les niveaux Q = 25, 15 et 5. Pour ce type de dégradation, on se place dans une optique "recherche" qui consiste à savoir si une dégradation de la qualité, qu'elle soit représentative des cas réels ou non, a une influence ou pas sur notre comportement. C'est dans cette idée que nous choisissons le MNRU 5 qui dégrade fortement le signal. Par ailleurs, nous avons sélectionné la voix de l'homme pour y appliquer les dégradations MNRU. En effet, comme l'explique la recommandation [UIT-T P.810, 1996] il s'avère que le MNRU s'applique sur un signal échantillonné à 16 kHz et après un filtre 70-7000 Hz. Comme nous souhaitons tester l'effet du bruit (et non pas celui du filtre), nous choisissons la voix sur laquelle le filtre s'entend le moins.

- La dégradation "Bande Passante du signal"(BP). Pour ce type de dégradation, on s'oriente davantage vers des dégradations qui correspondent à cas réels, avec des dégradations beaucoup plus fines. Les niveaux choisis correspondent aux bandes passantes utilisées aujourd'hui dans les services de communication : 50-14000 Hz (super wide band), 50-7000 Hz (wide band), et 300-3400 Hz (narrow band ou bande téléphonique). De la même façon, cette dégradation est appliquée à la voix féminine, pour un effet de

[1] dBov niveau relatif à la surcharge d'un système numérique

la réduction de bande plus prégnant.

Ainsi, nous disposons de quatre niveaux de qualité pour chaque type de dégradations, si l'on considère l'absence de dégradation (le signal à 48 kHz) comme le quatrième niveau.

Enfin, les signaux résultants sont sur-échantillonnés, si besoin, à 48 kHz pour le système de restitution des sons. Les divers processus de sous et sur échantillonnages, additions de bruits et réductions de bande passante ont été réalisés soit à l'aide de Matlab, soit à l'aide d'un exécutable issus de la bibliothèque d'algorithmes utilisée par l'Union Internationale des Télécommunications [ITU-T Software Tools Library, 2005].

3.4 Les sujets

Vingt-quatre sujets naïfs (dix femmes et quatorze garçons), volontaires, tous droitiers, âgés de 22 à 55 ans ont participé au test. Tous les sujets sont otologiquement sains (un audiogramme est réalisé avant le test).

3.5 La procédure

3.5.1 Le test

Le test comprend deux sessions de 10-15 minutes chacune, une pour le MNRU et une pour la bande passante, avec une pause entre les deux sessions. Une session est composée de quatre écoutes d'apprentissage et de huit écoutes test. Chaque écoute correspond à une liste de cinquante mots sur chaque oreille parmi lesquelles il faut détecter vingt mots cibles, dix sur chaque oreille.

Pour chaque session, *i.e.* chaque type de dégradation, les huit écoutes tests correspondent aux quatre niveaux de dégradations du signal, présentés de façon symétrique (on a le même niveau de dégradation dans chaque oreille) et asymétrique (la dégradation est présentée sur une seule oreille uniquement).

Les sujets sont divisés en deux groupes selon l'oreille sur laquelle est présentée la dégradation en situation asymétrique : pour le groupe A, les dégradations sont présentées sur l'oreille droite (OD), pour le groupe B, les dégradations sont présentées sur l'oreille gauche (OG).

Les notations employées sont les suivantes : OG pour Oreille Gauche, OD pour Oreille Droite, HQ pour Haute Qualité.

Les tableaux 3.1 et 3.2 donnent les huit conditions considérées, pour les deux groupes et les deux types de dégradations.

L'ordre de présentation des conditions est différent selon les sujets, suivant un carré latin d'ordre 6 (cf. tableau 3.3).

N° de la condition	1	2	3	4	5	6	7	8
OG pour A, OD pour B	HQ	HQ	HQ	HQ	MNRU 25	MNRU 15	MNRU 5	HQ
OD pour A, OG pour B	HQ	MNRU 25	MNRU 15	MNRU 5	MNRU 25	MNRU 15	MNRU 5	HQ

TAB. 3.1 – Présentation des conditions considérées pour le MNRU

N° de la condition	1	2	3	4	5	6	7	8
OG pour A, OD pour B	HQ	HQ	HQ	HQ	BP 50-14000	BP 50-7000	BP 300-3400	HQ
OD pour A, OG pour B	HQ	BP 50-14000	BP 50-7000	BP 300-3400	BP 50-14000	BP 50-7000	BP 300-3400	HQ

TAB. 3.2 – Présentation des conditions considérées pour la bande passante

Ordre	N° de la condition							
1	1	2	3	4	5	6	7	8
2	1	6	4	7	2	5	3	8
3	1	5	7	3	4	2	6	8
4	1	7	5	2	6	3	4	8
5	1	4	6	5	3	7	2	8
6	1	3	2	6	7	4	5	8

TAB. 3.3 – Agencement des huit conditions pour la session MNRU

Au sein d'un groupe, la moitié des sujets commence le test par la session MNRU et l'autre moitié par la session par la session bande passante. L'autre groupe suit la même démarche. L'organisation complète est donnée dans le tableau récapitulatif 3.4 ci-dessous.

Enfin, même si les listes de mots sont toutes fabriquées à partir des mêmes 100 mots, l'ordre de ces mots est différent pour toutes les écoutes d'un sujet et les 12 sujets d'un groupe.

Le groupe B suit la même démarche que le groupe A sauf qu'on inverse le casque. Ainsi, tout ce qu'a eu le sujet 1 du groupe A (Sujet 1A) sur l'oreille gauche, le sujet 13 (ou Sujet 1B) l'aura sur l'oreille droite. L'ordre des écoutes pour le sujet X du groupe A est le même que pour le sujet X du groupe B (sujet XB).

Notons que l'ordre des conditions n'est pas tout à fait le même d'une session à l'autre afin qu'un sujet X n'ait pas le même enchaînement en termes de niveau de qualité pour la session MNRU et session la bande passante.

	Session 1	Ordre des conditions	Session 2	Ordre des conditions
Sujet 1A	MNRU	1	BP	6
Sujet 2A	MNRU	2	BP	1
Sujet 3A	MNRU	3	BP	2
Sujet 4A	MNRU	4	BP	3
Sujet 5A	MNRU	5	BP	4
Sujet 6A	MNRU	6	BP	5
Sujet 7A	BP	6	MNRU	1
Sujet 8A	BP	1	MNRU	2
Sujet 9A	BP	2	MNRU	3
Sujet 10A	BP	3	MNRU	4
Sujet 11A	BP	4	MNRU	5
Sujet 12A	BP	5	MNRU	6
Sujet 1B	MNRU	1	BP	6
Sujet 2B	MNRU	2	BP	1
Sujet 3B	MNRU	3	BP	2
Sujet 4B	MNRU	4	BP	3
Sujet 5B	MNRU	5	BP	4
Sujet 6B	MNRU	6	BP	5
Sujet 7B	BP	6	MNRU	1
Sujet 8B	BP	1	MNRU	2
Sujet 9B	BP	2	MNRU	3
Sujet 10B	BP	3	MNRU	4
Sujet 11B	BP	4	MNRU	5
Sujet 12B	BP	5	MNRU	6

TAB. 3.4 – Récapitulatif de l'ordre des sessions et de l'écoute des conditions

3.5.2 L'apprentissage

L'apprentissage est formé des quatre écoutes suivantes, selon la session MNRU / BP :

	écoute 1	écoute 2	écoute 3	écoute 4
OG pour A, OD pour B	HQ	HQ	MNRU 5 / BP 300-3400	HQ
OD pour A, OG pour B	HQ	MNRU 5 / BP 300-3400	MNRU 5 / BP 300-3400	HQ

TAB. 3.5 – Présentation des conditions pour l'apprentissage du MNRU et de la bande passante

L'apprentissage de la session est le même pour les vingt-quatre sujets.

3.5.3 Les variables mesurées

Les variables mesurées sont le temps de réaction, qui est calculé entre la fin de la prononciation du mot cible et l'appui sur une touche du clavier, et les erreurs de détection. On enregistre la conductance électrodermale (GSR), la pression sanguine périphérique (BVP) et la fréquence cardiaque (HR) pour chaque écoute (cf. annexe 3).

3.6 Le matériel

L'écoute se fait via un casque. Une interface est créée sous Matlab pour afficher "en temps réel" les mots cibles à détecter.

La conductance électrodermale (GSR), la pression sanguine périphérique (BVP) et la fréquence cardiaque (HR) sont enregistrées grâce à un appareil développé par Thought Technology Ltd, le Procomp en liaison avec un logiciel, le Biograph. Deux pc tournent en parallèle, un pour les signaux physiologiques, un pour l'expérience proprement dite et nécessitent la présence de l'expérimentateur durant les tests.

3.7 Réalisation et contraintes

Les écoutes se font au casque, à 73 dB SPL. Le réglage se fait à l'aide d'un mannequin B&K dans lequel des microphones sont placés à l'entrée des conduits auditifs. Ils sont reliés pour l'étalonnage à un sonomètre B&K, cf. photos 3.1.

FIG. 3.1 – a) sonomètre B&K - b) mannequin B&K

Les listes de mots et l'interface sont créées sous Matlab. Plusieurs contraintes liées à ces listes surgissent ; elles sont regroupées dans l'annexe 2.

L'interface doit être synchronisée avec la liste de mots écoutés. Elle affiche le mot cible à l'avance et à partir du moment où le mot est prononcé, elle laisse au sujet un délai maximal de 800 ms pour répondre, c'est-à-dire appuyer sur 1 si le sujet a entendu le mot cible à gauche ou 2 s'il l'a entendu à droite. Lorsque le sujet appuie, le mot cible suivant apparaît.

Si le sujet n'appuie pas dans le temps imparti, le mot suivant apparaît automatiquement au bout des 800 ms. Si par mégarde le sujet se trompe de touche (il appuie par exemple sur 4 au lieu de 1), une alarme rouge l'avertit, la première réaction n'est pas prise en compte et il peut appuyer de nouveau sur la touche correcte mais toujours dans le temps qui lui est donné (800ms). L'interface récupère la touche sur laquelle appuie le sujet, s'il a commis ou non une erreur et son temps de réaction. Entre chaque écoute, l'interface affiche la mention suivante : " Appuyer sur une touche pour commencer " ; c'est donc l'utilisateur qui déclenche l'écoute. Cela lui laisse de plus un peu de répit entre chaque écoute.

3.8 Mesures électrophysiologiques : le principe

On enregistre les signaux physiologiques grâce à des petits capteurs qu'on place sur les doigts. Ces capteurs sont reliés au pc via un petit boîtier, le ProComp (cf. photos 3.2).

Selon [Wilson et Sasse, 2000], la conductance électrodermale, la pression sanguine périphérique et la fréquence cardiaque sont des indicateurs fiables du stress et de l'éveil. Détaillons leur principe.

On mesure la conductance électrodermale (SC ou GSR) au moyen des deux capteurs qui évaluent la propension de la peau à conduire l'électricité. Une infime tension électrique est appliquée entre les deux électrodes, attachées à deux doigts d'une main, afin d'établir un circuit électrique dans lequel la main devient une résistance variable. La variation en temps réel de la conductance (unité, le siemens), qui est simplement l'inverse de celle de la résistance, est calculée et affichée par le logiciel. La conduction électrodermale représente les changements de notre système sympathique, dont le rôle est de déclencher les réactions de notre organisme face à un danger éventuel, notamment par la libération de l'adrénaline. Le GSR est ainsi associé au stress et à l'éveil.

Quand une personne devient stressée, la conductance électrodermale croît. La raison précise de ce phénomène n'est pas connue mais on peut déjà avancer deux propositions d'explication : la première serait un durcissement de la peau qui la protègerait contre les dommages mécaniques. Il a été observé en effet qu'il est difficile couper la peau sous la transpiration abondante. La deuxième hypothèse consiste à dire que le GSR augmente pour refroidir le corps en vue de l'activité projetée. Le GSR est également connu pour être la mesure la plus rapide et la plus robuste de l'effort, la hausse du GSR étant aussi liée à l'effort. SC (Skin Conductance) et GSR (Galvanic Skin Response) sont deux appellations différentes pour la même mesure physiologique.

La pression sanguine périphérique (BVP) ou encore photoplethysmographie est une technique électro-optique pour mesurer l'impulsion cardio-vasculaire à travers le corps humain. Cette impulsion d'onde est due aux pulsations périodiques d'un volume de sang artériel. Elle est mesurée par le changement d'absorption optique que le sang artériel induit. Le capteur est muni d'une lumière infrarouge qui rebondit contre la surface de la peau et mesure la quantité de lumière réfléchie. Elle varie selon la quantité de sang présent sous la peau. A chaque battement de cœur, le sang afflue, réfléchit la lumière rouge, absorbe les autres couleurs et renvoie ainsi beaucoup de lumière. En revanche, entre chaque battement de cœur, la quantité

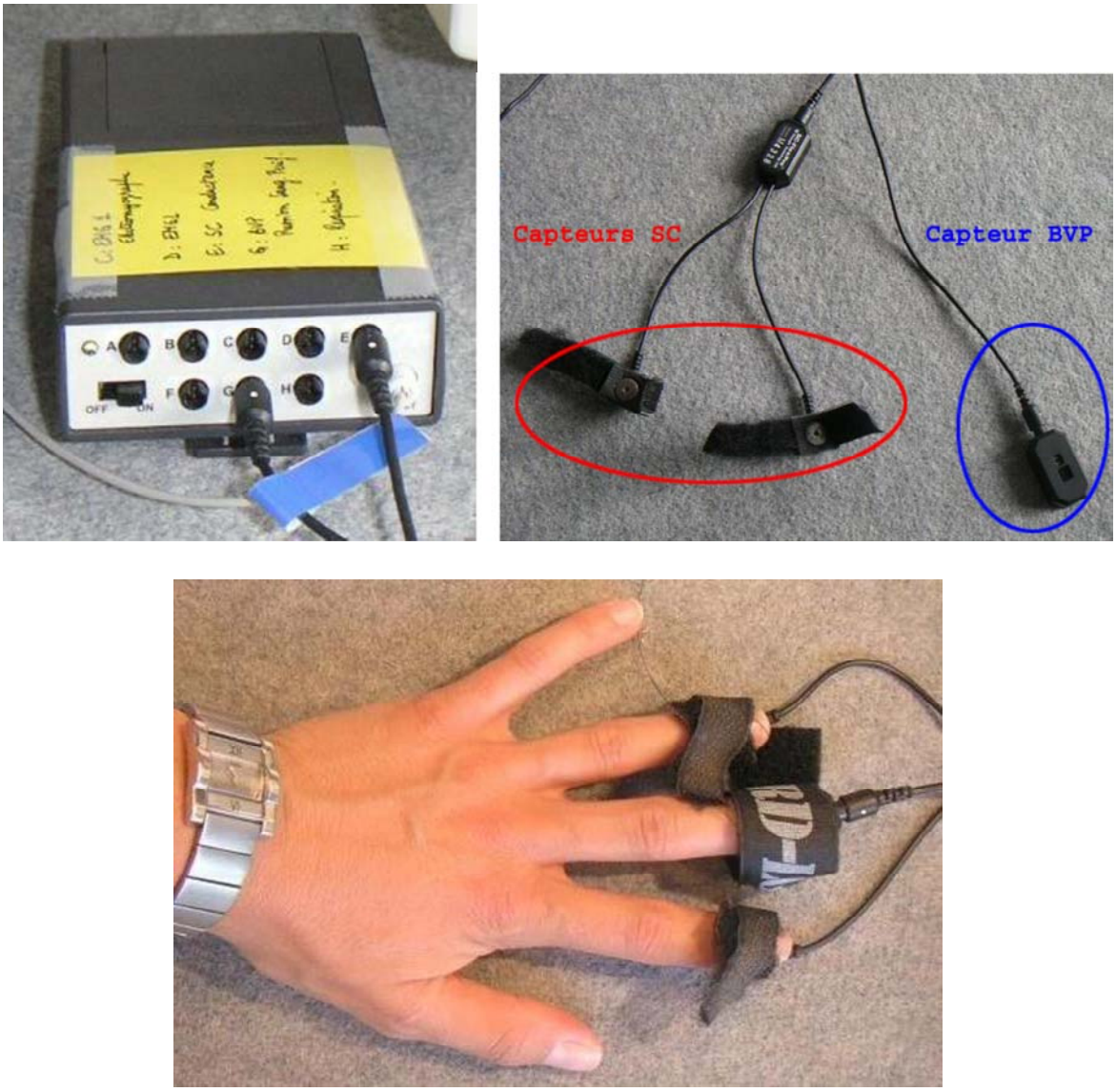

FIG. 3.2 – Le ProComp et ses capteurs SC et BVP

de sang diminue et la lumière rouge est absorbée. Autrement dit, le BVP est un indicateur de l'écoulement du sang.

BVP diminue sous l'effort et le stress puisque le sang est détourné vers les muscles qui travaillent afin de les irriguer et les préparer à une action imminente. Ceci signifie que l'écoulement de sang est réduit aux extrémités, et donc aux doigts.

Il n'est pas nécessaire d'employer un autre capteur pour mesurer la fréquence cardiaque (HR), elle se déduit de la mesure du BVP. Sous l'effet du stress, la fréquence cardiaque augmente, puisque le rythme du cœur s'accélère pour envoyer davantage de sang vers les muscles et les préparer à l'action.

Chapitre 4

Résultats et Analyses

4.1 Temps de réaction (TR) pour la session MNRU

Pour chaque sujet, et chaque condition, on dispose de 10 TR pour chaque oreille.

4.1.1 Données aberrantes

Avant toute analyse, on supprime certaines données[1] :

- les non réponses, c'est-à-dire, les temps de réaction pour lesquels les sujets n'ont pas détecté le mot cible en question. Ces temps de réaction valent par conséquent 800ms.
- les points aberrants : les temps de réaction des sujets qui ont répondu par erreur avant le début du mot cible en question.

Notons qu'il est tout à fait possible que le sujet réagisse de façon correcte avant la fin du mot cible. Le temps de réaction est alors négatif mais pas pour autant rejeté. Il faut distinguer les cas des sujets qui ont détecté le mot cible et répondu très rapidement, avant même que sa prononciation ne soit achevée, de ceux qui ont réagit alors que le mot était à peine prononcé. Dans ce dernier cas, la réponse relève du hasard ou du stress qui fait commettre une erreur au sujet, et ne peut être considérée au même titre qu'un sujet qui agit correctement. Pour juger de la validité du temps de réaction dans les cas critiques, l'arbitrage se fait manuellement et au cas par cas. Prenons un exemple : nous sommes face à un temps de réaction de -520 ms. La question est donc de savoir si le sujet a eu le temps d'entendre le début du mot cible et de le reconnaître avant d'avoir agi. Pour cela, on va tout d'abord rechercher la longueur du mot cible correspondant. Deux cas :

- le mot cible dure 400 ms. Comme le TR est mesuré à la fin du mot, cela signifie que le sujet a répondu 120ms avant le début du mot. Il s'est donc trompé et le résultat n'est pas pris en compte.
- le mot cible dure 600 ms. Cela signifie alors que le sujet aurait répondu 80 ms après le début du mot. On écoute sous Cool Edit les 80 premières ms du mot en question pour

[1] Le traitement des données aberrantes est vu dans section du MNRU mais il est bien entendu exactement le même pour la bande passante.

déterminer si dans ce temps le mot est reconnaissable. Si oui, on garde le TR, sinon, on le rejette. Dans un cas comme celui-ci, le TR ne serait pas conservé car ces vérifications ont montré qu'il fallait compter au moins 120 à 150 ms pour que la première syllabe soit dite et qu'un espoir de reconnaissance apparaisse.

Les données aberrantes retirées sont remplacées, pour l'analyse, par la moyenne du sujet pour la condition considérée.

4.1.2 Résultats pour la session MNRU

4.1.2.1 Variabilité inter-individuelle

La figure 4.1 ci-dessous présente les TR moyens par sujet. En moyenne, le TR moyen est de 0,144 ms. Mais l'on constate que certains sujets sont plus lents ou plus rapides que d'autres (TR max = 0,275 ms, TR du sujet A 4 et TR min = 0,013, TR du sujet B 7).

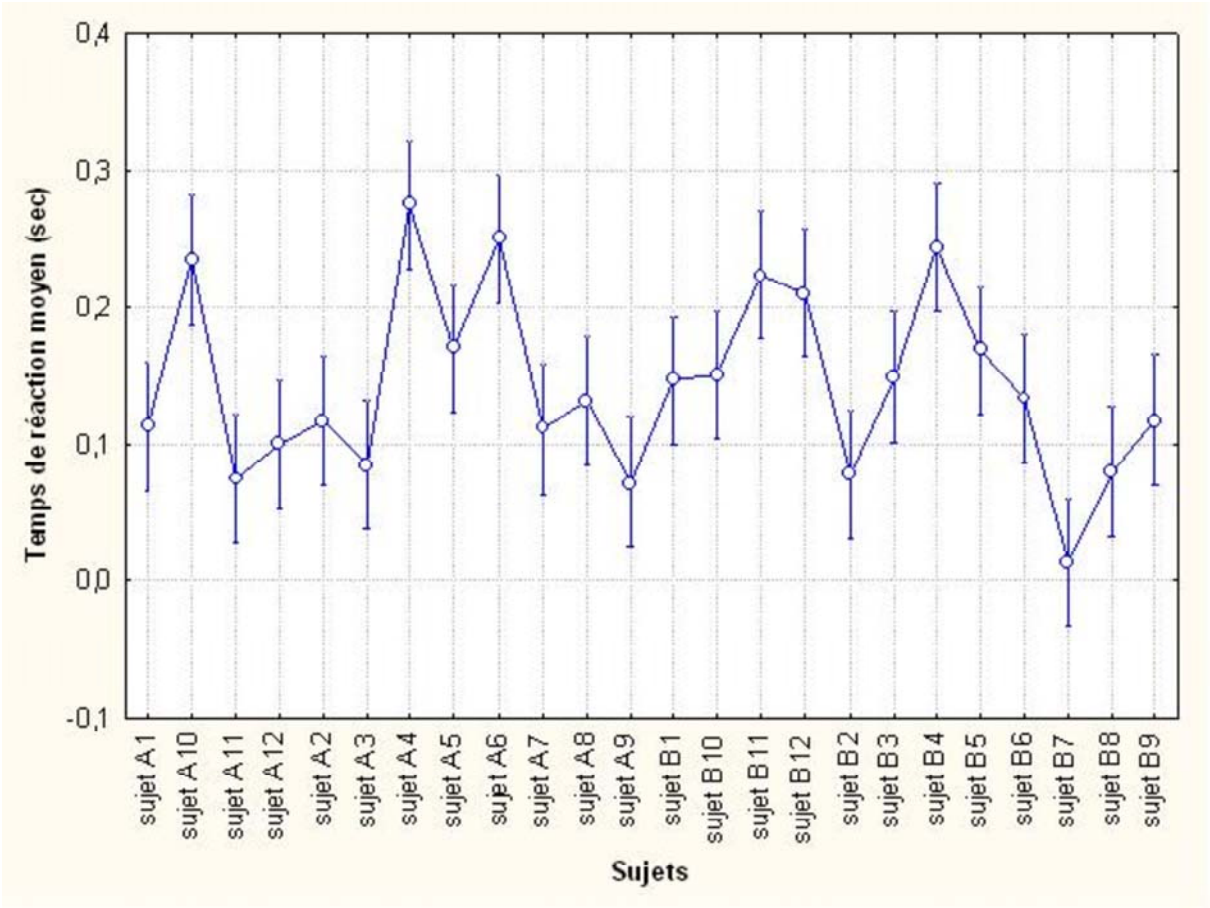

FIG. 4.1 – Temps de réaction moyens par sujet. Les barres représentent les intervalles de confiance à 0,95.

La variabilité inter-individuelle étant importante et susceptible de masquer d'autres effets plus riches en informations les données sont centrées et réduites selon la formule suivante :

$$Z = (\frac{X-\mu}{\sigma})$$

où Z est la valeur centrée réduite
X est la valeur non centrée réduite
μ est la moyenne du sujet sur toutes les conditions
σ est l'écart type du sujet sur toutes les conditions.

4.1.2.2 Effet de la dégradation MNRU sur les temps de réaction

La figure 4.2 montre les TR moyennés sur tous les sujets et sur toutes les conditions, en fonction du niveau de qualité. Il semblerait que globalement, plus la qualité se dégrade, plus le TR moyen s'allonge.

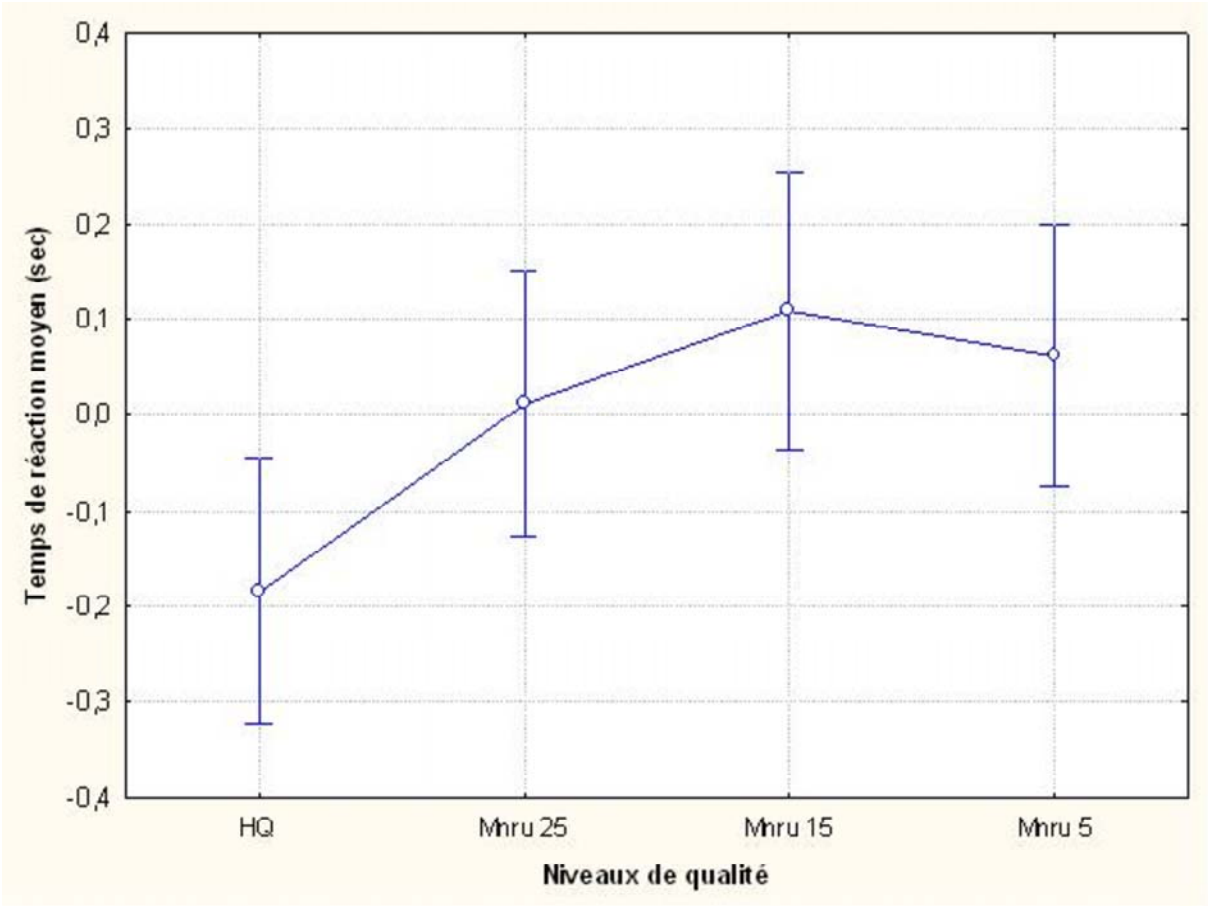

FIG. 4.2 – Temps de réaction moyens en fonction du niveau de qualité. Les barres représentent les intervalles de confiance à 0,95.

Afin de confirmer l'effet du niveau de qualité sur les TR, on réalise une ANOVA ou analyse de variance sur les TR individuels en considérant les facteurs suivants :

- Le facteur intra-groupe "Situation" à 2 niveaux :
  - **Asymétrique** : HQ/ HQ, HQ/ MNRU 25, HQ/ MNRU 15, HQ/ MNRU 5
  - **Symétrique** : HQ/ HQ, MNRU 25/ MNRU 25, MNRU 15/ MNRU 15, MNRU 5/ MNRU 5.

- Le facteur intra-groupe "Qualité" à 4 niveaux :
  - **sans dégradation (Haute Qualité)**,
  - **MNRU à 25**,
  - **MNRU à 15**,
  - **MNRU à 5**,

- Le facteur intra-groupe "Oreille" à 2 niveaux :
  - **Oreille 1** = oreille non dégradée en situation asymétrique uniquement
  - **Oreille 2** = oreille dégradée dans les deux situations.

- Le facteur Groupe à 2 niveaux :
  - **Groupe A** : groupe dans lequel la dégradation en situation asymétrique est présentée sur l'oreille droite.
  - **Groupe B** : groupe dans lequel la dégradation en situation asymétrique est présentée sur l'oreille gauche.

Le tableau 4.1 suivant récapitule les résultats de l'ANOVA. En rouge apparaissent les effets et interactions significatives :

Facteur	SC	dl1	dl2	MC	F	p
1 GROUPE	0	1	238	0	0	1
2 SITUATION	0,02	1	238	0,02	0,02	0,895
SITUATION * GROUPE	0,01	1	238	0,01	0,01	0,922
3 QUALITE	48,1	3	714	16,03	18,56	,000*
QUALITE * GROUPE	4,16	3	714	1,39	1,6	0,187
4 OREILLE	26,3	1	238	26,3	23,81	,000*
OREILLE * GROUPE	40,93	1	238	40,93	37,05	,000*
SITUATION * QUALITE	3,79	3	714	1,26	1,5	0,213
SITUATION * QUALITE * GROUPE	11,06	3	714	3,69	4,39	,005*
SITUATION * OREILLE	10,85	1	238	10,85	13,14	,000*
SITUATION * OREILLE * GROUPE	0,01	1	238	0,01	0,01	0,932
QUALITE * OREILLE	14	3	714	4,67	5,39	,001*
QUALITE * OREILLE * GROUPE	3,14	3	714	1,05	1,21	0,305
SITUATION * QUALITE * OREILLE	8,92	3	714	2,97	3,54	,014*
2 * 3 * 4 * 1	2,57	3	714	0,86	1,02	0,383

TAB. 4.1 – Récapitulatif de l'analyse pour la session MNRU

L'ANOVA confirme l'effet du niveau de qualité. Un HSD (Honestly Significant Difference) Tukey test montre que le TR moyen pour la condition HQ est significativement plus court ($p<0,05$) que les trois autres TR, similaires entre eux, ($p\geq 0,64$ pour les comparaisons entre les trois TR).

Par ailleurs, l'ANOVA révèle un effet de l'oreille sur les TR ($F(1, 238) = 23{,}810$, $p<0,000$). La figure 4.3 ci-dessous illustre cet effet. Les temps de réaction recueillis pour les mots apparaissant sur l'oreille dégradée dans toutes les conditions (oreille 2) sont significativement plus longs, en moyenne, que les TR recueillis pour les mots apparaissant sur l'oreille non dégradée pour la moitié des conditions et dégradée pour l'autre moitié (oreille 1).

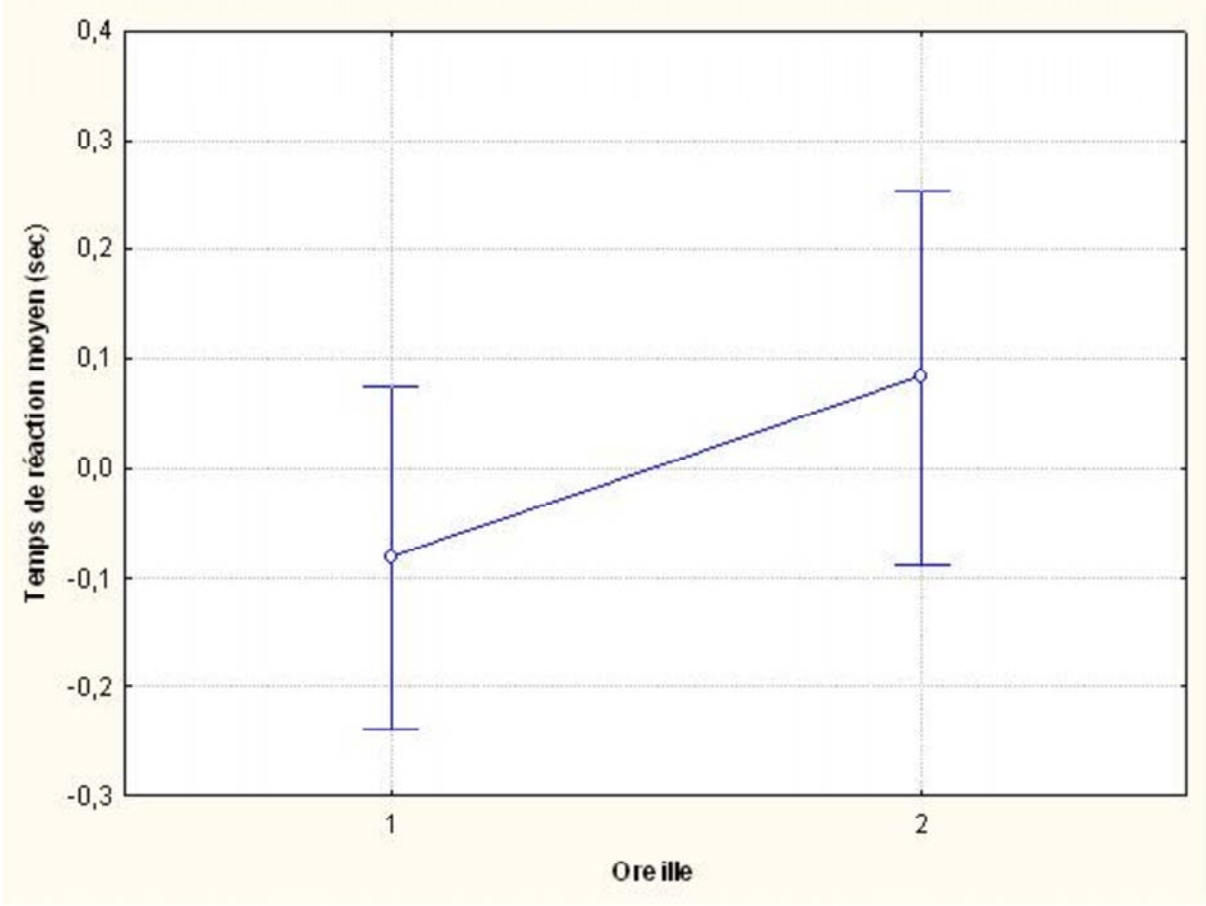

FIG. 4.3 – Effet de l'oreille sur les temps de réaction. Les barres représentent les intervalles de confiance à 0,95.

Si l'on introduit le facteur Situation (cf. interaction significative avec le facteur Oreille $F(1, 238) = 13{,}144$, $p<0,001$) on constate que dans la situation symétrique, la différence entre les deux oreilles semble disparaître, ce qui se conçoit aisément puisque les deux oreilles sont dégradées à l'identique. En revanche, la différence de comportement entre les deux oreilles est clairement présente dans la situation asymétrique (cf. figure 4.4).

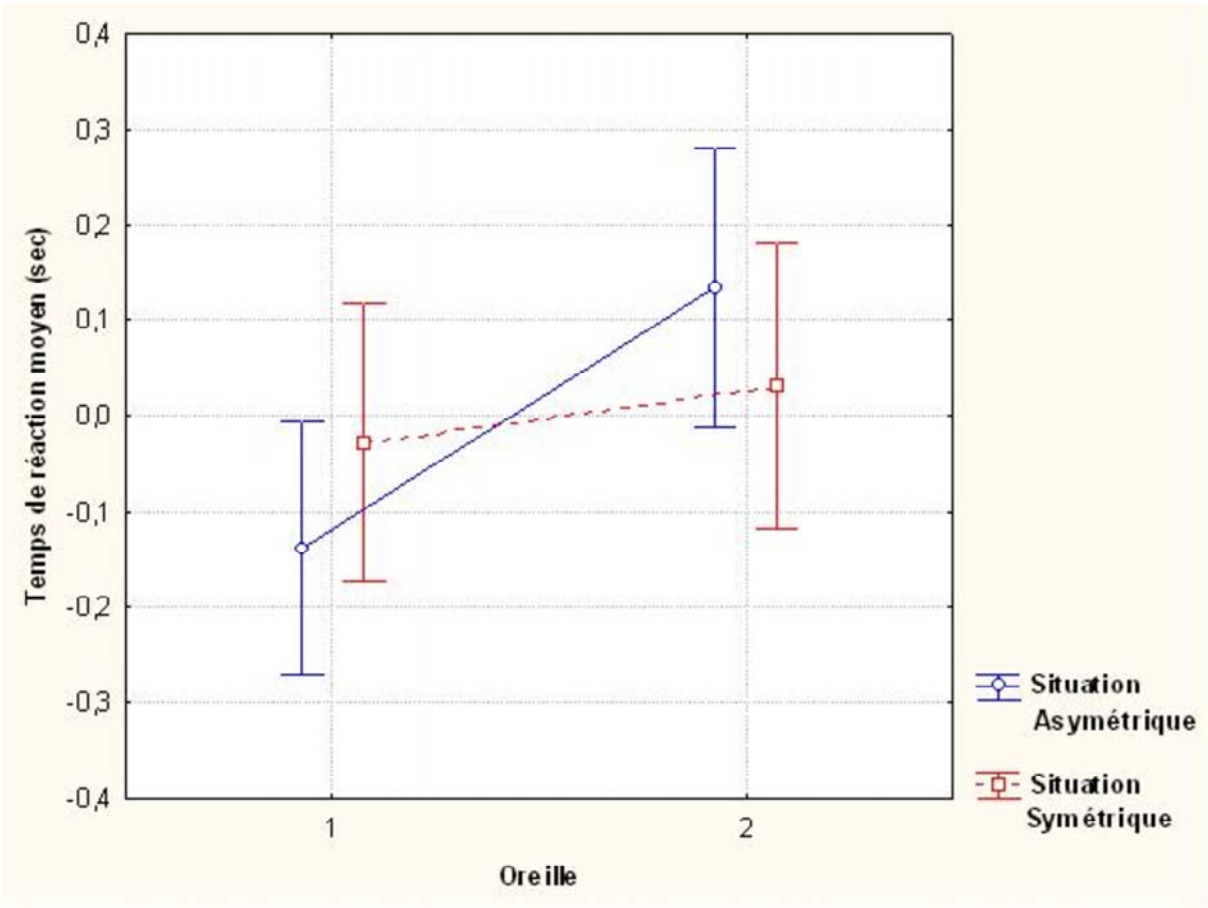

FIG. 4.4 – Interaction entre les facteurs situation et oreille.

Un HSD (Honestly Significant Difference) Tukey Test confirme que la différence significative entre les deux oreilles en situation asymétrique ($p<0,05$) disparaît en situation symétrique ($p=0,48$).

Ainsi, la présence de dégradations de type MNRU dans le signal de parole présentée à une oreille semble détériorer les performances de détection de mots cibles sur cette oreille, en termes de temps de réaction.

La figure 4.5 résume l'interaction significative entre les trois facteurs ($F(3,714) = 3,539$, $p<0,05$). Elle confirme les deux effets observés précédemment : plus la qualité se dégrade, plus les performances diminuent, mais seulement sur la ou les oreilles dégradées selon que l'on se trouve en situation asymétrique ou symétrique. On rappelle que l'oreille 1 est l'oreille non dégradée en situation asymétrique.

D'autre part, il est intéressant de constater sur la partie gauche de la figure qu'une dégradation sur une oreille, en l'occurrence l'oreille 2, ne semble pas avoir d'influence sur l'oreille non dégradée (oreille 1) ; les TR de l'oreille sont en effet sensiblement constants. Cette figure

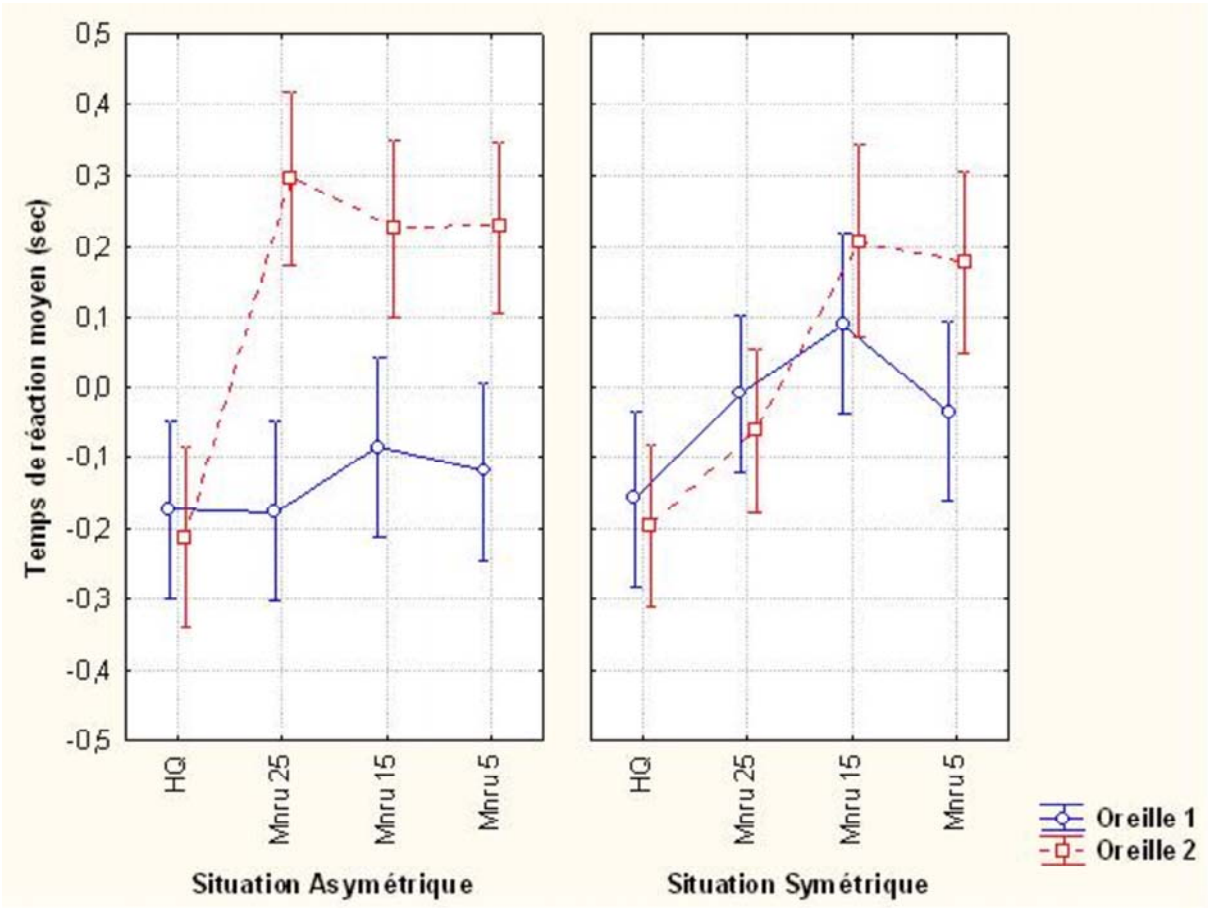

FIG. 4.5 – Interaction entre les facteurs situation, oreille et qualité.

met donc en évidence un phénomène d'indépendance des oreilles, de l'une part rapport à l'autre.

L'ANOVA sur les données centrées réduites révèle également une interaction significative entre les facteurs Groupe et Oreille ($F(1, 238) = 37{,}054$, $p<0{,}000$), illustrée par la figure 4.6. On rappelle que le groupe A est le groupe pour lequel la dégradation en situation asymétrique est présentée sur l'oreille droite, alors que le groupe B est le groupe pour lequel la dégradation en situation asymétrique est présentée sur l'oreille gauche.

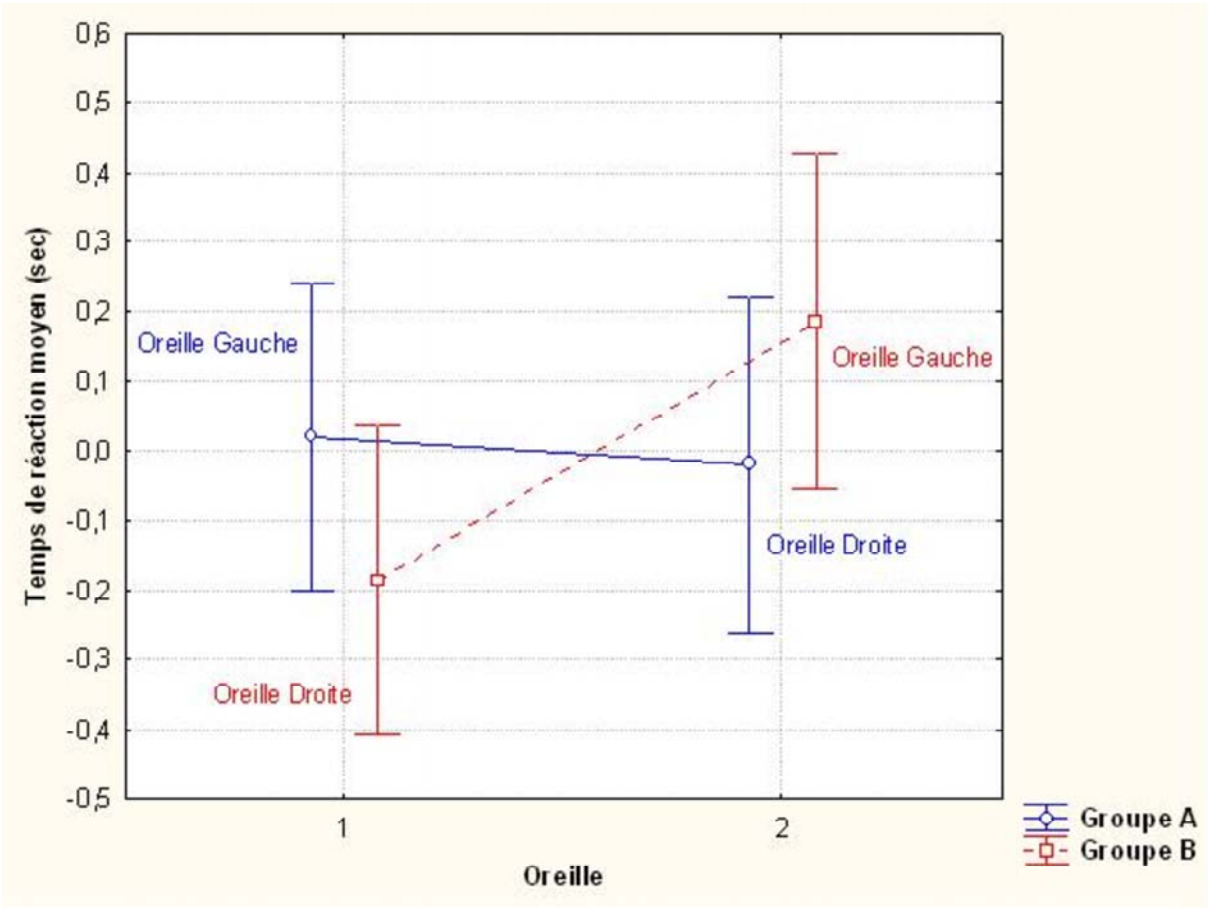

FIG. 4.6 – Interaction entre les facteurs oreille et groupe.

Il semblerait que en moyenne, les performances en termes de TR de détection de mots cibles sont meilleures sur l'oreille droite que sur l'oreille gauche pour le groupe B, et non pour le groupe A. Ces résultats sont appuyés par un HSD Tukey Test : pour le groupe B, les performances de l'oreille droite sont significativement meilleures que celle de l'oreille gauche ($p<0,05$) alors que pour le groupe A, cette distinction s'efface ($p = 0,829$).

En effet, si l'on regarde les figures 4.7a et b, qui illustrent l'effet combiné entre les quatre facteurs Oreille, Situation, Groupe et Qualité (interaction non significative) on constate, sur la figure 4.7b, en situation symétrique, que les performances en termes de TR de détection de mots cibles sont globalement meilleures sur l'oreille droite que sur l'oreille gauche, pour les deux groupes.

En situation asymétrique, l'effet est renforcé pour le groupe B puisque l'on dégrade l'oreille gauche la moins performante. Pour le groupe A, les performances sont meilleures sur l'oreille droite pour la condition HQ/HQ, et deviennent moins bonnes que l'oreille gauche, dès lors que l'on introduit des dégradations. Un HSD Tukey test confirme ces résultats en précisant que la chute des performances de l'oreille droite entre la condition HQ et les conditions MNRU 25, 15, 5 est significative ($p<0,05$).

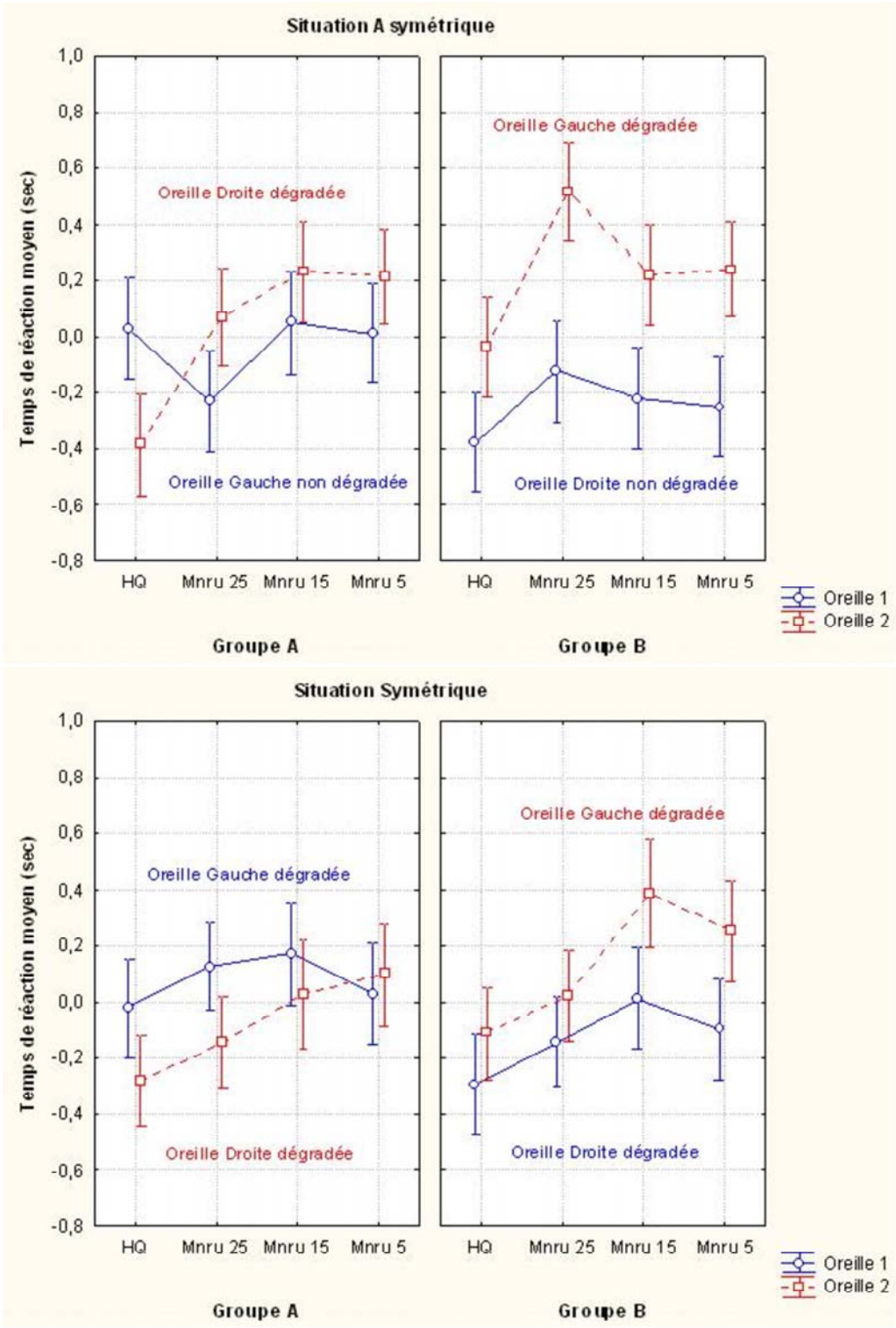

FIG. 4.7 – a) et b) Interaction entre les facteurs groupe, oreille, qualité et situation.

4.1.3 Discussion

On retiendra trois conclusions de ces premières analyses. Tout d'abord, on confirme notre hypothèse de départ : plus la qualité se dégrade, plus les temps de réaction de détection de mots cibles augmentent, mais cela, uniquement sur la ou les oreilles dégradées. Par ailleurs, il existe une sorte d'indépendance des oreilles qui leur permet de réagir de manière autonome sans être influencée par une dégradation sur l'oreille controlatérale. Enfin, on a observé une prédominance de l'oreille droite sur l'oreille gauche dans la détection de mots cibles, avec des sujets droitiers.

Revenons sur la première conclusion qui affirme que la qualité influence les temps de réaction. Précisons que cet effet est dû à la différence entre la condition HQ et les trois autres conditions MNRU. Autrement dit, nous n'avons pas d'information sur un éventuel seuil de dégradation à partir duquel le TR commencerait à s'allonger. L'absence de différence entre les trois niveaux de MNRU ne nous permet pas de quantifier le niveau dégradation. Comme dans les articles de [Shtyrov et al., 2000a], [Shtyrov et al., 2000b], [Zatorre et al., 2002] et [Hickok et Poeppel, 2000], la troisième observation vient corroborer à nouveau le modèle de WERNICKE GESCHWIND. Ce modèle donne un aperçu fonctionnel de l'organisation des traitements réalisés par le cerveau pour les tâches de reconnaissance et parole. En voulant soigner des aphasiques, BROCA d'abord, au XIXe siècle, puis WERNICKE et GESCHWIND, au XXe siècle localisent dans l'hémisphère gauche les aires du cerveau relatives à la parole : l'aire de BROCA et l'aire de WERNICKE (cf. figure 4.8 et [Bases neurologiques du langage]). Dans sa théorie, GESCHWIND, en 1979, met en évidence l'existence de tâches intermédiaires associées à des zones particulières du cortex.

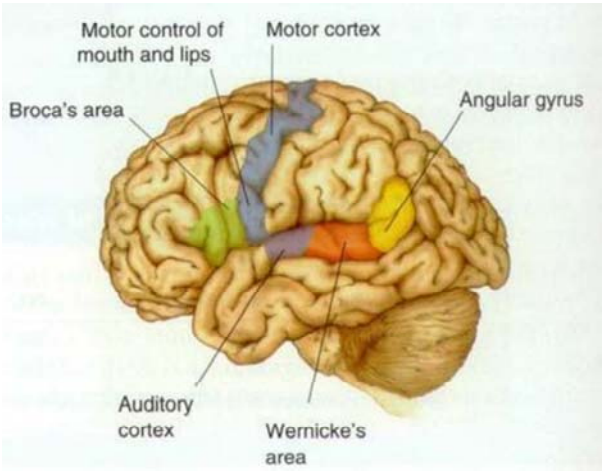

FIG. 4.8 – Cartographie des territoires cérébraux du langage chez un sujet normal.

Il décrit par exemple le trajet d'un mot entendu qui passe successivement par le cortex auditif puis l'aire de WERNICKE. Si celui-ci doit être prononcé, le message de l'aire de WERNICKE doit être envoyé à l'aire de BROCA puis au cortex moteur. L'aire de WERNICKE joue par ailleurs un rôle majeur dans les mécanismes de reconnaissance de la parole autant visuelle qu'auditive. Les quatre références cités ci-dessus sont des expérimentations récentes

qui confirment la prédominance de l'hémisphère gauche dans la traitement de la parole en mettant en avant, entre autres, une perception des transitions rapides associée à la parole localisée dans l'hémisphère gauche. De plus, l'hémisphère gauche serait doté d'une résolution temporelle plus précise qui permettrait de décoder les sons de parole alors que l'hémisphère droit serait pourvu une résolution spectrale meilleure.

Ainsi, l'hémisphère gauche ou l'oreille droite serait prédestinée à l'analyse de la parole. Ce point est particulièrement intéressant, et il nous amène à considérer que les conclusions tirées de cette expérience sont spécifiques aux signaux de parole, et qu'une expérience similaire pour l'évaluation de la qualité audio de signaux de musique pourrait conduire à des résultats différents.

Cependant, il semble que les auteurs s'accordent pour affirmer que la contribution des deux hémisphères est bien réelle pour le traitement de la parole mais que celle-ci est probablement différente pour chaque hémisphère. [Shtyrov, 2000] montre que la participation de l'hémisphère gauche décroît alors que celle de l'hémisphère droit s'accroît lorsque le signal de parole est présenté avec un bruit blanc. Il se pourrait que ce résultat explique l'allure des courbes de l'oreille gauche dégradée sur les figures 4.7a et b. Cette supposition serait à confirmer en réalisant une expérience similaire à la notre avec de la parole de fond sonore au lieu du MNRU.

4.2 Les taux d'erreurs de la session MNRU

Ces résultats ne sont que partiellement analysés ; nous ne pouvons les exposer dans ce mémoire.

4.3 Temps de réaction pour la session bande passante (BP)

4.3.1 Variabilité inter-individuelle

Comme pour le MNRU, on commence par regarder si un effet du sujet est présent. La figure suivante indique en effet une forte variabilité inter-individuelle. En moyenne, le TR moyen est de 0,162 ms. Mais l'on constate que certains sujets sont plus lents ou plus rapides que d'autres (TR max = 0,284 ms, TR du sujet B 12 et TR min = 0,072, TR du sujet A 3, cf. figure 4.9.

Pour éliminer cet effet indésirable, les données sont centrées et réduites dans la suite de l'analyse.

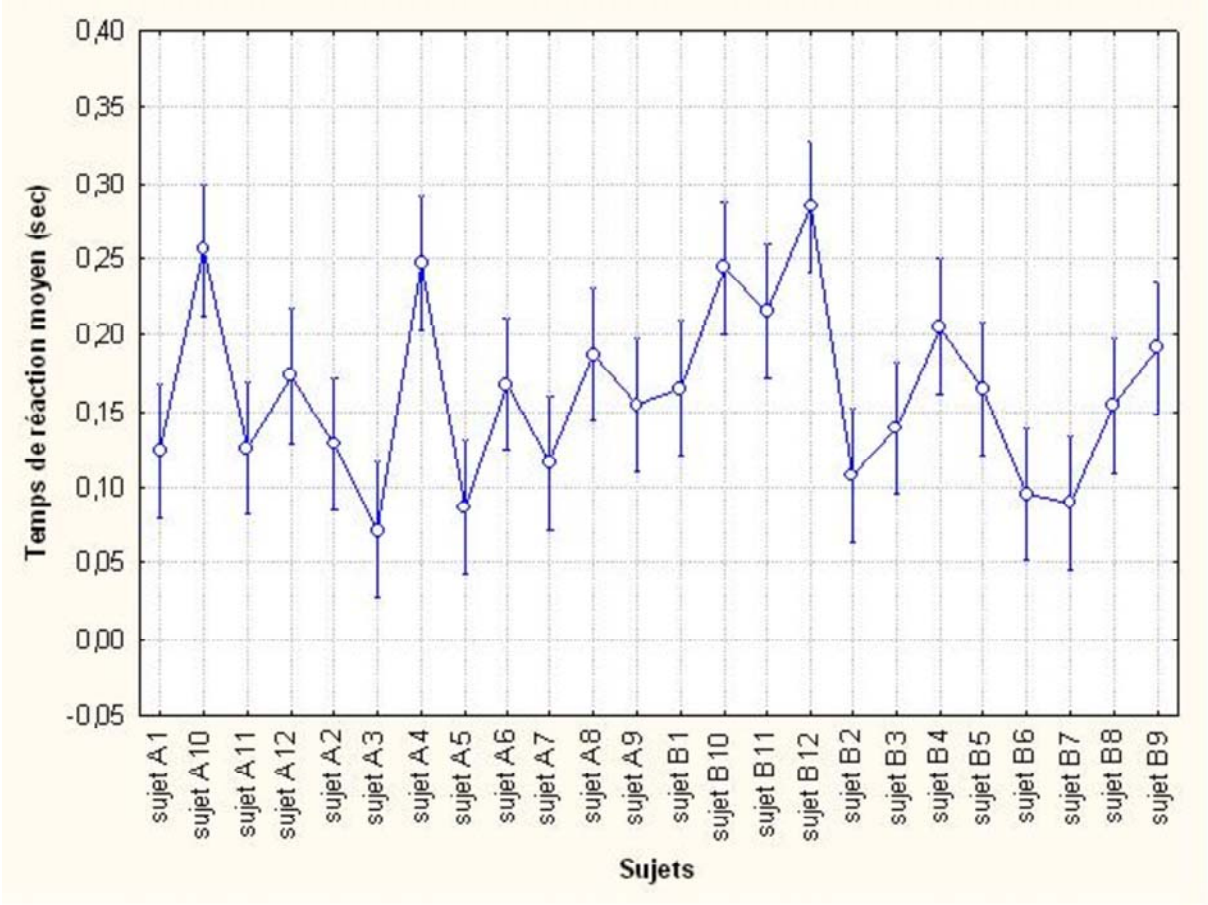

FIG. 4.9 – Temps de réaction moyens par sujet. Les barres représentent les intervalles de confiance à 0,95.

4.3.2 Effet de la réduction de la bande passante sur les temps de réaction

La figure 4.10 montre les TR moyens en fonction de la largeur de bande. Bien que le TR obtenu pour la bande téléphonique (300-3400 Hz) semble légèrement plus important que pour la condition HQ, l'effet de la dégradation BP n'est pas aussi évident que pour la dégradation MNRU.

Une ANOVA est réalisée sur les TR individuels en considérant les mêmes facteurs que ceux définis précédemment pour le MNRU. On les rappelle brièvement :

- Le facteur intra-groupe "Situation" à 2 niveaux :
  - **Asymétrique** : HQ/HQ, HQ/ BP 50-14000 Hz, HQ/ BP 50-7000 Hz, HQ/ BP 300-3400 Hz.
  - **Symétrique** : HQ/HQ, BP 50-14000 Hz / BP 50-14000 Hz, BP 50-7000 Hz/ BP 50-7000 Hz, BP 300-3400 Hz / BP 300-3400 Hz.

- Le facteur intra-groupe "Qualité" à 4 niveaux :

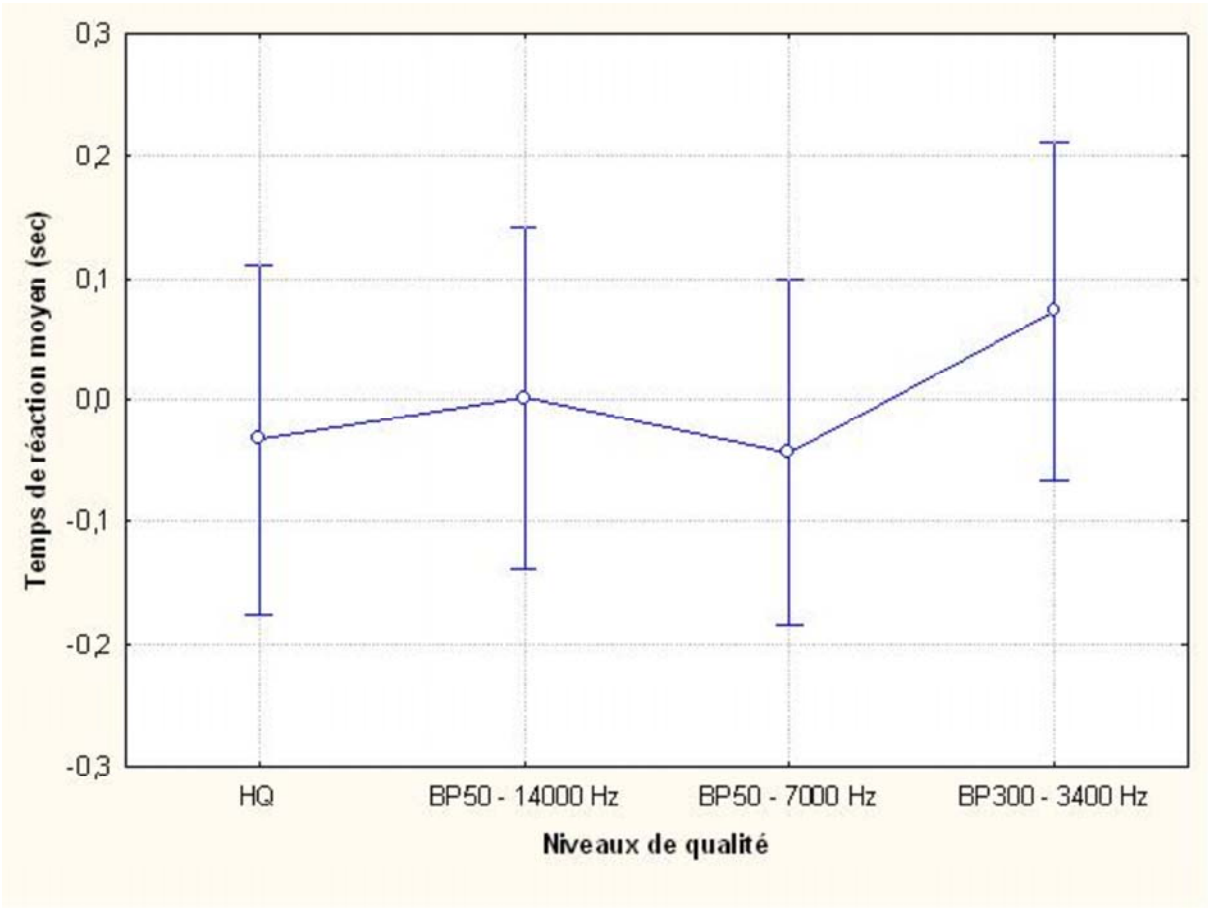

FIG. 4.10 – Effet de la qualité du signal sur les temps de réaction des sujets.

- sans dégradation (Haute Qualité),
- **BP 50-14000 Hz,**
- **BP 50-7000 Hz,**
- **BP 300-3400 Hz,**

- Le facteur intra-groupe "Oreille" à 2 niveaux : **Oreille 1, Oreille 2.**
- Le facteur Groupe à 2 niveaux : **Groupe A, Groupe B.**

Le tableau 4.2 suivant récapitule les résultats de l'ANOVA :

Facteur	SC	dl1	dl2	MC	F	p
1 GROUPE	0	1	238	0	0	1
2 SITUATION	0,88	1	238	0,88	0,97	0,326
SITUATION * GROUPE	1,02	1	238	1,02	1,12	0,292
3 QUALITE	7,94	3	714	2,65	3,03	,029*
QUALITE * GROUPE	2,38	3	714	0,79	0,91	0,437
4 OREILLE	6,22	1	238	6,22	5,99	,015*
OREILLE * GROUPE	33,62	1	238	33,62	32,36	,000*
SITUATION * QUALITE	2,09	3	714	0,7	0,75	0,523
SITUATION * QUALITE * GROUPE	3,91	3	714	1,3	1,4	0,242
SITUATION * OREILLE	15,69	1	238	15,69	16,16	,000*
SITUATION * OREILLE * GROUPE	0,64	1	238	0,64	0,66	0,416
QUALITE * OREILLE	8.02	3	714	2,67	3,14	,025*
QUALITE * OREILLE * GROUPE	3,58	3	714	1,19	1,4	0,241
SITUATION* QUALITE * OREILLE	5,69	3	714	1,9	2,32	0,074
2 * 3 * 4 * 1	1,95	3	714	0,65	0,8	0,497

TAB. 4.2 – Récapitulatif de l'analyse pour la session bande passante

L'ANOVA révèle un effet Qualité ($p=0,29$). Un HSD (Honestly Significant Difference) Tukey test montre cependant que seule la condition 300-3400 Hz est significativement supérieure à la condition BP 50-7000 Hz, ($p=0,03$).

Par ailleurs, on retrouve un effet de l'oreille sur les TR, certes moins marqué mais toujours significatif : F(1, 238)=5,991, $p=0,015$.

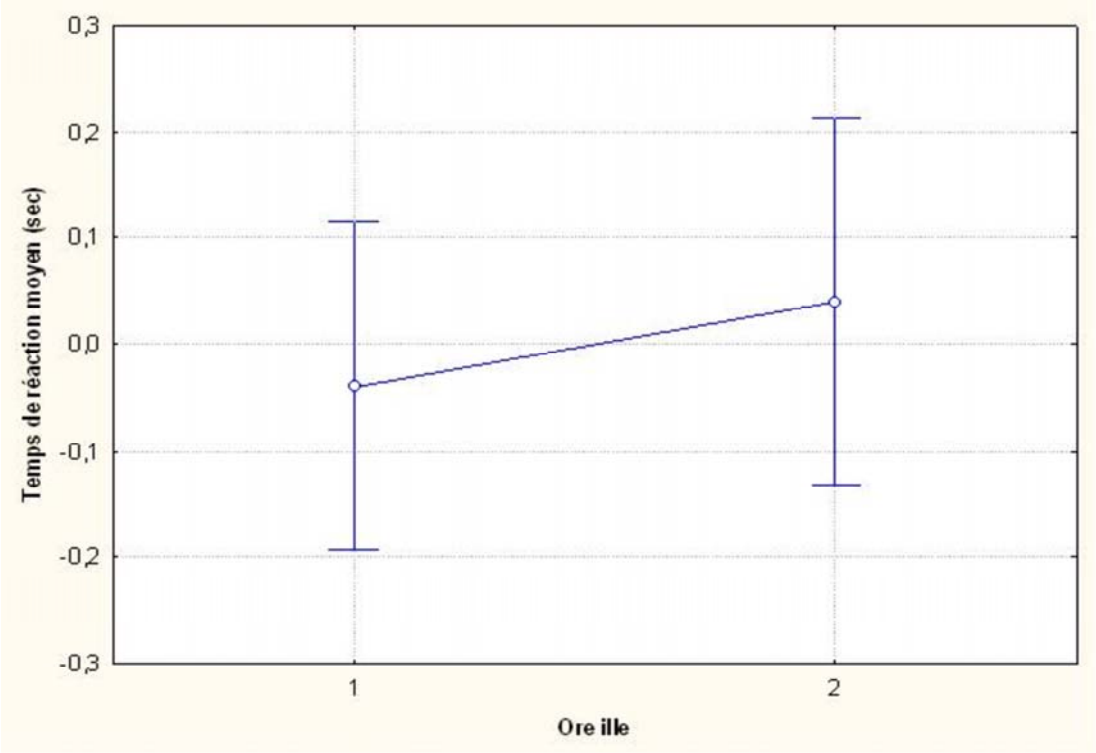

FIG. 4.11 – Effet de l'oreille sur les temps de réaction.

Les temps de réaction recueillis pour les mots apparaissant sur l'oreille dégradée dans toutes les conditions (oreille 2) sont significativement plus longs, en moyenne, que les TR recueillis pour les mots apparaissant sur l'oreille non dégradée pour la moitié des conditions (oreille 1).

Réapparaissent également sur la figure 4.12 (ci-dessous) les mêmes comportements selon les situations. Dans la situation symétrique, les deux oreilles sont à peu près équivalentes. Dans la situation asymétrique, la différence de comportement entre les deux oreilles se manifeste clairement ($F(1,238)=16,157$, $p=0,000$).

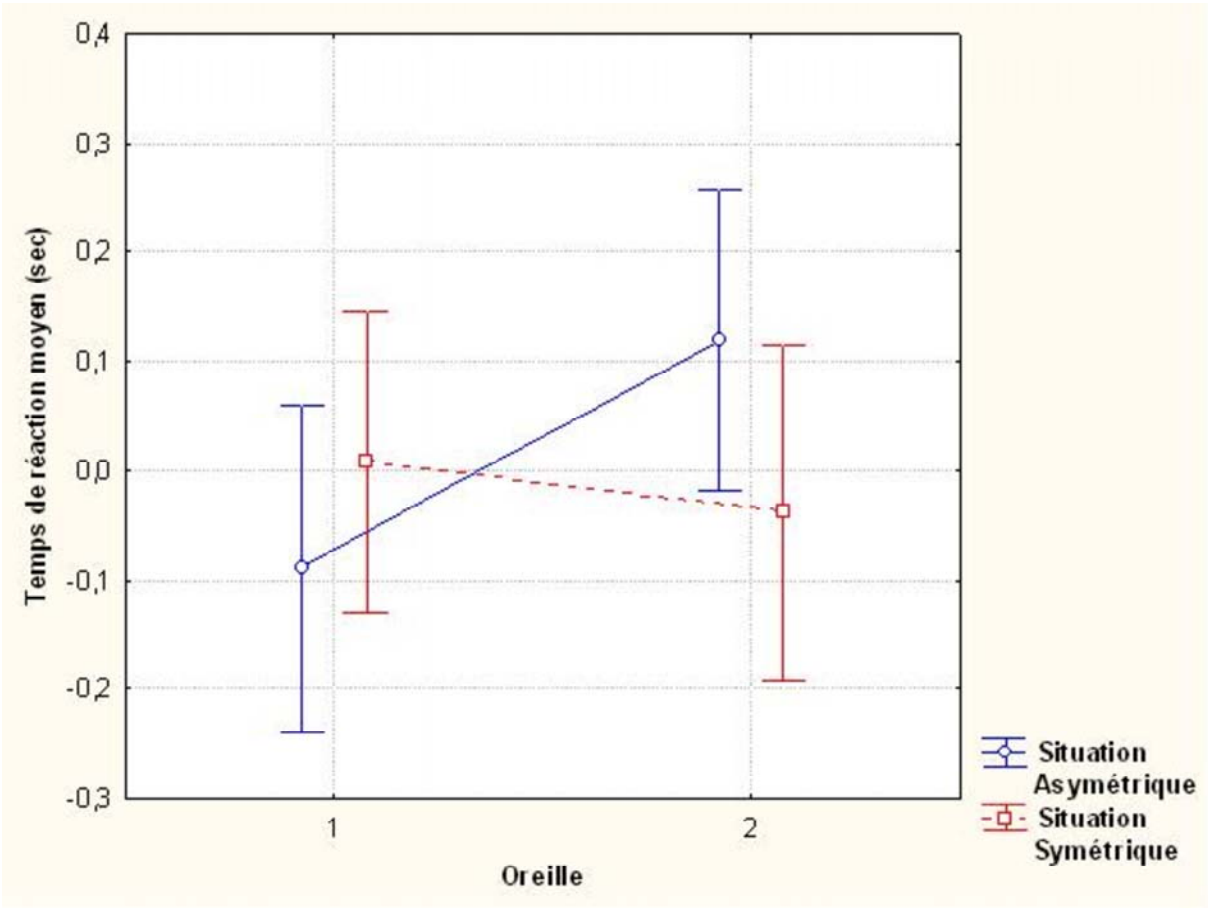

FIG. 4.12 – Interaction entre les facteurs situation et oreille.

Un HSD Test confirme que la différence significative entre les deux oreilles en situation asymétrique ($p<0,05$) disparaît en situation symétrique ($p=0,48$).

Ainsi, de la même façon que pour la dégradation MNRU, la présence de dégradations dans le signal de parole présentée à une oreille semble détériorer les performances, en termes de temps de réaction, de détection de mots cibles sur cette oreille.

La figure 4.13 montre l'effet combiné des trois facteurs Qualité, Oreille et Situation (interaction non significative : $F(3,714)=2,3203$, $p=0,074$). La courbe bleue en situation asymétrique (oreille non dégradée) est curieuse : au lieu d'être globalement horizontale, elle a tendance à chuter. Cependant, un HSD Tukey test révèle qu'aucune différence entre les quatre niveaux de qualité n'est significative ($p\geq 0,82$ pour toutes les comparaisons). Ceci signifie que les TR sur l'oreille non dégradée n'évoluent pas alors que ceux sur l'oreille dégradée augmentent dans l'ensemble. La notion d'indépendance déjà évoquée précédemment s'applique donc ici.

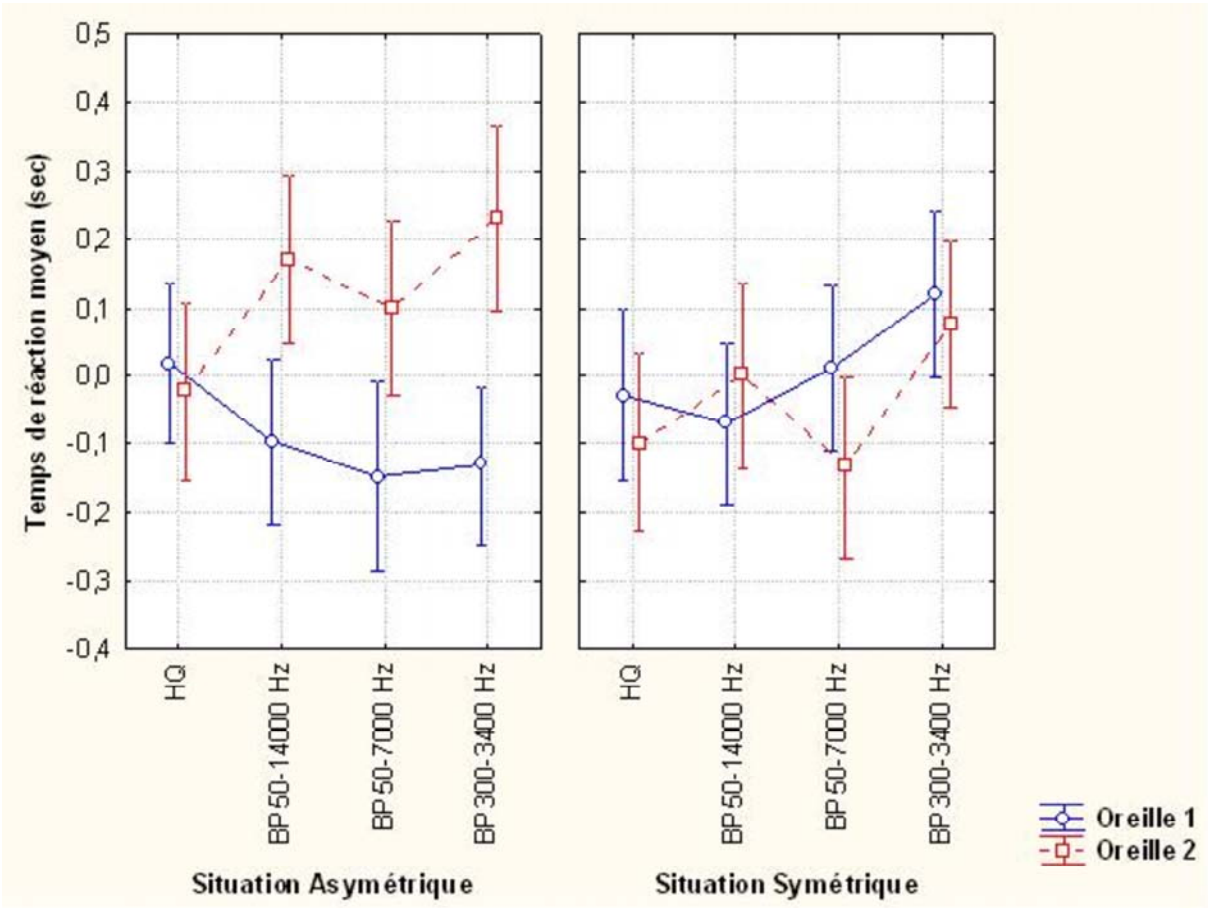

FIG. 4.13 – Interaction entre les facteurs situation, oreille et qualité.

De même que dans le cas du MNRU, l'ANOVA sur les données centrées réduites montre une interaction significative entre les facteurs Groupe et Oreille ($F(1, 238)=32{,}363$, $p=0{,}000$), (cf. figure 4.14).

Il semblerait que en moyenne, les performances en termes de TR de détection de mots cibles sont meilleures sur l'oreille droite que sur l'oreille gauche pour le groupe B, et non pour le groupe A. Ces résultats sont appuyés par un HSD Tukey Test : pour le groupe B, les performances de l'oreille droite sont significativement meilleures que celle de l'oreille gauche ($p<0{,}05$) alors que pour le groupe B, cette distinction s'efface ($p=0{,}1$).

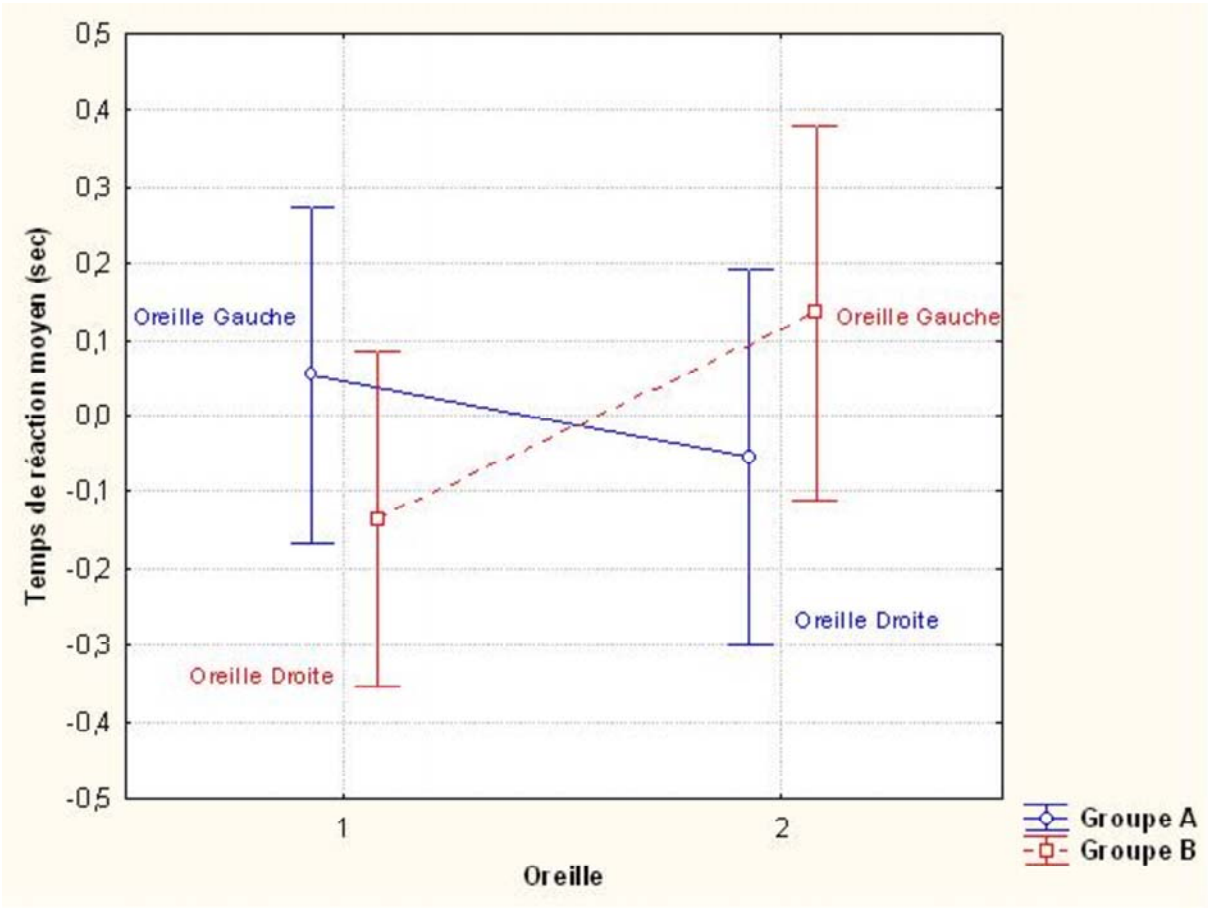

FIG. 4.14 – Interaction entre les facteurs groupe et oreille.

En effet, si l'on regarde les figures 4.15a et b, qui illustrent l'effet combiné entre les quatre facteurs Oreille, Situation, Groupe et Qualité, on constate qu'en situation symétrique les performances en termes de TR de détection de mots cibles sont meilleures sur l'oreille droite que sur l'oreille gauche, pour les deux groupes. En situation asymétrique, l'effet est renforcé pour le groupe B puisque l'on dégrade l'oreille gauche la moins performante.

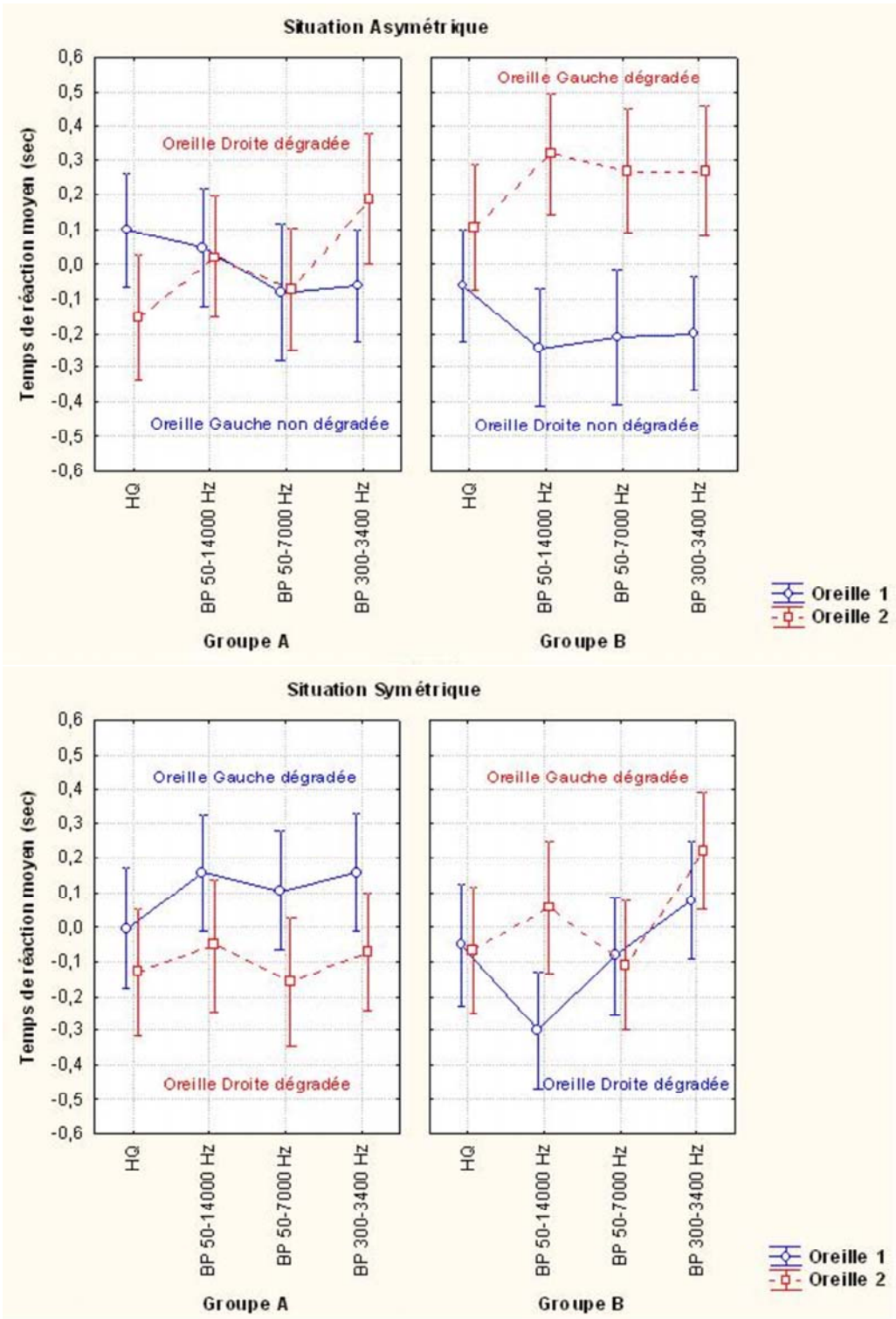

FIG. 4.15 – a) et b) Interaction entre les facteurs groupe, oreille, qualité et situation.

4.3.3 Discussion

Globalement, les conclusions de la dégradation par la réduction de la bande passante sont sensiblement les mêmes que celles obtenues avec la dégradation MNRU. Bien que moins marqué, on retrouve un effet global de la qualité : plus la qualité se dégrade, plus le ou les temps de réaction de ou des oreilles dégradées augmentent. Cet effet semble surtout du à la présence de la condition bande téléphonique, comparée aux autres conditions. Il est aussi probablement possible que notre oreille soit plus sensible (dans le sens plus perturbée) au bruit qu'à une réduction de bande passante. Nous écoutons d'ailleurs quotidiennement des sons dont la bande est réduite que ce soit par l'intermédiaire du téléphone, d'Internet, de codeurs...Cette remarque pourrait expliquer les différences rencontrées dans les forces des effets.

Enfin, on retrouve l'effet d'indépendance d'une oreille par rapport à l'autre ainsi que la prédominance de l'oreille droite sur l'oreille gauche pour la détection de mots cibles.

4.4 Les taux d'erreurs de la session de la bande passante

De même que pour le MNRU, ces résultats ne sont que partiellement analysés ; nous ne pouvons les exposer dans ce mémoire. Signalons de même que les analyses des signaux physiologiques ne sont pas achevées.

Conclusion

Guidés par les expérimentations antérieures, nous souhaitions étudier la possibilité de mesurer la qualité vocale à travers des critères comportementaux et électrophysiologiques. Plus précisément, notre travail a cherché à vérifier l'hypothèse selon laquelle les performances diminuent (*i.e.* les temps de réaction s'accroissent et / ou les erreurs se multiplient), quand la qualité se dégrade. Pour confirmer ou infirmer cette hypothèse, nous avons mené une expérimentation basée sur une tâche de détection de mots cibles en écoute dichotique avec l'introduction de dégradations, à différents degrés, sur l'une ou les deux oreilles.

L'analyse des premiers résultats permet de valider l'hypothèse proposée. Dans la mesure de dégradations suffisamment prononcées (comme l'introduction de bruit avec des rapports signal/bruit relativement importants), il en résulte un allongement significatif des temps de réaction. Avec la dégradation de type réduction de bande passante, cet effet, bien que moins marqué, est toujours présent. De ce point de vue, les conclusions du test sont encourageantes. Toutefois, si l'on obtient un allongement des temps de réaction en présence d'un signal dégradé par rapport à un signal non dégradé, cet allongement ne reflète pas encore les différents niveaux de dégradations introduites dans les signaux de parole. La mesure des temps de réaction dans une tâche de détection de mots cibles en écoute dichotique ne nous permet donc pas encore une mesure quantitative du niveau de dégradation, objectif de ce travail.

Les premiers résultats de cette expérience suggèrent de même une prédominance de l'oreille droite dans la détection de mots cibles en écoute dichotique, dont il faudrait tenir compte à l'avenir, dans des protocoles similaires.

Des analyses supplémentaires sont encore à mener, d'une part sur les taux d'erreurs commises par les sujets, d'autre part sur les signaux électrophysiologiques et à mettre en relation avec ces premiers résultats. Ce travail amorcera la thèse qui poursuit cette étude.

En conclusion, cette première expérience explorant la possibilité de mesurer la qualité vocale "indirectement" (non plus à travers des jugements de qualité émis par les sujets, mais à travers l'impact de la qualité vocale des signaux de parole mis en jeu dans une tâche spécifique sur les performances des sujets) nous encourage à poursuivre dans cette voie, avec des tâches mettant en jeu d'autres processus cognitifs (reconnaissance, mémoire) peut-être plus sensibles aux dégradations d'un signal de parole.

Bibliographie

[Bases neurologiques du langage] http ://schwann.free.fr/Langage01.html,
Bases neurologiques du langage, par le Pr. J. Lehouelleur

[Cheminée, 2004] Cheminée,
P. Vous avez dit "clair" ?
Communication aux 2èmes journées du Sensolier, Paris, Octobre 2004

[Fortin et Rousseau, 1993] Fortin, C., Rousseau, R.
Psychologie cognitive, une approche de traitement de l'information
pp74-87, Ed. Université du Québec Télé-université, 1993

[Gros, 2001] Gros, L.
Evaluation subjective de la qualité vocale fluctuante
Thèse, France Télécom & Université de la Méditerranée Aix Marseille II, 2001

[Gros et al., 2005] Gros, L., Château, N., Macé, A.
Assessing speech quality : a new approach
Forum Acusticum, Septembre, 29th-4th, 2005

[Hickok et Poeppel, 2000] Hickok, G., Poeppel, D.
Towards a functional neuroanatomy of a speech perception
TRENDS in Cognitive Sciences, Vol.4, No.4, April 2000

[Lai et al., 2001] Lai, J., Cheng, K., Green, P., Tsimhoni, O.
On the road and on the Web ? Comprehension of synthetic and human speech while driving
Proceedings of SIGCHI, Conference on human factors in computing systems, 2001

[Landström et al., 2002] Landström, U., Söderberg, L., Kjellberg, A., Nordström, B.
Annoyance and performance effects of nearby speech
Acta Acustica united with Acustica, vol.88 pp 549-553, 2002

[Mullin et al., 2001] Mullin, J., Smallwood, L., Watson, A., Wilson, G. M.
New techniques for assessing audio and video quality in real-time interactive communications,
In : J. Vanderdonckt, A. - Blandford & A. Derycke (eds)
Proceedings of IHM HCI, Lille, France, September, 10th - 14th, 2001

[Mullin et al., 2002] Mullin, J., Jackson, M., Anderson, A. H., Smallwood, L., Sasse, M. A., Watson, A., Wilson, G. M.
Assessment methods for assessing audio and video quality in real-time interactive communications
The ETNA Taxonomy February 2002
http ://wwwmice.cs.ucl.ac.uk/multimedia/projects/etna/taxonomy.pdf

[Pachiaudi, 2002] Pachiaudi, G.
Analyse des risques de l'utilisation du téléphone mobile en situation de conduite
Article ATEC, Février 2002

[Pashler, 1994] Pashler, H.
Dual-task interference in simple tasks : data and theory
Psychological Bulletin, 116, pp 220-244

[Shtyrov, 2000] Shtyrov, Y.
New aspects of the cerebral functional asymmetry in speech processing as revealed by auditory cortex evoked magnetic fields
Doctoral dissertation, University of Helsinki, Faculty of Arts, November 2000

[Shtyrov et al., 2000a] Shtyrov, Y., Kujala, T., Lyytinen, H., Kujala, J., Ilmoniemi, R.J., and Näätänen, R.
Lateralization of speech processing in the brain as indicated by mismatch negativity and dichotic listening
Brain and Cognition 43, pp 392-398, 2000

[Shtyrov et al., 2000b] Shtyrov, Y., Palva, S., Ilmoniemi, R.J., Näätänen, R.
Discrimination of speech and of complex nonspeech sounds of different temporal structure in the left and right cerebral hemispheres
NeuroImage 12, pp 657-663, 2000

[Sonntag et al., 1998] Sonntag, G. P., Portele, T., Haas, F.
Comparing the comprehensibility of different synthetic voices in dual task experiment
In : Proc. 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves,
Blue Mountains, Australia, November, 26th-29th, 1998

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

[UIT-T P.810, 1996] UIT-T Recommandation P. 810
Appareil de référence à bruit modulé, 1996

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

[UIT-T P.832, 2000] UIT-T Recommandation P. 832
Evaluation subjective des performances des terminaux mains-libres, 2000

[Wilson et Sasse, 2001] Wilson, G. M., Sasse, M. A.
Straight from the heart : Using physiological measurements in the evaluation of media quality
Proceedings of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour Convention, Symposium on Emotion, Cognition and Affective Computing, March 21st- 24th 2001, York, UK, pp 63-73, ISBN : 1-902956-19-7

[Wilson et Sasse, 2000] Wilson, G. M., Sasse, M. A.
Investigating the impact of audio degradations on users : Subjective vs. objective assessment methods
Proceedings of OZCHI, Sydney, Australia, December 4th- 8th, 2000

[Zatorre et al., 2002] Zatorre, R.J, Belin, P., Penhune, V.B.
Structure and function of auditory cortex : music and speech
TRENDS in Cognitive Sciences, Vol.6, No.1, January 2002

Annexes

Annexe 1

Cette annexe fournit les instructions données aux sujets lors du test.

Instructions :

Vous allez effectuer un test dans lequel votre tâche est de détecter des mots dans des listes présentées à vos oreilles. A chaque fois, les mots à détecter sont différents et la qualité sonore peut varier.

Pour chaque écoute, vous entendrez simultanément 2 listes de mots, une sur chaque oreille. Sur l'écran de l'ordinateur s'afficheront les mots que vous devrez détecter dans la liste écoutée. Pour cela, vous appuierez sur :

1 si vous avez entendu le mot à détecter à **gauche**
ou 2 si vous avez entendu le mot à détecter à **droite**

dans le temps qui vous est imparti.

Si vous ne répondez pas, le mot suivant apparaîtra automatiquement à l'écran. Si, lors d'une écoute, vous appuyez par erreur sur une autre touche que 1 ou 2, une alarme rouge vous en avertira et votre première réaction ne sera pas prise en compte.

Chaque écoute est précédée de la mention "appuyer sur une touche pour commencer".

Le but de ce test est d'être le plus performant, c'est-à-dire, de réagir **le plus rapidement possible** et de commettre **le moins d'erreurs possible.** Essayez donc de rester bien concentré tout au long de l'écoute.

Pour vous familiariser à ce test, une session d'apprentissage vous est proposée avant de commencer.

Merci de votre participation.

Prêt... ? C'est parti !

Annexe 2

Nous présentons dans cette annexe les listes de mots ainsi que les contraintes liées à celles-ci. Cette annexe comprend un cd.

Les cent mots utilisés pour créer les listes sont les suivants :

auto
accroc
revue
bévue
ballet
galet
jonction
appui
kiwi
bandit
plaisir
soupir
frisson
moisson
cristal
jovial
accent
argent
fureur
ardeur
balcon
façon
virus
cactus
vitrail
corail
bijou
hibou
époux
ego
vélo
menteur

meneur
papier
berger
pluriel
fusain
fuseau
autel
rappel
pompier
pommier
citron
siphon
miroir
peignoir
voisin
raisin
poumon
prénom
bateau
badaud
bidon
vison
capot
chapeau
lapsus
lapin
vaccin
taquin
tourment
tournant
bajoue
bagout
appeau
budget
suspect
anchois
effroi
maison
raison
poisson
potion
avion
camion
ballon
crayon
tableau

râteau
tapis
partie
persil
babil
casier
panier
levier
segment
parent
fonction
version
moteur
tracteur
bâtis
repli
relief
grief
client
patient
moignon
oignon

Comme un sujet entend huit listes test au cours d'une session, il est prudent d'élaborer des listes suffisamment variées afin que le sujet ne développe pas des attentes ou des points de repères. En effet, on souhaite éviter deux écueils : une alternance gauche droite des mots trop régulière, qui inciterait le sujet à se dire, "attention, le mot suivant va apparaître sur mon oreille gauche" ; et un repère tel que "le mot maison est toujours sur mon oreille droite et après le mot chapeau". On construit alors les listes selon le schéma suivant : parmi l'ensemble de 100 mots, on tire au hasard deux paquets de cinquante mots. Les mots du paquet 1 seront spatialisés à gauche ; ceux du paquet 2 à droite. Puisque pour les huit listes test d'un sujet, on tire à nouveau ces deux paquets, l'ordre des mots et leur spatialisation sont toujours différents, et il devient impossible au sujet de savoir quel mot se trouve sur quelle oreille et s'il se trouve après tel ou tel mot. Une fois les mots spatialisés à droite ou à gauche, on ajoute à chaque mot une pause aléatoire entre 50 et 200 ms. On se retrouve alors avec deux listes sur chaque oreille de 50 mots séparés par des pauses. On décale une des deux listes d'une pause aléatoire de 50 ms et 1 s, ainsi, en rassemblant les deux listes en un seul fichier son, on obtient la liste finale d'une écoute dont l'agencement des mots est variable (par exemple droite/ gauche / gauche /droite/ etc.).

Ci-joint, un cd présentant les listes et les trois niveaux de dégradations pour le MNRU et la bande passante.
La sélection des mots cibles doit, elle aussi, satisfaire à des contraintes : d'une part, on doit avoir dix mots cibles sur l'oreille droite et dix mots cibles sur l'oreille gauche. D'autre part, les mots cibles remplissent la condition suivante : la différence entre le début du mot cible i et la fin du mot cible i-1 doit être strictement supérieure à 800ms. (Par sécurité, on prend 850 ms.)

On procède alors de la manière suivante : on crée une permutation de 1 et de 2, de longueur vingt, comprenant dix 1 et dix 2. Cette permutation indique sur quelle oreille se trouvent les vingt mots cibles. Par ailleurs, on pioche des mots cibles dans la liste de cent mots que l'on a créée et on regarde s'ils vérifient les contraintes de temps et "d'oreille". La contrainte "d'oreille" est remplie si l'oreille sur laquelle se trouve le Xième mot cible choisi dans la liste corespond à l'indice X de la permutation. Si oui, on pioche le mot cible suivant, sinon, on pioche un autre mot cible et s'il n'y a pas d'issue, on modifie la permutation.
Prenons l'exemple suivant avec quatre mots cibles répartis sur les oreilles gauche (1) et droite(2) comme suit : [1 2 2 1], deux sont à droite et deux sont à gauche. Considérons que notre liste contient dix mots (on note à coté leur position sur les oreilles ; cinq sont à gauche et cinq sont à droite) : maison 1 ; chapeau 1 ; ballon 2 ; raison 2 ; rateau 2 ; vélo 1 ; partie 1 ; pompier 1 ; voisin 2 ; poumon 2 ;

On tire par exemple chapeau comme premier mot cible. Il convient puisqu'il est en 1 (*i.e.* sur l'oreille gauche, comme le demande le premier indice de la permutation [1 2 2 1]). On tire ensuite raison. Il convient lui aussi en considérant la contrainte de temps satisfaite. Ensuite il faut piocher un mot sur l'oreille 2. Si par exemple, rateau est exclu par la contrainte de temps et qu'on ne peut choisir le troisième mot cible au delà de pompier, nous nous apercevons qu'aucun mot de la liste ne satisfait aux deux conditions pour être mot cible. On modifie alors la permutation en [1 2 1 2], et alors on a chapeau, raison, partie et poumon comme mots cibles.

Annexe 3

FIG. 16 – Sujet passant le test.

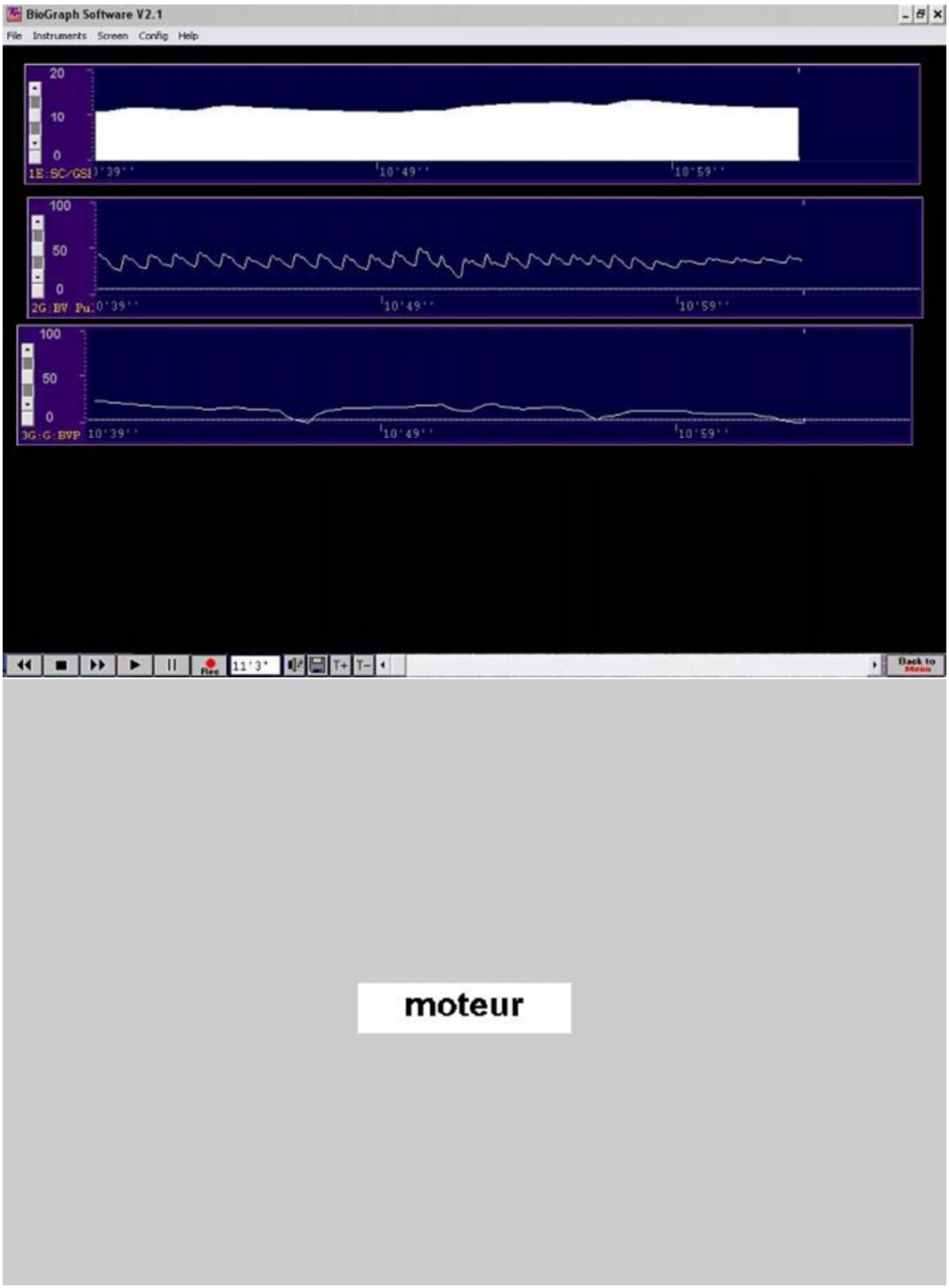

FIG. 17 – a) et b) Impressions d'écran des deux pc durant un test.