École Nationale Supérieure Louis-Lumière

Mémoire de fin d'études

Usage de l'analyse IDS pour une meilleure compréhension de la parole dans le bruit

Titouan Ralle

Promotion Son 2017

Directeur interne : Laurent Millot

Directeur externe : Bernard Auriol

Rapporteur : Gérard Pelé

Louis Lumière
école nationale supérieure

Résumé

Dans ce mémoire nous présentons une nouvelle approche pour améliorer la compréhension de la parole dans le bruit qui s'appuie sur le principe de l'analyse IDS. Nous expliquons tout d'abord comment des découpages fréquentiels adaptés à la voix française ont été obtenus à partir de corpus oraux préalablement constitués, puis nous décrivons comment ces découpages ont été utilisés pour mener une campagne de tests perceptifs grâce à l'« IDS *Speech Enhancer* » (IDSSE) élaboré à cette occasion. C'est à partir des résultats de ces tests que nous obtenons les filtres à réaliser pour améliorer la compréhension de la parole dans une ambiance bruitée. Comme ces traitements sont destinés à une utilisation quotidienne nous examinons, pour finir, les ressources nécessaires au fonctionnement de ces filtres en temps réel.

Mots clés : compréhension de la parole dans le bruit, analyse IDS, découpage fréquentiel, convolution en temps réel.

Abstract

In this research paper we present an alternative approach to improve speech recognition in noise based on IDS analysis. First we explain how « French-voice-adapted » frequency mappings have been obtained from spoken corpus built up beforehand, then we describe how these frenquency mappings were used to conduct a test campaign with the « IDS Speech Enhancer » (IDSSE) designed on this occasion. Using the tests results we can develop the filters improving speech recognition in a noisy environement. As these processings are designed on a daily-basis use, we examine finally the conditions to perform these processings in real time.

Key words : speech recognition in noise, IDS analysis, frequency mapping, real-time convolution.

Remerciements

Je remercie tout d'abord mes deux directeurs de mémoire, Bernard Auriol et Laurent Millot, pour m'avoir apporter leur soutien ainsi que de précieux conseils durant l'élaboration de ce travail.

Je tiens aussi à remercier Alan Blum, Mohammed Elliq, Jean Rouchouse, Pascal Spitz et Eric Urbain qui m'ont permis d'avoir accès au matériel et aux installations de l'ENS Louis-Lumière.

Pour finir, merci à toutes celles et à tous ceux qui ont participé aux enregistrements ainsi qu'aux tests perceptifs indispensables à la réalisation de ce mémoire.

Table des matières

Table des matières **5**

1 Introduction **6**

1.1 Comprendre la parole dans le bruit 6

1.2 Démarche 20

2 Constitution de corpus oraux représentatifs de la voix française **24**

2.1 Pourquoi constituer des corpus originaux ? 24

2.2 Composition des corpus 27

2.3 Enregistrement des corpus 34

2.4 Conclusion 41

3 IDS et découpages fréquentiels **42**

3.1 Description de l'IDS 42

3.2 Exemples de découpages fréquentiels existants 49

3.3 Méthode pour la détermination de découpages fréquentiels 52

3.4 Résultats pour les corpus étudiés 58

3.5 Conclusion 60

4 L'« IDS *Speech Enhancer* » (IDSSE) et tests perceptifs **61**

4.1 Présentation générale du test 61

4.2 Description de l'IDS *Speech Enhancer* 64

4.3 Résultats des premiers tests 76

4.4 Amendements et perspectives 79

4.5 Conclusion 85

5 Utilisation des traitements en temps réel **87**

5.1 Convolution en temps réel 88

5.2 Analyse IDS en temps réel 93

5.3 Conclusion 95

6 Conclusion générale **96**

Bibliographie **97**

Annexe 1 : Contenu des corpus oraux **103**

Corpus de voix parlées 103

Corpus de voix télévisuelles 107

Annexe 2 : Tests perceptifs **108**

Résultats généraux 108

Résultats pour les malentendants 118

Patch *Pure Data* de l'IDSSE 120

Annexe 3 : Spécifications techniques du matériel utilisé **121**

Microphones 121

Contrôleur MIDI 123

Audiomètre 124

Annexe 4 : Guide pour l'utilisation de l'IDSSE **125**

1 Introduction

1.1 Comprendre la parole dans le bruit

1.1.1 Préambule

L'ouïe est un sens d'une grande précision chez l'être humain. Elle offre la possibilité de percevoir un monde sonore remarquablement complexe. Cependant cette évidence est souvent négligée car nous entendons en permanence. La richesse de notre audition est devenue une habitude, et comme l'affirme Samuel Beckett : « L'habitude est une grande sourdine. » [1].

Il n'y a que certaines situations où le rôle de l'ouïe est explicitement indispensable pour nous en rappeler sa valeur : écoute d'une musique, de la radio, d'une annonce à travers un haut-parleur, d'un discours, d'un murmure... Dans ces situations nous prêtons attention à ce que nous entendons, nous *tendons l'oreille*[1], **nous écoutons**.

Les merveilles de l'audition sont donc reconnues lorsque l'on écoute, puis aussitôt oubliées lorsque l'on ne fait « qu'entendre ». Néanmoins dans cette attitude passive consistant à percevoir par l'oreille, nous recevons en fait une abondance d'informations très précises.

1. À l'origine, l'homme pouvait orienter son oreille pour diriger son audition et ainsi mieux se protéger des dangers environnants, mais il a perdu cette faculté au fil de l'évolution (la persistance d'une structure vestigiale à l'arrière de l'oreille prouve cet état antérieur [2]). Certains mammifères ont toujours cette capacité, comme les chiens par exemple.

L'audition effectue sans cesse des enquêtes qui nous permettent de distinguer des sons multiples, simultanés ou successifs, d'assigner leur provenance à des sources sonores différentes, d'évaluer les positions, voire les dimensions de ces sources, d'inférer les modes de production sonore... Un tour de force si l'on songe que chaque oreille ne reçoit qu'une variation de pression, une information bien mince. Le psychologue Albert Bregman propose une analogie éclairante : observons deux bouchons au bord d'un lac, mus par les ondes se propageant à la surface. Ne serait-ce pas un exploit d'en déduire les positions et les mouvements des poissons ou des autres êtres subaquatiques qui sont à la source de ces ondes ? [3, préface]

Prenons la situation d'un verre se brisant contre le sol. Le son résultant de cet événement nous indique qu'il est question d'un verre à pied assez fragile, ou bien d'un verre plus robuste ; qu'il est tombé depuis une table basse, ou qu'il a été jeté violemment ; qu'il a heurté un sol en carrelage, en parquet, en béton ; que l'incident a eu lieu dans une cuisine, un réfectoire, ou bien encore dans une église. Et toute cette analyse est réalisée **dans l'instant** sans que nous y prêtions attention, sans être forcément en situation d'écoute.

Ces informations auditives que nous filtrons en permanence sont fondamentales dans notre perception du monde, notamment dans notre perception de l'espace :

Contrairement à ce que l'on pense, l'espace est très marqué de traits acoustiques. Nous vivons dans un monde acoustique. Nous apprécions la distance des objets, la profondeur du champ perceptif, le volume de l'environnement à l'oreille. L'œil apporte la perspective, l'oreille la direction, la profondeur. [4]

Parmi tous les types de signaux sonores que nous captons constamment, il y en a un auquel nous accordons un intérêt particulier : **la parole**.

La voix parlée reste en effet le moyen de communication le plus utilisé chez l'être humain, et certainement l'un des plus subtils. En plus de véhiculer une information de sens qui serait une application logique et objective du langage, la parole est aussi vecteur d'émotions. Toutes les nuances sonores présentent dans la voix lui font dire beaucoup plus que ce qu'elle n'exprime : une même phrase énoncée avec une intonation différente sera perçue différemment par l'auditeur. Il est d'ailleurs possible de déceler l'humeur d'une personne en entendant sa voix, ou encore d'en contester la sincérité selon le ton qu'elle va prendre. « La voix ne trompe point même si les paroles trompent. », écrivait André Suarès [5].

Ici encore on ne s'étonne plus de la facilité et de la finesse avec laquelle nous sommes capables de communiquer oralement. On ne réalise pas vraiment la puissance des mécanismes mis en œuvre nous permettant de capter et d'analyser, en temps réel, des messages sonores complexes provenant d'un autre individu.

> « *You have probably never thought about it this way, but every time you talk to someone, you are effectively engaging in something that can only be described as a telepathic activity, as you are effectively "beaming your thoughts into the other person's head," using as your medium a form of "invisible vibrations."* » [2] [6, p. 1]

Même si ce rapprochement entre l'audition et la télépathie paraît étonant, il a le mérite de nous rappeler la commodité de la communication orale que nous tenons naturellement pour acquise. D'ailleurs c'est souvent lorsque l'on perd cette capacité que l'on en découvre l'importance.

Car si l'ouïe est extrêmement précise, **elle est aussi très fragile**.

2. [Traduction] Vous n'y avez peut-être jamais pensé de cette façon, mais à chaque fois que vous parlez à quelqu'un, vous procédez en fait à quelque chose qui pourrait être décrit comme une pratique télépathique, puisque vous « transmettez vos pensées dans la tête de l'autre personne », sous la forme de « vibrations invisibles ».

1.1.2 Vers une hausse des troubles auditifs

En vieillissant l'être humain subit une perte progressive de l'audition plus ou moins marquée selon les individus : la presbyacousie. La plupart du temps elle se manifeste d'abord dans les aigus. Si cette détérioration commence très tôt (dès la fin de la croissance) elle devient en générale sensible et donc gênante à partir de 50 ans.

Mais nos nouvelles pratiques d'écoute ainsi que l'évolution de notre environnement sonore pourraient accélérer cette dégénérescence et poser un problème sanitaire aux prochaines générations.

Une enquête *JNA* (Journée Nationale de l'Audition) - *Ifop* (Institut Français d'Opinion Publique) [7], publiée en mars 2017 et « réalisée sur un panel de 1 202 individus âgés de 15 ans et plus », indique que 54 % des 15-17 ans s'endorment en écoutant de la musique via des écouteurs[3]. **91 % d'entre eux écoutent de la musique dans les transports publics**, souvent à des niveaux plus élevés compte tenu du bruit important dans ces lieux. Enfin ils sont 71 % à écouter de la musique plus d'une heure par jour sur leur smartphone et encore une fois via des écouteurs. Si beaucoup comprennent le danger d'une exposition à un fort niveau sonore, peu connaissent les risques d'une exposition prolongée même à un niveau raisonnable.

Parallèlement, en France, l'espérance de vie ne cesse de croître et la population continue de vieillir. L'*INSEE* (Institut National de la Statistique et des Études Économiques) prévoit qu'en 2050 **un habitant sur trois** sera âgé de 60 ans ou plus [8].

Il est donc fort probable que le nombre de personnes mal-entendantes augmente dans les années à venir.

3. Ce moyen de diffusion est particulièrement dangereux car il implique l'introduction d'oreillettes directement dans le creux du conduit auditif.

Différentes instances, gouvernementales ou non, ont bien compris ce problème. Nous avons déjà cité l'association *JNA*, mais nous pouvons aussi évoquer *La Semaine du Son*[4] qui multiplie chaque année les campagnes de prévention auditive, notamment chez les plus jeunes.

Dans un même temps, la législation française liée à la prévention et la répression des nuisances sonores évolue. En 2013, une loi limitant la puissance sonore maximale de « tout appareil portable permettant l'écoute de sons par l'intermédiaire d'un dispositif d'écoute » est entrée en vigueur [10].

Des changements commencent aussi à voir le jour concernant l'accès aux soins pour les mal-entendants.

La question du remboursement des audioprothèses par la Sécurité Sociale est en effet au cœur de nombreux débats et a même été évoquée durant la campagne présidentielle de 2017.

Actuellement en France, le prix moyen des prothèses auditives est comparable à celui pratiqué dans les autres pays de l'Union Européenne mais les prises en charge par l'Assurance Maladie et les complémentaires santé sont beaucoup plus faibles [11]. Par conséquent, les restes à charge devant être payés par les patients sont encore très importants (environ 1000 euros par oreille) [12].

Néanmoins si des progrès sociaux ne semblent s'engager que depuis peu quant à l'accès aux audioprothèses, la technologie de ces appareils a, elle, considérablement évolué ces dix dernières années.

4. « La Semaine du Son sensibilise le public et tous les acteurs de la société à l'importance des sons et de la qualité de notre environnement sonore. » [9]

1.1.3 Aides auditives : des performances nuancées

Avec l'arrivée du numérique, il est devenu possible d'embarquer des microprocesseurs réalisant des calculs de plus en plus puissants dans des prothèses de plus en plus petites.

Cette miniaturisation est d'ailleurs un argument très utilisé pour motiver les individus présentant des déficiences auditives à s'appareiller. Même si comme nous l'avons vu, il est tout à fait normal de moins bien entendre avec l'âge, certaines personnes refusent de montrer aux autres cette défaillance. Contrairement au port de lunettes qui est largement démocratisé, le port d'appareils auditifs reste encore légèrement tabou.

Nous comprenons donc facilement pourquoi la communication et la publicité autour du secteur de l'audioprothèse a explosé ces dernières années, à coups d'appareils « 100 % invisibles » de « qualité sonore optimale » pour « entendre parfaitement en toute discrétion ».

Bien qu'il soit indéniable que de grands progrès aient été faits sur les appareils auditifs depuis dix ans, il faut tout de même faire attention à ne pas confondre innovation technologique et innovation marketing.

Aujourd'hui les fabricants d'audioprothèses garantissent tous une qualité de son et de compréhension quasi-parfaite, obtenue à l'aide d'une panoplie de nouveaux traitements qui offriraient presque la possibilité de ré-entendre comme avant. Or, dans la pratique, les performances de ces appareils sont beaucoup plus nuancées.

Dans son mémoire présenté en 2013 pour l'obtention du diplôme d'État d'audioprothésiste, Jérémy Ecalard expose les résultats qu'il a obtenus suite à une série de tests effectués avec l'audioprothèse *Micon* qui était alors la nouvelle génération d'aide auditive du groupe *Siemens* :

> *La réduction offre un confort au sujet présentant un déficit auditif* [...]. *Cependant, même si le confort auditif est un paramètre important pour les malentendants en milieu bruyant, la compréhension est d'autant plus recherchée. Nos mesures ont permis de montrer un apport de l'ordre de 2-3 dB*[5] *de rapport signal sur bruit en moyenne, qui se traduit par quelques points d'intelligibilité supplémentaires.* [13, p. 89]

Nous sommes ici très loin de l'annonce de certains revendeurs, qui nous assurent fièrement « une qualité sonore si exceptionnelle que vos patients oublieront qu'ils portent des aides auditives. » [14].

Et bien que Jérémy Ecalard fasse lui même la critique de son protocole de test (qu'il n'a pas pu optimiser faute de temps et de moyens), on retrouve des résultats similaires dans des enquêtes à plus grande échelle.

Une enquête de *Que Choisir* publiée en 2015 et réalisée sur 2721 personnes signale que : « Malgré les aides auditives, 38 % des répondants se trouvent confrontés à des situations d'écoute difficiles » [15].

Une enquête réalisée en 2014 par *Audio 2000* et *Senior Strategic* indique que, sur 1487 personnes, **seulement 35 %** sont satisfaites par leurs aides auditives [16, p. 17].

Dans une étude menée en 2013 par *Anovum*, en partenariat avec *Phonak* et *Amplifon*, on observe que sur les 1137 personnes interrogées, 88 % sont d'accord avec le fait d'avoir « de la peine à comprendre les conversations dans un environnement bruyant » **sans aide auditive**. Après avoir porté pendant « environ un mois » leur premier appareillage, la même question leur est posée mais cette fois **avec aide auditive**. Les auteurs sont alors ravis de nous montrer que le chiffre passe à 47 % [17, p. 24]. Le pourcentage a diminué nous en convenons bien, cependant encore près de la moitié des personnes

5. Le décibel (dB) est une grandeur sans dimension qui exprime un rapport entre deux puissances. Ici il s'agit du rapport entre la puissance du signal vocal et la puissance du bruit.

peinent à suivre une conversation dans le bruit malgré l'utilisation d'audioprothèses.

Que ce soit avec ou sans appareillage, la statistique des problèmes que pose le bruit dans la compréhension d'une conversation est élevée dans le plupart des études sur l'audition.

Les résultats d'une enquête nationale effectuée en mars 2013 par l'institut de sondages *Ipsos* nous révèle que 57 % des personnes âgées de 50 ans ou plus ont du mal à suivre une conversation dans le bruit [18, p. 9]. Un article publié dans le *Hearing Journal* en 2010 annonce que c'est lors d'utilisations en situations bruyantes que les sujets appareillés sont les moins satisfaits de leurs aides auditives [19, p. 24].

1.1.4 Le bruit et ses problématiques pour l'audition

- Le Dictionnaire de l'Académie française définit le bruit ainsi : « Son ou ensemble de sons qui se produisent en dehors de toute harmonie régulière » [20].

- Luigi Russolo fait lui aussi intervenir le concept d'irrégularité dans sa description : « On appelle son ce qui est dû à une succession régulière et périodique de vibrations ; bruit, au contraire, ce qui est dû à des mouvements irréguliers aussi bien pour le tempo que pour l'intensité » [21, p. 51].

- Les normes de l'acoustique définissent le bruit de deux façons :

  - norme NF S 30-101 : « Vibration acoustique erratique intermittente ou statistiquement aléatoire. » ;
  - norme NF S 30-105 : « Toute sensation auditive désagréable ou gênante. ».

- Dans la terminologie électronique, le bruit est considéré comme un signal de type aléatoire qui altère de manière plus ou moins importante le signal utile.

Finalement, à partir de ces définitions nous pouvons, **dans le cadre de la compréhension de la parole**, caractériser le bruit ainsi : « Son ou ensemble de sons qui altèrent un signal vocal et provoquent une gêne à sa compréhension. » [6]

Si le bruit est gênant pour la compréhension de la parole, c'est à cause d'une particularité psychoacoustique que l'on nomme *effet de masque* : la perception de certaines fréquences empêche la perception d'autres composantes.

> *L'effet de masque est défini par la norme NF S 30-105 comme la « diminution de la sonie*[7] *d'un son donné (son masqué) résultant de la présence d'un son différent (son masquant) ». En pratique, on constate que, lorsque l'on présente simultanément à la même oreille deux sons de fréquences différentes, dans certaines conditions le son le plus aigu n'est pas perçu.* [22, p. 98]

En d'autres termes, les basses fréquences masquent plus facilement les hautes fréquences que l'inverse.

Dans le cas du bruit, le masquage vient d'une multitude de sons évoluant de façon aléatoire. Parmi eux, certains vont même présenter des caractéristiques fréquentielles très proches de la voix, ce qui les rend d'autant plus problématiques pour la compréhension de la parole.

Pour remédier à ces difficultés nous possédons naturellement une capacité de *démasquage*.

Prenons l'exemple d'une salle pleine de monde. Des discussions sont en cours au

6. On notera que la parole elle-même peut devenir un bruit si ce n'est pas vers elle que se porte l'attention de l'auditeur et qu'elle le gêne dans l'écoute d'une autre voix.

7. La sonie est une intensité subjective du son. À la valeur « objective » du niveau sonore, mesurée au sonomètre par exemple, correspond une valeur de sensation qu'est la sonie. En effet la sensibilité de notre organe auditif n'est pas linéaire. Selon sa fréquence, un son d'une même intensité peut nous paraître d'un niveau sonore différent.

sein de chaque groupe, créant un brouhaha *a priori* insupportable. Néanmoins, nous n'avons aucun problème à suivre la conversation que nous entretenons avec notre voisin (dans le cas d'une audition « normale »). C'est ce qu'on appelle l'effet *cocktail-party* : la possibilité de focaliser notre attention sur un flux sonore (souvent vocal) et de le démasquer de l'ambiance bruyante environnante. De plus, même en étant concentré sur une source sonore précise, notre système auditif poursuit le traitement des sons périphériques et nous restons ainsi sensibles à la moindre alerte extérieure.

Les mécanismes impliquant ces capacités psychoacoustiques sont extrêmement complexes et nous commençons à bien les appréhender depuis une petite dizaine d'années seulement. Avant il n'était pas rare qu'un patient ayant des difficultés à comprendre des conversations dans le bruit soit renvoyé chez lui sans prise en charge médicale, suite à un audiogramme tonal[8] « normal ».

Depuis, des études ont montré qu'un audiogramme classique n'était pas suffisant pour la détection de ce type d'anomalies [23, 24].

Ces études révèlent l'importance de la contribution des cellules ciliées externes[9] dans la compréhension d'une conversation en milieu bruyant. Pour éviter le phénomène d'une « perte auditive cachée », il faut donc compléter le test d'audition classique par une mesure des *oto-émissions acoustiques* (OEA).

Les OEA peuvent [...] *être définies comme des sons émis par l'oreille. La genèse de ces OEA repose sur la normalité du fonctionnement des cellules*

8. Un audiogramme est une représentation visuelle de l'audition d'une personne. Avec l'audiogramme tonal, une courbe représentant le seuil d'audition pour différentes fréquences (souvent de 125 Hz à 8000 Hz) est tracée pour chaque oreille. On compare ensuite ces résultats au seuil « normal » de perception pour détecter, le cas échéant, une déficience auditive. C'est le test le plus fréquemment pratiqué lors d'un dépistage

9. les cellules ciliées externes ou CCE, sont des cellules sensorielles situées du côté externe de l'organe de Corti, à l'intérieur de l'oreille interne. « Elles ont un rôle d'amplification ou, au contraire, d'amortissement selon les sons qui leur parviennent. Si la fréquence du son extérieur et celle des cils coïncident, il y a résonance entre les deux et très forte amplification. Au contraire, lorsque la fréquence propre de la cellule ciliée externe diffère de celle du son, en particulier si ce dernier est plus grave, un amortissement se produit » [3].

ciliées externes de l'organe de Corti. Il existe [...] *des sons émis spontanément par l'oreille nommés OEA spontanées, tandis que les autres s'appellent OEA provoquées transitoires. Ce sont essentiellement les OEA provoquées transitoires qui sont utilisées dans l'examen clinique. Les OEA sont enregistrées par une petite sonde placée dans le conduit auditif externe. Leur mesure nécessite un appareillage particulier.* [25]

Les oto-émissions acoustiques étant provoquées par les cellules ciliées externes, on peut donc, en effectuant la mesure décrite ci-dessus, déterminer le bon fonctionnement de ces cellules. Si elles présentent un dysfonctionnement nous saurons que le patient est atteint de troubles auditifs et qu'il possède des difficultés à comprendre la parole dans le bruit.

D'autre part, plusieurs études ont prouvé que l'aspect temporel de la parole possédait une forte influence sur sa compréhension dans le bruit [26, 27]. Pour illustrer les conclusions auxquelles ces chercheurs ont abouti, appliquons-les à l'exemple de l'effet *cocktail-party*.

De nombreuses personnes discutent dans une salle et toutes leurs voix (du moins une grande majorité) parviennent jusqu'à notre cerveau. Pour les traiter, nos neurones des régions sensorielles s'organisent en groupes et synchronisent leurs activités sur le rythme de différents signaux correspondant à chacune des voix. Chaque groupe de neurones identifie donc **un flux** de parole grâce à son rythme, ce qui permet de fractionner le brouhaha général en autant de signaux vocaux qu'il y a de locuteurs dans la pièce. Par la suite, d'autres régions cérébrales effectuent le tri entre tous ces signaux pour ne garder que les informations utiles.

Les neurones en charge de ce tri auraient la capacité d'«apprendre» très rapidement la signature rythmique de la voix à écouter. C'est ce rythme qui leur permet de sélectionner quasi instantanément le signal à retenir au milieu

de tout ce qui leur est envoyé par les neurones des zones sensorielles. [28]

D'autre part ces informations temporelles existent parallèlement à des informations de hauteur, d'intensité, de timbre, de spatialisation, etc. C'est l'ensemble de ces indications sonores qui permet de faire émerger la parole du bruit.

Cependant, dans certaines situations, il n'est pas (ou plus) possible d'utiliser cette capacité de démasquage. Soit parce qu'il existe une déficience auditive, soit parce que le niveau sonore du bruit est tellement élevé que le masquage devient alors trop important même pour une personne normo-entendante [10], soit parce que l'outil de captation sonore n'est pas notre appareil auditif mais un microphone par exemple, dans le cas d'une prise de son cinématographique.

On peut alors tenter de recréer artificiellement cette faculté naturelle en utilisant un *réducteur de bruit*.

1.1.5 Introduction aux techniques de réduction de bruit

La majorité des réducteurs de bruit sont construits sur le même schéma que notre capacité psychoacoustique de démasquage : analyser l'environnement sonore puis effectuer un traitement pour ne retenir que le signal utile. Mais effectuer ces étapes de manière artificielle est beaucoup plus complexe puisqu'il s'agit de définir **objectivement** ce que l'on considère comme de la parole et comme du bruit.

10. Une personne normo-entendante possède, par définition, une audition « normale » : « Votre audition est considérée comme normale si votre perte auditive moyenne se situe entre 0 et 20 décibels. » [29]. Comme nous l'avons déjà vu, le décibel est un rapport entre deux puissances. En acoustique, ce rapport est souvent celui de la grandeur mesurée sur une valeur de référence : 2.10^{-5} Pa (20 micropascals). On l'appelle alors dB SPL (*Sound Pressure Level*). Mais comme la sensibilité de notre système auditif n'est pas linéaire en fréquence, nous sommes souvent amenés à utiliser des **pondérations**. Dans le cadre de l'audiométrie, on utilise le dB HL (*Hearing Loss*) :

> *Cette échelle tient compte des différentes sensibilités de l'oreille en fonction des fréquences sonores. A titre d'exemple, une personne normo-entendante détecte un son de 500 Hz à partir de 10 dB SPL, un son de 2000 Hz à partir de 0 dB SPL et un son de 8000 Hz à partir de 20 dB SPL. A des fins audiologiques ces seuils d'audition aux différentes fréquences audibles sont ramenés à 0 dB HL.* [30]

Ainsi, une personne est considérée comme normo-entendante tant que sa perte auditive ne dépasse pas, en moyenne, 20 dB HL sur les différentes fréquences audibles.

Pour ce faire, on caractérise souvent en amont ces deux signaux en leur attribuant des paramètres fréquentiels et temporels [11]. De cette façon quand le processeur parvient à les différencier, il peut appliquer différents types de filtres (selon la marque, le modèle, etc.) pour atténuer le bruit et/ou amplifier la parole. Bien entendu ces systèmes ne sont pas parfaits et il faut trouver un équilibre entre la distorsion induite sur le signal vocal et la quantité de bruit que l'on souhaite supprimer [12].

Pour reprendre l'exemple du son au cinéma, ce genre de traitement est toujours effectué sur un ordinateur après le tournage avec suffisamment de temps pour parfaire les réglages et donc obtenir un résultat satisfaisant.

Dans le cas de la déficience auditive cependant, on utilise des audioprothèses et une difficulté apparaît : tous les traitements doivent être réalisés en temps réel depuis l'appareil. En effet, il n'est pas question d'ajouter un retard à l'information sonore transmise par les prothèses. On risquerait sinon de rendre toute conversation encore plus inaudible et incompréhensible.

C'est pourquoi bien que le numérique ait permis le développement de traitements plus puissants sur des appareils plus petits, la taille des processeurs présents sur les appareils auditifs et la vitesse à laquelle ils doivent réaliser les calculs impliquent l'utilisation d'algorithmes optimisés.

Nous retrouvons donc souvent des techniques d'*atténuation spectrale à court terme* dans les prothèses auditives.

La famille des approches par atténuation spectrale à court terme regroupe pratiquement l'ensemble des solutions utilisées dans les équipements

11. Ces paramètres sont généralement obtenus à partir d'analyses fréquentielles préalables et d'hypothèses statistiques. Dans certains cas, on estime que le bruit est un signal stationnaire.

12. Ainsi dans le cas où le niveau sonore du bruit est trop élevé et masque toute conversation, même pour une personne normo-entendante, les outils actuels ne permettent pas une amélioration de la compréhension de la parole. En effet les traitements à appliquer pour atténuer suffisamment le bruit détruiraient par la même occasion le signal vocal.

industriels en raison de la simplicité des concepts mis en jeu et de la grande disponibilité d'outils de base (notamment la FFT [*Fast Fourier Transform* ou Transformée de Fourier Rapide[13]*) nécessaires à la programmation de ces techniques*. [31, p. 6]

Le principe d'atténuation spectrale à court terme repose sur l'hypothèse que le bruit est stationnaire. Pour résumer, après avoir effectué une FFT sur le signal bruité, on identifie le bruit comme étant le signal ne variant pas, ou peu, à long terme (quelques secondes). Une fois l'identification effectuée, on atténue les composantes spectrales du bruit en essayant de minimiser les distorsions que ce traitement va provoquer sur le signal vocal. Le problème de cette méthode repose sur l'estimation du bruit. En effet, même s'il peut généralement être considéré comme « plus stationnaire » que la parole, il se révèle parfois de nature non-stationnaire.

Reprenons à nouveau l'exemple de l'effet *cocktail-party*. Le brouhaha général produit par tous les individus peut être décrit comme du bruit stationnaire car il ne va pas évoluer de manière significative. En revanche, la voix de la personne X parlant à quelques mètres de nous, qui gêne notre conversation et pouvant donc être considérée comme du bruit, n'est en aucun cas stationnaire. Dans cette situation comment faire la différence entre la voix de notre interlocuteur (signal utile) et celle de la personne X (bruit), la prothèse auditive les captant de la même manière ? L'appareil ne répond qu'en fonction de paramètres objectifs d'un point de vue physique. Le risque est donc que l'audioprothèse retransmette les deux signaux vocaux amplifiés l'un et l'autre. Cela ne va nous être d'aucune aide si nous avons perdu notre capacité de démasquage, et risque même de nous gêner puisque la voix de la personne X sera encore plus audible que sans appareillage.

13. La transformée de Fourier est un outil mathématique permettant, en théorie, de caractériser n'importe quel signal analogique par son spectre de fréquences. Son équivalent discret, la TFD (Transformée de Fourier Discrète), est utilisé dans le domaine numérique. La TFD est efficace pour déterminer le spectre d'un signal mais nécessite un temps de calcul très important. L'intérêt de la FFT est de réduire significativement ce temps de calcul. La TFD reste donc la transformée la plus utilisée dans les techniques d'atténuation spectrale à court terme (et donc dans les aides auditives) puisque qu'il existe un algorithme puissant permettant d'effectuer les calculs (la FFT), que ne possèdent pas les autres transformées exploitables (entre autres les transformées en ondelettes).

Cette situation où une parole en masque une autre est extrêmement délicate à traiter, car le signal et le bruit présentent alors des caractéristiques similaires.

Une solution proposée par les fabricants d'audioprothèses est d'augmenter le nombre de microphones dans l'appareil afin de permettre une discrimination spatiale. Si l'on considère le signal capté à l'arrière et/ou de manière latérale comme étant du bruit, on peut facilement l'identifier et le soustraire au signal frontal, considéré comme utile. Néanmoins, ce traitement suppose de rester face à son interlocuteur pour suivre une conversation et implique qu'une information importante provenant par exemple de l'arrière ne soit pas forcément perçue.

Pour optimiser les prothèses actuelles, plusieurs techniques sont utilisées de manière simultanée. On y associe traitement temporel, fréquentiel et spatial. Cependant nous l'avons vu, les performances proposées par ces appareils ne sont pas encore assez satisfaisantes pour les patients. Beaucoup ont en effet du mal à suivre une conversation dans le bruit malgré ces aides auditives.

1.2 Démarche

Nous avons souhaité étudier dans le cadre de ce mémoire de fin d'études comment, *in fine*, contribuer à l'amélioration de ces dispositifs en nous concentrant sur l'aspect **fréquentiel** des traitements à réaliser et en veillant à s'assurer du respect des points suivants :

- faire émerger la voix d'un environnement sonore bruyant tout en conservant sa clarté ;
- éviter ou sinon minimiser les distorsions spectrales liées aux traitements nécessaires ;
- étudier comment appliquer ces traitements en temps réel.

Pour ce faire, nous nous sommes appuyés sur le principe de l'analyse/re-synthèse IDS (Intégration de Densité Spectrale) présentée par Émile Leipp à la fin des années 1970 et proposée en version numérique par Laurent Millot.

Cet outil permet d'effectuer une analyse de l'énergie fréquentielle (ou spectrale) d'un son. Contrairement à un analyseur de spectre classique, cette analyse s'effectue dans la durée en procédant à une sommation temporelle (d'où la référence à l'intégration). Dans sa version de 1977, Leipp propose un découpage fréquentiel du spectre sonore audible par l'être humain [14] en 8 sous-bandes de fréquences.

> *L'extension de l'analyse IDS originelle tient à un ajustement du découpage proposé par Leipp, à un changement du mode de représentation des données et surtout à l'ajout de la possibilité d'écouter les composantes de l'analyse ou toute reconstruction, partielle ou totale, du signal sonore étudié* [...] [33, p. 163]

L'avantage de l'analyse IDS dans sa version numérique est qu'elle offre la possibilité de faire émerger puis d'utiliser un découpage fréquentiel **adapté à l'objet d'étude**, contrairement à la FFT (généralement utilisée dans les réducteurs de bruit classiques) qui analyse le signal en suivant une échelle linéaire et uniforme en fréquence.

De plus pour un jeu de fréquences données, la FFT ne calcule qu'une **approximation** de la valeur du spectre. En effet elle ne permet pas la connaissance de ce qui se produit entre deux fréquences calculées, tandis que l'analyse IDS donne une information (poids relatif) pour l'ensemble des composantes présentes dans chaque sous-bande.

Par ailleurs l'analyse IDS permet d'écouter les composantes du signal analysé (puisqu'il s'agit de signaux en sous-bandes) là ou les autres analyses classiques (par exemple transformées de Fourier et transformées en ondelettes) ne proposent qu'une apprécia-

14. Le spectre audible est l'ensemble des fréquences sonores pouvant être perçues par l'être humain. En théorie il s'étend de 20 Hz à 22000 Hz [32].

tion visuelle. S'agissant de signaux audio, il semble logique de vouloir vérifier les résultats obtenus en les comparant à notre perception auditive.

L'IDS donne aussi les moyens de choisir le poids de chaque composante à prendre en compte pour une re-synthèse, partielle à totale, du signal.

Enfin cette analyse n'induit pas de distorsions et la reconstruction du signal se fait sans erreur sensible.

Pour comprendre la démarche que nous avons adoptée afin d'aboutir à la réalisation de traitements permettant une meilleure compréhension de la parole dans le bruit, il est intéressant d'observer les recommandations de Roland Carrat quant au fonctionnement des aides auditives :

> *L'amplification des fréquences déficitaires ou perdues ne provoquent pas pour autant une régénération des capteurs neurosensoriels manquants (tout au plus la perception d'harmoniques inférieurs).* [...] *La logique serait, inversement, non pas de transmettre un signal de très large spectre avec compensation du déficit auditif par une amplification sonore sélective, mais de procéder à une adaptation du signal à la capacité du récepteur (l'oreille), en effectuant une analyse de parole puis la synthèse d'un nouveau signal codé compatible avec le canal auditif du sujet, en d'autres termes de réduire le flux d'information pour l'adapter au débit du canal auditif.* [34, p. 211]

La première étape du travail a donc été de caractériser la parole en obtenant, à l'aide de l'analyse IDS, un découpage fréquentiel adapté à la voix française. À cet effet, nous avons constitué deux corpus de voix : un premier composé de voix humaines « réelles » et un second rassemblant des voix issues de programmes télévisuels. Suite à l'analyse IDS de ces corpus, plusieurs découpages ont été retenus.

Ces découpages fréquentiels ont ensuite été utilisés pour réaliser des re-synthèses IDS à l'aide d'un outil que nous avons nommé l'« IDS *Speech Enhancer* » (IDSSE). Lors de tests perceptifs, des sujets ont pu ainsi régler le poids des différentes sous-bandes d'un signal vocal pour le faire émerger d'une ambiance bruyante. Les informations récoltées à l'aide de l'IDSSE doivent permettre d'obtenir les filtres à réaliser pour améliorer la compréhension de la voix dans chacune des ambiances testées. Comme ces traitements sont destinés à une utilisation quotidienne nous avons, pour finir, étudié les moyens et les outils nécessaires au fonctionnement de ces filtres en temps réel.

La rédaction de ce mémoire suivra notre démarche.

1. Tout d'abord nous allons détailler la façon dont les corpus de voix ont été composés ainsi que l'élaboration du protocole de leur enregistrement.

2. Puis nous expliquerons plus en détails les principes de l'IDS et la façon dont nous avons retenu des découpages fréquentiels adaptés à la voix française.

3. Nous décrirons ensuite les premiers tests perceptifs mis en place grâce à l'IDS *Speech Enhancer* et nous apporterons des amendements au protocole de test en fonction des résultats de l'expérience.

4. Enfin nous étudierons les compromis et les outils nécessaires pour envisager l'utilisation des traitements obtenus à partir de l'IDSSE en temps réel.

2 Constitution de corpus oraux représentatifs de la voix française

Dans ce chapitre nous allons retracer la manière dont deux corpus oraux (voix parlées et voix télévisuelles) ont été constitués pour obtenir un découpage fréquentiel adapté à la voix française. Nous allons tout d'abord expliquer pourquoi il s'est trouvé nécessaire d'établir de tels corpus, nous étudierons ensuite les différents choix qui ont guidé leur composition puis nous détaillerons l'élaboration du protocole de leur enregistrement.

2.1 Pourquoi constituer des corpus originaux ?

Dans l'introduction nous avons vu que pour réaliser des traitements permettant l'émergence de la parole dans le bruit, il faut en amont attribuer à ces deux types de signaux des caractéristiques objectives. Dans le cadre de cette étude nous avons tout d'abord cherché à définir des caractéristiques fréquentielles pour la parole. À cet effet, il a fallu constituer des corpus oraux représentatifs de la voix française, dont nous avons ensuite fait l'analyse IDS afin d'en faire émerger des découpages fréquentiels appropriés.

Mais pourquoi ne pas utiliser des corpus déjà existants ?

Premièrement, parce que la majorité des corpus oraux sont à destination d'études linguistiques et n'ont pas forcément de contraintes quant à la qualité sonore des enregistrements. En effet ces corpus sont généralement constitués pour travailler sur le contenu **sémantique** de la parole. L'enregistrement en lui même n'a pas de valeur tant que la voix est audible et compréhensible, il s'agit juste de retransmettre ce qui est dit.

Nous pouvons par exemple citer le projet TCOF (Traitement de Corpus Oraux en Français) du laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française) qui propose un grande variété d'enregistrements de la parole française dans des contextes interactifs (conversations, débats, etc.).

> *Le corpus mis à disposition comporte deux grandes catégories : des enregistrements d'interactions adultes-enfants (enfants jusqu'à 7 ans) et des enregistrements d'interactions entre adultes. Les enregistrements sont de durées diverses : de 5 à 45 minutes.* [...] *Il s'agit, en l'absence de corpus de référence du français parlé, de faciliter l'accès à des données qui restent encore rares, en particulier en ce qui concerne les interactions adultes-enfants, et de compléter les données existantes mises à disposition au travers d'un certain nombre de sites* [...] [35]

Bien que ce corpus soit d'une grande richesse au niveau du contenu, la qualité sonore des enregistrements (globalement mauvaise) n'en permet pas, *a priori*, l'utilisation pour une analyse fréquentielle précise. De plus les prises de son mises à disposition n'ont pas été réalisées pour le corpus mais dans divers contextes. Elles sont donc fortement hétérogènes car elles ont été effectuées par différentes personnes, avec différents outils et dans différents lieux.

Or dans le cadre de notre étude, il est important de connaître et de ne jamais modifier les conditions d'enregistrement du corpus, du moins dans un premier temps. En effet en appliquant un protocole identique à chaque prise de son, on s'assure que les divergences potentielles observées dans les résultats de l'analyse IDS ne peuvent dépendre que des *stimuli* (ici de la parole). Dans un deuxième temps, on aurait pu modifier un paramètre du protocole et observer les conséquences de cette correction, puis un deuxième, etc. Peut être d'ailleurs qu'il n'y aurait eu aucun changement. Peut être que les découpages fréquentiels obtenus auraient été semblables qu'on analyse des enregistrements réalisés avec un *smartphone* ou avec un microphone de mesure.

En tout cas il fallait partir d'un protocole d'enregistrement connu qui allait servir de référence.

Et c'est le deuxième problème des corpus oraux déjà existants : même si leur contenu est d'une bonne qualité sonore, les protocoles d'enregistrement ne sont jamais (ou très peu) décrits.

Prenons par exemple les prises de son réalisées pour l'audiométrie vocale [1]. Le Guide des Bonnes Pratiques en Audiométrie de l'Adulte [36], constitué au sein de la Société Française d'Audiologie, indique :

> *Conformément à la norme ISO 8253-3, le matériel vocal doit être enregistré ce qui a pour avantage de garantir le niveau (pression acoustique) et la qualité sonore du message vocal, indépendamment de l'opérateur. L'enregistrement doit cependant obéir à des règles précises, et comporter notamment un signal d'étalonnage et des signaux permettant de contrôler la distorsion harmonique de l'audiomètre vocal.*

Pas une recommandation n'est formulée sur le type de microphone ou d'enregistreur à utiliser, ainsi que sur la distance à respecter entre le locuteur et le microphone. Concernant le « signal d'étalonnage » et les « signaux permettant de contrôler la distorsion harmonique », aucune information supplémentaire n'est donnée et il faut avoir accès à la norme ISO 8253-3 pour en savoir plus.

La volonté de constituer des corpus oraux originaux vient donc du besoin de disposer d'enregistrements de bonne qualité sonore, réalisés en suivant un protocole connu et respecté pouvant être réutilisé et/ou amendé.

1. L'audiométrie vocale est un examen permettant de mesurer la capacité à reconnaître la parole. Le test consiste à faire répéter par le patient des mots émis à différents niveaux sonores, puis à comptabiliser le nombre d'erreurs commises. Elle est complémentaire de l'audiométrie tonale qui ne donne que des informations sur la perception de sons tonaux.

2.2 Composition des corpus

Nous avons décidé de constituer deux corpus composés chacun de *stimuli* différents. Le premier présente des voix parlées « réelles » enregistrées par des locuteurs physiques, et le deuxième regroupe des voix issues de programmes télévisuels diffusées depuis une télévision (TV).

Le premier corpus apparaît spontanément comme indispensable car chaque jour nous échangeons de manière vocale avec nos proches ou divers interlocuteurs. Cette parole « réelle » demeure encore aujourd'hui prédominante dans notre façon de communiquer.

Le deuxième corpus a été constitué pour deux raisons. D'une part, pour observer s'il existe des différences notables dans les découpages fréquentiels obtenus à partir de voix provenant de sources sonores différentes. D'autre part la télévision est encore le média le plus consommé en France [37]. En outre, ce média est particulièrement utilisé chez les 50 ans et plus [38] qui sont nombreux à déclarer se sentir « gênés au niveau de leur audition » lorsqu'ils regardent la TV justement [18, p. 9]. Il paraît donc intéressant d'intégrer la « parole télévisuelle » à notre étude.

Pour les deux corpus nous n'avons utilisé que des enregistrements en français. En effet pour éviter d'ajouter des variables supplémentaires liées à l'usage de plusieurs langues, nous avons décidé dans un premier temps de n'en étudier qu'une.

Dans les pages qui suivent, nous allons décrire le contenu de ces deux corpus et exposer les choix qui ont guidé leur composition.

2.2.1 Composition du corpus de voix parlées

Le corpus composé de voix parlées se divise en 3 enregistrements différents.

Premier enregistrement

Durant ce premier enregistrement le locuteur doit prononcer quatre listes de mots. Ces listes reprennent le principe de la liste cochléaire[2] du professeur Lafon, qu'il a décrit ainsi :

> *La liste contient une série d'éléments de 17 mots, elle est destinée à mesurer les déformations apportées par les surdités. Elle doit donc représenter un large éventail phonétique, chaque élément de la liste contient presque tous les phonèmes*[3] *de la langue.* [40, p. 127]

Le terme « élément » désigne ici une sous-liste de 17 mots, formés chacun de 3 phonèmes. Pour l'enregistrement du corpus nous avons utilisé quatre de ces sous-listes que nous avons toutefois modifiées.

En effet le but de la liste cochléaire, telle que Jean-Claude Lafon l'a conçue, est de mettre en évidence des difficultés auditives. Elle n'a donc pas été créée de façon à être phonétiquement représentative de la parole française, bien au contraire. Les phonèmes peu utilisés dans le français parlé ont été mis en avant car ils étaient plus susceptibles de poser des problèmes d'audition et de compréhension. Étant moins habitués à les entendre, les patients sont potentiellement moins sensibles à leurs sonorités et donc moins à même de les identifier.

2. Liste « cochléaire » car elle est utilisée pour la mesure des déformations acoustiques provoquées par la cochlée.

3. Phonème : « la plus petite unité sonore d'une langue donnée, caractérisée par des traits distinctifs, et qui, combinée à d'autres, sert à former des unités signifiantes telles que les morphèmes, les mots, les phrases, etc. Le français compte trente-six phonèmes. » [39]

Cependant pour notre étude, le corpus devait être représentatif de la voix française. Nous avons donc réorganisé quatre « éléments » de cette liste cochléaire, en s'appliquant à respecter la fréquence d'occurrence des phonèmes de la langue française **parlée**.

Si l'on insiste sur le terme « parlée », c'est parce qu'il existe plusieurs travaux qui ont été effectués sur ces statistiques, à l'écrit comme à l'oral [41, 42, 43, 44]. Toutes ces études aboutissent à des résultats plus ou moins différents car la nature des corpus utilisés pour l'analyse est très variable (romans, poésies, essais, traités scientifiques, ouvrages scolaires, sous-titres de films, retranscriptions de discours, de conversations, de débats, etc.).

Nous avons décidé d'utiliser les travaux menés par François Wioland sur le français **parlé** et ses fréquences obtenues « à partir d'échantillons représentatifs qui totalisent 200 000 phonèmes » [44, p. 30]. Ses résultats sont retranscrits dans les tableaux ci-dessous avec, entre parenthèses, des mots dont les lettres en gras et majuscules correspondent à la prononciation du phonème correspondant :

pour les consonnes

1 - / R / 7,25 (**R**ouge)	11 - / j / 2,00 (**Y**ahourt)
2 - / s / 6,00 (**S**ite)	12 - / ʒ / 1,66 (**J**our)
3 - / l / 5,63 (**L**oup)	13 - / z / 1,535 (**Z**inc)
4 - / t / 5,335 (**T**u)	14 - / f / 1,40 (**F**ête)
5 - / k / 4,06 (**C**alme)	15 - / w / 1,40 (**W**att)
6 - / d / 4,035 (**D**on)	16 - / b / 1,31 (**B**elle)
7 - / m / 3,845 (**M**atin)	17 - / ʃ / 0,535 (**CH**aud)
8 - / p / 3,715 (**P**ile)	18 - / ɥ / 0,515 (**HU**ile)
9 - / n / 3,095 (**N**ez)	19 - / g / 0, 475 (**G**alop)
10 - / v / 2,755 (**V**ase)	

pour les voyelles

1 - / E / 10,60 mélange de / e / (pr**É**) et / ɛ / (p**È**re)
2 - / a / 8, 55 (b**A**l)
3 - / i / 5,115 (par**I**)
4 - / Œ / 4,31 mélange de / ø / (p**EU**) et / œ / (**ŒU**f)
5 - / O / 3,36 mélange de / o / (s**O**t) et / ɔ / (s**O**rt)
6 - / ɑ̃ / 3,09 (ava**AN**t)
7 - / u / 2,43 (hib**OU**)
8 - / ɔ̃ / 2,255 (b**ON**)
9 - / y / 1,90 (v**U**)
10 - / ɛ̃ / 1,845 (p**AIN**)

(en %)

FIGURE 2.1 – Fréquence d'occurrence des phonèmes dans le discours [Ibid.]

Pour modifier la liste cochléaire du professeur Lafon en fonction de ces paramètres, nous avons procédé ainsi.

- Tout d'abord nous avons sélectionné quatre « éléments » de cette liste, rassemblant donc 68 mots de 3 phonèmes chacun, c'est à dire un total de 204 phonèmes.

- Nous avons ensuite comptabilisé le nombre d'occurrences de chaque phonème, tels qu'ils ont été décrits dans la Figure 1[4].

- Enfin nous avons modifié certains mots de façon à nous rapprocher le plus possible des fréquences d'utilisation des phonèmes annoncées par François Wioland. Prenons par exemple le phonème / s /. Sa fréquence d'occurrence est de 6 % d'après la Figure 1. Nous avons donc fait en sorte qu'il soit présent 12 fois, puisque la base de données était de 200 phonèmes[5].

Comme pour des raisons de réalisme nous avons décidé d'utiliser des mots existants[6], nous n'avons pas pu atteindre exactement les résultats obtenus par le professeur Wioland. Toutefois nous avons cherché à nous en éloigner le moins possible, en limitant le taux d'erreur pour chaque phonème à un maximum de 1 %.

Les sous-listes cochléaires du professeur Lafon, les modifications que nous y avons apportées et un tableau comparant les fréquences d'occurrence des phonèmes selon Wioland à celles que nous avons obtenues, sont disponibles en annexe.

4. On remarque qu'il n'y a pas de distinction entre le / e / et le / ɛ /, le / ø / et le / œ /, ainsi qu'entre le / o /, et le / ɔ /. En effet ces *voyelles d'aperture moyenne*, telles qu'on les appelle, posent problème à l'oral car pour un mot donné elles ne vont pas être prononcées de la même façon d'une personne à l'autre. C'est sans doute pour cette raison qu'elles ont été associées, s'agissant d'une étude sur le discours.

5. En arrondissant les 204 phonèmes à 200 on s'octroie une légère marge d'erreur.

6. En audiométrie vocale, il existe des tests où des mots sans significations (logatomes) sont utilisés afin de supprimer les effets de la suppléance mentale (identification par contexte). Une fois encore, nous ne cherchons pas ici à diagnostiquer des défaillances auditives mais à nous rapprocher le plus possible de la langue parlée. Il est donc préférable d'utiliser des mots réels pour être dans une situation assez homogène avec des conditions de vie « réelle ».

Deuxième enregistrement

Le deuxième enregistrement consistait en la lecture d'un texte. Il s'agissait de choisir un document assez court dont la durée ne devait pas excéder 3 minutes. Cela à la fois pour ne pas fatiguer le locuteur afin qu'il garde la même prononciation du début à la fin, mais aussi de façon à enchaîner les enregistrements assez rapidement pour en avoir un nombre significatif.

D'autre part, le texte ne devait pas employer de manière excessive des figures de style comme la répétition, l'allitération ou l'assonance. Il fallait éviter que certaines sonorités prennent explicitement le dessus dans l'ensemble de la lecture, toujours dans le but d'avoir un corpus représentatif de la voix française.

Nous avons sélectionné une nouvelle de l'écrivain français Didier Daeninckx : *Coupe-Coupe* [45, p. 111-113]. La durée de lecture de ce texte oscillait entre 2 minutes 30 et 3 minutes selon le débit de parole du locuteur. De plus les caractéristiques phonétiques correspondaient à nos attentes. Une transcription écrite de ce texte est disponible en annexe.

Troisième enregistrement

Le troisième et dernier enregistrement était une forme d'*interview* où le locuteur était amené à parler librement.

Pour ce faire, nous lui avons posé plusieurs questions à partir desquelles il pouvait développer des réponses assez longues. Par exemple : « Pouvez-vous vous présenter, nous dire d'où vous venez et quelles études vous avez faites ? » ; « Avez-vous/Allez-vous voter aux $1^{er}/2^{ème}$ tour des élections présidentielles ? Pourquoi ? » ; « Qu'avez-vous prévu de faire cet été ? Stage ? Vacances ? Travail ? » ; etc. Ces questions servaient à instaurer un dialogue avec la personne afin de la faire parler d'une manière plus naturelle, plus conversationnelle, durant 2 à 3 minutes.

Conclusion

Nous avons donc composé 3 sous-corpus de voix parlées. Chacun possède des caractéristiques différentes :

- **pour la liste de mots :** une parole assez « artificielle » présentant des sonorités qui se veulent les plus proches de statistiques « objectives » ;
- **pour le texte :** une parole plus fluide, qui a du sens, qui raconte une histoire ;
- **pour l'*interview* :** une parole plus détendue, vivante, proche de la conversation de tous les jours.

Nous ne pouvions pas savoir à l'avance si l'analyse IDS de ces 3 types d'enregistrements allait ou non nous donner des résultats différents, mais l'idée était d'obtenir potentiellement plusieurs découpages fréquentiels et de séléctionner ensuite celui ou ceux à retenir pour les tests perceptifs.

2.2.2 Composition du corpus de voix télévisuelles

Pour composer ce corpus nous avons dû utiliser des programmes télévisuels où la voix était particulièrement mise en avant. C'est pourquoi nous avons sélectionné des journaux télévisés (JT) ainsi que des bulletins météorologiques (météos). Nous avons choisi des programmes issus de *TF1*, *France 2*, *France 3* et *M6*, qui sont encore les quatre chaînes de télévision les plus visionnées en France [46].

L'idée a ensuite été de récupérer ces programmes dans une qualité identique à celle de leur diffusion.

Depuis le 5 avril 2016, la télévision numérique terrestre (TNT) est diffusée en haute-définition (HD).

La norme de codage utilisée pour les programmes est le MPEG-4[7] et les signaux audio sont compressés avec l'algorithme AAC-LC[8].

Les programmes utilisés pour la constitution du corpus ont donc été téléchargés en MPEG-4, avec une compression du son en AAC-LC, directement depuis les différents sites de *replay* des chaînes.

Concernant la propriété intellectuelle, la loi indique :

> *Lorsque l'œuvre a été divulguée, l'auteur ne peut interdire :*
>
> 1. *Les représentations privées et gratuites effectuées exclusivement dans un cercle de famille ;*
> 2. *Les copies ou reproductions réalisées à partir d'une source licite* [...] [48]
>
> *Les bénéficiaires des droits ouverts au présent titre ne peuvent interdire :*
>
> 1. *Les représentations privées et gratuites effectuées exclusivement dans un cercle de famille ;*
> 2. *Les reproductions réalisées à partir d'une source licite* [...] [49]

Ici nous nous trouvons bien dans un cadre régi par la loi puisque :

- la **représentation** n'a eu lieu que pendant l'enregistrement dans une pièce où seuls des microphones étaient présents ;
- les **copies** ont été réalisées à partir des sites internets officiels des chaînes de télévision qui sont donc des sources licites.

7. Le MPEG-4 est une norme de codage du contenu audiovisuel. Comparé au MPEG-2, qui était l'ancienne norme utilisée pour la TNT, le MPEG-4 permet d'obtenir un débit beaucoup plus important. Sur une bande de fréquences moins large, il est donc possible de faire passer autant, voire plus d'informations qu'avec le MPEG-2. Ce changement a permis la vente de la bande des 700 Mhz aux opérateurs de téléphonie mobile pour qu'ils déploient la 5G, et la TNT a pu passer au « tout HD » [47].

8. L'AAC-LC est un format de compression audio qui assure un très bon rapport qualité/débit. C'est la raison pour laquelle il est utilisé dans le MPEG-4.

Un descriptif détaillé des programmes télévisuels utilisés pour la composition de ce corpus est disponible en annexe.

Nous venons d'exposer les différentes étapes de la composition des deux corpus oraux, nous allons maintenant décrire l'élaboration du protocole de leur enregistrement.

2.3 Enregistrement des corpus

2.3.1 Élaboration du protocole d'enregistrement pour les voix parlées

Matériel

Pour l'enregistrement de ce corpus, deux microphones ont été utilisés :

- un microphone de mesure *Behringer ECM8000* : électrostatique[9] et omnidirectionnel[10] ;
- un microphone *Neumann TLM103* : électrostatique avec une directivité cardioïde[11].

Nous avons choisi des microphones distincts pour observer si les résultats obtenus après l'analyse IDS étaient nettement différents selon le microphone, ou si au contraire leur influence était trop faible pour être prise en compte.

Le *Behringer ECM8000* étant un microphone de mesure et possédant donc une réponse en fréquence quasiment plate[12], permet d'avoir une prise de son peu altérée par

9. Un microphone électrostatique (ou statique) a besoin d'une alimentation externe (alimentation fantôme ou pile) pour fonctionner. Il possède une sensibilité très importante.

10. Avec un microphone omnidirectionnel, aucune source sonore n'est privilégiée. En théorie le microphone capte les sons sur 360° d'une manière uniforme. En pratique il est moins sensible aux hautes fréquences lorsqu'elles proviennent de l'arrière ou de manière latérale.

11. Un microphone cardioïde privilégie les sources sonores frontales. Cette appellation vient de l'apparence de son diagramme directionnel car il a la forme d'un cœur. Le son d'un microphone cardioïde est légèrement moins réaliste que celui d'un omnidirectionnel.

12. La courbe de réponse en fréquence d'un microphone est la représentation graphique de

ses caractéristiques techniques. Le *Neumann TLM103*, microphone généralement utilisé pour l'enregistrement de voix (musique, radio, etc.), est susceptible de « colorer » davantage les prises de son puisque sa réponse en fréquence est plus irrégulière (surtout dans les aigus). Les courbes de réponse en fréquence ainsi que les diagrammes directionnels de ces deux microphones sont disponibles en annexe.

La carte son employée est une Mbox2 de chez *Avid* et les prises de son sont réalisées sur le logiciel *Reaper* (version 5.40) à une fréquence d'échantillonnage de 44100 Hz et à une résolution de 32 bits à virgule flottante (ou « 32 bits float »).

Le lieu dans lequel les prises de son sont effectuées est une pièce d'une dizaine de mètres carrés. Il s'agit d'une « cabine *speak* » normalement destinée à la captation radiophonique. Elle présente une bonne absorption acoustique et est assez bien isolée [13].

Position des microphones par rapport au locuteur

Nous avons décidé que le locuteur devait être debout lors des enregistrements. Cette position permet d'empêcher la présence de bruits parasites liés à l'utilisation d'un siège (repositionnements, balancements, etc.) et évite l'effet soporifique qu'une posture assise pourrait provoquer, surtout lors de la lecture assez monotone de la liste de mots.

Nous avons naturellement choisi de placer les microphones face à la bouche du locuteur. Le *TLM103* ayant une directivité cardioïde et le *Behringer* n'étant pas tout à fait omnidirectionnel dans la pratique, c'est de manière frontale que ces microphones captent le mieux les signaux sonores. En prenant une taille moyenne de 1,70 m pour les Français [50] et en situant la bouche environ 10 cm plus bas, nous avons donc positionné les microphones à une hauteur de 1,60 m, face au locuteur (le corps des microphones pointant vers le sujet).

---

sa capacité à restituer toutes les fréquences qu'il doit enregistrer. En pratique les microphones ne sont pas sensibles à toutes les fréquences de la même façon (comme notre oreille). Chaque microphone possède donc une courbe de réponse en fréquence plus ou moins différente qui est normalement fournie par le fabricant.

13. L'**absorption** acoustique est la correction acoustique à l'intérieur d'une pièce. L'**isolation** acoustique est la protection du bruit venant des pièces avoisinantes.

Le but était d'obtenir une prise de son qui soit la plus représentative d'une situation de parole réelle. En diffusant l'enregistrement à travers des enceintes, nous devions avoir l'impression que quelqu'un était vraiment en train de parler. Le locuteur ne pouvait donc pas être trop près des microphones à cause des effets de proximité [14]. Il ne fallait pas non plus trop l'éloigner car nous risquions alors de donner beaucoup d'importance à l'acoustique de la salle et d'ajouter de la réverbération au signal vocal.

Nous avons donc procédé à des tests où nous écoutions un sujet parler à différentes distances des microphones afin de déterminer l'éloignement idéal. Finalement, nous avons constaté que la sensation d'avoir un vrai locuteur face à nous était la plus exacerbée lorsque le microphone était placé à 1 m de la personne que l'on enregistrait. Nous avons donc décidé que la distance bouche/microphones devait être d'1 m pour toutes les prises de son.

Des marques au sol ont été installées de façon à éviter de re-mesurer la distance pour chaque personne. Évidemment, selon la taille du locuteur et les légers mouvements effectués lors de la lecture, la distance n'était jamais exactement d'un mètre. Mais cette distance ayant été trouvée empiriquement, d'après un avis subjectif, nous avons considéré que de faibles variations sur le placement étaient perceptivement acceptables.

Déroulement des enregistrements

Les prises de son sont effectuées individuellement. La liste de mots, le texte, et l'« *interview* », sont enregistrés dans cet ordre par chaque locuteur.

- Pour la liste de mots, nous donnons au locuteur la consigne de laisser une petite pause entre chaque mot afin de garder une prononciation claire et précise.

- Pour le texte, il est indiqué que la lecture doit se faire en mettant le ton afin de produire des intonations variées.

14. L'effet de proximité est un phénomène qui se traduit par un renforcement des fréquences basses lorsque la soucre est très proche du microphone.

- Lors de l'« *interview* », le locuteur est situé dans la même pièce que l'« interviewer ». Nous avons fait l'hypothèse que ce dernier n'était pas un élément parasite pour l'enregistrement. En effet il ne prononçait qu'une faible quantité de parole comparée à celle de l'interrogé et les *stimuli* qu'il produisait étaient eux aussi de la voix parlée. De plus la distance à laquelle il était situé des microphones atténuait son influence sur le signal enregistré [15].

Pour nommer les enregistrements du corpus de voix parlées nous avons adopté la convention suivante :

- en premier figure le **microphone utilisé** : « Behringer » ou « TL103 » ;

- on indique ensuite la **nature du *stimulus*** : « ListeMots », « Texte », ou « Itw » ;

- vient après le **sexe** du sujet, son **âge** et sa **nationalité** [16] : si c'est une femme française de 21 ans, on notera « F21fra », si c'est un homme espagnol de 25 ans, « H25esp ». Étant donné que notre étude porte sur le français et qu'elle n'a rassemblé que des sujets de nationalité française, tous les enregistrements du corpus possèdent la mention « fra ». Néanmoins cette convention pourrait être utile si l'on étend la recherche à d'autres langues ou si le locuteur effectuant les enregistrements est d'origine étrangère ;

- pour finir, on ajoute un **numéro** pour différencier deux personnes du même sexe, du même âge et de même nationalité. Ce numéro correspond à l'ordre de passage du sujet. Par exemple, on attribue à la première femme française de 21 ans le numéro « 00 », si une deuxième femme française du même âge effectue aussi les enregistrements on lui donne le numéro « 01 », etc. [17] ;

15. L'« interviewer » était à environ 4 mètres des microphones, face au locuteur.
16. Le système d'abréviation utilisé pour le critère de nationalité est celui des événements sportifs internationaux (jeux olympiques, coupe du monde, etc.), c'est-à-dire, pour la plupart des nations, l'utilisation des 3 premières lettres du pays en question. Exemples : « fra » pour une nationalité française, « ita » pour italienne, etc.
17. Nous avons adopté un principe présent en programation qui est de définir 0 comme point initial d'une numérotation plutôt que 1 afin de simplifier l'utilisation potentielle de ce corpus sur différents loiciels.

- tous ces éléments sont séparés par un ***underscore*** : « _ ».

Exemple : si le fichier correspond à un enregistrement réalisé au *Neumann TLM103* et qu'il s'agit de la lecture du texte effectuée par le troisième homme français de 22 ans à passer, alors on lui attribuera le nom : « TLM103_Texte_H22fra_02 ».

Nous avons enregistré au total 43 personnes (24 hommes et 19 femmes) âgées entre 20 et 37 ans. La durée d'une session d'enregistrement était de 5 à 10 minutes par locuteur selon son débit de parole. La durée complète d'enregistrement pour chaque microphone est de 4 heures 43 minutes et 20 secondes. La durée totale de ce corpus est donc de 9 heures 26 minutes et 40 secondes.

2.3.2 Élaboration du protocole d'enregistrement pour les voix télévisuelles

L'idée pour ce protocole d'enregistrement était de se rapprocher au mieux du protocole décrit précédemment. En effet l'une des raisons pour lesquelles nous avons constitué deux corpus était d'observer si, après analyse IDS, nous trouverions des différences significatives entre les résultats obtenus pour le corpus de voix parlées et le corpus de voix télévisuelles. Il fallait donc que les deux protocoles soient les plus similaires possible de manière à ce qu'ils ne puissent pas être la cause de ces différences.

Matériel

Tout d'abord nous utilisons les mêmes microphones (*Behringer ECM8000* ; *Neumann TLM103*) et le même lieu (« cabine *speak* ») que pour les enregistrements précédents. Cette fois cependant nous réalisons les enregistrements avec une console *Yamaha DM2000* sur le logiciel *Protools* (version 12). Nous gardons la fréquence d'échantillonnage à 44100 Hz et la résolution à 32 bits *float*.

Ici la source sonore n'est plus un locuteur mais une télévision sur laquelle des programmes préalablement téléchargés sont diffusés. Cette télévision est une LG 47LA6205 *Full HD* (Haute Définition), d'une taille de 119 cm de diagonale et possédant deux hauts-parleurs (HP) « invisibles » [18] aux deux extrémités de l'écran.

Position des microphones par rapport au téléviseur

Pour positionner les microphones par rapport à l'écran, nous sommes partis du principe qu'ils devaient être situés au même endroit que les oreilles d'une personne regardant un programme télévisé dans un contexte traditionnel (assis sur un canapé/fauteuil dans un salon).

D'un point de vue vertical nous avons donc positionné les microphones à une hauteur correspondant au centre de l'écran, en supposant que l'angle entre notre regard et la télévision faisait généralement 90°.

Pour la distance horizontale nous avons suivi les recommandations stipulant qu'avec un téléviseur *Full HD*, l'écart conseillé entre les yeux et l'écran est de 2,6 fois la diagonale de l'écran [51]. Par conséquent l'éloignement optimal était de 309 cm pour notre télévision. Nous avons donc placé les microphones à environ 3,10 m du centre de l'appareil.

Déroulement de l'enregistrement

Il fallait définir un niveau sonore auquel nous allions diffuser les programmes. Plusieurs sources indiquent que le niveau sonore d'un téléviseur en fonctionnement se situe entre 65 et 75 dB(A) [19] [52, 53]. Nous avons donc choisi un niveau de diffusion de 70 dB(A) que nous avons mesuré à l'aide d'un sonomètre.

Tous les programmes à enregistrer étaient contenus sur un disque dur que nous avons

18. Les HP des téléviseurs sont souvent invisibles puisqu'ils sont intégrés à l'écran.
19. Comme nous l'avons déjà vu, il existe plusieurs pondérations du dB SPL. Ces différentes pondérations (A, B, C, HL, etc.) permettent de mesurer les niveaux sonores **en fonction de notre sensibilité**. Le choix d'une pondération va dépendre du contexte d'étude et du niveau sonore que l'on souhaite mesurer. La pondération A est utilisée pour mesurer des niveaux faibles à moyens.

branché sur l'une des prises USB du téléviseur. Il suffisait ensuite de lancer la lecture d'une nouvelle vidéo au début de chaque prise de son.

Que ce soit pour les JT ou les météos, nous avons démarré l'enregistrement après les « génériques » d'introduction. Ces génériques contenaient souvent de la musique ou un tapis sonore trop présents par rapport au signal vocal.

Pour nommer les enregistrements du corpus de voix télévisuelles nous avons adopté la convention suivante :

- en premier figure le **microphone utilisé** : « Behringer » ou « TL103 » ;
- on indique ensuite la **chaîne** dont est issu le programme enregistré : « tf1 », « fr2 », « fr3 », ou « m6 » ;
- vient après le **nom du programme**. Pour les JT, le nom dépend de la chaîne :

TF1 :	France 2 :	France 3 :	M6 :
- « 13h »	- « 13h »	- « 1213 »	- « 1245 »
- « 20h »	- « 20h »	- « 1920 »	- « 1945 »

  Pour les bulletins météorologiques, on indique simplement : « meteo » ;
- pour finir, on ajoute la **date** de diffusion du programme sous la forme « jjmmaaaa », par exemple « 05042017 » pour le 5 avril 2017 ;
- tous ces éléments sont séparés par un ***underscore*** : « _ ».

Exemple : si le fichier correspond à un enregistrement réalisé au *Behringer ECM8000* et qu'il s'agit du JT 19.45 du 4 avril 2017 diffusé sur M6, alors on lui attribuera le nom : « Behringer_m6_1945_04042017 ».

Nous avons enregistré au total 15 journaux télévisés et 7 météos. La durée complète d'enregistrement pour chaque microphone est de 7 heures 58 minutes et 24 secondes (pour rappel, un descriptif détaillé des programmes télévisuels utilisés dans la composition

de ce corpus est disponible en annexe). La durée totale de ce corpus est donc de 15 heures 56 minutes et 48 secondes.

2.4 Conclusion

Nous venons de retracer la manière dont nous avons constitué deux corpus oraux (voix parlées et voix télévisuelles) en étudiant les différents choix qui ont guidé leur composition ainsi que l'élaboration du protocole de leur enregistrement. Ces corpus ont été créés pour être analysés par IDS et ainsi nous permettre d'obtenir des découpages fréquentiels adaptés à la voix française.

3 IDS et découpages fréquentiels

Dans ce chapitre nous allons décrire plus en détail les principes de l'analyse/resynthèse IDS. Nous expliquerons ensuite la méthode permettant d'obtenir un découpage fréquentiel adapté pour un corpus d'enregistrements donné puis nous commenterons les résultats obtenus à partir des corpus évoqués dans le chapitre 2.

3.1 Description de l'IDS

3.1.1 Version analogique

Comme il l'indique dans son bulletin du GAM (Groupe d'Acoustique Musicale) publié en décembre 1977, Émile Leipp a développé l'Intégrateur de Densité Spectrale suite à des réflexions autour de la qualité sonore [54].

Il cherchait à comprendre les mécanismes psychoacoustiques nous permettant d'évaluer la qualité d'un instrument de musique, d'une chaîne d'écoute, ou encore de l'acoustique d'une salle.

Partant du principe que pour saisir des informations esthétiques à l'origine de cette appréciation notre système auditif exploitait sûrement la mémoire à long terme, il en a conclu que les mécanismes mis en place dans notre cerveau devaient être **intégratifs**. Cependant comme ces mécanismes (notamment ceux responsables des sensations de « coloration » qui l'intéressaient particulièrement) étaient inobservables en direct, Émile Leipp a donc tenté de les simuler afin de pouvoir espérer les comprendre.

À cette époque il utilisait beaucoup le sonagraphe : un appareil permettant de représenter sur un diagramme l'évolution de la fréquence d'un son en fonction du temps. Cet outil était en fait utilisé pour relever les lignes mélodiques d'un chant ou d'une pièce instrumentale. Voici comment il explique avoir trouvé les bases de l'IDS à partir de ces relevés :

> *J'ai alors eu l'idée fort simple de recopier la ligne mélodique (d'une berceuse vietnamienne dans l'exemple de la figure 1a*[1]*) sur du papier quadrillé, millimétré. Ensuite j'ai tout simplement additionné sur chaque ligne millimétrée horizontale les points d'intersection avec la ligne mélodique. En reportant le total des points de chaque ligne sur un diagramme séparé (petit diagramme en bas à droite de la figure 1a), on obtenait une courbe de « taux d'occurrence » de chaque fréquence, d'où émergeaient des « bosses à allure gaussienne » dont les sommets représentent en fait les « notes » de la gamme que cerne le musicien (voir figure 1a).*
>
> *On peut bien entendu faire toutes les opérations (fastidieuses...) avec l'aide d'un ordinateur, ce que nous avons effectivement fait (figure 1b). Mais au lieu de ne considérer que la fondamentale d'un sonagramme de chant ou de musique, il suffit de prendre un sonagramme complet comportant les sons et tous leurs harmoniques ou partiels. On obtient alors, en procédant comme précédemment, la courbe de densité spectrale d'une séquence de musique indiquant la répartition statistique de l'énergie dans cette séquence.*
>
> *L'idée de base était trouvée : il suffisait de la développer, de concevoir et réaliser les moyens matériels pratiques pour faire ces relevés de densité spectrale. C'est à quoi nous nous sommes employés avec M. Sole et M. Sapaly dans le cadre de l'E.R.A. 537 du C.N.R.S.* [Ibid.]

1. Les figures évoquées dans cette citation (1a et 1b) sont disponibles à la page suivante.

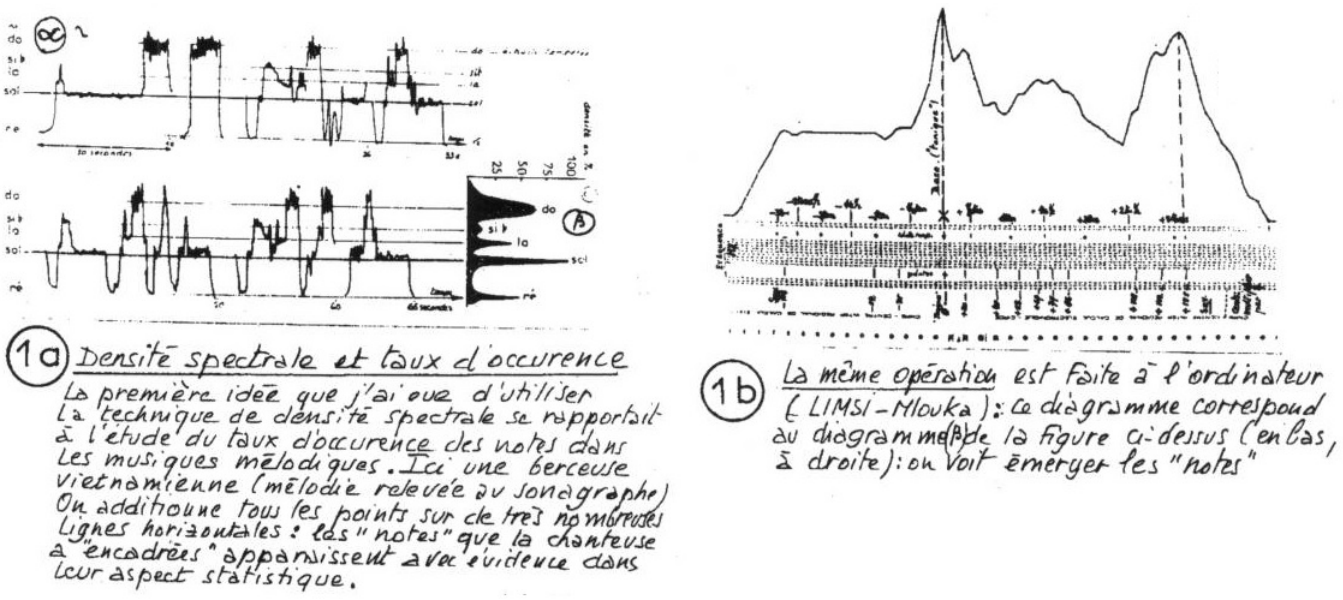


FIGURE 3.1 – Figures 1a et 1b évoquées par Émile Leipp dans le bulletin du GAM n°94, publié en 1977.

Émile Leipp a ensuite décidé de simplifier ces relevés en découpant la gamme des fréquences audibles par l'être humain selon 8 sous-bandes de fréquences. Pour ce faire il a constitué un auditoire de 30 sujets auquel il faisait écouter des œuvres orchestrales qu'il filtrait durant la lecture[2]. Il appliquait des filtres de réjection qu'il élargissait petit à petit jusqu'à ce que l'ensemble des sujets réagissent, en signalant par exemple l'absence de graves, d'aigus ou bien la présence d'un « creux » fréquentiel. Il notait ensuite la fréquence pour laquelle l'auditoire s'était manifesté et continuait le processus. Voici le découpage obtenu suite à ces expériences :

- **50 - 200 Hz** : bande des « basses » ;
- **200 - 400 Hz** : bande des « graves » ;
- **400 - 800 Hz** : bande du « médium grave » ;
- **800 - 1200 Hz** : bande du « médium » ;
- **1200 - 1800 Hz** : bande du « médium aigu » ;
- **1800 - 3000 Hz** : bande de l'« aigu » ;
- **3000 - 6000 Hz** : bande du « sur-aigu » ;
- **6000 - 15000 Hz** : bande de « stridence ».

2. Dans son bulletin Émille Leipp explique l'expérience en prenant pour exemple la « Grande Valse » du *Chevalier à la rose* de Richard Strauss [Ibid.].

Comme l'idée de départ était de travailler sur les mécanismes psychoacoustiques en œuvre dans l'évaluation de la qualité sonore, ce découpage permettait de se rapprocher d'une sensation auditive humaine puisque des tests perceptifs expérimentaux avaient servi à son élaboration.

De plus normaliser tous les relevés en utilisant les mêmes sous-bandes de fréquences permettait d'interpréter les mesures rapidement et d'en faciliter la comparaison. En effet Émile Leipp avait prévu de calculer la grandeur de l'énergie dans une sous-bande donnée en pourcentage de l'énergie totale. Pour comparer deux relevés il suffisait donc d'examiner les différences de pourcentage sous-bande par sous-bande. Leipp utilisait même du papier calque pour superposer différents relevés et les comparer de ce fait instantanément.

L'IDS ainsi développé permettait donc d'associer à n'importe quel signal audio un « portrait » de la répartition de son énergie spectrale, décomposé d'après les 8 sous-bandes de fréquences évoquées.

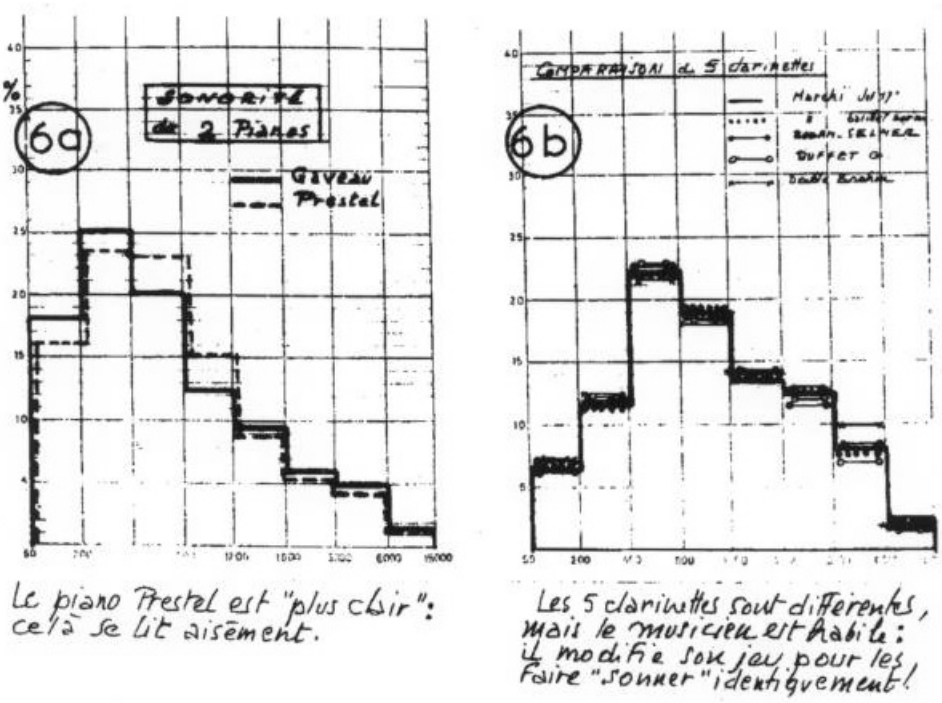


FIGURE 3.2 – « Portraits » IDS issus du bulletin du GAM n°94. À gauche, comparaison du profil IDS de 2 pianos ; à droite, même comparaison pour 5 clarinettes [Ibid.].

Expliquons plus précisément le fonctionnement de l'analyse IDS développée par Leipp, Sole et Sapaly.

- L'opération consistait à filtrer le signal analysé en suivant un jeu de filtres dont les fréquences de coupure correspondaient aux fréquences du découpage cité précédemment. Ces filtres étaient des filtres de *Butterworth*[3] analogiques possédant une pente de -12 dB par octave.

- Il s'agissait ensuite de calculer l'énergie cumulée des signaux de chaque sous-bande et l'énergie cumulée du signal total (signal étudié). On obtenait donc 9 valeurs d'énergie :
  - 1 pour chaque sous-bande, appelons les « e_1 », « e_2 », ..., « e_8 » ;
  - 1 pour le signal étudié, appelons la « E ».

- Enfin pour calculer l'énergie de toutes ces sous-bandes en pourcentage de l'énergie totale, il suffisait de faire le rapport e_n/E pour chacune d'entre elles. Par exemple e_1/E pour la bande des basses, e_2/E pour la bande des graves, e_3/E pour la bande du médium grave, etc.

Une fois l'IDS opérationnel Émile Leipp a effectué de nombreuses mesures et a trouvé plusieurs domaines d'application pour lesquels cet outil se révélait pertinent : étude de la sonorité des instruments de musique, de la coloration de la voix humaine, ou encore de l'acoustique des salles.

Le principe initial de l'IDS tel qu'il est décrit par Leipp en 1977 a été partiellement conservé dans la version numérique proposée par Laurent Millot. Les modifications de cette extension reposent sur les possibilités qu'apporte le numérique : filtrage précis facilement modifiable, capacité de stockage volumineuse, puissance de calcul importante, etc. [33, 55]

3. Un filtre de *Butterworth* est conçu pour posséder un gain constant dans sa bande passante.

3.1.2 Version numérique

Dans la version numérique de l'IDS, le découpage fréquentiel utilisé par Leipp a été étendu. Les 8 sous-bandes initiales ont été sauvegardées mais deux sous-bandes ont été rajoutées :

- **0 - 50 Hz** : pour analyser des phénomènes de très basses fréquences pouvant par exemple se manifester au niveau (voire à l'intérieur) des sources ;

- **15000 Hz - Fréquence de Nyquist**[4] **(F_N)** : pour observer ce qui peut exister au-delà de 15 kHz (voire au dessus de 20 kHz) et avoir, entre autres, une plus grande précision dans l'analyse des transitoires.

Le découpage adopté ici correspond donc à une extension du découpage proposé par Leipp en 1977, même si en principe n'importe quel découpage peut être défini puis utilisé. C'est d'ailleurs un point qui nous a particulièrement intéressé dans le cadre de cette étude et sur lequel nous allons revenir.

Les filtres utilisés pour l'analyse sont des filtres passe-bande RIF[5] (Réponse Impulsionnelle Finie) à phase nulle, ce qui permet de s'affranchir de toute distorsion de phase mais implique un fonctionnement en temps différé. Ils comportent un nombre important de coefficients et proposent donc une atténuation rapide (minimum - 80 dB en 5 Hz dans les bandes de transition) qui permet de limiter au maximum les recouvrements entre deux bandes consécutives.

Ces filtres sont calculés à partir de l'algorithme suivant.

4. La fréquence de Nyquist (F_N) est égale à la moitié de la fréquence d'échantillonnage. Pour garantir la restitution fidèle d'un signal dans le cas d'un échantillonnage, toutes les fréquences contenues dans ce signal doivent être inférieures ou égales à F_N. En prenant une fréquence d'échantillonnage de 44100 Hz (comme nous l'avons fait pour l'enregistrement des corpus, cf. chapitre 2) la fréquence de Nyquist est donc égale à 22050 Hz.

5. Les filtres RIF sont des filtres numériques stables, peu sensibles aux erreurs de quantification et simples à implémenter dans un système numérique de traitement.

- On part d'un filtre fréquentiel idéal, noté $\delta[n]$, qui possède un gain unitaire sur la gamme de fréquences [0, F_N]. On retranche à $\delta[n]$ la réponse impulsionnelle du filtre passe-bas à phase nulle dont la fréquence de coupure est égale à la limite inférieure de la dernière sous-bande (f_K pour un découpage utilisant K sous-bandes). Ce filtre passe-bas est noté h_{PB,f_K}. On obtient pour $h_K[n]$ le filtre de sous-bande K :

$$h_K[n] = \delta[n] - h_{PB,f_K}[n],$$

$h_K[n]$ correspond donc bien au filtre passe-bande ne contenant idéalement que la gamme de fréquences $[f_K, F_N]$ [6].

- Le même processus est itéré pour chacune des autres sous-bandes excepté la première. Ainsi pour la sous-bande $K-1$ délimitée par la gamme de fréquences $[f_{K-1}, f_K]$, on calcule le filtre passe-bande :

$$h_{K-1}[n] = h_K[n] - h_{PB,f_{K-1}}[n].$$

- On a donc une dérivation itérative des filtres passe-bande. Pour le calcul du filtre de sous-bande $h_k[n]$ correspondant à la gamme de fréquences $[f_k, f_{k+1}]$, on obtient :

$$h_k[n] = h_{k+1}[n] - h_{PB,f_k}[n],$$

excepté pour $k = 1$.

- En effet pour la première sous-bande correspondant à la gamme de fréquences $[0, f_2]$, on a :

6. En reprenant le découpage de Leipp étendu cette gamme de fréquences correspond à la sous-bande [15000 Hz - F_N].

$$h_1[n] = h_{PB,f_2}[n].$$

Comme les filtres utilisés sont complémentaires, il est possible de re-synthétiser le signal analysé par simple sommation de toutes les sous-bandes[7].

La re-synthèse (partielle à totale) s'écoute, ce que ne permettait pas la version analogique de l'IDS. On peut par exemple choisir de n'écouter que la 5$^{\text{ème}}$ sous-bande, ou bien un mixage de la 2$^{\text{ème}}$ et de la 6$^{\text{ème}}$, ou encore le signal complet. De plus on peut modifier le niveau de ces sous-bandes et donc leur poids dans la re-synthèse, de la même manière que l'on ferait varier le niveau des pistes d'une console de mixage.

Le poids relatif de chacune des sous-bandes n'est plus indiqué en pourcentage mais en décibels et ces informations sont complétées par l'affichage du niveau moyen du signal étudié. Cette présentation des poids relatifs en dB permet de mettre en évidence des comportements qui auraient pu être masqués avec l'échelle linéaire des pourcentages.

L'appréciation de l'analyse n'est donc plus uniquement visuelle (« portraits » IDS) puisque l'on peut entendre les résultats. S'agissant de signaux sonores, il semble intéressant de pouvoir comparer de manière auditive la perception que nous avons du poids relatif obtenu pour chaque sous-bande aux résultats « réels » donnés en décibels.

3.2 Exemples de découpages fréquentiels existants

Jusqu'ici nous avons mentionné le découpage fréquentiel présenté par Émile Leipp ainsi que son extension proposée par Laurent Millot. Mais il en existe bien d'autres et nous allons maintenant les évoquer (de manière non exhaustive).

7. Du fait des contraintes imposées lors de la détermination des filtres (filtres RIF à phase nulle opérant en temps différé) cette re-synthèse donne une erreur théoriquement nulle, non audible en pratique.

- Eberhard Zwicker, qui était un acousticien allemand, a par exemple proposé un découpage fréquentiel du spectre sonore en 24 « bandes critiques », spécialement défini pour les mesures de la sonie. En voici les fréquences de coupure :

  100 ; 200 ; 300 ; 400 ; 510 ; 630 ; 770 ; 920 ; 1080 ; 1270 ; 1480 ; 1720 ; 2000 ; 2320 ; 2700 ; 3150 ; 3700 ; 4400 ; 5300 ; 6400 ; 7700 ; 9500 ; 12000 ; et 15500 Hz [56].

- Nous pouvons aussi citer les découpages en bandes d'octave ou bandes de tiers d'octave qui sont notamment utilisés pour les mesures du bruit (dans des domaines comme la protection auditive ou l'isolation acoustique exemple) et dont les fréquences centrales sont les suivantes :

  - 16 ; 31,5 ; 63 ; 125 ; 250 ; 500 ; 1000 ; 2000 ; 4000 ; 8000 ; et 16000 Hz pour les analyses par bandes d'octave ;
  - 13 ; 20 ; 25 ; 31,5 ; 40 ; 50 ; 63 ; 80 ; 100 ; 125 ; 160 ; 200 ; 250 ; 320 ; 400 ; 500 ; 640 ; 800 ; 1000 ; 1250 ; 1600 ; 2000 ; 2500 ; 3200 ; 4000 ; 5000 ; 6400 ; 8000 ; 10000 ; 12500 ; et 16000 Hz pour les analyses par bandes de tiers d'octave [57].

- En musique et particulièrement en mixage, nous retrouvons ces découpages dans les outils de type *equalizer*[8] (EQ) mais aussi dans le langage commun, afin de pouvoir mettre des mots précis sur des sensations auditives de hauteur.

  Voici par exemple une description « psychoacoustique » du spectre audible (diminué) effectuée en suivant les bandes d'octave et proposée par Pascal Spitz dans un cours sur l'analyse sonore à l'ENS Louis-Lumière :

  · **30 - 60 Hz** : extrême grave, sensation tripale (le son est plus perçu par le corps que par les oreilles) ;

8. Un *equalizer* est un logiciel ou un appareil de traitement sonore qui permet de filtrer et/ou d'amplifier différentes bandes de fréquences d'un signal audio (souvent en bandes d'octave ou de tiers d'octave).

- **60 - 120 Hz** : centre-grave, sensation de rondeur ;
- **120 - 250 Hz** : haut-grave, « corps » des sources sonores ;
- **250 - 500 Hz** : bas-médium, sensation de résonance ;
- **500 - 1000 Hz** : centre-médium, octave de transition (passage des sensations associées à la partie basse aux sensations associées à la partie haute) ;
- **1000 - 2000 Hz** : haut-médium, sensation de clarté ;
- **2000 - 4000 Hz** : bas-aigu, sensation d'agressivité ;
- **4000 - 8000 Hz** : centre-aigu, sensation de présence, de proximité ;
- **8000 - 16000 Hz** : extrême-aigu, sensation de précision.

Dans le cadre de cette étude il nous fallait un découpage fréquentiel caractéristique de la voix française. Or parmi les découpages que nous avons cités jusqu'à présent aucun n'est conçu selon ce paramètre.

Toutefois comme nous l'avons déjà évoqué, n'importe quel découpage peut être défini puis utilisé afin d'effectuer une analyse IDS. Compte tenu de l'algorithme employé pour faire les calculs, on peut en effet envisager l'utilisation d'un découpage adapté à la situation testée.

C'est pourquoi nous avons fait le choix de ne pas emprunter des découpages pré-existants mais d'obtenir ceux dont nous avions besoin, en constituant des corpus oraux dans un premier temps puis en les analysant. Nous allons donc expliquer la méthode qui a permis de retenir des découpages fréquentiels à partir des corpus décrits dans le chapitre 2.

3.3 Méthode pour la détermination de découpages fréquentiels

3.3.1 Principe des spectres cumulés

Définition du spectre cumulé

La recherche de découpages fréquentiels adaptés à l'objet d'étude s'appuie sur le calcul préalable des spectres cumulés pour chacun des éléments constituant le corpus étudié [58, 59]. Afin d'obtenir la définition du spectre cumulé on part de l'expression théorique de la TFTD pour un signal numérique $s[n]$ échantillonné à la fréquence F_N :

$$S[\nu] = \sum_{n=-\infty}^{+\infty} s[n] \ \exp -2\pi \mathrm{j}\nu n,$$

expression où ν correspond à la fréquence réduite définie par $\nu = f/F_s$ avec f la fréquence en Hertz.

En pratique les signaux constituant le corpus ont un nombre d'échantillons fini (potentiellement variable d'un signal à l'autre) que l'on notera N dans la suite. La définition pratique de $S[\nu]$ change puisque l'on a maintenant :

$$S[\nu] = \sum_{n=0}^{N-1} s[n] \ \exp -2\pi \mathrm{j}\nu n.$$

On sait que $\nu = f/F_s$, on peut donc réécrire la définition pratique de $S[\nu]$ en adoptant la variable f au lieu de ν :

$$S[f] = \sum_{n=0}^{N-1} s[n] \ \left(\exp -2\pi \mathrm{j}\frac{n}{F_s} \right)^f.$$

Comme $\exp -2\pi \mathrm{j}\frac{n}{F_s}$ est périodique de période F_s, il s'avère intéressant d'exprimer $s[n]$ en fonction de F_s dès lors que N est bien plus grand que F_s. Si l'on note M l'entier directement supérieur à N/F_s on peut ainsi écrire pour $s[n]$:

$$s[n] = s[k + m.F_s], \quad k \in [0, F_s - 1],\ m \in [0, M[.$$

Si l'on rajoute au besoin le nombre nécessaire de zéros à la fin du signal $s[n]$ pour considérer que celui-ci comporte en fait $M.F_s$ échantillons (opération de *zero-padding*[9] pour le dernier bloc), on peut alors réécrire $S[f]$ sous la forme :

$$S[f] = \sum_{m=0}^{M-1} \sum_{k=0}^{F_s-1} s[k + m.F_s] \left(\exp -2\pi \mathrm{j}\frac{k + m.F_s}{F_s} \right)^f.$$

Comme m est un entier, $\exp -2\pi \mathrm{j}\frac{k+m.F_s}{F_s} = \exp -2\pi \mathrm{j}\frac{k}{F_s}$. On peut donc encore simplifier l'expression définissant $S[f]$:

$$S[f] = \sum_{m=0}^{M-1} \sum_{k=0}^{F_s-1} s[k + m.F_s] \left(\exp -2\pi \mathrm{j}\frac{k}{F_s} \right)^f.$$

On constate alors que $S[f]$ correspond à la somme de M TFTD consécutives sur des blocs de F_s échantillons, c'est-à-dire au cumul de M TFTD successives de taille F_s.

La fréquence d'échantillonnage est imposée par les fichiers à analyser, ici on aura $F_s = 44100$ Hz. Le nombre de blocs M dépend de la longueur du fichier.

Le seul paramètre que nous pouvons choisir est donc la gamme des fréquences calculées, soit l'ensemble des fréquences f.

9. Le *zero-padding* est une technique consistant à ajouter des « zéros » à la fin d'un signal pour augmenter sa longueur. En ajoutant des échantillons on augmente la précision spectrale. Les maxima et minima du spectre sont donc mieux localisés après analyse.

Puisque nous voulons *in fine* obtenir des découpages fréquentiels avec des fréquences entières, nous avons choisi d'opter pour les fréquences allant de 0 à la fréquence de Nyquist par pas de 1 Hz.

Nous aurions pu choisir un pas plus faible (0,5 voire 0,1 Hz par exemple) mais cela aurait augmenté le nombre de fréquences à calculer puis à étudier pour n'en retenir de toute façon qu'une dizaine.

Finalement, on mettra en œuvre la définition suivante pour calculer ce que l'on nomme dans la suite un « spectre cumulé » :

$$S[f] = \sum_{m=0}^{M-1} \sum_{k=0}^{F_s-1} s[k + m.F_s] \left(\exp -2\pi \mathrm{j} \frac{k}{F_s} \right)^f .$$

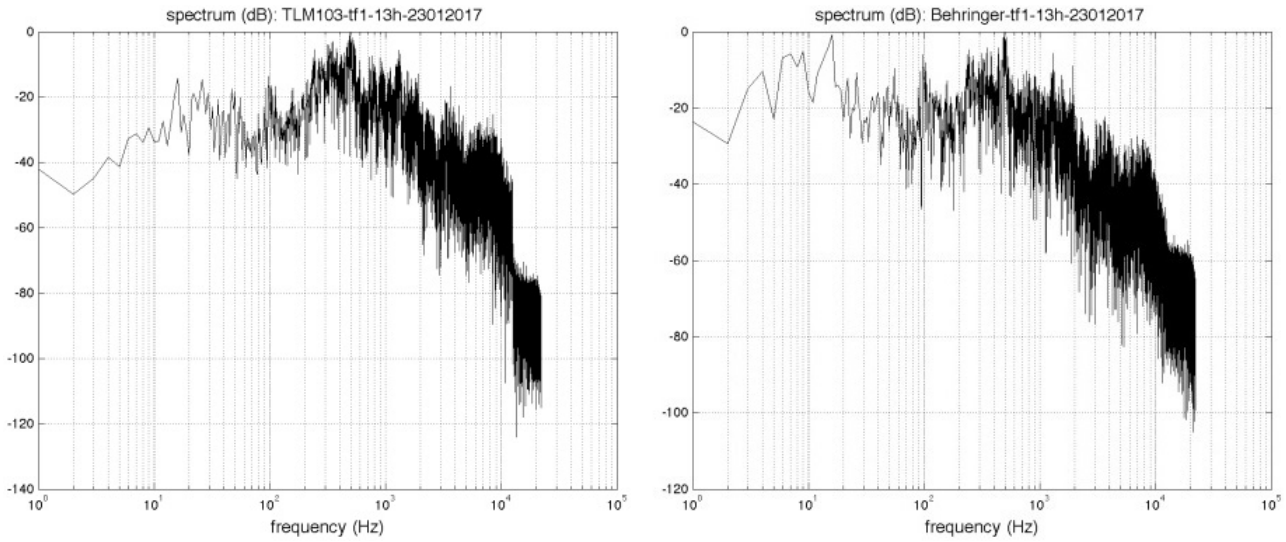


FIGURE 3.3 – Exemples de spectres cumulés pour l'enregistrement : du JT de 13h (TF1) du 23 janvier 2017 avec le *TLM103* à gauche ; de ce même JT mais avec le *Behringer* à droite.

Spectres cumulés d'amplitude

L'idée pour la recherche d'un découpage fréquentiel adapté à un corpus donné consiste à calculer le spectre cumulé $S[f]$ pour chaque mesure puis le **spectre cumulé d'amplitude** $\left|S[f]\right|$ associé (cf. par exemple les tracés de la figure 3.3). On fait ensuite la somme de tous ces spectres cumulés d'amplitude que l'on appelle spectres cumulés de corpus. Cette somme va faire émerger les creux spectraux qui sont pertinents à l'échelle du corpus et faire disparaître les « petits » creux qui ne sont valables que pour un nombre très limité de mesures (voire une seule mesure).

C'est à partir de l'étude et la sélection des creux spectraux ayant résisté que l'on va pouvoir définir les limites des sous-bandes fréquentielles constituant le découpage fréquentiel : on va donc opérer une sélection dans les minima locaux du spectre cumulé de corpus (exemples de spectres cumulés de corpus en figure 3.4). Toutefois comme nous partons de spectres cumulés de corpus calculés pour 22050 fréquences, même en décidant de ne sélectionner que les minima locaux on aboutirait à un nombre trop conséquent de sous-bandes potentielles.

Nous allons donc chercher à regrouper des sous-bandes potentielles consécutives afin de constituer des sous-bandes plus larges et ainsi simplifier la description spectrale des signaux étudiés.

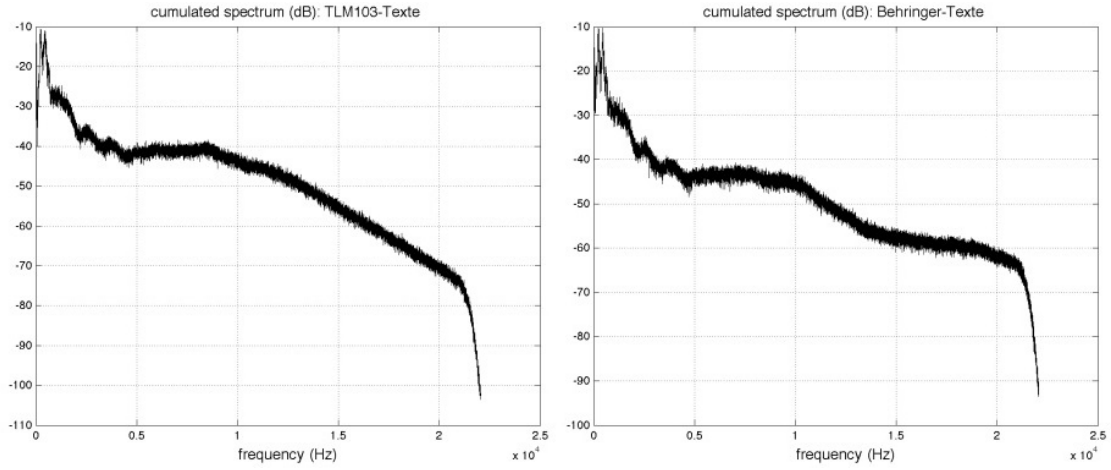


FIGURE 3.4 – Exemples de la somme des spectres cumulés d'amplitude pour : le sous-corpus « Texte » (issu du corpus de voix parlées) enregistré au *TLM103* à gauche ; ce même sous-corpus mais enregistré au *Behringer* à droite.

Redressement des spectres cumulés

Quand on considère les spectres de corpus que l'on obtient (cf. par exemple figure 3.4), on constate qu'ils présentent une allure globale faisant apparaître de larges bosses gênantes pour déterminer et sélectionner les minima locaux.

Afin de réduire au maximum la courbure globale du spectre cumulé de corpus, on filtre celui-ci sur la gamme des fréquences étudiées à l'aide d'un filtre moyenneur glissant prenant en compte 9 échantillons consécutifs. On obtient alors des portions de spectre cumulé redressées, à l'image des exemples de la figure 3.5, qui permettent l'émergence des minima locaux. De manière abusive on nomme ces portions de spectre : « spectres redressés ».

3.3.2 Recherche des minima locaux

Pour sélectionner les minima locaux on travaille par sous-octave en commençant par la bande 0-63 Hz. Pour la bande de fréquences étudiée on affiche sur le spectre redressé local les minima locaux déterminés automatiquement parmi un nombre limité de minima (entre 12 et 16 par sous-octave pour un spectre cumulé de corpus, moins pour un spectre cumulé d'un seul élément).

Parmi ces minima locaux on ne retient au plus que ceux qui constituent des minima pour l'ensemble des 3 et 5 points pertinents : le minima et les deux points l'entourant ; le minima et les deux couples de points l'entourant. Cette procédure permet d'éviter d'obtenir tous les minima spectraux pour la gamme des fréquences testées et ainsi de simplifier un peu la recherche des limites de sous-bandes (cf. exemples de la figure 3.5).

Une fois tous les minima locaux identifiés pour chacun des sous-octaves, on réalise une analyse IDS du corpus en utilisant le découpage fréquentiel issu de ce travail : les fréquences de coupure correspondent aux minima locaux retenus. Dans le cas du corpus de voix télévisuelles par exemple, nous avons obtenu 38 minima locaux à cette étape.

Cela donnait donc une analyse IDS en 40 sous-bandes.

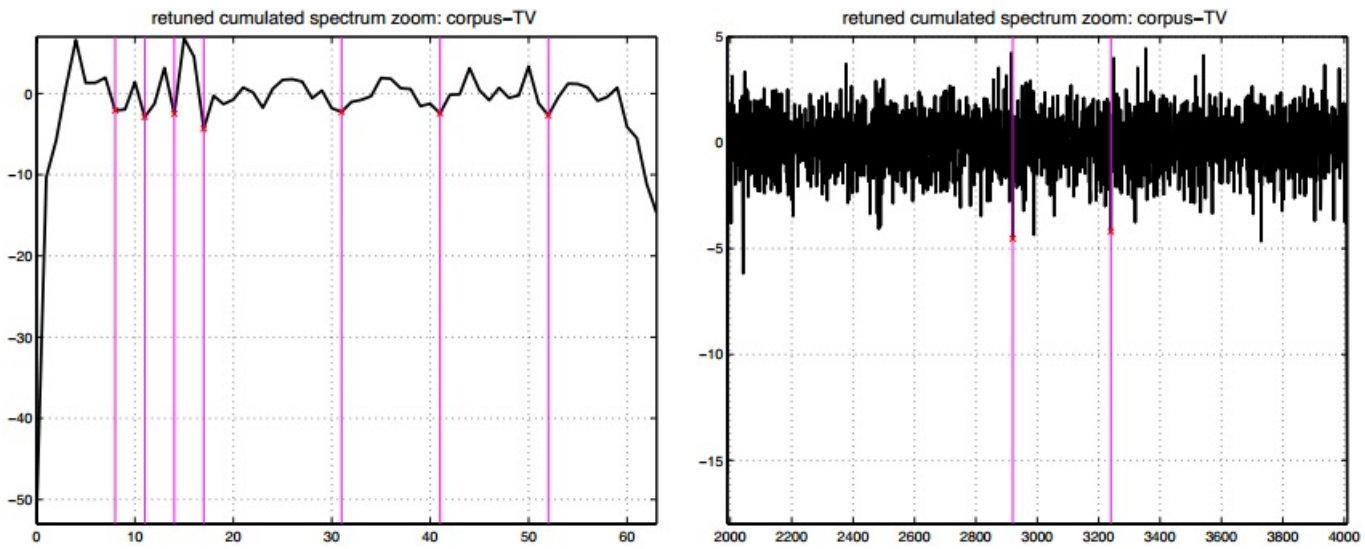


FIGURE 3.5 – Exemples des premiers minima locaux retenus pour le corpus de voix télévisuelles, dans le sous-octave : 0 - 63 Hz à gauche et 1990 - 4010 Hz à droite.

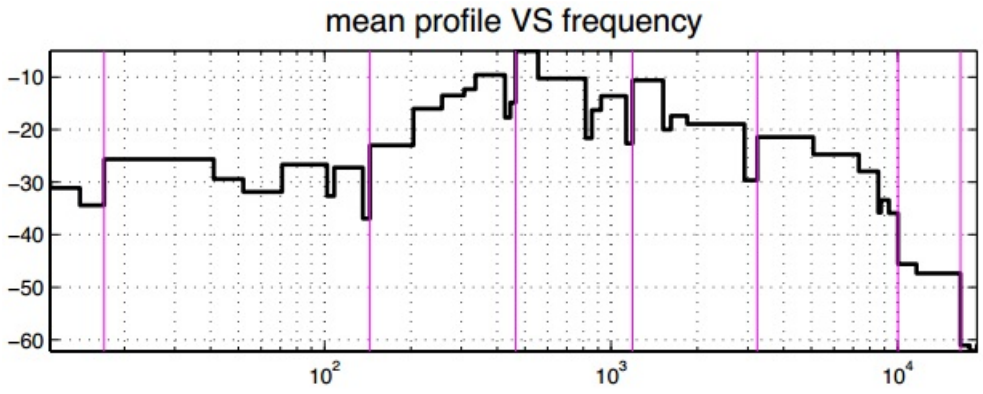


FIGURE 3.6 – Profil IDS du corpus de voix télévisuelles, obtenu à partir de la première analyse en 40 sous-bandes, et sélection finale des sous-bandes.

À partir du profil IDS obtenu on procède de nouveau à une recherche de minima locaux pour obtenir un découpage fréquentiel composé d'une dizaine de sous-bandes environ. Pour reprendre l'exemple du corpus de voix télévisuelles nous avons obtenu au final 8 sous-bandes (cf. figure 3.6).

3.4 Résultats pour les corpus étudiés

Voici tout d'abord les découpages fréquentiels retenus pour les deux corpus.

Corpus de voix parlées :

- **0 - 40 Hz** ;
- **40 - 83 Hz** ;
- **83 - 229 Hz** ;
- **229 - 625 Hz** ;
- **625 - 867 Hz** ;
- **867 - 2667** ;
- **2667 - 7237** ;
- **7237 - 16087 Hz** ;
- **16087 - 22050 Hz**.

Corpus de voix télévisuelles :

- **0 - 17 Hz** ;
- **17 - 144 Hz** ;
- **144 - 464 Hz** ;
- **464 - 1189 Hz** ;
- **1189 - 3240 Hz** ;
- **3240 - 10038 Hz** ;
- **10038 - 16565 Hz** ;
- **16565 - 22050 Hz**.

La fréquence d'échantillonnage pour l'enregistrement des corpus étant de 44100 Hz, la fréquence de Nyquist est donc de 22050 Hz. Nous avons décidé de garder cette fréquence comme valeur maximale des découpages et donc de considérer des portraits IDS prenant en compte toutes les fréquences entre 0 et 22050 Hz réparties en 8 à 9 sous-bandes.

Au vu de ces résultats nous pouvons déjà confirmer que la nature de la source sonore enregistrée a une influence sur le découpage fréquentiel retenu. Compte tenu des différences très marquées entre le son d'une télévision et le son d'une voix pour notre perception auditive, c'est une conclusion à laquelle nous pouvions naturellement nous attendre.

Observons maintenant les résultats obtenus pour les sous-corpus de voix parlées en fonction des microphones et des *stimuli* utilisés. Les découpages retenus sont rassemblés dans le tableau 3.7.

	Liste de mots	Texte	Interview
TLM 103	· 0 – 39 Hz · 39 – 109 Hz · 109 – 299 Hz · 299 – 403 Hz · 403 – 739 Hz · 739 – 1441 Hz · 1441 – 3072 Hz · 3072 – 4509 Hz · 4509 – 13210 Hz · 13210 – 22050 Hz	· 0 – 47 Hz · 47 – 108 Hz · 108 – 163 Hz · 163 – 331 Hz · 331 – 798 Hz · 798 – 1048 Hz · 1048 – 3388 Hz · 3388 – 5205 Hz · 5205 – 14307 Hz · 14307 – 22050 Hz	· 0 – 14 Hz · 14 – 47 Hz · 47 – 84 Hz · 84 – 168 Hz · 168 – 345 Hz · 345 – 740 Hz · 740 – 1931 Hz · 1931 – 3267 Hz · 3267 – 6811 Hz · 6811 – 14506 Hz · 14506 – 18658 Hz · 18658 – 22050 Hz
Behringer	· 0 – 29 Hz · 29 – 99 Hz · 99 – 109 Hz · 109 – 221 Hz · 221 – 403 Hz · 403 – 758 Hz · 758 – 1444 Hz · 1444 – 4378 Hz · 4378 – 17578 Hz · 17578 – 22050 Hz	· 0 – 47 Hz · 47 – 110 Hz · 110 – 331 Hz · 331 – 347 Hz · 347 – 575 Hz · 575 – 1558 Hz · 1558 – 3388 Hz · 3388 – 4769 Hz · 4769 – 15294 Hz · 15294 – 22050 Hz	· 0 – 40 Hz · 40 – 338 Hz · 338 – 651 Hz · 651 – 1213 Hz · 1213 – 2166 Hz · 2166 – 4490 Hz · 4490 – 12017 Hz · 12017 – 22050 Hz

FIGURE 3.7 – Découpages fréquentiels retenus pour les sous-corpus de voix parlées.

Tout d'abord comparons les résultats obtenus pour les deux microphones. Nous retrouvons parfois quelques similitudes sur certaines fréquences de sous-bandes pour un même *stimulus* mais dans l'ensemble les découpages sont quand même plutôt différents. Ce n'est pas particulièrement surprenant étant donné les disparités notables entre un microphone à directivité cardioïde et un microphone omnidirectionnel.

Nous pouvons donc confirmer que le dispositif d'enregistrement possède une influence sur le découpage fréquentiel retenu.

Comparons maintenant les *stimuli*. De la même façon que pour les microphones les découpages sont assez différents selon le *stimulus* étudié. On observe notamment que les résultats obtenus pour l'*interview* sont très hétérogènes et possèdent même un nombre de sous-bandes différent des autres découpages, que ce soit pour le *Behringer* ou le *TLM103*.

Ici le fait de trouver des variations entre les découpages n'était pas aussi prévisible qu'avec les critères de sources et de microphones. Même si l'on proposait une palette d'intonations assez variée d'un enregistrement à l'autre, la nature du *stimulus* étudié restait la voix humaine dans les 3 cas.

Nous pouvons maintenant affirmer que le « type » de parole enregistré a lui aussi une influence sur le découpage fréquentiel retenu.

Finalement nous avons obtenu autant de découpages qu'il y avait de sous-corpus, ce qui prouve que non seulement la source sonore a une influence sur le découpage fréquentiel adapté mais qu'il en est de même pour le microphone utilisé et pour la nature du *stimulus* enregistré.

3.5 Conclusion

Nous avons décrit la version analogique de l'IDS présentée par Émile Leipp en 1977 puis nous avons exposé les modifications apportées à cet outil dans sa version numérique développée par Laurent Millot. Nous avons ensuite étudié la méthode permettant d'obtenir un découpage fréquentiel adapté à un objet d'étude donné pour enfin commenter les découpages obtenus à partir des corpus évoqués dans le chapitre 2.

De ces résultats nous avons déduit que lorsque l'on souhaite obtenir un découpage fréquentiel à partir d'un corpus donné, les enregistrement doivent être effectués en appliquant un protocole qui retranscrit au mieux la situation réelle pour laquelle ce découpage est nécessaire. Par exemple dans le cadre de notre étude, il ne faudrait donc pas utiliser les microphones *Behringer* ou *TLM103* pour les enregistrements des corpus mais plutôt des microphones directement embarqués dans le réducteur de bruit final (prothèse auditive notamment).

4 L'« IDS *Speech Enhancer* » (IDSSE) et tests perceptifs

À partir des découpages fréquentiels retenus dans le chapitre 3 nous avons conçu un outil dont le but est de faire émerger un signal de parole lorsqu'il est plongé dans une ambiance bruitée. Nous avons nommé cet outil l'« IDS *Speech Enhancer* » (IDSSE). Il nous a permis de faire passer des tests perceptifs à un certain nombre de sujets, afin d'obtenir les filtres à réaliser pour améliorer la compréhension de la parole dans une ambiance sonore donnée.

Dans ce chapitre nous exposerons tout d'abord le protocole des tests perceptifs effectués par les sujets. Nous analyserons ensuite les différents constituants de l'IDSSE pour mieux en comprendre le fonctionnement. Enfin nous mentionnerons les conclusions de la première session de tests et nous en déduirons les amendements à apporter pour optimiser le déroulement de l'expérience et la pertinence de ses résultats.

4.1 Présentation générale du test

Tout d'abord avant d'effectuer le test en lui-même, le sujet réalise un examen audiométrique. Cet examen possède deux intérêts :

- distinguer les sujets présentant une déficience auditive des normo-entendants ;
- observer la façon dont cette déficience influence ou non les résultats obtenus pour les malentendants.

Pour réaliser l'examen nous utilisons un appareil *Interacoustics AS208* et un casque audiométrique TDH39 pour lequel il a été étalonné. Cet audiomètre permet d'effectuer des mesures d'audiométrie tonale de 125 Hz à 8000 Hz avec un aténuateur de -10 à 100 dB HL [1]. Pour procéder à l'examen nous avons suivi les indications du Guide des Bonnes Pratiques en Audiométrie de l'Adulte [36] :

> *Le test commence* [...] *à 1 000 Hz, se poursuit par les fréquences aiguës en ordre croissant, confirme le seuil à 1 000 Hz et se termine par les fréquences graves en ordre décroissant.*
>
> *Il est recommandé d'effectuer la recherche des seuils selon la méthode des niveaux ascendants. Des pas de 5 dB sont utilisés en pratique courante.*

En applicant cette méthode à notre audiomètre, l'ordre des fréquences pour l'examen a donc été celui-ci :

1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, 8000 Hz, 1000 Hz (vérification du seuil), 750 Hz, 500 Hz, 250 Hz, 125 Hz.

Pour chaque fréquence on débute avec une atténuation de -10 dB HL (plus petite valeur fournie par l'appareil) puis on augmente le niveau de cette fréquence par pas de 5 dB jusqu'à ce que le sujet entende le son (pour se manifester il appuie sur un bouton). On note alors la valeur d'atténuation correspondant à la fréquence étudiée sur un graphe. En réitérant ce processus pour les 11 fréquences on obtient une courbe audiométrique. L'examen s'effectue séparément pour les oreilles. Il faut donc le réaliser deux fois pour obtenir une courbe par oreille.

En temps normal il est conseillé de faire plusieurs mesures successives pour valider le seuil de chaque fréquence testée. Cependant nous ne voulions pas fatiguer le sujet avant même que le « vrai » test ait commencé. Un parcours unique durant déjà 8 minutes environ (4 minutes par oreille), il n'était pas question d'en effectuer davantage.

1. Les caractéristiques techniques de cet audiomètre sont disponibles en annexe.

Cet examen devait permettre d'avoir un aperçu sommaire de la perte auditive du patient et non pas d'en connaître tous les détails. Pour cela il aurait fallu ajouter un test d'audiométrie vocal, une mesure des oto-émissions acoustiques, un examen en conduction osseuse, etc. Ici il était plutôt question de réaliser un « dépistage auditif ».

Une fois l'examen préliminaire effectué, le sujet peut débuter le test avec l'IDS *Speech Enhancer*. Nous allons le présenter dans sa forme générale puis nous reviendrons en détails sur chaque élément dans la section suivante.

- Le sujet s'assoit face à un ordinateur et inscrit dans l'interface affichée à l'écran certaines informations personnelles : son sexe, son âge, le département dans lequel il a grandi, le département dans lequel il vit actuellement et son niveau d'expérience concernant les outils de traitement du son (sujet « candide » ou « expérimenté »). Ces informations permettent d'analyser les résultats selon différents critères et d'attribuer un « nom de code » au sujet.

- Le sujet s'équipe ensuite d'un casque audio. On lui propose alors d'écouter 5 « scènes » pour lesquelles des signaux vocaux ont été intégrés dans des ambiances sonores bruyantes. Au préalable, des analyses IDS de ces scènes ont été effectuées en suivant 3 découpages fréquentiels adaptés aux *stimuli* de parole étudiés. Le sujet écoute donc en fait la re-synthèse IDS de ces scènes.

- À l'aide d'une interface munie de *faders*, le sujet peut modifier le poids des sous-bandes de ces scènes. Il lui est demandé de faire varier le niveau de ces sous-bandes jusqu'à ce que le signal vocal soit pour lui le plus compréhensible possible.

- Une fois cette étape terminée il indique une note de 0 à 5 correspondant à sa compréhension de la voix « *post* réglage » pour la scène testée et enregistre la « balance spectrale » obtenue. Il répète ensuite ce processus pour les autres scènes.

L'idée est de comparer les balances spectrales effectuées par tous les sujets pour chaque scène afin d'observer si certaines tendances se dessinent. Si c'est le cas nous pouvons alors en déduire l'allure du filtre à réaliser pour améliorer la compréhension de la parole dans l'ambiance en question.

4.2 Description de l'IDS *Speech Enhancer*

4.2.1 Diffusion

Le son diffusé pendant les tests avec l'IDSSE représente le signal de sortie qu'auraient des prothèses auditives en situation réelle et non l'audition naturelle du patient. C'est pourquoi il a fallu trouver un moyen de diffusion se rapprochant de manière convaincante d'un appareillage auditif.

L'utilisation d'enceintes a tout de suite été écartée. En effet la diffusion sonore aurait été en champ libre alors que nous souhaitions rapprocher la source au plus près des oreilles.

Nous aurions pu utiliser des écouteurs de type « intra-auriculaires » qui correspondaient bien à ces attentes. Mais porter des écouteurs sur toute la durée du test aurait pu causer une gêne pour le sujet étant donné la qualité de confort très moyenne que propose ce genre de matériel.

Nous avons donc opté pour l'utilisation d'un casque audio fermé (isolant) qui alliait les critères de proximité et de confort.

Le niveau de diffusion est le même pour tous les sujets afin qu'ils effectuent les réglages en partant d'une base commune. C'est un point essentiel si l'on veut pouvoir ensuite comparer les résultats.

4.2.2 *Stimuli*

Bruit

Nous avons décidé que pour les tests effectués avec l'IDSSE, le bruit devait être une ambiance sonore.

Notre volonté était de se rapprocher au maximum d'une expérience réelle où la compréhension de la parole est rendue difficile par le bruit. Or dans la vie de tous les jours, le bruit que l'on rencontre est rarement un sinus pur, un bruit blanc, un bruit rose, etc. Nous avons donc sélectionné cinq environnements sonores représentatifs de situations problématiques pour l'écoute :

- une ambiance de **café**, où les voix des clients sont clairement distinctes ;
- une ambiance de **marché**, avec un brouhaha plus important et des voix très présentes par moment ;
- une ambiance de **gare**, où les voix ne sont plus du tout compréhensibles mais avec un niveau de bruit de fond élevé et réverbéré ;
- une ambiance de **ville**, avec une circulation moyenne, sans aucune voix ;
- une ambiance de **métro**, où le point d'écoute est à l'intérieur d'un wagon durant un trajet, sans aucune voix.

L'idée était d'avoir un panel d'ambiances présentant des caractéristiques variées de manière à examiner si ces variations se retrouvaient ensuite dans les balances spectrales effectuées par les sujets.

De plus nous voulions observer si pour une ambiance donnée, les sujets s'accordaient sur une balance similaire ou au contraire effectuaient des réglages discordants.

Parole

Les signaux de parole utilisés proviennent du corpus de voix parlées décrit dans le chapitre 2[2]. Il s'agit en effet de deux enregistrements effectués au *Neumann TLM103* de la lecture du texte de Didier Daeninckx, par un homme et par une femme.

Les prises de son ont été sélectionnées en fonction de la qualité de diction du locuteur, de manière à ce que la compréhension des sujets ne puisse pas être altérée par des problèmes de prononciation durant les tests. Les fichiers retenus sont :

- **TLM103_Texte_F21fra_03** pour la voix féminine ;
- **TLM103_Texte_H24fra_02** pour la voix masculine.

Le fait que le même texte soit utilisé à plusieurs reprises ne pose pas de problème dans le cadre de notre étude. Nous ne demandons pas au sujet de nous dire s'il comprend la parole dans une situation donnée mais de faire en sorte que cette compréhension soit la meilleure possible. Certes, connaître déjà le texte lui permet de comprendre plus facilement la voix d'une manière générale. Toutefois il reste en mesure d'évaluer si la qualité de cette compréhension s'améliore ou se dégrade en fonction des réglages qu'il effectue.

Si nous avons choisi d'utiliser les *stimuli* du sous-corpus « Texte » plutôt que les « *interviews* » (qui semblent pourtant plus proches d'une situation réelle) c'est parce le flux de parole est constant avec la lecture du texte. Dans un premier temps il est plus facile pour les sujets d'effectuer le réglage du niveau des sous-bandes avec une voix « continue » qu'avec une parole hachée, saccadée, pouvant présenter de longs moments de silence. Nous aurions pu réaliser des points de montage à l'intérieur de ces *interviews* afin de les linéariser mais nous aurions alors perdu l'aspect « naturel » que l'on recherchait.

2. Nous avons décidé de limiter l'étude aux voix parlées dans un premier temps et donc de ne pas introduire de voix télévisuelles. Nous ne voulions pas compliquer les tests dans le cadre de cette première approche. L'idée était d'obtenir un protocole stable et efficace que l'on pourrait ensuite élargir à d'autres *stimuli*.

Par ailleurs nous avons choisi d'utiliser les fichiers provenants des enregistrements du *TLM103* parce que la voix est plus claire et plus précise qu'avec le *Behringer*. Comme il fallait un signal vocal le plus « propre » possible pour les tests perceptifs, nous avons opté pour le *TLM103*. De plus il possède un directivité cardioïde que l'on retrouve sur les microphones des prothèses auditives et qui correspond donc bien à notre étude.

Parole dans le bruit

Dans une situation réelle la parole et le bruit ne parviennent jamais séparément jusqu'au réducteur de bruit. Si c'était le cas cet outil serait inutile puisque c'est justement l'une des opérations qu'il doit effectuer.

C'est pour cette raison évidente qu'avec l'IDS *Speech Enhancer*, les deux *stimuli* (ambiance et voix) sont rassemblés sur un même fichier. La balance spectrale effectuée par le sujet doit se faire sur l'ensemble « parole + bruit ». Cela n'aurait pas été intéressant de faire des réglages uniquement sur le signal vocal.

L'ambiance sonore est stéréophonique et la voix monophonique (diffusée au centre de l'image stéréo)[3]. Nous sommes partis du principe que le locuteur virtuel devait être placé face au sujet durant les tests afin, encore une fois, de ne pas compliquer le protocole dans cette première approche.

Les deux signaux sont normalisés à -20 dB FS (*Full Scale*) RMS (*Root Mean Square*)[4]. En plaçant la voix et l'ambiance au même niveau sonore, on obtient une sensation de gêne qui commence à être importante et donc à poser problème pour la compréhension de ce qui est dit (même pour un sujet normo-entendant).

3. La stéréophonie est une méthode de reproduction sonore utilisant deux canaux (gauche et droit) qui permet de reconstituer la répartition spatiale des sources dans l'espace et d'obtenir une impression de relief acoustique. Par exemple avec un casque audio, il est possible de donner l'impression qu'un son vient du centre alors que les deux hauts-parleurs sont placés sur nos oreilles gauche et droite. C'est ce que l'on appelle une image (ou source) fantôme.

4. Le dB FS est utilisé dans le domaine numérique. Il exprime le niveau d'un signal numérique par rapport au niveau de saturation. 0 dB FS correspond donc au niveau maximal du signal avant saturation numérique. Le dB FS peut se référer à deux grandeurs : le niveau crête et le niveau RMS. Le niveau crête est la grandeur maximale du signal détectée sur un laps de temps donné. Le niveau RMS est un niveau moyen, c'est lui qui donne la sensation de sonie.

Cet équilibre est un point de départ qui nous semble logique par rapport à l'exercice demandé aux sujets.

En audiométrie vocale lorsque l'audioprothésiste souhaite tester l'audition d'un patient dans le bruit il lui est conseillé d'utiliser des bruits « cliniques ». Ces bruits sont obtenus en filtrant des ambiances sonores de type *cocktail-party* (composées de plusieurs voix incompréhensibles) par rapport au signal vocal étudié. Il s'agit de faire en sorte que le rapport signal sur bruit (RSB) soit le même à toute les fréquences. Cette normalisation permet d'harmoniser les mesures entre audioprothésistes et de pouvoir ensuite comparer les résultats de façon correcte. Néanmoins il en résulte une perte de réalisme dans le processus liée au filtrage [60].

Avec l'IDSSE nous ne souhaitions pas obtenir des résultats d'audiométrie vocale normalisés mais des indications permettant de réaliser des filtres pour améliorer la compréhension de la parole dans le bruit. Il n'a donc pas été question de filtrer les ambiances sonores utilisées en fonction du signal vocal. Nous cherchions en effet à nous rapprocher le plus possible d'une expérience réaliste.

Pour nommer les fichiers des différentes « scènes » constituées nous avons adopté la convention suivante :

- en premier figure le **nom de l'ambiance** : « Cafe », « Gare », « Marche », « Metro », ou « Ville » ;
- on indique ensuite le **niveau sonore auquel l'ambiance a été normalisée** : « 20 » (pour -20 dB FS). Ici toutes les ambiances ont été normalisées à -20 dB FS mais cette convention pourrait servir si l'on décide de faire des tests à différents niveaux ;
- vient après le **nom du fichier** utilisé pour le signal vocal :
  - **TLM103_Texte_F21fra_03** pour la voix féminine ;
  - **TLM103_Texte_H24fra_02** pour la voix masculine ;

- pour finir on note le **niveau de normalisation du signal vocal** (même raison que pour l'ambiance) : « 20 » (pour - 20 dB FS) ;

- tous ces éléments sont séparés par un **underscore** : « _ ».

Exemple : si le fichier correspond à la voix féminine intégrée dans l'ambiance de gare et que ces deux signaux ont été normalisés à - 20 dB FS, alors on lui attribuera le nom : « Gare_20_TLM103_Texte_F21fra_03_20 ».

En tout nous avons constitué 10 scènes à partir des 5 ambiances et des 2 voix. Les fichiers composés des voix féminines durent 3 minutes et 12 secondes, ceux composés des voix masculines : 2 minutes et 43 secondes. Cela fait donc une durée totale de 29 minutes et 35 secondes.

Sachant que la durée de l'examen audiométrique est de 8 minutes minimum et qu'il faut ajouter à cela un temps d'explication et une première prise en main de l'outil avant de commencer les réglages, la durée totale du test serait trop longue avec les 10 scènes. Cela pourrait induire des résultats faussés liés à une baisse de la concentration du sujet au fur et à mesure de l'expérience.

Nous avons donc décidé de réduire le nombre de scènes testées à 5 par sujet.

4.2.3 Découpages fréquentiels utilisés

L'IDSSE permet au sujet de régler la balance spectrale des scènes que nous avons évoquées ci-dessus, à partir de leur re-synthèse IDS. Or nous savons que pour réaliser la re-synthèse d'un fichier il faut auparavant en avoir fait l'analyse IDS selon un découpage fréquentiel.

Dans le cadre de cette étude nous avons finalement décidé d'utiliser trois découpages fréquentiels différents pour l'analyse des scènes testées :

- un découpage « général » correspondant au découpage retenu pour le sous-corpus « Texte/TLM103 » (cf. figure 3.7) ;
- deux découpages « spécialisés » correspondant aux découpages retenus après l'analyse des fichiers **TLM103_Texte_F21fra_03** pour la voix féminine et **TLM103_Texte_H24fra_02** pour la voix masculine.

L'idée est de comparer les balances spectrales obtenues selon les découpages, le « général » et chaque « spécialisé », afin de déterminer si l'utilisation d'un découpage adapté à la voix du locuteur possède une influence dans les résultats.

Pour le découpage « général » nous avons donc conservé l'un des découpages retenus à la suite des analyses de corpus (Texte/TLM103). Pour les découpages « spécialisés » nous avons effectué une nouvelle recherche non plus pour un corpus mais pour deux fichiers distincts.

Voici les trois découpages fréquentiels retenus pour l'analyse/re-synthèse IDS des scènes testées[5] :

Découpage G :	Découpage F :	Découpage H :
· **0 - 47 Hz** ;	· **0 - 19 Hz** ;	· **0 - 48 Hz** ;
· **47 - 108 Hz** ;	· **19 - 52 Hz** ;	· **48 - 80 Hz** ;
· **108 - 163 Hz** ;	· **52 - 123 Hz** ;	· **80 - 204 Hz** ;
· **163 - 331 Hz** ;	· **123 - 336 Hz** ;	· **204 - 537 Hz** ;
· **331 - 798 Hz** ;	· **336 - 692 Hz** ;	· **537 - 1020 Hz** ;
· **798 - 1048 Hz** ;	· **692 - 2443 Hz** ;	· **1020 - 1795 Hz** ;
· **1048 - 3388 Hz** ;	· **2443 - 4323 Hz** ;	· **1795 - 2723 Hz** ;
· **3388 - 5205 Hz** ;	· **4323 - 10872 Hz** ;	· **2723 - 6041 Hz** ;
· **5205 - 14307 Hz** ;	· **10872 - 13941 Hz** ;	· **6041 - 15331 Hz** ;
· **14307 - 22050 Hz**.	· **13941 - 22050 Hz**.	· **15331 - 22050 Hz**.

5. « Découpage G » correspond au découpage général, « Découpage H » au découpage spécialisé pour la voix masculine et « Découpage F » au découpage spécialisé pour la voix féminine.

Le nombre de scènes testées double car chaque fichier est analysé par le découpage général et le découpage adapté au *stimulus* de parole utilisé : découpage F pour les scènes avec une voix féminine et découpage H pour celles avec une voix masculine[6].

Nous obtenons donc 10 scènes avec un découpage général et 10 scènes avec des découpages spécialisés.

L'idée a été de faire deux groupes de sujets : l'un testant le découpage G et l'autre les découpages S (spécialisés)[7].

Comme le nombre de scènes testées est limité à 5 par sujet voici les règles que nous avons fixées :

- chaque sujet doit tester **toutes les ambiances** ;
- chaque sujet doit tester **au moins 2 scènes avec une voix féminine et 2 scènes avec une voix masculine**.

Nous aurions pu faire tester aux sujets toutes les ambiances avec la même voix mais cela n'aurait pas, *a priori*, apporté de résultats supplémentaires. Par ailleurs alterner voix féminine et voix masculine permet au sujet de ne pas s'enfermer dans une forme de monotonie.

6. Les scènes présentant une voix féminine ne sont pas analysées avec le découpage H et inversement.
7. Les découpages S correspondent à l'ensemble « découpage H + découpage F ».

4.2.4 Outil logiciel

Toute la partie logicielle de l'IDS *Speech Enhancer* est développée sur *Pure Data*[8] :

> « *Pure Data (Pd) is an open source visual programming language that can run on anything from personal computers and Raspberry Pis to smartphones* [...] *It is a major branch of the family of patcher programming languages known as Max (Max/FTS, ISPW Max, Max/MSP, jMax), originally developed by Miller Puckette at IRCAM.*
>
> *Pd enables musicians, visual artists, performers, researchers, and developers to create software graphically without writing lines of code.* »[9] [61]

Le patch *Pure Data* utilisé dans l'IDSSE est en fait issu du patch de re-synthèse IDS développé par Laurent Millot et Romain Vuillet qui est présenté dans le livre *Traitement du signal audiovisuel : Applications avec Pure Data* [33].

Expliquons brièvement son fonctionnement.

Comme nous l'avons déjà évoqué il faut effectuer une analyse IDS du fichier que l'on souhaite re-synthétiser. Si le découpage utilisé lors de cette analyse est composé de 10 sous-bandes nous devons alors obtenir 10 fichiers audio qui correspondent chacun à l'une des sous-bandes du découpage.

Dans *Pure Data* on indique ensuite où se trouvent ces fichiers et il suffit de lancer la lecture des 10 fichiers en simultané pour obtenir le signal re-synthétisé. Nous avons donc 10 lecteurs synchronisés ensemble correspondant à chaque sous-bande.

8. Le patch *Pure Data* général de l'IDSSE est disponible en annexe.

9. [Traduction] Pure Data (Pd) est un langage de programmation visuel open-source qui peut fonctionner sur n'importe quel appareil des PC et *Raspberry Pis* aux *smartphones* [...] Il est une figure emblématique de la famille des langages de programmation par *patchs* tels que Max (Max/FTS, ISPW Max, Max/MSP, jMax), développés à l'origine par Miller Puckette à l'IRCAM. Pd permet aux musiciens, artistes, performeurs, chercheurs, et développeurs de créer graphiquement des programmes sans écrire de lignes de code.

Pour modifier le poids-relatif de l'une de ces sous-bandes il faut faire varier le volume de son lecteur associé. Par une opération de conversion entre une échelle linéaire et une échelle logarithmique, on obtient le niveau de ce poids en dB FS.

Les sujets peuvent régler le poids des sous-bandes de -30 à +20 dB FS. Nous avons volontairement laissé une grande plage de dynamique pour qu'ils puissent effectuer des réglages « généreux ». Nous souhaitons dans un premier temps obtenir des tendances **globales** pour les balances spectrales et non des réglages précis au décibel près.

L'initialisation du niveau des sous-bandes pour chaque scène est faite à -30 dB FS pour pousser le sujet à agir dès le début du test. Bien qu'il subsiste un faible signal sonore à -30 dB FS, le sujet est obligé d'élever les *faders* pour commencer à percevoir le signal vocal et de ce fait entamer les réglages.

Même si les découpages fréquentiels que nous souhaitons utiliser pour les tests ne correspondent pas à celui employé par Laurent Millot (découpage de Leipp étendu) dans le patch de re-synthèse IDS initial, nous avons pu garder la partie traitement du signal intacte. Étant donné que le choix du découpage fréquentiel n'intervient qu'en amont, lors de l'analyse, on peut en fait utiliser **n'importe quel découpage avec le même programme**. Il faut juste préciser à *Pure Data* le dossier dans lequel aller chercher les fichiers de sous-bandes.

Les modifications introduites viennent du besoin d'enregistrer les réglages de balances spectrales ainsi qu'un certain nombre d'informations personnelles sur le sujet. C'est donc l'interface utilisateur qui a été remaniée.

Tout d'abord pour plus de confort et d'efficacité, le réglage du poids des sous-bandes se fait via un contrôleur MIDI (*Musical Instrument Digital Interface*)[10] muni

10. Le MIDI est un protocole de communication qui était à l'origine dédié à la musique. Il permet d'envoyer des messages depuis un instrument ou contrôleur MIDI vers des logiciels capables de reconnaitre ce protocole. Aujourd'hui le MIDI est toujours utilisé en musique mais aussi dans d'autres domaines (scénographie, danse, traitement vidéo en temps réel, etc.)

de 8 *faders* : la *Behringer BCF2000*. Il est toujours possible de modifier les niveaux à la souris directement sur le patch mais l'utilisation d'une surface de contrôle facilite l'élaboration d'une balance spectrale.

Comme ce contrôleur ne dispose que de 8 *faders* nous avons supprimé la première et la dernière sous-bande des réglages. Ce choix a été motivé par des tests préliminaires qui ont montré l'absence d'influence de ces 2 sous-bandes dans la compréhension de la voix. Les caractéristiques techniques de la *Behringer BCF2000* sont disponibles en annexe.

FIGURE 4.1 – Sujet effectuant un test perceptif avec l'IDS *Speech Enhancer*

Comme nous l'avons évoqué une section « informations » a été rajoutée au patch.

Tout d'abord un numéro de sujet est attribué à chaque nouvel utilisateur. Ce dernier peut ensuite indiquer son nom, son prénom, son sexe, son âge, le département dans lequel il a grandi, le département dans lequel il vit actuellement et la nature de son rapport au son. Pour ce dernier critère nous lui avons proposé deux paramètres :

- **sujet « candide »** : pour les sujets n'ayant jamais (ou très peu) utilisé d'outils de traitement du son ;
- **sujet « expérimenté »** : pour les sujets étant habitués à ces pratiques.

Pour indiquer la note (de 0 à 5) qu'un sujet attribue à sa compréhension de la voix après les réglages, 6 boutons horizontaux ont été ajoutés. Nous avons aussi mis en place un bouton « Enregistrer » qui permet de sauvegarder sur un fichier texte toutes les informations que le sujet a indiquées, le poids des différentes sous-bandes de la balance obtenue ainsi que la date à laquelle le test a été réalisé.

Pour nommer ces fichiers textes nous avons adopté la convention suivante :

- en premier figure le **nom de la scène** pour laquelle le sujet a effectué le réglage : « Gare_20_TLM103_Texte_F21fra_03_20 » par exemple, comme nous l'avons cité plus haut ;
- on indique ensuite le **numéro du sujet**[11] : « num0 » pour le premier sujet, « num1 » pour le deuxième, etc. ;
- pour finir on note le **sexe** et l'**âge** du sujet : « F21 » pour une femme de 21 ans, « H32 » pour un homme de 32 ans ;
- tous ces éléments sont séparés par un **underscore** : « _ ».

Exemple : si le 15ème sujet à réaliser les tests est un homme de 22 ans et qu'il souhaite enregistrer sa balance spectrale obtenue pour la scène :

- « Gare_20_TLM103_Texte_F21fra_03_20 » ;

alors le nom de ce fichier sera :

- « Gare_20_TLM103_Texte_F21fra_03_20_num14_H22 ».

11. Ce numéro est obtenu grâce à un compteur qui incrémente sa valeur de 1 pour chaque nouveau sujet.

4.3 Résultats des premiers tests

Nous avons réalisé une première session de tests perceptifs tels qu'ils ont été décrits ci-dessus. 19 sujets ont pris part à l'expérience et chaque scène a été testée 5 fois au total [12].

Voici les conclusions que nous pouvons d'ors et déjà proposer suite à l'analyse des résultats obtenus. Des graphiques illustrants ces résultats sont disponibles en annexe.

- D'une manière générale on observe que la sous-bande 1 [13] est très atténuée par les sujets. Pour l'ambiance « Gare » par exemple, 15 des 20 sujets ont placé cette sous-bande à -30 dB FS (le niveau minimal). Cette tendance se retrouve pour toutes les ambiances avec les deux voix et les deux découpages. Il semblerait donc que cette sous-bande ne soit pas nécessaire à la compréhension de la voix.

- On remarque aussi pour la majorité des tests que les sujets préfèrent retirer du niveau plutôt que d'en ajouter. Cela pourrait être dû au fait qu'élever le niveau d'une sous-bande pour renforcer la voix implique l'augmentation du niveau de bruit présent dans cette sous-bande. Or les trois découpages utilisés pour les tests possèdent tous des sous-bandes assez larges qui contiennent donc des informations de bruit importantes. Les sujets semblent préférer ne pas ajouter cette quantité de bruit même si cela permet de remonter le niveau de la parole. Ils agissent plutôt par soustraction.

- D'après les sujets les ambiances posant le plus de problèmes sont celles présentant des flux de paroles : café et marché. Cela corrobore ce que l'on peut trouver dans la littérature. Pour ces deux ambiances les résultats sont assez chaotiques et

12. Le premier sujet a testé les 10 scènes du découpage G. C'est après ses retours sur le déroulement du test que nous avons décidé de réduire le nombre de scènes à 5 par sujet.

13. Pour rappel, nous avons supprimé la première et la dernière sous-bande des découpages initiaux pour les tests. Ici la sous-bande 1 correspond donc à la sous-bande 2 des découpages complets

aucune tendance ne semble se dessiner (si ce n'est celle de l'atténuation de la sous-bande 1). Avec une ambiance trop complexe à gérer les sujets se mettent à agir de manière « aléatoire » et il n'y a donc pas de cohérence entre les résultats.

- Pour les autres ambiances des tendances dans les balances spectrales obtenues se dessinent plus ou moins. En observant les graphiques disponibles en annexe on remarque que pour certaines scènes les tracés représentant les balances spectrales réalisées par les sujets peuvent parfois être cohérents pour quelques sous-bandes. Nous considérons qu'il y a « cohérence » dans le réglage du poids d'une sous-bande si l'écart maximal entre les niveaux proposés par les différents sujets pour cette sous-bande tient dans 10 dB de dynamique. En appliquant cette définition aux graphiques disponibles en annexe, voici les sous-bandes « cohérentes » obtenues :

  - la 3ème pour l'ambiance de ville avec le découpage G et la voix masculine ;
  - la 5ème et la 6ème pour l'ambiance de ville avec le découpage G et la voix féminine ;
  - la 2ème pour l'ambiance de métro avec le découpage H et donc la voix masculine ;
  - la 5ème pour l'ambiance de métro avec le découpage G et la voix masculine ;
  - la 4ème pour l'ambiance de gare avec le découpage F et donc la voix féminine ;
  - la 6ème pour l'ambiance de gare avec le découpage G et la voix féminine.

  Tout d'abord on peut noter que ces résultats sont trop lacunaires pour pouvoir en tirer des informations permettant de réaliser les filtres que nous recherchons. Ensuite les sujets ne semblent pas mieux s'accorder selon la voix ou le type de découpage utilisé. Cela recoupe les données récoltées à partir des notes qui étaient à peu près égales pour les scènes avec la voix féminine et les scènes avec la voix masculine ainsi qu'entre le découpage G et les découpages S [14].

14. Nous apportons un intérêt modéré aux notes obtenues car elles dépendent de nombreuses variables que nous ne pouvons pas maîtriser. Toutefois elles peuvent devenir intéressantes lorsque des tendances sont observées dans les extrêmes. On peut par exemple relever un phénomène de notes très basses pour l'ambiance de marché qui correspond aux difficultés de réglages évoquées par les sujets pour cette scène.

- Il ne semble pas y avoir de disparités entre les sujets « candides » et les sujets « expérimentés » dans les résultats obtenus. La seule différence se trouve dans le temps d'adaptation à l'outil qui est moins important chez les sujets expérimentés. Au cas par cas on peut observer chez certains sujets candides des balances spectrales « exotiques » avec parfois un gain à + 20 dB FS pour une sous-bande donnée. Chez les sujets expérimentés il y a peut être un phénomène d'autocensure lié à des habitudes de pratiques conventionnelles.

- 5 sujets sur les 19 présentent des troubles auditifs et 2 parmi eux ont des déficiences assez marquées. Pour les 5 sujets ces troubles sont des « accidents » dans les aigus (4000 Hz, 6000 Hz et 8000 Hz) plus ou moins importants.

  Au vu des résultats obtenus il ne semble pas y avoir de corrélation entre ces problèmes auditifs et les réglages effectués. Cela semble assez logique si l'on reprend l'argument cité plus haut. Comme les sous-bandes de fréquences utilisées sont larges, en augmentant la voix dans l'une de ces sous-bandes on fait aussi remonter beaucoup de bruit. Un sujet malentendant n'a donc pas plus d'intérêt qu'un sujet sain à relever le niveau de la sous-bande pour laquelle il présente une déficience car il ajouterait du bruit dans une zone qu'il a déjà du mal à cerner. On peut aussi rappeler la remarque de Roland Carrat évoquée en introduction qui correspond bien à cette situation : « L'amplification des fréquences déficitaires ou perdues ne provoquent pas pour autant une régénération des capteurs neurosensoriels manquants [...] »[34, p.211]

  En annexe se trouve les audiogrammes des deux sujets présentant la plus forte déficience auditive ainsi que les balances spectrales qu'ils ont effectuées.

Bien sûr il faut prendre du recul avec ces résultats qui n'ont été obtenus qu'à partir d'un faible échantillon de sujets. C'est la raison pour laquelle d'autres tests perceptifs vont être effectués avant la soutenance de ce mémoire afin de tester la pertinence des tendances exposées ci-dessus.

Avec les résultats que nous avons actuellement nous ne pouvons pas encore réaliser de filtres (même génériques) permettant d'améliorer la compréhension de la parole dans les ambiances étudiées. Les futurs tests doivent donc nous permettre d'obtenir les informations manquantes pour pouvoir lancer une première élaboration de ces filtres.

4.4 Amendements et perspectives

Suite à l'observation des résultats de la première session de tests et à des réflexions menées de manière parallèle, nous allons évoquer un certain nombres de perspectives et d'amendements potentiels pour l'IDS *Speech Enhancer* qui permettraient l'otpimisation des futurs tests perceptifs. Comme il ne sera pas possible de tout réaliser dans le cadre de ce mémoire, nous résumerons à la fin de cette section les amendements que nous allons conserver pour les tests à mener avant la soutenance.

4.4.1 Diffusion

Pour se rapprocher au maximum d'une situation réelle il faudrait diffuser les *stimuli* de paroles et d'ambiances à travers des enceintes tout en faisant écouter au sujet ces mêmes signaux dans un casque audio, le réglage de la balance spectrale n'étant effectué que pour le son diffusé au casque. Le sujet serait ainsi dans des conditions plus proches de la réalité car il pourrait percevoir :

- le son transmis par les prothèses (représentées par le casque audio) ;
- le résidu du son « réel » (correspondant au son provenant des enceintes) qui n'a pas été atténué par les prothèses (casque).

Il est possible que la perception de ce signal résiduel influence le réglage du poids des sous-bandes.

4.4.2 *Stimuli*

Bruit

Les ambiances utilisées pour les premiers tests sont issues de banques de sons déjà existantes. L'idéal serait de réaliser nous-même les prises de son avec un protocole précis pour que l'enregistrement soit similaire à chaque fois et n'influe pas sur la balance spectrale.

Dans le cadre d'une application pour des patients « réels » on pourrait même penser à enregistrer ces ambiances dans les lieux que l'utilisateur à l'habitude de fréquenter afin de créer des filtres « personnalisés » qui correspondent vraiment à ses attentes.

Parole

Concernant le signal vocal nous pourrions envisager d'utiliser un texte que les sujets connaissent bien, soit parce qu'ils l'auraient lu en amont, soit parce qu'il s'agirait d'un texte connu (hymne national, fable ou poème de notre culture commune, etc.). Nous pourrions même leur faire passer les tests avec des voix qu'ils ont l'habitude d'entendre, comme la voix d'un proche. Nous observerions alors si la connaissance préalable du contenu de la parole améliore la vitesse et/ou la précision de réglage de la balance spectrale.

Prenons l'exemple des domaines du mixage et de la sonorisation. Les ingénieurs du son utilisent tout le temps les mêmes musiques pour tester un système de diffusion. En comparant la manière dont ces musiques « sonnent » sur le système testé à celle qu'ils ont l'habitude d'entendre, ils savent très rapidement repérer les qualités et les défauts de ce système.

Par analogie nous pourrions faire l'hypothèse qu'en connaissant le contenu du *stimulus*, le sujet affinerait en fait son jugement et donc la précision de ses réglages.

L'idée serait aussi d'utiliser un discours plus naturel que la lecture d'un texte. Cependant comme nous l'avons vu le flux de parole des *interviews* est trop saccadé. Un compromis serait de demander au locuteur de raconter une histoire sans support de texte (une aventure qui lui est arrivée par exemple). La voix serait plus naturelle que lors d'une lecture et beaucoup moins irrégulière qu'avec les *interviews*.

En partant du principe que les personnes avec qui nous parlons le plus sont généralement les mêmes au quotidien, nous pourrions proposer au patient la possibilité d'effectuer une « empreinte » de la voix de ses proches. On déterminerait les découpages fréquentiels associés à ces voix et les tests avec l'IDSSE se feraient à partir de ces données.

Nous serions ensuite en mesure de lui offrir une « banque de caractéristiques vocales » dans laquelle il irait piocher le découpage fréquentiel qu'il souhaite utiliser selon la situation. Supposons qu'il soit par exemple en train de discuter avec sa fille, il pourrait choisir le réglage « fille », de même pour les autres membres de sa famille ou ses collègues qu'il voit tous les jours, etc.

Par ailleurs les découpages obtenus à partir du corpus de voix parlées seraient toujours utiles dans le cas où la conversation n'aurait pas lieu avec un proche.

Il faudrait donc nourrir ce corpus avec de nouveaux enregistrements afin d'augmenter son aspect générique. On pourrait ensuite le diviser en plusieurs sous-corpus :

- selon le sexe : homme, femme ;
- et/ou selon l'âge : jeunes enfants, adolescents, jeunes adultes, adultes d'une quarantaine/cinquantaine d'années, personnes âgées.

Il pourrait aussi être intéressant de constituer des corpus pour les accents les plus représentés en France : haut-nordien, sudiste, grand-estien, breton, « bordelais », corse, accents étrangers, etc.

Parole dans le bruit

Il pourrait être intéressant d'observer l'influence du niveau sonore de l'ambiance sur la balance spectrale obtenue. Pour l'instant nous avons fait les tests avec des niveaux d'ambiance et de voix identiques : -20 dB FS.

Cette valeur était peut être trop élevée pour les ambiances, notamment celles du café et du marché qui ont destabilisé les sujets. Pour les prochains tests nous diminuerons le niveau des ambiances à -25 ou -30 dB FS.

4.4.3 Découpages fréquentiels

Lors des premiers tests nous avons supprimé la 1ère sous-bande des réglages car elle ne comporte qu'un son continu très grave, fatigant pour l'écoute et ne contenant pas d'informations fréquentielles associées au signal vocal. Nous avons aussi supprimé la 10ème sous-bande car les composantes de la voix n'atteignent que très peu cette zone fréquentielle, qui est en plus potentiellement inaudible pour les malentendants.

On pourrait ajouter la 2ème sous-bande à cette élimination car les résultats des premiers tests ont montré qu'elle n'était pas utile pour notre étude.

Par ailleurs il faudrait utiliser un découpage fréquentiel plus segmenté dans le médium et le « bas-aigu » car c'est dans ces zones que se trouvent les principales informations d'intelligibilité de la parole. Cela permettrait d'être plus précis dans le traitement du signal vocal et de ne pas ajouter un grande quantité de bruit lorsque l'on cherche à remonter le niveau d'une composante de la voix qui nous semble précieuse.

Pour définir ce découpage fréquentiel on repart du découpage initial à 10 sous-bandes. On supprime la 1ère, 2ème et 10ème sous-bande comme nous l'avons évoqué mais aussi la 3ème et la 9ème. Bien qu'il existe quelques informations fréquentielles associées à la voix dans ces deux sous-bandes, elles sont très peu nombreuses et ne représentent pas des

enjeux important pour la compréhension de la parole. On les élimine donc du réglage pour pouvoir ajouter 3 nouvelles sous-bandes dans la zone de 1000 à 4000 Hz où se trouve les fréquences d'intelligibilité de la voix.

En suivant cette méthode voici ce que l'on obtient par exemple pour le découpage adapté à la voix masculine :

Découpage H intermédiaire :

- **204 - 537 Hz** ;
- **537 - 1020 Hz** ;
- **1020 Hz** - F_4 ;
- F_4 - **1795 Hz** ;
- **1795 Hz** - F_6 ;
- F_6 - **2723 Hz** ;
- **2723 Hz** - F_8 ;
- F_8 - **6041 Hz**.

On détermine ensuite F_4, F_6 et F_8 à partir du portrait IDS en 40 sous-bandes que l'on avait réalisé après la première recherche des minima locaux. Finalement, voici les 3 nouveaux découpages retenus pour les prochains tests :

Découpage G2 :

- **163 - 331 Hz** ;
- **331 - 511 Hz** ;
- **511 - 798 Hz** ;
- **798 - 1048 Hz** ;
- **1048 - 1558 Hz** ;
- **1558 - 2208 Hz** ;
- **2208 - 3388 Hz** ;
- **3388 - 5205 Hz**.

Découpage F2 :

- **336 - 692 Hz** ;
- **692 - 851 Hz** ;
- **851 - 1159 Hz** ;
- **1159 - 2443 Hz** ;
- **2443 - 2599 Hz** ;
- **2599 - 4323 Hz** ;
- **4323 - 5096 Hz** ;
- **5096 - 10872 Hz**.

Découpage H2 :

- **204 - 537 Hz** ;
- **537 - 1020 Hz** ;
- **1020 - 1319 Hz** ;
- **1319 - 1795 Hz** ;
- **1795 - 2101 Hz** ;
- **2101 - 2723 Hz** ;
- **2723 - 4584 Hz** ;
- **4584 - 6041 Hz**.

4.4.4 Outil logiciel

Durant la première session de tests nous avons pu observer que plusieurs sujets avaient du mal à gérer de manière simultanée les 8 sous-bandes proposées. Ils plaçaient certaines sous-bandes à un niveau « aléatoire » au début de la scène et n'y revenaient plus jusqu'à la fin du réglage.

Par conséquent il faudrait guider le sujet sur la conduite à tenir pour optimiser le mixage de la balance spectrale.

Lors des prochains tests nous expliquerons aux sujets qu'il faut commencer par régler les 4 sous-bandes centrales de façon à ce que la voix soit bien compréhensible puis ajouter progressivement les autres sous-bandes afin d'optimiser l'intelligibilité et le confort d'écoute.

Il serait intéressant d'ajouter une fonction « solo » pour donner la possibilité au sujet de n'écouter qu'une seule sous-bande lorsqu'il souhaite la régler précisément. De la même manière on pourrait mettre en place une fonction « mute » pour enlever une ou plusieurs sous-bandes de la balance spectrale durant l'écoute et ainsi déterminer leur influence dans le mixage général.

4.4.5 Amendements retenus pour les prochains tests

Voici la liste des amendements conservés pour les tests à réaliser avant la soutenance.

- Diminuer le niveau des ambiances à **-25 ou -30 dB FS** ;
- Utiliser les **nouveaux découpages fréquentiels** qui présentent un nombre plus important de sous-bandes dans les zones de fréquences associées à l'intelligibilité de la voix ;

- **Guider les sujets** en leur indiquant la conduite à tenir pour optimiser les réglages de la balance spectrale ;

- Ajouter les fonctions « **solo** » et « **mute** » à l'IDSSE.

4.5 Conclusion

Dans ce chapitre nous avons présenté le protocole des tests perceptifs effectués avec l'IDS *Speech Enhancer*. Nous avons ensuite décrit point par point cet outil : en justifiant l'utilisation d'un casque audio comme moyen de diffusion, en évoquant ensuite la façon dont les *stimuli* présentés aux sujets ont été constitués, en mentionnant les découpages fréquentiels utilisés pour l'analyse/re-synthèse IDS de ces *stimuli* et en étudiant l'interface logicielle de l'IDSSE développée sur *Pure Data*.

Par la suite nous avons analysé les résultats obtenus à partir de la première session de tests menés. Quelques tendances ont pu être observées :

- la **2$^{\text{ème}}$ sous-bande** des découpages fréquentiels utilisés ne semble **pas avoir d'influence** sur la compréhension de la voix tout comme la 1$^{\text{ère}}$ et la 10$^{\text{ème}}$;
- pour améliorer la compréhension de la parole les sujets **agissent plutôt par soustraction** : ils préfèrent enlever du bruit plutôt que d'augmenter les composantes du signal vocal ;
- les ambiances posant le plus de **problèmes** sont celles qui présentent des **flux de parole** (ambiances de type *cocktail-party*) ;
- on peut observer des **cohérences** dans les réglages des sujets pour **quelques sous-bandes isolées** ;
- il ne semble **pas y avoir de discordances** entre les résultats obtenus pour les sujets « **candides** » et les sujets « **expérimentés** » ;
- la déficience auditive d'un sujet **n'influence pas**, *a priori*, les réglages qu'il effectue pour faire émerger la voix de l'ambiance bruitée.

Cependant le nombre de sujets interrogés n'a pas été assez important pour valider ces conclusions.

De plus le protocole de test présente des faiblesses et nous avons donc proposé un certain nombre d'amendements et de perspectives potentiels permettant d'améliorer l'IDSSE.

Comme toutes ces modifications sont impossibles à mettre en place dans le cadre du mémoire, nous avons sélectionné les plus pertinentes pour notre étude. Ces amendements doivent permettre d'optimiser les tests réalisés avant la soutenance afin de confirmer ou d'infirmer les premières tendances et d'obtenir des informations pertinentes sur les filtres à réaliser pour améliorer la compréhension de la parole dans les ambiances étudiées.

5 Utilisation des traitements en temps réel

Même si nous n'avons pas encore obtenu d'informations suffisantes pour élaborer les filtres évoqués dans le chapitre précédent, nous pouvons d'ors et déjà préparer l'étude des outils nécessaires pour optimiser leur fonctionnement et leur utilisation.

En effet une fois que nous aurons collecté les données permettant la réalisation de ces filtres, il faudra ensuite les intégrer dans la chaîne de traitement d'un réducteur de bruit.

Notre projet final étant de proposer une meilleure compréhension de la parole dans le bruit au quotidien, cette intégration est soumise à un certain nombre de contraintes qui sont liées à l'utilisation en temps réel de ce réducteur de bruit.

Dans ce chapitre nous allons donc étudier les compromis et les outils nécessaires pour permettre un fonctionnement des traitements obtenus grâce à l'IDS *Speech Enhancer* en temps réel.

Nous expliquerons tout d'abord les caractéristiques que devrait posséder le convolueur idéal dans un tel contexte, puis nous mentionnerons les compromis à effectuer sur ces paramètres en fonction des outils pratiques mis à notre disposition. Nous indiquerons ensuite les ressources nécessaires à l'utilisation de l'analyse IDS en temps réel et nous évoquerons les perspectives que pourrait apporter cette technologie à notre projet.

5.1 Convolution en temps réel

5.1.1 Convolueur temps réel idéal

Un convolueur est un algorithme qui s'insère dans une chaîne de traitement du signal. Dans le cas de notre étude le convolueur utilisé doit « traiter » le signal sonore capté en lui appliquant les filtres issus des balances spectrales obtenues avec l'IDSSE lorsque les résultats accumulés seront suffisamment pertinents.

Actuellement on arrive très bien à exécuter cette opération lorsque le signal à convoluer est enregistré dans un fichier audio. Il suffit d'entrer dans le convolueur les valeurs du filtre sélectionné et le fichier son, de lancer la convolution, et on récupère ensuite le fichier traité.

Cependant nous avons besoin d'aller plus loin pour l'utilisation que nous voulons faire de ces filtres obtenus à la suite des tests perceptifs.

En effet nous souhaitons proposer des traitements pour améliorer la compréhension de la parole dans des situations de vie quotidienne. Or durant une conversation nous n'avons pas le temps d'enregistrer le discours de notre locuteur, de le passer dans un convolueur avec un filtrage adapté à la situation, et d'écouter le résultat avant de lui répondre. Il faut effectuer toutes ces opérations en **temps réel**.

Dans l'idéal voici comment fonctionnerait ce convolueur « temps réel ».

- Le signal capté (par un ou plusieurs microphones) est instantanément convolué, sans aucune latence. Il n'y a donc aucun retard entre le signal d'entrée correspondant au son capté (parole du locuteur + ambiance sonore) et le signal de sortie qui correspond au son traité favorisant la compréhension de la parole.

- De plus cette convolution se fait sans distorsion spectrale et permet d'obtenir un signal traité d'une qualité sonore identique au signal original.

- Pour finir les filtres utilisés par ce convolueur idéal peuvent être mis à jour de manière immédiate si l'utilisateur souhaite les modifier.

5.1.2 Compromis pratiques

En pratique un tel convolueur n'existe pas encore et il faut donc trouver des compromis réalistes.

Il y a deux tendances vers lesquelles on peut se diriger. Soit on décide de mettre l'accent sur la qualité sonore mais on perd alors en réactivité (augmentation de la latence). Soit à l'inverse on essaye d'obtenir une latence très courte mais on prend le risque d'introduire des distorsions dans le signal traité.

Dans les outils à notre disposition il existe deux grandes familles de convolutions :

- les convolutions basés sur la FFT ;
- les convolutions temporelles.

Les convolutions basées sur la FFT partent du principe qu'une multiplication dans le domaine fréquentiel correspond à une convolution dans le domaine temporel. Le signal d'entrée est transformé dans le domaine fréquentiel avec la FFT, multiplié par la réponse fréquentielle du filtre, puis re-transformé dans le domaine temporel grâce à la FFT inverse.

Les convolutions de ce type ont une limitation liée au nombre de fréquences à calculer. Pour obtenir un signal d'une bonne qualité il faut effectuer le calcul sur un grand nombre de fréquences. Le problème c'est qu'en utilisant un grand nombre de fréquences on ralentit la vitesse de calcul et on augmente donc la latence.

Pour les convolutions temporelles nous allons nous intéresser plus spécialement à la convolution par algorithme des taches présentée par Laurent Millot et Gérard Pelé lors de la 124ème conférence de l'AES (*Audio Engineering Society*) [62] :

> « *Taches-algorithm principle : each input value generates a scaled and delayed version of the impulse response and the global output is the result of linear superposition of all the intermediate outputs.* » [1]

L'algorithme des taches permet donc de réduire le temps de latence sans diminuer la qualité du signal car il donne un accès immédiat au nouvel échantillon de sortie à partir des impulsions constituant le signal d'entrée.

Par contre, il faut quand même calculer le produit de la valeur courante de l'entrée ($e[n]$) avec l'ensemble de la réponse impulsionnelle h (à la manière d'une multiplication vectorielle) et additionner $e[n].h$ avec le résidu des convolutions précédentes.

La convolution par algorithme des taches est d'autant plus intéressante pour notre projet qu'elle offre la possibilité de changer la réponse impulsionnelle utilisée en temps réel [Ibid.]. Ce principe a d'ailleurs fait l'objet d'une démonstration lors de la présentation du papier à l'AES.

Si le sujet souhaite modifier le filtre appliqué au convolueur en cours d'utilisation ou le transformer totalement parce qu'il rencontre une nouvelle ambiance sonore (passage du métro à la rue par exemple), il pourra effectuer la modification de manière instantanée avec cet algorithme, dès lors que nous saurons optimiser, pour le réaliser en temps réel, le calcul des réponses impulsionnelles pendant la phase de transition (« cross fade » linéaire entre deux réponses impulsionnelles par exemple) [2].

1. [Traduction] Principe de l'algorithme des taches : chaque valeur d'entrée génère une version échelonnée et retardée de la réponse impulsionnelle et le signal de sortie général est le résultat de la superposition linéaire de toutes les valeurs de sorties intermédiaires.
2. Pour permettre au sujet ce genre de réglage on pourrait réaliser un programme transportable sur des appareils mobiles (smartphone, tablette, montre connectée, etc.) qui communiquerait avec le convolueur (sûrement grâce au standard Bluetooth).

On pourrait aussi décider de basculer brutalement d'un filtre à l'autre, ce qui ne nécessiterait alors qu'un chargement rapide du nouveau filtre associé vraisemblablement à l'oubli des résidus de convolution du filtre précédent (principe à valider en amont).

Finalement c'est donc la convolution par algorithme des taches qui semble la plus adaptée à nos besoins. Toutefois la version temps réel de cet algorithme n'est pas encore disponible et il faudra donc se pencher dans un premier temps sur une « convolution FFT ».

Dans la version numérique de l'analyse IDS les filtres ont une atténuation rapide qui permet de réduire la présence du contenu des sous-bandes adjacentes de manière conséquente.

Pour le convolueur utilisé il faut redéfinir la raideur de ces filtres. En effet l'usage de filtres coupants avec une bande de transition très courte (filtres « katana ») peut introduire des temps de calcul assez longs qui ne sont pas compatibles avec l'utilisation en temps réel que nous souhaitons faire de ce convolueur (d'autant plus s'il est embarqué sur un appareil mobile). Cependant utiliser des filtres trop courts présentant donc une atténuation faible peut dénaturer le signal de sortie, notamment puisque les bandes de transition ont alors une largeur importante.

L'idée pour sélectionner la taille minimale et/ou maximale de ces filtres consiste à générer plusieurs jeux de filtres et plusieurs jeux de sous-bandes afin d'effectuer des tests et d'obtenir expérimentalement la raideur minimale à adopter.

Les traitements que nous souhaitons apporter pour améliorer la compréhension de la parole dans le bruit sont uniquement fréquentiels. Nous avons en effet choisi de nous focaliser sur l'aspect spectral des réducteurs de bruit mais cela ne peut pas suffire si l'on souhaite élaborer un outil efficace.

Il faut, comme nous l'avons évoqué dans l'introduction, associer à ces traitements fréquentiels des traitements de discrimination spatiale et des traitements temporels. Nous n'étudierons pas ici ces deux dernières techniques mais nous pouvons réfléchir à la manière d'implémenter notre convolueur fréquentiel au sein de cette chaîne de traitement.

Avec l'IDSSE nous avons observé qu'augmenter le niveau d'une sous-bande pour renforcer une information importante du signal vocal faisait fatalement remonter le niveau du bruit de cette sous-bande. Même si nous souhaitons réduire la largeur des sous-bandes utilisées pour les fréquences associées à la voix afin d'atténuer ce phénomène, renforcer les composantes de la parole amplifiera toujours en conséquence le niveau du bruit.

À partir de ce constat nous pouvons faire deux hypothèses sur la place du convolueur fréquentiel dans la chaîne de traitement du signal d'un réducteur de bruit.

Soit on fonctionne par amplification du signal vocal comme décrit ci-dessus et il faut alors avoir réduit au maximum le bruit avant d'effectuer les traitements associés. Dans ce cas le convolueur fréquentiel doit se trouver à la fin de la chaîne de traitement (*post* traitement temporel et spatial).

Soit on fonctionne par soustraction de bruit comme les sujets avaient tendance à le faire durant les tests et le convolueur pourrait alors venir se placer en tête de chaîne.

L'idée pour optimiser le réducteur de bruit serait d'avoir en fait deux convolueurs fréquentiels en parallèle : un « soustracteur de bruit » en tête de chaîne et un « amplicateur vocal » qui clôturerait tous les traitements.

5.2 Analyse IDS en temps réel

5.2.1 Ressources nécessaires

Pour aller plus loin dans l'optimisation de notre dispositif, que ce soit au niveau du réducteur de bruit utilisé dans la vie quotidienne ou de l'IDS *Speech Enhancer* permettant d'effectuer des tests en conditions « cliniques », il faudrait être en mesure de réaliser des analyses IDS en temps réel.

Le problème c'est que les moyens dont nous devons disposer pour effectuer de tels traitements sont très lourds.

En utilisant une convolution basée sur la FFT, un « cœur »[3] permet de faire la convolution de deux voies en même temps. Avec une convolution temporelle (comme la convolution par algorithme des taches) il faut un cœur par sous-bande et par voie.

Réaliser une analyse IDS stéréophonique à partir d'un découpage fréquentiel à 10 sous-bandes demanderait donc un fonctionnement en parallèle de 10 cœurs pour une convolution FFT et de 20 cœurs pour une convolution temporelle[4].

Il ne s'agit donc pas simplement d'avoir tous ces cœurs à disposition, il faut ensuite réussir à les faire fonctionner de manière fiable et synchronisée.

Au vu des ressources nécessaires et des moyens à mettre en œuvre pour aboutir à ce résultat, nous ne pouvons pas encore effectuer des analyses IDS en temps réel.

3. En informatique un cœur (ou *core* en anglais) est une zone de calcul qui est capable d'exécuter des programmes de façon autonome. Les processeurs actuels sont souvent « multi-cœurs » c'est-à-dire qu'ils possèdent plusieurs cœurs fonctionnant en simultané.

4. Même si la convolution FFT nécessite moins de ressources que la convolution temporelle, elle possède les limitations que nous avons déjà évoquées.

5.2.2 Perspectives

Bien qu'il soit actuellement impossible d'utiliser l'analyse IDS en temps réel, cette technologie pourrait avoir un certain nombre d'applications dans le cadre de notre étude.

Tout d'abord pour les tests perceptifs avec l'IDSSE, les *stimuli* vocaux enregistrés pourraient êtres remplacés par un locuteur réel qui parlerait face au sujet. Ce locuteur pourrait être un proche du sujet et les réglages des niveaux de sous-bandes se feraient en temps réel.

Ainsi pour une application médicale nous n'aurions plus besoin d'enregistrer les voix avant d'effectuer les tests. La totalité du processus se déroulerait lors d'une séance où le patient pourrait discuter avec ses proches et optimiser une balance spectrale correspondant à chacun d'entre eux.

Dans un second temps il s'agirait de transporter ce système en dehors des conditions de « laboratoire » et d'en permettre une application mobile. On offrirait par exemple la possibilité au patient d'effectuer ses réglages de balances spectrales à partir de son téléphone ou d'une tablette. Il pourrait ainsi mettre à jour sa « bibliothèque » de filtres quand il le souhaite sans avoir à aller chez un spécialiste.

Une autre idée serait de faire fonctionner en continu une analyse IDS dans les prothèses auditives du patient. Elle analyserait en temps réel l'évolution de l'environnement sonore et pourrait donc modifier les filtres à appliquer en fonction des situations. L'ajustement des paramètres de filtrage lors d'un changement d'ambiance ne serait plus effectué par le patient à l'aide d'un appareil connecté mais serait automatisé.

Néanmoins dans toutes les perspectives que nous avons évoquées le découpage fréquentiel utilisé pour cette analyse IDS « temps réel » reste à déterminer en amont et est donc fixé à l'avance.

L'idée pour optimiser notre système serait d'effectuer une recherche de découpage fréquentiel automatisée. L'appareil, après un temps d'adaptation le plus court possible, calibrerait lui même ses filtres sur un découpage fréquentiel adapté à la voix du locuteur.

Pour permettre ce type d'automatisation il faudrait sans doute avoir recours à des algorithmes neuronaux avec apprentissage.

5.3 Conclusion

De façon à pouvoir utiliser les filtres obtenus à partir de l'IDSSE en situation « réelle » lorsque les résultats des tests nous aurons fourni les informations nécessaires pour les réaliser, nous avons préparé l'étude des outils essentiels au fonctionnement de ces filtres en temps réel.

Pour ce faire nous avons tout d'abord décrit les caractéristiques que devrait avoir un convolueur temps réel idéal, puis nous avons mentionné les compromis pratiques liés aux outils dont nous disposons actuellement :

- jusqu'à ce qu'une version temps réel de la convolution par algorithme des taches soit mise en place nous utiliserons un convolueur basé sur la FFT pour notre réducteur de bruit ;
- concernant la taille des filtres, il faut mener des tests pour obtenir expérimentalement la raideur à appliquer.

Nous avons ensuite indiqué les ressources qu'il faudrait mettre en place pour permettre une utilisation de l'analyse IDS en temps réel et pour finir, nous avons évoqué les perspectives que pourrait apporter cette technologie à notre projet.

6 Conclusion générale

À travers ce mémoire nous avons cherché à déterminer et à appliquer des traitements fréquentiels permettant une meilleure compréhension de la parole dans le bruit.

Pour ce faire nous avons tout d'abord constitué deux corpus oraux (voix parlées et voix télévisuelles) représentatifs de voix françaises dénuées d'accent sensible. Nous avons examiné les différents choix qui ont guidé leur composition ainsi que l'élaboration du protocole de leur enregistrement. Ces corpus ont été créés pour être analysés par IDS et ainsi nous permettre d'obtenir des découpages fréquentiels adaptés à la voix française.

Dans un deuxième temps nous avons décrit la version analogique de l'IDS présentée par Émile Leipp en 1977, puis nous avons exposé les modifications apportées à cet outil dans sa version numérique développée par Laurent Millot. Nous avons ensuite étudié la méthode permettant d'obtenir un découpage fréquentiel adapté à un objet d'étude et nous avons commenté les découpages retenus pour les corpus évoqués dans le paragraphe précédent.

De ces résultats nous avons déduit que lorsque l'on souhaite obtenir un découpage fréquentiel adapté à un corpus donné, les enregistrement doivent être effectués en appliquant un protocole qui retranscrit au mieux la situation réelle pour laquelle ce découpage est nécessaire.

À partir des découpages fréquentiels retenus nous avons conçu un outil dont le but est de faire émerger un signal de parole lorsqu'il est plongé dans une ambiance bruitée.

Nous avons nommé cet outil l'IDS *Speech Enhancer* (IDSSE). Il nous a permis de faire passer des tests perceptifs à un certain nombre de sujets, afin d'obtenir les filtres à réaliser pour améliorer la compréhension de la parole dans une ambiance sonore donnée.

Nous avons présenté le protocole de ces tests puis nous avons décrit point par point cet outil. Par la suite nous avons analysé les résultats obtenus à partir de la première session de tests menés. Quelques tendances ont pu être observées mais le nombre de sujets interrogés n'a pas été assez important pour valider ces conclusions.

De plus le protocole de test présentait des faiblesses et nous avons donc proposé un certain nombre d'amendements et de perspectives potentiels permettant d'optimiser l'IDSSE. Comme toutes ces modifications étaient impossibles à mettre en place dans le cadre du mémoire, nous avons sélectionné les plus pertinentes pour notre étude.

Ces amendements doivent permettre d'optimiser les tests réalisés avant la soutenance afin de confirmer ou d'infirmer les premières tendances et d'obtenir des informations pertinentes sur les filtres à réaliser pour améliorer la compréhension de la parole dans les ambiances étudiées.

De façon à pouvoir utiliser les filtres obtenus à partir de l'IDSSE en situation « réelle » lorsque les tests nous aurons fourni les informations nécessaires pour les réaliser, nous avons préparé l'étude des outils essentiels au fonctionnement de ces filtres en temps réel.

Nous avons tout d'abord décrit les caractéristiques que devrait avoir un convolueur temps réel idéal, puis nous avons mentionné les compromis pratiques liés aux outils dont nous disposons actuellement.

Nous avons ensuite indiqué les ressources qu'il faudrait mettre en place pour permettre une utilisation de l'analyse IDS en temps réel et pour finir, nous avons évoqué les perspectives que pourrait apporter cette technologie à notre projet.

Bibliographie

[1] BECKETT S., *En attendant Godot*, Paris, Éditions de Minuit, 1952.

[2] OSMAN HILL W., *Primates : Comparative Anatomy and Taxonomy, Vol.1*, Édimbourg, University Press, 1953.

[3] AURIOL B., *La Clef des Sons*, préface de J.-C. Risset, Toulouse, Érès, 1991 (première édition).

[4] LAFON J.-C., « Handicap sensoriel et personnalité », *Psychologie Médicale*, 21, (13), p. 1934-1945, 1989.

[5] SUARÈS A., *Remarques*, Paris, Gallimard, Les Cahiers De La Nrf, 2000.

[6] SCHNUPP J., NELKEN I. et KING A., *Auditory Neuroscience*, Cambridge (Massachusetts), MIT Press, 2012.

[7] Journée Nationale de l'Audition, « Communiqué de presse », mars 2017. `www.journee-audition.org/pdf/cp-resultats-enquete-2017.pdf`.

[8] ROBERT-BOBÉE I., « Projections de population pour la France métropolitaine à l'horizon 2050 », *Insee Première n°1089*, juillet 2006.

[9] La Semaine du Son, page d'accueil du site. `http://www.lasemaineduson.org`.

[10] Code de la Santé Publique. Article L5232-1.

[11] PERRIN G., « Audioprothèses : comment l'Unsaf compte réagir aux récentes conclusions de l'Autorité de la Concurrence », janvier 2017. `http://www.argusdelassurance.com`.

[12] ROUCOUS D., « Le reste à charge d'un appareil auditif visant à compenser un handicap est trop lourd », mars 2017. `http://www.humanite.fr`.

[13] ECALARD J., *Étude de l'efficacité des réducteurs de bruit de la nouvelle génération de prothèse auditive Siemens : Micon*, Mémoire (sous la direction de Mr. Bouchet), Diplôme d'État d'Audioprothésiste, Université de Lorraine, Faculté de Pharmacie de Nancy, 2013.

[14] Best Sound Technology, « Les solutions auditives Siemens, micon ». `https://www.bestsound-technology.fr`.

[15] Que Choisir, « Prothèse auditive, audioprothèse (Infographie). Vous et votre audition », septembre 2015. `https://www.quechoisir.org`.

[16] SERRIÈRE F., « Résultat du baromètre de l'audition Audio 2000 et Senior Strategic », mars 2014. `http://www.acuite.fr`.

[17] Anovum, Phonak et Amplifon, « Étude nationale sur l'audition : Résultats », février 2013. `http://www.phonak.com`.

[18] Ipsos, « Les seniors et l'audition », mars 2013. `http://www.ipsos.fr`.

[19] KOCHKIN S., « MarkeTrak VIII : Consumer satisfaction with hearing aids is slowly increasing », *Hearing Journal*, 63, (1), p. 19-32, janvier 2010.

[20] « Bruit. » Def. 1. Dictionnaire de l'Académie française, 9ème Édition.

[21] RUSSOLO L., *L'Art des bruits*, trad. Nina Sparta, Paris, l'Âge d'homme, 1975.

[22] GELIS C., *Biophysique de l'environnement sonore*, Paris, Ellipses, 2002.

[23] KAMIYA K., MICHEL V. et al., « An unusually powerful mode of low-frequency sound interference due to defective hair bundles of the auditory outer hair cells », *PNAS*, 111, (25), p. 9307-9312, juin 2014.

[24] HOBEN R., EASOW G. et al., « Outer Hair Cell and auditory nerve function in speech recognition in quiet and in background noise », *Frontiers in Neuroscience*, avril 2017, 11 :157. doi : 10.3389/fnins.2017.00157.

[25] MUDRY A., « Les otoémissions acoustiques (OEA) », juillet 2016. `https://www.oreillemudry.ch`.

[26] AUBANEL V., DAVIS C. et KIM J., « Exploring the role of brain oscillations in speech perception in noise : Intelligibility of isochronously retimed

speech », *Frontiers in Human Neuroscience*, août 2016, 10 :430. doi : 10.3389/fnhum.2016.00430.

[27] ZION GOLUMBIC E. et al., « Mechanisms underlying selective neuronal tracking of attended speech at a "Cocktail Party" », *Neuron*, 77, (5), p. 980-991, mars 2013.

[28] GARDIER S., « Comment notre cerveau filtre les bruits ambiants », mars 2013. http://sante.lefigaro.fr.

[29] Ooreka, « Perte auditive : à partir de quand doit-on agir? ». https://appareil-auditif.ooreka.fr.

[30] LORENZI A. et CHAIX B., « Représentation du son. Échelle des Bels », décembre 2016. http://www.cochlea.eu.

[31] PLAPOUS C., *Traitements pour la réduction de bruit. Application à la communication parlée*, Thèse, Traitement du signal et de l'image, Université Rennes 1, 2005.

[32] ZWAARDEMAKER H., « Sur la sensation de l'oreille aux différentes hauteurs des sons », *L'Année Psychologique*, 10, (1), p. 161-178, 1903.

[33] MILLOT L., *Traitement du signal audiovisuel*, Paris, Dunod, 2008.

[34] CARRAT R., *L'oreille numérique*, Les Ulis, EDP Sciences, 2009.

[35] CNRTL, « TCOF », 2012. http://www.cnrtl.fr/corpus/tcof.

[36] Société Française d'Audiologie, « Guide des Bonnes Pratiques en Audiométrie de l'Adulte », novembre 2010. http://sfaudiologie.fr/Drupal.

[37] Le Monde, « La télévision, média le plus consommé en france », mars 2013. http://www.lemonde.fr.

[38] WOJCIAK T., « Media in Life : 43,9 contacts médias et multimédias par jour et par personne », avril 2016. http://www.cbnews.fr.

[39] « Phonème. » Def. 1. Dictionnaire de l'Académie française, 9ème Édition.

[40] LAFON J.-C., *Le test phonétique et la mesure de l'audition*, Paris, Dunod, 1964.

[41] GROMER B. et WEISS M., *Lire, Tome 1 : apprendre à lire*, Paris, Armand Colin, 1990.

[42] NEW B., PALLIER C., BRYSBAERT M. et FERRAND L., « Lexique 2 : A New French Lexical Database », *Behavior Research Methods, Instruments, & Computers*, 36, (3), p. 516-524, 2004.

[43] MALHERBE M., *L'euphonie des romances sans paroles de Paul Verlaine*, Amsterdam, Rodopi, 2004.

[44] WIOLAND F., *Prononcer les mots du français*, Paris, Hachette, 1991.

[45] DAENINCKX D., *L'espoir en contrebande*, « Coupe-Coupe », Paris, Gallimard, 2013.

[46] Médiametrie, Médiamat, « Communiqué de presse. L'audience de la télévision du 17 au 23 avril 2017 ». `http://www.mediametrie.fr/television/communiques`.

[47] CSA, « Questions sur la bande 700 MHz et son transfert aux opérateurs de téléphonie mobile ». `http://www.csa.fr`.

[48] Code de la Propriété Intellectuelle. Article L122-5.

[49] Code de la Propriété Intellectuelle. Article L211-3.

[50] DELOZIER T., « Dans quel pays est-on le plus grand ? », juillet 2016. `http://sante.lefigaro.fr`.

[51] Fnac, « À quelle distance regarder votre télévision ? », août 2015. `http://www.fnac.com`.

[52] La Croix, « La législation concernant les nuisances sonores », juillet 2010. `http://www.la-croix.com`.

[53] Gérer son Audition, « Échelle du bruit au quotidien ». `http://www.gerersonaudition.com`.

[54] LEIPP É., « L'intégrateur de densité spectrale (IDS) et ses applications », *Bulletin du Groupe d'Acoustique Musicale (GAM) n°94*, Laboratoire d'Acoustique Musicale, Université Paris 6, décembre 1977.

[55] MILLOT L. et PELÉ G., « An Objective and Subjective Alternative Audio Sounds and Scenes Analysis : the IDS », *International Symposium on Musical Acoustics*, september 2007, Barcelone.

[56] ZWICKER E. et FASTL H., *Psychoacoustics : Facts and Models*, Berlin, Springer, 2007 (troisième édition).

[57] Norme ISO 266 :1997.

[58] MILLOT L., « Analyse IDS : Principe des spectres cumulés », cours d'Acoustique pour l'ENS Louis-Lumière, avril 2015.

[59] MILLOT L., « Analyse IDS : Recherche de découpages fréquentiels », cours de Traitement du Signal pour la Formation Supérieure aux Métiers du Son, avril 2017.

[60] DELERCE X., « Faire du bruit pendant le test ou un test dans le bruit ? », avril 2015. `http://www.blog-audioprothesiste.fr`.

[61] Pure Data, « More About Pure Data ». `https://puredata.info`.

[62] MILLOT L. et PELÉ G., « An alternative approach for the convolution in time-domain : the taches-algorithm », *124th Convention of the Audio Engineering Society, Paper 7412*, mai 2008, Amsterdam.

Annexe 1 : Contenu des corpus oraux

Corpus de voix parlées

Liste de mots

Ci-dessous, quatre « éléments » de la liste cochléaire du professeur Lafon, présentés sous la forme de tests pour l'audiométrie vocale, et proposés par le Collège National d'Audioprothèse [1] :

Date		Voix M F E		Voix M F E		Voix M F E		Voix M F E		Voix M F E	
Audiomètre											
Opérateur		CD 1	piste								
Observations		**1**	5 25 35	**2**	6 26 36	**3**	7 27 37	**4**	8 28 38	**5**	9 29 39
		buée		bile		rôde		abbé		balle	
		ride		dors		fente		sud		soude	
		foc		sage		tige		fausse		mur	
		agis		gaine		grain		joute		nef	
		vague		fil		cave		dogue		change	
		croc		cru		bulle		acquis		gage	
		lobe		boule		somme		ville		trou	
		mieux		cale		maine		mare		mal	
		natte		bonne		preux		noce		tonne	
		col		rive		bord		appas		peur	
		fort		sol		rouille		route		rampe	
		soupe		tempe		oser		cil		puce	
		tonte		fauve		site		fête		cor	
		vêle		phase		bouée		veule		vite	
		nage		mule		sauve		chaise		rance	
		souche		chatte		chance		bâche		mouche	
		rogne		règne		gagne		souille		fille	
		/ 50		/ 50		/ 50		/ 50		/ 50	

FIGURE 1 – Les quatre premières sous-listes cochléaires de Jean-Claude Lafon, par le Collège National d'Audioprothèse

1. Cf. `http://www.college-nat-audio.fr/fichiers/img85a.pdf`

Puis, les sous-listes obtenues après modification de ces « éléments » selon la fréquence d'occurrence des phonèmes dans la langue française parlée, issue des travaux de François Wioland :

Noé	l'œil	raide	abbé
ride	dors	site	sud
boule	gueuse	tige	hautain
agis	salle	Alain	joute
plomb	heureux	cave	fille
crin	cru	fête	acquis
led	aisé	somme	tonte
mieux	cale	mène	nappe
natte	hareng	preux	mince
conque	sel	épais	appas
mère	bouée	rouille	phase
soupe	tempe	errer	cieux
ville	fauve	tente	mule
sensé	ronde	rive	veule
nage	mule	sauve	aîné
souche	date	danse	aqueux
raid	même	happant	souille

FIGURE 2 – Liste de mots à prononcer par le locuteur lors du premier enregistrement pour la constitution du corpus de voix parlées.

Voici un tableau comparant les fréquences d'occurrence des phonèmes dans le discours selon le professeur Wioland, à celles que nous avons obtenues dans la liste de mots :

	Fréquence d'après F. Wioland (en %)	Fréquence obtenue dans la liste (en %)
/ E /	10,6	10
/ a /	8,55	8,5
/ R /	7,25	7
/ s /	6	6,5
/ l /	5,63	6
/ t /	5,335	6
/ l /	5,115	5
/ Œ /	4,31	4,5
/ k /	4,06	4
/ d /	4,035	4,5
/ m /	3,845	4,5
/ p /	3,715	4
/ O /	3,36	3
/ ɑ̃ /	3,09	3

/ n /	3,095	3
/ v /	2,755	3
/ u /	2,43	1,5
/ ɔ̃ /	2,255	2
/ j /	2	2
/ y /	1,9	2
/ ɛ̃ /	1,845	2
/ ʒ /	1,66	2
/ z /	1,535	1,5
/ f /	1,4	2
/ b /	1,31	1,5
/ ʃ /	0,535	0,5
/ ɥ /	0,515	0
/ g /	0,475	0,5

FIGURE 3 – Fréquence d'occurence des phomènes dans le discours selon François Wioland, et d'après la liste de mots de la Figure 2.

Texte

Ci-dessous, la nouvelle de Dider Daeninckx devant être lue pour le deuxième enregistrement de ce corpus :

La légende prétend qu'aussitôt décapité, Denis, premier évêque de Paris, prit sa tête dans ses mains, traversa ce qui allait devenir le Quartier latin, escalada la colline de Montmartre, puis redescendit droit sur un village qui bien des temps plus tard fut baptisé du nom de ce martyr chrétien : Saint-Denis. C'est à l'endroit même où la vie le quitta que fut édifiée une orgueilleuse basilique. Elle renferme les dépouilles des rois de France dont celle, ironie de l'histoire, d'un autre décapité célèbre, Louis le seizième. Si l'on consulte un plan du secteur, on constate que Denis, lesté de son chef, a foulé les terres du Cornillon et qu'il est probablement passé sur l'ellipse aujourd'hui dévolue au Stade de France. Peut-être même que pour reprendre, sinon son souffle, du moins ses forces, il a un instant posé sa tête dans l'herbe et que, mille ans plus tard, une puissance obscure a conduit géomètres et architectes à faire de cet emplacement le point d'engagement de la finale de la Coupe du monde !

On sait que les dernières décennies ne furent pas tendres pour le paysage de ma ville natale. Tous les minerais de France, tous les produits chimiques convergeaient sur ce coin de banlieue. Toutes les misères aussi. Un ciel lourd naissait en permanence des cheminées qui hérissaient le quartier Pleyel, le quartier du Landy. En lieu et place du stade s'élevaient d'autres enceintes arrondies, celles des gazomètres. Les voies du secteur avaient puisé leur nom dans leur ombre : rue du Gaz, impasse du Gaz... Le monde entier fournissait des bras aux industries dionysiennes. Au début du siècle, c'est par milliers que les jeunes désertaient les villages du Morbihan, du Morvan et de l'Aubrac pour venir peupler les ateliers des usines automobiles Delaunay-Belleville ou Hotchkiss. Les Bretons étaient les plus nombreux. Les plus

rejetés aussi à cause de leur langue, étrangère d'apparence. Les registres policiers gardent la trace des chasses au faciès dont ils étaient victimes, le samedi soir, et un nom fut même inventé par un fonctionnaire pour les désigner : les « bretonnades ».

Plus tard, ce furent les Italiens qui prirent la relève, puis les cohortes espagnoles et républicaines, vaincues par les armées noires de Franco. Le quartier se transforma en « petite Espagne » où résonnèrent longtemps les accords de guitare rageurs de Paco Ibañez. D'autres exilés, affamés de pain ou de liberté, trouvèrent le chemin à leur tour. Portugais, Algériens, Maliens, Yougoslaves, s'installèrent par vagues successives dans les pauvres maisons laissées vacantes par ceux qui avaient enfin réussi à gagner le cœur des villes.

Aujourd'hui, c'est d'un pas volontaire et le sourire aux lèvres que de tous les pays du monde on se presse vers Saint-Denis pour encourager les dieux du Stade de France, pour applaudir les vainqueurs de la Coupe. Toutes les couleurs, tous les drapeaux, toutes les musiques. Comme si les hommes s'étaient enfin aperçus qu'avant d'être le nom d'un martyr, Denis désignait le dieu grec du vin et de la vigne, le dieu de la fête : Dionysos.

DAENINCKX, Didier. *L'espoir en contrebande*, « Coupe-Coupe », p. 111-113, Paris, Gallimard, 2013.

Corpus de voix télévisuelles

Voici tous les programmes TV ayant été utilisés pour la composition du corpus :

- **TF1** (`https://www.tf1.fr/tf1`) :
  - JT de 20h du 22/01/2017
  - JT de 13h du 23/01/2017
  - JT de 20h du 23/01/2017
  - JT de 13h du 24/01/2017
  - JT de 13h du 25/01/2017
  - Météo du 22/01/2017
  - Météo du 23/01/2017
  - Météo du 24/01/2017
- **France 2** (`http://pluzz.francetv.fr/a-z/france2`) :
  - JT de 13h du 26/01/2017
  - JT de 13h du 04/04/2017
  - JT de 20h du 04/04/2017
  - JT de 13h du 05/04/2017
  - Météo du 05/04/2017
- **France 3** (`http://pluzz.francetv.fr/a-z/france3`) :
  - JT 19/20 du 01/04/2017
  - JT 12/13 du 05/04/2017
  - JT 19/20 du 05/04/2017
  - Météo du 26/01/2017
  - Météo du 04/04/2017
- **M6** (`http://www.6play.fr/m6`) :
  - JT 12.45 du 04/04/2017
  - JT 19.45 du 04/04/2017
  - JT 12.45 du 05/04/2017
  - Météo du 05/04/2017

Annexe 2 : Tests perceptifs

Résultats généraux

Voici les résultats des premiers tests perceptifs réalisés avec l'IDS *Speech Enhanceur*. On peut observer une comparaison des balances spectrales effectuées par les sujets pour les différentes scènes [2].

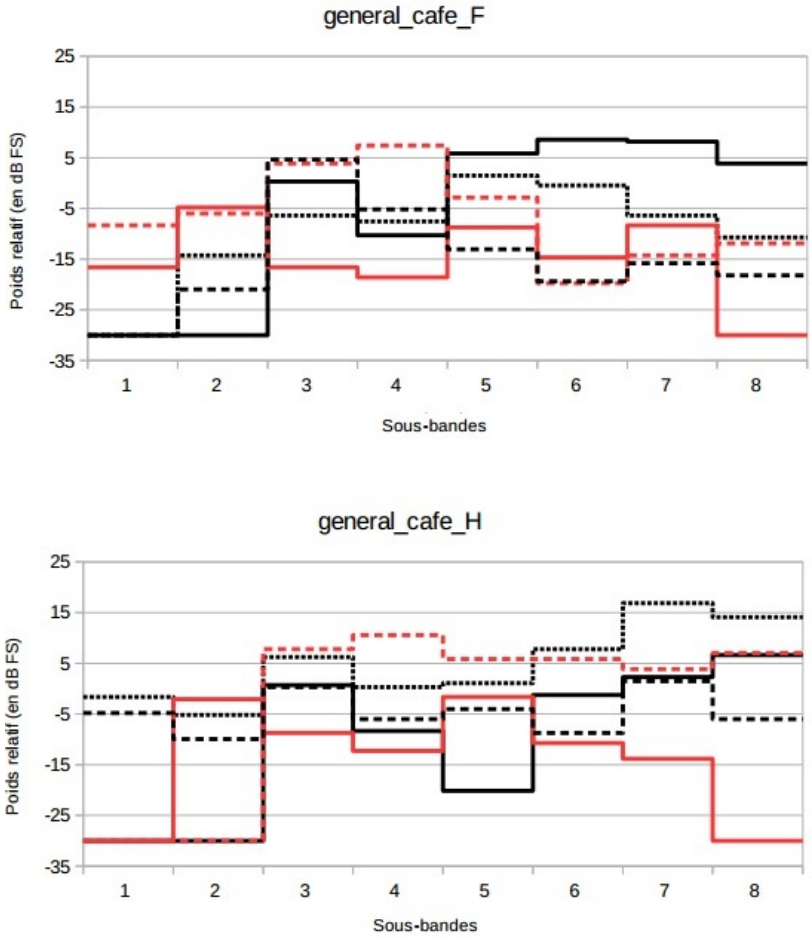


FIGURE 4 – Résultats des premiers tests pour l'ambiance de café avec le découpage G

2. Les titres des graphiques sont formés avec : le type de découpage, l'ambiance étudiée, et le *stimulus* de parole utilisé (F pour voix féminine et H pour voix masculine).

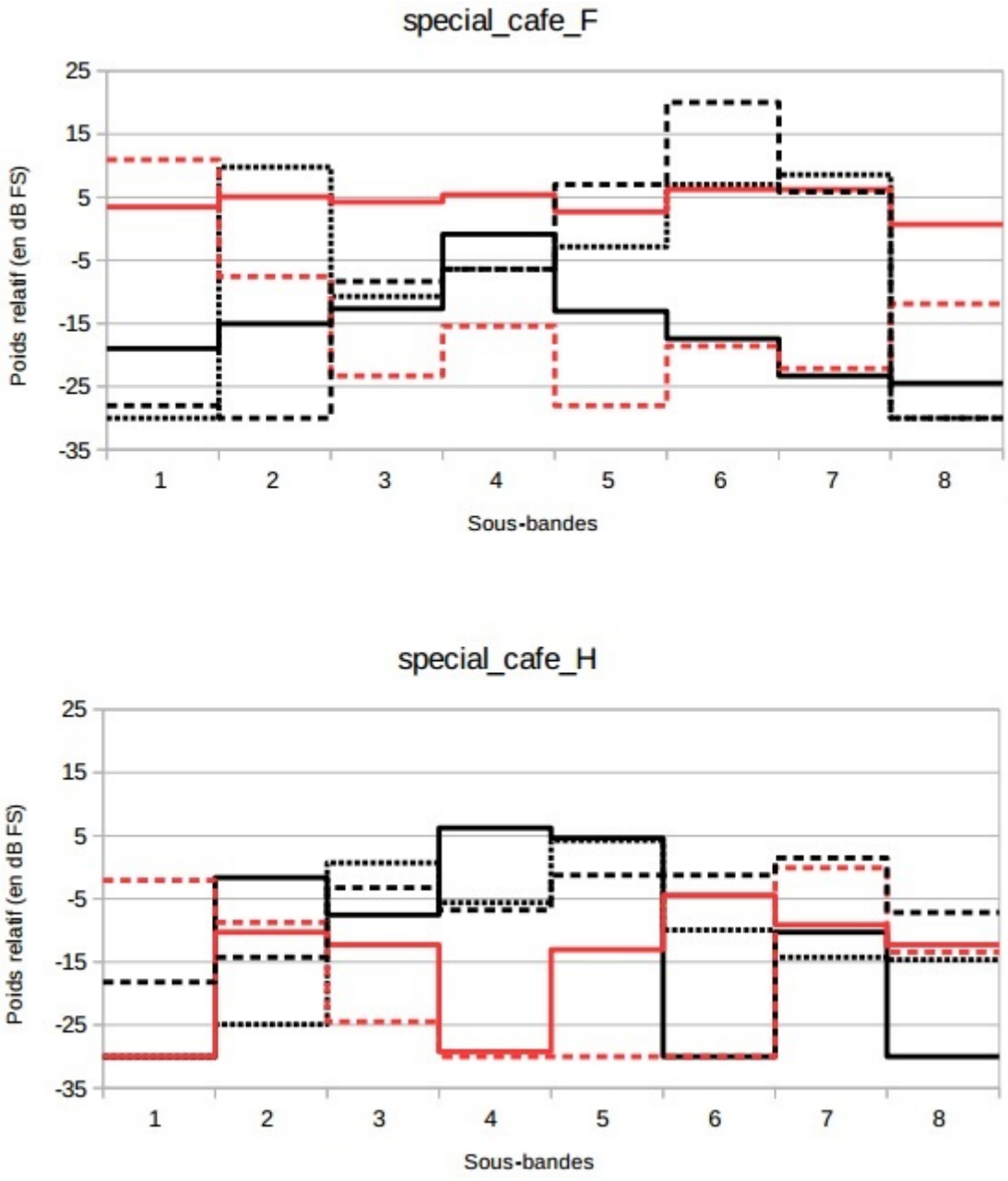


FIGURE 5 – Résultats des premiers tests pour l'ambiance de café avec les découpages S

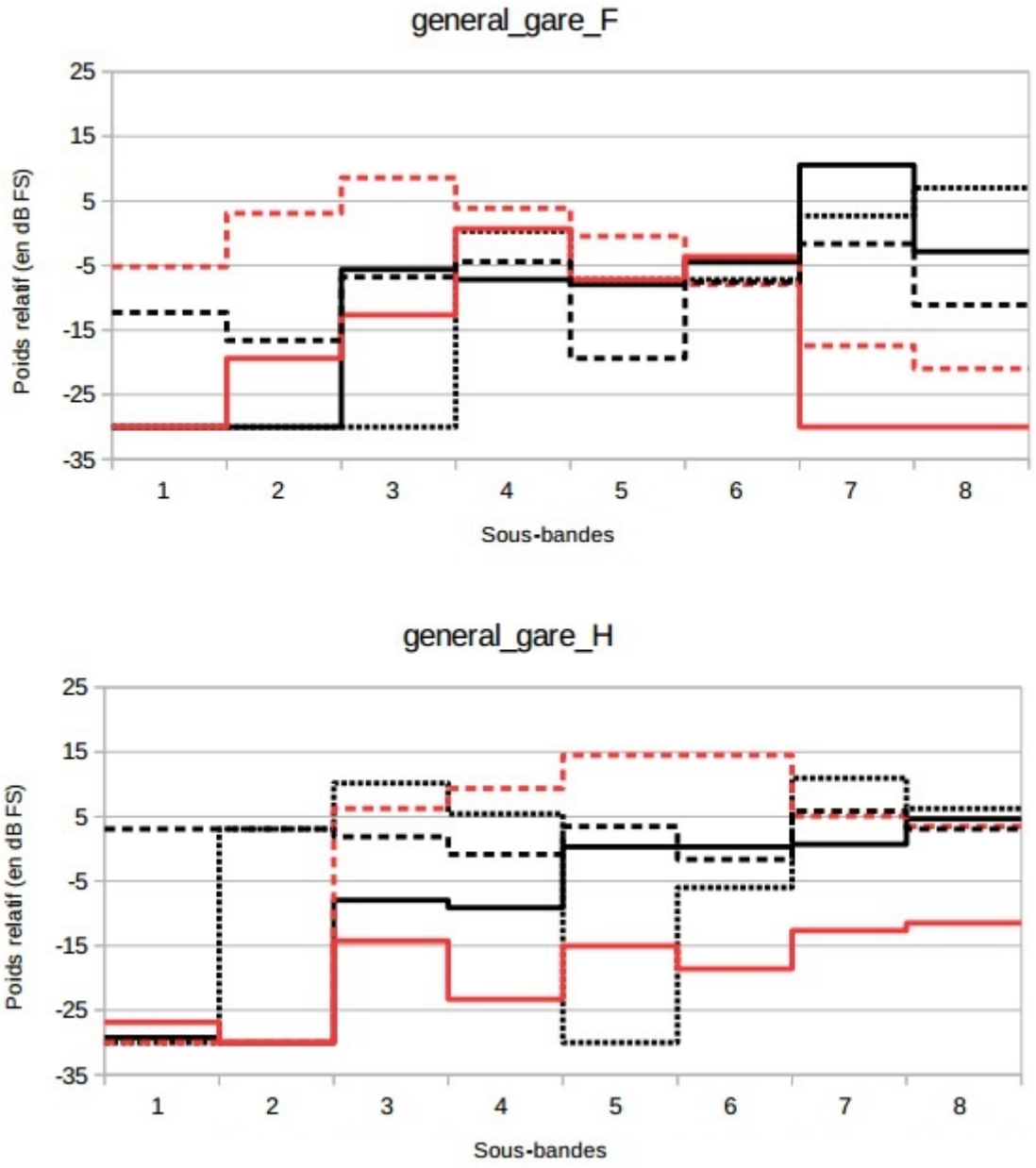


FIGURE 6 – Résultats des premiers tests pour l'ambiance de gare avec le découpage G

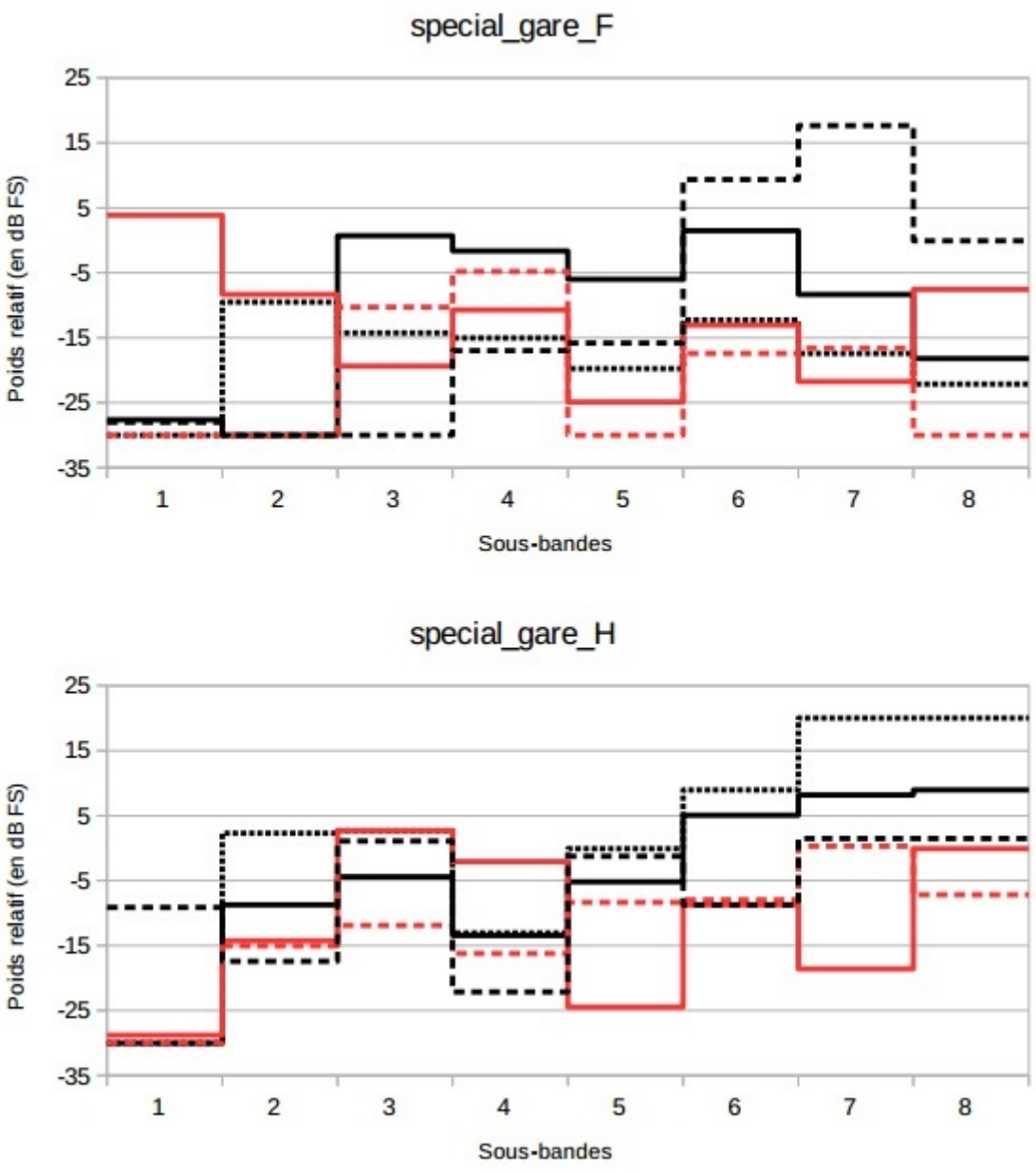


FIGURE 7 – Résultats des premiers tests pour l'ambiance de gare avec les découpages S

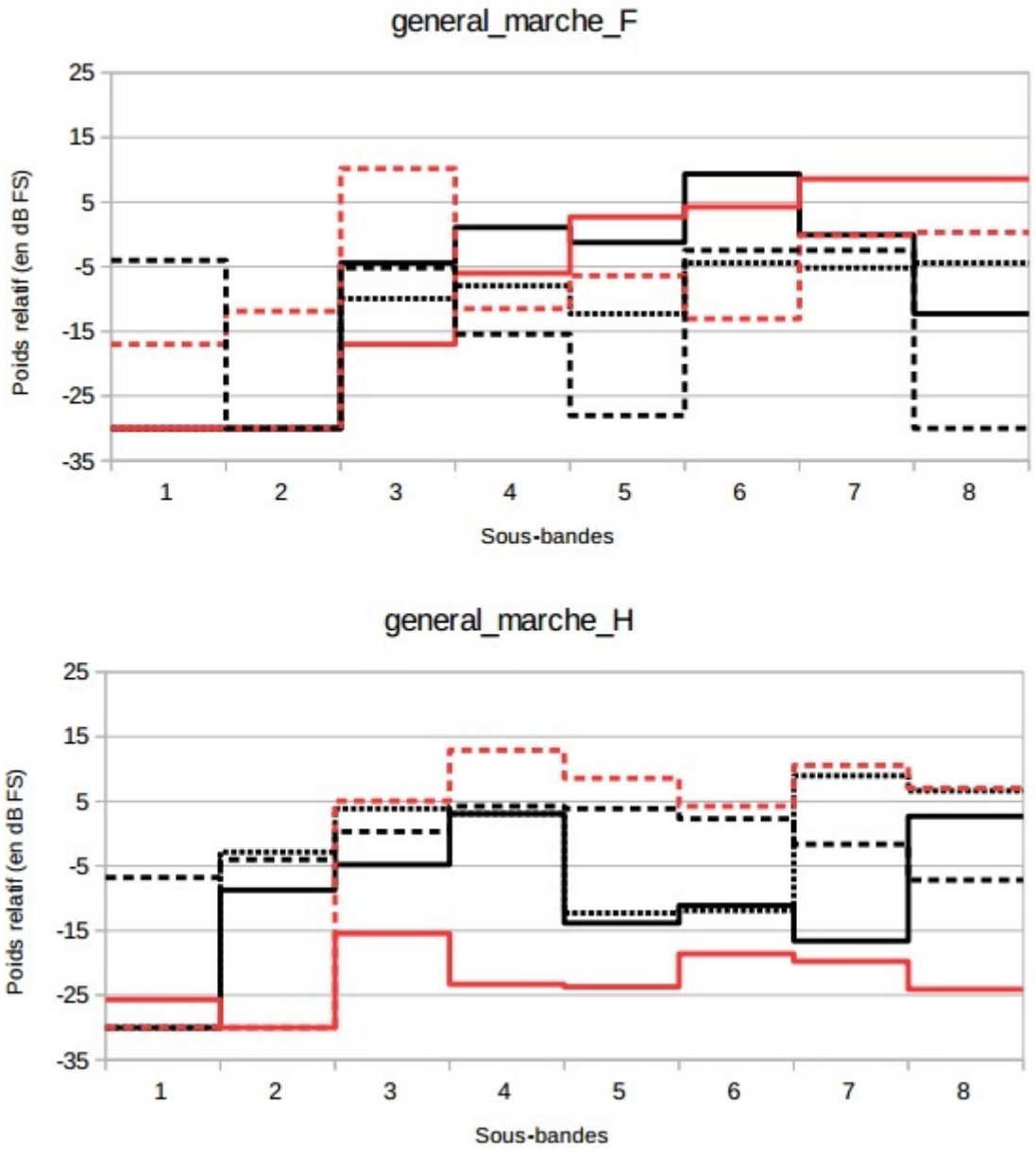


FIGURE 8 – Résultats des premiers tests pour l'ambiance de marche avec le découpage G

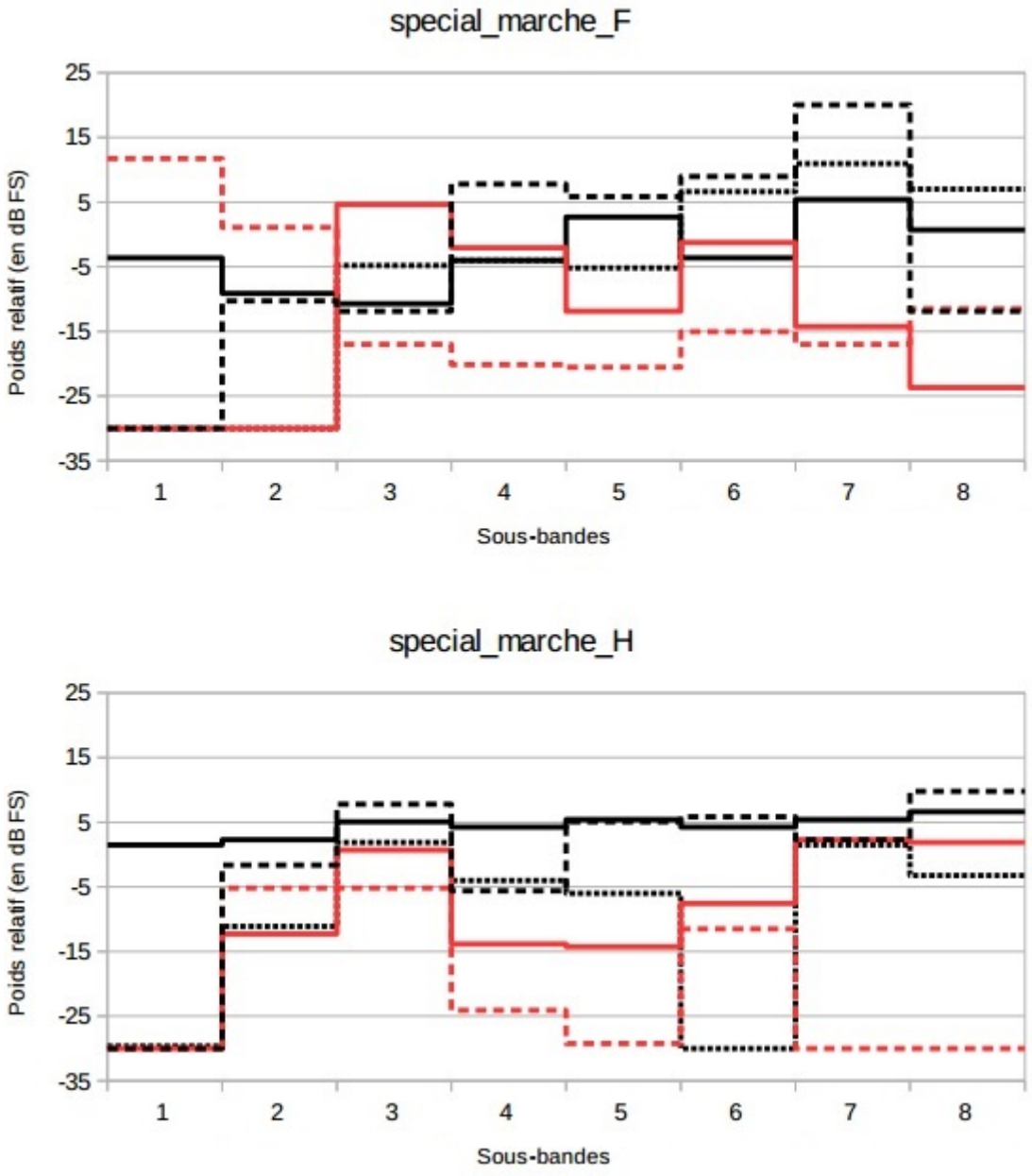


FIGURE 9 – Résultats des premiers tests pour l'ambiance de marché avec les découpages S

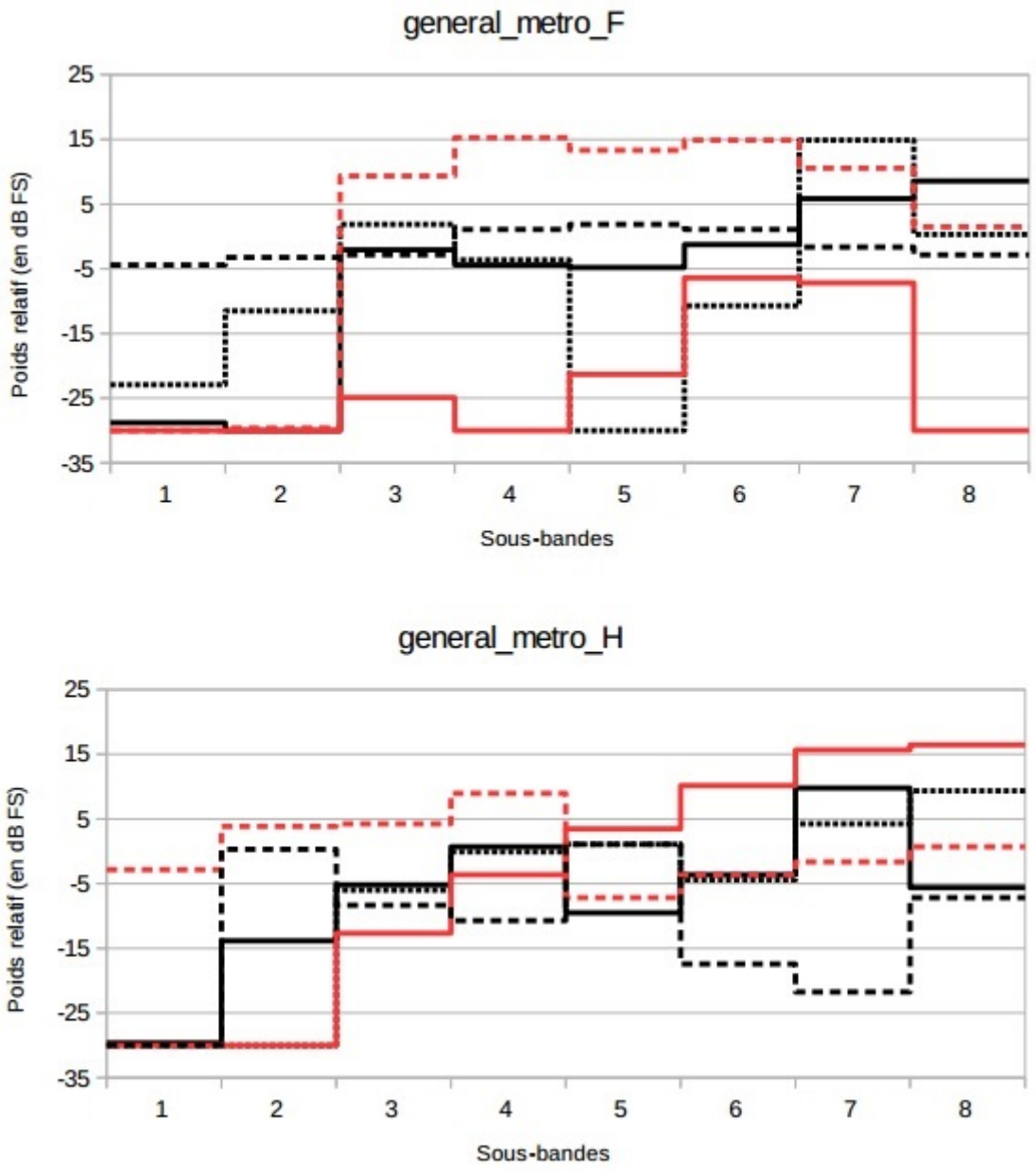


FIGURE 10 – Résultats des premiers tests pour l'ambiance de métro avec le découpage G

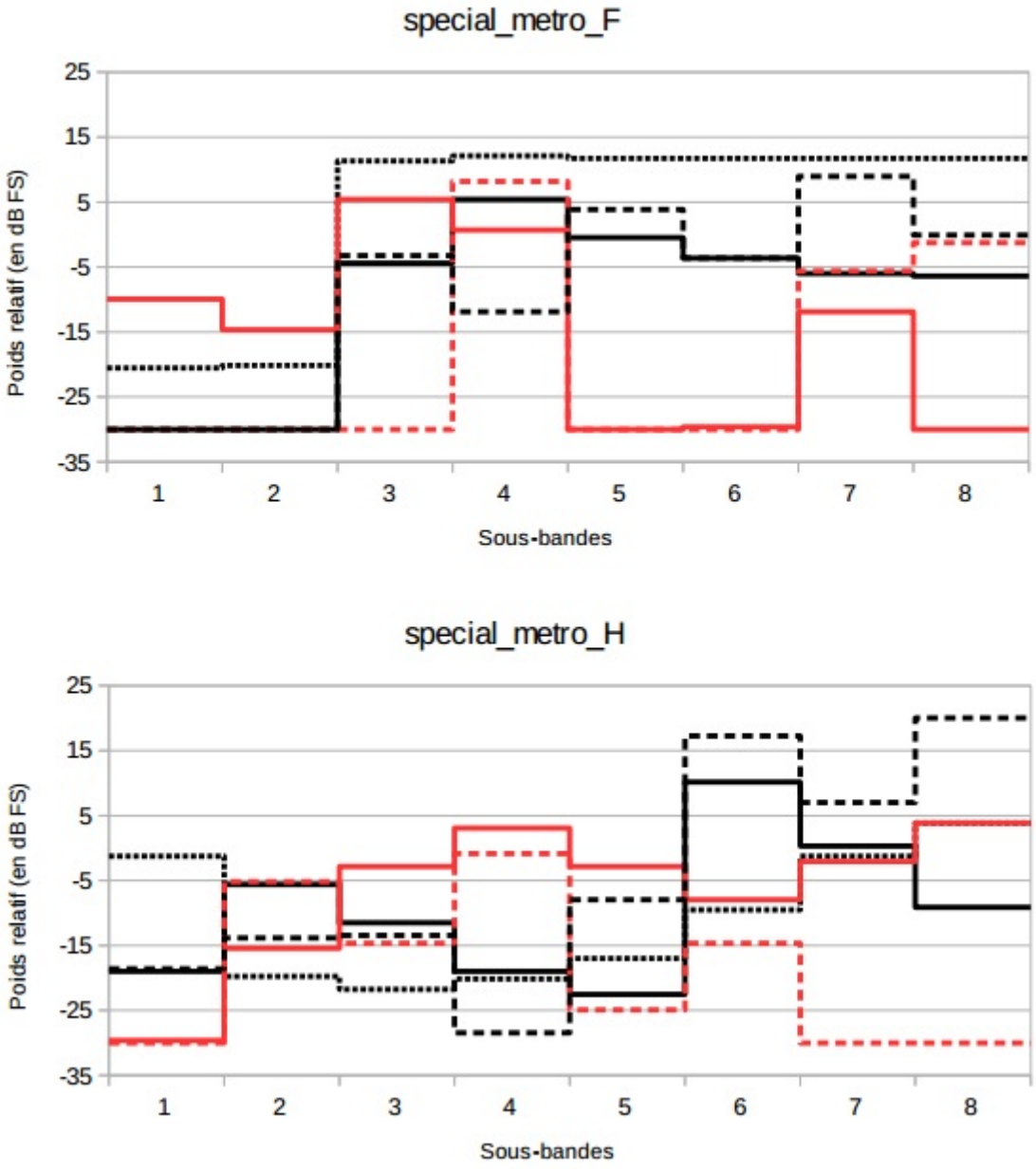


FIGURE 11 – Résultats des premiers tests pour l'ambiance de métro avec les découpages S

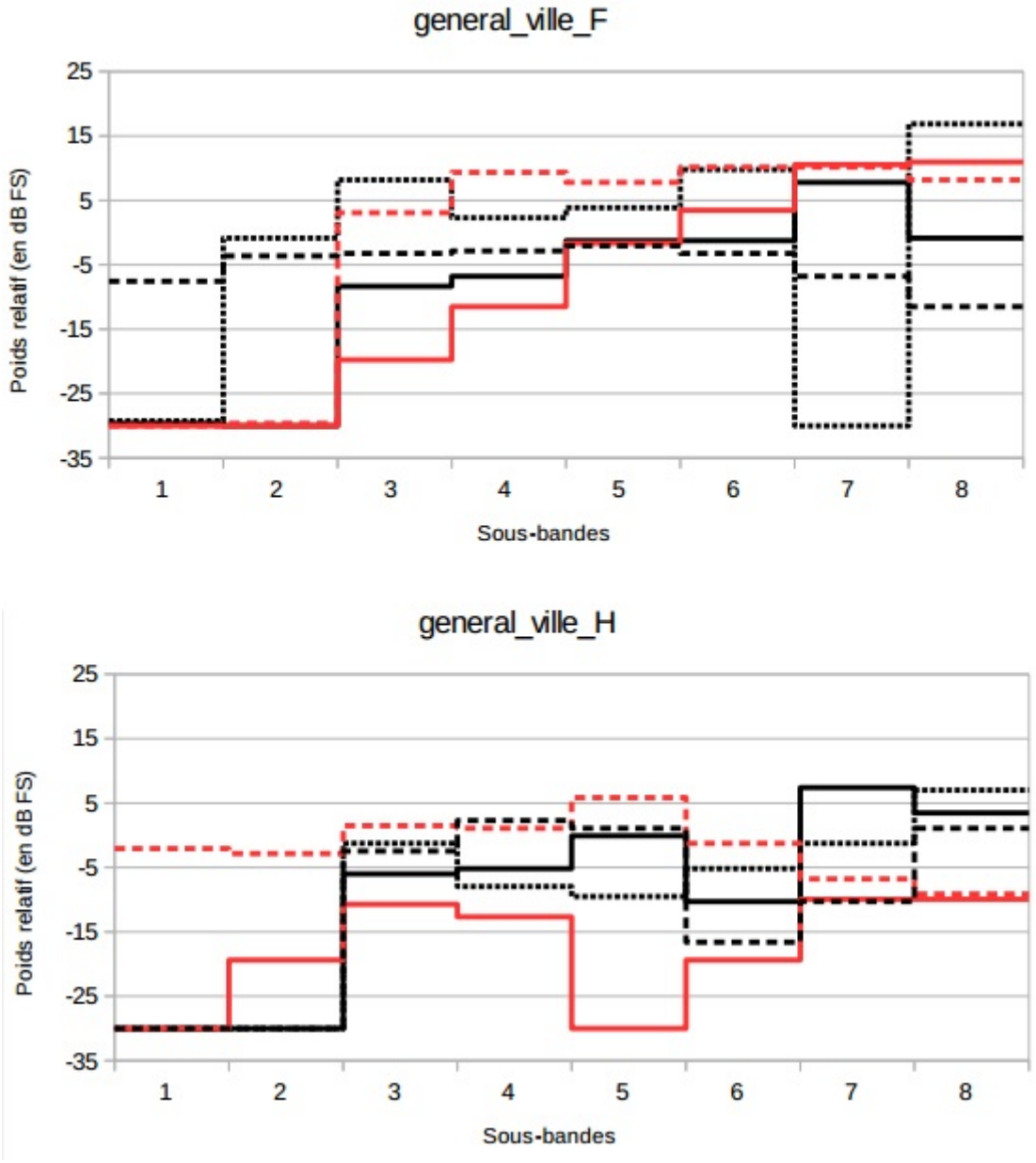


FIGURE 12 – Résultats des premiers tests pour l'ambiance de ville avec le découpage G

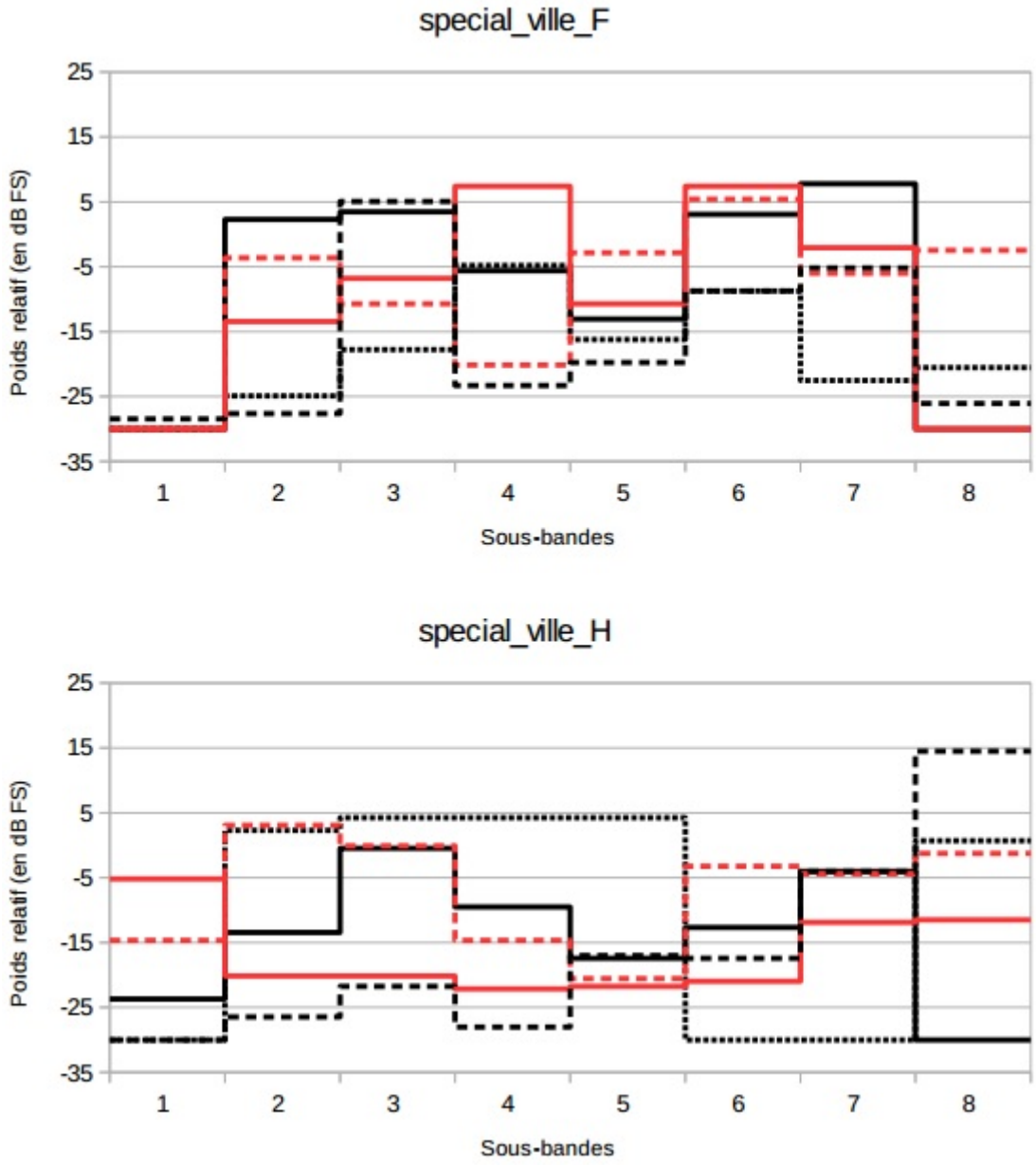


FIGURE 13 – Résultats des premiers tests pour l'ambiance de ville avec les découpages S

Résultats pour les malentendants

Voici l'audiogramme du sujet n°1 (le sujet nous a indiqué que cette déficience provenait d'un accident de *Jet Ski*) :

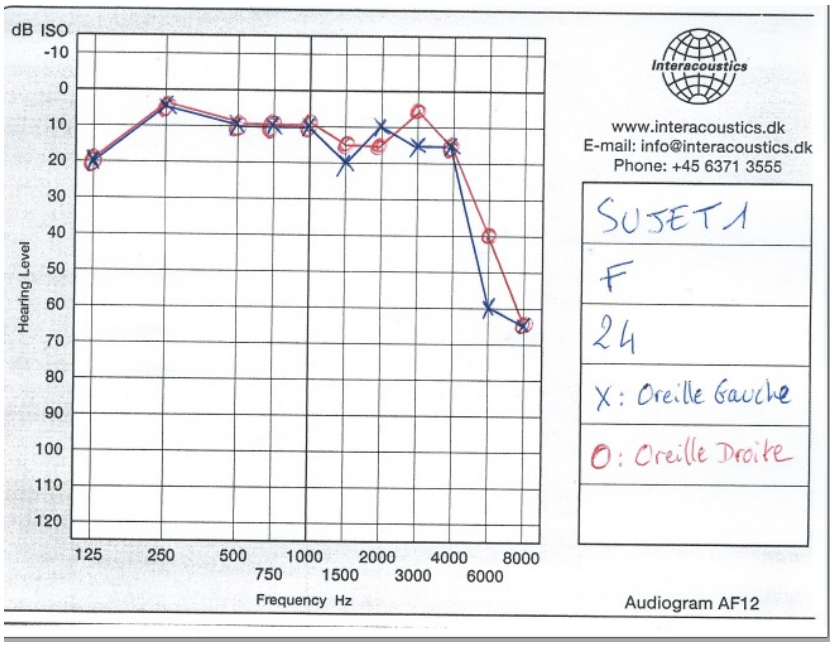


FIGURE 14 – Audiogramme du sujet n°1

Et les balances spectrales qu'il a effectuées avec les découpages S :

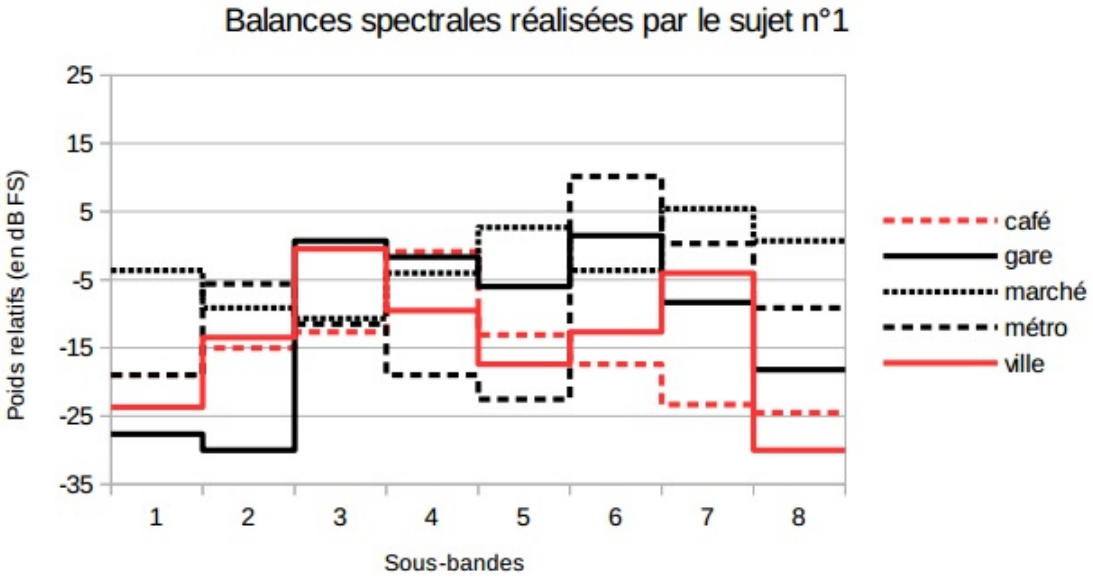


FIGURE 15 – Balances spectrales obtenues pour le sujet n°1

Voici l'audiogramme du sujet n°17 (le sujet nous a indiqué que cette déficience provenait d'acouphènes permanents) :

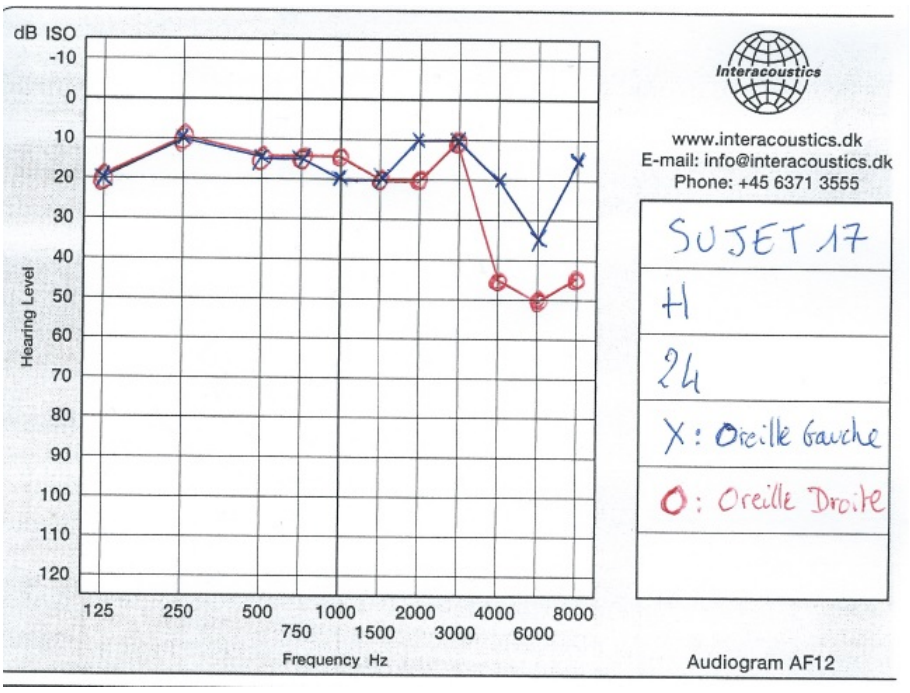


FIGURE 16 – Audiogramme du sujet n°17

Et les balances spectrales qu'il a effectuées avec le découpage G :

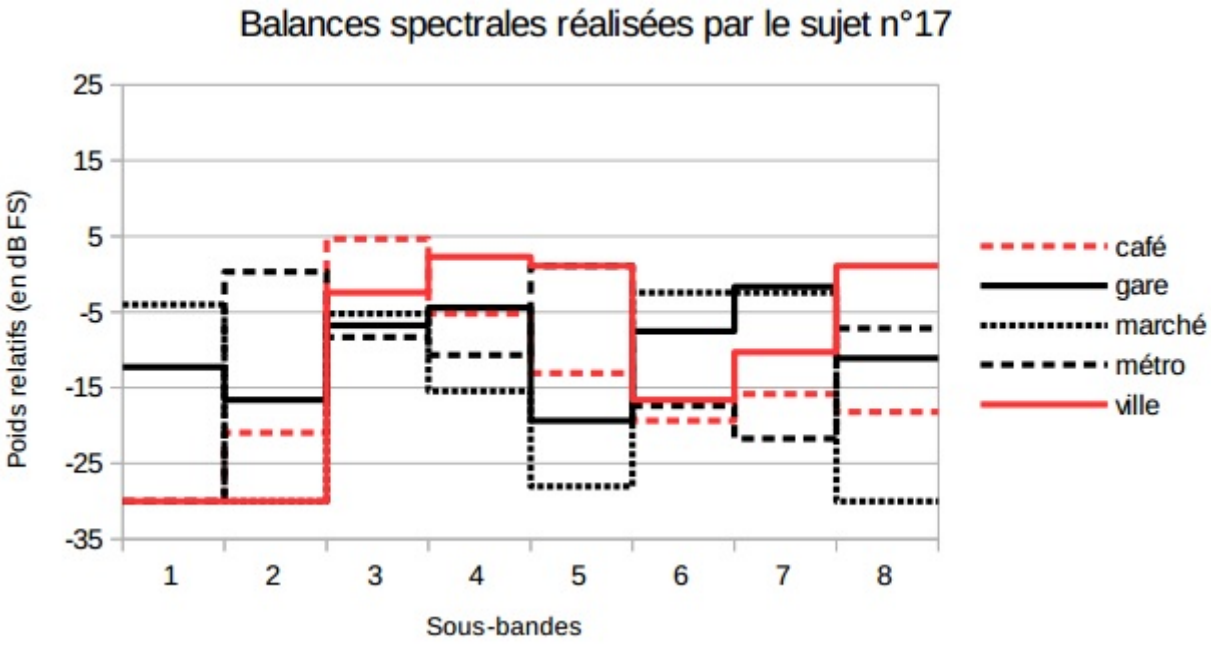


FIGURE 17 – Balances spectrales obtenues pour le sujet n°17

Patch *Pure Data* de l'IDSSE

Voici le patch général de l'IDSSE « en fonctionnement » :

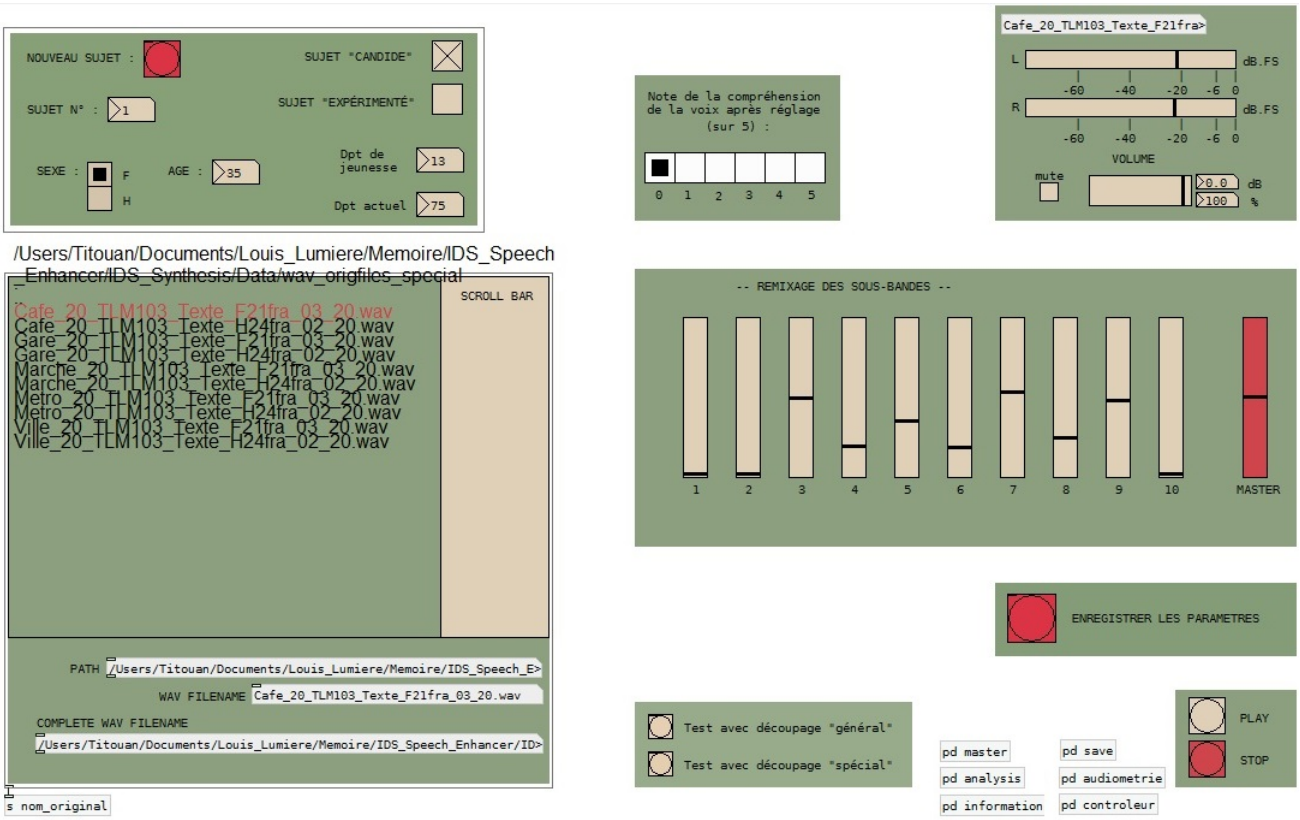


FIGURE 18 – Patch général de l'IDSSE

Annexe 3 : Spécifications techniques du matériel utilisé

Microphones

Voici les courbes de réponse en fréquence des deux microphones utilisés pour l'enregistrement des corpus :

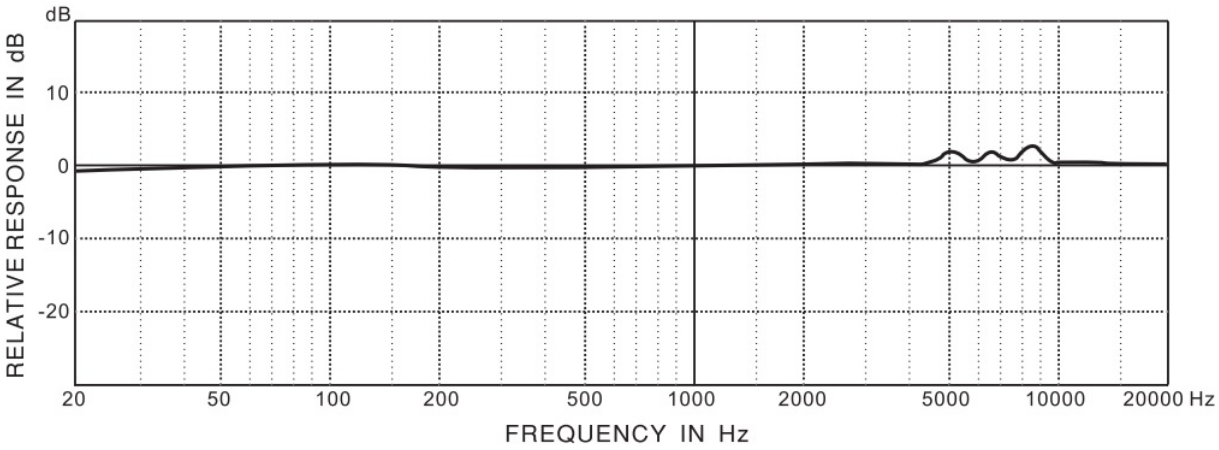


FIGURE 19 – Courbe de réponse en fréquence du *Behringer ECM8000*.

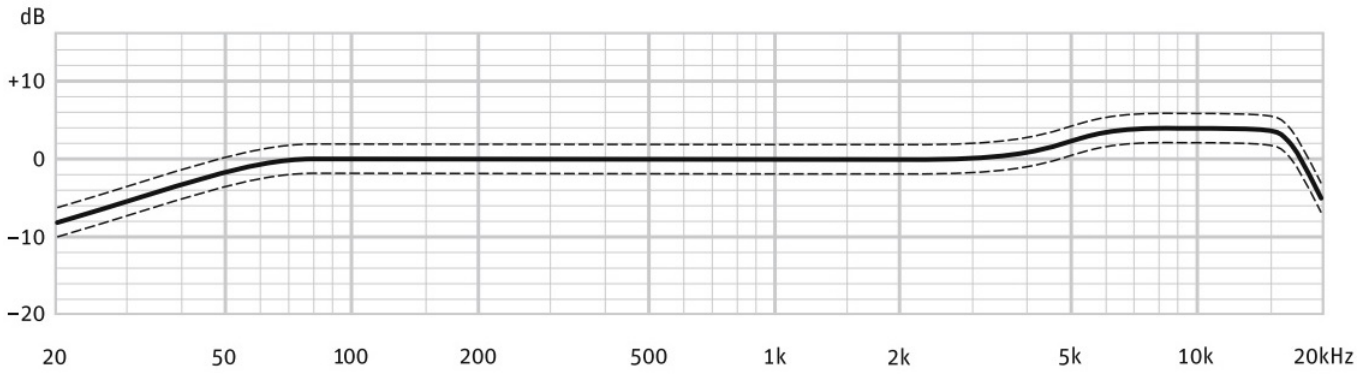


FIGURE 20 – Courbe de réponse en fréquence du *Neumann TLM103*.

Et leurs diagrammes directionnels :

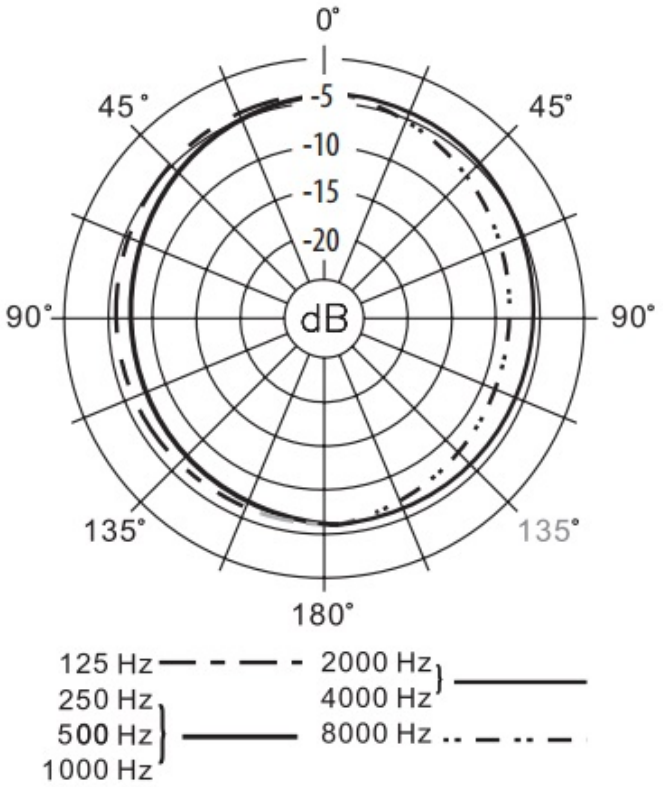


FIGURE 21 – Diagramme directionnel du *Behringer ECM8000*.

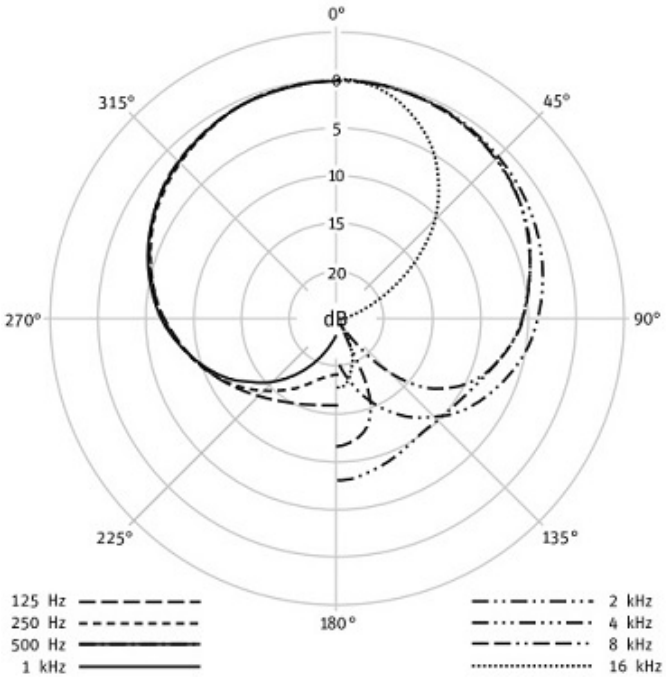


FIGURE 22 – Diagramme directionnel du *Neumann TLM103*.

Contrôleur MIDI

Voici les caractéristiques techniques de la *Behringer BCF2000* :

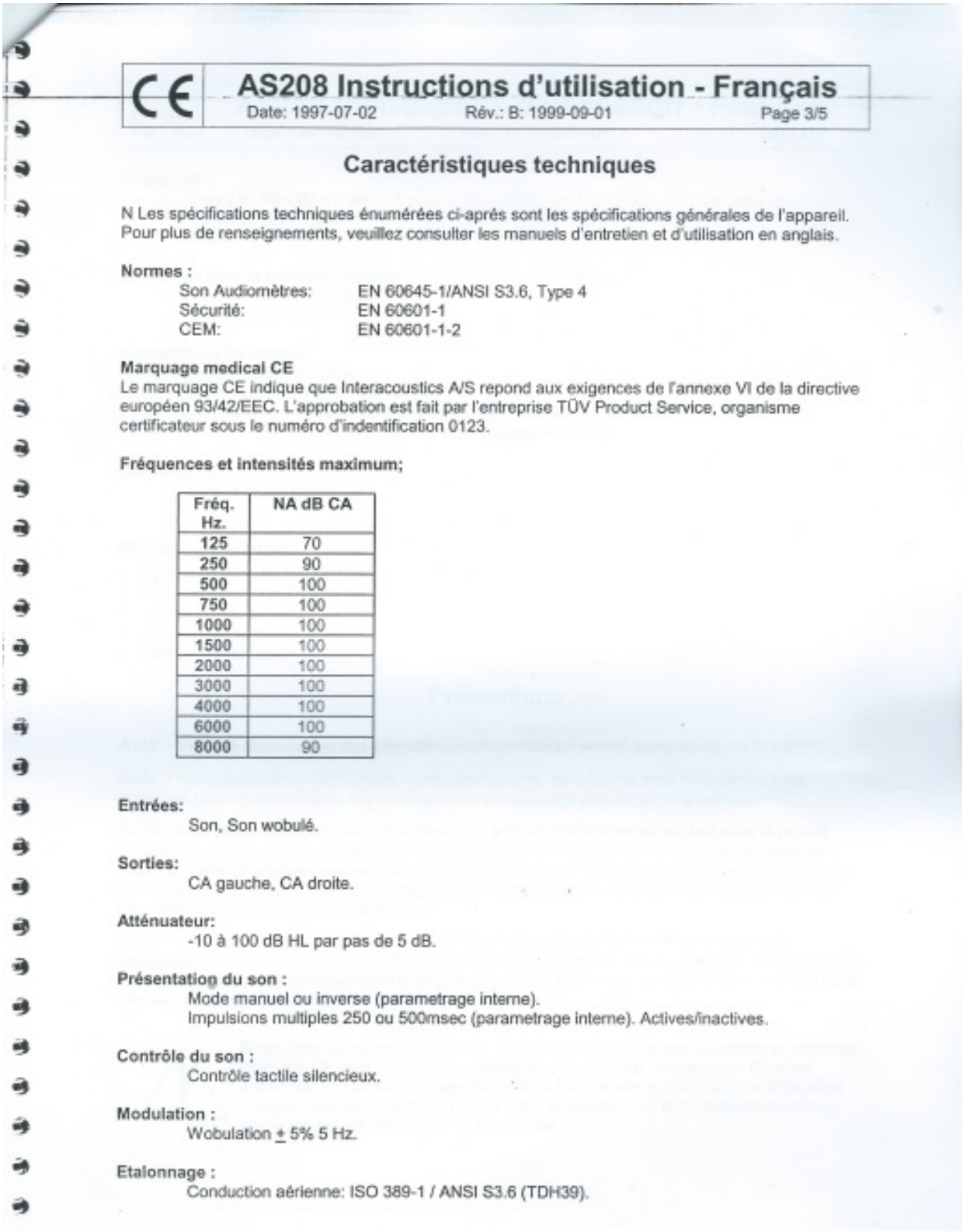
Specifications

USB Interface	
Type	Full-speed 12 MBit/sec. USB MIDI class-compliant
MIDI Interface	
Type	5-pin DIN connectors IN, OUT A, OUT B / THRU
Control Elements	
Controls	8 motorized 100-mm faders 8 infinitely variable push encoders with LED rings
Keys	20 keys 10 system keys (4x Encoder Group, 4x programming, 2x Preset)
Display	
Type	4-digit 7-segment LED display
Switched Inputs	
Footswitch	1 x ¼" TS connector with automatic polarity detection
Foot pedal	1 x ¼" TRS connector
Power Supply	
Voltage	100 to 240 V~, 50/60 Hz
Power Consumption	max. 10 W
Fuse	T 1 A H 250 V
Mains Connection	Standard IEC receptacle
Dimensions / Weight	
Dimensions (H x W x D)	approx. 330 x 100 x 300 mm (13 x 3.94 x 11.8")
Weight	approx. 2.7 kg (5.9 lbs)

FIGURE 23 – Caractéristiques techniques de la *Behringer BCF2000*

Audiomètre

Voici les caractéristiques techniques de l'audiomètre *Interacoustics AS208* :



CE **AS208 Instructions d'utilisation - Français**
Date: 1997-07-02 Rév.: B: 1999-09-01 Page 3/5

Caractéristiques techniques

N Les spécifications techniques énumérées ci-aprés sont les spécifications générales de l'appareil. Pour plus de renseignements, veuillez consulter les manuels d'entretien et d'utilisation en anglais.

Normes :

Son Audiomètres:	EN 60645-1/ANSI S3.6, Type 4
Sécurité:	EN 60601-1
CEM:	EN 60601-1-2

Marquage medical CE
Le marquage CE indique que Interacoustics A/S repond aux exigences de l'annexe VI de la directive européen 93/42/EEC. L'approbation est fait par l'entreprise TÜV Product Service, organisme certificateur sous le numéro d'indentification 0123.

Fréquences et intensités maximum;

Fréq. Hz.	NA dB CA
125	70
250	90
500	100
750	100
1000	100
1500	100
2000	100
3000	100
4000	100
6000	100
8000	90

Entrées:
Son, Son wobulé.

Sorties:
CA gauche, CA droite.

Atténuateur:
-10 à 100 dB HL par pas de 5 dB.

Présentation du son :
Mode manuel ou inverse (parametrage interne).
Impulsions multiples 250 ou 500msec (parametrage interne). Actives/inactives.

Contrôle du son :
Contrôle tactile silencieux.

Modulation :
Wobulation ± 5% 5 Hz.

Etalonnage :
Conduction aérienne: ISO 389-1 / ANSI S3.6 (TDH39).

FIGURE 24 – Caractéristiques techniques de l'*Interacoustics AS208*

Annexe 4 : Guide pour l'utilisation de l'IDSSE

Avec la version électronique de ce mémoire il est possible de télécharger le patch *Pure Data* de l'IDS *Speech Enhancer*[3]. Voici une explication synthétique de son utilisation.

- Au préalable il faut avoir une version *Pure Data* à jour installée sur l'ordinateur utilisé. Il est possible de télécharer le logiciel à cette adresse : `https://puredata.info/downloads/pd-extended`.

- Une fois cette étape effectuée, télécharger puis décompresser le fichier de l'IDSSE disponible avec la version électronique du mémoire : *IDS_Speech_Enhancer.rar*.

- Ouvrir le dossier *IDS_Speech_Enhancer* puis *Application_PureData* et enfin le fichier *IDS_Speech_Enhancer.pd*. Une fenêtre avec le patch *Pure Data* de l'IDSSE s'ouvre alors.

- Pour commencer il faut choisir le découpage à utiliser. Dans la zone en haut à droite cliquer sur « Test avec découpage "général" » ou « Test avec découpage "spécial" ».

- Pour jouer une scène sonore cliquer sur le nom du fichier .wav souhaité dans la grande zone à gauche.

3. La version proposée au téléchargement est une version « allégée » et seulement deux scènes sonores sont disponibles. Pour obtenir la version complète de l'IDSSE vous pouvez en faire la demande à l'adresse : `titouan.ra@gmail.com`

- Monter ensuite les *faders* et faire varier leur niveau pour effectuer les réglages.

- En haut à gauche se trouve une zone pour indiquer les informations concernant le sujet.

- Enfin, appuyer sur « ENREGISTRER LES PARAMETRES » crée un fichier dans lequel est stocké toutes les données indiquées ainsi que le poids des différentes sous-bandes. Ce fichier est disponible dans le dossier *Balances_Spectrales*.