



HAL
open science

Correction centralisée des distorsions spectrales de la parole sur les réseaux téléphoniques

Gaël Mahé

► **To cite this version:**

Gaël Mahé. Correction centralisée des distorsions spectrales de la parole sur les réseaux téléphoniques. Traitement du signal et de l'image [eess.SP]. Université Rennes 1, 2002. Français. NNT: . tel-00114668

HAL Id: tel-00114668

<https://theses.hal.science/tel-00114668v1>

Submitted on 17 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 2794

THÈSE

présentée

DEVANT L'UNIVERSITÉ DE RENNES 1

pour obtenir

le grade de : **DOCTEUR DE L'UNIVERSITÉ DE RENNES 1**

Mention : Traitement du Signal

PAR

Gaël MAHÉ

Équipe d'accueil : France Télécom R&D, Labo Interactions par la Parole et le Son, Lannion

École Doctorale : Mathématiques, Informatique, Signal, Électronique et Télécommunications

Composante universitaire : Structure et Propriétés de la Matière

**Correction centralisée
des distorsions spectrales de la parole
sur les réseaux téléphoniques**

SOUTENUE LE 19 décembre 2002 devant la commission d'Examen :

M. Gérard FAUCON	Président
Mme Madeleine BONNET	Rapporteur
M. Phillip A. REGALIA	Rapporteur
M. Jean-Marc BOUCHER	Directeur
M. André GILLOIRE	Examinateur
Mme Régine LE BOUQUIN-JEANNÈS	Examinateur

Remerciements

Mes remerciements s'adressent en premier lieu à André Gilloire, qui a encadré cette thèse. Son enthousiasme et ses compétences, ainsi que la confiance et l'autonomie qu'il m'a accordées dans l'orientation et la réalisation de ces travaux de thèse, ont été des atouts précieux.

Je tiens à remercier Jean-Marc Boucher, directeur de cette thèse, pour l'attention qu'il a portée au déroulement de celle-ci malgré la centaine de kilomètres entre Brest et Lannion.

Merci à Madeleine Bonnet et Phillip Regalia pour l'intérêt dont ils ont témoigné pour ces travaux en acceptant d'en être les rapporteurs.

Merci à Gérard Faucon et Régine Le Bouquin-Jeannès pour leur participation à l'évaluation de cette thèse.

Ces travaux de thèse ont bénéficié du cadre du laboratoire Interactions par la Parole et le Son de France Télécom R&D : merci à Jean-Pierre Petit, responsable de ce laboratoire, pour son accueil, ainsi qu'à Dominique Massaloux, qui m'a accueilli au sein de son équipe et soutenu dans mes travaux.

Je suis profondément reconnaissant aux collègues des laboratoires IPS et EQS des conseils et coups de main divers qu'ils m'ont apportés. Merci notamment à Claude Marro et Pascal Scalart pour leurs lumières en traitement du signal, à Alain Le Guyader pour les longues et enrichissantes discussions sur le masquage du bruit, à Lætitia Gros pour son aide précieuse dans la préparation et l'analyse des tests subjectifs (les sciences "molles" sont parfois bien coriaces), à Martine Apperry pour la mise en œuvre de ces tests, à Delphine Charlet pour ses conseils sur la classification et à Alain Curti, dont la maîtrise du DSP aura rendu moins terrible la confrontation de ces algorithmes à la réalité du réseau téléphonique. Merci enfin à Janine, qui sait effacer d'un sourire tous les tracés administratifs.

Merci à celles et ceux qui ont consacré un peu de leur temps et de leurs oreilles à l'évaluation subjective de mes algorithmes, malgré un emploi du temps parfois chargé.

Ce mémoire doit beaucoup à l'œil vigilant et expert de mes relecteurs : merci à André, Claude, Emmanuelle et Lætitia d'avoir fait de ce document leur livre de chevet.

Merci enfin à toutes celles et tous ceux, présents physiquement ou électroniquement, qui m'ont accompagné dans et hors de la thèse durant ces trois ans.

Résumé

Ces travaux ont pour objet la correction des distorsions spectrales subies par la parole sur les réseaux téléphoniques, en premier lieu le réseau fixe (terrestre) dans sa partie analogique. Ces distorsions sont dues aux fonctions de transfert des terminaux téléphoniques en émission et en réception, et aux lignes téléphoniques analogiques correspondantes. Le but est de restaurer, en aveugle, un "timbre" le plus proche possible de la voix originale du locuteur, au moyen d'un traitement du signal centralisé dans un équipement du réseau.

Nous proposons un algorithme d'égalisation spectrale aveugle consistant à aligner, sur une bande de fréquences limitée (200-3150 Hz), le spectre à long terme du signal traité sur un spectre de référence (spectre de la recommandation P.50 de l'UIT-T). Des évaluations subjectives mettent en évidence une restauration satisfaisante du timbre original des locuteurs, dans la limite de la bande d'égalisation choisie.

Il apparaît toutefois que la quantification en loi A des échantillons de sortie de l'égaliseur induit un bruit gênant en réception. Deux approches sont donc proposées pour masquer perceptivement ce bruit par un reformage spectral. L'une est fondée sur la réinjection à l'entrée du quantificateur de l'erreur de quantification filtrée. L'autre explore selon un algorithme de type Viterbi les séquences temporelles des niveaux de quantification possibles, de manière à maximiser un critère probabiliste de masquage du bruit. Une évaluation subjective montre finalement d'une part que le bruit non reformé est préféré au bruit reformé, plus sporadique mais plus "rauque", d'autre part qu'une voix dont le timbre a été corrigé, au prix de ce bruit de quantification, est préférée à la même voix en réception d'une liaison téléphonique sans correction de timbre (et non bruitée).

Afin d'améliorer l'adéquation du spectre de référence de l'égaliseur aux différents locuteurs, une classification des locuteurs selon leur spectre, en deux ou quatre classes, est étudiée, et des critères de classement robustes aux distorsions de la liaison téléphonique sont définis. Cette classification permet d'utiliser non plus un spectre de référence unique, mais un spectre de référence par classe. Il en résulte une réduction de la distorsion spectrale induite par l'égaliseur, ce qui se traduit, pour certains locuteurs, par une amélioration significative de la correction de timbre.

Abstract

The aim of this thesis is to compensate for spectral distortions of voice on telephone networks, particularly on the analog parts of the terrestrial network. These distortions are generated by the non-flat transfer functions of the sending and receiving terminals and of the corresponding analog lines. Our purpose is to restore a "timbre" as close as possible to the original voice of the speaker, using a blind equalizer centralized in the network.

We propose a spectral equalization algorithm, which consists in matching the long-term spectrum of the processed signal to a reference spectrum (spectrum of the ITU-T Recommendation P.50) in a limited frequency bandwidth (200-3150 Hz). Subjective evaluations show a satisfying restoration of the timbre of the speakers, within the limits of the chosen equalization band.

The A-law quantization of the output samples of the equalizer induces however a disturbing noise at the reception end. Two methods are proposed to mask this noise, using a perceptual spectral shaping. The first one is based on the feed-back of the filtered quantization error to the input of the quantizer. The second one explores the temporal sequences of the possible quantization levels, in order to maximize a probabilistic criterion of noise masking, using a Viterbi-like algorithm. A subjective evaluation finally shows on the one hand that the non-reshaped noise is preferred to the reshaped noise, on the other hand that voices with a corrected timbre, even with quantization noise, are preferred to the same voices at the output of a telephone link without timbre correction (and without noise).

In order to make the reference spectrum more appropriate to the various speakers' voices, we define two or four classes of speakers, based on their long-term spectra. Classification criteria robust to telephone link distortions are defined. This classification allows using one reference spectrum for each class, instead of the same reference spectrum for the whole population. This leads to a decrease of the spectral distortion induced by the equalizer and, as a consequence, to a significant improvement of the timbre correction for a part of the speakers, in a perceptual point of view.

Table des matières

INTRODUCTION.....	1
CHAPITRE I CONTEXTE ET OBJECTIFS.....	3
I.1. Sources de dégradation du timbre de la parole sur les réseaux téléphoniques.....	3
I.1.1. Réseau Téléphonique Commuté (RTC, ie réseau filaire classique).....	3
I.1.2. Réseau Numérique à Intégration de Services (RNIS) et réseau mobile GSM	6
I.2. Objectifs de la correction de timbre.....	7
CHAPITRE II ÉGALISATION SPECTRALE AVEUGLE.....	9
II.1. État de l'art	9
II.1.1. Égalisation fixe	9
II.1.2. Égalisation adaptative	9
II.2. L'égalisation adaptée : principes et mise en œuvre.....	13
II.2.1. Approche retenue	13
II.2.2. Principes.....	14
II.2.3. Bande de fréquences d'égalisation	17
II.2.4. Nécessité d'une pré-égalisation	17
II.2.5. Mise en œuvre.....	19
II.2.6. Mise en œuvre en temps réel	22
II.3. Simulations et résultats.....	25
II.3.1. Conditions expérimentales.....	25
II.3.2. Outils d'évaluation	26
II.3.3. Rapidité de convergence de l'égaliseur	27
II.3.4. Distorsion spectrale finale	29
II.3.5 Limite de l'égalisation : le bruit de quantification	33
II.3.6. Évaluation subjective	38
II.3.7. Validation de la version temps réel.....	52
II.4. Conclusion.....	53

CHAPITRE III ÉGALISATION ET BRUIT DE QUANTIFICATION : APPROCHES PERCEPTIVES.....	55
III.1. Principes du masquage du bruit et application au codage	55
III.1.1. Le masquage fréquentiel du bruit	55
III.1.2. Calcul du seuil de masquage : méthode de Johnston	56
III.1.3. Application au masquage du bruit de quantification.....	58
III.2. Méthode de réinjection de l'erreur de quantification	60
III.2.1. Principe.....	60
III.2.2. Structure du filtre de boucle	61
III.2.2. Résultats	62
III.3. Méthode probabiliste	64
III.3.1. Principes	64
III.3.2. Mise en œuvre	65
III.3.3. Résultats	68
III.3.4. Influence des paramètres de l'algorithme	69
III.4. Comparaison des deux méthodes.....	75
III.4.1. Complexité	75
III.4.2. Performances de masquage	75
III.5. Évaluation de la perception conjointe du bruit et du timbre.....	76
III.5.1. Objectifs et méthode.....	76
III.5.2. Plan de test	78
III.5.3. Résultats	79
III.6. Conclusion.....	85
CHAPITRE IV ÉGALISATION DIFFERENCIÉE PAR CLASSES DE LOCUTEURS. 87	
IV.1. Classification des locuteurs.....	87
IV.1.1. Corpus.....	87
IV.1.2. Définition de l'individu : le cepstre partiel.....	88
IV.1.3. Classification hiérarchique ascendante [Lebart, 2000a]	88
IV.1.4. Algorithme de classification	90
IV.1.5. Agrégation selon le critère du saut minimal	90
IV.1.6. Agrégation selon le critère de Ward généralisé	91
IV.1.7. Consolidation de la partition.....	92
IV.2. Classement des locuteurs	94
IV.2.1. Stratégie de classement	94
IV.2.2. Calcul des fonctions linéaires discriminantes	95
IV.2.3. Affectation d'une nouvelle observation	96
IV.2.4. Application au classement en deux classes hommes / femmes	97
IV.2.5. Application au classement en quatre classes.....	103

IV.3. Égalisation adaptée multiréférences	106
IV.3.1. Mise en œuvre dans le domaine des cepstres partiels	106
IV.3.2. Application à la classification hommes / femmes	107
IV.3.3. Application à la classification en quatre classes	112
IV.4. Conclusion	117
CONCLUSION	119
Annexe A : Consignes du test d'évaluation de l'égaliseur	121
Annexe B : Résultats des Tukey tests du chapitre II	125
Annexe C : Principes du masquage fréquentiel	127
Annexe D : Consignes du test de comparaison par paires	131
Annexe E : Consignes du test de comparaison de dégradations	133
annexe F : Évaluation du bruit de quantification	135
Annexe G : Significativité de l'écart entre deux pourcentages	137
annexe H : Construction d'une échelle de Thurstone	139
Annexe J : Calcul des fonctions linéaires discriminantes	141
Références bibliographiques	145

Introduction

De nombreuses méthodes ont été développées jusqu'à présent pour corriger les dégradations les plus critiques de la parole téléphonique : bruit [Davis, 2002], écho [Gritton, 1984][Naylor, 1994] et, dans une moindre mesure, niveaux non optimaux [Mahé, 1998].

L'écho électrique [Gritton, 1984], provenant de la désadaptation d'impédance des jonctions 2 fils - 4 fils des liaisons téléphoniques filaires, devient perceptible sur les liaisons longue distance, lorsque le délai de propagation aller-retour dépasse 30 ms. L'écho acoustique [Gilloire, 1994] résulte de la transmission acoustique du signal de réception du haut-parleur vers le microphone : transmission solidoportée par la coque du terminal, notamment lorsque le microphone et le haut-parleur sont proches, transmission aérienne lors de l'utilisation du téléphone en mode mains libres.

Au-delà du bruit de codage, peu perceptible sur les liaisons classiques, le bruit perçu dans une communication téléphonique résulte principalement de l'utilisation du terminal d'émission dans des conditions dégradées : mode main-libres et / ou milieu ambiant bruyant, notamment dans le cas de terminaux mobiles.

Des spécifications strictes sur les équipements garantissaient autrefois un niveau sonore satisfaisant pour toutes les communications [UIT-T/G.121, 1993]. Le nouveau contexte de dérégulation rend toutefois plus délicat le contrôle du niveau de manière réglementaire. La multiplication des réseaux et l'interconnexion entre réseaux d'opérateurs concurrents, ainsi que la diversité croissante des types de terminaux téléphoniques, conduisent ainsi à une augmentation de la disparité des niveaux de parole sur les réseaux. A cette diversité des matériels s'ajoute celle du niveau de la voix des locuteurs à l'émission, qu'elle soit d'origine physiologique ou qu'elle résulte de la variété des conditions d'émission : environnement calme ou bruyant, combiné classique ou mode mains-libres sont autant de sources de disparité de niveaux. Ainsi est apparue la nécessité d'un contrôle automatique de niveau.

Ces trois types de dégradation (écho, bruit et niveaux non optimaux) font l'objet de traitements correctifs pour deux raisons. D'une part, ces dégradations peuvent perturber le fonctionnement des équipements du réseau : le bruit et l'écho, modifiant les propriétés du signal de parole, altèrent le fonctionnement des codeurs et décodeurs utilisant ces propriétés ; les disparités de niveau empêchent d'exploiter correctement la dynamique des équipements. D'autre part, ces dégradations peuvent nuire de manière critique à la qualité de la communication [Gilloire, 1994], rendant la parole inaudible, saturée (selon le niveau), ou incompréhensible (bruit et / ou écho).

Le timbre de la parole téléphonique est également dégradé, du fait des distorsions spectrales introduites par les parties analogiques des liaisons : outre la limitation de la bande passante qui prive la voix de ses harmoniques d'ordre élevé et de ses basses fréquences, la voix manque de présence et semble parfois étouffée. Cette dégradation est cependant peu traitée, sans doute parce qu'elle est moins critique que les précédentes. L'altération du timbre ne perturbe pas le fonctionnement des équipements, ne nuit pas à l'intelligibilité de la communication et peut même être un moyen de l'améliorer. Il en est ainsi du débruitage, qui procède à une atténuation du signal dans les bandes de fréquences où le signal parasite est trop fort.

La correction du timbre de la parole téléphonique apparaît donc comme un traitement "de confort". Ceci implique d'une part que son action ne doit pas entraver celle des traitements prioritaires que sont le débruitage et l'annulation d'écho. D'autre part, elle doit améliorer la qualité vocale de manière sensible, sans ajouter de défauts tels qu'une modification du niveau ou des bruits supplémentaires.

Une première approche visant à associer un algorithme de correction de timbre à une fonction de correction de niveau a été proposée dans [Mahé, 1998]. Dans ce travail préliminaire, nous avons montré que les deux traitements peuvent cohabiter sans que l'action de l'un ne perturbe celle de l'autre. Une étude de la combinaison de cet algorithme avec un traitement de réduction du bruit [Lanoë, 1999] a donné lieu à des premiers résultats encourageants sur la capacité des deux traitements à conjuguer leurs effets de manière satisfaisante.

Au cours des travaux de thèse décrits dans le présent mémoire, nous nous sommes essentiellement attachés à étudier en profondeur et sous ses principaux aspects le problème de la correction des distorsions spectrales subies par le signal de parole sur un réseau téléphonique. Cette correction vise, par un dispositif centralisé dans le réseau, à restaurer le timbre de la voix perçue en réception d'une liaison téléphonique.

Le mémoire est organisé comme suit. Les types de distorsions spectrales à corriger sont recensés dans le chapitre I, ce qui permet de préciser les objectifs que nous assignons à notre étude de la correction de timbre. Compte tenu de ces distorsions, un algorithme d'égalisation spectrale est proposé dans le chapitre II et évalué à la fois par des mesures objectives et subjectivement, à l'aune de l'objectif de restauration du timbre de la voix originale. L'algorithme présenté dans le chapitre II se révélant efficace mais source d'artefacts audibles (bruit de quantification) et mal adapté à la diversité des locuteurs, nous avons consacré une part importante de notre étude au traitement de ces défauts ; les travaux correspondants sont décrits dans les chapitres III et IV. Dans le chapitre III, nous proposons une approche perceptive du bruit induit par l'égalisation, et tentons de remédier à ce bruit en le masquant. Les travaux présentés dans le chapitre IV visent à affiner la correction de timbre en effectuant une correction différenciée par classes de locuteurs.

Les compléments utiles à la compréhension de l'exposé figurent dans les annexes du mémoire.

Chapitre I

Contexte et objectifs

Ce chapitre étudie la nature et la place dans le réseau des distorsions spectrales à l'origine des dégradations du timbre de la parole sur les réseaux téléphoniques, ce qui permet de préciser les objectifs que nous assignons à notre étude de la correction de timbre.

I.1. Sources de dégradation du timbre de la parole sur les réseaux téléphoniques

I.1.1. Réseau Téléphonique Commuté (RTC, ie réseau filaire classique)

La Figure 1.1 présente une liaison RTC schématisée : chaque correspondant est relié par une ligne analogique (paire torsadée) au central téléphonique le plus proche, et la liaison entre les centraux emprunte un réseau entièrement numérique. Nous considérerons que la transmission numérique est sans erreur et que les distorsions spectrales proviennent uniquement des éléments de transmission analogique du signal de parole en bande de base. Dans ces conditions, le spectre de la voix est affecté par deux types de distorsions.

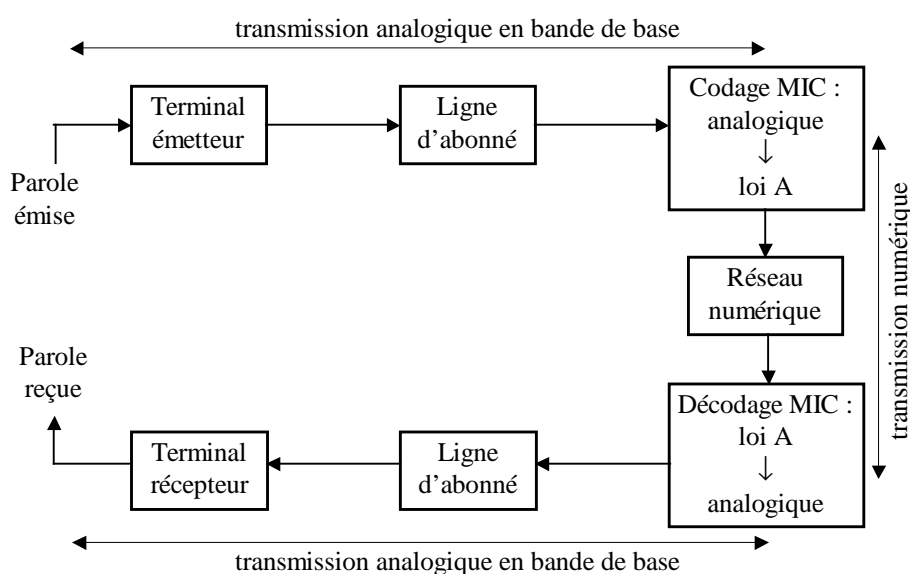


Figure 1.1 : Liaison téléphonique schématisée sur le RTC

Le premier type de distorsion est le filtrage passe-bande des terminaux et des points d'accès à la partie numérique du réseau. Les caractéristiques typiques de ce filtrage sont décrites par l'UIT-T sous le nom de "système de référence intermédiaire" (SRI) [UIT-T/P.48, 1988], système dont la réponse nominale en fréquence et le gabarit sont représentés sur la Figure 1.2 pour la partie émission et sur la Figure 1.3 pour la partie réception. Nous appellerons respectivement, selon la terminologie de l'UIT-T, "système émetteur" et "système récepteur" les parties émission et réception du SRI.

Ces caractéristiques fréquentielles, issues de mesures réalisées dans les années 70, tendent cependant à devenir obsolètes. D'une part, elles reflètent les liaisons longue distance intégralement analogiques qui existaient alors. Dans le cadre du multiplexage analogique, limiter la bande passante à la fois dans les hautes et les basses fréquences permettait d'accroître la capacité des porteuses. Dès lors que le signal est transmis sous forme numérique, une aussi forte atténuation des basses fréquences perd son intérêt. D'autre part, une partie des terminaux de l'époque utilisaient encore des microphones à charbon, peu efficaces dans les basses fréquences ; les terminaux actuels atténuent moins fortement celles-ci.

C'est pourquoi l'UIT-T préconise depuis 1996 d'utiliser un SRI "modifié" [UIT-T/P.830, 1996], dont la caractéristique nominale est représentée sur la Figure 1.4 pour la partie émission, et sur la Figure 1.5 pour la partie réception. Entre 200 et 3400 Hz, la tolérance est de $\pm 2,5$ dB ; en dessous de 200 Hz, la décroissance de la caractéristique du système global doit être d'au moins 15 dB par octave.

La seconde distorsion affectant le spectre de la voix est l'atténuation des lignes d'abonné. Dans un modèle simple de la ligne analogique locale [Cadoret, 1983], on considère que celle-ci introduit un affaiblissement du signal dont la valeur en dB dépend de sa longueur et est proportionnelle à la racine carrée de la fréquence. L'affaiblissement est de 3 dB à 800 Hz pour une ligne moyenne (environ 2 km), de 9,5 dB à 800 Hz pour les lignes les plus longues (jusqu'à 10 km). Selon ce modèle, l'affaiblissement d'une ligne, représenté sur la Figure 1.6, a pour expression :

$$A_{dB}(f) = A_{dB}(800\text{Hz}) \cdot \sqrt{\frac{f}{800}} \quad (1.1)$$

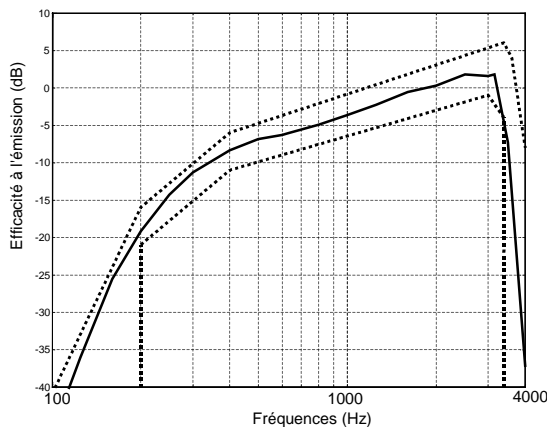


Figure 1.2 : Gabarit (pointillés) et caractéristique nominale (trait plein) du SRI en émission

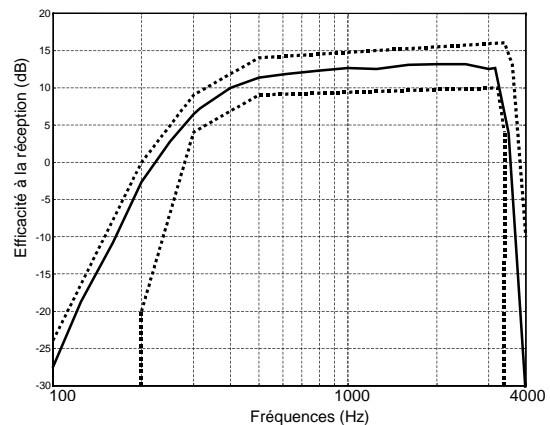


Figure 1.3 : Gabarit (pointillés) et caractéristique nominale (trait plein) du SRI en réception

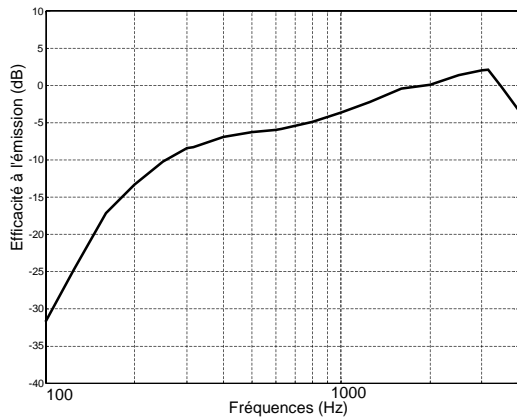


Figure 1.4 : Réponse fréquentielle en émission du SRI modifié

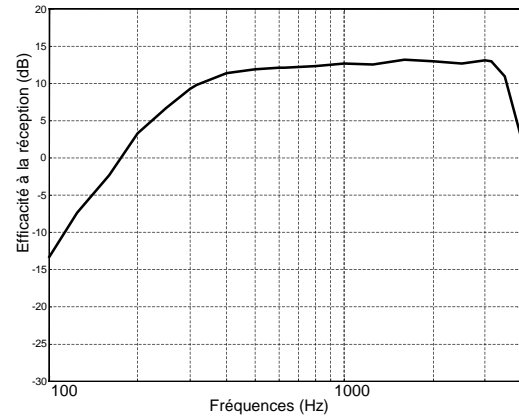


Figure 1.5 : Réponse fréquentielle en réception du SRI modifié

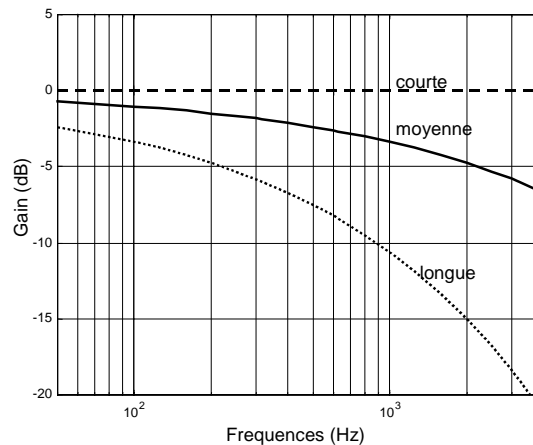


Figure 1.6 : Réponses en fréquence des lignes d'abonnés selon leur longueur

A ces distorsions s'ajoute le filtrage anti-repliement du codeur MIC. Nous considérerons un filtrage selon le gabarit des cofidec utilisés sur le réseau France Télécom [National Semiconductor, 1994], représenté sur la Figure 1.7. C'est un filtre passe-bande 200-3400 Hz avec une réponse presque plate sur la bande passante. La coupure des basses fréquences vise à éliminer la composante continue du signal et les signaux parasites à 50 Hz résultant de l'alimentation électrique.

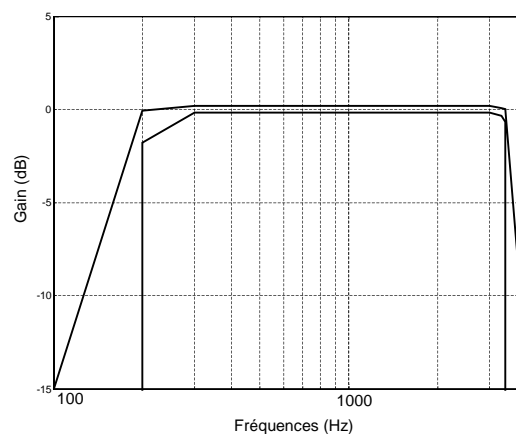


Figure 1.7 : Gabarit du filtre anti-repliement du codeur MIC

Au final, la voix subit une distorsion spectrale telle que représentée sur la Figure 1.8 pour les différentes combinaisons de trois types de ligne analogique en émission et en réception (soit 6 distorsions), sous l'hypothèse d'équipements respectant la caractéristique nominale du SRI modifié. La voix apparaît ainsi étouffée si une des lignes analogiques est longue et souffre dans tous les cas d'un manque de "présence" dû à l'affaiblissement des composantes basse fréquence.

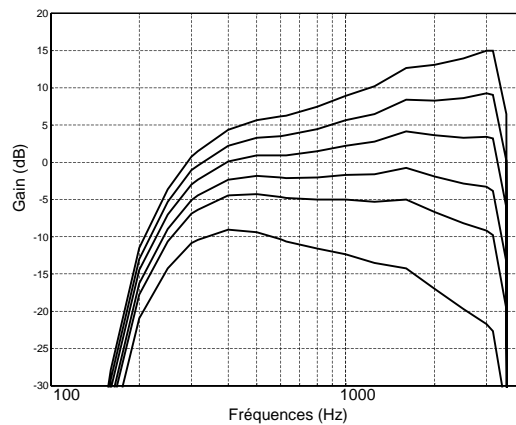


Figure 1.8 : Distorsions spectrales subies par la parole sur le RTC avec un SRI moyen et différentes combinaisons de lignes analogiques

1.1.2. Réseau Numérique à Intégration de Services (RNIS) et réseau mobile GSM

Dans le RNIS et le réseau GSM, le signal est numérisé dès le terminal. Les seules parties analogiques sont les transducteurs en émission et en réception associés à leurs chaînes d'amplification et de conditionnement respectives. L'UIT-T a défini des gabarits d'efficacité en fréquence à l'émission (Figure 1.9) et à la réception (Figure 1.10), valables à la fois pour les téléphones numériques filaires [UIT-T/P.310, 2000] et les terminaux numériques mobiles ou sans fil [UIT-T/P.313, 2000]. Ces gabarits, bien que moins contraignants que ceux du SRI, sont toutefois peu respectés par les constructeurs, comme l'ont montré les mesures de l'Observatoire des Mobiles de France Télécom R&D.

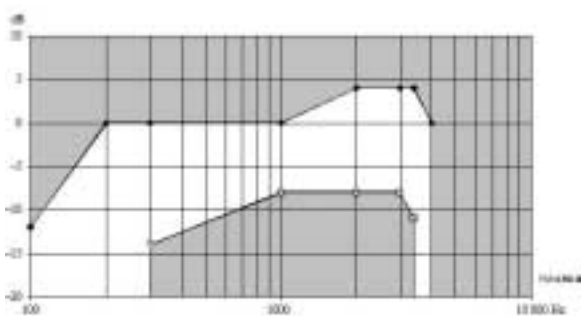


Figure 1.9 : Gabarit à l'émission pour les terminaux numériques

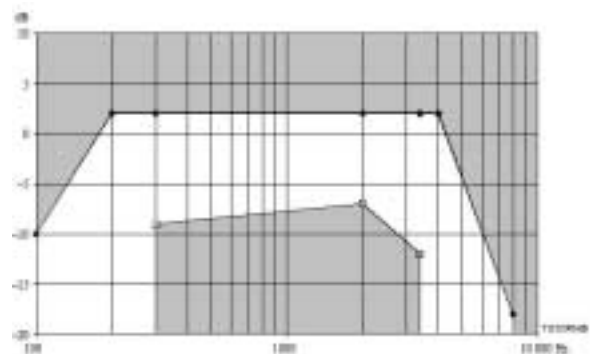


Figure 1.10 : Gabarit à la réception pour les terminaux numériques

Nous considérerons que la transmission numérique est sans erreur et n'introduit pas de distorsion spectrale. La partie numérique de la liaison n'est toutefois pas exempte de distorsion pour les réseaux GSM : le codage et le décodage modifient légèrement l'enveloppe spectrale du signal.

Cette altération est représentée sur la Figure 1.11 pour un bruit rose codé puis décodé en mode EFR (*Enhanced Full Rate*) [Com. int., 2002].

L'effet de ces filtrages sur le timbre est principalement un affaiblissement des composantes basse fréquence, moins marqué cependant que dans le cas du RTC.

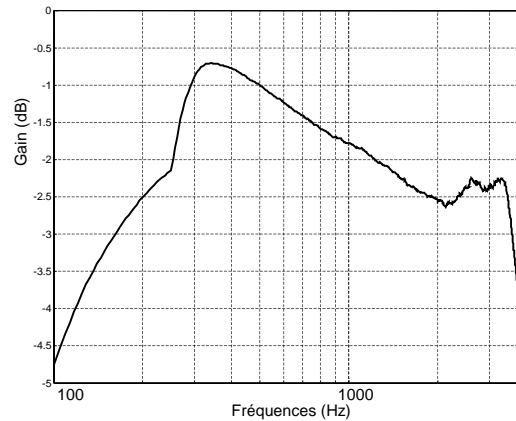


Figure 1.11 : Distorsion spectrale introduite par le codage-décodage GSM en mode EFR

I.2. Objectifs de la correction de timbre

Précisons que, par la suite, nous utiliserons le terme de "timbre" sans aborder toute la richesse contenue dans cette notion. Au vu des distorsions présentées ci-dessus, il s'agira essentiellement de corriger ces modifications de l'enveloppe spectrale du signal, de manière à améliorer le naturel et la présence de la voix. L'objectif est que le timbre de la voix en réception soit le plus proche possible de celui de la voix émise.

Les traitements sont réalisés "en aveugle", c'est-à-dire qu'*a priori* aucune information sur le signal original, sur le locuteur ou sur les caractéristiques exactes des équipements à l'origine des distorsions spectrales n'est disponible. La seule hypothèse est la connaissance du type de liaison : nous nous plaçons essentiellement dans le cas d'une liaison empruntant uniquement le RTC, pour lequel les distorsions sont les plus fortes, et, partant, la correction la plus intéressante.

La correction de timbre sera centralisée dans le réseau, c'est-à-dire effectuée au cœur du réseau numérique indiqué sur la Figure 1.1. L'intérêt d'une telle position est multiple :

- elle permet une allocation dynamique des ressources de calcul aux communications, ce qui est plus économique qu'un traitement dédié à chaque liaison ;
- elle est déjà utilisée par les annuleurs d'écho électrique [UIT-T/G.168, 1999] et les autres dispositifs de rehaussement de la parole [UIT-T/G.VED, 2002], ce qui permet de limiter les besoins matériels de l'implantation ;
- pour un opérateur, elle permet de maîtriser la qualité du service fourni. La conception des terminaux échappe en effet largement au contrôle des opérateurs (si ce n'est par le biais des normalisations qui imposent des spécifications plancher), alors qu'elle influe sur la qualité perçue par les abonnés. Effectuer les traitements dans le réseau permet à

l'opérateur d'offrir la qualité vocale souhaitée indépendamment des choix techniques des constructeurs de terminaux. Les améliorations de qualité peuvent ainsi être déployées très rapidement vers le public le plus large, n'étant pas soumises à la vitesse de renouvellement des terminaux.

Il est souhaitable que les traitements soient compatibles avec les traitements des dégradations de la parole existants – débruitage, annulation d'écho, correction de niveau. Une première approche du problème, dans le cadre d'un projet de France Télécom R&D de plate-forme centralisée combinant les différents traitements correctifs de la parole, a fait apparaître que la résolution du problème de la correction du timbre en aveugle nécessite en soi une exploration approfondie. Cela nous a conduit à l'étudier isolément, sans toutefois perdre de vue la nécessaire liaison avec les autres traitements des dégradations. Cette approche justifie une fois de plus le choix prioritaire de l'hypothèse d'une liaison sur le RTC : ce réseau étant le moins sujet à des utilisations dans des conditions dégradées, les traitements autres que la correction de timbre sont généralement inutiles, laissant le champ libre à un traitement exclusivement tourné vers la restauration du timbre original des locuteurs.

Chapitre II

Égalisation spectrale aveugle

Après une étude des dispositifs d'égalisation existants visant à corriger les distorsions spectrales évoquées au chapitre I, le présent chapitre propose une nouvelle méthode d'égalisation spectrale aveugle, l'*égalité adaptée*. Nous avons étudié cette méthode par simulations et nous l'avons évaluée par des mesures objectives de distorsion spectrale, par des tests subjectifs formels et par une mise en œuvre en temps réel.

II.1. État de l'art

La compensation des distorsions spectrales introduites dans le signal de parole par les divers éléments de la liaison téléphonique est réalisée par des dispositifs à base d'égalisation. Celle-ci peut être fixe ou s'adapter en fonction des conditions de transmission.

II.1.1. Égalisation fixe

Des dispositifs d'égalisation centralisée ont été proposés dans [Bowker, 1993] et [Ho, 1993]. Ces égaliseurs sont des filtres fixes qui amplifient les basses fréquences atténuées par l'émetteur. Bowker propose par exemple un gain de 10 à 15 dB sur la bande 100-300 Hz [Bowker, 1993]. Ces dispositifs, tels qu'ils sont mis en œuvre, présentent plusieurs inconvénients :

- L'égaliseur ne compense que le filtrage de l'émetteur, de sorte qu'à la réception, les composantes basse-fréquence restent fortement affaiblies par le filtrage de réception modélisé par le SRI.
- La non-adaptabilité de l'égalisation ne peut pas permettre une correction satisfaisante dans tous les cas. Si les conditions réelles de transmission sont trop différentes de celles corrigées par le dispositif, celui-ci peut amplifier insuffisamment, ou au contraire exagérément, les composantes basse fréquence. D'autre part, l'égaliseur laisse subsister d'éventuelles autres altérations du timbre dues à des modifications des parties moyennes et hautes fréquences du spectre de la voix par la liaison analogique.

II.1.2. Égalisation adaptative

D'autres dispositifs permettent d'égaliser le signal de parole de manière adaptative, pour améliorer soit la qualité vocale, soit les performances de systèmes de reconnaissance

automatique de la parole. Le principe général de ces égaliseurs est de rapprocher le spectre de la parole traitée d'un spectre de référence.

- **Blanchiment – reformage**

Le dispositif décrit dans [De Jaco, 1997] vise à corriger la réponse fréquentielle non idéale d'un transducteur de téléphone mobile. L'égaliseur est décrit comme étant placé entre le convertisseur analogique-numérique et le codeur CELP, mais peut être situé aussi bien dans le réseau que dans le terminal. L'une des méthodes consiste à blanchir le signal puis le reformer spectralement selon un spectre cible pré-défini.

Les coefficients du filtre blanchisseur sont actualisés à chaque trame selon la procédure suivante. La première étape est le calcul de "*coefficients d'autocorrélation à long terme*" R_{LT} :

$$R_{LT}(n, i) = \alpha R_{LT}(n-1, i) + (1-\alpha) R(n, i), \quad (2.1)$$

avec $R_{LT}(n, i)$ $i^{\text{ème}}$ coefficient d'autocorrélation à long terme à la $n^{\text{ème}}$ trame, $R(n, i)$ $i^{\text{ème}}$ coefficient d'autocorrélation (à court terme) spécifique à la $n^{\text{ème}}$ trame, et α constante de lissage fixée par exemple à 0,995, ce qui correspond à une constante de temps de 10 s pour une fréquence d'échantillonnage de 8 kHz. De ces coefficients sont déduits, selon l'algorithme de Levinson, les "*coefficients LPC à long terme*", qui sont les coefficients du filtre blanchisseur. C'est donc le spectre à long terme du signal qui est blanchi par ce premier filtre.

À la sortie de ce filtre, le signal est filtré par un filtre fixe qui lui imprime les caractéristiques spectrales à long terme "*idéales*", *i.e.* celles qu'il aurait à la sortie d'un transducteur ayant la réponse fréquentielle "*idéale*". Ces deux filtres sont complétés par un gain multiplicatif égal au rapport entre les énergies à long terme de l'entrée du blanchisseur et de la sortie du deuxième filtre.

L'intérêt de cette méthode est d'exploiter les coefficients d'autocorrélation (à court terme), qui sont déjà calculés dans le codeur. Cet intérêt disparaît dans notre cas, où la méthode devra être généralisable à tous types de réseaux et adaptée de préférence au RTC.

- **Adaptation du niveau par sous-bande**

Une autre méthode consiste à diviser le signal en sous-bandes et, pour chaque sous-bande, appliquer un gain multiplicatif de manière à atteindre une énergie cible, ce qui revient à considérer le spectre de référence comme une distribution d'énergies de sous-bande.

Dans la réalisation présentée dans [De Jaco, 1997], le signal est filtré par un banc de filtres. On calcule l'énergie à long terme E_i de chaque $i^{\text{ème}}$ sortie de celui-ci par un lissage de l'énergie à court terme s_i^2 du signal de sous-bande, selon l'équation (2.2) :

$$E_i(n) = \alpha E_i(n-1) + (1-\alpha) s_i^2(n), \quad (2.2)$$

où α est une constante de lissage correspondant à une constante de temps de 10 s. Le gain à appliquer dans la sous-bande est alors défini comme le rapport entre l'énergie cible de la sous-bande et l'énergie à long terme ainsi calculée.

Dans le cadre de l'amélioration de la robustesse des systèmes de reconnaissance vocale, une autre réalisation a été étudiée par C. Mokbel *et al* [Mokbel, 1996]. Les performances de la reconnaissance de la parole à travers le réseau téléphonique sont en effet sensiblement dégradées

par le filtrage de la ligne téléphonique, ce qui a conduit au développement de plusieurs techniques de compensation de l'effet de celle-ci. L'une de ces méthodes est l'égalisation adaptative aveugle dans le domaine spectral.

La reconnaissance de parole utilise typiquement les coefficients cepstraux calculés selon l'échelle des fréquences MEL, les Mel Frequency Cepstral Coefficients (MFCC). Ceux-ci sont obtenus selon les étapes suivantes :

- transformée de Fourier rapide de la trame courante du signal ;
- regroupement des raies spectrales en 24 bandes critiques [Zwicker, 1981] réparties selon une échelle perceptuelle de fréquences MEL et calcul de l'énergie $V_y(i)$ de chaque $i^{\text{ème}}$ bande critique ;
- transformée de Fourier inverse du logarithme du vecteur V_y .

L'effet convolutif de la liaison téléphonique se traduit par une translation des vecteurs cepstraux. Cette translation, variable suivant les appels, réduit la capacité de discrimination dans l'espace cepstral.

L'égalisation adaptative est réalisée dans le domaine spectral suivant le schéma de la Figure 2.1. L'énergie $V_y(i)$ de chaque bande critique i est multipliée par un gain adaptatif $W(i)$ pour donner l'énergie de bande critique égalisée $V_n(i)$. Le gain est adapté selon l'algorithme du gradient stochastique, par minimisation de l'erreur quadratique moyenne, l'erreur $E(i)$ étant définie comme la différence entre $V_n(i)$ et une énergie de référence $R(i)$ définie pour chaque sous-bande.

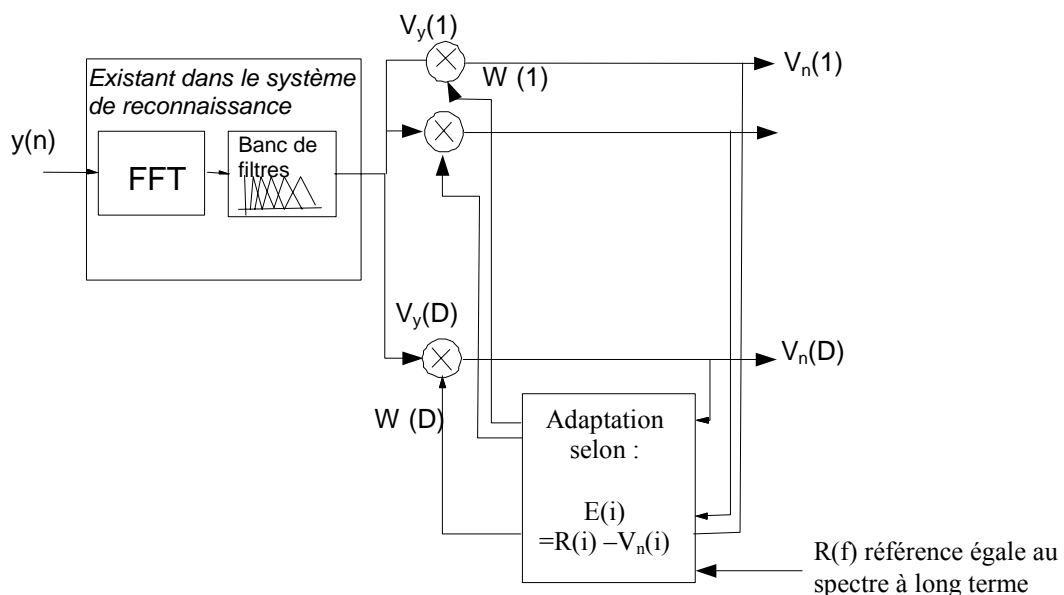


Figure 2.1 : Égalisation aveugle dans le domaine spectral

Différents algorithmes du gradient stochastique ont été évalués [Mauuary, 1996] pour la mise en œuvre de l'adaptation. Le plus performant, permettant une convergence à la même vitesse sur toutes les sous-bandes, est l'algorithme du gradient normalisé, qui s'écrit, pour chaque sous-bande i :

$$W_{n+1}(i) = W_n(i) + \mu \left(\frac{R(i)}{V_y(i)} - W_n(i) \right) \quad (2.3)$$

où n est l'indice de trame et μ est le pas d'adaptation.

Quelques précautions doivent être prises pour assurer la convergence :

- l'énergie du spectre de référence doit être modulée par celle de la trame, de manière à respecter les variations naturelles du niveau à court terme de la parole ;
- le pas d'adaptation choisi doit être assez grand pour permettre une convergence rapide et assez petit pour ne pas perturber les variations locales du spectre liées au son prononcé.

Le filtre adaptatif converge en 2 s environ et permet une nette amélioration des performances en reconnaissance de la parole [Mokbel, 1996].

• Soustraction cepstrale

La déviation des MFCC peut être également corrigée par la méthode de soustraction cepstrale [Mokbel, 1993, 1996]. Si l'on pose $s(t)$ le signal de parole original, $n(t)$ le bruit de fond à l'émission, $h(t)$ la fonction de transfert du canal téléphonique (considéré comme stationnaire), $p(t)$ celle du filtre de préaccentuation qui précède le système de reconnaissance et $y(t)$ le signal reçu, on a :

$$y(t) = (s(t) + n(t)) * h(t) * p(t). \quad (2.4)$$

Le filtre de préaccentuation est un filtre RIF passe-haut d'ordre 1 compensant la pente de -6 dB / octave du spectre moyen de la parole. En notant $x(t)$ le signal original préaccentué :

$$x(t) = s(t) * p(t), \quad (2.5)$$

l'équation (2.4) s'exprime :

$$y(t) = (x(t) + n(t)) * h(t). \quad (2.6)$$

Le rapport signal à bruit étant supposé élevé, on néglige l'influence du bruit. En appelant C_x , C_h et C_y les cepstres respectifs de x , h et y , l'équation (2.6) se traduit dans le domaine cepstral par :

$$C_y(\tau) = C_x(\tau) + C_h(\tau). \quad (2.7)$$

Sur la Figure 2.2, les vecteurs acoustiques de trois appels différents ont été projetés sur le plan des deux premiers coefficients cepstraux, de même que les cepstres moyens de ces appels [Mokbel, 1993]. Il apparaît que la translation des vecteurs cepstraux de chaque appel correspond au vecteur du cepstre moyen de l'appel. Cette expérience montre que le cepstre moyen du signal reçu préaccentué constitue une bonne estimation du cepstre du canal.

Ce résultat est à l'origine de la méthode de soustraction cepstrale : en notant $\hat{C}_x(\tau)$ le cepstre estimé du signal de parole original préaccentué et $\overline{C}_y(\tau)$ la moyenne temporelle de C_y ,

$$\hat{C}_x(\tau) = C_y(\tau) - \overline{C}_y(\tau). \quad (2.8)$$

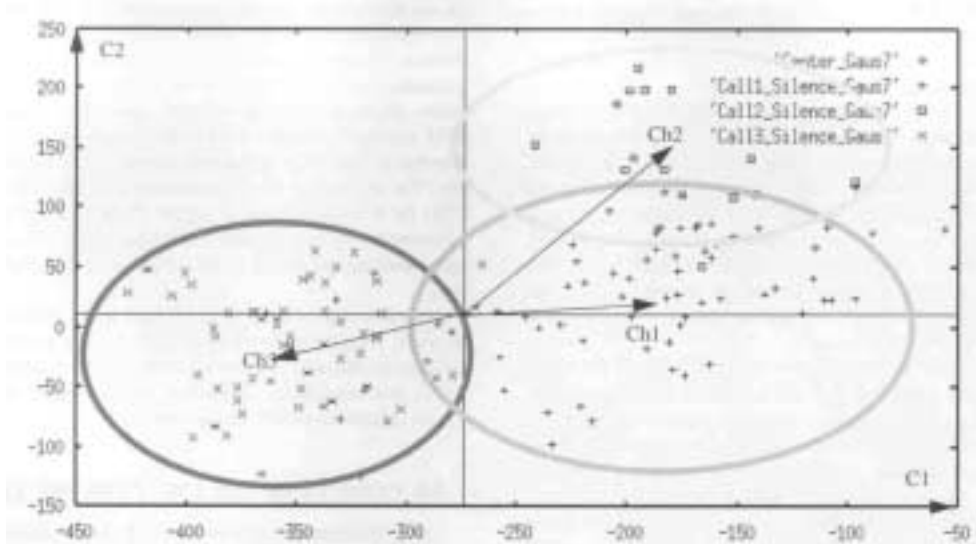


Figure 2.2 : Projection des vecteurs acoustiques de trois appels différents et des cepstres moyens de ces appels sur le plan des deux premiers coefficients cepstraux [Mokbel, 1993]

Cette méthode simple permet une nette amélioration des performances en reconnaissance de la parole. Les résultats expérimentaux sont même légèrement supérieurs à ceux de l'égalisation adaptative. Elle est cependant considérée comme une méthode « *off-line* », puisque 2 à 4 secondes de parole sont nécessaires pour estimer le cepstre du canal. Il est à noter que les performances en reconnaissance sont meilleures quand :

- le cepstre du canal est estimé uniquement sur de la parole (sans qu'une estimation sur un mélange parole+silence ne soit rédhitoire) ;
- on soustrait au cepstre uniquement les 6 premiers coefficients cepstraux moyens, les coefficients d'indice élevé étant dépendants du locuteur.

II.2. L'égalisation adaptée : principes et mise en œuvre

II.2.1. Approche retenue

Dans l'optique de l'association de l'égaliseur à un dispositif de réduction de bruit, il peut être avantageux de privilégier une méthode calculant l'égaliseur dans le domaine fréquentiel, puisque la réponse fréquentielle des débruiteurs que nous envisageons d'utiliser (méthode d'Ephraïm et Mallat notamment [Cappé, 1994]) est calculée à partir des transformées de Fourier des trames successives de signal.

A cet égard, nous ne retenons pas la méthode de blanchiment-reformage : l'utilisation des coefficients d'autocorrélation dans le calcul du filtre blanchisseur, avantageuse dans le cas d'un codeur CELP (qui utilise ces coefficients pour le calcul des coefficients LPC), perd son intérêt dans le cas d'un codeur MIC, où ces coefficients devraient être calculés spécialement pour l'égalisation.

De même, la correction du gain par sous-bande en utilisant les sorties temporelles d'un banc de filtres, envisageable dans le cas d'un égaliseur isolé, n'est pas retenue. Si une correction par sous-bande doit être appliquée, le regroupement des raies spectrales du signal, qui sont éventuellement déjà disponibles si l'on réalise un débruitage conjoint, est *a priori* plus avantageux en termes de complexité.

La méthode d'égalisation adaptative aveugle dans le domaine spectral pourrait être utilisée. Son application est cependant délicate : la modulation du spectre de référence par l'énergie de trame et le choix du pas d'adaptation nécessitent un réglage fin, sous peine de dégrader sensiblement l'égalisation.

Au vu des bons résultats et de la simplicité de la méthode de soustraction cepstrale, c'est finalement de cette méthode que nous nous sommes inspirés, en en transposant le principe dans le domaine fréquentiel.

II.2.2. Principes

La chaîne de traitement envisagée dans la méthode de soustraction cepstrale en reconnaissance vocale est représentée sur la Figure 2.3. Selon cette méthode, le cepstre du canal h peut être estimé par :

$$\hat{C}_h(\tau) = \overline{C_r(\tau)}, \quad (2.9)$$

où $\overline{C_r(\tau)}$ est la moyenne temporelle du cepstre du signal r reçu par le système de reconnaissance (après pré-accentuation). Dans le domaine fréquentiel, ce résultat se traduit par :

$$|\hat{H}(f)|^2 = \overline{|R(f)|^2}^g \quad (2.10)$$

où \hat{H} est la réponse fréquentielle estimée du canal et $|R(f)|^2$ est le spectre à court terme de r , \bar{a}^g désignant pour une variable a sa moyenne temporelle géométrique.

Dans le cas où aucun filtre de pré-accentuation ne serait utilisé, l'estimation du canal devrait se fonder sur la sortie r' du canal. L'équation (2.10) deviendrait alors :

$$|\hat{H}(f)|^2 = |P(f)|^2 \overline{|R'(f)|^2}^g, \quad (2.11)$$

où $P(f)$ est la réponse fréquentielle du filtre de pré-accentuation p et $|R'(f)|^2$ est le spectre à court terme de r' .

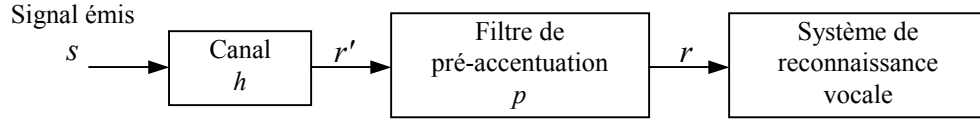


Figure 2.3 : Reconnaissance vocale en réception d'une liaison téléphonique

Généralisons cette méthode à une liaison téléphonique complète telle que représentée sur la Figure 1.1 et, de manière plus schématique, sur la Figure 2.4. La réponse fréquentielle globale $G(f)$ du canal analogique (système d'émission, système de réception et lignes analogiques) peut ainsi être estimée par :

$$|\hat{G}(f)|^2 = |P(f)|^2 \overline{|Y(f)|^2}^g, \quad (2.12)$$

avec $|Y(f)|^2$ le spectre à court terme du signal de réception y .

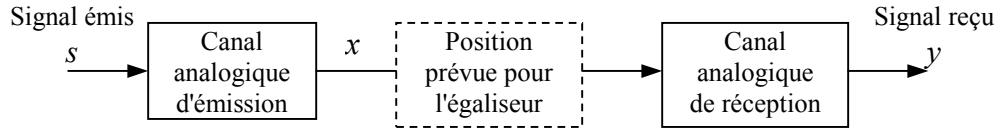


Figure 2.4 : Filtrages d'une liaison téléphonique et place de l'égaliseur

Ce résultat peut également être trouvé simplement de la manière suivante :

$$|Y(f)|^2 = |G(f)|^2 |S(f)|^2, \quad (2.13)$$

où $|S(f)|$ est le spectre à court terme du signal émis s . Si le canal analogique est supposé invariant dans le temps,

$$\overline{|Y(f)|^2}^{(g)} = |G(f)|^2 \overline{|S(f)|^2}^{(g)}, \quad (2.14)$$

la moyenne temporelle pouvant être soit arithmétique soit géométrique. Sous l'hypothèse

$$\overline{|S(f)|^2}^{(g)} \approx \frac{1}{|P(f)|^2}, \quad (2.15)$$

on retrouve bien le résultat (2.12). Le spectre moyen original du locuteur courant n'étant pas connu, une approximation de ce type est nécessaire. Cette approximation par l'inverse de la réponse du filtre de pré-accentuation est toutefois très grossière. Pour limiter l'erreur de l'approximation de la moyenne du spectre du signal d'émission, nous approcherons celle-ci par le spectre moyen de la parole défini par l'UIT [UIT-T/P.50, 1998]. Ce spectre moyen a été calculé à partir de mesures sur un grand nombre d'échantillons de parole prononcés par différents locuteurs dans 20 langues.

Nous appellerons ce spectre *spectre de référence* et le noterons $\chi_{\text{ref}}(f)$. Par ailleurs, nous appellerons désormais *spectre à long terme* d'un signal de parole x , noté $\chi(f)$, la moyenne

temporelle (arithmétique) de son spectre à court terme $|X(f)|^2$. Ainsi, la réponse fréquentielle de l'égaliseur compensant le canal analogique G est définie par :

$$|EQ(f)| = \sqrt{\frac{\gamma_{\text{ref}}(f)}{\gamma_y(f)}} \quad (2.16)$$

Cette formule est valable pour un égaliseur placé en réception, après le transducteur électro-acoustique, ce qui d'une part est irréaliste et d'autre part ne correspond pas au cas de figure envisagé ici, où l'égaliseur doit être centralisé dans le réseau comme indiqué sur la Figure 2.5. La grandeur γ_y est donc en fait le spectre à long terme du signal de réception s'il n'y avait pas d'égaliseur dans le réseau. Cette valeur n'étant pas directement accessible, on l'exprime en fonction de γ_x , spectre à long terme de l'entrée x de l'égaliseur :

$$\gamma_y(f) = |L_{RX}(f)|^2 |S_{RX}(f)|^2 \gamma_x(f), \quad (2.17)$$

avec L_{RX} la réponse fréquentielle de la ligne de réception et S_{RX} la réponse fréquentielle du système de réception. Ainsi,

$$|EQ(f)| = \frac{1}{|S_{RX}(f) \cdot L_{RX}(f)|} \sqrt{\frac{\gamma_{\text{ref}}(f)}{\gamma_x(f)}} \quad (2.18)$$

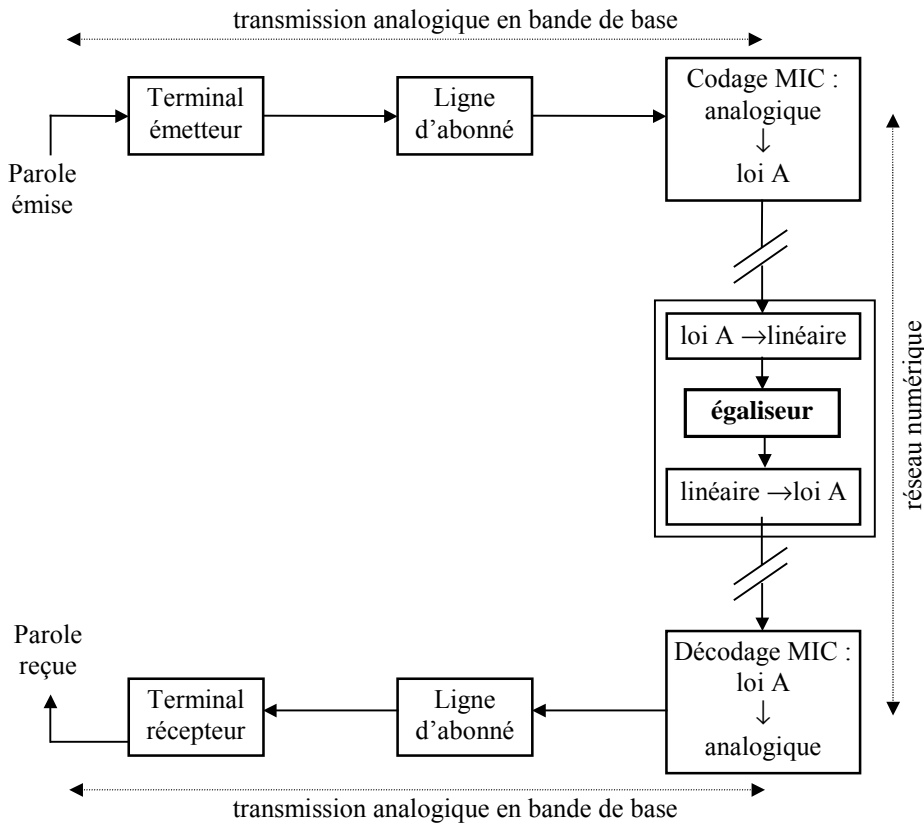


Figure 2.5 : Position de l'égaliseur dans le réseau

Les grandeurs L_{RX} et S_{RX} sont inconnues *a priori*. On peut les remplacer par des valeurs moyennes, auquel cas on perd cependant l'avantage du caractère adaptatif de l'égaliseur. On peut aussi imaginer que, l'égaliseur étant destiné à améliorer le confort d'écoute de l'abonné de réception, les caractéristiques de sa ligne et de son terminal ont été préalablement mesurées, de sorte que dans ce cas L_{RX} et S_{RX} sont connues. L'égalisation n'est donc aveugle que pour la partie de la liaison en amont de l'égaliseur. Quelle que soit la solution retenue, **nous ferons désormais l'hypothèse d'une ligne de réception moyenne et d'un système de réception respectant la caractéristique nominale du SRI modifié.**

Nous appellerons cet égaliseur *égaliseur adapté*, en ce qu'il s'adapte automatiquement au canal de transmission. L'adaptation étant fondée sur la simple estimation de la moyenne du spectre du signal traité, nous utilisons ce terme plutôt que le qualificatif "adaptatif", qui désigne habituellement une correction utilisant une boucle de contrôle rétroactif.

II.2.3. Bande de fréquences d'égalisation

L'atténuation des composantes du signal en dehors de la bande 200-3400 Hz par le système d'émission et par le filtre anti-repliement du codeur MIC est telle que le rapport signal à bruit de quantification est faible pour ces composantes. Ainsi, la méthode de rehaussement présentée n'est pas envisageable en deçà de 200 Hz et au-delà de 3400 Hz : elle conduirait à une amplification du bruit haute et basse fréquence.

Par ailleurs, comme la pente de la caractéristique des systèmes d'émission et de réception peut être forte entre 200 et 300 Hz ainsi qu'entre 3150 et 3400 Hz, la réponse fréquentielle de l'égaliseur devra *a priori* avoir une pente raide dans ces bandes. De manière à limiter le nombre de coefficients du filtre, nous ne chercherons pas à compenser l'atténuation de 3150 à 3400 Hz, bande de fréquence dont la restauration est moins critique, d'un point de vue perceptif, que celle des basses fréquences.

Au final, l'égaliseur corrigera les distorsions spectrales du canal analogique sur la bande $[F_c-3150 \text{ Hz}]$, avec F_c une fréquence de coupure basse comprise entre 200 et 300 Hz, dont le choix sera discuté dans la section II.2.5.

II.2.4. Nécessité d'une pré-égalisation

Dans l'exposé des principes de l'égaliseur adapté, nous avons fait l'hypothèse que, pour tous les locuteurs, le spectre à long terme du signal original est identique au spectre moyen défini par l'UIT. En réalité, l'allure générale du spectre à long terme d'un signal de parole pour un locuteur quelconque est proche de celle du spectre de référence, mais elle est bien moins lisse. La réponse fréquentielle de l'égaliseur adapté comportera donc une erreur correspondant à l'écart entre ces deux spectres exprimés en dB. La Figure 2.6 représente cette erreur pour un locuteur masculin et une locutrice. Pour chaque locuteur, le spectre à long terme a été calculé comme la moyenne, sur un texte lu d'une durée totale d'activité vocale de 20 s environ, des spectres de puissance des trames successives d'activité vocale, chaque trame représentant 32 ms et recouvrant la précédente de 50 %.

Une telle erreur spectrale se traduit par une nette dégradation du timbre (effet de tonneau). Toutefois, l'allure générale de la courbe d'erreur est plate. Par conséquent, **en lissant fortement la réponse fréquentielle de l'égaliseur donnée par l'équation (2.18), il doit être possible d'obtenir une courbe d'erreur spectrale presque plate, correspondant à une distorsion non**

perceptible. Ce lissage correspond à la limitation de la soustraction cepstrale aux premiers coefficients dans [Mokbel, 1993].

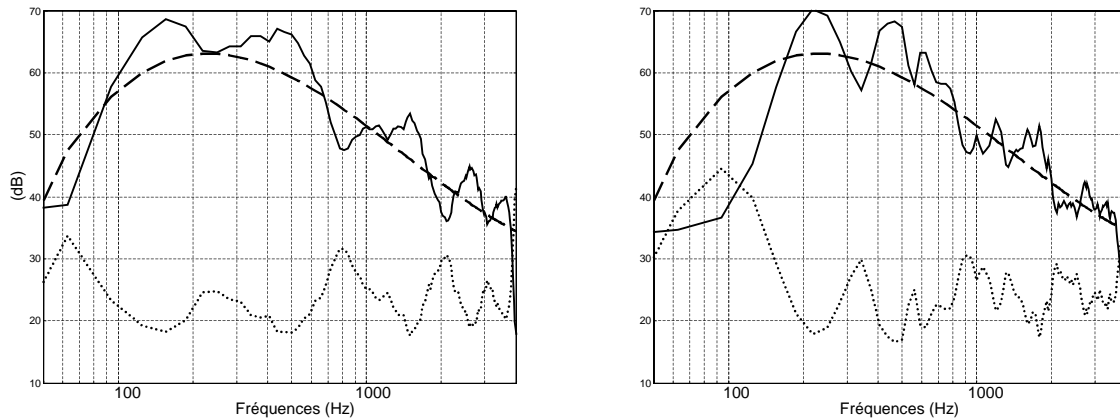


Figure 2.6 : Spectres à long terme d'un locuteur (à gauche, trait plein) et d'une locutrice (à droite, trait plein) vs spectre moyen UIT (tirets). Erreur d'approximation du spectre du locuteur par le spectre moyen de l'UIT (pointillés)

Le lissage est cependant contradictoire avec la raideur du filtre nécessaire à la compensation des distorsions de la liaison téléphonique. La Figure 2.7 représente la réponse fréquentielle, sur la bande 200-3150 Hz, que devrait avoir un égaliseur corrigeant sur la bande 200-3150 Hz une liaison téléphonique moyenne. Nous définissons comme "moyenne" une liaison dont la partie analogique est composée de systèmes d'émission et de réception conformes aux caractéristiques nominales du SRI modifié [UIT-T/P.830, 1996], ainsi que de deux lignes d'abonné moyennes (cf. Figure 1.6). Un lissage assez fort pour atténuer des fluctuations telles que celles de la Figure 2.6 est incompatible avec la raideur de la réponse fréquentielle de la Figure 2.7 dans les basses fréquences.

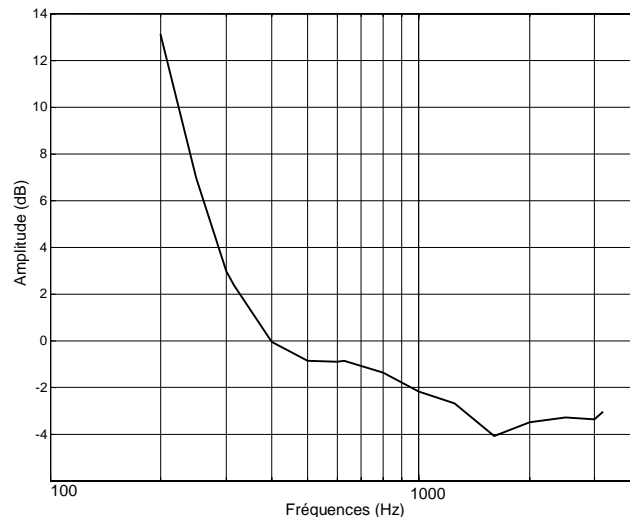


Figure 2.7 : Réponse fréquentielle idéale d'un égaliseur adapté à une liaison téléphonique moyenne

Cette contradiction est résolue de la manière suivante. L'UIT ayant défini un gabarit dont la largeur n'excède pas 5 dB sur la bande 200-3150 Hz, le filtrage SRI est approximativement connu sur cette bande, en admettant que les terminaux respectent ce gabarit. D'autre part, la réponse en fréquence des lignes analogiques fluctue autour de celle d'une ligne moyenne. Par

conséquent, il n'est pas nécessaire de compenser le filtrage du canal analogique entièrement par une égalisation aveugle : **nous proposons donc d'effectuer d'abord une pré-égalisation fixe inversant à la fois les caractéristiques nominales du SRI modifié et celles de deux lignes d'abonné moyennes, puis de compléter l'égalisation de manière adaptée selon les principes décrits dans la section II.2.2. Le rôle de l'égaliseur adapté consiste ainsi à corriger la désadaptation entre le pré-égaliseur fixe et les conditions réelles de transmission, ce qui implique une réponse en fréquence adoucie, compatible avec le lissage évoqué ci-dessus.** Notons que les équations (2.16) à (2.18) restent valables, en notant x la sortie du pré-égaliseur. Nous désignerons désormais sous le terme *égaliseur* (respectivement *égalisation*) la combinaison du pré-égaliseur et de l'égaliseur adapté (respectivement de la *pré-égalisation* et de l'*égalisation adaptée*).

II.2.5. Mise en œuvre

Le fonctionnement de la double structure pré-égaliseur / égaliseur adapté que nous proposons est décrit ci-après et schématisé sur la Figure 2.9.

- **Pré-égaliseur**

Le pré-égaliseur est un filtre fixe, dont la réponse fréquentielle, sur la bande $[F_c-3150 \text{ Hz}]$, est l'inverse de la réponse globale de la partie analogique du canal moyen défini en II.2.4. La raideur de la réponse fréquentielle de ce filtre implique une réponse impulsionnelle longue ; c'est pourquoi, de manière à limiter le retard introduit par le traitement, le pré-égaliseur est réalisé sous forme d'un filtre RII d'ordre 20 par la méthode de Yule-Walker. La Figure 2.8 représente la réponse fréquentielle du pré-égaliseur pour trois valeurs de F_c . La dispersion des retards de groupe est inférieure à 2 ms, de sorte que la distorsion de phase résultante n'est pas perceptible.

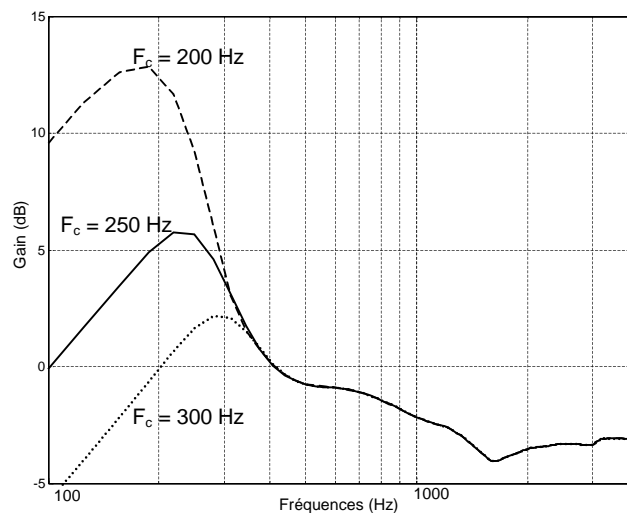


Figure 2.8 : Réponse fréquentielle du pré-égaliseur pour différentes F_c .

- **Égaliseur adapté**

La sortie x du pré-égaliseur est analysée par trames de 32 ms, avec un recouvrement inter-trames de 50 %. L'égaliseur adapté est un filtre RIF dont les coefficients sont adaptés à chaque trame d'activité vocale selon l'équation (2.18), comme décrit ci-après et représenté sur la Figure 2.9.

D'après [Mokbel, 1993], 2 à 4 secondes d'activité vocale sont nécessaires pour estimer le canal. Par conséquent, le spectre à long terme de x , γ_x , est d'abord calculé (à partir de l'instant initial de fonctionnement) sur une fenêtre temporelle croissant de 0 à 4 s d'activité vocale, puis ajusté récursivement à chaque trame d'activité vocale, ce qui se traduit par la formule générique suivante :

$$\gamma_x(f, n) = \alpha(n) |X(f, n)|^2 + (1 - \alpha(n)) \gamma_x(f, n-1), \quad (2.19)$$

où $\gamma_x(f, n)$ est le spectre à long terme de x à la $n^{\text{ème}}$ trame d'activité vocale, $X(f, n)$ la transformée de Fourier de la $n^{\text{ème}}$ trame d'activité vocale, et $\alpha(n)$ est défini par l'équation (2.20). En notant N le nombre de trames dans 4 s,

$$\alpha(n) = \frac{1}{\min(n, N)}. \quad (2.20)$$

La réponse fréquentielle de l'égaliseur est alors calculée selon l'équation (2.18) pour les fréquences comprises entre F_c et 3150 Hz. Comme le pré-égaliseur n'effectue aucune compensation de l'affaiblissement introduit par la liaison en dehors de cette bande, appliquer l'équation (2.18) en deçà de F_c et au delà de 3150 Hz reviendrait à faire réaliser cette compensation par l'égaliseur adapté, avec des valeurs de $|EQ|$ très élevées en dehors de ces limites, alors que nous avons choisi de restreindre l'égalisation à la bande F_c -3150 Hz. C'est pourquoi les valeurs de $|EQ|$ hors de cette bande de fréquences sont calculées par extrapolation linéaire de la valeur en dB de $|EQ|_{[F_c-3150 \text{ Hz}]}$, notée EQ_{dB} par la suite, de la manière suivante. Pour chaque indice de fréquence k , l'approximation linéaire de EQ_{dB} s'exprime par :

$$\tilde{EQ}_{\text{dB}}(k) = a_1 + a_2 k \quad (2.21)$$

Les coefficients a_1 et a_2 sont choisis de manière à minimiser l'erreur quadratique de l'approximation sur l'intervalle F_c -3150 Hz, définie par

$$e = \sum_{k=k_1}^{k_2} \left(EQ_{\text{dB}}(k) - \tilde{EQ}_{\text{dB}}(k) \right)^2 \quad (2.22)$$

où k_1 et k_2 sont les indices de fréquence correspondant respectivement à F_c et 3150 Hz. Les coefficients a_1 et a_2 sont donc définis par :

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} k_2 - k_1 + 1 & \sum_{k=k_1}^{k_2} k \\ \sum_{k=k_1}^{k_2} k & \sum_{k=k_1}^{k_2} k^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=k_1}^{k_2} EQ_{\text{dB}}(k) \\ \sum_{k=k_1}^{k_2} k EQ_{\text{dB}}(k) \end{pmatrix} \quad (2.23)$$

Les valeurs de $|EQ|$, en dB, hors de la bande F_c -3150 Hz, sont alors calculées à partir de la formule (2.21).

La caractéristique en fréquence ainsi obtenue doit être lissée. Comme le filtrage doit être réalisé dans le domaine temporel, le moyen le plus simple est de multiplier par une fenêtre étroite la réponse impulsionnelle correspondante. Celle-ci est obtenue par une IFFT de $|EQ|$ suivie d'une

symétrisation, de manière à obtenir un filtre causal à phase linéaire. La fenêtre utilisée est typiquement une fenêtre de Hamming de longueur 15 centrée sur le pic de la réponse impulsionnelle.

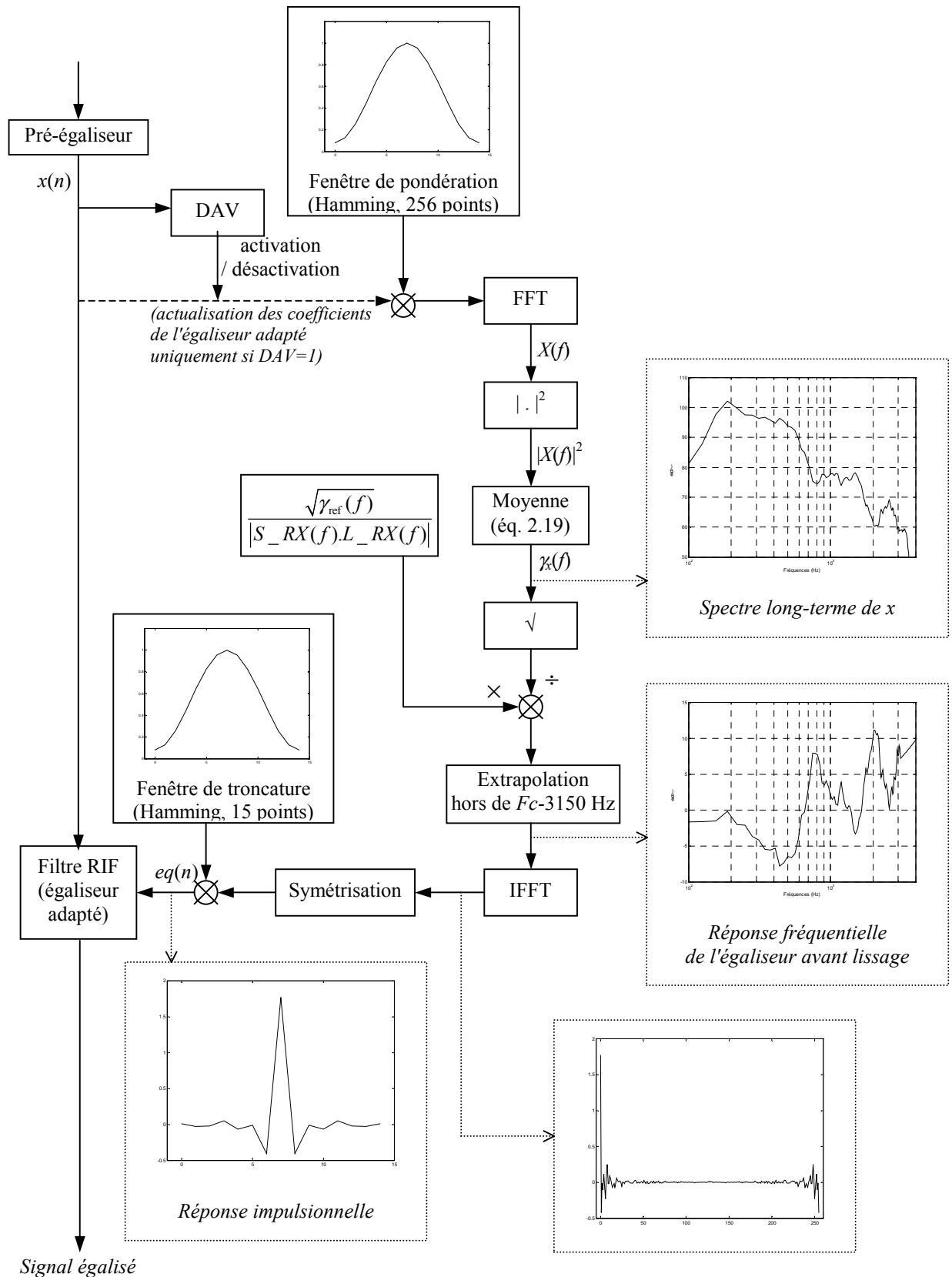


Figure 2.9 : Schéma de principe de l'égaliseur adapté

II.2.6. Mise en œuvre en temps réel

Nous présentons ici l'implantation en temps réel de l'égaliseur décrit ci-dessus dans le cadre d'une plate-forme expérimentale de traitements de la parole centralisés dans le réseau téléphonique, "Mainate" (Machine A Intégrer de Nouveaux Algorithmes de Traitement du signal en Exploitation). Cette plate-forme, développée à France Télécom R&D, intègre divers traitements : correction de niveau, correction de timbre, réduction de bruit et annulation d'écho.

- **Architecture matérielle**

Comme illustré sur la Figure 2.10, la plate-forme est installée sur un PC relié par une liaison RNIS 30 voies à l'autocommutateur expérimental du laboratoire, lui-même relié à celui du site lannionnais de France Télécom R&D. L'établissement d'une communication utilisant les fonctionnalités de *Mainate* se fait en appelant la plate-forme, qui établit alors entre l'appelant et le correspondant demandé une liaison passant par elle. Les traitements sont effectués par le processeur de traitement du signal (DSP).

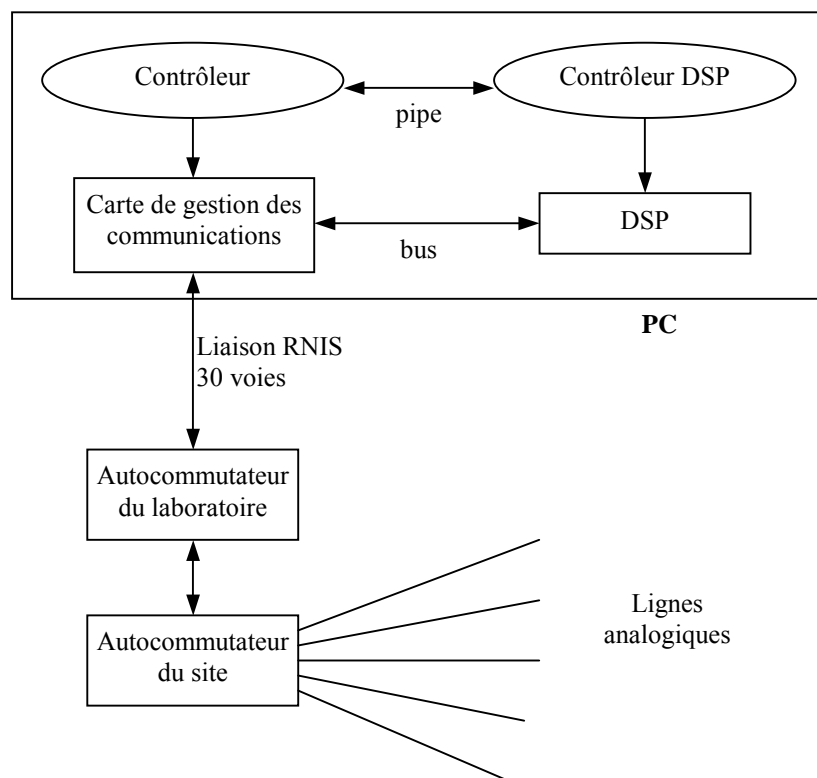


Figure 2.10 : Architecture matérielle de la plate-forme "Mainate"

- **Intégration algorithmique des traitements**

Au moment de l'intégration de la fonction de correction de timbre à la plate-forme, seule la fonction de réduction de bruit était implantée. Nous avons donc étudié la combinaison du débruitage et de l'égalisation. La fonction de réduction de bruit est implantée sous forme d'un filtre RIF à 65 coefficients, dont la réponse fréquentielle est calculée selon les principes du filtrage de Wiener [Scalart, 2001]. La réponse impulsionnelle est calculée à partir de cette réponse fréquentielle selon la même méthode que celle utilisée pour notre égaliseur adapté

(TFD inverse, symétrisation de la réponse impulsionnelle obtenue, puis troncature par une fenêtre de longueur 65 centrée sur le pic de la réponse impulsionnelle).

La première combinaison étudiée a été une mise en cascade du pré-égaliseur, de l'égaliseur adapté et du filtre de débruitage, selon le schéma de la Figure 2.11.

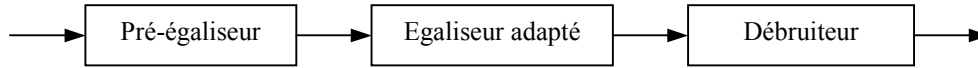


Figure 2.11 : Première combinaison de traitements

Le signal en réception d'une liaison simulée avec cette combinaison de traitements au centre du réseau est affecté de "clics", qui correspondent à des discontinuités de la forme d'onde lors de changements de trame d'analyse dans l'algorithme de débruitage. Ces discontinuités apparaissent également lorsque l'égaliseur n'est pas présent, mais elles sont moins marquées et non perceptibles.

Elles s'expliquent par ce que la réponse du filtre de débruitage varie assez rapidement d'une trame à la suivante. Si l'on applique à un son pur les gains successifs du débruiteur à la fréquence de ce son, les discontinuités sont d'autant plus perceptibles que le son est grave. Or le pré-égaliseur suramplifie les composantes basse fréquence de la parole, de manière à compenser de manière anticipée leur atténuation en réception. Cette prédominance des composantes basse fréquence dans le signal à l'entrée du débruiteur explique les discontinuités perceptibles.

Ce phénomène est évité en plaçant le débruiteur avant le pré-égaliseur. Ainsi, les discontinuités sont même lissées par l'effet passe-bas du pré-égaliseur.

Nous souhaitons par ailleurs que l'égaliseur adapté et le débruiteur soient accolés, en vue d'un traitement combiné. Nous plaçons donc le pré-égaliseur après l'égaliseur adapté, comme indiqué sur la Figure 2.12. L'équation (2.18) devient alors :

$$|EQ(f)| = \frac{1}{|PRE_EQ(f).S_RX(f).L_RX(f)|} \sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}}, \quad (2.24)$$

où PRE_EQ est la réponse fréquentielle du pré-égaliseur et γ_x est le spectre à long terme de l'entrée x du dispositif.

Le débruiteur et l'égaliseur adapté utilisent les mêmes fonctions d'analyse du signal et de construction de la réponse impulsionnelle du filtre. Par conséquent, nous proposons de les combiner en un filtre unique selon le schéma de la Figure 2.13. Le signal est analysé par trames de 32 ms se recouvrant de 50 %. A chaque trame,

- si une activité vocale est détectée ($DAV = 1$), la réponse impulsionnelle de l'égaliseur adapté eq est actualisée selon la procédure décrite dans la section II.2.5 ;
- la réponse impulsionnelle du débruiteur deb est calculée selon la méthode décrite dans [Scalart, 2001], la densité spectrale du bruit γ_b étant actualisée si aucune activité vocale n'est détectée ($DAV = 0$).

Les réponses eq et deb sont alors convoluées pour obtenir les coefficients du filtre de débruitage et d'égalisation adaptée. Il est à noter que selon cette procédure, l'égaliseur adapté et le débruiteur

sont calculés indépendamment l'un de l'autre et qu'aucun ne perturbe le fonctionnement de l'autre : leurs effets respectifs sont simplement superposés.

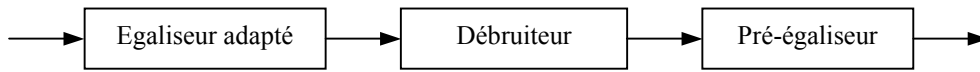


Figure 2.12 : Combinaison de traitements évitant les discontinuités du signal

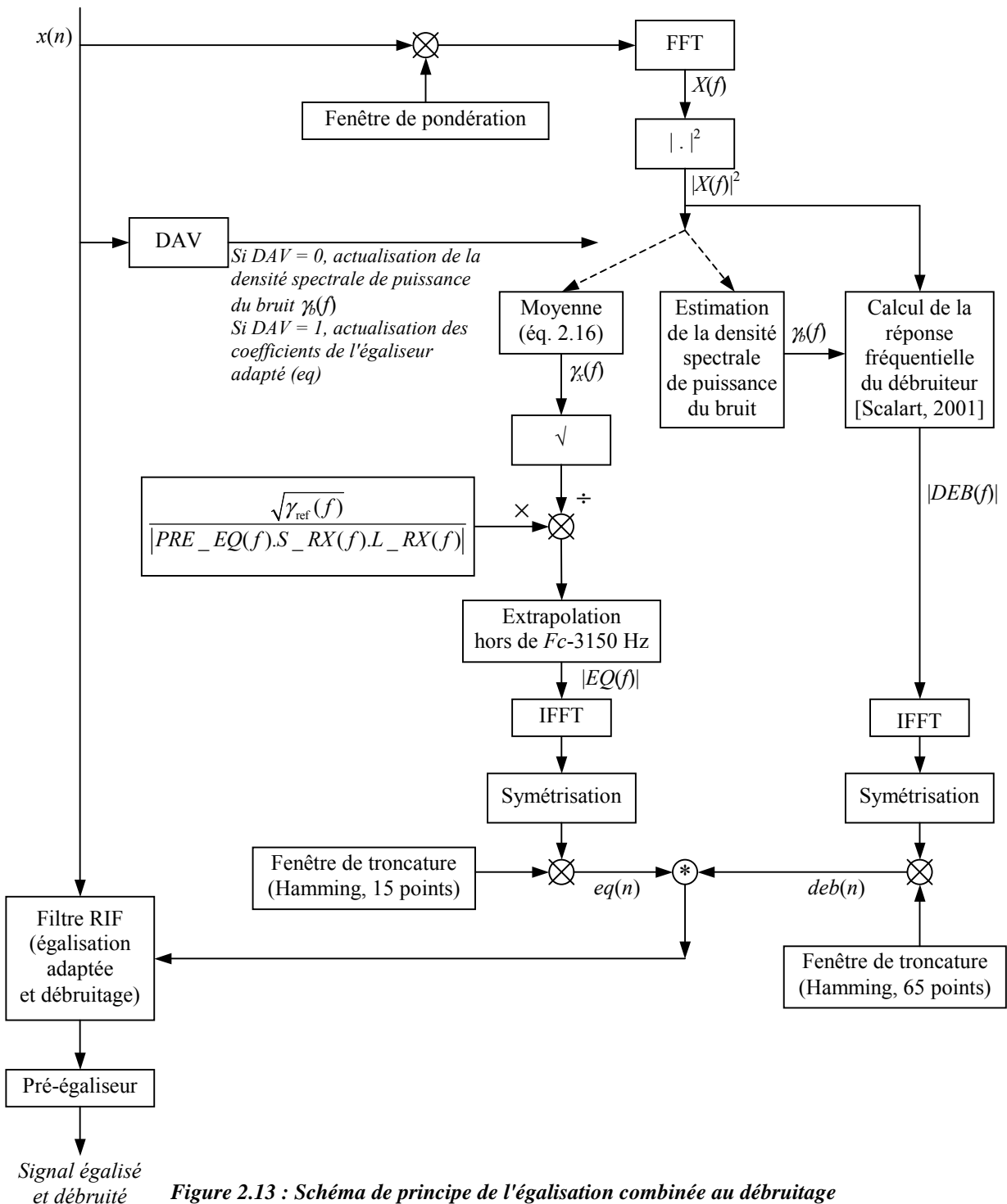


Figure 2.13 : Schéma de principe de l'égalisation combinée au débruitage

II.3. Simulations et résultats

II.3.1. Conditions expérimentales

Les signaux d'émission utilisés proviennent d'enregistrements réalisés à France Télécom R&D en studio, dans les mêmes conditions d'enregistrement que celles du corpus utilisé par l'UIT pour établir le spectre moyen de la parole [UIT-T/P.50, 1998]. Nous disposons d'un corpus de 34 locuteurs (17 hommes et 17 femmes), prononçant le même texte, représentant une vingtaine de secondes d'activité vocale : *"La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort, quand ils virent un voyageur s'avancer enveloppé dans son manteau. Ils tombèrent d'accord que celui qui arriverait le premier à lui faire enlever son manteau serait le plus fort. Alors la bise se mit à souffler de toutes ses forces ; mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui, et à la fin la bise renonça à le lui faire enlever."*

Nous considérons une liaison RTC telle que représentée sur la Figure 2.5. Le système d'émission, le système de réception et les lignes d'abonnés sont simulés par des filtres RIF réalisés selon la méthode du fenêtrage à partir de leurs réponses fréquentielles respectives. Six conditions de transmission sont simulées, correspondant aux combinaisons suivantes :

- Nous utilisons deux systèmes d'émission. Le premier respecte la caractéristique nominale du SRI modifié [UIT-T/P.830, 1996]. Le second a une réponse en fréquence, représentée sur la Figure 2.14, qui respecte le masque de la partie émission du SRI modifié, mais diffère de la caractéristique nominale de celui-ci. Cette réponse a été choisie arbitrairement de manière à tester un écart de la caractéristique réelle du système d'émission par rapport à la caractéristique nominale prise comme référence.
- Trois types de lignes analogiques d'émission sont testés : très courte, moyenne et longue (réponse fréquentielle sur la Figure 1.6).

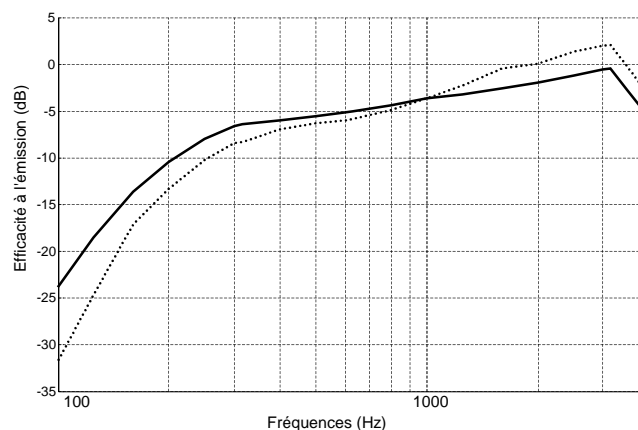


Figure 2.14 : Réponse fréquentielle du système d'émission simulé (trait plein) et caractéristique nominale de celui du SRI modifié (pointillés)

La ligne analogique de réception et le système de réception sont choisis respectivement moyenne (voir Figure. 1.6) et conforme à la caractéristique nominale en réception du SRI modifié, représentée sur la Figure 1.5. Le choix de ces éléments importe peu, puisque la caractéristique de la liaison en aval de l'égaliseur est supposée connue dans l'algorithme d'égalisation adaptée.

La liaison numérique étant considérée comme transparente, nous en simulons uniquement les interfaces avec les liaisons analogiques et avec le dispositif d'égalisation. Le filtre anti-repliement du codeur MIC est simulé par deux filtres de Butterworth d'ordre 6 – un passe-bas et un passe-haut – en cascade, dont la réponse fréquentielle globale respecte le gabarit de la figure 1.7. Nous considérons ici un codage MIC en loi A [UIT-T/G.711, 1988]. Les traitements étant réalisés sur des valeurs linéaires, l'égaliseur est précédé d'une conversion des échantillons loi A en valeurs linéaires. De même, les échantillons traités sont convertis en loi A après l'égalisation, puis reconvertis en valeurs linéaires pour simuler le décodage MIC (voir Figure 2.5).

Les caractéristiques en fréquence de ces six liaisons simulées sont représentées sur la Figure 2.15.

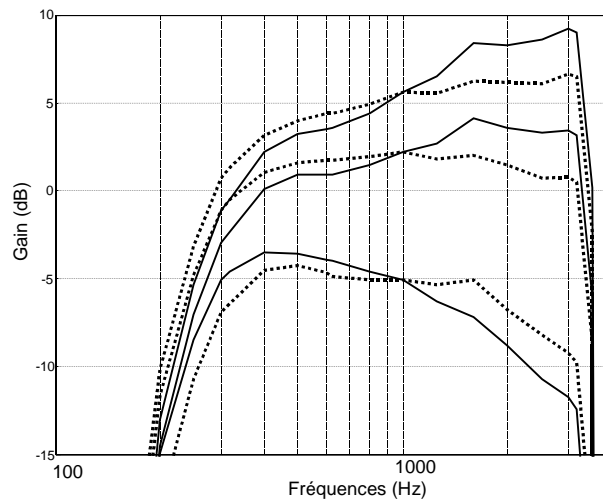


Figure 2.15 : Caractéristiques spectrales des liaisons simulées

Le dispositif d'égalisation est composé du pré-égaliseur et de l'égaliseur décrits dans la section II.2. La fréquence de coupure F_c est fixée à 200 Hz. Afin d'évaluer à la fois le dispositif complet d'égalisation et l'apport de l'égaliseur adapté, nous considérerons les trois conditions suivantes d'égalisation : aucun traitement ; pré-égaliseur seul ; égalisation complète.

II.3.2. Outils d'évaluation

L'objectif étant de corriger les distorsions spectrales introduites par le canal analogique, les performances de l'égalisation peuvent être évaluées par la donnée de la caractéristique fréquentielle de la liaison égalisée. Ainsi, l'égalisation sera d'autant meilleure que celle-ci sera constante sur l'espace des fréquences le plus large possible. En tenant compte de la nécessaire limitation de la bande d'égalisation (200-3150 Hz), une autre méthode d'évaluation consiste à comparer la réponse fréquentielle de l'égaliseur après convergence avec celle de l'*égaliseur idéal*, que nous définirons comme celui dont la réponse fréquentielle est l'inverse de celle du canal analogique sur la bande d'égalisation considérée, 200-3150 Hz. Dans les simulations, l'égaliseur idéal est constitué du même pré-égaliseur et d'un filtre complétant celui-ci de telle sorte que la réponse fréquentielle des deux en cascade soit celle souhaitée. Nous appellerons ce filtre *égaliseur adapté idéal*. L'égalisation sera ainsi d'autant meilleure que la différence, en dB, entre la réponse fréquentielle de l'égaliseur adapté et celle de l'égaliseur adapté idéal sera constante sur l'espace des fréquences et au cours du temps.

Cette proximité de forme entre l'égaliseur adapté et l'égaliseur idéal adapté nécessite d'être quantifiée, afin de comparer aisément les performances de l'égaliseur selon les conditions d'utilisation. C'est pourquoi nous introduisons une deuxième mesure : l'erreur cepstrale, ou distance cepstrale entre ces réponses fréquentielles [Faucon, 1993]. Celle-ci est définie par :

$$e = \sqrt{\sum_{i=1}^{20} (C_{\text{eq}}^i - C_{\text{eq_id}}^i)^2}, \quad (2.25)$$

où C_{eq}^i et $C_{\text{eq_id}}^i$ désignent les $i^{\text{èmes}}$ coefficients cepstraux de l'égaliseur adapté et de l'égaliseur adapté idéal, respectivement. Les premiers coefficients cepstraux, C_{eq}^0 et $C_{\text{eq_id}}^0$, ne sont pas pris en compte dans le calcul de la distance, de sorte que celle-ci reflète uniquement la différence de forme des deux réponses fréquentielles, mais pas leur différence de niveau éventuelle. Ainsi, à une différence entre les réponses fréquentielles uniforme sur l'espace des fréquences correspondra une distance cepstrale nulle. D'autre part, la distance est calculée uniquement à partir des 20 premiers coefficients cepstraux, parce que les suivants sont a priori négligeables, au vu du lissage des réponses fréquentielles considérées.

L'erreur cepstrale est calculée à chaque actualisation de l'égaliseur adapté, soit toutes les 16 ms. Son évolution permet d'évaluer la rapidité de convergence de l'égaliseur : celle-ci a lieu lorsque l'erreur cepstrale a atteint une valeur minimale autour de laquelle elle varie peu.

Ni la constance de l'erreur spectrale sur l'espace des fréquences ni la valeur de l'erreur cepstrale ne permettent toutefois d'évaluer l'atteinte de l'objectif de restauration du timbre de la voix originale (ne serait-ce que limitée à la bande F_c -3150 Hz), même s'ils en donnent une image. L'erreur spectrale peut ne pas être exactement constante, ou l'erreur cepstrale ne pas être nulle, sans que cette différence soit perceptible. Reste à déterminer jusqu'à quel point. On peut imaginer que la perceptibilité de la distorsion spectrale correspond à un seuil d'erreur cepstrale. D'après les écoutes informelles réalisées, ce seuil serait autour de 0,2. Cependant, aucun test formel n'a été mené pour établir ce seuil de manière rigoureuse. En outre, à une même erreur cepstrale peuvent correspondre des distorsions spectrales de formes variées et diversement localisées dans l'espace des fréquences. Il paraît hasardeux de fixer la perceptibilité de ces distorsions en fonction d'une seule variable scalaire. Enfin, l'erreur cepstrale définie par la formule (2.25) permet d'évaluer la correction des distorsions spectrales sur la bande F_c -3150 Hz mais ne permet pas d'évaluer l'égaliseur en termes de restauration du timbre original.

Au final, eu égard à l'imperfection des mesures objectives, la restauration du timbre de la parole originale sera évaluée par des tests subjectifs (section II.3.6). Cette restauration sera évaluée d'une part par comparaison du signal de réception à celui d'émission, d'autre part en tenant compte de la nécessaire limitation de la bande d'égalisation : il s'agira alors de comparer le signal de réception après égalisation adaptée au signal de réception obtenu avec un égaliseur idéal.

II.3.3. Rapidité de convergence de l'égaliseur

La simulation ayant été réalisée pour les 34 locuteurs dans les 6 conditions de transmission envisagées, la Figure 2.16 présente, pour une des liaisons test, l'évolution de l'erreur cepstrale au cours des 21 premières secondes de parole active. L'égaliseur n'étant pas actualisé pendant les moments d'inactivité vocale, ces derniers, non pertinents du point de vue de la variation de l'erreur cepstrale, ont été supprimés. La liaison test considérée ici est celle comportant un

système d'émission différent de la caractéristique nominale de [UIT-T/P.830, 1996] et une ligne analogique longue à l'émission.

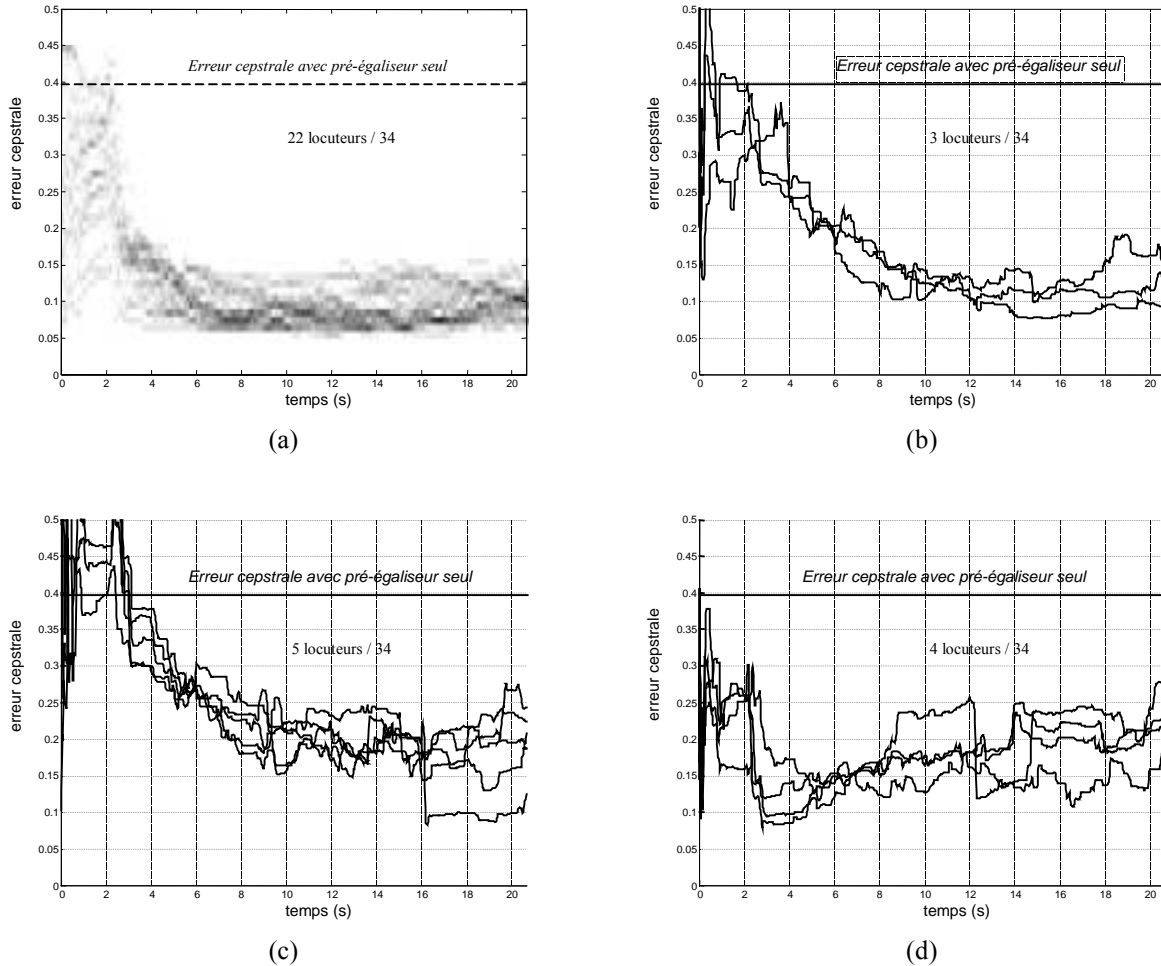


Figure 2.16 : Evolution de l'erreur cepstrale au cours des 21 premières secondes de parole, pour 34 locuteurs.

Pour 22 des 34 locuteurs testés, la convergence de l'égaliseur, marquée par l'atteinte d'un niveau minimal d'erreur cepstrale, est réalisée en 2 à 5 secondes. L'évolution de l'erreur cepstrale pour ces locuteurs est représentée par une série d'histogrammes sur la Figure 2.16(a). Chaque $n^{\text{ième}}$ ligne verticale représente l'histogramme des erreurs cepstrales à la $n^{\text{ème}}$ trame d'activité vocale, un pixel de coordonnées (n,e) étant d'autant plus sombre que le nombre de locuteurs ayant une erreur cepstrale autour de e à la $n^{\text{ème}}$ trame est important.

Pour 8 autres locuteurs, la convergence est plus lente, l'erreur minimale étant atteinte en 10 secondes environ. Pour 3 d'entre eux, celle-ci est proche de 0,1 comme dans le premier groupe : les trajectoires des erreurs de ces locuteurs sont représentées sur la Figure 2.16(b). L'erreur cepstrale des 5 autres locuteurs, représentée sur la Figure 2.16(c), converge vers une valeur plus élevée (0,2).

Enfin, pour 4 locuteurs (Figure 2.16(d)), l'évolution de l'erreur cepstrale est plus atypique, avec une première décroissance rapide (moins de 3 secondes), suivie d'un comportement assez irrégulier. Ce phénomène s'explique par des variations brusques du spectre à long terme, que l'on peut percevoir comme des modifications du timbre au cours de l'élocution.

En moins de 3 secondes d'activité vocale, pour tous les locuteurs, l'erreur cepstrale est inférieure à celle obtenue avec le pré-égaliseur seul, représentée par une ligne horizontale à l'ordonnée 0.4.

Les résultats obtenus avec les autres liaisons sont très proches, comme l'illustre la Figure 2.17, représentant l'évolution de l'erreur cepstrale pour 4 des locuteurs test et pour les 6 liaisons test. Les locuteurs ont été choisis de la manière suivante :

- H1 : locuteur masculin, erreur cepstrale finale forte ;
- H2 : locuteur masculin, erreur cepstrale finale faible ;
- F1 : locutrice, erreur cepstrale finale faible ;
- F2 : locutrice, erreur cepstrale finale forte.

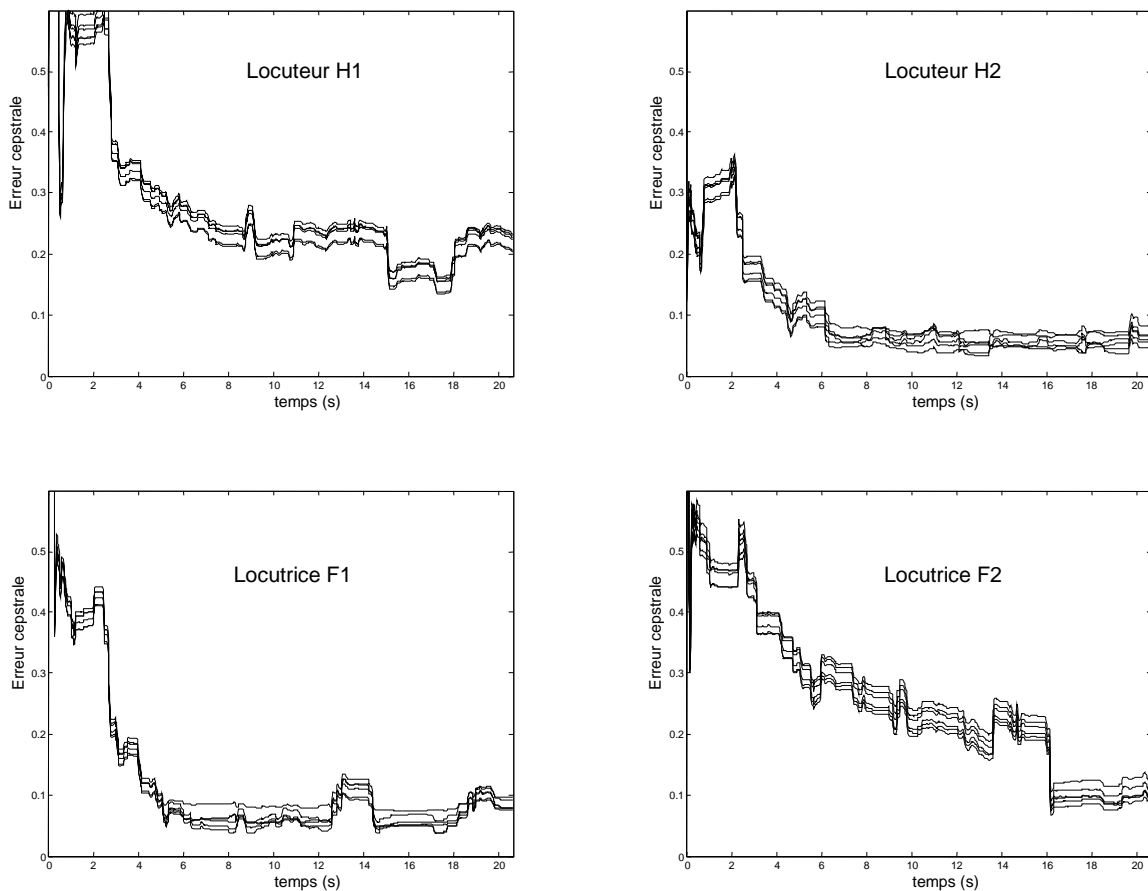


Figure 2.17 : Evolution de l'erreur cepstrale au cours des 21 premières secondes de parole, pour 4 locuteurs et 6 liaisons.

II.3.4. Distorsion spectrale finale

La Figure 2.18 présente, pour les mêmes locuteurs, la réponse fréquentielle de l'égaliseur adapté (trait plein) comparée à celle de l'égaliseur adapté idéal (pointillés), après 10 s d'activité vocale, c'est-à-dire après la convergence, pour 3 des 6 liaisons test :

- liaison 1 : système d'émission ayant la caractéristique nominale du SRI modifié [UIT-T/P.830, 1996] et ligne d'émission très courte ;

- liaison 2 : système d'émission ayant la caractéristique nominale du SRI modifié et ligne d'émission moyenne ;
- liaison 3 : système d'émission différent de la caractéristique nominale du SRI modifié mais respectant le masque (voir Figure 2.14) et ligne d'émission longue.

Ces trois liaisons correspondent aux distorsions médianes et extrémales de la Figure 2.15, représentées en trait plein.

La différence, en dB, entre la réponse fréquentielle de l'égaliseur adapté et celle de l'égaliseur adapté idéal correspond, sur la bande d'égalisation 200-3150 Hz, à la distorsion spectrale entre le signal de réception et le signal émis. L'amplitude des variations de cette différence sur l'espace des fréquences n'excède pas 3 dB pour H2 et F1, représentatifs des deux tiers des locuteurs du corpus, quelle que soit la liaison. Pour les quatre locuteurs, la distorsion spectrale après égalisation dépend très peu de la liaison.

Cette réponse fréquentielle de l'égaliseur adapté se traduit par les distorsions spectrales entre le signal original et le signal de réception représentées sur la Figure 2.19 pour les mêmes locuteurs et les mêmes liaisons.

Pour H2 et F1, quelle que soit la liaison, la caractéristique fréquentielle de la liaison égalisée est nettement plus plate que celles de la liaison sans traitement ou simplement pré-égalisée, et ce sur une bande de fréquence plus large. La seule exception est évidemment la liaison 2, pour laquelle le pré-égaliseur correspond à l'égaliseur idéal. Subjectivement, le signal de réception de la liaison ainsi égalisée a un timbre très proche de celui du signal égalisé par l'égaliseur idéal. Pour H2, la différence avec le signal original est nettement perceptible, puisque la voix de ce locuteur masculin possède des composantes en deçà de 200 Hz, qui ne sont pas restaurées. En revanche, le signal de réception égalisé est perceptivement proche de l'original pour F1, dont le pitch moyen est supérieur à 200 Hz.

Pour H1 et F2, la réponse fréquentielle de la liaison égalisée présente des variations importantes, jusqu'à 9 dB, sur la bande 200-3150 Hz. Cela se traduit par une voix plus étouffée que la voix émise ou que celle en réception de la même liaison égalisée par l'égaliseur idéal. A cette distorsion s'ajoute la non-restauration des composantes basses-fréquences, très sensible pour H1.

Au vu de ces résultats, l'égalisation atteint donc son but pour la majorité des locuteurs : le canal analogique est compensé sur la bande 200-3150 Hz avec une erreur inférieure à 3 dB dans l'estimation de la réponse fréquentielle, et le timbre de la voix est restauré au moins pour la voix restreinte à la bande 200-3150 Hz. Ces premières conclusions seront validées dans la section II.3.6 par des tests subjectifs formels, qui permettront en outre de préciser dans quelle mesure l'égalisation échoue pour des locuteurs tels que H1 et F2.

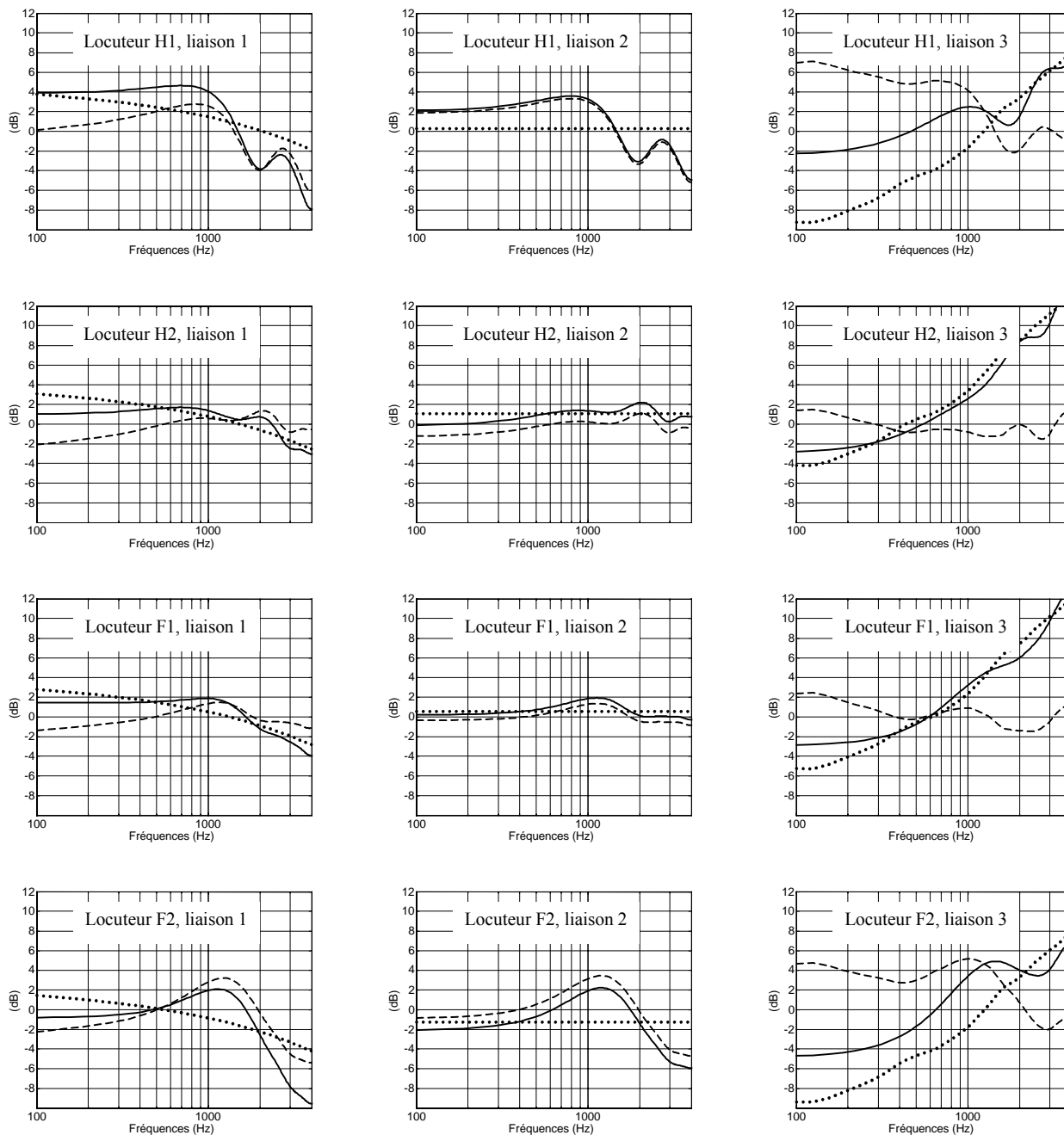


Figure 2.18 : Réponses fréquentielles de l'égaliseur adapté (trait plein) et de l'égaliseur adapté idéal (pointillés), différence entre les deux (tirets)

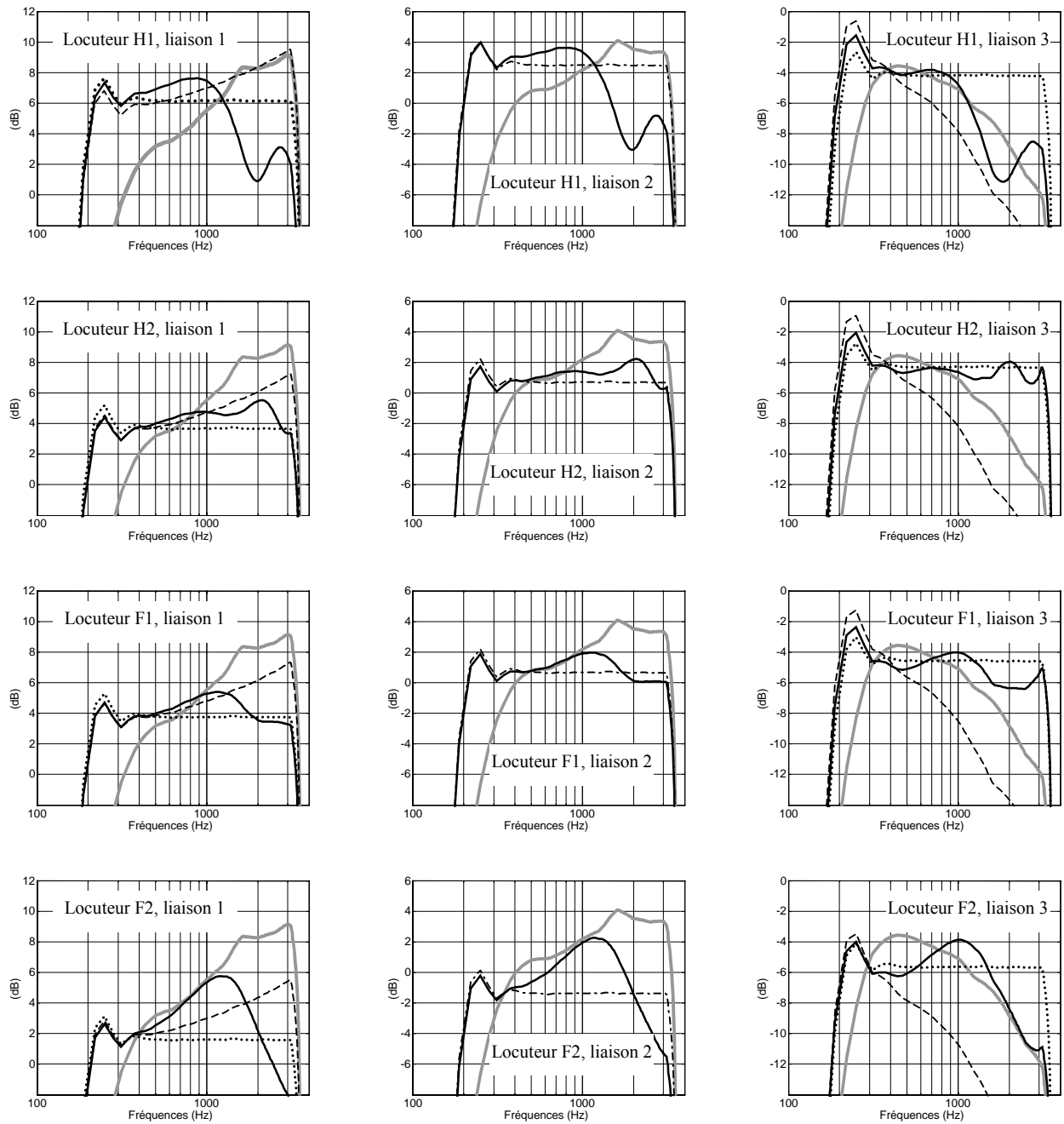


Figure 2.19 : Pour 3 liaisons et 4 locuteurs, réponse fréquentielle de la liaison

- sans traitement (trait épais gris)
- égalisée (trait plein noir)
- égalisée par l'égaliseur idéal (pointillés)
- pré-égalisée (tirets)

Remarque : pour la liaison 2, les signaux pré-égalisé et égalisé idéalement sont les mêmes

II.3.5 Limite de l'égalisation : le bruit de quantification

Lors des simulations et de la validation en temps réel du système proposé, nous avons observé à l'écoute que le signal en réception de la liaison égalisée est affecté d'un bruit blanc, de niveau irrégulier et souvent supérieur à celui du bruit de quantification du signal de réception de la même liaison sans égaliseur. Recherchant la cause de ce bruit en excès, nous avons étudié le rapport signal à bruit des signaux égalisés dans différentes conditions. Prenant en compte le caractère irrégulier du bruit additif observé, nous définissons le *RSB instantané* d'une trame de parole bruitée comme le rapport entre l'énergie de la trame de signal de parole et celle de la trame de bruit correspondante, ces trames ayant une durée de 32 ms et un recouvrement inter-trames de 50 %. La Figure 2.20 représente l'évolution du RSB instantané du signal de réception pour une phrase de 1,7 secondes ("*A-t-il eu froid cette nuit ?*") transmise par une liaison moyenne, dans deux cas : sans égaliseur et avec un égaliseur corrigeant parfaitement la liaison sur la bande 200-3150 Hz. Ces mesures confirment le résultat subjectif : le RSB en réception est dégradé par l'égalisation, de manière irrégulière.

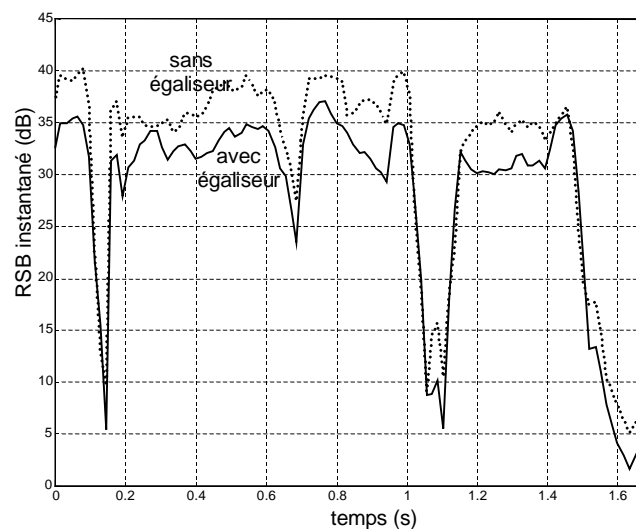


Figure 2.20 : RSB instantané de réception sur une liaison RTC avec ou sans égaliseur

Le RSB instantané a été calculé dans les mêmes conditions de liaison avec un corpus de 10 locuteurs (5 hommes et 5 femmes) prononçant 10 phrases phonétiquement équilibrées [Combescure, 1981], représentant pour chaque locuteur 22,6 secondes d'activité vocale en moyenne. Les résultats se traduisent par la répartition des dégradations de RSB indiquée sur la Figure 2.21 : les valeurs correspondent à l'écart entre les RSB de réception d'une liaison égalisée (avec $F_c = 200$ Hz) et ceux d'une liaison non égalisée.

Une première explication est que, dans la liaison avec égaliseur, le signal subit deux quantifications en loi A : celle du codeur MIC et celle à la sortie de l'égaliseur, celui-ci traitant les signaux dans un format linéaire. Ceci est illustré par le schéma de liaison de la Figure 2.22, où l'on a modélisé par une addition de bruit blanc :

- la mise en cascade du codeur MIC, de la liaison numérique et du convertisseur loi A \rightarrow format linéaire (bruit q_0) ;
- la mise en cascade de la conversion format linéaire \rightarrow loi A, de la liaison numérique et du décodage MIC (bruit q_1).

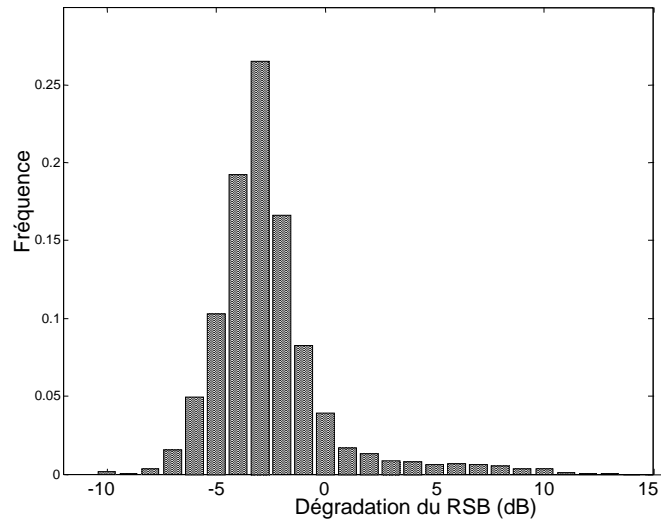


Figure 2.21 : Distribution des dégradations du RSB de réception dues à l'égalisation

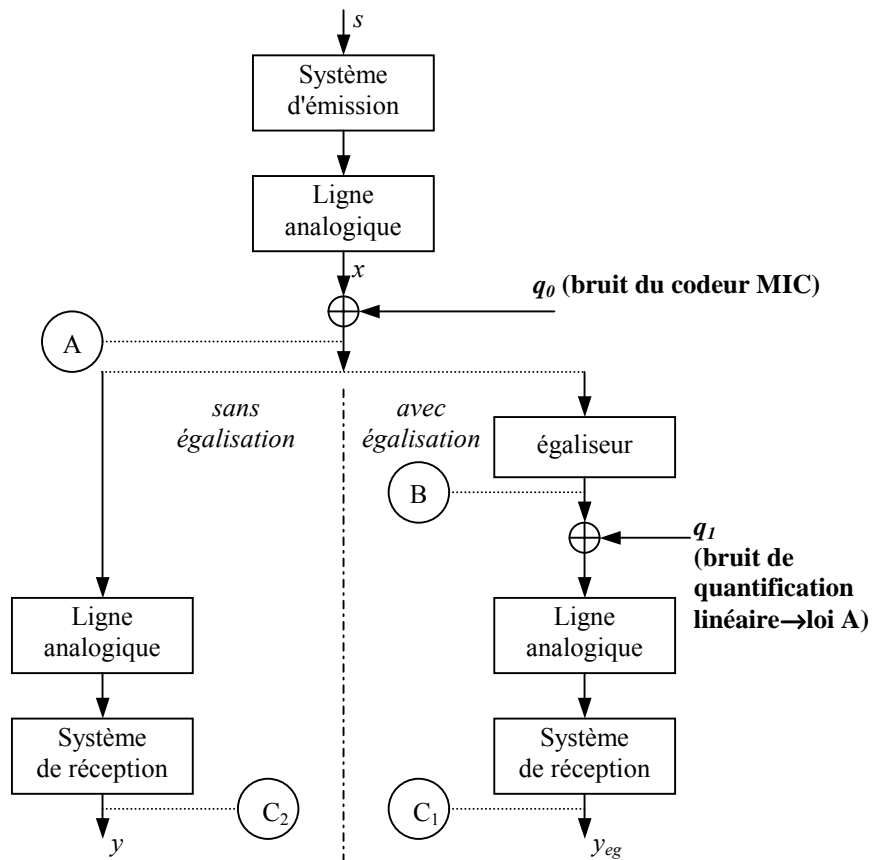


Figure 2.22 : Bruits de quantification sur une liaison RTC avec ou sans égaliseur

Cette double addition de bruit de quantification ne suffit cependant pas à expliquer les dégradations de RSB observées : seule, elle ne justifierait qu'une dégradation constante de 3 dB. Comme l'égaliseur est placé avant le système de réception dont il corrige l'atténuation des composantes basse fréquence, il suramplifie celles-ci. La Figure 2.23 représente, pour différentes valeurs de F_c , la réponse globale en fréquence du filtrage appliqué à la parole entre l'extrémité

"émission" de la liaison et la sortie de l'égaliseur, c'est à dire juste avant la deuxième conversion en loi A (point B sur la Figure 2.22). Les composantes fréquentielles du signal de sortie de l'égaliseur sont ainsi d'autant plus déséquilibrées que F_c est faible et que le spectre original du phonème prononcé est riche dans la bande 200-300 Hz. **Or le quantificateur en loi A superpose au signal un bruit blanc avec un RSB de 38,16 dB dans le meilleur des cas. Selon les phonèmes prononcés, du fait de la prédominance des basses fréquences, le niveau de ce bruit de quantification peut être proche de celui des composantes hautes et moyennes fréquences. Ainsi, après atténuation des composantes basses fréquences à la réception, le RSB est dégradé, et ce d'autant plus que F_c est basse.**

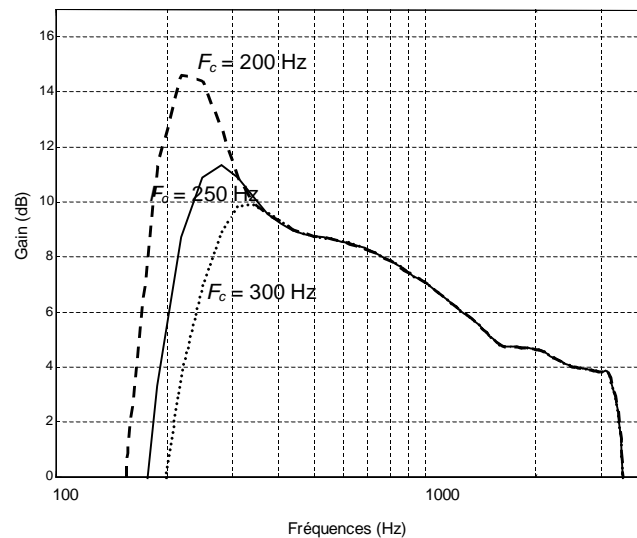


Figure 2.23 : Filtrage appliqué à la parole entre l'extrémité "émission" de la liaison et la sortie de l'égaliseur

Examinons plus en détail ce phénomène sur deux phonèmes de la phrase précédente, le [Y] de "nuit" et le [a] de "froid", pour expliquer la variabilité de la dégradation du RSB. Les Figure 2.24 et Figure 2.25 représentent, pour ces deux phonèmes respectivement, les enveloppes spectrales du signal de parole et des bruits en trois points successifs de la liaison schématisée sur la Figure 2.22 : après le codage MIC (A) ; après l'égaliseur le cas échéant (B) ; en réception pour une liaison avec ou sans égaliseur (respectivement C1 et C2). Les bruits q_0 et q_1 résultent respectivement du codeur MIC et de la quantification en loi A à la sortie de l'égaliseur. Pour la Figure 2.24 (phonème [Y]), les enveloppes spectrales ont été calculées sur une trame où l'égaliseur dégrade le RSB instantané en réception de 5,5 dB. La dégradation est de 2,5 dB seulement pour la trame de calcul des enveloppes spectrales de la Figure 2.25 (phonème [a]). De manière à pouvoir comparer le niveau des bruits de quantification dans les cas avec et sans égaliseur, le gain de l'égaliseur a été ajusté de telle manière que l'énergie de chaque trame en réception soit la même qu'en réception de la liaison sans égaliseur.

Le phonème [Y] possède un premier formant vers 200 Hz, atténué par le système d'émission. Après suramplification de la bande 200-300 Hz par l'égaliseur, les composantes en dessous de 300 Hz sont donc très énergétiques (B). Elles déterminent le niveau du bruit de quantification du convertisseur linéaire-loi A (q_1), qui est alors nettement supérieur à celui du bruit du codeur MIC (q_0) et dépasse le deuxième formant. A la réception (C₁), l'énergie cumulée des deux bruits est supérieure de 5 dB environ à celle du bruit blanc de la liaison sans traitement (C₂).

Le premier formant de [a] est situé vers 600 Hz. Ainsi la suramplification des basses fréquences par l'égaliseur (B) ne modifie-t-elle pas notablement l'équilibre des composantes

fréquentielles. Le niveau du bruit du convertisseur linéaire-loi A (q_1) est donc proche de celui du codeur MIC (q_0). A la réception (C_1), l'énergie cumulée des deux bruits est proche de celle du bruit blanc de la liaison sans traitement (C_2).

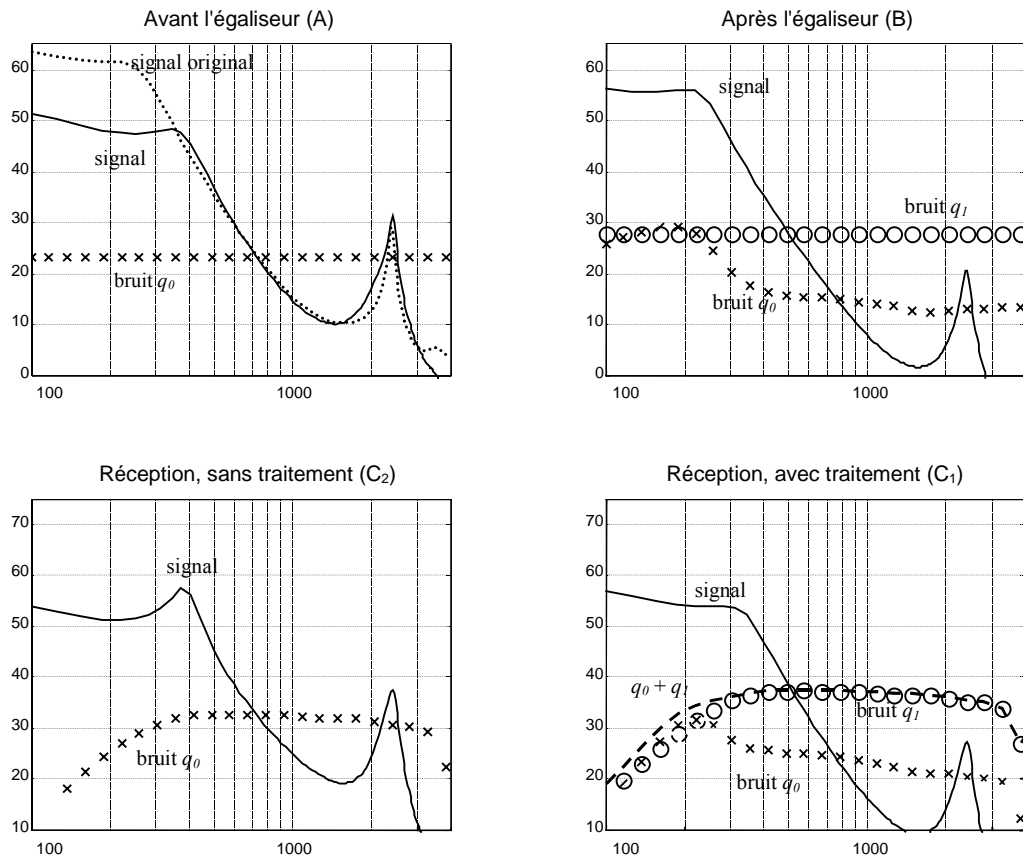


Figure 2.24 : Enveloppe spectrale de $[Y]$ et des bruits associés en différents points de la liaison

Une manière de traiter cette amplification du bruit de quantification due à l'égalisation peut être d'augmenter F_c . L'effet de cette augmentation est illustré sur la Figure 2.26, qui représente, pour la même phrase que précédemment, la différence entre le RSB instantané de réception d'une liaison avec égalisation sur F_c -3150 Hz et celui d'une liaison sans égalisation, pour deux valeurs de F_c . L'augmentation de F_c permet d'éviter les dégradations de RSB les plus fortes. Les basses fréquences étant une part importante du timbre perçu, l'inconvénient de cette solution est que le timbre n'est plus aussi bien restauré. Il faut donc trouver un compromis entre restauration du timbre et niveau du bruit ou agir sur le convertisseur linéaire-loi A pour réduire le bruit de quantification à F_c constant. Cette deuxième solution fera l'objet du chapitre III.

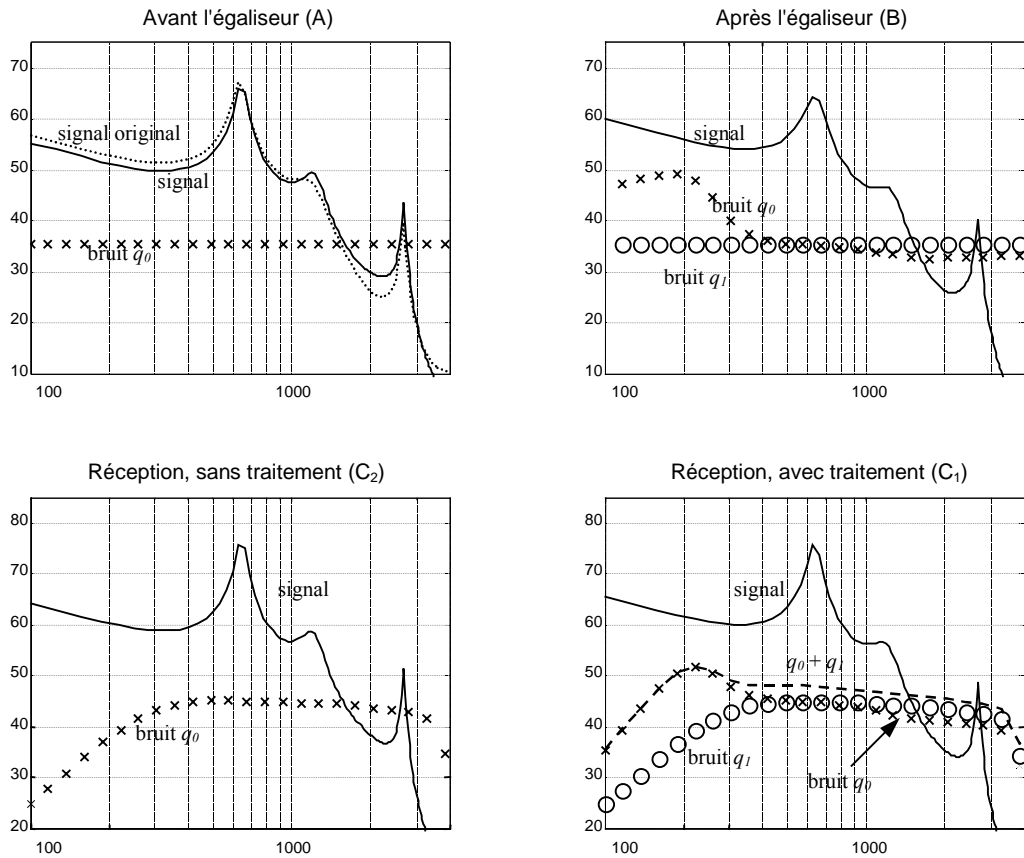


Figure 2.25 : Enveloppe spectrale de [a] et des bruits associés en différents points de la liaison

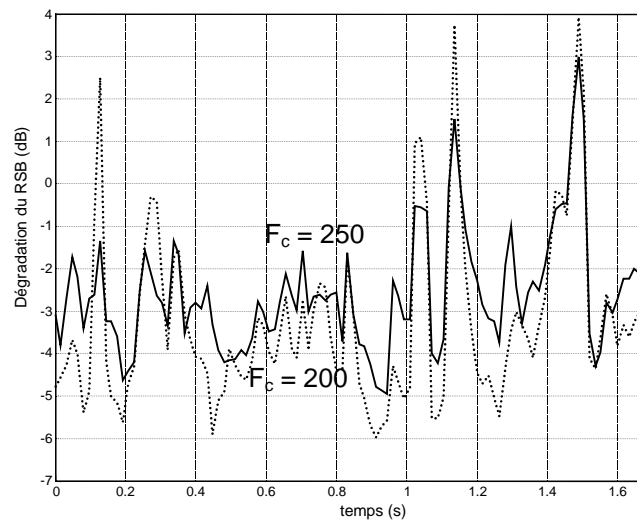


Figure 2.26 : Dégradation du RSB pour $F_c = 200$ et 250 Hz

II.3.6. Évaluation subjective

- **Objectifs**

L'objectif du traitement étant que le "timbre" de la voix en réception soit le plus proche possible de celui de la voix originale émise, l'atteinte de cet objectif sera évaluée, pour différents types de liaisons téléphoniques et différents locuteurs, en demandant à des auditeurs de noter la proximité de timbre entre le signal de réception égalisé et le signal original. Une telle note aurait peu de sens seule, notamment si l'auditeur perçoit une nette différence de timbre entre les deux signaux. C'est pourquoi la même comparaison doit être effectuée entre le signal original et le signal de réception non traité, ainsi qu'entre ce dernier et le signal de réception égalisé.

Par ailleurs, l'évaluation doit tenir compte de la nécessaire limitation de l'égalisation à la bande 200-3150 Hz. Il convient donc de comparer le signal de réception égalisé avec le signal de réception traité par l'égaliseur idéal précédemment évoqué. Si aucune différence n'est perceptible entre les deux, cela signifie que l'égaliseur adapté corrige au mieux le timbre dans les limites imposées par la restriction de la bande.

Enfin, il est intéressant d'évaluer l'apport de la partie adaptée de l'égaliseur. Pour ce faire, les mêmes comparaisons doivent être effectuées en remplaçant le signal de réception égalisé par le signal de réception de la même liaison dans laquelle l'égaliseur adapté aurait été supprimé (le pré-égaliseur seul subsistant).

- **Application de la méthode MUSHRA [UIT-R/BS.1534, 1996]**

Cette évaluation nécessite typiquement un test de comparaison par paires. Cependant, eu égard au nombre de comparaisons à effectuer, cette méthode conduirait à un test d'une durée inacceptable si l'on souhaite évaluer les performances de l'égaliseur pour différents locuteurs et différentes conditions de transmission. C'est pourquoi nous nous inspirons d'une méthode plus simple, récemment normalisée à l'UIT, la méthode "multi stimuli avec référence et repère cachés" (MUSHRA, *MULTI Stimuli test with Hidden Reference and Anchor*) [UIT-R/BS.1534, 1996].

La méthode MUSHRA a été conçue pour évaluer des codecs audio de qualité intermédiaire, *i.e.* introduisant des dégradations moyennes ou fortes. Elle s'applique typiquement aux systèmes tels que la diffusion sur l'Internet, la radio mondiale numérique (DRM, *digital radio mondiale*) ou la radiodiffusion numérique par satellite, pour lesquels des sons de moindre qualité sont inévitables ou acceptables.

La méthode consiste à présenter simultanément au sujet un signal de référence et des signaux tests, qu'il peut écouter librement, et à demander de noter, par comparaison à la référence, la qualité des signaux tests, sur une échelle continue de 0 à 100, divisée en 5 intervalles de même longueur étiquetés "*mauvais*" (0 à 20), "*médiocre*" (20 à 40), "*assez bon*" (40 à 60), "*bon*" (60 à 80) et "*excellent*" (80 à 100). Les signaux tests comprennent à la fois les signaux testés (par exemple les signaux issus des différents codecs que l'on souhaite évaluer), le signal de référence ("*référence cachée*") et un signal repère ("*repère caché*" ou "*point d'ancrage*"). Ce signal repère correspond à un niveau de qualité audio bien connu (par exemple le signal de référence filtré par un passe-bas de fréquence de coupure 3500 Hz), ce qui offre une base de comparaison des notes attribuées aux signaux testés.

Dans le cas de dégradations fortes ou moyennes, si l'on comparait uniquement chaque signal test à la référence, deux signaux ayant des niveaux de dégradation assez proches

risqueraient d'obtenir la même note alors que le sujet aurait trouvé l'un meilleur que l'autre s'il avait comparé ces deux signaux tests entre eux. Pour surmonter cette difficulté, l'auditeur, dans le test MUSHRA, peut commuter à volonté entre la référence et tous les signaux test. Ainsi, le test est équivalent à un test complet de comparaison par paires, tout en étant moins long, l'auditeur adoptant une stratégie intelligente de notation : généralement, les participants estiment d'abord grossièrement la qualité de chaque signal test, puis affinent la note en comparant les échantillons de qualité proche. La méthode MUSHRA permet ainsi d'obtenir une résolution élevée des notes.

Le test doit comporter une phase de familiarisation, au cours de laquelle les sujets apprennent à utiliser l'interface de test et peuvent écouter les différentes dégradations qui affecteront les signaux testés. Cette phase de familiarisation permet de fiabiliser les résultats.

La norme [UIT-R/BS.1534, 1996] recommande d'utiliser des signaux de durée inférieure à 20 secondes, et de ne pas présenter plus de 15 signaux tests (référence et repère cachés inclus) au cours de chaque séquence de test.

Les sujets sont choisis de préférence expérimentés, c'est-à-dire ayant l'habitude d'écouter des sons de manière critique. Il est possible de les présélectionner – *i.e.* d'éliminer des sujets insuffisamment expérimentés ou inaptes auditivement – ainsi que de les post-sélectionner – *i.e.* de rejeter les sujets dont les résultats sont incohérents soit entre eux soit avec ceux du groupe. Une vingtaine de sujets suffit à obtenir des résultats fiables ; il est conseillé d'en retenir plus si les sujets sont peu expérimentés.

Bien que la méthode MUSHRA ait été conçue pour tester des codecs audio, elle nous a paru, par ses principes, être adaptée à l'évaluation à mener. La référence est le signal original (dans la bande 0-4000 Hz) et les signaux à tester, dans chaque séquence, sont les signaux de réception :

- sans traitement (TRANSP) ;
- égalisé selon l'algorithme proposé (EG) ;
- seulement pré-égalisé (*i.e.* égalisation fixe corrigeant des conditions moyennes de transmission) (PRE) ;
- égalisé avec l'égaliseur idéal (*i.e.* égaliseur corrigeant parfaitement la liaison sur la bande F_c -3150 Hz) (ID).

L'échelle de notation est adaptée à notre cas de la manière suivante : comme il ne s'agit pas de noter la qualité du signal connaissant la référence, mais la proximité de timbre entre chaque signal test et la référence, c'est celle-ci qui est notée sur une échelle de 0 à 100, avec les appréciations suivantes :

- timbre identique : 100 ;
- timbre très proche : 80 à 100 ;
- timbre assez proche : 60 à 80 ;
- timbre moyennement proche : 40 à 60 ;
- timbre assez différent : 20 à 40 ;
- timbre très différent : 0 à 20.

Ainsi, si la note de EG est nettement supérieure à celle de TRANSP, l'égalisation corrige le timbre. Si la note de EG est supérieure à celle de PRE, on en déduit que la partie adaptée de l'égalisation améliore la correction. Si la note de EG est proche de 100, la correction est parfaite. Enfin, si les notes de EG et ID sont proches, l'égaliseur adapté corrige au mieux le timbre sur la bande F_c -3150 Hz.

Les signaux tests comprendront à la fois les signaux à tester et la référence. Cependant, il ne nous est pas possible d'introduire de repère caché : ce test étant le premier du genre, nous ne disposons pas de signal repère dont la note serait connue. Cela ne remet pas en cause la validité de l'utilisation de la méthode MUSHRA, puisque l'introduction d'un repère caché n'a qu'un rôle de comparaison des systèmes testés à un repère connu, et ne sert en aucun cas à un réajustement des notes. La référence constitue alors le seul repère caché.

- **Plan de test**

Nous souhaitons évaluer l'égaliseur, avec $F_c = 200$ Hz, pour différentes combinaisons de liaisons et de locuteurs. La réponse fréquentielle de l'égaliseur dépendant uniquement du spectre à long terme du locuteur, la phrase prononcée a peu d'influence sur le fonctionnement de l'égaliseur une fois que la convergence est atteinte. Par conséquent, la liaison égalisée sera simulée sur tout le texte *"la bise et le soleil"*, mais seule la dernière phrase, prononcée après 13 à 16 secondes d'activité vocale selon les locuteurs (donc après convergence de l'égaliseur), sera présentée aux auditeurs : *"Alors la bise se mit à souffler de toutes ses forces; mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui, et à la fin la bise renonça à le lui faire enlever."* (10 s environ)

Un minimum de 4 locuteurs est nécessaire. Nous choisissons d'une part H2 et F1, qui correspondent aux performances typiques de l'égaliseur en terme d'erreur cepstrale après convergence (25 locuteurs sur 34, voir Figure 2.16 (a) et (b)), d'autre part H1 et F2, qui correspondent aux performances les plus basses. L'évolution de l'erreur cepstrale au cours de la phrase test est représentée pour ces quatre locuteurs sur la Figure 2.27. Le choix de F2 est certes particulier, eu égard à la brusque variation d'erreur cepstrale au milieu de la phrase. Cette locutrice a cependant été choisie comme représentante des locuteurs à forte erreur cepstrale, d'une part parce que nous pensons que la partie la plus dégradée de la phrase sera déterminante dans le jugement des auditeurs, d'autre part parce que F2 est la seule locutrice du corpus présentant à la fois une forte erreur cepstrale et une voix assez agréable.

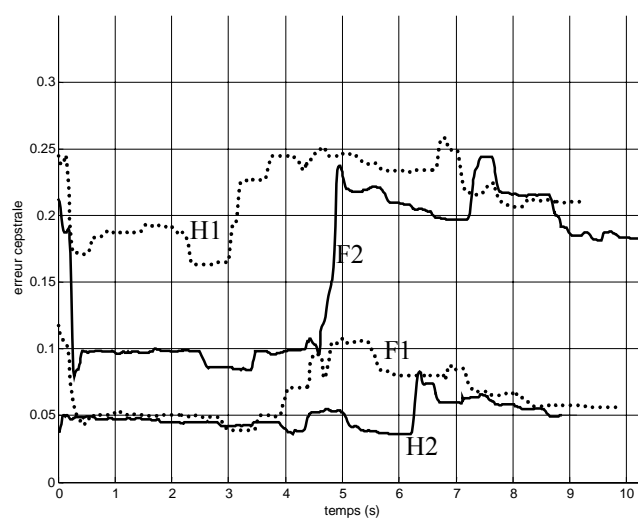


Figure 2.27 : Erreur cepstrale de l'égaliseur sur la phrase test pour les quatre locuteurs choisis

Il a été souhaité d'évaluer l'égaliseur sur les combinaisons suivantes d'éléments de liaisons téléphoniques, la partie émission de la liaison étant sur le RTC (voir Figure 2.30) :

- les 2 types de terminaux d'émission évoqués dans la section II.3.1 ;
- 3 types de lignes analogiques d'émission (très courte / moyenne / longue) ;
- 3 types de liaison en réception : liaison RTC moyenne / liaison RNIS / liaison GSM avec codage EFR. Dans le cas de la liaison GSM, l'égaliseur est supposé placé au niveau du transcodage MIC-EFR.

Il n'est pourtant pas envisageable de tester ces 18 conditions de liaison, qui, combinées aux 4 locuteurs à tester, conduiraient à 72 séquences MUSHRA. L'expérience montre qu'une séquence telle que celle envisagée (5 signaux tests de 10 s) représente environ 5 mn de test. Comme le test ne peut excéder 1 h (pour ne pas fatiguer les sujets), il est impératif de réduire à 12, et si possible moins, le nombre de séquences. Le nombre de locuteurs ne pouvant être plus restreint, nous avons réduit le nombre de liaisons en comparant leurs signaux de réception respectifs, égalisés d'une part, sans traitement d'autre part, de la manière suivante.

Les simulations correspondant aux 18 liaisons ont été réalisées pour les 4 locuteurs tests. Pour les liaisons RNIS et GSM, l'expression de la réponse fréquentielle de l'égaliseur adapté donnée par l'équation (2.18) est remplacée respectivement par :

$$|EQ(f)| = \frac{1}{|RNIS_RX(f)|} \sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}} \quad (2.26)$$

$$|EQ(f)| = \frac{1}{|GSM_RX(f).EFR(f)|} \sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}} \quad (2.27)$$

où $RNIS_RX$ est l'efficacité en réception supposée connue du terminal $RNIS$, GSM_RX désigne l'efficacité en réception supposée connue du mobile et EFR est la modification d'enveloppe spectrale du codage EFR (voir Figure 1.11). Les caractéristiques fréquentielles du terminal $RNIS$ et du mobile simulés sont choisies de manière arbitraire dans le gabarit défini par [UIT-T/P.310, 2000] et [UIT-T/P.313, 2000] (cf. Figure 1.10) et représentés sur les Figures 2.28 et 2.29 respectivement.

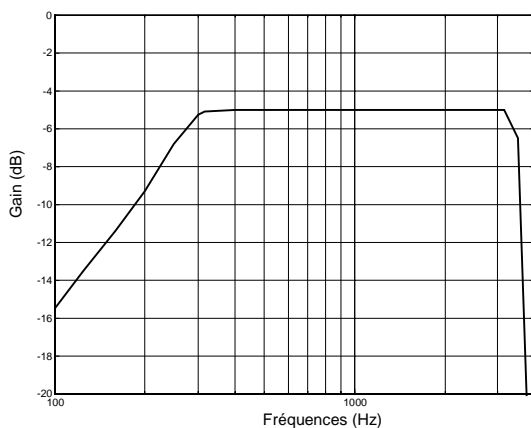


Figure 2.28 : Efficacité en réception du terminal $RNIS$ simulé

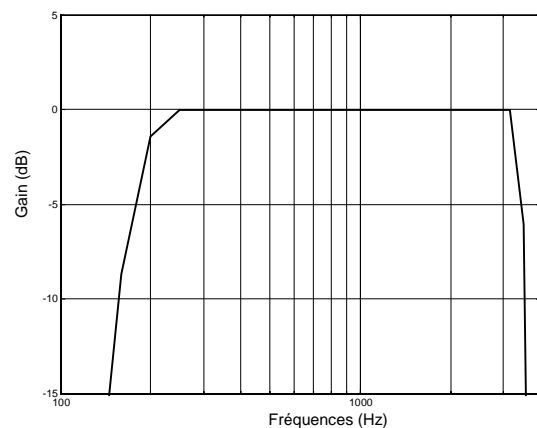


Figure 2.29 : Efficacité en réception du terminal GSM simulé

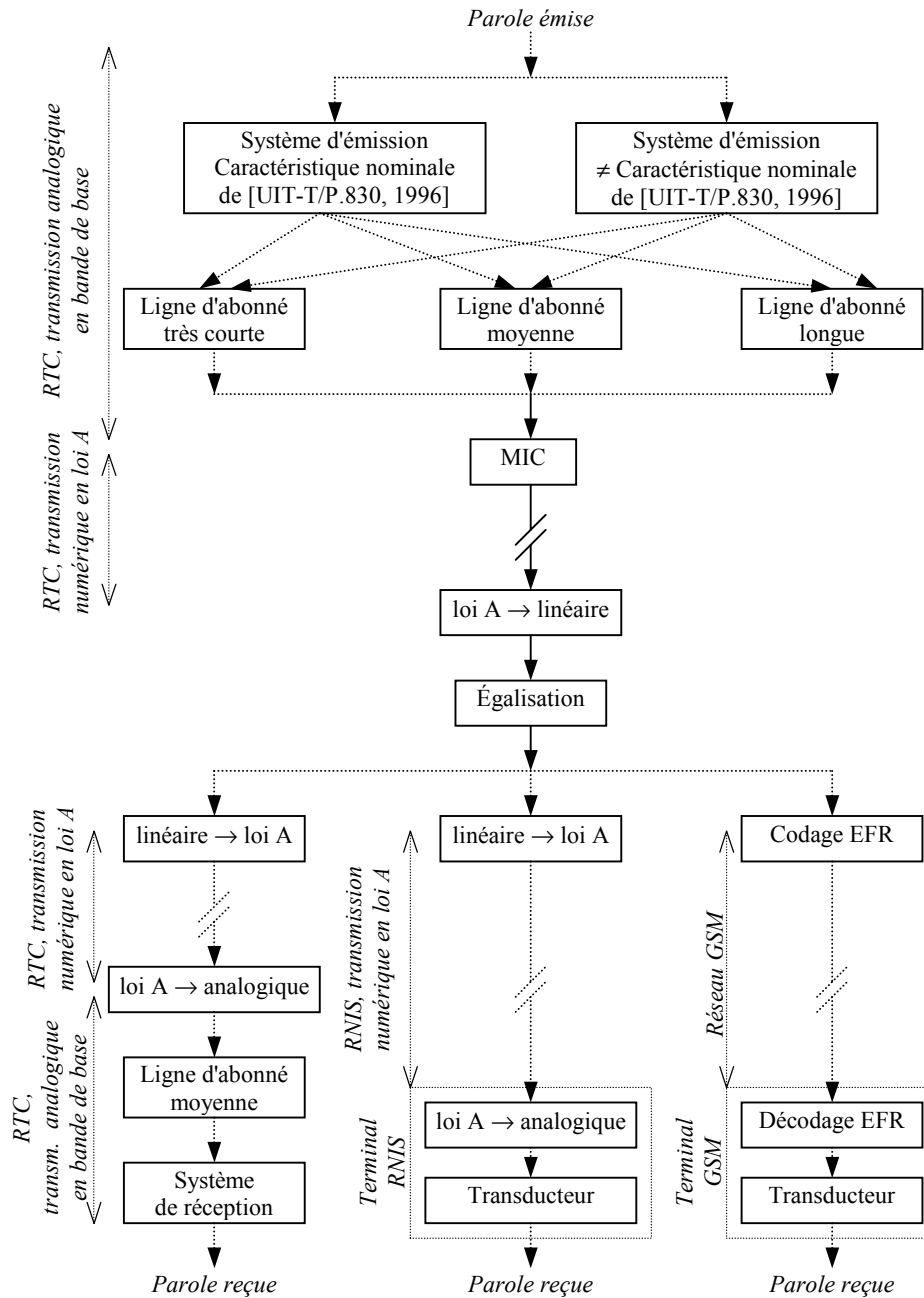


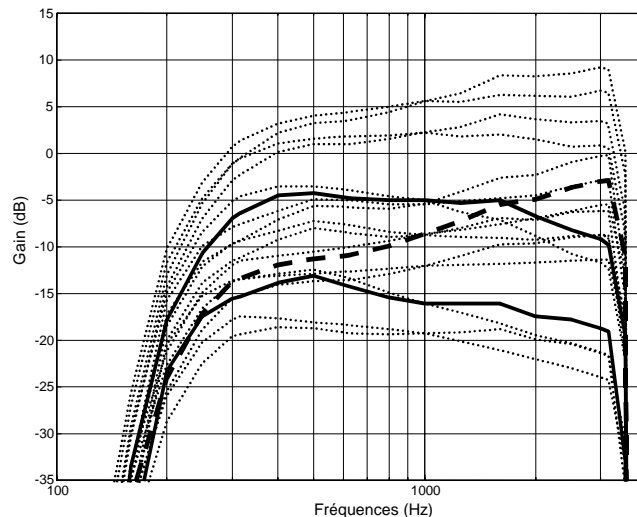
Figure 2.30 : Liaisons téléphoniques initialement envisagées pour le test

Comme le montrent les résultats de la section II.3.4, la distorsion spectrale finale après égalisation est peu sensible, dans le cas de la liaison RTC, aux caractéristiques des terminaux et des lignes d'émission (sauf caractéristiques très éloignées des caractéristiques moyennes). Les tests informels effectués ne révèlent pas de différence de timbre perceptible entre les signaux de réception égalisés correspondants. Si la réception RTC est remplacée par une réception RNIS ou mobile, le résultat est identique, puisque la partie réception de la liaison n'a aucune influence sur l'adaptation de l'égaliseur. La seule différence de timbre, clairement perceptible uniquement pour les locutrices, est la coloration propre au codage EFR (sonorité robotique). Mis à part cet aspect, on peut donc prédire que les notes de EG seront peu différentes d'une liaison à l'autre.

Les différences entre les liaisons résident principalement entre les signaux de réception sans traitement. La Figure 2.31 représente les modifications d'enveloppe spectrale correspondant

aux 18 liaisons envisagées. En termes perceptifs, ces filtrages correspondent à un continuum de timbres du plus clair au plus sourd, avec peu de différence entre deux plus proches voisins. On se restreint donc aux liaisons extrémales, les plus démonstratives. A cela s'ajoute une deuxième dimension dans la variation du timbre : avec ou sans la coloration du codage EFR. Ainsi, selon ces deux dimensions, il reste 4 liaisons envisageables :

- 2 liaisons entièrement fixes (RTC ou RNIS) : sourd / clair ;
- 2 liaisons RTC-GSM : sourd / clair.



*Figure 2.31 : Distorsions spectrales introduites par les 18 liaisons.
En gras, celles des trois liaisons retenues dans le plan de test*

Ce nombre de conditions doit être encore réduit, et peut l'être selon la simplification suivante. Les écoutes informelles font apparaître que l'effet de la coloration GSM est nettement moins perceptible que le choix d'une liaison "sourde" ou "claire", et quasiment imperceptible pour les locuteurs masculins. Le codage EFR n'a par ailleurs pas d'influence sur le fonctionnement de l'algorithme, puisqu'il a lieu après l'égalisation dans le cas d'une communication fixe vers mobile. Le seul effet qu'il peut avoir est de modifier légèrement la perception de l'amélioration introduite par la correction de timbre. Par conséquent, il ne nous apparaît pas nécessaire de tester pour le GSM les 2 conditions "sourde" et "claire". Les éventuelles variations des résultats entre les conditions "liaison fixe sourde" et "liaison RTC-GSM sourde" seront extrapolables au passage "liaison fixe claire" → "liaison RTC-GSM claire". **On se contentera donc des conditions suivantes : "liaison fixe sourde", "liaison fixe claire" et "liaison RTC-GSM sourde", cette dernière condition étant testée uniquement pour les locutrices.** On a ainsi réduit à 10 le nombre de conditions (liaison, locuteur), ce qui représente 50 mn de test.

Nous évaluons au final les performances de l'égaliseur dans les conditions suivantes :

- pour les 4 locuteurs, liaison RTC intégrale comprenant un système d'émission respectant la caractéristique nominale de [UIT-T/P.830, 1996], une ligne d'émission analogique longue, une ligne de réception analogique moyenne et un système de réception analogique respectant la caractéristique nominale de [UIT-T/P.830, 1996] (liaison 1) ;
- pour les 4 locuteurs, liaison RTC-RNIS comprenant un système d'émission analogique respectant la caractéristique nominale de [UIT-T/P.830, 1996], une ligne d'émission

analogique très courte, et une liaison de réception RNIS avec un terminal respectant la caractéristique en fréquence de la Figure 2.28 (liaison 2) ;

- pour les 2 locutrices, liaison RTC-GSM comprenant un système d'émission respectant la caractéristique nominale de [UIT-T/P.830, 1996], une ligne d'émission analogique longue et une liaison de réception GSM avec un codage EFR et un terminal mobile respectant la caractéristique en fréquence de la Figure 2.29 (liaison 3).

Les distorsions spectrales introduites par ces trois liaisons sont représentées en gras sur la Figure 2.31.

Dans chacune de ces conditions, le sujet compare, sur la phrase test, le timbre des signaux TRANSP, PRE, EG et ID à celui du signal original, et attribue à chacun une note de proximité de timbre avec l'original, entre 0 (timbre très différent) et 100 (timbre identique).

Dans le cas des liaisons RTC et RNIS, le signal de réception égalisé est affecté d'un fort bruit de quantification. Même si l'on précise au sujet de noter uniquement la proximité de timbre avec l'original, la présence de ce bruit risque de perturber le jugement de l'auditeur. C'est pourquoi, de manière à s'assurer d'évaluer uniquement la capacité de l'égaliseur à restaurer le timbre original, les signaux tests sont obtenus en supprimant la quantification en loi A qui suit l'égaliseur dans la liaison simulée. Il ne s'agit pas de négliger les limites réelles de l'égaliseur, mais d'isoler la question du timbre pour l'évaluer indépendamment des autres dégradations du signal, le traitement conjoint du timbre et du bruit faisant l'objet de travaux et d'évaluations ultérieurs décrits dans le chapitre III.

La norme [UIT-R/BS.1534, 1996] recommande d'utiliser des sujets expérimentés. Compte tenu des difficultés à trouver 20 sujets experts, nous avons constitué un panel intermédiaire : 24 sujets ont participé à l'expérience, dont :

- 13 experts, *i.e.* auditeurs ayant l'habitude d'écouter des sons de manière critique ;
- 11 naïfs habitués à utiliser des interfaces informatiques.

Les sujets sont placés seuls dans une pièce calme devant une interface graphique sur ordinateur, telle que représentée sur la Figure 2.32 (logiciel CRC-SEAQ, du Communication Research Center, Ottawa). L'échelle "excellent" ... "bad" n'étant pas modifiable, il est précisé à chaque sujet (à la fois oralement et dans les consignes écrites) de ne pas en tenir compte et de se référer à l'échelle de proximité de timbre précédemment évoquée. Le PC est équipé d'une carte son haute qualité Digigram VX222, dont le niveau, initialement à une valeur "confortable", est ajustable pour ceux qui le désirent. L'écoute se fait en binaural sur casque fermé Sony MDR CD1000. Il s'agit de conditions d'écoute plus discriminantes et plus sensibles que l'écoute téléphonique habituelle. Dix séquences sont présentées au sujet, correspondant à la phrase "*Alors la bise ...*" prononcée dans les 10 conditions (liaison, locuteur) retenues.

Le bouton REF correspond à la référence (signal original), les boutons A, B, C, D, E aux fichiers tests (ORI, ID, EG, PRE et TRANSP), selon un ordre aléatoire mais identique pour tous les sujets. Au-dessus de chaque bouton se trouve un curseur permettant de juger sur une échelle continue la proximité de timbre avec l'original. L'auditeur indique sa note en déplaçant le curseur avec la souris. Le fichier noté est celui couramment sélectionné (en cliquant sur le bouton) et affiché en rouge.

On peut écouter la phrase en boucle (bouton boucle) et commuter entre les différents fichiers comme on le désire en cliquant sur les boutons REF, A, B, ... On peut également sélectionner des parties du signal à l'aide des curseurs du bas et effectuer l'évaluation en ayant

isolé ces parties. Cette présentation de l'interface permet au sujet non seulement d'évaluer les signaux à tester par rapport à la référence, mais aussi de les comparer entre eux. Une fois les fichiers notés, le sujet appuie sur la touche "Next Trial" : le logiciel demande alors la confirmation des votes et passe à la séquence suivante.

Le test est précédé de la lecture des consignes reproduites dans l'annexe A, accompagnées de quelques indications orales. La durée totale du test, incluant la lecture des consignes et la phase d'apprentissage, est de 50 mn en moyenne.

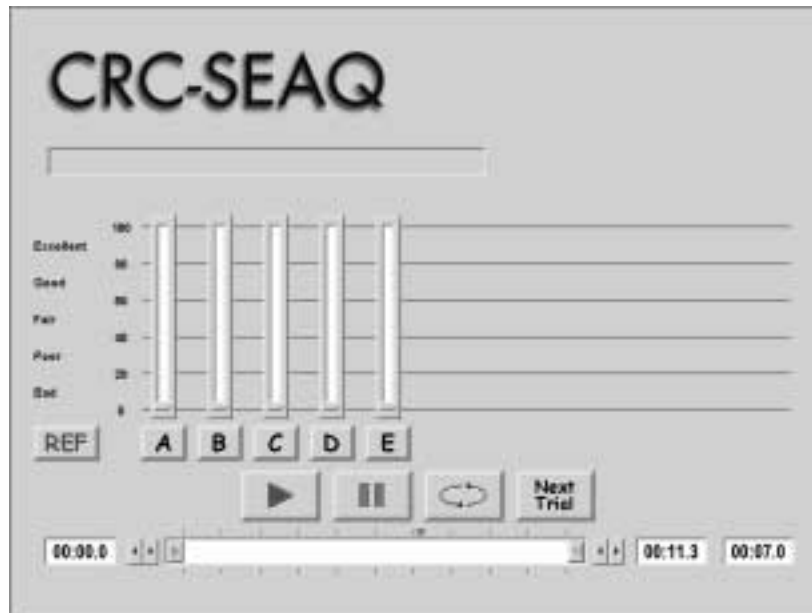


Figure 2.32 : Interface du logiciel CRC-SEAQ utilisée pour les tests

• Résultats

Les Figures 2.33 à 2.36 représentent, pour les quatre locuteurs et les trois liaisons test, les résultats des tests subjectifs, en regard avec les distorsions spectrales correspondantes. Pour H1, H2 et F1, la distorsion de la liaison égalisée varie peu au cours de la phrase, à l'instar de l'erreur cepstrale correspondante (voir Figure 2.27). Pour F2, la distorsion de la liaison égalisée est représentée en double, chaque représentation correspondant à un niveau d'erreur cepstrale (environ 0,1 dans la première moitié de la phrase, environ 0,2 dans la deuxième). Pour chaque liaison et chaque locuteur sont représentées les notes moyennes des 5 signaux tests, ainsi que les intervalles de confiance à 95 % associés. Pour une note moyenne \bar{u} , l'intervalle de confiance à 95 % est défini par :

$$I = \left[\bar{u} - t_{0,05} \frac{\sigma}{\sqrt{N}} ; \bar{u} + t_{0,05} \frac{\sigma}{\sqrt{N}} \right], \quad (2.28)$$

où $t_{0,05}$ est le quantile de la loi gaussienne normale associé à la probabilité de 95 %, σ est l'estimée de l'écart-type de la note et N est le nombre de sujets. La probabilité que la note moyenne réelle (celle que l'on obtiendrait avec un nombre de sujets infini) soit dans cet intervalle est de 95 %.

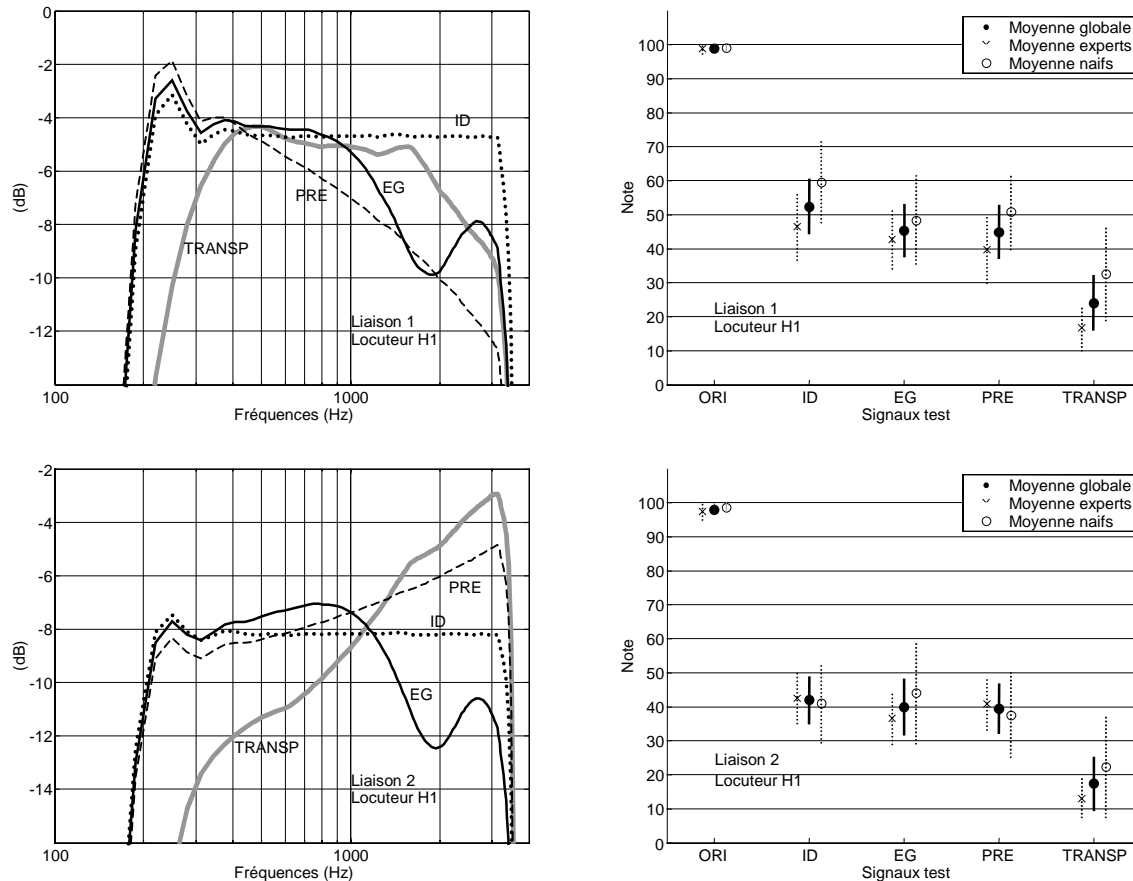


Figure 2.33 : Pour le locuteur H1 et les liaisons 1 et 2,
 - à gauche, distorsion de la liaison sans traitement (trait épais gris) ; égalisée (trait fin noir);
 égalisée par l'égaliseur idéal (pointillés) ; seulement pré-égalisée (tirets)
 - à droite, notes de leurs signaux de réception respectifs

Pour toutes les liaisons et tous les locuteurs, la note de EG est significativement supérieure à celle de TRANSP, ce qui confirme l'efficacité de la correction de distorsion spectrale indiquée par les résultats objectifs.

En revanche, pour les locuteurs masculins, les différences objectives entre EG, ID et PRE ne se retrouvent pas dans les notes moyennes. Les auditeurs jugent certes meilleure la correction apportée par l'égaliseur idéal, mais l'amélioration est très faible et les intervalles de confiance de ID, EG et PRE se recouvrent largement. Par ailleurs, la note moyenne de EG est quasiment identique à celle de PRE, alors que, selon les courbes de distorsion spectrale, l'égaliseur adapté devrait apporter une nette amélioration du timbre pour le locuteur H2.

Les notes de la locutrice F1 sont un peu plus conformes aux résultats objectifs. Aux bonnes performances de l'égaliseur correspondent des notes moyennes très proches de celles de l'égaliseur idéal, tandis que le filtrage passe-bas introduit par le pré-égaliseur seul dans les liaisons 1 et 3 se traduit par des notes inférieures, avec cependant un recouvrement de 50 % entre les intervalles de confiance. La note de PRE est proche de celles de ID et EG pour la liaison 2, ce qui s'explique par une distorsion spectrale moins prononcée que pour les liaisons 1 et 3.

Les notes de la locutrice F2 sont, de manière plus nette, conformes à ce que les courbes de distorsion spectrale laissaient prévoir. La hiérarchie des notes correspond à celle des déséquilibres spectraux introduits par les différents égaliseurs, en supposant que des écarts de 3 à

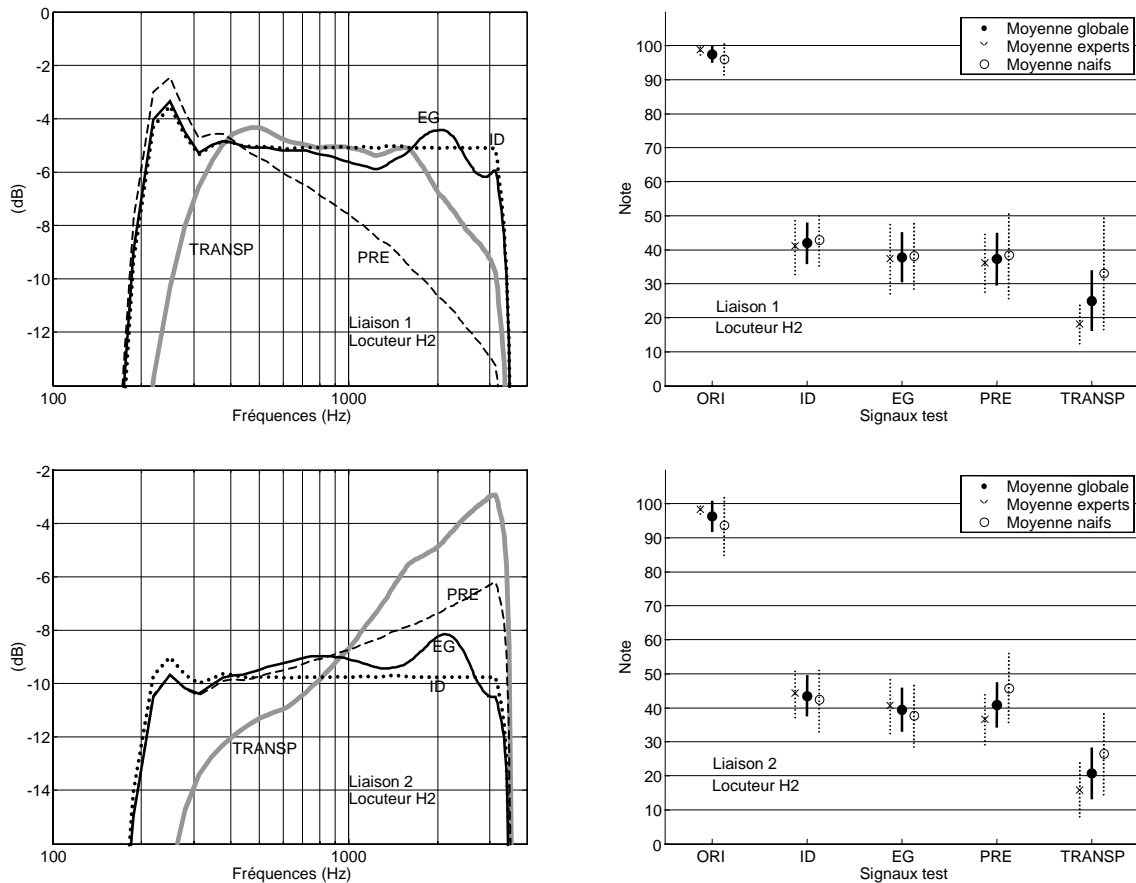


Figure 2.34 : Pour le locuteur H2 et les liaisons 1 et 2,
 - à gauche, distorsion de la liaison sans traitement (trait épais gris) ; égalisée (trait fin noir) ;
 égalisée par l'égaliseur idéal (pointillés) ; seulement pré-égalisée (tirets)
 - à droite, notes de leurs signaux de réception respectifs

4 dB ne sont pas perceptibles : ID>EG>PRE pour les liaisons 1 et 3 ; ID≈PRE>EG pour la liaison 2.

La comparaison entre les résultats de la liaison 3 et ceux de la liaison 1 pour les locutrices conforte partiellement les choix effectués dans la préparation du plan de test. La hiérarchie des notes de ID, EQ et PRE est conservée, avec une dégradation globale des notes pour la liaison 3 qui peut s'expliquer par l'altération du timbre par le codage EFR. La note attribuée à TRANSP ne subit pas cette dégradation, notamment pour F2, ce qui s'explique par ce que les transducteurs GSM simulés atténuent un peu moins que le SRI modifié les composantes basse fréquence, part importante du timbre perçu.

Notons que les résultats diffèrent peu entre les auditeurs experts et naïfs : dans chaque séquence, soit les notes moyennes des deux groupes d'auditeurs sont très proches, soit elles diffèrent du même nombre de points pour tous les signaux testés, la notation des experts étant généralement plus sévère. Le choix d'intégrer des auditeurs naïfs altère donc assez peu la fiabilité des résultats.

Ces résultats mettent en évidence la correction du timbre introduite par l'égaliseur, mais sont globalement très proches pour les trois égaliseurs, malgré les distorsions très fortes introduites dans certains cas par l'égaliseur (non idéal) ou le pré-égaliseur seul, et indépendamment de la variabilité des performances de l'égaliseur selon le locuteur. Une première

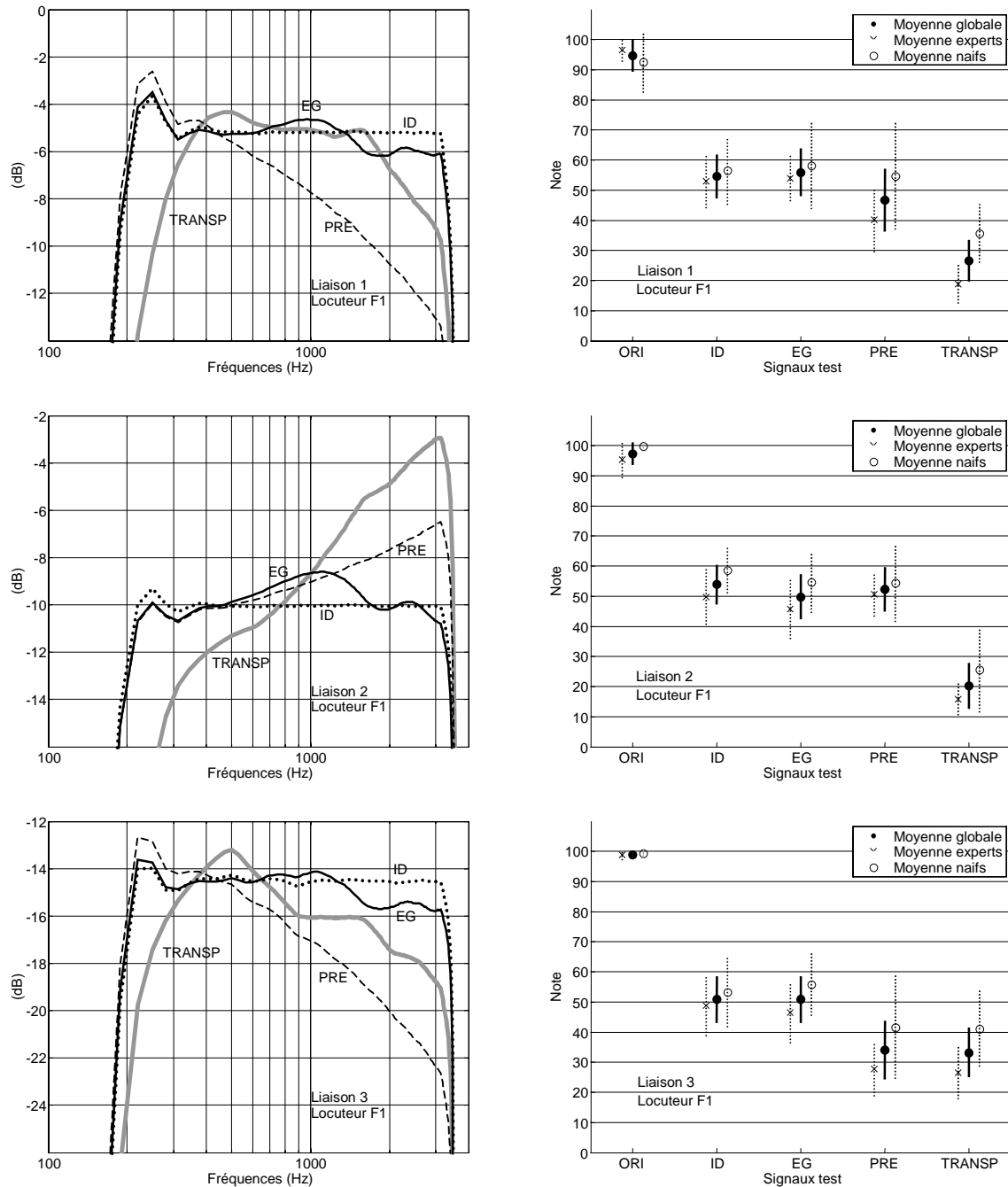


Figure 2.35 : Pour la locutrice F1 et les trois types de liaisons,
 - à gauche, distorsion de la liaison sans traitement (trait épais gris) ; égalisée (trait fin noir) ;
 égalisée par l'égaliseur idéal (pointillés) ; pré-égalisée seulement (tirets)
 - à droite, notes de leurs signaux de réception respectifs

conclusion pourrait être que notre égaliseur rapproche certes le timbre de celui du signal original, puisque la note de EG est toujours nettement supérieure à celle de TRANSP, mais est inutilement complexe, puisqu'un filtre fixe corrigeant une liaison moyenne obtient des résultats subjectifs moyens très proches pour la plupart des locuteurs. Les analyses complémentaires qui suivent permettent cependant de nuancer cette conclusion.

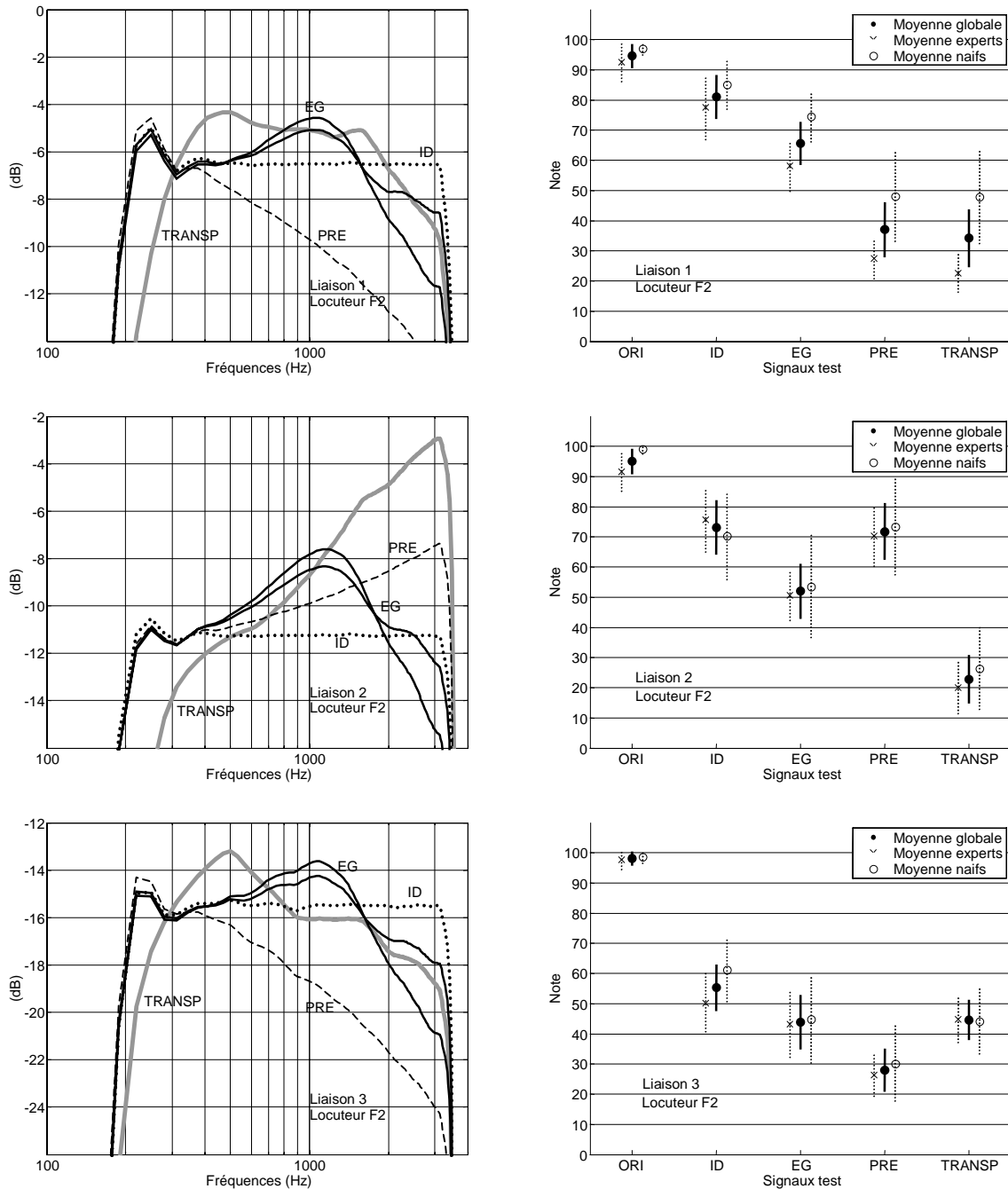


Figure 2.36 : Pour la locutrice F2 et les trois types de liaisons,
- à gauche, distorsion de la liaison sans traitement (trait épais gris) ; égalisée (trait fin noir) ;
égalisée par l'égaliseur idéal (pointillés) ; pré-égalisée seulement (tirets)
- à droite, notes de leurs signaux de réception respectifs

Une analyse de variance (ANOVA) est effectuée pour chaque liaison séparément. Pour chaque ANOVA, on considère le facteur Traitement à 5 niveaux (ORI, ID, EG, PRE, TRANSP) ainsi que le facteur Locuteur à 4 niveaux (H1, H2, F1 et F2) pour les liaisons 1 et 2 et 2 niveaux (F1, F2) pour la liaison 3.

La Figure 2.37 représente, pour les liaisons 1 à 3, les notes moyennes de chaque traitement (ORI, ID, EG, PRE, TRANSP) tous locuteurs confondus, ainsi que les intervalles de confiance associés. Un test de Tukey [Tukey, 1953] effectué sur le facteur Traitement uniquement (*i.e.* tous locuteurs confondus) permet de mesurer le caractère significatif des différences de notes

entre ID, EG et PRE. Ce test montre que pour les liaisons 1 et 3, la différence entre ID et EG n'est pas significative (indice de significativité $p = 0,26$, nettement supérieur au seuil de significativité 0,05), alors qu'elle est significative entre ID et PRE ainsi qu'entre EG et PRE ($p < 0,05$). Les résultats détaillés figurent dans l'annexe B. En d'autres termes, **EG est jugé, en moyenne sur l'ensemble des sujets et des locuteurs, comme à même "distance" de l'original que ID, PRE étant quant à lui jugé plus éloigné.**

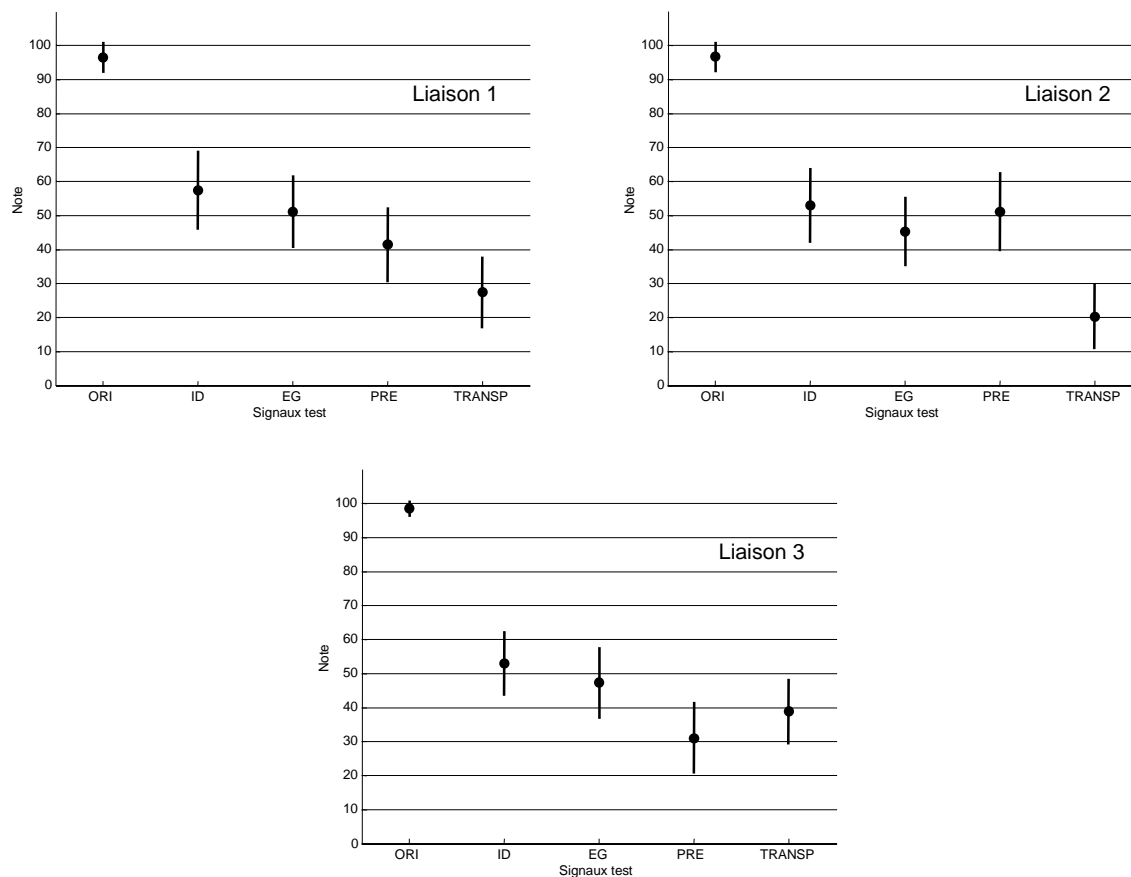


Figure 2.37 : Notes moyennes des signaux test sur tous locuteurs et intervalles de confiance associés

Toutefois, le même test réalisé locuteur par locuteur ne fait apparaître ce résultat que pour la locutrice F1 dans la liaison 3. Pour cette locutrice avec les liaisons 1 et 2, le test n'indique pas de différence significative entre ID, EG et PRE, malgré la forte distorsion spectrale introduite par le pré-égaliseur seul dans le cas de la liaison 1. De même, pour les deux locuteurs masculins, aucune différence significative entre ID, EG et PRE n'est révélée par le test, indépendamment des distorsions spectrales plus ou moins fortes des signaux. Exception faite des conditions (F1, liaison 3), le test ne donne de résultats conformes aux résultats objectifs que pour la locutrice F2 (comme dans la précédente analyse des résultats) : EG, ID et PRE sont jugés significativement différents pour les liaisons 1 et 3, tandis que pour la liaison 2, ID et EG sont jugés équivalents entre eux et significativement différents de PRE.

L'importance des composantes basses fréquences dans le jugement des auditeurs permet d'expliquer ces résultats. Les locuteurs H1, H2, F1 et F2 ont des pitchs moyens respectifs de 120 Hz, 160 Hz, 210 Hz et 240 Hz. Le fait que les résultats subjectifs de F2, et, dans une moindre mesure, de F1, soient les plus conformes aux résultats objectifs est significatif de cette influence de la restauration des composantes basses fréquences. Pour les locuteurs masculins, la faible discrimination du test subjectif entre EG, ID et PRE peut s'expliquer par la limitation de

l'égalisation à la bande 200-3150 Hz. La non-restauration des composantes en deçà de 200 Hz introduit une telle différence de timbre avec l'original que les différences entre les notes moyennes de EG, ID et PRE sont comprimées.

Nous nous proposons de préciser cet effet en exploitant différemment les résultats du test. Pour chaque auditeur, on calcule d'une part la différence Δ_{ID-EG} entre la note de EG et celle de ID, d'autre part la différence Δ_{ID-PRE} entre la note de PRE et celle de ID. La Figure 2.38 représente, pour le locuteur H2 et la liaison 1, les distributions de Δ_{ID-EG} et Δ_{ID-PRE} . Ces distributions font apparaître une nette séparation entre EG et PRE, qui ne pourrait pas être mise en évidence par une analyse de la variance, les deux classes étant concentriques. Ainsi, en moyenne, Δ_{ID-EG} et Δ_{ID-PRE} sont égaux, mais cette moyenne cache le fait que la majorité des auditeurs jugent EG nettement plus proche de ID que PRE (écart inférieur à 5), ce qui est conforme au résultat objectif observé. La différence de timbre entre ORI d'une part et PRE, EG et ID d'autre part est telle que les auditeurs sont incapables de juger avec certitude si PRE est plus ou moins éloigné de ORI que ID et EG : sur les 24 auditeurs, 10 répondent "moins", 14 répondent "plus".

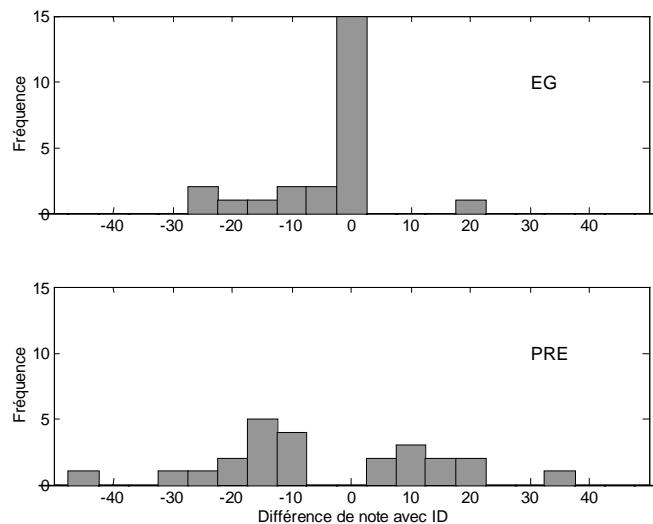


Figure 2.38 : Distribution des différences de notes entre EG et ID d'une part, entre PRE et ID d'autre part, pour le locuteur H₂ avec la liaison 1.

La comparaison de PRE et EG à ID doit donc se faire non seulement par comparaison des moyennes de Δ_{ID-EG} et Δ_{ID-PRE} , mais aussi par comparaison des variances de ces valeurs. Les moyennes et écarts-types de Δ_{ID-EG} et Δ_{ID-PRE} sont représentés pour les dix combinaisons de liaisons et de locuteurs sur la Figure 2.39. Pour les liaisons 1 et 2 et les locuteurs H1, H2 et F1, les moyennes de Δ_{ID-EG} et Δ_{ID-PRE} sont très proches, indépendamment des distorsions spectrales observées. Mais dans le cas des locuteurs H2 et F1, pour lesquels les résultats objectifs indiquent une plus grande proximité entre EG et ID qu'entre PRE et ID, l'écart-type de Δ_{ID-EG} est deux fois plus faible que celui de Δ_{ID-PRE} . Cela signifie que **la proximité des notes subjectives observée en moyenne entre ID et EG a une probabilité nettement plus forte que celle entre ID et PRE**. Pour le locuteur H1, aux mauvaises performances objectives de l'égaliseur correspond une variance de Δ_{ID-EG} similaire à celle de Δ_{ID-PRE} avec la liaison 1, inférieure avec la liaison 2.

Si cette analyse des résultats ne permet de conclure, pour H2 et F1, à une plus grande proximité entre ID et EG qu'entre ID et PRE qu'en termes de probabilité, cette faiblesse du résultat tient principalement à la consigne donnée aux auditeurs. Le test ne pouvant comporter

qu'une référence, il n'était pas demandé à ceux-ci de noter EG et PRE par rapport à ID mais par rapport à ORI.

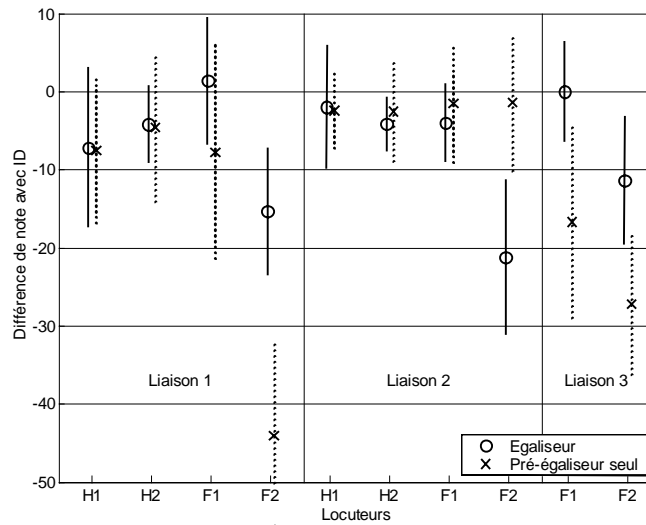


Figure 2.39 : Pour les dix combinaisons liaison-locuteur, moyennes et écarts-types des différences entre EG et ID d'une part, entre PRE et ID d'autre part.

• Bilan du test

La limitation de la bande d'égalisation dans les basses fréquences réduit la capacité de l'égaliseur à restaurer le timbre perçu : les notes attribuées à ID et EG sont proches de 50 seulement pour les trois premiers locuteurs, contre 20 environ pour TRANSP. Pour les locutrices à pitch élevé, comme F2, les notes de ID et EG sont plus élevées, mais celles de TRANSP le sont d'autant. Cependant, dans cette limite, les résultats témoignent d'une part de l'atteinte de l'objectif de distorsion nulle sur la bande d'égalisation, d'autre part de la supériorité de l'égaliseur proposé sur un filtre fixe compensant une liaison moyenne.

II.3.7. Validation de la version temps réel

La fonction de correction de timbre implantée sur la plate-forme *Mainate* a été validée de manière informelle lors de démonstrations. L'expérience consistait à établir une communication au cours de laquelle l'utilisateur pouvait commuter à volonté entre différents modes de fonctionnement de *Mainate* – sans traitement ; débruitage seul ; débruitage et correction de timbre – en appuyant sur les touches de son téléphone, de manière à comparer les effets correspondants. Les utilisateurs ont ainsi observé la restauration du timbre de la voix originale, aussi bien en ambiance bruitée que non bruitée, sur des liaisons fixes ou mobiles, le timbre étant jugé généralement "plus naturel" lorsque la correction de timbre était activée.

II.4. Conclusion

A partir d'un principe simple, aligner le spectre à long terme du signal de parole traité sur le spectre moyen de la parole défini par l'UIT-T, nous avons conçu et implanté en temps réel un égaliseur qui corrige le filtrage d'une liaison téléphonique sur la bande 200-3400 Hz de manière très satisfaisante pour la plupart des locuteurs. Subjectivement, la limitation de la bande de fréquences d'égalisation ne permet pas de restaurer pleinement le timbre original des locuteurs, mais le signal en réception d'une liaison égalisée s'en approche nettement plus que celui de la même liaison sans égaliseur. D'autre part, pour les trois quarts des locuteurs, l'amplitude de la distorsion spectrale entre le signal original et le signal en réception de la liaison égalisée n'excède pas 3 à 4 dB sur la bande d'égalisation. Cette faible erreur spectrale se traduit subjectivement par un timbre très proche de celui qu'on obtiendrait avec un égaliseur corrigeant parfaitement la liaison téléphonique sur cette bande. L'objectif d'une distorsion nulle sur la bande d'égalisation est donc atteint.

Cependant, la restauration des composantes basse fréquence de la parole par l'égaliseur amplifie le bruit de quantification perçu en réception, de manière particulièrement gênante pour certains locuteurs. C'est pourquoi nous nous attacherons, dans le chapitre III, à réduire perceptuellement ce bruit et à déterminer si une égalisation "bruyante" est préférable à l'absence d'égalisation.

D'autre part, pour une minorité de locuteurs, les performances de l'égaliseur restent en deçà de celles d'un égaliseur assurant une distorsion nulle sur la bande 200-3150 Hz. Ces locuteurs étant ceux dont le spectre à long terme s'éloigne le plus du spectre de référence sur lequel se fonde l'égalisation, nous tenterons de remédier à cette limitation des performances dans le chapitre IV, en adaptant l'égaliseur à la variété des spectres des locuteurs.

Chapitre III

Égalisation et bruit de quantification : approches perceptives

Nous avons montré au chapitre II comment la restauration des composantes basse fréquence de la parole par l'égaliseur peut amplifier sensiblement le bruit perçu en réception, bruit résultant de la quantification en loi A des échantillons de sortie de l'égaliseur. Dans ce chapitre, nous étudions la possibilité de réduire perceptivement ce bruit, en utilisant les propriétés de masquage fréquentiel de la parole. Nous proposons pour cela deux méthodes de reformage spectral du bruit de quantification, l'une fondée sur un filtrage de l'erreur de quantification, l'autre consistant à trouver une quantification optimale selon une approche probabiliste. Enfin, nous évaluons par des tests formels la perception subjective du bruit, selon qu'il est reformé ou non. Cette évaluation est menée en tenant compte de la problématique de la correction de timbre étudiée plus spécifiquement au chapitre II.

De manière à étudier le bruit de quantification indépendamment des erreurs d'estimation de la réponse fréquentielle de l'égaliseur, nous nous placerons dans le cas d'une liaison fixe moyenne telle que définie au chapitre II, corrigée par un égaliseur compensant parfaitement la distorsion spectrale subie par la parole sur la bande 200-3150 Hz.

III.1. Principes du masquage du bruit et application au codage

III.1.1. Le masquage fréquentiel du bruit

Lorsqu'un bruit et un signal de parole sont présentés simultanément, le bruit peut être inaudible, selon ses caractéristiques spectrales, celles du signal de parole et le caractère harmonique de ce dernier. Les principes de ce phénomène, appelé masquage fréquentiel, sont précisés en annexe C.

Une modélisation des propriétés de masquage de la parole selon ces principes permet de calculer, pour chaque trame de signal de parole, un seuil de masquage, tel que le bruit est inaudible si son spectre est en dessous de ce seuil. Parmi les méthodes existantes, nous avons retenu celle de Johnston [Johnston, 1988], pour sa simplicité de mise en œuvre. Comme nous le verrons dans les sections suivantes, l'objet de ce chapitre est plus de chercher une méthode de reformage spectral du bruit en fonction d'une courbe de masquage donnée que de rechercher la courbe de masquage la plus précise.

III.1.2. Calcul du seuil de masquage : méthode de Johnston

Le masque est calculé selon les étapes suivantes :

- Analyse en bandes critiques
- Convolution du spectre en Bark par la fonction d'étalement
- Soustraction d'un seuil de correction
- Normalisation du seuil de masquage
- Comparaison au seuil d'audition absolu
- Conversion du masque dans le domaine fréquentiel

• Analyse en bandes critiques

A l'instar du mécanisme de perception auditive, le spectre du signal est divisé en bandes critiques, et l'énergie du signal est calculée dans chaque bande critique, de manière à obtenir le spectre discret sur une échelle en Bark de 1 à 18 (pour un signal échantillonné à 8 kHz). Ainsi, si l'on note X la transformée de Fourier discrète du signal et B la densité spectrale sur l'échelle des Bark,

$$B(i) = \sum_{k \in b_i} |X(k)|^2 \quad (3.1)$$

où k désigne le $k^{\text{ème}}$ indice de fréquence et b_i la $i^{\text{ème}}$ bande critique.

• Convolution du spectre en Bark par la fonction d'étalement

Cette deuxième étape permet de tenir compte de l'étalement d'une excitation dans une bande critique sur les bandes critiques proches. La distribution de cette excitation autour de la bande critique considérée, représentée sur la Figure 3.1, est appelée fonction d'étalement. Une expression analytique de cette fonction est donnée dans [Schroeder, 1979] :

$$10 \cdot \log(f_E(i)) = 15,81 + 7,5 \cdot (i + 0,474) - 17,5 \cdot \sqrt{1 + (i + 0,474)^2}, \quad (3.2)$$

où i est le numéro de bande critique.

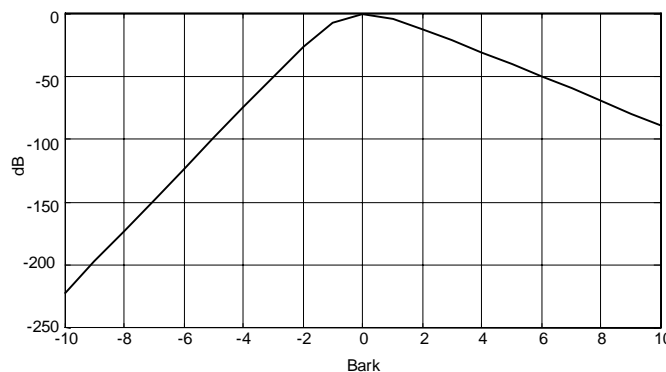


Figure 3.1 : Fonction d'étalement [Schroeder, 1979]

L'excitation globale du signal est considérée comme l'addition des excitations provoquées dans chaque bande critique, ce qui se traduit par la convolution du spectre en Bark par la fonction d'étalement :

$$E(i) = B(i) * f_E(i) \quad (3.3)$$

Le résultat de cette convolution est appelé *spectre étalé*.

- **Soustraction d'un seuil de correction**

La courbe de masquage correspond au spectre étalé à un écart près, qui dépend de la nature du signal. Dans ce modèle, pour une tonale masquant un bruit, le masque est situé à $14,5 + i$ dB sous le spectre étalé, avec i la fréquence en Bark. Pour un bruit masquant une tonale, l'écart est de 5,5 dB quelle que soit la fréquence. Il est en effet plus facile de masquer une tonale par un bruit que l'inverse.

La parole étant constituée schématiquement d'harmoniques et de bruit, c'est une combinaison de ces seuils de correction qui sera appliquée, dépendant de l'harmonicité du signal. A cet effet, on calcule une *mesure de platitude spectrale (spectral flatness measure, SFM)*, définie comme le rapport entre la moyenne géométrique et la moyenne arithmétique du spectre de puissance du signal. La valeur en dB de cette mesure, SFM_{dB} , est alors utilisée pour calculer un *coefficient de tonalité* α :

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dBmax}}, 1\right), \quad (3.4)$$

où SFM_{dBmax} vaut -60 dB et correspond par convention à la mesure de platitude spectrale d'un signal tonal pur. SFM vaut 0 dB pour un bruit blanc. Pour les signaux de parole, SFM est compris entre -20 et -30 dB. Le seuil de correction de masquage pour chaque bande critique est alors défini par :

$$O(i) = \alpha \cdot (14,5 + i) + (1 - \alpha) \cdot 5,5. \quad (3.5)$$

et le seuil de masquage vaut :

$$M(i) = E(i) - O(i). \quad (3.6)$$

- **Normalisation du seuil de masquage**

La fonction d'étalement accroît l'énergie estimée dans chaque bande critique. C'est pourquoi le seuil de masquage doit être normalisé par l'énergie de la fonction d'étalement.

- **Comparaison au seuil d'audition absolu**

Le masque normalisé est comparé au seuil d'audition absolu : pour chaque bande critique, le niveau final du masque est égal à la plus grande des valeurs entre le masque normalisé et le seuil d'audition.

• **Conversion du masque dans le domaine fréquentiel**

Le masque défini par la série de ses niveaux dans les bandes critiques est converti dans le domaine fréquentiel : pour toute fréquence f , si $f \in i$,

$$M(f) = M(i). \tag{3.7}$$

III.1.3. Application au masquage du bruit de quantification

La Figure 3.2 rappelle la structure de la liaison en aval de l'égaliseur et du quantificateur en loi A qui la suit. Nous souhaitons reformer le spectre du bruit de quantification, de manière à ce qu'il soit sous la courbe de masquage de la parole. Celle-ci sera calculée, à chaque trame, selon la méthode de Johnston, avec les adaptations suivantes.

D'une part, le signal étant traité dans le réseau, il est difficile de connaître le niveau absolu de réception. Par conséquent, l'étape de comparaison du masque au seuil d'audition absolu est supprimée.

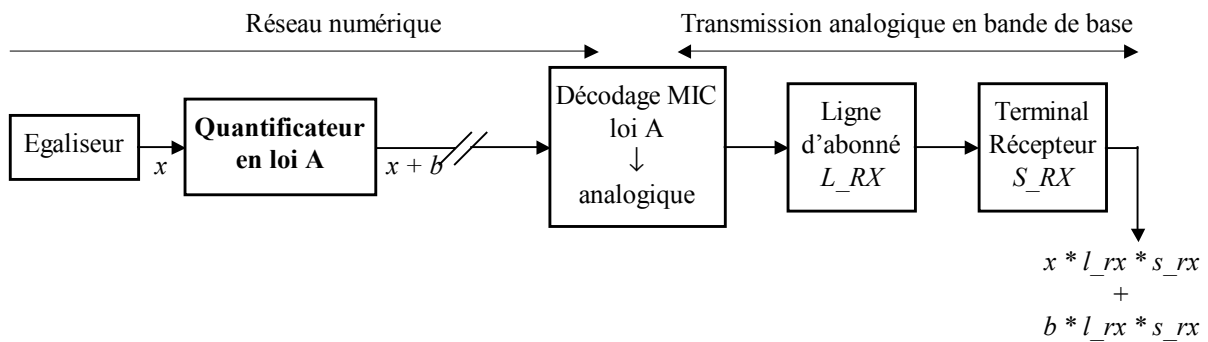


Figure 3.2 : Schéma de la liaison en aval de l'égaliseur

D'autre part, comme c'est le bruit perçu à la réception qui doit être subjectivement réduit, nous ne calculerons pas le seuil de masquage du signal de sortie de l'égaliseur (signal à quantifier), mais celui du signal de réception. Si l'on note x le signal de sortie de l'égaliseur, dans l'équation (3.1), $|X|$ est remplacé par $|X|S_{RX}|L_{RX}|$ (voir Figure 3.2). La densité spectrale du bruit de quantification reçu devra être de la forme :

$$\gamma_b^R(f) = \lambda^2 \cdot \text{Masque}(|L_{RX}(f)||S_{RX}(f)||X(f)|), \tag{3.8}$$

avec λ^2 un facteur inférieur à 1, de telle sorte que le spectre du bruit de quantification reçu soit sous la courbe de masquage. Par ailleurs,

$$\gamma_b^R(f) = |L_{RX}(f)|^2 |S_{RX}(f)|^2 \gamma_b(f), \tag{3.9}$$

où γ_b est la densité spectrale du bruit de quantification b .

Ainsi, le bruit de quantification reformé b devra avoir pour densité spectrale :

$$\gamma_b(f) = \lambda^2 \frac{\text{Masque}(|L_{RX}(f)||S_{RX}(f)||X(f)|)}{|L_{RX}(f)|^2 |S_{RX}(f)|^2} = \lambda^2 \gamma_{\text{Masque}}(f), \quad (3.10)$$

avec :

$$\gamma_{\text{Masque}}(f) = \frac{\text{Masque}(|L_{RX}(f)||S_{RX}(f)||X(f)|)}{|L_{RX}(f)|^2 |S_{RX}(f)|^2}. \quad (3.11)$$

Les dispositifs usuels de reformage spectral du bruit de quantification dans le domaine temporel ont typiquement la structure représentée sur la Figure 3.3 [Boite, 1987][Makhoul, 1979], où $A(z)$ est le polynôme prédicteur du signal s à un ordre donné.

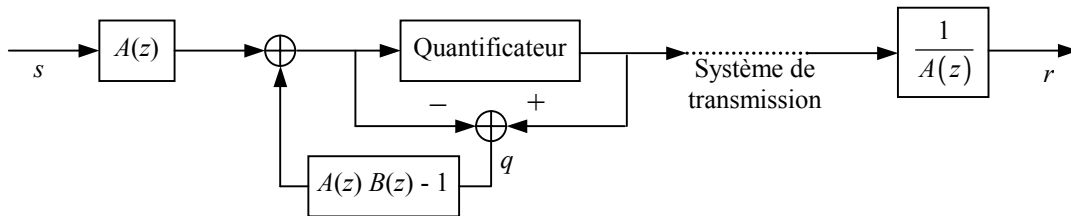


Figure 3.3 : Reformage du bruit de quantification d'un codeur temporel

De cette structure se déduit la relation suivante :

$$R(z) = S(z) + B(z)Q(z) \quad (3.12)$$

Dans [Boite, 1987] et [Makhoul, 1979], il n'est pas fait explicitement référence au masquage, mais l'objectif est de reformer le spectre du bruit selon une forme proche de celle du spectre du signal, de manière à le rendre moins perceptible. A cet effet sont étudiés différents filtres B .

[Makhoul, 1979] montre que le choix de $B = 1$ permet de minimiser le rapport signal à bruit, mais observe que le bruit blanc ainsi obtenu reste très perceptible dans les hautes fréquences. Choisir $B(z)=1/A(z)$, c'est-à-dire supprimer la boucle de reformage du bruit de quantification, conduit à un bruit de spectre parallèle à celui du signal, de niveau élevé, et perçu comme trop "rugueux". L'optimum perceptuel est atteint en donnant au bruit un spectre de forme intermédiaire entre le spectre du signal et celui du bruit blanc. Le filtre B approprié est une approximation tout zéro de $1/A$, d'ordre 2. [Boite, 1987] propose une solution proche, avec $B(z) = A(\gamma z)/A(z)$, où γ est compris entre 0,8 et 0,9.

Plus qu'un reformage du bruit, ces dispositifs réalisent un blanchiment du signal (par le filtre A), dont le spectre est reformé en réception. Dans notre cas, il n'est pas envisageable d'introduire un filtre en réception, de sorte que ces méthodes ne sont pas applicables. Nous proposons de reformer le bruit de quantification sans changer le système de réception, sans blanchir le signal, simplement en utilisant différemment la quantification en loi A après l'égalisation. La section III.2 présente une méthode utilisant un filtrage récursif de l'erreur de quantification. Nous présentons dans la section III.3 une deuxième méthode, fondée sur une approche probabiliste.

III.2. Méthode de réinjection de l'erreur de quantification

III.2.1. Principe

Nous reprenons la structure proposée par [Boite, 1987] et [Makhoul, 1979], en remplaçant $A(z)$ par 1 (la prédiction sur le signal est supprimée). **Il s'agit, selon le schéma de la Figure 3.4, d'injecter à l'entrée du quantificateur l'erreur de quantification filtrée, de telle sorte que le bruit de quantification final soit masqué.**

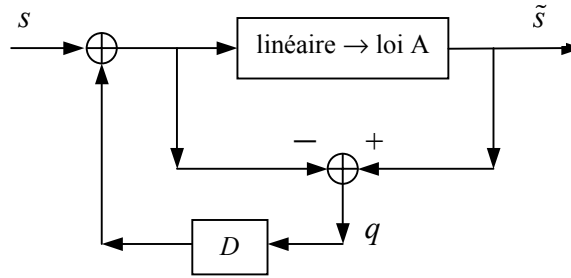


Figure 3.4 : Reformage du bruit de quantification par réinjection de l'erreur

D'après la Figure 3.4, en notant \tilde{s} le signal quantifié en loi A et q l'erreur de quantification,

$$\tilde{S}(z) = S(z) + (1+D(z))Q(z). \quad (3.13)$$

Le bruit de quantification reformé a donc pour densité spectrale de puissance :

$$\gamma_b(f) = |1+D(f)|^2 \sigma_q^2, \quad (3.14)$$

où σ_q^2 désigne la variance de q . D'après l'équation (3.10), le reformage spectral du bruit de quantification doit se traduire par :

$$|1+D(f)|^2 \sigma_q^2 = \lambda^2 \gamma_{\text{Masque}}(f), \quad (3.15)$$

avec λ^2 le rapport entre la densité spectrale de puissance du bruit et le seuil de masquage, que l'on souhaite inférieur à 1. Nous remplaçons cette condition par la condition suffisante :

$$|1+D(z)|^2 \sigma_q^2 = \lambda^2 |H(z)|^2, \quad (3.16)$$

avec H un filtre dont la réponse fréquentielle correspond à la courbe de masquage. Cette égalité est vérifiée dès lors que :

$$D(z) = \frac{\lambda}{\sigma_q} H(z) - 1 \quad (3.17)$$

Comme la boucle doit contenir un retard, $d(0) = 0$. Ainsi, λ est complètement déterminé par :

$$\lambda = \frac{\sigma_q}{h(0)} \quad (3.18)$$

et D est défini par :

$$D(z) = \frac{H(z)}{h(0)} - 1 \quad (3.19)$$

III.2.2. Structure du filtre de boucle

La définition de la structure du filtre doit être guidée par la nécessité de stabilité de la boucle de rétroaction. Le masque changeant à chaque trame de signal, nous cherchons une définition générique du filtre qui garantisse cette stabilité. L'étude de la stabilité selon le lieu des pôles ne peut être réalisée qu'en contournant l'opération de quantification, c'est-à-dire en considérant \tilde{s} non pas comme la sortie du quantificateur, mais comme l'entrée de l'additionneur dont est issu q :

$$Q(z) = \tilde{S}(z) - (S(z) + D(z)Q(z)) \quad (3.20)$$

Ainsi,

$$Q(z) = \frac{\tilde{S}(z) - S(z)}{1 + D(z)} = \left(\frac{h(0)}{H(z)} \right) (\tilde{S}(z) - S(z)) \quad (3.21)$$

Le système sera donc stable si $1/H$ est stable. Si **nous construisons par l'algorithme de Levinson un modèle AR $\{(a_i)_{1 \leq i \leq p}; \sigma\}$ correspondant à l'inverse de la courbe de masquage**, le filtre H , dont la réponse fréquentielle doit suivre la courbe de masquage, peut être choisi tel que :

$$H(z) = \frac{1 + \sum_{i=1}^p a_i z^{-i}}{\sigma}, \quad (3.22)$$

de sorte que :

$$Q(z) = \frac{1}{1 + \sum_{i=1}^p a_i z^{-i}} (\tilde{S}(z) - S(z)) \quad (3.23)$$

Le modèle AR étant stable par construction, on est assuré de la stabilité de la boucle. D est alors un filtre RIF défini, d'après (3.17), par :

$$D(z) = \sum_{i=1}^p a_i z^{-i} \quad (3.24)$$

La structure proposée est représentée sur la Figure 3.5.

L'ordre de modélisation AR de l'inverse du masque doit être choisi assez grand pour approcher avec suffisamment de précision celui-ci, sans atteindre des valeurs qui mettraient en péril la stabilité des algorithmes de calcul du modèle en précision finie. Un ordre de 20 permet d'approcher le masque avec une erreur inférieure à 3 dB, tout en permettant un calcul en virgule fixe selon l'algorithme de Schür [Proakis, 1996], avec une précision de 20 bit après la virgule.

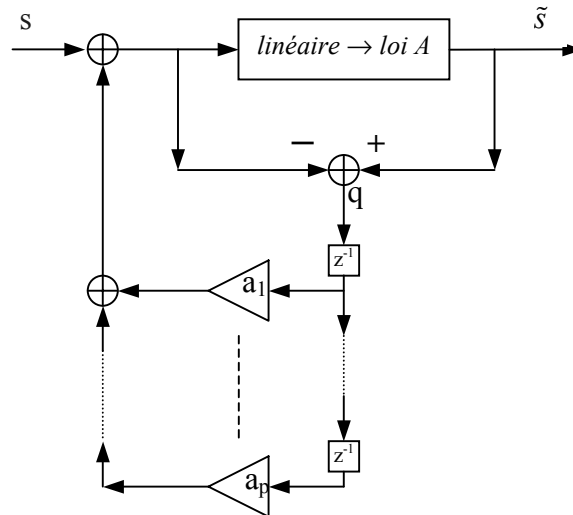


Figure 3.5 : Structure de la boucle de rétroaction

III.2.2. Résultats

Cette méthode est simulée sur des doubles phrases phonétiquement équilibrées [Combesure, 1981] prononcées par quatre locuteurs (masculins) et quatre locutrices. La Figure 3.6 représente, pour deux trames particulières du signal de réception d'une des locutrices, le spectre du bruit de quantification reçu, avec et sans reformage, comparé à la courbe de masquage du signal ainsi qu'à la modélisation MA (filtre H) du masque atténué dans les basses fréquences. Dans les deux cas, le spectre du bruit reformé suit comme prévu la forme du modèle du masque. Pour la première trame, le bruit reformé est bien sous le masque, tandis que le bruit non reformé (blanc) est au-dessus du seuil de masquage entre 1000 et 2000 Hz. En revanche, le masquage échoue pour la seconde trame. L'observation de ces courbes pour les trames successives des signaux de réception des différents locuteurs montre que le spectre du bruit suit toujours la forme fixée, mais que le niveau du bruit par rapport au masque est extrêmement variable, résultat de l'absence de contrôle du paramètre λ dans l'algorithme. C'est pour cette raison que nous avons choisi d'atténuer les sommets du masque. Le dépassement d'un seuil de masquage élevé dans les basses fréquences se traduisait en effet par un bruit "rauque" très gênant. Lorsque le spectre du bruit est défini à partir d'un masque dont les maxima ont été atténués, le bruit est certes masqué moins facilement. Mais dans les cas où le masquage échoue que les maxima aient été atténués ou non, ce bruit non masqué est moins désagréable si la forme de son spectre est celle du masque dont les maxima ont été atténués.

Nous pouvons évaluer de manière objective la capacité de masquage de notre méthode par l'observation de la valeur λ , qui représente l'écart entre le spectre du bruit reformé et la courbe de masquage. Cette valeur est représentée en dB sur la Figure 3.7 pour quatre locuteurs (deux hommes : H1 et H2 ; deux femmes : F1 et F2) prononçant chacun une double phrase. Les zones de stabilité de λ correspondent aux pauses entre les phrases. Idéalement, λ devrait rester

inférieur à 0 dB pour que le bruit de quantification reformé soit masqué. Il apparaît que le bruit dépasse occasionnellement le seuil de masquage, avec une fréquence qui dépend des locuteurs : sur l'ensemble des locuteurs testés, le masquage échoue plus fréquemment pour les locutrices.

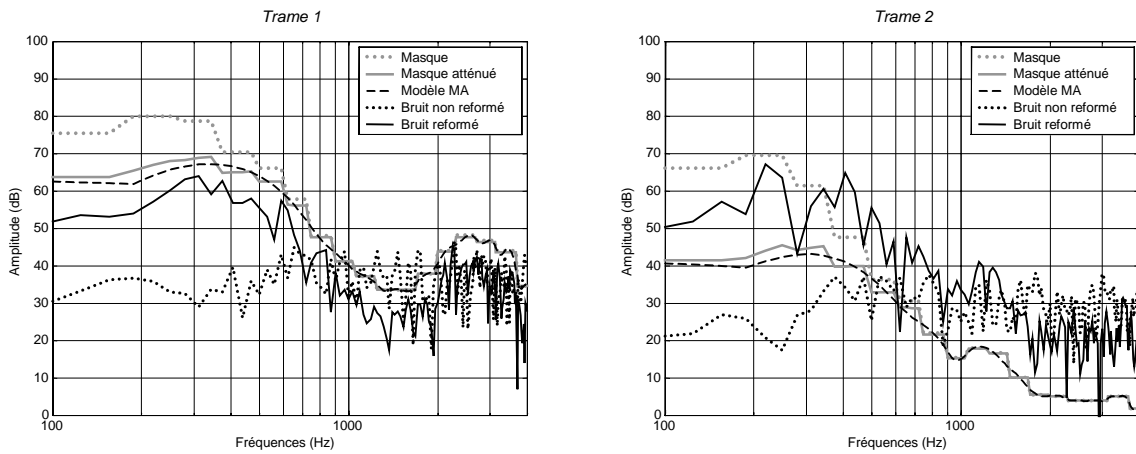


Figure 3.6 : Pour deux trames de signal, comparaison du bruit de quantification au masque

Subjectivement, cette irrégularité du masquage se traduit par un bruit "rauque" apparaissant plus ou moins fréquemment selon les locuteurs. Ce bruit est moins souvent audible que le bruit de quantification non reformé, mais est plus désagréable. La préférence des auditeurs pour l'un ou l'autre sera évaluée dans la section III.5. Notons que ce bruit apparaît sur les mêmes phonèmes que ceux pour lesquels le bruit non reformé est le plus audible.

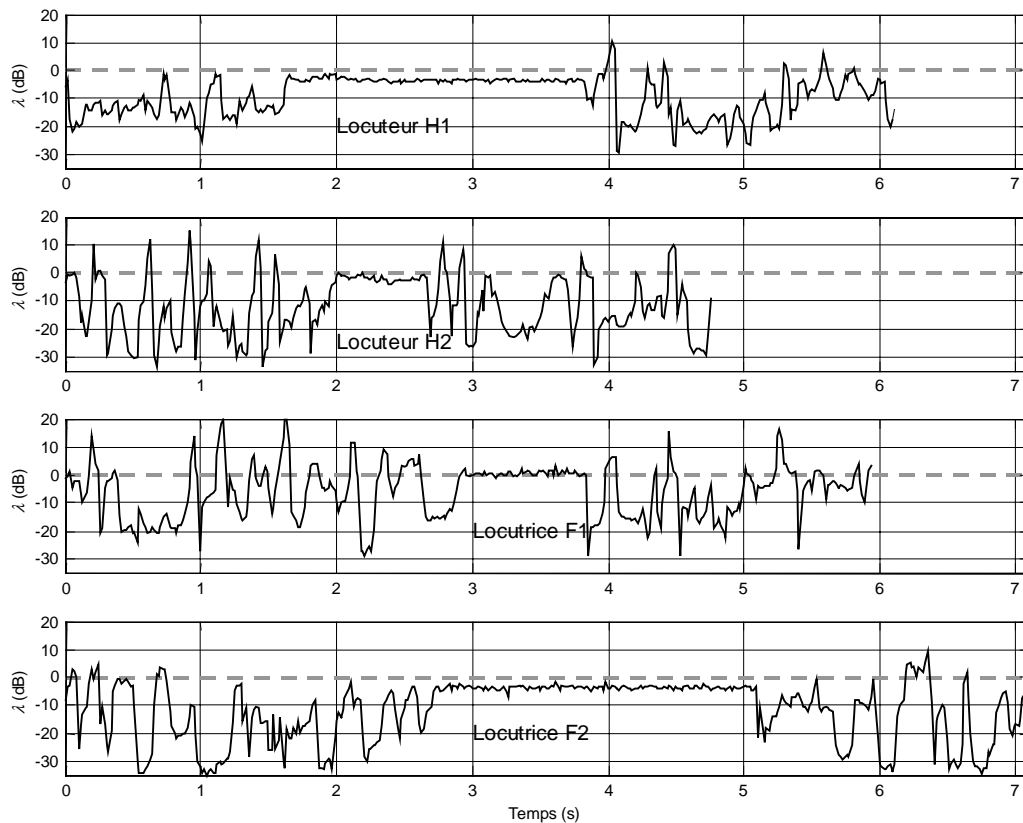


Figure 3.7 : Ecart entre le spectre du bruit reformé et le modèle du masque

III.3. Méthode probabiliste

III.3.1. Principes

Au lieu de quantifier chaque échantillon par le niveau de quantification le plus proche, nous proposons une méthode originale [Mahé, 2002] consistant à quantifier globalement une séquence d'échantillons, *a priori* infinie, de manière à ce que le spectre du bruit de quantification respecte la condition de masquage définie par l'équation (3.10). Cette quantification est effectuée selon une approche probabiliste : il s'agit de chercher la séquence d'échantillons quantifiés la plus probable connaissant le spectre que doit avoir le bruit de quantification. La quantification devant être effectuée en temps réel avec un retard et une taille mémoire limités, nous verrons plus loin comment effectuer une quantification au fil de l'eau tout en conservant cette approche globale. Nous appellerons *chemin* une séquence d'échantillons quantifiés, *niveau de quantification* une des 256 valeurs de sortie d'un codeur MIC, et *valeur quantifiée* les valeurs linéaires, en entier signé codé sur 16 bit, correspondant à ces niveaux de quantification.

Pour chaque séquence de quantification des instants 0 à n , notée $C(0\dots n)$, la probabilité de ce chemin peut s'exprimer par :

$$P(C(0\dots n)) = P(C(0\dots n-1))P(Q(n) | C(0\dots n-1)) \quad (3.25)$$

avec $Q(n)$ la valeur quantifiée à l'instant n . On progresse ainsi de proche en proche : connaissant la probabilité de chaque chemin à l'instant $n-1$, on en déduit, pour chaque valeur de quantification $Q(n)$ possible à l'instant n , les probabilités respectives des chemins composés des chemins précédents concaténés aux différents $Q(n)$.

Si nous considérons la quantification comme l'ajout d'un bruit b ayant le spectre souhaité, la probabilité conditionnelle de quantifier un échantillon $x(n)$ par une valeur quantifiée Q_k est définie par :

$$\begin{aligned} &P(Q(n) = Q_k | C(0\dots n-1), \text{spectre de } b) \\ &= P(S_k < x(n) + b(n) < S_{k+1} | C(0\dots n-1), \text{spectre de } b) \end{aligned} \quad (3.26)$$

où S_k et S_{k+1} sont les seuils inférieur et supérieur associés à Q_k .

Le spectre de b peut être défini par un modèle ARMA :

$$b(n) = w(n) - \sum_{i=1}^p a_i b(n-i) + \sum_{j=1}^q d_j w(n-j) \quad (3.27)$$

avec w bruit blanc centré de variance σ^2 , p et q les ordres de modélisation AR et MA, respectivement. Ainsi, connaissant $C(0\dots n-1)$, $x(n)+b(n)$ est une variable aléatoire de même distribution que $w(n)$ autour de la valeur moyenne :

$$x(n) - \sum_{i=1}^p a_i b(n-i) + \sum_{j=1}^q d_j w(n-j) \quad (3.28)$$

Connaissant la densité de probabilité de w , on peut donc calculer la probabilité (3.26), et déterminer, selon l'équation (3.25), la probabilité de toute séquence quantifiée connaissant le spectre souhaité pour le bruit de quantification.

Il n'est pas possible en pratique de calculer les probabilités de tous les chemins passant par les 256 valeurs de quantification : cela supposerait, pour N échantillons, de considérer 256^N chemins. On sélectionne donc les chemins possibles selon un algorithme de type Viterbi.

En notant \circ l'opération de concaténation entre deux chemins, pour toute suite $C(n+1\dots N)$ du chemin $C(0\dots n)$,

$$P(C(0\dots n) \circ C(n+1\dots N)) = P(C(0\dots n))P(C(n+1\dots N) | C(0\dots n)) \quad (3.29)$$

Or, d'après ce qui précède (notamment les équations (3.26) et (3.28)), le deuxième facteur du membre de droite de l'équation (3.29) ne dépend que des échantillons $x(i)$ et des valeurs quantifiées $Q(i)$ postérieurs à l'instant $n-L$, avec $L = \max(p, q)$. Par conséquent, pour tous les chemins finissant par le même sous-chemin $C(n-L+1\dots n)$, ce facteur est le même. On ne garde donc, à l'instant n , pour chaque sous-chemin $C(n-L+1\dots n)$, que le chemin de plus forte probabilité $P(C(0\dots n))$ finissant par ce sous-chemin. L'algorithme nécessite ainsi, dans une première approche, de mémoriser et actualiser à chaque échantillon 256^L chemins, avec les probabilités et bruits correspondants.

III.3.2. Mise en œuvre

- **Modélisation du bruit de quantification**

D'après l'équation (3.10), la densité spectrale de puissance du bruit doit être proportionnelle à celle de γ_{Masque} :

$$\gamma_b(f) = \lambda_0^2 \gamma_{\text{Masque}}(f), \quad (3.30)$$

où γ_{Masque} est la courbe de masquage du signal de réception divisée par la réponse fréquentielle de la partie de la liaison en aval du quantificateur. Par conséquent, γ_{Masque} est approchée par un modèle ARMA, qui sera celui du bruit au facteur λ_0 près. L'ordre de modélisation doit être le plus faible possible, de manière à limiter la complexité de l'algorithme de recherche du chemin optimal. D'après nos observations, pour les phonèmes les plus critiques en terme de bruit de quantification, γ_{Masque} a une pente descendante et possède deux maxima locaux. Par conséquent, un modèle ARMA d'ordres $p=5$ et $q=4$ permet de modéliser ce masque avec un bon compromis précision-complexité.

L'algorithme proposé présente l'inconvénient de ne pas chercher le bruit de variance minimale connaissant sa forme spectrale, mais le bruit le plus probable, pour une forme spectrale et une variance fixées. La variance doit donc être fixée, à quelques dB sous celle du modèle ARMA du masque. Nous choisissons une marge de 5 dB, ie : $20\log(\lambda_0) = -5$ dB.

Enfin, les calculs mis en œuvre nécessitent de connaître le type de distribution de w . Ce bruit n'ayant pas la réalité physique d'un bruit de quantification et ne constituant qu'un paramètre de l'algorithme, nous sommes libres de choisir sa distribution. Nous choisissons pour w une distribution gaussienne.

Ainsi, la probabilité de l'équation (3.26) s'écrit :

$$P(Q(n) = Q_k | C(0..n-1), \gamma_b) = \frac{1}{\sigma\sqrt{2\Pi}} \int_{S_k}^{S_{k+1}} \exp \left[-\frac{\left(q - x(n) + \sum_{i=1}^p a_i b_C(n-i) - \sum_{j=1}^q d_j w_C(n-j) \right)^2}{2\sigma^2} \right] dq \quad (3.31)$$

avec w_C et b_C les bruits w et b associés au chemin C . Cette probabilité est illustrée sur la Figure 3.8. L'observation des valeurs effectives de w pour une séquence quantifiée montre bien, *a posteriori*, que ce bruit présente une distribution gaussienne pour chaque trame de signal.

• **Simplifications**

En tenant compte de la forte décroissance de la densité de probabilité gaussienne autour de sa moyenne, on peut simplifier la recherche des chemins, en termes de mémoire et de nombre de calculs de probabilités. Pour un chemin donné $C(0..n-1)$, nous ne considérerons pas les 256 terminaisons possibles $Q(n)$, mais uniquement K valeurs quantifiées autour du centre de la gaussienne, comme indiqué pour un chemin sur la Figure 3.8, pour $K = 4$, valeur que nous utiliserons dans les simulations. On remplace ainsi le treillis à 256 états par un arbre dont chaque nœud donne naissance à K branches.

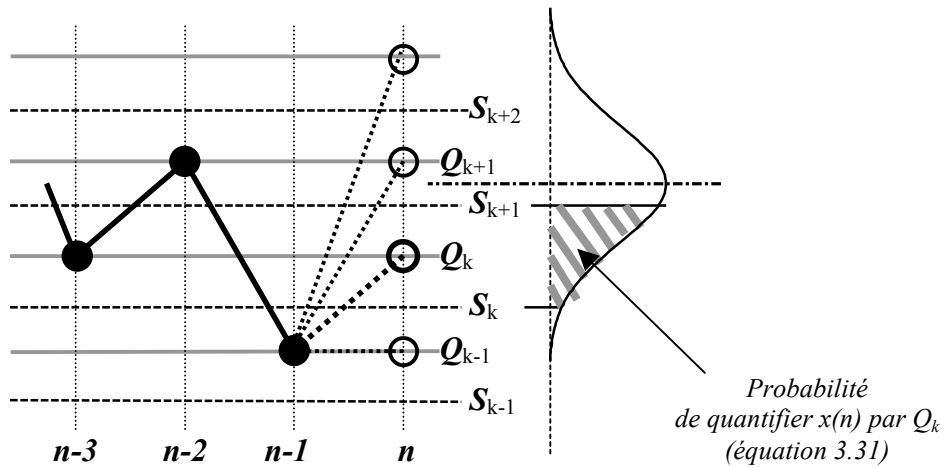


Figure 3.8 : Terminaisons d'un chemin

Par ailleurs, le calcul de l'intégrale de l'équation (3.31) peut être simplifié en approchant la densité de probabilité sur chaque intervalle $[S_k; S_{k+1}]$ par sa valeur en Q_k :

$$P(Q(n) = Q_k | C(0..n-1), \gamma_b) = \frac{S_{k+1} - S_k}{\sigma\sqrt{2\Pi}} \exp \left[-\frac{\left(Q_k - x(n) + \sum_{i=1}^p a_i b_C(n-i) - \sum_{j=1}^q d_j w_C(n-j) \right)^2}{2\sigma^2} \right] \quad (3.32)$$

- **Délai de décision**

Dans les algorithmes de Viterbi utilisés en communications numériques pour décoder les codes convolutifs, les chemins conservés convergent, en remontant le temps de quelques échantillons, vers un même chemin. En d'autres termes, tous les chemins $C(0..n)$ possèdent le même sous-chemin $C(0..n-M)$. Pour un code de rendement $1/2$, par exemple, il est établi empiriquement que M vaut 5 à 6 fois la longueur de contrainte du code [Glavieux, 1996]. Un tel fonctionnement permettrait, dans notre cas, d'attribuer à chaque échantillon une valeur de quantification unique avec un retard M .

L'expérience montre qu'il n'en est pas ainsi : des groupes de chemins aux origines distinctes peuvent survivre longtemps, comme l'illustre la Figure 3.9. Cela peut s'expliquer par la non-stationnarité du signal de parole : du fait du changement de modèle ARMA à chaque trame, un chemin qui aurait dépéri avec un modèle constant peut reprendre de la vigueur si le nouveau modèle lui est favorable. L'application de l'algorithme sur une phrase complète montre certes une convergence vers un chemin de quantification unique après cessation de l'activité vocale, mais il n'est évidemment pas envisageable, dans une application en temps réel, d'imposer un tel délai de décision dans la quantification.

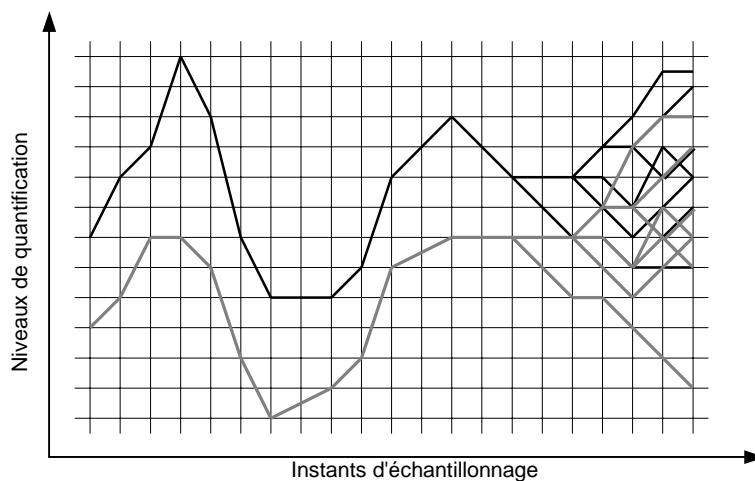


Figure 3.9 : *Persistence de chemins de quantification multiples*

Pour limiter le retard dû à l'algorithme de quantification, le délai de décision M est donc fixé de manière arbitraire, et l'arbre des chemins est élagué de la manière suivante. A chaque instant n , on ne conserve que les chemins passant par la valeur de quantification $Q(n-M)$ qui appartenait, lorsque $n-M$ était l'instant discret courant, au chemin de plus forte probabilité. Cette valeur est alors choisie comme valeur de quantification de $x(n-M)$. Dans une première approche, nous fixons M à 128, ce qui correspond à un délai de décision de 16 ms.

- **Réduction du nombre de chemins**

Malgré toutes les simplifications introduites dans l'algorithme, le nombre de chemins à conserver en mémoire dépasse rapidement la capacité des machines utilisées pour les simulations. Nous avons donc introduit une simplification supplémentaire, source de sous-optimalité dans la sélection du meilleur chemin, mais nécessaire : à chaque échantillon, parmi tous les chemins conservés selon la procédure décrite ci-dessus, seuls les N plus probables sont conservés. Nous choisissons ici $N = 500$.

• **Algorithme**

Chaque chemin est représenté par le vecteur de ses M dernières valeurs quantifiées. A chaque vecteur chemin correspondent le vecteur de ses probabilités successives et les vecteurs des p (respectivement q) derniers échantillons du bruit b (respectivement w) associé. L'algorithme se déroule selon les étapes suivantes :

Pour chaque échantillon $x(n)$,

- 1) Si changement de trame, calcul du modèle ARMA $\{(a_i)_{1 \leq i \leq 5}; (b_j)_{1 \leq j \leq 4}; \sigma_{\text{masque}}\}$ approchant le masque de la nouvelle trame et, b étant défini par l'équation (3.27), détermination de la variance de w :

$$\sigma^2 = \lambda_0^2 \sigma_{\text{masque}}^2 \quad (3.33)$$

- 2) Le 1^{er} élément de tous les chemins conservés est le même : $Q(n-M)$ prend cette valeur.
- 3) Décalage des vecteurs chemins et probabilités.
- 4) Pour chaque chemin,
 - Définition du centre de la densité de probabilité de $x(n)+b(n)$, selon (3.28).
 - Construction des K nouveaux chemins en complétant le chemin par les K valeurs de quantification autour de ce centre.
 - Calcul des probabilités de ces chemins, selon (3.31) (ou (3.32))
 - Calcul des nouveaux échantillons des bruits associés à ces chemins :

$$b(n) = Q(n) - x(n) \quad (3.34)$$

$$w(n) = \sum_{i=0}^p a_i b(n-i) - \sum_{j=1}^q d_j w(n-j) \quad (3.35)$$

- 5) Ordonnancement des nouveaux chemins selon leur dernier niveau de quantification. Ainsi, les chemins sont toujours classés selon le dernier niveau, puis l'avant-dernier, *etc.*
- 6) Pour chaque sous-chemin ($Q(n-L+1) \dots Q(n)$), sélection du chemin de plus forte probabilité se finissant par ces valeurs et élimination des autres.
- 7) Élimination des vecteurs chemins dont la 1^{ère} valeur $Q(n-M+1)$ n'appartenait pas, lorsqu'elle était valeur courante, au chemin de plus forte probabilité.
- 8) Sélection, parmi ces chemins, des N chemins les plus probables.

III.3.3. Résultats

Cette méthode est simulée sur les mêmes doubles phrases phonétiquement équilibrées que précédemment. La Figure 3.10 représente, pour les deux mêmes trames du signal de réception que celles considérées sur la Figure 3.6, le spectre du bruit de quantification reçu, avec et sans reformage, comparé à la courbe de masquage du signal ainsi qu'à la modélisation ARMA du masque atténué dans les basses fréquences. De manière générale, les résultats sont similaires à

ceux de la méthode de réinjection du bruit de quantification : le spectre du bruit reformé suit bien la forme du modèle du masque mais son niveau par rapport au masque est très variable. Ainsi, la variance σ^2 du modèle du bruit n'est pas respectée, l'algorithme fixant celle-ci de manière non contrôlée, comme dans la méthode précédente.

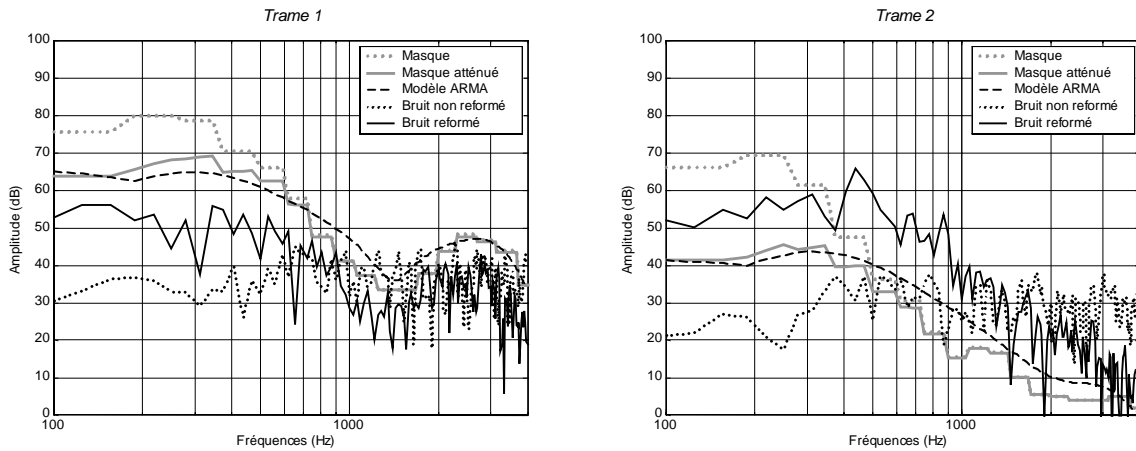


Figure 3.10 : Pour deux trames de signal, comparaison du bruit de quantification au masque

La capacité de masquage de cet algorithme peut être évaluée par la valeur λ de l'équation (3.10), finalement différente du paramètre λ_0 (constante de -5 dB) initialement imposé. En reprenant les notations de la section précédente, pour chaque trame de signal,

$$\lambda = \frac{\sigma_{w_c}}{\sigma_{\text{masque}}},$$

avec w_c le bruit w associé au chemin retenu, calculé selon l'équation (3.35).

Cette valeur λ est représentée en dB sur la Figure 3.11 pour les quatre mêmes doubles phrases que celles de la Figure 3.7, en parallèle avec la valeur λ obtenue par la méthode de réinjection du bruit. Selon cette mesure, les performances des deux méthodes sont donc assez proches. Notons toutefois que sur les parties les plus critiques, à savoir les pics de λ , la méthode probabiliste limite le dépassement du seuil de masquage (voir notamment la locutrice F1). La méthode de réinjection n'est meilleure que pour les valeurs les plus faibles de λ , ce qui ne présente pas d'intérêt puisque ces valeurs correspondent dans les deux cas à un bruit masqué.

Subjectivement, le même bruit "rauque" apparaît lorsque le masquage échoue. Par ailleurs, pour certains locuteurs, les signaux sont affectés d'un léger bruit musical haute fréquence.

III.3.4. Influence des paramètres de l'algorithme

Les résultats qui précèdent ont été obtenus avec les valeurs de paramètres et les simplifications de l'algorithme proposées dans la section III.3.2, dont le choix peut être source de sous-optimalité :

- limitation du nombre de chemins conservés à 500 ;
- prise de décision avec un délai de 16 ms ;
- calcul simplifié de l'intégrale ;

- limitation du nombre de terminaisons d'un chemin à 4 ;
- choix de $\lambda_0 = -5$ dB.

Nous examinons l'effet du choix de ces paramètres sur la capacité de masquage de l'algorithme, mesurée par la valeur λ . Compte-tenu de la lenteur de l'algorithme, cette étude a été effectuée uniquement pour les locutrices, pour lesquelles le dépassement du seuil de masquage est le plus sensible. Nous en présentons graphiquement les résultats pour la locutrice F1, sur la première phrase de la double phrase traitée précédemment. Les résultats sont similaires pour la locutrice F2.

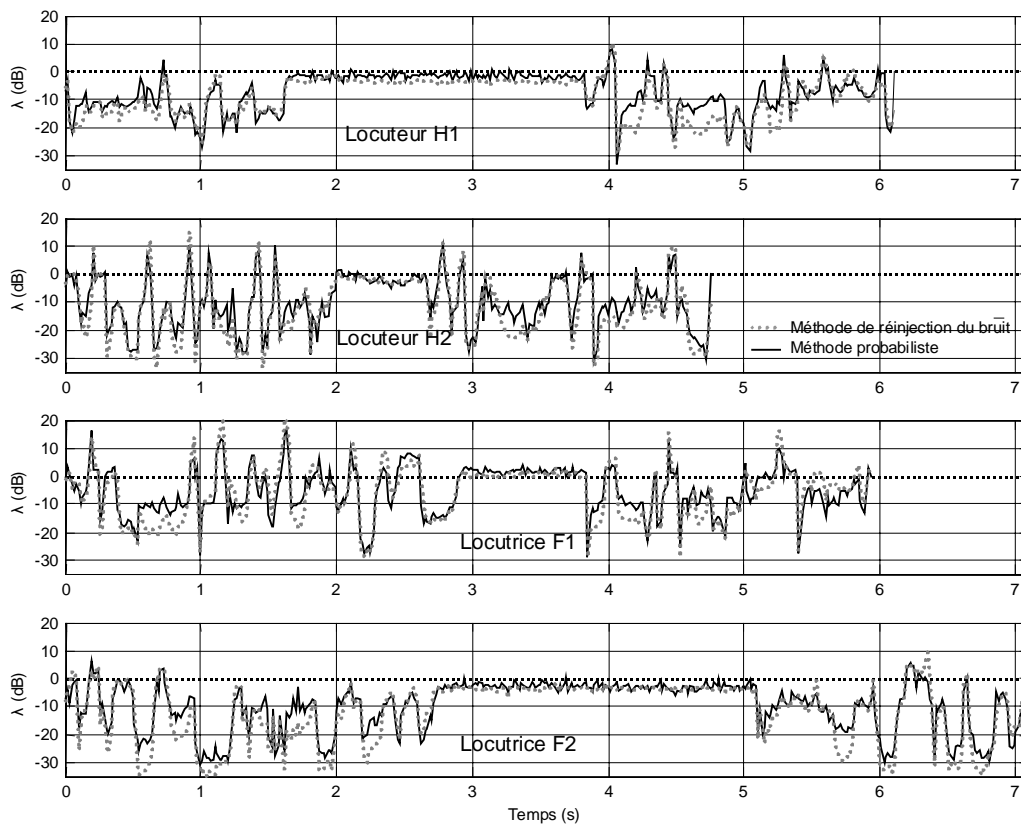


Figure 3.11 : Ecart entre le spectre du bruit reformé et le modèle du masque

- **Effet du nombre maximum de chemins N**

La limitation du nombre maximum de chemins N est *a priori*, avec celle du délai de décision, la principale source de sous-optimalité, le choix des autres paramètres ne constituant que des approximations (calcul de l'intégrale, nombre de terminaisons d'un chemin) ou ne semblant pas avoir d'influence sur λ (choix de λ_0).

La Figure 3.12 présente l'évolution de λ selon que N est fixé à 500 ou 1000, ainsi que la différence entre la valeur λ pour $N=1000$ et la valeur λ pour $N=500$. Globalement, cette différence fluctue autour de zéro, mais présente des chutes dans les parties critiques, c'est-à-dire pour les fortes valeurs de λ . L'augmentation de N permet donc d'améliorer le masquage, mais cette amélioration, obtenue au prix d'un doublement de la complexité de l'algorithme, reste

limitée à moins de 4 dB, ce qui maintient le bruit très au dessus du seuil de masquage dans les parties les plus bruitées. Subjectivement, l'amélioration est imperceptible.

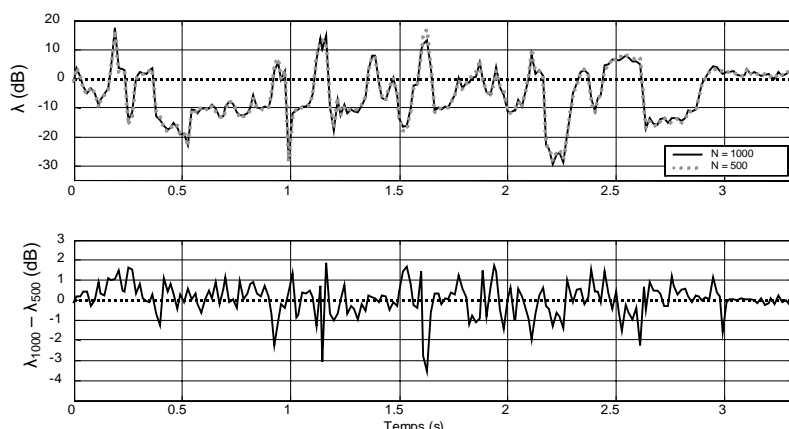


Figure 3.12 : λ pour $N = 500$ ou 1000 ; écart entre les valeurs de λ pour ces deux valeurs de N

- Effet du délai de décision

La Figure 3.13 présente les pourcentages d'abandons à l'étape 7 de l'algorithme, selon que le délai de décision d est fixé à 16 ou 32 ms. Même avec un délai de 16 ms, les abandons restent peu fréquents et isolés. L'évolution de λ dans ces deux cas, ainsi que la différence entre λ pour $d = 32$ ms et λ pour $d = 16$ ms, sont illustrés sur la Figure 3.14. L'allongement du délai permet de réduire légèrement la valeur de λ lors des plus forts dépassements du seuil de masquage, mais λ reste élevé dans ces parties du signal. Par ailleurs, ces 32 ms de retard de décision s'ajoutent aux 16 ms de retard dues à l'analyse du signal par trames (trames de 32 ms se recouvrant à 50 %). Il en résulte un retard inacceptable pour une communication téléphonique.

Les mêmes résultats sont illustrés sur les Figures 3.15 et 3.16 pour $d = 16$ ou 8 ms. La réduction du délai de décision induit de nombreux abandons de chemins à l'étape 7 de l'algorithme, mais cela n'affecte pas les performances de l'algorithme : lorsque $\lambda > 0$ avec $d = 16$ ms (bruit non masqué), la valeur de λ pour $d = 8$ ms n'augmente pas, voire diminue.

Pour ces trois valeurs du délai de décision, 8, 16 et 32 ms, les écoutes informelles ne font apparaître aucune différence entre les signaux traités.

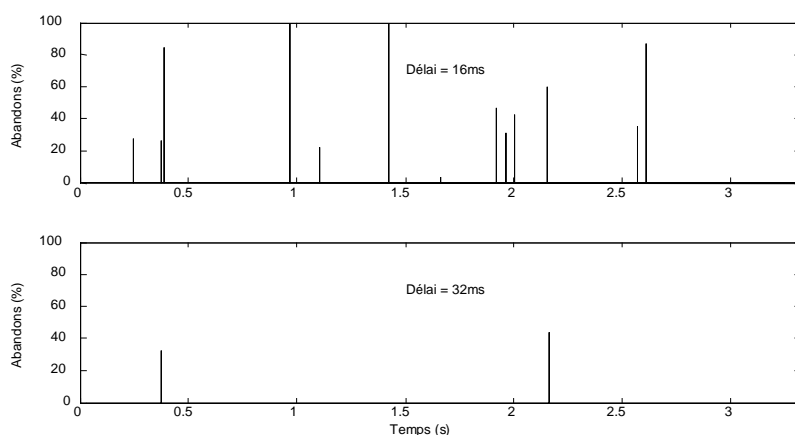


Figure 3.13 : Pourcentages de chemins abandonnés à l'étape 7 de l'algorithme

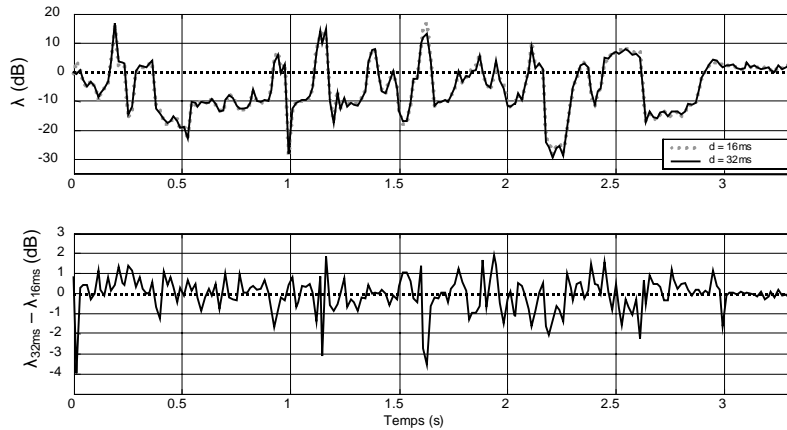


Figure 3.14 : λ pour $d = 16$ ou 32 ms ; écart entre les valeurs de λ pour ces deux valeurs de d

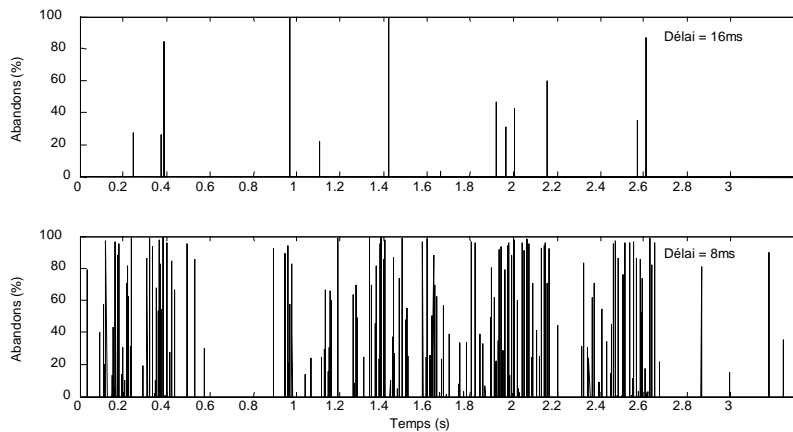


Figure 3.15 : Pourcentages de chemins abandonnés à l'étape 7 de l'algorithme

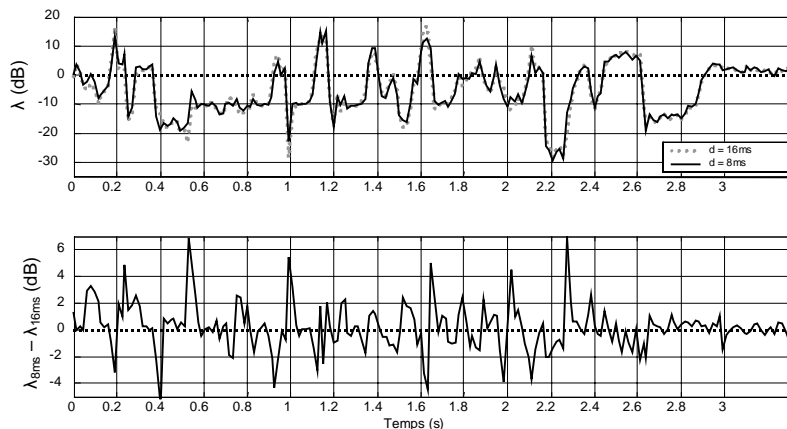


Figure 3.16 : λ pour $d = 16$ ou 8 ms ; écart entre les valeurs de λ pour ces deux valeurs de d

- Effet du calcul approché de l'intégrale

Le calcul de l'intégrale (3.31) selon la méthode des trapèzes, avec un pas de 1 (pour une dynamique maximale de 2^{16}), plutôt que selon la formule simplifiée (3.32) ne permet pas de

réduire la valeur de λ . Celle-ci est même le plus souvent supérieure, y compris dans les zones de dépassement du masque.

- **Effet du nombre de terminaisons des chemins**

La Figure 3.17 présente l'évolution de λ selon que le nombre de terminaisons K de chaque chemin est fixé à 4 ou 8, ainsi que la différence entre la valeur λ pour $K = 8$ et la valeur λ pour $K = 4$. L'augmentation de K permet de réduire λ dans les parties critiques, mais cette amélioration est limitée et subjectivement imperceptible.

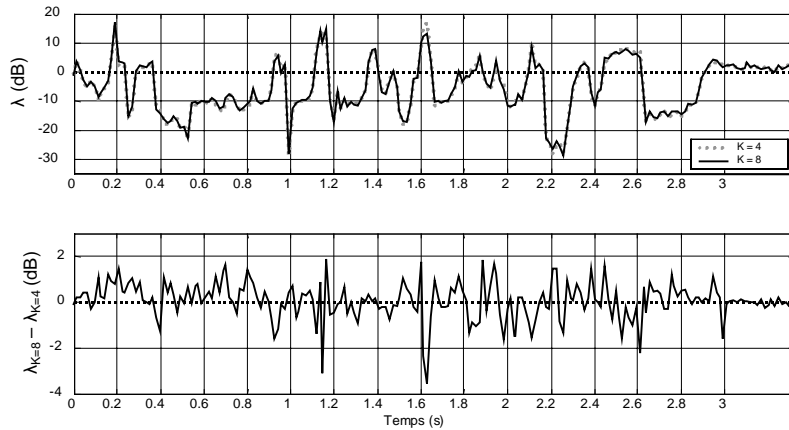


Figure 3.17 : λ pour $K = 4$ ou 8 ; écart entre les valeurs de λ pour ces deux valeurs de K

- **Effet de λ_0**

Nous avons observé que la valeur effective de λ est apparemment indépendante du paramètre λ_0 fixé dans l'algorithme. La Figure 3.18 représente la valeur λ obtenue en fixant λ_0 à -10 dB, comparée à celle obtenue précédemment avec $\lambda_0 = -5$ dB. Il ne résulte de cet abaissement de λ_0 aucune amélioration du masquage dans les parties les plus critiques du signal. Les seules parties où λ est nettement réduit (jusqu'à 6 dB) sont celles où λ était déjà faible.

Ces résultats peuvent s'expliquer de la manière suivante. L'algorithme cherche le chemin le plus probable connaissant une variance de bruit donnée. S'il existe des chemins tels que la variance du bruit correspond à $\lambda = -10$ dB, ces derniers seront sélectionnés si l'on a fixé λ_0 à -10 dB et ne le seront pas nécessairement si l'on a fixé λ_0 à -5 dB, puisqu'il existera *a fortiori* des chemins dont le bruit a une variance correspondant à $\lambda = -5$ dB. Ainsi les valeurs de λ seront-elles plus faibles dans ces cas pour $\lambda_0 = -10$ dB. En revanche, s'il n'existe pas de chemin tel que la variance du bruit corresponde à une valeur négative (en dB) de λ , fixer λ_0 à -5 dB ou -10 dB changera peu la variance effective du chemin choisi : le chemin "le plus probable connaissant λ_0 " sera choisi et sera dans les deux cas peu probable.

La Figure 3.19 représente la valeur λ obtenue en fixant λ_0 à 0 dB, comparée à celle obtenue avec $\lambda_0 = -5$ dB. Comme précédemment, l'abaissement de λ_0 réduit la valeur de λ uniquement dans les zones où celle-ci était déjà faible. Dans les zones de dépassement du seuil de masquage ($\lambda > 0$ dB), la valeur de λ est plus faible pour $\lambda_0 = 0$ dB. Il peut donc être néfaste de fixer une valeur de λ_0 trop faible. S'il n'existe pas de chemin respectant cette valeur, l'algorithme choisit le chemin le plus probable parmi des chemins de faible probabilité, ce qui peut conduire à un choix sous-optimal en termes de variance du bruit associé.

A l'écoute, aucune différence n'est perceptible entre les signaux de réception correspondant à ces trois valeurs de λ_0 .

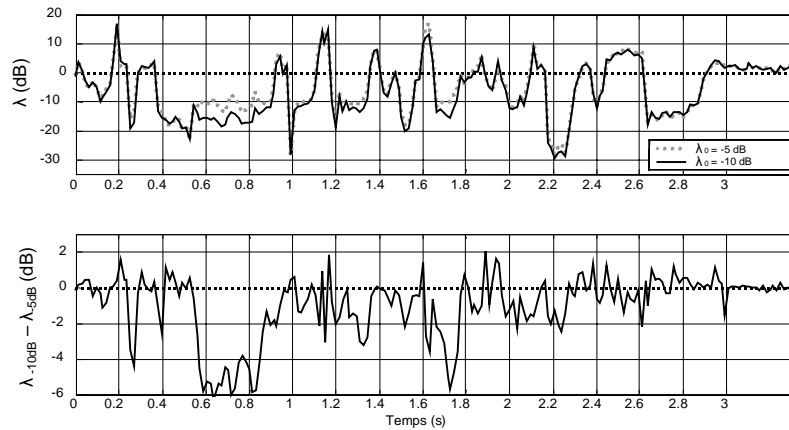


Figure 3.18 : λ pour $\lambda_0 = -5$ ou -10 dB ; écart entre les valeurs de λ pour ces deux valeurs de λ_0

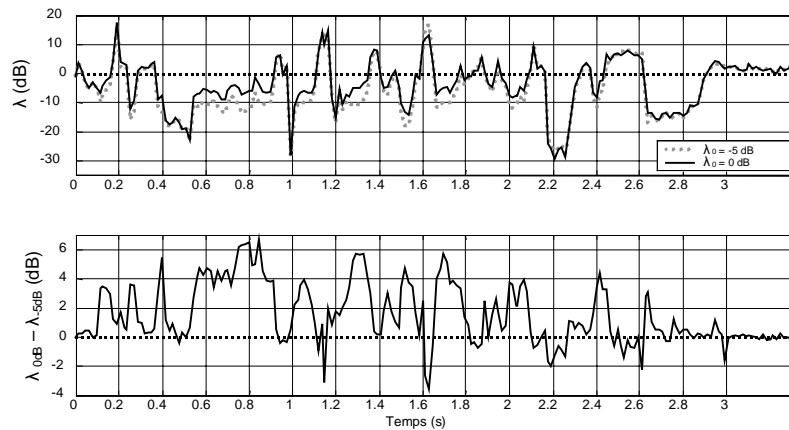


Figure 3.19 : λ pour $\lambda_0 = -5$ ou 0 dB ; écart entre les valeurs de λ pour ces deux valeurs de λ_0

• Conclusion

Au vu des résultats qui précèdent, l'algorithme proposé est peu sensible aux variations de ses paramètres. Retenons toutefois que l'augmentation du nombre de chemins maximal N est susceptible de réduire les valeurs les plus élevées de λ (mais augmente d'autant la complexité) et qu'il est préférable de ne pas chercher à donner à λ_0 une valeur trop faible.

III.4. Comparaison des deux méthodes

III.4.1. Complexité

La méthode de réinjection du bruit de quantification présente l'avantage d'être peu complexe. Elle nécessite de calculer à chaque trame le seuil de masquage et le modèle AR de l'inverse de celui-ci, et met en œuvre un filtre à 20 coefficients.

La méthode probabiliste est en revanche d'une complexité qui peut être rédhibitoire pour une application en temps réel. Outre le calcul du masque et de son modèle ARMA à chaque trame, cette méthode implique, à chaque échantillon, de :

- calculer la probabilité de chacun des KN nouveaux chemins, soit 2000 probabilités avec les valeurs proposées ;
- réordonner 1 fois (étape 5) et sélectionner 3 fois (étapes 6, 7 et 8) les différents chemins conservés, ainsi que les probabilités et les séquences de bruit associées, ces valeurs étant stockées dans des matrices de dimensions respectives $\alpha N \times M$, $\alpha N \times M$, $\alpha N \times 5$ (pour b) et $\alpha N \times 4$ (pour w). La valeur α est un facteur égal à K à la première sélection et décroissant au fil des sélections suivantes.

III.4.2. Performances de masquage

Nous ne considérerons ici que la performance objective mesurée par la valeur λ . D'un point de vue perceptif, le bruit "rauque" occasionnel est le même pour les deux méthodes, mais la méthode probabiliste est, en l'état actuel de son étude, disqualifiée par le léger bruit musical haute fréquence qu'elle induit pour certaines locutrices, et dont nous ne sommes jusqu'à présent pas parvenus à trouver l'origine.

La méthode de réinjection du bruit est *a priori* sous-optimale, dans la mesure où la minimisation de λ , c'est-à-dire la maximisation de $h(0)$ d'après l'équation (3.18), est contrainte par la stabilité de la boucle. De fait, dans l'algorithme proposé, λ n'est pas contrôlé, $h(0)$ étant entièrement déterminé par la variance du modèle AR de l'inverse du masque.

La méthode probabiliste n'est pas limitée par cette contrainte, ce qui permet *a priori* d'atteindre la valeur de λ la plus faible possible. La définition du critère de recherche du chemin optimal de quantification nous écarte cependant de cette potentialité. L'algorithme ne cherche en effet pas la minimisation de la variance du bruit connaissant la forme de son spectre, mais le chemin de probabilité maximale *sachant un spectre donné de bruit*. La difficulté réside dans le choix de la variance de ce bruit (*i.e.* du paramètre λ_0). Comme nous l'avons vu, fixer une variance très faible est illusoire et conduit à un bruit qui est certes le plus probable, mais sans être pour autant très probable, ce bruit étant au final d'une variance supérieure à ce qu'elle aurait pu être en fixant une valeur de λ_0 plus grande.

Les résultats de la section III.3.4 ont montré que pour les plus hautes valeurs de λ , correspondant à un dépassement du seuil de masquage, le dépassement est moins fort avec la méthode probabiliste. On pourrait objecter que la modélisation du masque n'est pas la même, de sorte que les valeurs de λ ne sont pas tout à fait comparables. Il n'est cependant pas envisageable d'utiliser un modèle MA d'ordre 20 pour la méthode probabiliste, pour des raisons de complexité. D'autre part, l'utilisation, dans la méthode de réinjection du bruit, d'un filtre de boucle RII

correspondant au modèle ARMA d'ordres 5 et 4 peut conduire à des dysfonctionnements : pour certains locuteurs, la sortie du dispositif reste bloquée à une valeur constante. Nous nous proposons de comparer expérimentalement les deux méthodes sur la même base, en nous affranchissant des difficultés liées à la modélisation de la courbe de masquage.

Nous considérons un signal stationnaire x résultant du filtrage d'un bruit blanc w de variance $\sigma_w^2 = 2000^2$ par un filtre RIF d'ordre 2 ayant un zéro double valant 0,99. Les échantillons de x sont définis par :

$$x(n) = w(n) + a_1 w(n-1) + a_2 w(n-2). \quad (3.36)$$

Pour simplifier la modélisation de la courbe de masquage, nous utilisons la méthode d'injection de bruit de Paillard [Paillard, 1992], selon laquelle le spectre du bruit maximum injectable est celui du signal atténué de 13 dB. Ainsi, la courbe de masquage peut être modélisée par H tel que :

$$H(z) = 10^{\frac{-13}{20}} \sigma_w (1 + a_1 z^{-1} + a_2 z^{-2}). \quad (3.37)$$

Pour la méthode de réinjection du bruit, le filtre de boucle est défini comme indiqué dans la section III.2.2. Pour la méthode probabiliste, le bruit b est défini par :

$$b(n) = \lambda_0 \sigma_w (w'(n) + a_1 w'(n-1) + a_2 w'(n-2)), \quad (3.38)$$

avec w' un bruit blanc de variance 1 et λ_0 une constante dont la valeur en dB est inférieure à -13.

La méthode de réinjection du bruit conduit à une valeur de λ de -30 dB environ. La limitation à 2 de l'ordre du modèle du masque permet de simuler la méthode probabiliste de manière optimale, c'est-à-dire sans élimination de chemins à l'étape 8. Si l'on fixe λ_0 à -35 dB, la valeur finale de λ est inférieure de 1 à 2 dB (selon les trames considérées) à celle obtenue avec la méthode de réinjection du bruit. Ainsi, en fixant convenablement la valeur de λ_0 , la méthode probabiliste se révèle plus performante.

III.5. Évaluation de la perception conjointe du bruit et du timbre

III.5.1. Objectifs et méthode

Les deux méthodes de reformage du bruit proposées dans les sections précédentes conduisent à un bruit de quantification équivalent à la fois en termes de spectre et de niveau. Lors des écoutes informelles réalisées, il n'apparaît pas de différence entre les deux bruits, si ce n'est la présence, pour certains locuteurs, d'un léger sifflement dans les hautes fréquences si l'on utilise la méthode probabiliste. Ce sifflement est suffisamment audible pour qu'une comparaison des deux méthodes par des tests formels soit sans objet. Nous évaluerons donc uniquement le bruit reformé par la première méthode (réinjection de l'erreur filtrée à l'entrée du quantificateur).

Le reformage sera évalué par comparaison des signaux en réception d'une liaison égalisée sans et avec reformage du bruit de quantification. Ce bruit étant lié à l'égalisation, ces deux

signaux doivent également être comparés au signal en réception de la même liaison sans égaliseur, afin de déterminer si une voix au timbre corrigé mais affectée d'un bruit de quantification, reformé ou non, est préférable à la voix téléphonique habituelle. **Ainsi seront évalués conjointement l'intérêt du reformage spectral du bruit de quantification et l'intérêt de l'égalisation compte-tenu du bruit qu'elle induit.**

Afin d'alléger les notations, nous notons désormais, pour une liaison et un signal donnés :

- O le signal original ;
- A le signal en réception de la liaison non égalisée ;
- B le signal en réception de la liaison égalisée, sans reformage du bruit ;
- C le signal en réception de la liaison égalisée, avec reformage du bruit selon la première méthode.

Dans le cas où la voix originale du locuteur n'est pas connue de l'auditeur, l'intérêt de l'égalisation n'est pas de rapprocher le timbre de la voix en réception du timbre original du locuteur, mais simplement de lui donner un caractère plus "naturel". La voix originale n'a pas à être présentée aux auditeurs, qui jugent A, B et C en fonction du "naturel" du timbre et de la gêne due au bruit. Un test de comparaison par paires [Bonnet, 1986] suffit donc, dans lequel chaque auditeur donne, pour chaque paire de $\{A,B,C\}$, sa préférence pour l'un ou l'autre des deux éléments.

Dans le cas où la voix originale du locuteur est connue de l'auditeur, la question à résoudre est "quel signal, entre A, B et C, l'auditeur préfère-t-il, connaissant la voix originale O ?". La difficulté à mettre en œuvre un test permettant de répondre à cette question tient à la condition "connaissant la voix originale". Il est en effet peu réaliste d'envisager un test où tous les auditeurs connaîtraient la voix originale de chaque locuteur test. Cette condition doit donc être simulée. Il s'agit alors d'effectuer un test de comparaison de dégradations par paire, c'est-à-dire de comparer la dégradation de X par rapport à O à la dégradation de Y par rapport à O, avec $\{X,Y\}$ une paire de $\{A,B,C\}$.

Cette méthode se heurte cependant au problème de mémoire du sujet si l'on présente la séquence OX OY : après avoir écouté OY, ne risque-t-on pas d'avoir déjà oublié OX ? C'est pourquoi nous proposons de présenter la séquence OX OY OX. Ainsi, l'auditeur :

- effectue un premier jugement sur la dégradation de X par rapport à O lors de la première présentation de OX ;
- juge la dégradation de Y par rapport à O lors de la présentation de OY ;
- effectue une première comparaison entre les deux dégradations, peu sûre car le souvenir de OX s'est estompé.

La deuxième présentation de OX permet alors de conforter ou modifier son jugement. La séquence est également présentée dans l'ordre OY OX OY au cours du test, afin d'annuler un éventuel effet d'ordre.

Une autre solution pourrait être de remplacer cette comparaison par paires par un test de catégories de dégradation [UIT-T/P.800, 1996] : pour chaque $X \in \{A,B,C\}$, on présente au sujet OX et on lui demande de donner une note de dégradation entre 1 et 5, 1 correspondant à une dégradation très gênante, 5 à une dégradation imperceptible. La comparaison des notes moyennes de A, B et C permettrait ainsi de déterminer la préférence "connaissant l'original". Cependant, les dégradations par rapport à l'original étant à la fois assez fortes, comparables en niveau et différentes dans leur forme, il est à craindre que ce test soit peu discriminant. C'est pourquoi nous n'avons pas retenu cette solution.

Effectuer toutes les comparaisons deux à deux entre A, B et C dans les deux tests serait redondant. La connaissance de O ne change en effet *a priori* pas la préférence entre B et C, puisqu'en termes de timbre, ces deux signaux sont identiques, leurs caractéristiques spectrales étant très proches. D'autre part, une fois que le bruit préféré (bruit de quantification blanc ou reformé) aura été déterminé dans le premier test, il n'est pas utile d'évaluer le signal affecté par l'autre bruit dans le deuxième test. Le deuxième test consistera donc simplement à comparer A à B ou C (suivant les résultats du premier test) selon la méthode décrite ci-dessus.

III.5.2. Plan de test

- **Premier test : préférence sans connaissance de l'original**

Nous utilisons un corpus de doubles phrases phonétiquement équilibrées, d'une durée de 8 s environ chacune. Les résultats objectifs et les écoutes informelles témoignant d'une grande sensibilité du niveau de bruit au locuteur et aux phonèmes prononcés, nous souhaitons évaluer les signaux A, B et C pour différents locuteurs, et plusieurs doubles phrases par locuteur. Notons N_{loc} le nombre de locuteurs et N_{phr} le nombre de doubles phrases. Par ailleurs, nous souhaitons mesurer l'influence du bruit de fond sur la perception du bruit de quantification. Nous nous placerons donc dans N_{br} ambiances sonores, dont une est silencieuse.

Pour chaque condition (locuteur, double phrase, bruit de fond), les 3 paires de {A,B,C} doivent être présentées chacune dans les deux sens, de manière à ce que l'effet de l'ordre de présentation n'affecte pas le résultat. Au total, $N_{loc} \times N_{phr} \times N_{br} \times 3 \times 2$ paires doivent être présentées. Le temps de présentation et de notation d'une paire étant de 22 s, les conditions retenues sont les suivantes, de manière à donner au test une durée raisonnable :

- 2 ambiances sonores : silencieuse et brouhaha avec un RSB de 25 dB ;
- 4 locuteurs (2 hommes : H1 et H2 ; 2 femmes : F1 et F2) en ambiance silencieuse, 2 locuteurs (H1 et F1) en ambiance bruitée ;
- 2 doubles phrases par locuteur.

Ces conditions représentent 72 paires à tester, soit 26 mn de test. La limitation à 2 du nombre de locuteurs test en ambiance bruitée se justifie par le fait que le bruit de quantification y est *a priori* moins gênant et, d'après nos écoutes informelles, partiellement masqué autant pour B que pour C.

L'évaluation est effectuée par 24 sujets, dont deux groupes de 8 naïfs et un groupe de 8 experts. Les sujets sont placés dans une salle de test isolée des bruits extérieurs et écoutent les sons avec des casques binauraux. Les consignes reproduites dans l'annexe D sont d'abord lues individuellement, puis répétées oralement par l'expérimentateur. Le test comprend une séance d'apprentissage composée de 8 paires et deux séances de test de 13 mn chacune séparées par une pause de 5 mn.

Au cours de ces deux séances, les 72 paires sont présentées dans un ordre aléatoire différent pour chaque groupe d'auditeurs (de manière à annuler l'effet d'ordre). Après l'écoute de chaque paire, les sujets disposent de 5 s pour indiquer leur préférence en appuyant sur le bouton 1 s'ils préfèrent le premier échantillon, sur le bouton 2 s'ils préfèrent le deuxième échantillon.

- **Deuxième test : préférence connaissant l'original**

Dans ce deuxième test, nous testons les mêmes conditions (locuteur, double phrase, bruit de fond) que précédemment.

Pour chacune de ces conditions sont présentées aux auditeurs les séries de 3 paires OA OD OA et OD OA OD, avec $D = B$ ou C , selon le bruit préféré par les auditeurs du premier test. Les 24 séries sont présentées dans un ordre aléatoire. Les sujets sont informés de la structure paire 1 – paire 2 – paire 1 des séries, une paire étant décrite comme l'échantillon (double-phrase) de référence (O) suivi d'un échantillon traité (A ou D). Après l'écoute de chaque série, les auditeurs ont 5 s pour indiquer dans quelle paire (1 ou 2) la modification de l'échantillon traité par rapport à la référence est la moins gênante. De manière à faciliter le repérage au cours de l'écoute, nous fixons les temps de pause suivants :

- 300 ms entre les deux phrases d'une double phrase ;
- 600 ms entre les deux échantillons d'une paire ;
- 1200 ms entre deux paires.

Le test est effectué dans les mêmes conditions matérielles par trois groupes de 8 sujets, constitués comme précédemment. Les consignes reproduites dans l'annexe E sont d'abord lues individuellement, puis répétées oralement par l'expérimentateur. Ces consignes s'inspirent de l'exemple donné dans [UIT-T/P.800, 1996] pour les tests de catégorie de dégradation. Le test comprend une séance d'apprentissage composée de 6 séries de paires et de deux séances de test de 7 mn chacune (12 séries) séparées par une pause de 3 mn.

III.5.3. Résultats

- **Premier test : préférence sans connaissance de l'original**

Nous rappelons les notations utilisées :

- A le signal en réception de la liaison non égalisée ;
- B le signal en réception de la liaison égalisée, sans reformage du bruit ;
- C le signal en réception de la liaison égalisée, avec reformage du bruit selon la première méthode.

Les figures 3.20 à 3.26 présentent les pourcentages de préférence d'un traitement X par rapport à un traitement Y. Pour chaque comparaison de deux traitements X et Y et pour chaque combinaison (locuteur, bruit de fond) le pourcentage de préférence attribué à X est la proportion de préférences du traitement X dans un ensemble de 96 jugements : 24 auditeurs \times 2 doubles phrases \times 2 sens de présentation de chaque paire. Les résultats détaillés pour chaque combinaison (bruit de fond, locuteur, phrase, ordre de présentation de la paire) figurent en annexe F. Notons que, à deux exceptions près (comparaisons AB pour la locutrice F2 et BC pour le locuteur H2), les résultats ne font pas apparaître de différence significative entre les jugements des auditeurs sur deux phrases d'un même locuteur. Le mode de détermination de la *significativité* de la différence entre deux pourcentages est détaillé dans l'annexe G.

De ces pourcentages de préférences nous déduisons les positions relatives de A, B et C sur une échelle de préférence de Thurstone [Bonnet, 1986], calculées selon la méthode détaillée en annexe H. Ces échelles sont représentées sur les figures 3.20 à 3.26 en regard des pourcentages correspondants.

En parallèle avec ces résultats subjectifs, les Figures 3.20 à 3.25 présentent également l'évolution de la valeur λ , représentative de l'écart entre le spectre du bruit et le masque : une valeur de λ positive correspond à un bruit au dessus du seuil de masquage.

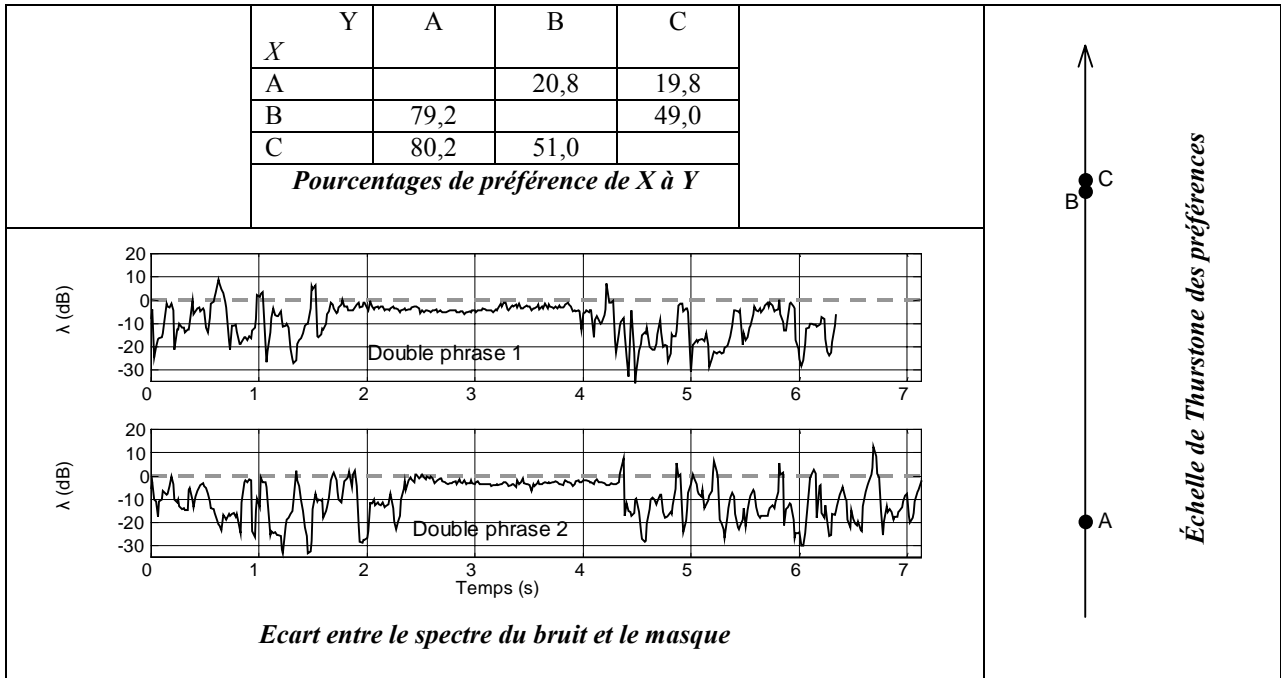


Figure 3.20 : Résultats locuteur H1, ambiance silencieuse

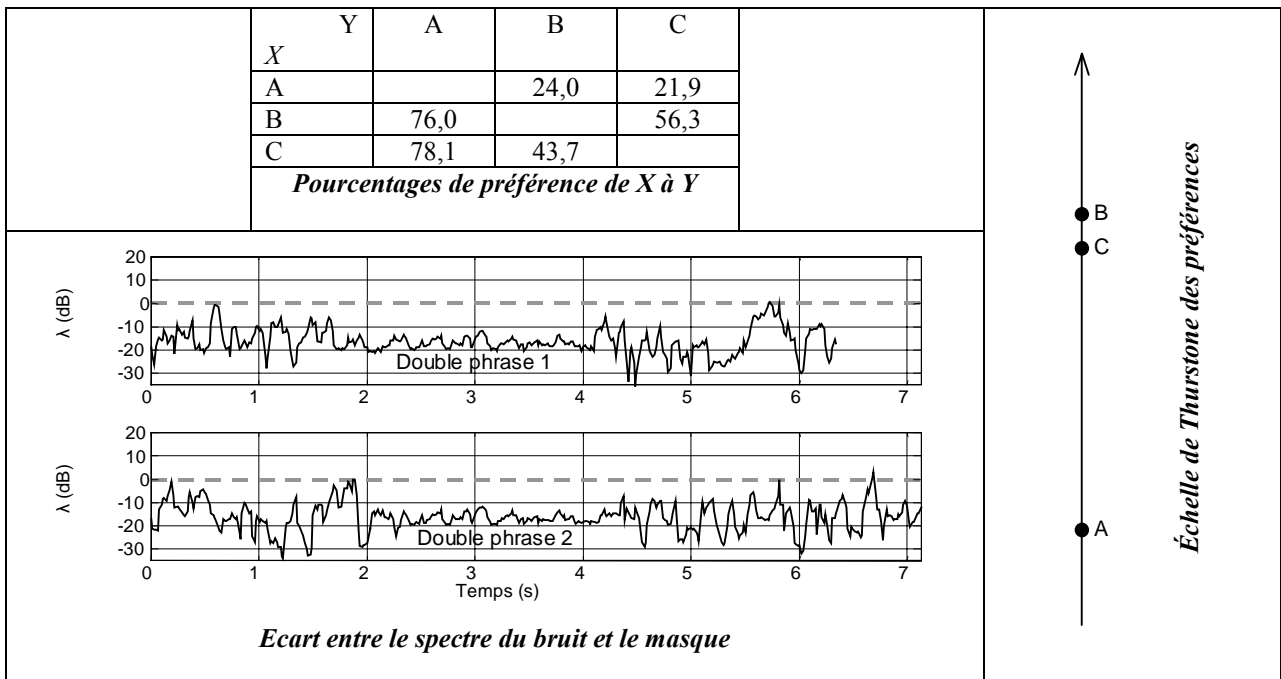


Figure 3.21 : Résultats locuteur H1, ambiance bruitée

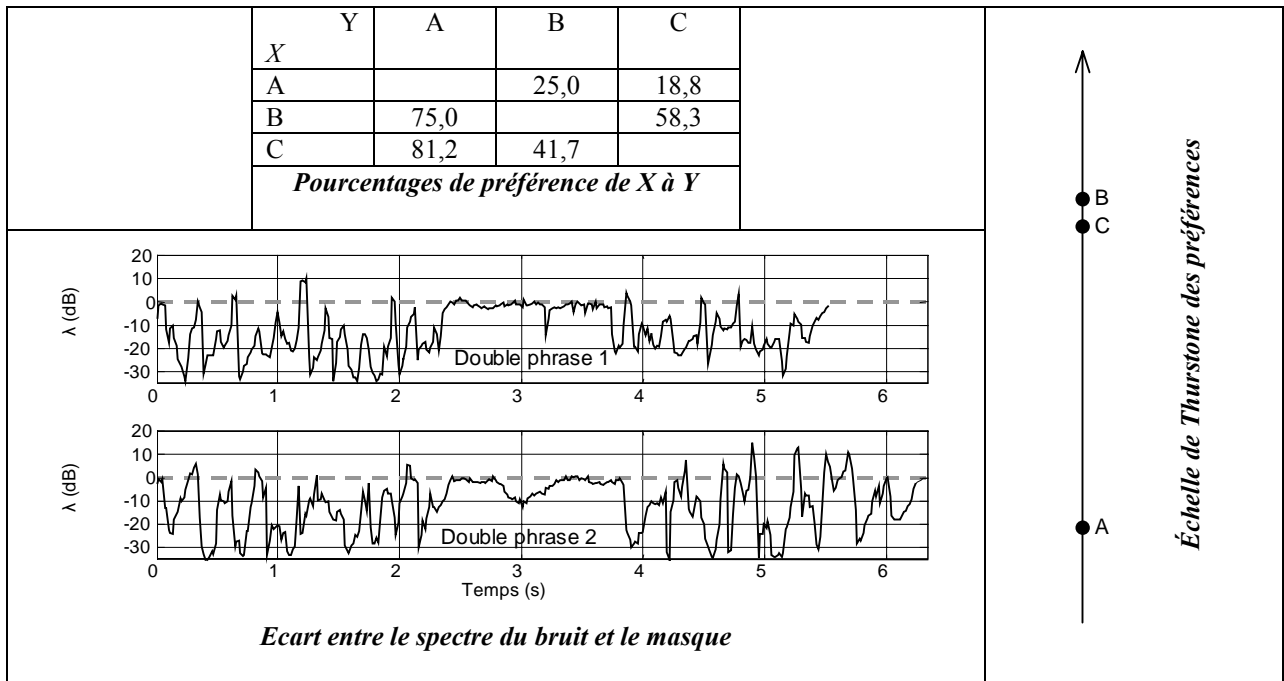


Figure 3.22 : Résultats locuteur H2, ambiance silencieuse

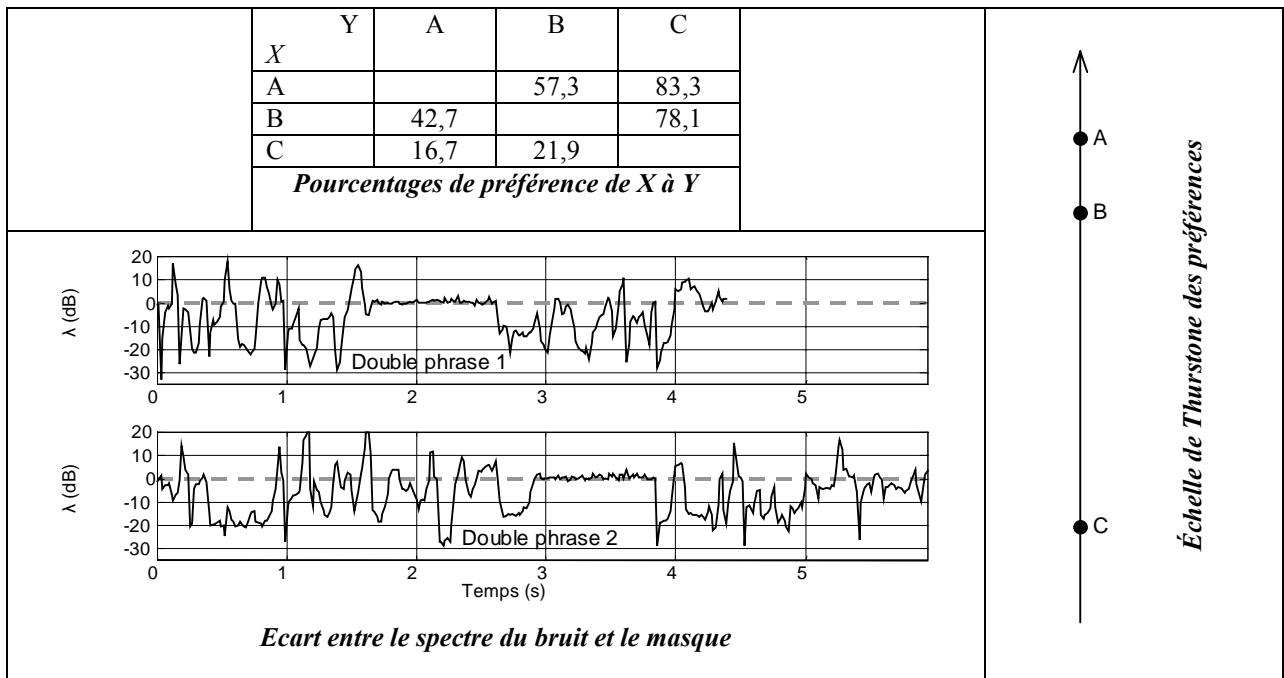


Figure 3.23 : Résultats locuteur F1, ambiance silencieuse

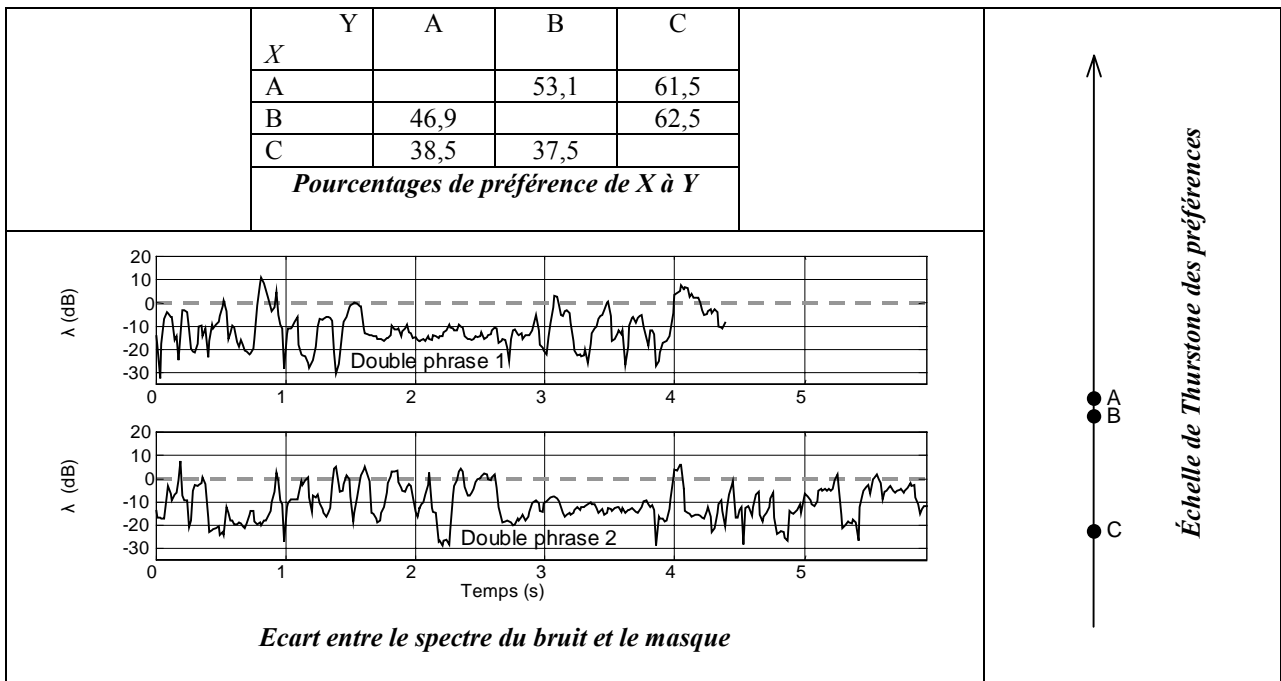


Figure 3.24 : Résultats locuteur F1, ambiance bruitée

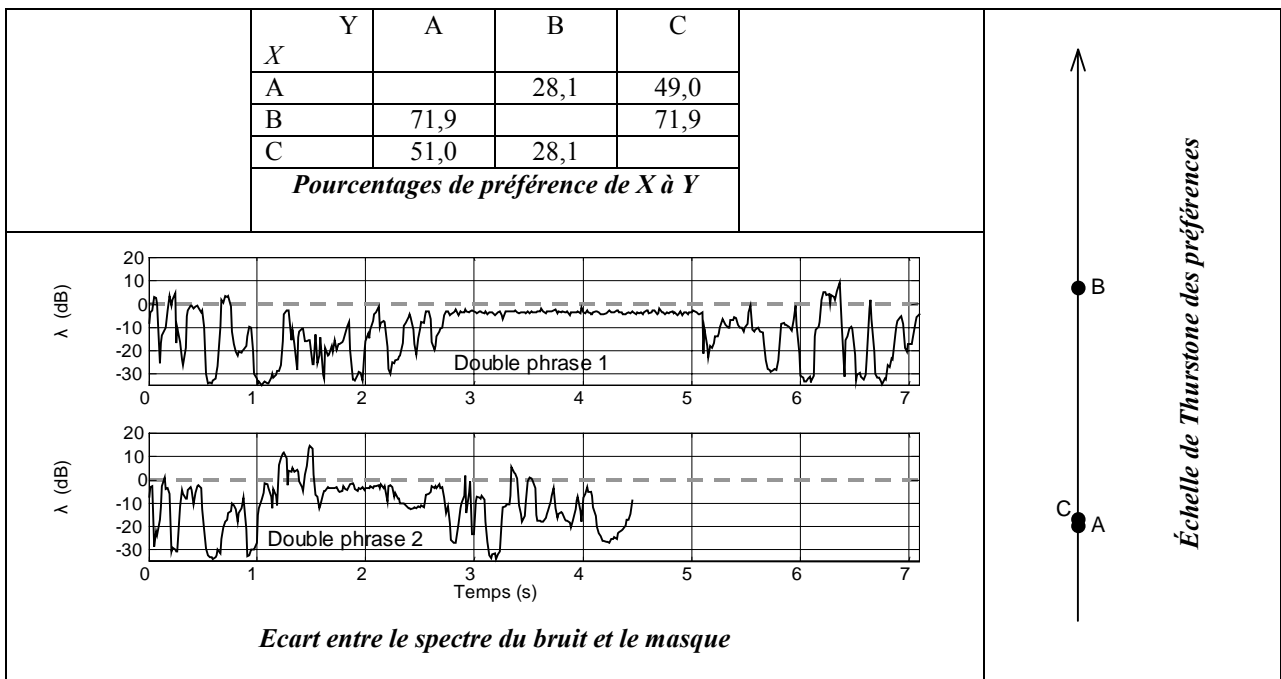
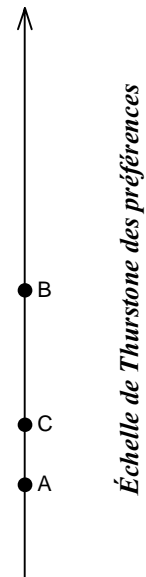


Figure 3.25 : Résultats locuteur F2, ambiance silencieuse

X	Y	A	B	C
A			32,8	42,7
B		67,2		64,3
C		57,3	35,7	
<i>Pourcentages de préférence de X à Y</i>				

Figure 3.26 : Résultats globaux, ambiance silencieuse



Les résultats moyens sur l'ensemble des locuteurs font apparaître de manière nette une hiérarchie de préférences $A < C < B$ en ambiance silencieuse.

Pour les locuteurs masculins, les auditeurs n'expriment pas de préférence nette entre B et C, et ces deux traitements sont préférés à A à 80 % environ. Les différences entre les résultats en ambiance bruitée et ceux en ambiance silencieuse ne sont pas significatives.

La préférence entre B et A pour la locutrice F2 est proche des résultats obtenus pour les locuteurs masculins. En revanche, le bruit reformé est moins bien perçu qu'avec les locuteurs masculins : aucune préférence nette n'est exprimée entre A et C, et B est largement préféré à C (72 %). La locutrice F1 est la seule des quatre locuteurs pour laquelle les auditeurs préfèrent A, cette préférence étant de 57 % quand A est comparé à B, de 83 % quand A est comparé à C. Comme pour F2, le bruit blanc de quantification est largement préféré au bruit reformé (78 %). Notons que les résultats de C sont significativement meilleurs en ambiance bruitée qu'en ambiance silencieuse, ce qui peut s'expliquer par un masquage partiel du bruit de quantification reformé par le bruit de fond. La valeur de λ est en effet plus faible en ambiance bruitée que silencieuse.

Ces différences entre les locuteurs tiennent à la fois :

- à la nature de leurs voix respectives ;
- à la dégradation du RSB de réception par l'égaliseur lorsque le bruit de quantification n'est pas reformé ;
- aux performances du reformage du bruit.

Même si la voix originale n'est pas connue, les échantillons A, privés de composantes basse fréquence, manquent de naturel pour les locuteurs masculins, et la correction de timbre est d'autant plus appréciable pour ces locuteurs que le niveau du bruit de quantification reste assez faible (échantillons B) et que le bruit reformé est presque toujours masqué (échantillons C). Pour les locutrices, l'atténuation des basses fréquences dans les échantillons A est d'autant moins gênante que leur voix est aiguë (notamment pour la locutrice F1). En outre, les résultats objectifs montrent que le bruit de quantification reformé est fréquemment au-dessus de la courbe de masquage, ce qui explique la mauvaise appréciation de C. Enfin, si la dégradation du RSB entre A et B est du même ordre de grandeur pour tous les locuteurs, la locutrice F1 est la seule pour

laquelle il existe de fortes chutes de RSB entre A et B aux instants où le RSB de A est déjà faible. Ainsi le bruit de quantification qui affecte B est-il particulièrement gênant pour F1 : il en résulte une légère préférence des auditeurs pour A.

Il ressort de ces résultats que :

- le reformage du bruit de quantification, tel que réalisé, est dans le meilleur des cas inutile, mais les auditeurs préfèrent une voix masculine égalisée et entachée de bruit reformé à la même voix non égalisée et non bruitée ;
- sans même connaître la voix originale des locuteurs, les auditeurs préfèrent une voix dont le timbre a été corrigé par l'égaliseur, malgré le bruit de quantification qui entache celle-ci.

Le deuxième test vise alors à déterminer dans quelle mesure la connaissance du timbre original des locuteurs modifie la préférence exprimée par les auditeurs entre un signal égalisé bruité et un signal non égalisé et non bruité. Dans ce deuxième test, nous considérons uniquement le bruit de quantification non reformé, préféré au bruit reformé.

• **Deuxième test : préférence connaissant l'original**

Le tableau 3.1 présente les pourcentages de préférence de B à A, comparés à ceux du premier test, pour les différents locuteurs et conditions de bruit de fond.

Pour tous les locuteurs, la préférence pour B dans le deuxième test est supérieure ou égale à celle exprimée dans le premier test. Cette augmentation est significative pour la moyenne sur l'ensemble des locuteurs, ainsi que pour la locutrice F1 (en ambiances silencieuse et bruitée). Ainsi, la connaissance de la voix originale renforce la préférence pour une voix égalisée bruitée. Pour tous les locuteurs, les auditeurs préfèrent largement une voix bruitée mais proche de l'originale à une voix dont le timbre est déformée par la liaison téléphonique. On retrouve dans le test 2 la même hiérarchie que dans le test 1 entre les résultats des différents locuteurs, liée comme précédemment à leurs voix respectives et à la dégradation du RSB dans les échantillons B.

Bruit de fond	Locuteur	Test 1	Test 2
<i>Silence</i>	H1	79,2	91,2
	H2	75,0	75,0
	F1	42,7	68,7
	F2	71,9	79,2
	Tous	67,2	76,0
<i>Brouhaha</i>	H1	76,0	76,0
	F1	46,9	70,9

Tableau 3.1 : Pourcentages de préférence de B à A

III.6. Conclusion

L'objectif premier de ce chapitre était de réduire perceptivement le bruit induit par la combinaison de l'égalisation et de la quantification en loi A associée à celle-ci, en masquant ce bruit par un reformage spectral. Les deux méthodes proposées permettent de donner au spectre du bruit de quantification une forme proche de celle de la courbe de masquage. Le bruit ainsi reformé est le plus souvent inaudible, contrairement au bruit de quantification blanc, mais dépasse occasionnellement le seuil de masquage, avec une fréquence qui dépend des locuteurs et des phonèmes. Subjectivement, le reformage spectral du bruit de quantification revient à remplacer un bruit blanc quasi-permanent par un bruit "rauque" sporadique.

L'évaluation subjective formelle du reformage indique finalement une préférence des auditeurs pour le bruit de quantification non reformé. Cette évaluation montre par ailleurs que la voix en réception d'une liaison égalisée, bien qu'entachée de bruit de quantification (reformé ou non), est préférée à celle, non bruitée, en réception de la même liaison sans égaliseur. Cette préférence est encore plus nette lorsque la voix du locuteur original est connue.

Ainsi, nous avons montré qu'à défaut de pouvoir être masqué de manière satisfaisante, le bruit de quantification induit par l'égalisation est largement toléré et ne remet pas en cause l'intérêt d'une correction de timbre par notre égaliseur.

Chapitre IV

Égalisation différenciée par classes de locuteurs

La méthode d'égalisation adaptée présentée au chapitre I, qui consiste à aligner le spectre à long terme du signal de parole traité sur le spectre moyen de la parole défini par l'UIT-T [UIT-T/P.50/App. I, 1998], permet de restaurer un timbre proche de l'original, du moins sur la bande 200-3150 Hz, pour la majorité des locuteurs testés. Cependant, l'adaptation de l'algorithme peut être assez lente (10 s d'activité vocale) pour certains locuteurs. D'autre part, pour quelques locuteurs dont le spectre à long terme est trop éloigné du spectre de référence choisi, le timbre original ne peut être restauré de manière suffisamment fidèle.

Nous nous proposons donc de prendre en considération cette variété des spectres des locuteurs en établissant des classes de locuteurs possédant chacune son propre spectre de référence. L'algorithme d'égalisation sera ainsi modifié de manière à déterminer la classe du locuteur et à égaliser suivant le spectre de référence de la classe. Celui-ci étant plus proche des spectres à long terme des membres de la classe que le spectre de référence unique utilisé dans le chapitre I, l'erreur d'approximation du spectre à long terme du locuteur par le spectre de référence devrait s'en trouver réduite.

Par ailleurs, cette réduction de l'erreur d'approximation devrait permettre de lisser moins fortement la réponse fréquentielle de l'égaliseur adapté, le rendant apte à corriger des distorsions spectrales plus fines.

Nous examinerons d'abord la pertinence d'une classification des locuteurs selon leur spectre à long terme. Les classes étant définies, des critères de classement des locuteurs seront établis, selon le nombre de classes. Enfin, l'intérêt de la classification pour la correction du timbre devra être vérifié : l'alignement du spectre à long terme sur le spectre de la classe plutôt que sur le spectre moyen de tous les locuteurs permet-il une restauration du timbre au moins aussi bonne pour tous les locuteurs, et meilleure pour certains locuteurs ?

IV.1. Classification des locuteurs

IV.1.1. Corpus

Le corpus de 34 locuteurs précédemment utilisé est trop petit pour permettre une classification pertinente en plus de deux classes. Nous utilisons un second corpus, de 29 locuteurs (16 hommes et 13 femmes) [GRECO-PRC, 1990], enregistré dans des conditions similaires à celles du premier. Ces 29 locuteurs prononcent le même texte que ceux du premier corpus, augmenté

d'une phrase d'une dizaine de secondes. Nous disposons au final d'un corpus de 63 locuteurs, dont 33 hommes et 30 femmes, prononçant chacun un texte de 23 à 52 secondes.

IV.1.2. Définition de l'individu : le cepstre partiel

La classification des locuteurs se fonde usuellement sur des statistiques sur les coefficients cepstraux calculés selon une échelle MEL [Reynolds, 1995]. L'objectif ici étant de disposer, dans chaque classe, d'un spectre de référence le plus proche possible du spectre à long terme de chaque membre de la classe, c'est sur cette base que doivent être agrégés les locuteurs. Cependant, seule la partie du spectre comprise entre F_c et 3150 Hz est prise en compte dans l'algorithme d'égalisation adaptée. Les classes doivent donc être constituées selon le spectre à long terme restreint à cette bande. D'autre part, la comparaison entre deux spectres doit être effectuée à un niveau assez bas de résolution spectrale, de manière à ne refléter que l'enveloppe spectrale. C'est pourquoi il est préférable de se placer dans l'espace des premiers coefficients cepstraux d'ordre supérieur à 0 (le coefficient d'ordre 0 représentant l'énergie), le choix du nombre de coefficients dépendant de la résolution spectrale souhaitée.

Nous définissons donc le "*cepstre partiel à long terme*", que nous noterons C^p , comme la représentation cepstrale du spectre à long terme restreint à une bande de fréquence. Si l'on note k_1 et k_2 les indices de fréquence correspondant respectivement aux fréquences F_1 et F_2 bornant cette bande (valant respectivement F_c et 3150 Hz dans le cas de la bande d'égalisation), et γ le spectre à long terme de la parole, le cepstre partiel est défini par :

$$C^p = \text{TFD}^{-1} \left(10 \log \left(\gamma(k_1 \dots k_2) \circ \gamma(k_2 - 1 \dots k_1 + 1) \right) \right) \quad (4.1)$$

où \circ désigne l'opération de concaténation. La TFD inverse est calculée par IFFT après interpolation des échantillons du spectre tronqué de manière à atteindre un nombre d'échantillons puissance de 2. En choisissant la bande d'égalisation 187-3187 Hz, correspondant aux indices fréquentiels 5 à 101 pour une représentation du spectre (symétrisé) sur 256 points (de 0 à 255), l'interpolation se fait simplement en intercalant une raie fréquentielle (interpolée linéairement) toutes les trois raies dans le spectre restreint à 187-3187 Hz. Les étapes du calcul du cepstre partiel sont représentées sur la Figure 4.1.

De manière à ce que les coefficients cepstraux reflètent l'enveloppe spectrale mais pas l'influence de la structure harmonique du spectre de la parole sur les spectres à long terme, nous ne conservons pas les coefficients d'ordre élevé. Les fréquences fondamentales moyennes des locuteurs du corpus sont inférieures à 300 Hz, soit $300 / 8000$ en fréquence normalisée. Cette fréquence fondamentale moyenne maximale est multipliée par $4/3$ lors de la troncature et de l'interpolation du spectre. Elle vaut alors $1/20$. Les locuteurs à classer sont donc représentés par les coefficients d'ordres 1 à 20 de leur cepstre partiel à long terme.

IV.1.3. Classification hiérarchique ascendante [Lebart, 2000a]

La classification hiérarchique ascendante consiste à créer, à partir des N individus disjoints, une hiérarchie de partitions selon le processus suivant : à chaque étape, on agrège les deux éléments les plus proches, un élément étant soit un individu non agrégé, soit un agrégat d'individus constitué lors d'une précédente étape. La proximité entre deux éléments est déterminée par une mesure de dissimilarité que nous appellerons *distance*. Le processus se poursuit jusqu'à l'agrégation de toute la population. La *hiérarchie de partitions* ainsi créée peut se représenter

sous la forme d'un arbre (ou *dendrogramme*) contenant $N-1$ partitions imbriquées, tel que celui de la Figure 4.2. Chaque coupure de l'arbre fournit une partition, d'autant plus fine que l'on coupe bas.

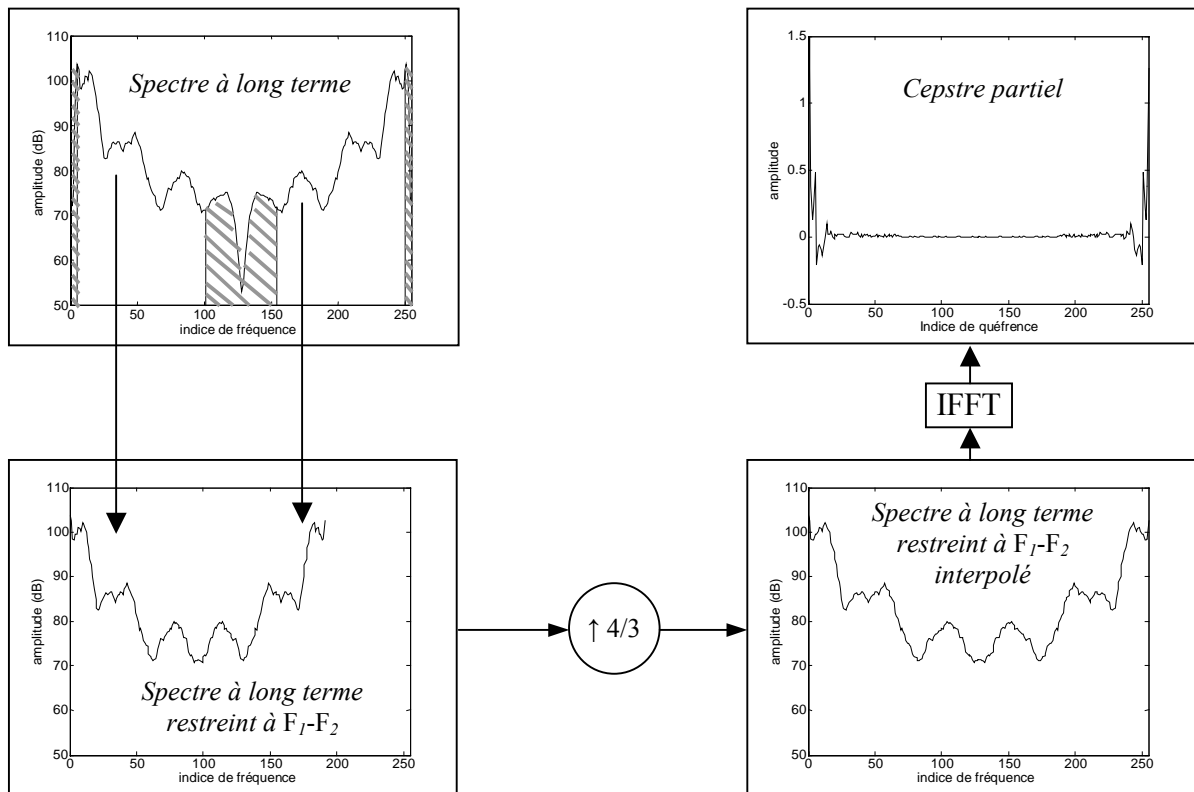


Figure 4.1 : Calcul du cepstre partiel

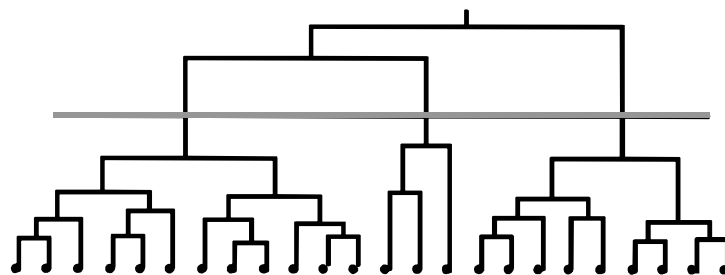


Figure 4.2 : Exemple de dendrogramme.
En gris, coupure de l'arbre au niveau optimal

La hiérarchie est dite *indicée* si à chacune de ses parties h est associée une valeur numérique $v(h)$ telle que, quelles que soient les parties h_1 et h_2 :

$$\text{si } h_1 \subset h_2 \text{ alors } v(h_1) < v(h_2)$$

Une hiérarchie peut être indicée de manière naturelle par les distances entre les éléments agrégés à chaque étape de l'algorithme. Dans la représentation sous forme d'arbre, la hauteur d'un segment horizontal agrégeant deux éléments est choisie proportionnelle à l'indice, ce qui permet de visualiser la proximité des éléments regroupés dans une même classe.

Cette représentation facilite le choix du niveau de coupure de l'arbre, et donc des classes retenues. La coupure doit être faite au-dessus des agrégations d'indice faible, qui regroupent des individus proches, et en dessous des agrégations d'indice élevé, qui associent des groupes d'individus bien distincts. Une partition de bonne qualité est donc obtenue en coupant l'arbre au niveau d'un saut important d'indice d'agrégation, comme l'illustre la barre horizontale sur la Figure 4.2.

IV.1.4. Algorithme de classification

Nous désignons sous le terme d'*élément* soit un individu soit un agrégat d'individus. Initialement, on dispose de N éléments, qui sont les individus à classer, et de la matrice D (de dimension $N \times N$) des distances entre ces éléments. A chaque étape k ($k = 1$ à $N-1$),

- on cherche les deux éléments les plus proches, que l'on agrège en un nouvel élément. On a alors une partition en $N-k$ éléments ;
- on construit la nouvelle matrice des distances D , de dimensions $(N-k) \times (N-k)$. Seule la distance entre le nouvel élément et les autres éléments est à calculer, les autres distances restant inchangées.

Le processus est réitéré jusqu'à ce que $k = N-1$, c'est-à-dire jusqu'à n'avoir plus qu'un élément regroupant tous les individus.

IV.1.5. Agrégation selon le critère du saut minimal

L'agrégation des deux plus proches éléments à chaque étape de la classification nécessite de définir une distance. Une distance simple à calculer est le *saut minimal*, défini comme suit. Si x , y et z sont trois éléments et si x et y sont agrégés en un élément h , la distance de h à z est définie par :

$$d(h, z) = \text{Min}(d(x, z), d(y, z)). \quad (4.2)$$

En d'autres termes, la distance entre deux classes est la distance entre leurs individus les plus proches, la distance entre deux individus étant définie, dans notre cas, comme la distance euclidienne dans l'espace des 20 premiers coefficients du cepstre partiel. Ainsi, si deux individus sont représentés par leurs cepstres partiels respectifs C^p et C^m , la distance entre eux est définie par :

$$d(C^p, C^m) = \sqrt{\sum_{i=1}^{20} (C_i^p - C_i^m)^2}. \quad (4.3)$$

La classification de notre corpus selon ce critère est représentée par l'arbre hiérarchique de la Figure 4.3. Une classe assez distincte des autres locuteurs apparaît (partie de l'arbre encadrée en pointillés), composée essentiellement de locutrices, et grossit par agrégation d'individus un à un, sans que ces derniers ne forment entre eux une autre classe visible. Il semble donc finalement difficile d'identifier clairement des classes dans cet arbre.

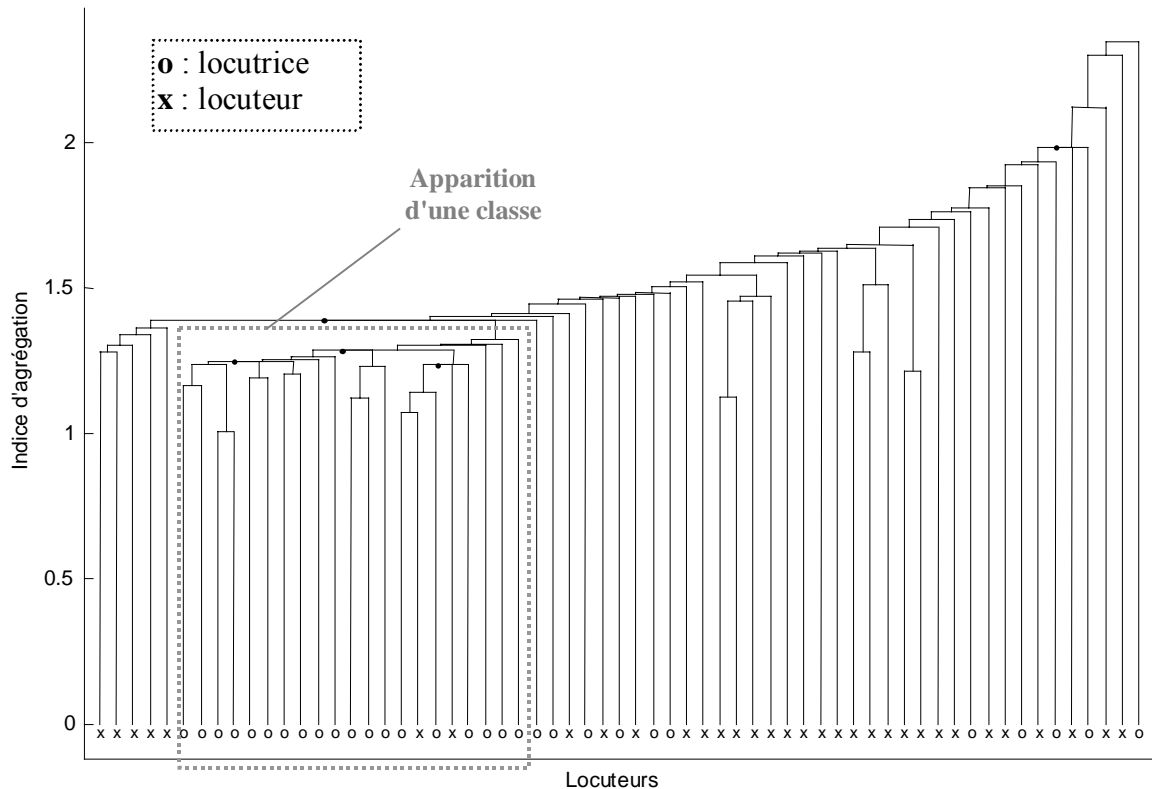


Figure 4.3 : Classification des cepstres partiels selon le critère du saut minimal

Ce phénomène est typique de l'agrégation selon le critère du saut minimal lorsque deux classes ne sont pas clairement disjointes. Il se produit un *effet de chaîne*, illustré sur la Figure 4.4. Sur cet exemple, alors que les groupes *A* et *B* sont visuellement discernables, ils ne le sont pas dans l'arbre hiérarchique, leurs sommets respectifs étant agrégés aux niveaux les plus bas. Cet effet peut être évité en utilisant un critère lié à l'inertie des classes constituées : le critère de Ward généralisé.

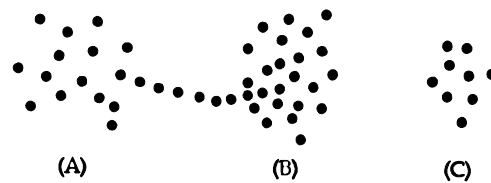


Figure 4.4 : Effet de chaîne d'une partition selon le critère du saut minimal.

IV.1.6. Agrégation selon le critère de Ward généralisé

Une partition est d'autant meilleure que les classes créées sont homogènes, c'est-à-dire que l'inertie intra-classes est faible. Dans le cas d'un nuage de points x_i de masses respectives m_i , répartis en classes q de centres de gravités respectifs g_q , l'inertie intra-classes est définie par :

$$I_{\text{intra}} = \sum_q \sum_{i \in q} m_i \|x_i - g_q\|^2. \quad (4.4)$$

L'inertie intra-classes, nulle à l'étape initiale de l'algorithme, s'accroît inévitablement à chaque agrégation. Le principe de l'agrégation selon la variance consiste à rechercher, à chaque étape de l'algorithme, les deux éléments dont l'agrégation produit l'augmentation d'inertie intra-classes la plus faible.

À l'initialisation de l'algorithme décrit dans la section IV.1.4, chaque élément est le vecteur des 20 premiers coefficients du cepstre partiel d'un locuteur et sa masse vaut 1. À chaque étape, deux éléments x_i et x_j de masses respectives m_i et m_j sont agrégés en un nouvel élément x de masse m tel que :

$$x = \frac{m_i x_i + m_j x_j}{m_i + m_j} \text{ et } m = m_i + m_j, \quad (4.5)$$

c'est-à-dire en leur barycentre. On montre [Lebart, 2000a] que l'accroissement de variance intra-classe résultant de cette agrégation vaut :

$$\Delta I_{ij} = \frac{m_i m_j}{m_i + m_j} \|x_i - x_j\|^2. \quad (4.6)$$

Cet accroissement définit la distance entre deux éléments évoquée dans l'algorithme. À chaque étape, on agrège comme indiqué ci-dessus les deux éléments x_i et x_j de masses respectives m_i et m_j tels que ΔI_{ij} est minimal.

Le dendrogramme de la classification de notre corpus selon ce critère, appelé critère de Ward généralisé, est représenté sur la Figure 4.5. En coupant l'arbre au niveau des sauts d'indice, on obtient de manière nette quatre classes. Notons que ces classes sont assez homogènes du point de vue du sexe des locuteurs, et qu'une coupure de l'arbre en deux classes fait apparaître à peu près une classe *hommes* et une classe *femmes*.

IV.1.7. Consolidation de la partition

La partition en quatre classes ainsi obtenue peut être améliorée par une procédure d'agrégation autour des centres mobiles, qui permet de réduire la variance intra-classes. L'algorithme est le suivant.

Initialement, les quatre classes sont définies par leurs centres respectifs $\{g_q^0 \mid q = 1 \text{ à } 4\}$, qui sont les barycentres des classes obtenues par classification ascendante. À chaque itération k (à partir de $k = 0$),

- Les quatre centres $\{g_q^k \mid q = 1 \text{ à } 4\}$ induisent une partition P^k de l'ensemble des individus en quatre classes $\{I_q^k \mid q = 1 \text{ à } 4\}$. Un individu est affecté à la classe I_q^k s'il est plus proche (en distance euclidienne) de g_q^k que des trois autres centres.
- Cette nouvelle partition permet de définir les nouveaux centres $\{g_q^{k+1} \mid q = 1 \text{ à } 4\}$, qui sont les barycentres respectifs des classes $\{I_q^k \mid q = 1 \text{ à } 4\}$.

L'algorithme s'arrête lorsque deux itérations successives conduisent à la même partition.

L'application de cet algorithme à la partition de notre corpus par agrégation selon le critère de Ward généralisé aboutit à quatre classes de cardinaux 18, 18, 16 et 11, plus homogènes que précédemment du point de vue du sexe : seuls un homme et deux femmes sont affectés à des

classes ne correspondant pas à leur sexe. L'homme a un pitch assez élevé et une des deux femmes a un pitch relativement bas. Les spectres restreints à la bande 187-3187 Hz correspondant aux centres de ces classes sont représentés sur la Figure 4.6 pour les classes hommes et femmes ainsi que pour leurs sous-classes respectives. A l'intérieur des grandes classes hommes et femmes, il est peu aisé de mettre en parallèle la classification selon le cepstre partiel avec des caractéristiques subjectives. Toutefois, des écoutes informelles ont permis de relever que les deux sous-classes de femmes se distinguent par la hauteur des voix. Quant aux hommes, les voix de la sous-classe 1 semblent plus "sonnantes", tandis que celles de la sous-classe 2 paraissent plus "blanches".

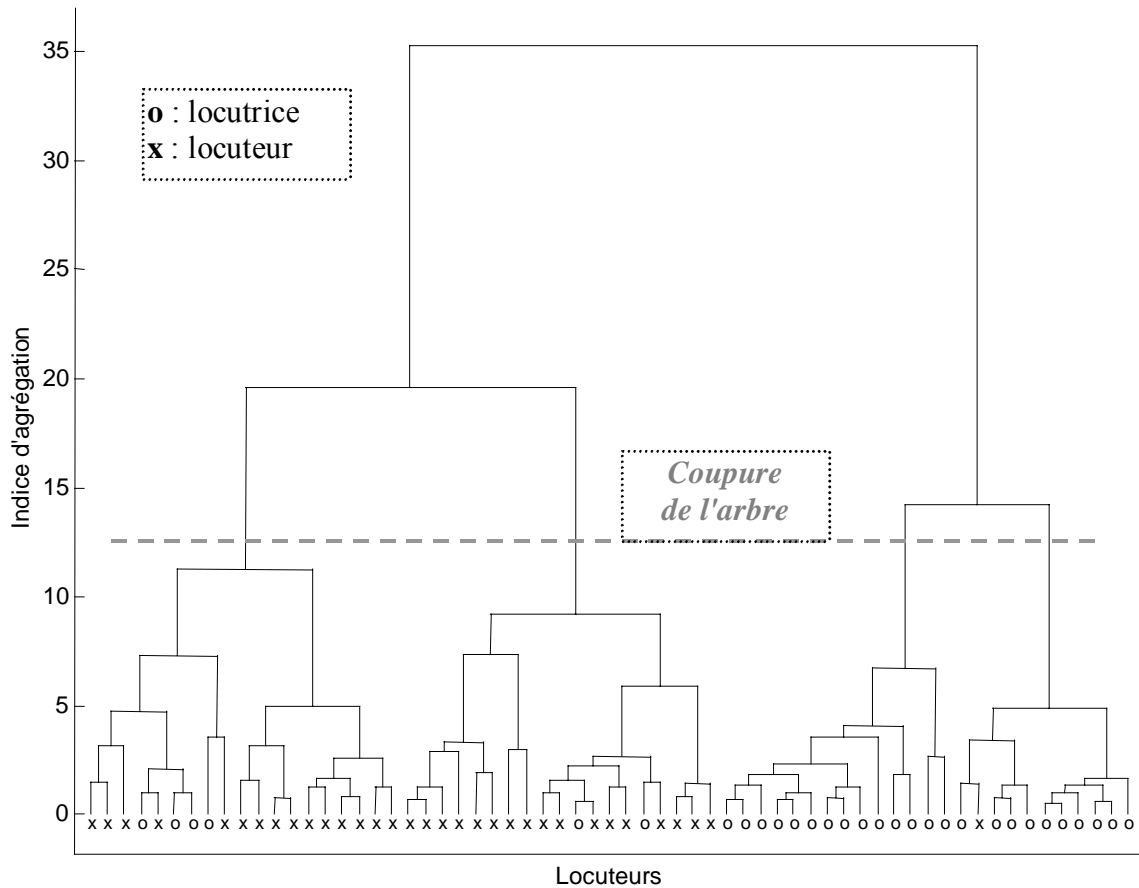


Figure 4.5 : Classification des cepstres partiels selon le critère de Ward généralisé

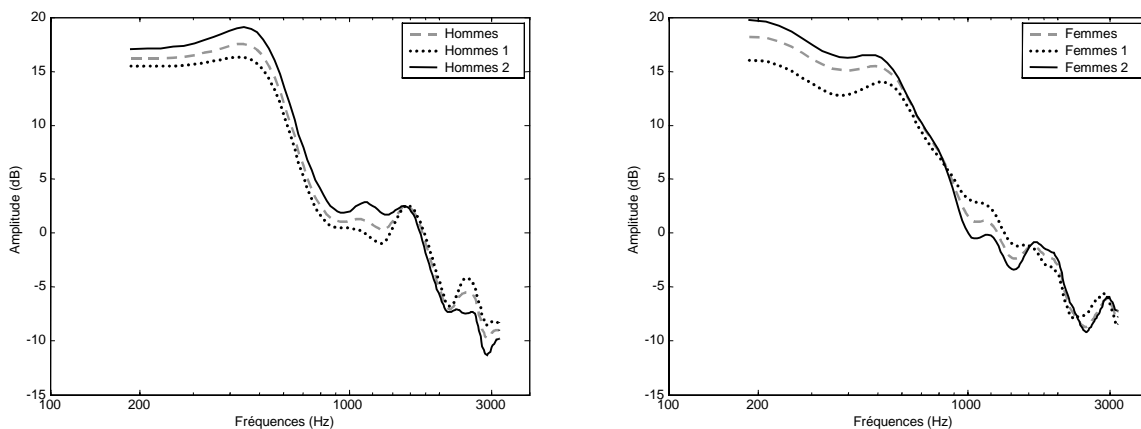


Figure 4.6 : Spectres à long terme correspondant aux centres des classes de locuteurs

IV.2. Classement des locuteurs

IV.2.1. Stratégie de classement

Les classes de locuteurs étant définies, il reste à déterminer selon quels paramètres et quels critères un locuteur sera affecté à l'une ou l'autre des classes. Cette affectation ne peut être réalisée simplement selon la proximité du cepstre partiel avec un des centres de classes, puisque ce cepstre est dévié par la partie de la liaison téléphonique en amont de l'égaliseur. Si le cepstre partiel est utilisé comme paramètre de classement, un critère robuste aux déviations par la liaison téléphonique devra être trouvé.

Que nous options pour une classification en deux ou quatre classes, les classes sont homogènes du point de vue du sexe. Le pitch étant à la fois assez discriminant pour un classement homme / femme et insensible aux distorsions spectrales envisagées, il serait pertinent de l'utiliser comme paramètre. Il pourrait être utilisé seul dans le cas d'une classification hommes / femmes, conjointement avec le cepstre partiel dans les deux classifications envisagées. H. Ezzaidi *et al.* [Ezzaidi, 2001] ont montré l'intérêt de l'utilisation conjointe du pitch et des coefficients cepstraux dans les systèmes d'identification du locuteur, en mettant en évidence la dépendance entre ces deux types de paramètres. Cette dépendance permet d'affecter les locuteurs à des classes fondées sur leurs coefficients cepstraux malgré la déviation de ceux-ci, grâce à la robustesse du pitch à ces déviations.

Nous appliquons à ces paramètres la technique usuelle de l'analyse linéaire discriminante. Si l'on dispose de N individus décrits par des vecteurs de dimension p et répartis *a priori* en K classes, l'analyse linéaire discriminante consiste :

- dans un premier temps, à chercher les $K-1$ fonctions linéaires indépendantes qui séparent au mieux les K classes. Il s'agit de déterminer quelles sont les combinaisons linéaires des p composantes des vecteurs qui minimisent la variance intra-classes et maximisent la variance interclasses.
- dans un deuxième temps, à déterminer la classe d'un nouvel individu par application des fonctions linéaires discriminantes au vecteur le représentant.

Il s'agit donc de transformer le problème de l'affectation d'un individu d'un espace à p dimensions à l'une des K classes mal séparées en un problème d'affectation d'un vecteur de dimension $K-1$ (généralement inférieur à p) à l'une des K classes bien séparées du sous-espace engendré par les $K-1$ fonctions linéaires discriminantes.

Dans notre cas, les vecteurs représentatifs des individus auront pour composantes le pitch et les coefficients 1 à 20 du cepstre partiel. La robustesse des fonctions discriminantes à la déviation des coefficients cepstraux sera assurée à la fois par la présence du pitch dans les paramètres et par le choix du corpus d'apprentissage. Celui-ci devra être composé d'individus dont la voix originale aura subi une grande diversité de filtrages représentatifs des distorsions occasionnées par les liaisons téléphoniques. Nous discuterons plus en détail du choix du corpus d'apprentissage dans les sections IV.2.4 et IV.2.5.

IV.2.2. Calcul des fonctions linéaires discriminantes

Soit $a(i)$ une combinaison linéaire des composantes x_j^i préalablement centrées d'un individu i représenté par le vecteur x^i :

$$a(i) = \sum_{j=1}^p a_j \cdot (x_j^i - \bar{x}_j), \quad (4.7)$$

où $(\bar{x}_j)_{1 \leq j \leq p} = \bar{x}$ est le centre de la population.

Soit T la matrice de covariance des p variables, d'élément générique $(t_{jk})_{1 \leq j, k \leq p}$ tel que :

$$t_{j,k} = \frac{1}{N} \sum_{i=1}^N (x_j^i - \bar{x}_j)(x_k^i - \bar{x}_k), \quad (4.8)$$

où N est le nombre d'individus de la population. La matrice T se décompose en $T = D + E$, avec D la matrice d'inertie intra-classes et E la matrice d'inertie inter-classes, d'éléments génériques respectifs d_{jk} et e_{jk} tels que :

$$d_{jk} = \frac{1}{N} \sum_{q=1}^K \sum_{i \in q} (x_j^i - \bar{x}_j^q)(x_k^i - \bar{x}_k^q) \quad (4.9)$$

$$e_{jk} = \sum_{q=1}^K \frac{N_q}{N} (\bar{x}_j^q - \bar{x}_j)(\bar{x}_k^q - \bar{x}_k), \quad (4.10)$$

où \bar{x}^q désigne le centre de la $q^{\text{ème}}$ classe et N_q le cardinal de la $q^{\text{ème}}$ classe.

Minimiser la variance intra-classes et maximiser la variance inter-classes de a revient à choisir a comme le vecteur propre de $T^1 E$ relatif à la plus grande valeur propre (voir détails de la démonstration dans l'annexe J). Cette valeur propre est appelée le *pouvoir discriminant* de la fonction linéaire a . Pour une partition en K classes, les $K-1$ fonctions linéaires discriminantes correspondront aux vecteurs propres associés aux $K-1$ plus grandes valeurs propres de $T^1 E$.

Dans ce cas particulier du classement en deux classes, hommes (H) et femmes (F), on montre (cf. annexe J) que la fonction linéaire discriminante a est définie par :

$$a = T^{-1} c, \quad (4.11)$$

avec c un vecteur colonne de composantes c_j telles que :

$$c_j = \frac{\sqrt{N_H N_F}}{N} \begin{pmatrix} -\bar{x}_j^H & -\bar{x}_j^F \end{pmatrix}. \quad (4.12)$$

où N_H et N_F sont les nombres respectifs d'hommes et de femmes, \bar{x}^H et \bar{x}^F sont les moyennes de x respectivement chez les hommes et chez les femmes.

IV.2.3. Affectation d'une nouvelle observation

Une fois définies les $K-1$ fonctions discriminantes $(a^k)_{1 \leq k \leq K-1}$, on souhaite affecter une nouvelle observation x en fonction de $(a^k(x))_{1 \leq k \leq K-1}$.

Jambu [Jambu, 1999] propose d'affecter x selon sa distance aux différents centres de classes dans le sous-espace engendré par les fonctions discriminantes. L'observation x est affectée à la classe q telle que la distance entre son centre \bar{x}^q et x :

$$d(x, q) = \sqrt{\sum_{k=1}^{K-1} \left(a^k(x) - a^k(\bar{x}^q) \right)^2} \quad (4.13)$$

soit minimale.

Cette approche purement géométrique ne tient cependant pas compte des variances des classes ni de leurs probabilités respectives. C'est pourquoi Lebart *et al.* [Lebart, 2000b] proposent de classer une nouvelle observation selon un critère bayésien d'affectation : un individu x est affecté à la classe q si la probabilité conditionnelle de q sachant x , notée $P(q|x)$ est maximale. Selon le théorème de Bayes,

$$P(q|x) = \frac{P(x|q)P(q)}{P(x)}. \quad (4.14)$$

Par conséquent, $P(q|x)$ est proportionnelle à $P(x|q)P(q)$. Lebart *et al.* expriment cette probabilité dans le cas de classes à distribution normale en restant dans l'espace initial à p dimensions, sans utiliser la fonction discriminante calculée. Nous calculons ce critère bayésien en nous plaçant plutôt dans le sous-espace engendré par les $K-1$ fonctions discriminantes.

Sous l'hypothèse d'une distribution multi-gaussienne des individus dans chaque classe, si l'on note $f_k(x)$ la densité de probabilité de $a(x)$ (vecteur de composantes $(a^k(x))_{1 \leq k \leq K-1}$) à l'intérieur de la classe q :

$$f_q(x) = \frac{1}{(2\pi)^{\frac{K-1}{2}} \sqrt{|S_q|}} \exp\left(-\frac{1}{2} \left(a(x) - a(\bar{x}^q) \right)' S_q^{-1} \left(a(x) - a(\bar{x}^q) \right)\right), \quad (4.15)$$

où S_q est la matrice des covariances de a à l'intérieur de la classe q , d'élément générique σ_{jk}^q que l'on peut estimer par :

$$\sigma_{jk}^q = \frac{1}{N_q} \sum_{i=1}^{N_q} \left(a^j(x^i) - a^j(\bar{x}^q) \right) \left(a^k(x^i) - a^k(\bar{x}^q) \right). \quad (4.16)$$

L'individu x sera affecté à la classe q qui maximise $f_q(x)P(q)$, ce qui revient à minimiser sur q la fonction $s_q(x)$ appelée score discriminant :

$$s_q(x) = \left(a(x) - a(\bar{x}^q) \right)' S_q^{-1} \left(a(x) - a(\bar{x}^q) \right) + \log(|S_q|) - 2 \log(P(q)). \quad (4.17)$$

IV.2.4. Application au classement en deux classes hommes / femmes

- **Choix du corpus d'apprentissage**

Il est souhaitable de disposer d'un corpus d'apprentissage pour déterminer les critères d'affectation et d'un corpus de test distinct pour vérifier le classement des locuteurs par application de ces critères. Le corpus d'apprentissage doit être représentatif de la multiplicité des filtrages subis par les signaux reçus par l'égaliseur, de sorte que les critères de classement soient robustes à la déviation des coefficients du cepstre partiel.

Nous disposons pour cela d'un nouveau corpus, enregistré par Vecsys pour France Télécom R&D à des fins d'authentification vocale du locuteur à travers le réseau téléphonique [Vecsys, 1994]. La procédure d'enregistrement consistait en ce que les locuteurs téléphonent, en utilisant leur poste RTC personnel, à un serveur placé en réception d'une liaison RNIS, selon le schéma de la Figure 4.7. Les phrases prononcées étaient alors enregistrées sur le serveur. La liaison RNIS n'introduisant aucun filtrage en aval de la position prévue pour notre égaliseur, les signaux du corpus peuvent être considérés comme représentatifs du type de signaux à l'entrée de l'égaliseur. Le corpus compte 129 locuteurs, lisant chacun 1 à 5 phrases phonétiquement équilibrées (le jeu de 5 phrases étant le même pour tous) issues du journal *Le Monde*.

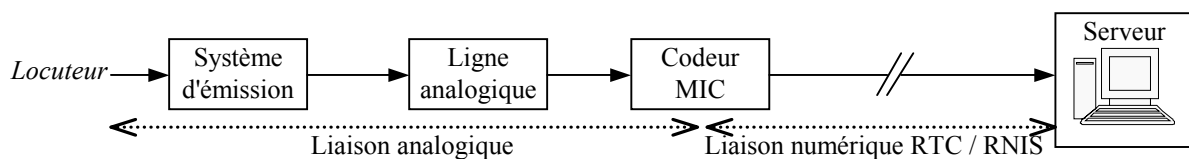


Figure 4.7 : Chaîne d'enregistrement du corpus utilisé pour l'apprentissage des critères de classement

Les classes respectives des locuteurs de ce corpus d'apprentissage doivent être connues *a priori*. Comme ce corpus sert précisément à définir les critères d'affectation, le seul moyen de connaître les classes est d'écouter les phrases et de déterminer à partir de cette écoute le sexe des locuteurs. Ainsi, nous éliminons de ce corpus 2 locuteurs dont l'écoute ne permet pas de déterminer le sexe.

Les paramètres utilisés pour définir les critères de classement étant le pitch moyen et les coefficients 1 à 20 du cepstre partiel à long terme, la durée d'activité vocale des phrases doit permettre une estimation fiable du spectre à long terme. D'après les résultats du chapitre II, 5 s d'activité vocale sont suffisantes pour la grande majorité des locuteurs. Nous sélectionnons donc dans ce corpus les phrases de plus de 5 s d'activité vocale.

Il reste au final 463 phrases prononcées par 126 locuteurs, dont 61 hommes et 65 femmes. Notons que pour chaque locuteur, les phrases ont généralement été enregistrées au cours de communications différentes, avec des conditions d'enregistrement variables d'un appel à l'autre : les écoutes effectuées indiquent que pour un même locuteur, le terminal, la distance de la bouche au micro et l'acoustique du lieu de l'appel varient entre deux appels. Ainsi, nous disposons d'un corpus de 463 individus représentatifs de conditions (locuteur, filtrage de la liaison) très diverses.

- **Définition des critères de classement**

Nous étudions ici les performances de deux critères de classement homme / femme : d'une part le pitch seul, d'autre part une combinaison linéaire (selon l'analyse linéaire discriminante) du pitch et des coefficients 1 à 20 du cepstre partiel.

Le pitch moyen étant à la fois caractéristique du sexe d'un locuteur et insensible aux distorsions spectrales introduites par une liaison téléphonique, il peut être utilisé seul comme critère de classement homme / femme.

La Figure 4.8 représente la distribution des pitches moyens des 463 individus du corpus, ainsi que la modélisation gaussienne de cette distribution. Pour chaque individu, le pitch F_0 est calculé sur chaque trame d'activité vocale (trames de 32 ms se recouvrant à 50 %) par la méthode de l'autocorrélation [UIT-T/G.729, 1996]. Le pitch moyen est la moyenne du pitch sur toutes les trames voisées, une trame étant considérée comme voisée ou non par comparaison de l'autocorrélation normalisée en $1/F_0$ à un seuil. Si l'on note Γ la fonction d'autocorrélation, l'autocorrélation normalisée, notée Γ_N , est définie par :

$$\Gamma_N(\tau) = \frac{\Gamma(\tau)}{\Gamma(0)} \quad (4.18)$$

La valeur $\Gamma_N(1/F_0)$, comprise entre 0 et 1, est d'autant plus proche de 1 que la trame est voisée. Le seuil de voisement est fixé à 0,5.

Nous examinons maintenant l'erreur de classement *apparente* (i.e. calculée sur le corpus d'apprentissage) si le pitch est utilisé comme critère unique de classement. D'après l'équation (4.14), la probabilité d'appartenance d'un locuteur à la classe q ($q = H$ ou F) connaissant son pitch moyen $\overline{F_0}$ est proportionnelle à la densité de probabilité de $\overline{F_0}$ sachant q , puisque les deux classes hommes / femmes ont *a priori* la même probabilité. Sous l'hypothèse d'une distribution gaussienne des pitches moyens dans chaque classe, cette densité de probabilité s'exprime par :

$$f_q(\overline{F_0}) = \frac{1}{\sigma_q \sqrt{2\pi}} \exp\left(-\frac{\left(\overline{F_0} - \overline{\overline{F_0}}^q\right)^2}{2\sigma_q^2}\right), \quad (4.19)$$

où $\overline{\overline{F_0}}^q$ et σ_q^2 sont respectivement la moyenne et la variance des pitches moyens sur la classe q . Le seuil de décision est la solution de l'équation :

$$f_H(\overline{F_0}) = f_F(\overline{F_0}). \quad (4.20)$$

Ainsi, un locuteur est classé parmi les hommes si son pitch moyen est inférieur à 193 Hz, parmi les femmes s'il est supérieur à ce seuil. Nous notons $\overline{F_0}^{\text{seuil}}$ ce seuil.

La probabilité de classer un homme parmi les femmes est donnée par :

$$e_H = 1 - \frac{1}{2} \operatorname{erfc}\left(\frac{\overline{F_0}^{\text{seuil}} - \overline{\overline{F_0}}^H}{\sigma_H \sqrt{2}}\right); \quad (4.21)$$

celle de classer une femme parmi les hommes est définie par :

$$e_F = \frac{1}{2} \operatorname{erfc} \left(\frac{\overline{F_0^{\text{seuil}}} - \overline{F_0^F}}{\sigma_F \sqrt{2}} \right), \quad (4.22)$$

où erfc désigne la fonction d'erreur complémentaire.

On obtient ainsi une erreur prévisible de classement de 6,4 % pour les hommes et 6,0 % pour les femmes.

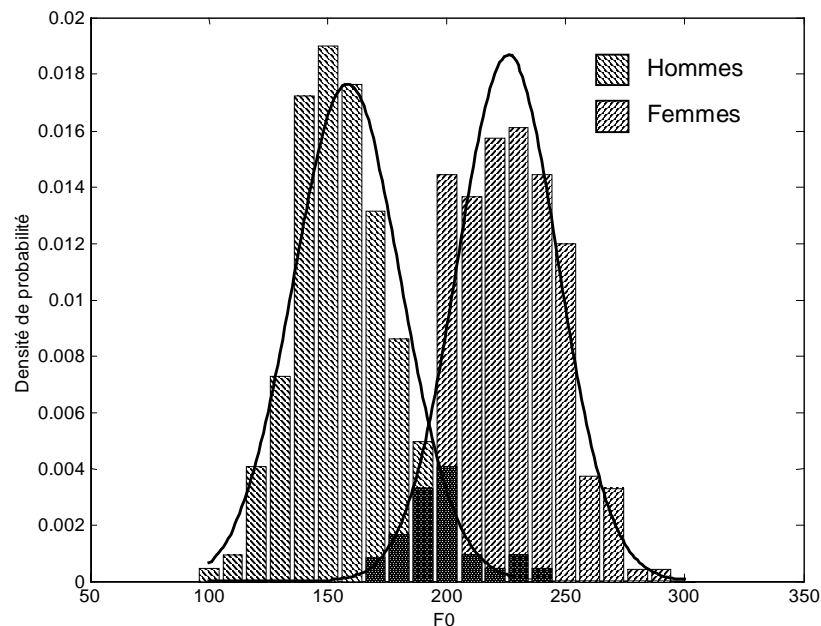
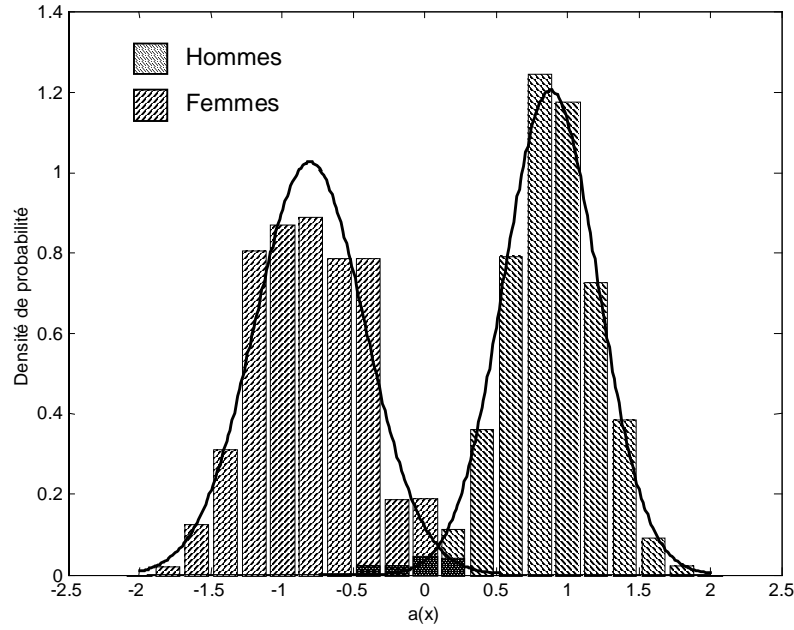


Figure 4.8 : Distribution des pitchs moyens ; modélisation gaussienne de cette distribution

Nous souhaitons maintenant évaluer l'apport des 20 premiers coefficients du cepstre partiel à la discrimination homme / femme, en utilisant de manière conjointe le pitch moyen et les coefficients du cepstre partiel comme paramètres de discrimination. Il s'agit de chercher la fonction linéaire qui, appliquée au vecteur $x = [\overline{F_0} ; C^p(1) ; \dots ; C^p(20)]$, sépare au mieux les deux classes.

Une analyse linéaire discriminante est menée comme indiqué dans la section IV.2.2 sur le corpus de 463 individus précédemment défini et conduit à la distribution des $a(x)$ représentée sur la Figure 4.9 avec la modélisation gaussienne de la distribution de chaque classe. La comparaison avec la Figure 4.8 indique une meilleure séparation des classes. Le pitch garde un rôle prépondérant dans la discrimination, puisque si l'on considère les coefficients de a calculés à partir du vecteur x normalisé (par sa variance), le coefficient multiplicatif du pitch vaut 0,51, contre 0,18 pour le plus grand des coefficients multiplicatifs des coefficients cepstraux.



**Figure 4.9 : Distribution de $a(x)$;
modélisation gaussienne de cette distribution**

De manière similaire à la discrimination selon le pitch, la probabilité d'appartenance d'un locuteur à la classe q ($q = H$ ou F) connaissant $a(x)$ est proportionnelle à la densité de probabilité de $a(x)$ sachant q . Sous l'hypothèse d'une distribution gaussienne de $a(x)$ dans chaque classe, cette densité de probabilité s'exprime par :

$$f_q(a(x)) = \frac{1}{\sigma_q^a \sqrt{2\pi}} \exp\left(-\frac{(a(x) - a(\bar{x}^q))^2}{2\sigma_q^{a^2}}\right), \quad (4.23)$$

où σ_q^a désigne l'écart-type de $a(x)$ dans la classe q . Le seuil de décision, que nous noterons a^{seuil} , est la solution de l'équation :

$$f_H(a) = f_F(a). \quad (4.24)$$

La probabilité de classer un homme parmi les femmes est donnée par :

$$e_H = \frac{1}{2} \operatorname{erfc}\left(\frac{a^{\text{seuil}} - a(\bar{x}^H)}{\sigma_H^a \sqrt{2}}\right); \quad (4.25)$$

celle de classer une femme parmi les hommes est définie par :

$$e_F = 1 - \frac{1}{2} \operatorname{erfc}\left(\frac{a^{\text{seuil}} - a(\bar{x}^F)}{\sigma_F^a \sqrt{2}}\right). \quad (4.26)$$

On obtient ainsi une erreur *apparente* de classement de 0,9 % pour les hommes et 1,0 % pour les femmes, ce qui permet d'augurer une efficacité du critère $a(x)$ meilleure que celle du pitch seul.

• **Application des critères de classement au corpus de test**

Nous utilisons comme corpus de test le corpus de 63 locuteurs présenté dans la section IV.1, que nous filtrons par différentes liaisons RTC en amont de l'égaliseur. Nous testons les trois mêmes combinaisons de systèmes d'émission et de lignes analogiques d'émission que dans la section II.3.4. Le vecteur x précédemment défini est calculé au fil de l'eau comme décrit ci-dessous.

Le spectre à long terme est ajusté à chaque trame d'activité vocale comme dans l'égaliseur adapté décrit au chapitre II (voir Figure 2.9), selon la formule suivante :

$$\gamma_x(f, n) = \alpha(n) |X(f, n)|^2 + (1 - \alpha(n)) \gamma_x(f, n-1), \quad (4.27)$$

où $\gamma_x(f, n)$ est le spectre à long terme de la sortie x du pré-égaliseur à la $n^{\text{ème}}$ trame d'activité vocale, $X(f, n)$ la transformée de Fourier de la $n^{\text{ème}}$ trame d'activité vocale, et

$$\alpha(n) = \frac{1}{\min(n, N)}, \quad (4.28)$$

où N est le nombre de trames dans 4 s. On déduit de ce spectre à long terme le cepstre partiel, selon la procédure définie à la section IV.1.2.

Le pitch moyen $\overline{F_0}$ est estimé à chaque trame voisée selon la formule :

$$\overline{F_0}(m) = \alpha(m) F_0(m) + (1 - \alpha(m)) \overline{F_0}(m-1), \quad (4.29)$$

où $F_0(m)$ est le pitch de la $m^{\text{ème}}$ trame voisée.

Ainsi, à chaque trame d'activité vocale, on dispose d'un nouveau vecteur x de composantes le pitch moyen et les coefficients 1 à 20 du cepstre partiel, auquel on applique la fonction discriminante a définie à partir du corpus d'apprentissage. On compare alors :

- le pitch moyen au seuil de 193 Hz ;
- $a(x)$ au seuil a^{seuil} .

Dans chacun des cas, on décide alors de l'affectation du locuteur dans une des deux classes.

Les Figures 4.10 et 4.11 représentent les erreurs de classement résultant de cette procédure, respectivement selon le critère du pitch et selon celui de la fonction discriminante appliquée au vecteur x (pitch moyen et cepstre partiel). Chaque ligne correspond à un locuteur : les lignes 1 à 33 (en partant du haut) représentent les locuteurs masculins, les lignes 34 à 63 les locutrices. Pour une ligne donnée, chaque pixel représente l'erreur de classement pour une trame : le pixel est gris clair si le locuteur est bien classé, noir sinon. Seuls les résultats de la deuxième liaison simulée (système d'émission ayant la caractéristique nominale du SRI modifié et ligne d'émission moyenne) sont représentés, ceux des autres liaisons étant similaires. Le Tableau 4.1 indique les taux d'erreur de classement pour les trois liaisons, calculés selon deux méthodes. La première consiste à considérer qu'un locuteur est mal classé si, après 10 s de parole (temps maximal d'estimation du spectre à long terme d'après les résultats du chapitre II), une erreur de classement est commise pour plus de 25 % des trames. La seconde méthode consiste à calculer, sur l'ensemble des locuteurs, le pourcentage de trames mal classées après 10 s de parole.

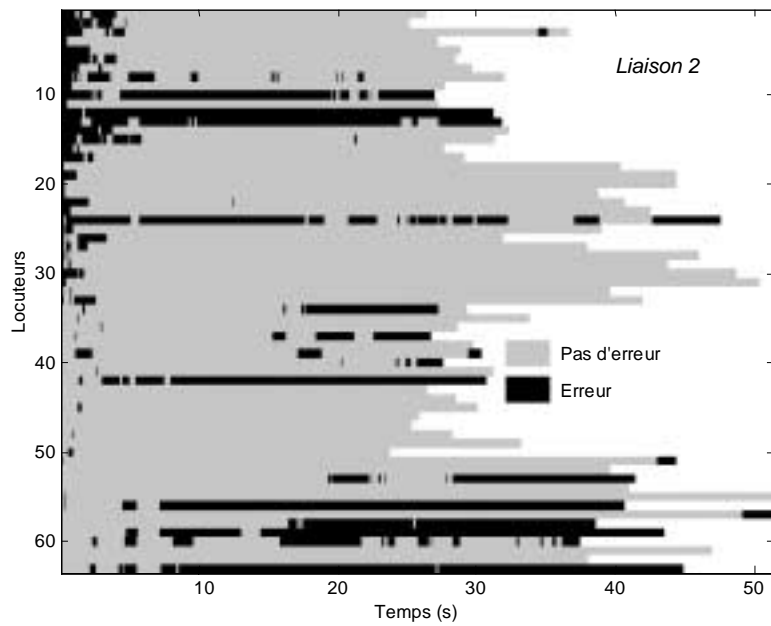


Figure 4.10 : Erreurs de classement selon le pitch pour chaque locuteur à chaque trame de signal

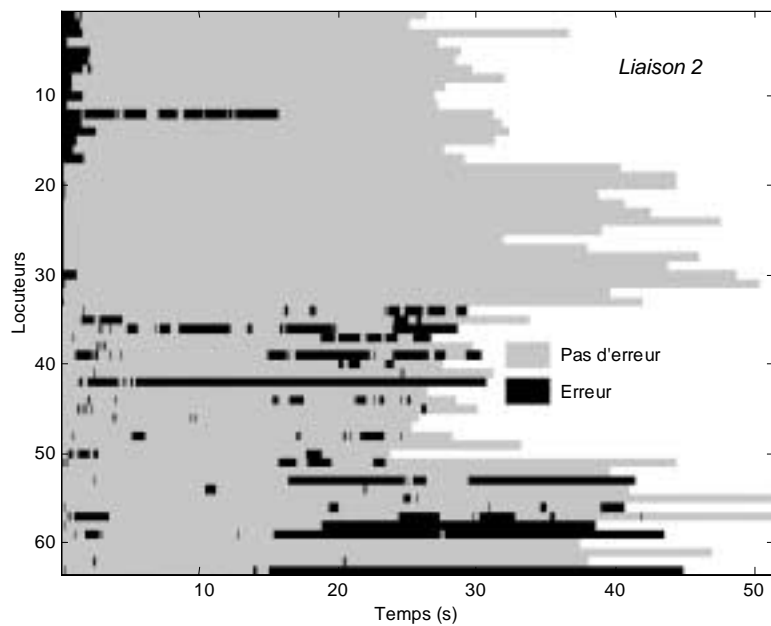


Figure 4.11 : Erreurs de classement selon $a(x)$ (combinaison du pitch et du cepstre partiel) pour chaque locuteur à chaque trame de signal

Ces résultats confirment que l'utilisation conjointe du pitch et du cepstre partiel dans la détermination du sexe du locuteur permet de réduire l'erreur de classement commise en se fondant sur le seul pitch. Selon la définition du taux d'erreur adoptée, celui-ci passe de 20 % à 15 % ou de 17 % à 13 % en moyenne. L'erreur reste cependant nettement supérieure à l'erreur théorique calculée sur le corpus d'apprentissage. Notons que les erreurs de classement sont plus particulièrement concentrées sur les locutrices (locuteurs 34 à 63), ce qui peut s'expliquer par une translation de la distribution de $a(x)$ pour le corpus de test par rapport au corpus d'apprentissage, ayant pour résultat de déplacer le seuil de décision vers le centre de la distribution de la classe

femmes. La fonction a étant linéaire, cette déviation correspond à une déviation cepstrale globale du corpus de test par rapport au corpus d'apprentissage.

Liaison	Locuteurs mal classés (1)			Trames mal classées (2)		
	1	2	3	1	2	3
Classement selon F_0 seul	21 %	21 %	19 %	17 %	17 %	17 %
Classement selon F_0 et $C_{1 \rightarrow 20}^p$	14 %	14 %	17 %	11 %	12 %	15 %
(1) locuteurs pour lesquels une erreur de classement est commise sur plus de 25 % des trames après 10 secondes de parole.						
(2) Sur l'ensemble des locuteurs, pourcentage de trames mal classées postérieures à 10 secondes						

Tableau 4.1 : Taux d'erreur de classement des locuteurs

IV.2.5. Application au classement en quatre classes

- **Choix des corpus d'apprentissage et de test**

Le corpus d'apprentissage utilisé pour l'analyse linéaire discriminante dans le cas d'un classement homme / femme ne peut être utilisé ici. L'analyse linéaire discriminante nécessite en effet de connaître *a priori* la classe de chaque locuteur du corpus d'apprentissage. Si le sexe peut être déterminé par une simple écoute, il n'en est pas de même pour l'appartenance à l'une des quatre classes définies dans la section IV.1. Avant définition des fonctions discriminantes, cette appartenance ne peut être déterminée qu'à partir de la parole originale (non filtrée par une liaison téléphonique), selon la distance entre le cepstre partiel à long terme et les différents centres de classes. Ne connaissant ni les signaux d'émission du corpus d'apprentissage précédent, ni les caractéristiques des liaisons téléphoniques des locuteurs, nous ne pouvons pas déterminer *a priori* la classe de ceux-ci.

Ce corpus ne peut être non plus utilisé comme corpus de test, puisque les classes des locuteurs de celui-ci doivent être également connues *a priori*, pour pouvoir calculer les taux d'erreur de reconnaissance.

Nous utilisons donc le corpus de 63 locuteurs utilisé dans la section IV.1 à la fois comme corpus d'apprentissage et comme corpus de test, en le dédoublant de la manière suivante. Les critères de classement seront appris sur ce corpus modifié par une grande variété de filtrages et seront testés sur ce corpus filtré comme précédemment par différentes liaisons RTC en amont de l'égaliseur, les caractéristiques de ces liaisons différant de celles des filtres utilisés pour construire le corpus de test.

La construction du corpus d'apprentissage consiste à définir un ensemble de M biais cepstraux qui s'ajouteront chacun à chaque cepstre partiel représentatif d'un locuteur du corpus original, ce qui permet d'obtenir un nouveau corpus de $63M$ individus. Ces biais dans le domaine du cepstre partiel doivent correspondre à une large gamme de distorsions spectrales sur la bande 187-3187 Hz. L'idéal serait de disposer de statistiques sur la caractéristique fréquentielle du système d'émission et de la ligne pour toute la population d'abonnés, de manière à définir un ensemble de biais cepstraux représentatif de cette distribution de réponses fréquentielles. Ne disposant pas de telles données, nous proposons, de manière arbitraire, l'ensemble de réponses fréquentielles représentées sur la Figure 4.12. pour la bande 187-3187 Hz : chaque réponse

fréquentielle correspond à un chemin de gauche à droite dans le treillis. L'amplitude de leurs variations sur cette bande n'excède pas 20 dB, à l'instar des caractéristiques extrémales des systèmes d'émission et lignes présentées au chapitre I.

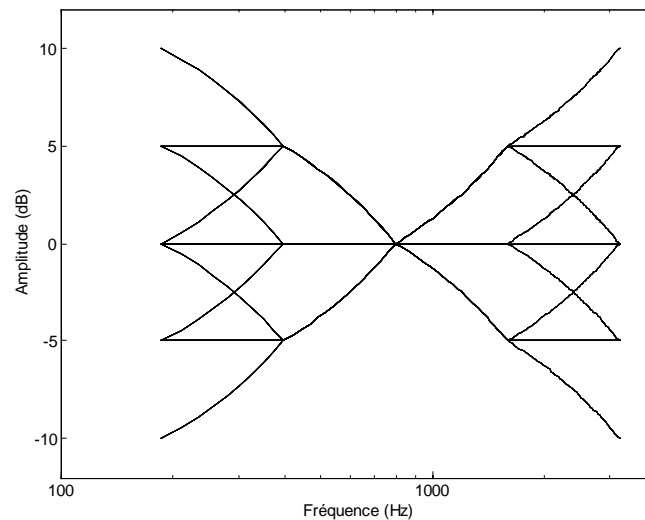


Figure 4.12 : Caractéristiques fréquentielles des filtrages appliqués au corpus initial pour définir le corpus de test

A partir de ces 81 caractéristiques fréquentielles sont calculés les 81 biais correspondants dans le domaine du cepstre partiel, selon la procédure définie en IV.1.2. Par addition de ces biais au corpus initial, on obtient un corpus d'apprentissage comptant 5103 individus représentatifs de diverses conditions (locuteur, filtrage de la liaison).

Notons que dans l'espace considéré, celui du cepstre partiel, les corpus d'apprentissage et de test constituent bien deux corpus différents. D'une part, les biais cepstraux introduits par les filtrages dans les corpus de test et dans le corpus d'apprentissage sont du même ordre de grandeur que les écarts entre locuteurs dans le corpus original. D'autre part, les filtrages utilisés pour construire le corpus d'apprentissage ne correspondent à aucune des liaisons simulées dans les corpus de test.

- **Définition des critères de classement**

Comme pour le classement en deux classes, nous menons une analyse linéaire discriminante en utilisant comme paramètre le vecteur $x = [\overline{F}_0 ; C^p(1) ; \dots ; C^p(20)]$. Trois fonctions linéaires discriminantes $(a^k)_{k=1,2,3}$ sont calculées selon la procédure de la section IV.2.2. Leurs pouvoirs discriminants respectifs valent 0,95, 0,50 et 0,43. Le pitch garde un rôle prépondérant dans la discrimination, son coefficient multiplicatif normalisé étant le plus grand dans chacune des trois fonctions discriminantes. La visualisation des nuages de points de coordonnées $(a^1(x); a^2(x); a^3(x))$ fait apparaître une nette séparation des classes hommes et femmes ainsi que des deux sous-classes de femmes (aucun recouvrement). En revanche, les deux sous-classes d'hommes se recouvrent légèrement.

Le calcul de l'erreur apparente de manière analytique est plus délicat que dans le cas du classement en deux classes. Le taux d'erreur par classe est donc déterminé par le calcul des scores discriminants sur toute la population d'apprentissage. Pour un individu d'une classe donnée, il y a erreur si le score discriminant associé à cette classe n'est pas inférieur à ceux associés aux trois autres classes. On obtient les taux d'erreurs indiqués dans le tableau 4.2.

Classes	Hommes_1	Hommes_2	Femmes_1	Femmes_2
Erreur sur le sexe	0 %	0 %	0 %	0 %
Erreur sur la sous-classe	4,5 %	11,5 %	0 %	0 %
Erreur totale	4,5 %	11,5 %	0 %	0 %

Tableau 4.2 : Taux d'erreurs de classement apparents

- **Application des critères de classement au corpus de test**

Les fonctions discriminantes sont appliquées aux corpus de test de la même manière que pour le classement homme / femme. La décision à chaque trame ne se fait plus par comparaison des résultats à un seuil, mais par calcul du score discriminant associé à chaque classe. La classe d'affectation est celle minimisant le score.

La Figure 4.13 représente les erreurs de classement résultant de cette procédure pour la liaison 2. Chaque ligne correspond à un locuteur : les lignes 1 à 33 (en partant du haut) représentent les locuteurs masculins, les lignes 34 à 63 les locutrices. Pour une ligne donnée, chaque pixel représente l'erreur de classement pour une trame : le pixel est gris clair si le locuteur est bien classé, gris foncé si l'on se trompe de sous-classe mais pas de sexe, noir si le locuteur est affecté à une sous-classe du sexe opposé. Seuls les résultats de la deuxième liaison simulée sont représentés, ceux des autres liaisons étant similaires. Le tableau 4.3 indique les taux d'erreur de classement pour les trois liaisons, calculés selon les deux méthodes présentées dans la section IV.2.4. Ces taux d'erreurs sont assez élevés, mais correspondent essentiellement à des erreurs de détermination de la sous-classe du locuteur, les erreurs de détermination du sexe du locuteur étant assez rares.

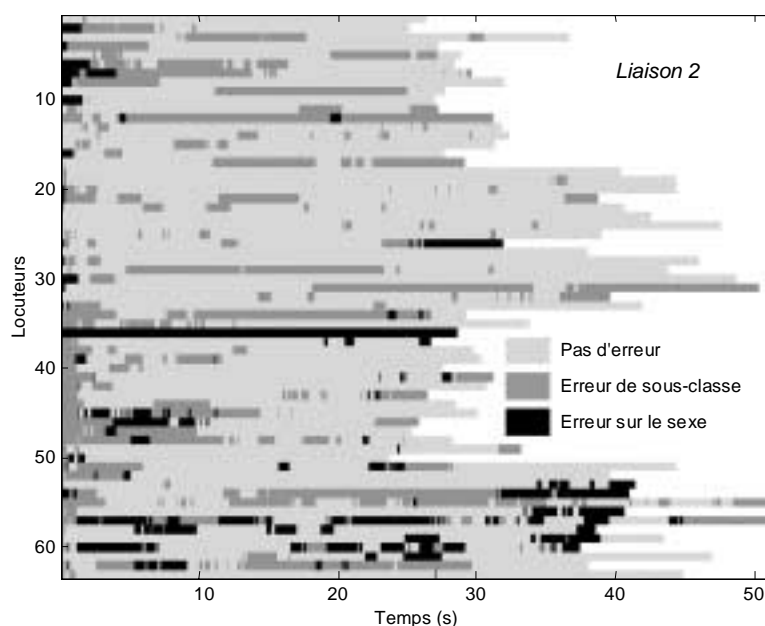


Figure 4.13 : Erreurs de classement pour chaque locuteur à chaque trame d'activité vocale

Liaison	Locuteurs mal classés (1)			Trames mal classées (2)		
	1	2	3	1	2	3
Classement selon F_0 et $C^p_{1 \rightarrow 20}$	32 %	33 %	27 %	24 %	23 %	24 %
<p>(1) Locuteurs pour lesquels une erreur de classement est commise sur plus de 25 % des trames après 10 secondes de parole.</p> <p>(2) Sur l'ensemble des locuteurs, pourcentage de trames mal classées postérieures à 10 secondes</p>						

Tableau 4.3 : Taux d'erreur de classement des locuteurs en quatre classes

IV.3. Égalisation adaptée multiréférences

IV.3.1. Mise en œuvre dans le domaine des cepstres partiels

Selon l'algorithme défini dans le chapitre II, la réponse fréquentielle de l'égaliseur adapté est calculée selon l'équation 2.18, rappelée ici :

$$|EQ(f)| = \frac{1}{|S_{RX}(f) \cdot L_{RX}(f)|} \sqrt{\frac{\gamma_{ref}(f)}{\gamma_x(f)}} \quad (4.30)$$

avec L_{RX} la réponse fréquentielle de la ligne de réception, S_{RX} la réponse fréquentielle du système de réception, γ_{ref} le spectre de référence et γ_x le spectre à long terme de la sortie x du pré-égaliseur. Cette réponse est calculée ainsi uniquement pour les fréquences entre F_c et 3150 Hz, et la valeur $|EQ(f)|_{[F_c-3150 \text{ Hz}]}$ est extrapolée linéairement pour les fréquences inférieures à F_c et supérieures à 3150 Hz.

Cet algorithme peut être transcrit dans le domaine du cepstre partiel, puisque nous disposons du cepstre partiel de la sortie x du pré-égaliseur, nécessaire au classement du locuteur. Ainsi, l'équation (4.30) devient :

$$C^p_{eq} = C^p_{ref} - C^p_x - C^p_{s_{rx}} - C^p_{l_{rx}}, \quad (4.31)$$

où C^p_{eq} , C^p_x , $C^p_{s_{rx}}$ et $C^p_{l_{rx}}$ sont les cepstres partiels respectifs de l'égaliseur adapté, de la sortie x du pré-égaliseur, du système de réception et de la ligne de réception, C^p_{ref} étant le cepstre partiel de référence. Les cepstres partiels sont calculés comme indiqué dans la section IV.1.2, en sélectionnant la bande de fréquences 187-3187 Hz. Ce calcul est effectué uniquement pour les coefficients 1 à 20, les coefficients suivants étant inutiles car représentatifs d'une finesse spectrale qui sera éliminée par la troncature de la réponse impulsionnelle de l'égaliseur adapté.

Pour chacune des classifications envisagées (2 ou 4 classes), nous souhaitons comparer les performances obtenues avec deux cepstres de référence, calculés sur le corpus de la section IV.1:

- cepstre partiel moyen (centre) de tout le corpus ;
- cepstre partiel moyen (centre) de la classe d'affectation du locuteur. Cette classe a été préalablement déterminée par application de la (ou des) fonction(s) linéaire(s)

discriminante(s) au vecteur ayant pour composantes le pitch moyen et les coefficients 1 à 20 de C_x^p .

Les 20 coefficients du cepstre partiel de l'égaliseur adapté calculés selon l'équation (4.31) sont complétés par des zéros de manière à obtenir une représentation sur 256 points. On en déduit le module en dB de la réponse fréquentielle de l'égaliseur adapté restreinte à la bande 187-3187 Hz :

$$EQ_{\text{dB}[187-3187 \text{ Hz}]} = \text{TFD}^{-1}(C_{eq}^p). \quad (4.32)$$

Cette grandeur est décimée d'un facteur 3/4 puis extrapolée en dehors de cette bande comme décrit dans la section II.2.5.

L'égaliseur adapté est ensuite calculé dans le domaine temporel selon la procédure décrite dans la section II.2.5. Comme l'approximation du cepstre partiel du locuteur par le cepstre partiel de référence de sa classe est moins grossière que ne l'était celle de son spectre à long terme par un spectre moyen unique, la réponse fréquentielle de l'égaliseur adapté nécessite un lissage moins fort. Ainsi, le nombre de coefficients de l'égaliseur adapté peut être plus élevé, de manière à corriger des distorsions spectrales plus fines. Nous fixons ce nombre à 31.

Ces différentes étapes de l'égalisation adaptée différenciée par classes de locuteurs sont résumées par le schéma de la Figure 4.14.

IV.3.2. Application à la classification hommes / femmes

La méthode d'égalisation précédemment exposée est simulée dans les conditions expérimentales décrites dans la section II.3.1. Le corpus utilisé est celui défini dans la section IV.1.

Nous noterons e_1 l'erreur cepstrale introduite par l'égalisation lorsque la référence est le cepstre moyen de toute la population, e_2 celle résultant de l'égalisation lorsque la référence est le cepstre moyen de la classe du locuteur. L'erreur cepstrale est définie (*cf.* chapitre II) comme la distance cepstrale entre l'égaliseur adapté et l'égaliseur adapté idéal.

La Figure 4.15 compare, pour la liaison 3 et pour chaque locuteur (les résultats des autres liaisons étant très proches), les moyennes des erreurs cepstrales e_1 et e_2 , notées respectivement \bar{e}_1 et \bar{e}_2 . L'erreur cepstrale moyenne est calculée à partir de 10 s d'activité vocale, de manière à être assuré que la convergence de l'égaliseur a eu lieu. Chaque locuteur est représenté par un point de coordonnées (\bar{e}_1, \bar{e}_2) . Pour la grande majorité des locuteurs, $\bar{e}_2 < \bar{e}_1$, ce qui signifie que l'utilisation du centre de la classe comme cepstre de référence permet bien de réduire l'erreur cepstrale.

Cependant, à cette réduction de l'erreur cepstrale ne correspond aucune amélioration perceptible lors des écoutes informelles pratiquées avec les locuteurs dont nous disposons. Selon les locuteurs, le timbre du signal de réception est dans les deux cas soit semblablement différent de celui du signal en réception de la même liaison égalisée par l'égaliseur idéal, soit très proche de ce signal. La Figure 4.16 représente, pour un locuteur présentant l'une des plus fortes réductions d'erreur cepstrale moyenne (marqué d'une croix sur la Figure 4.15), l'évolution de l'erreur cepstrale au cours du temps. Pour le même locuteur, la différence, en dB, entre la réponse fréquentielle de l'égaliseur adapté et celle de l'égaliseur adapté idéal, après stabilisation de l'erreur cepstrale (*i.e.* après convergence de l'égaliseur), est représentée sur la Figure 4.17, selon le cepstre de référence utilisé. Les deux erreurs spectrales ainsi illustrées sont très proches, ce qui explique la proximité de timbre des deux signaux.

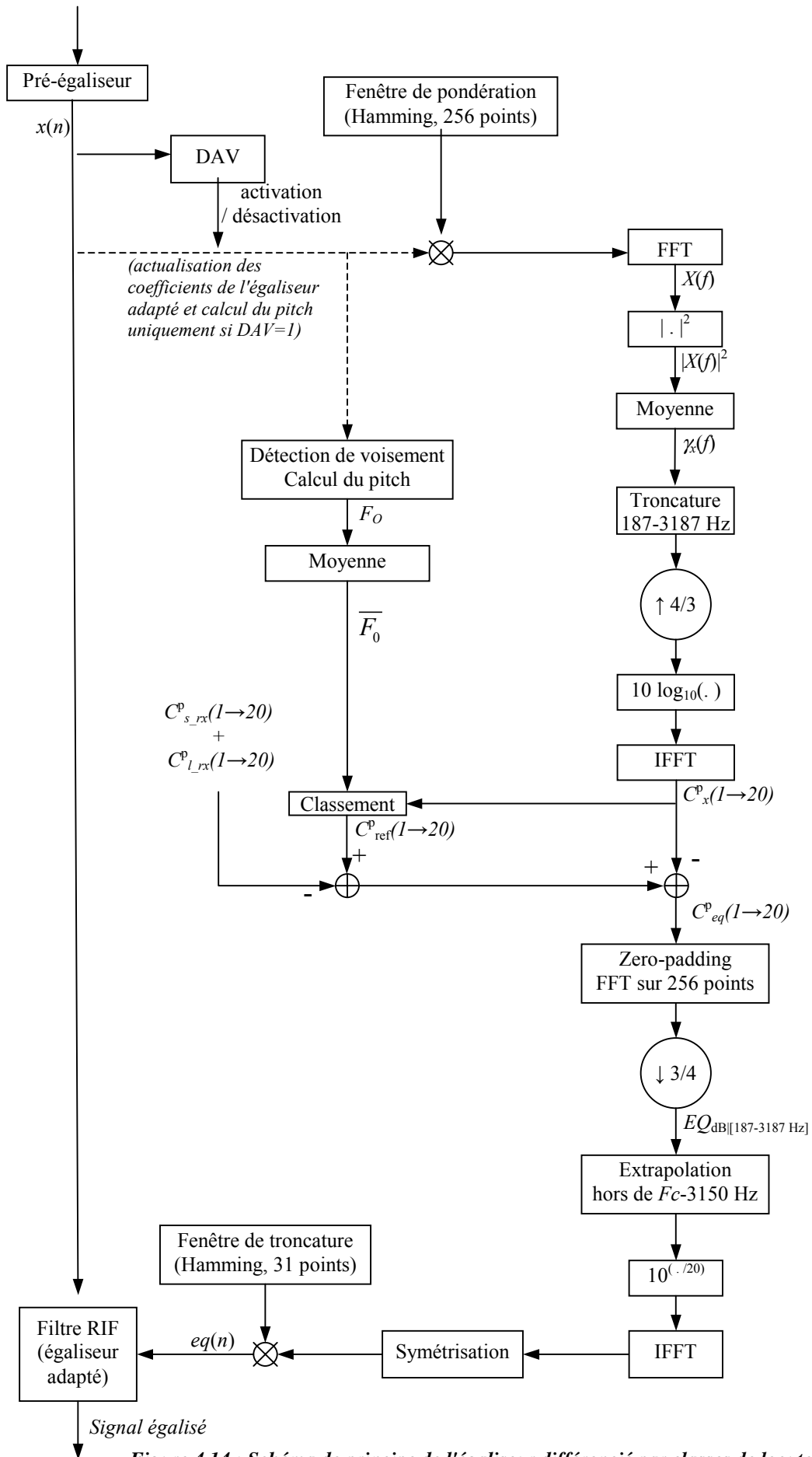


Figure 4.14 : Schéma de principe de l'égaliseur différencié par classes de locuteurs

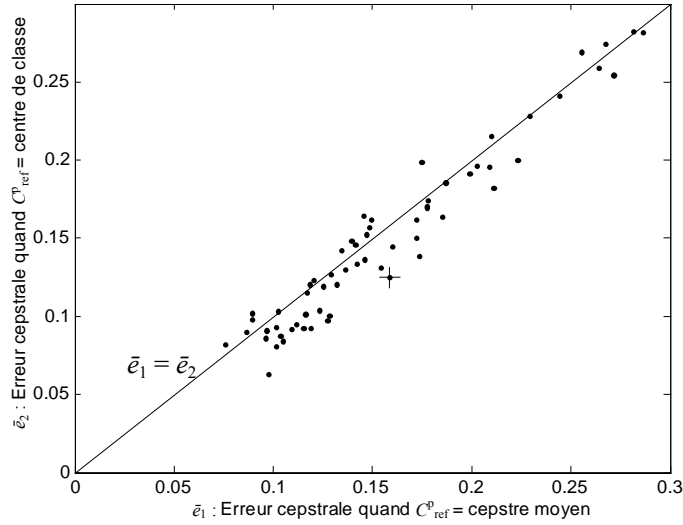


Figure 4.15 : Variation de l'erreur cepstrale moyenne lorsque l'égaliseur adapté utilise comme cepstre partiel de référence le centre de la classe du locuteur au lieu de celui de toute la population

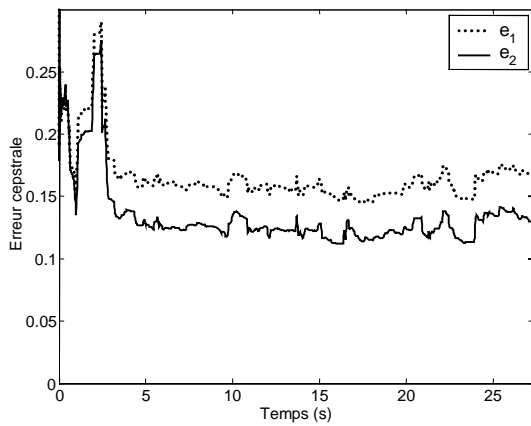


Figure 4.16 : Évolution de l'erreur cepstrale pour un locuteur

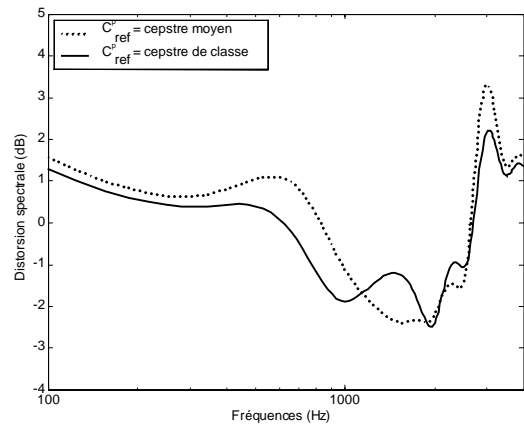


Figure 4.17 : Distorsion spectrale correspondante à la 10^{ème} seconde

Nous cherchons à définir dans quelles conditions la réduction d'erreur cepstrale permise par le recours à un cepstre de référence propre à chaque classe pourrait être perceptible. La première condition est de maximiser l'écart entre e_1 et e_2 . Comme l'erreur cepstrale introduite par l'égaliseur correspond à la distance entre le cepstre original du locuteur et le cepstre de référence utilisé comme approximation de celui-ci, un locuteur appartenant à une classe q remplit cette condition si son cepstre partiel C_v^p vérifie :

$$C_v^p = \underset{C^p}{\text{Argmax}} \left(\left\| C^p - \overline{C^p} \right\| - \left\| C^p - \overline{C^p}^q \right\| \right), \quad (4.33)$$

où $\overline{C^p}$ désigne le centre de la population et $\overline{C^p}^q$ le centre de la classe q ($q = H$ pour les hommes, F pour les femmes). La condition nécessaire et suffisante est que le cepstre partiel C_v^p soit sur la demi-droite $[\overline{C^p} \overline{C^p}^q)$, à l'extérieur du segment $[\overline{C^p} \overline{C^p}^q]$, comme représenté sur la Figure 4.18

pour le plan des deux premiers coefficients cepstraux. L'écart entre $\|C^p - \overline{C^p}\|$ et $\|C^p - \overline{C^p}^q\|$ est alors maximal et est égal à la distance entre ces deux centres.

Il reste à définir à quelle distance de $\overline{C^p}^q$ doit se situer C_v^p . Nous avons comparé, pour les locuteurs du corpus, la différence subjective de timbre entre le signal égalisé et le signal égalisé idéalement, à la distance entre le cepstre partiel du locuteur et le cepstre de référence. La différence de timbre nous a semblé perceptible à partir d'une certaine distance seuil d_{seuil} , approximativement 0,3. Le cepstre partiel C_v^p doit donc vérifier :

$$\begin{cases} \|C_v^p - \overline{C^p}^q\| = d_{seuil} - \varepsilon, \varepsilon > 0 \\ \|C_v^p - \overline{C^p}\| = d_{seuil} - \varepsilon + \|\overline{C^p}^q - \overline{C^p}\| > d_{seuil} \end{cases} \quad (4.34)$$

Ainsi l'erreur cepstrale sera peu perceptible lorsque l'égalisation utilise le centre de classe comme cepstre de référence et le sera plus lorsque l'égalisation utilise le cepstre moyen de la population.

Les locuteurs ainsi définis n'existent pas dans notre corpus et, eu égard aux conditions restrictives que doit vérifier leur cepstre partiel (appartenance au voisinage d'un segment d'un espace à 20 dimensions), il est peu probable de trouver de tels locuteurs dans la réalité. Nous les appellerons donc *locuteurs virtuels* (d'où l'indice "v" dans la notation du cepstre partiel) et les construisons à partir des locuteurs réels du corpus de la manière suivante. Ayant défini pour chaque classe le cepstre partiel des locuteurs virtuels conformément aux conditions (4.34), nous définissons pour chaque locuteur réel appartenant à la classe q un vecteur de *liffrage* dans l'espace du cepstre partiel, $C_{r \rightarrow v}^p$, par :

$$C_{r \rightarrow v}^p = C_v^p - C_r^p, \quad (4.35)$$

où C_r^p est le cepstre partiel du locuteur réel. A chaque locuteur réel correspond un locuteur virtuel, dont le signal de parole est construit par filtrage du signal de parole original du locuteur réel par un filtre temporel $f_{r \rightarrow v}$ issu du vecteur de liffrage $C_{r \rightarrow v}^p$ tel que représenté sur la Figure 4.18. La transformation de $C_{r \rightarrow v}^p$ en $f_{r \rightarrow v}$ est réalisée selon la même procédure que celle décrite dans la section IV.3.1 pour l'obtention du filtre RIF d'égalisation adaptée à partir de C_{eq}^p .

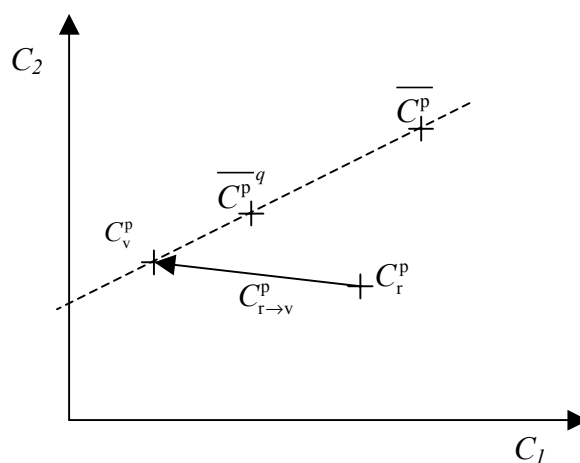


Figure 4.18 : Construction d'un locuteur virtuel à partir d'un locuteur réel

Il est à noter que, pour être virtuels, ces nouveaux locuteurs n'en ont pas moins une voix tout aussi "réelle" et naturelle que celle des locuteurs réels du corpus.

Nous avons testé la perceptibilité de l'amélioration de la correction de timbre pour les locuteurs virtuels, par des écoutes informelles menées par quatre auditeurs experts. Quatre locuteurs virtuels sont testés : 3 hommes (H1_v, H2_v, H3_v) et 1 femme (F_v). Pour chaque locuteur, trois fichiers de 3 à 5 s sont soumis à l'écoute, correspondant à une portion de la phrase test traitée par trois égaliseurs différents :

- égaliseur idéal (sur la bande 187-3187 Hz) (ID)
- égaliseur adapté avec comme référence le cepstre moyen (EG1)
- égaliseur adapté avec comme référence le centre de la classe du locuteur (EG2)

Le nombre de coefficients de l'égaliseur adapté a été porté à 41 au lieu de 31, de manière à accentuer l'écart entre EG1 et EG2. L'évolution de l'erreur cepstrale des signaux test EG1 et EG2 est représentée sur la Figure 4.19 pour les quatre locuteurs virtuels testés, comparée à celle obtenue avec les locuteurs réels correspondants. Parallèlement sont représentées les distorsions spectrales de EG1 et EG2.

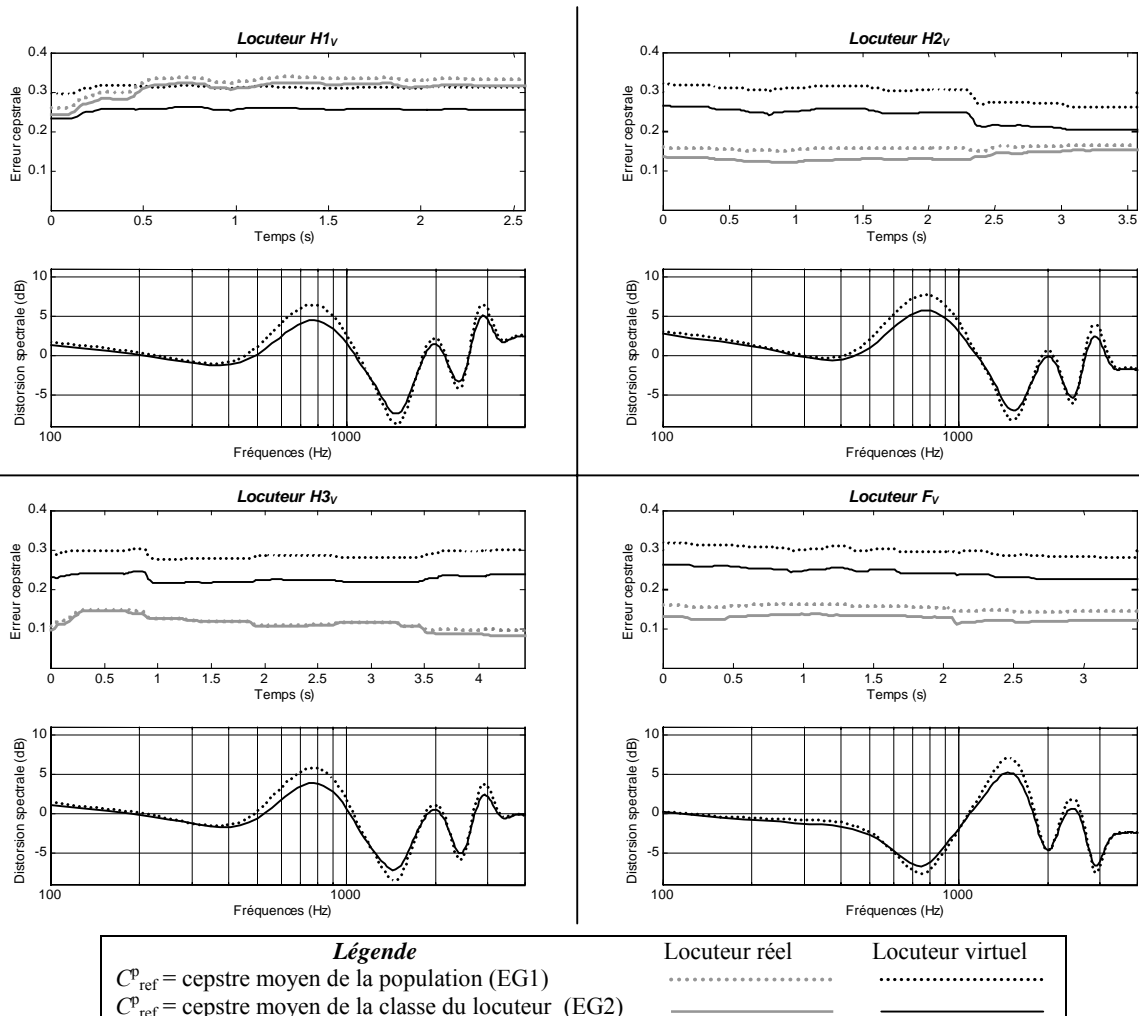


Figure 4.19 : Erreurs cepstrale et spectrale des signaux test EG1 et EG2

Le test consiste, après une libre écoute des fichiers, à comparer ID et EG1 d'une part, ID et EG2 d'autre part, et à indiquer :

- pour quelle paire la différence de timbre est la plus nette ;
- éventuellement, si aucune différence de timbre n'apparaît dans une des paires.

Dans toutes les paires, les auditeurs perçoivent une différence entre les deux échantillons. Un éventuel seuil de perception de l'erreur cepstrale se situerait donc en dessous de 0,2 (d'après les erreurs cepstrales de la Figure 4.19). Pour les trois locuteurs masculins, les quatre sujets jugent que EG2 est plus proche de ID que EG1, ce qui est conforme aux résultats objectifs. En revanche, deux des quatre auditeurs jugent que pour la locutrice F_V, EG1 est plus proche de ID que EG2.

Il semble donc, d'après ce test, que, dans certaines conditions de position du cepstre partiel du locuteur par rapport au centre de la population et au centre de la classe du locuteur, l'utilisation d'une référence par classe plutôt que d'une référence unique permet d'améliorer la correction de timbre par l'égaliseur adapté. Cependant, les auditeurs experts ont exprimé une réelle difficulté à exercer leur jugement : les différences de timbre entre EG1 et EG2 sont très peu perceptibles, ce qui limite l'intérêt d'une classification des locuteurs en deux classes seulement, déjà restreint par la rareté des conditions à vérifier sur le cepstre partiel du locuteur.

IV.3.3. Application à la classification en quatre classes

Comme pour la classification en deux classes, la méthode d'égalisation exposée dans la section IV.3.1 est simulée dans les conditions expérimentales décrites dans la section II.3.1, avec le corpus défini dans la section IV.1.

Nous noterons e_1 l'erreur cepstrale introduite par l'égalisation lorsque la référence est le cepstre moyen de toute la population, e_4 celle résultant de l'égalisation lorsque la référence est le cepstre moyen de la classe du locuteur.

La Figure 4.20 compare, pour la liaison 3 et pour chaque locuteur, les moyennes des erreurs cepstrales e_1 et e_4 , notées respectivement \bar{e}_1 et \bar{e}_4 . Chaque locuteur est représenté par un point de coordonnées (\bar{e}_1, \bar{e}_4) . Pour la grande majorité des locuteurs, \bar{e}_4 est inférieure à \bar{e}_1 , et ce de manière beaucoup plus nette que le résultat $\bar{e}_2 < \bar{e}_1$ dans le cas d'une classification en 2 classes.

Cette réduction de l'erreur cepstrale nous a semblé perceptible pour certains locuteurs. Nous proposons d'évaluer cette perceptibilité de l'amélioration par un test subjectif, en utilisant la méthode MUSHRA [UIT-R/BS.1534, 1996], modifiée comme indiqué dans la section II.3.6.

Pour chaque locuteur à tester, la liaison égalisée est simulée sur tout le texte "*la bise et le soleil*", dont une portion de 6 à 7 s d'activité vocale, prononcée après la convergence de l'égaliseur, est présentée aux auditeurs ; le traitement est effectué par trois égaliseurs différents :

- égaliseur idéal (sur la bande 187-3187 Hz) (ID) ;
- égaliseur adapté avec comme référence le cepstre moyen de toute la population (EG1) ;
- égaliseur adapté avec comme référence le centre de la classe du locuteur (EG4).

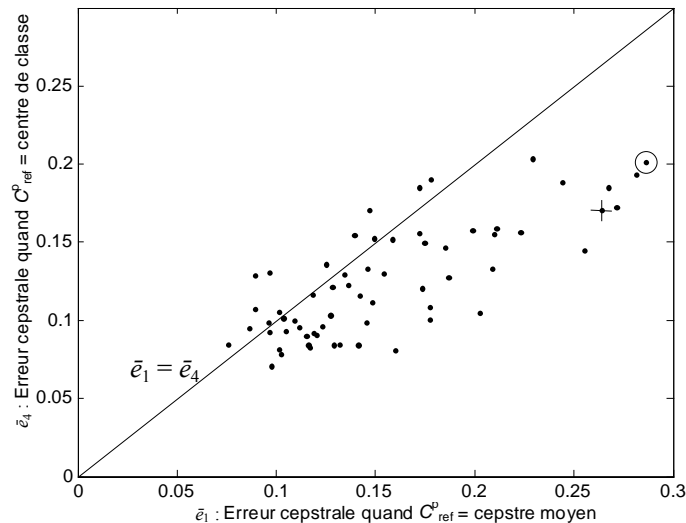


Figure 4.20 : Variation de l'erreur cepstrale moyenne lorsque l'égaliseur adapté utilise comme cepstre partiel de référence le centre de la classe du locuteur au lieu de celui de toute la population

La référence est le signal ID, les signaux tests sont les signaux ID, EG1 et EG4. Les auditeurs attribuent à chaque signal test une note de proximité de timbre avec la référence, entre 0 (timbre très différent) et 100 (timbre identique).

Le choix des locuteurs à tester est guidé par les objectifs suivants :

- Il s'agit d'abord de vérifier qu'il existe des locuteurs pour lesquels EG4 est significativement plus proche de ID que EG1. Nous choisirons donc des locuteurs pour lesquels à la fois les résultats objectifs (en termes d'écart entre \bar{e}_1 et \bar{e}_4) et les écoutes informelles préliminaires permettent d'augurer de tels résultats de test. Les deux locuteurs représentés respectivement par une croix et un cercle sur la Figure 4.20 correspondent à ce cas. Ces locuteurs ont un intérêt particulièrement démonstratif, puisque, présentant une erreur moyenne \bar{e}_1 assez forte, ce sont ceux pour lesquels une réduction d'erreur cepstrale est le plus souhaitable si elle correspond à une réduction de l'erreur de restauration de timbre.
- Par ailleurs, nous souhaitons étudier comment la perception de la différence de timbre évolue selon l'erreur cepstrale, afin notamment de répondre aux questions suivantes : une même réduction de l'erreur cepstrale entre EG1 et EG4 conduit-elle dans tous les cas à la même perception de rapprochement du timbre de celui de ID ? Existe-t-il un seuil d'erreur cepstrale en deçà duquel aucune différence de timbre n'est perceptible ? L'égalisation différenciée par classes perdrait en effet singulièrement de son intérêt si pour la plupart des locuteurs \bar{e}_1 était déjà inférieure à ce seuil. Pour étudier ces relations entre le timbre et l'erreur cepstrale, nous utilisons 3 locuteurs virtuels. L'emploi d'un locuteur virtuel permet en effet de fixer arbitrairement sa position dans l'espace du cepstre partiel et, partant, les erreurs cepstrales respectives de EG1 et EG4. Ces trois locuteurs virtuels sont construits à partir d'un seul locuteur réel masculin selon la procédure décrite dans la section IV.3.2, et placés à des distances respectives de 0,01, 0,10 et 0,20 du centre de la classe du locuteur réel. La transformation du locuteur réel en locuteur virtuel change peu la voix du locuteur, de sorte que cette construction des trois locuteurs virtuels à partir d'un locuteur réel unique permet de s'affranchir d'un éventuel effet de la voix du locuteur dans l'étude de la relation timbre-erreur cepstrale.

Nous testons donc 5 locuteurs : 2 locuteurs réels (masculins) R1 et R2 ; 3 locuteurs virtuels (masculins) V1, V2 et V3.

Le test est effectué par dix-huit auditeurs experts, dans les mêmes conditions matérielles que le test présenté au chapitre II (en particulier écoute binaurale sur casque fermé de haute qualité et niveau d'écoute confortable ajustable par chaque sujet). Cinq séquences de test, correspondant aux cinq locuteurs, sont présentées aux sujets. Elles sont précédées de la lecture de consignes similaires à celles de l'annexe A, ainsi que d'une séquence d'apprentissage permettant aux sujets de se familiariser avec l'interface et le type de dégradations rencontrées dans le test. La durée totale de l'expérience est de 15 à 20 mn par sujet.

Les résultats du test sont présentés sur les Figures 4.21 et 4.22, en regard des erreurs cepstrales et distorsions spectrales des échantillons notés. Les résultats sont conformes aux résultats objectifs : EG4 est jugé plus proche de ID que EG1, avec un écart de 20 points environ entre les notes moyennes de EG1 et EG4. Malgré la faible perceptibilité de la différence de timbre entre EG1 et EG4, cette différence observée au cours de l'expérience est significative, dans la mesure où les intervalles de confiance ne se recouvrent pas.

Pour tous les locuteurs, EG1 et EG4 sont jugés significativement différents de ID, sauf pour le locuteur V1, pour lequel EG4 obtient la même note moyenne que ID. Il existe donc vraisemblablement un seuil de perception de l'erreur cepstrale compris entre 0,05 et 0,1. Le signal EG4 du locuteur V1 présente en effet une erreur cepstrale de 0,05, tandis que l'erreur cepstrale est de 0,1 pour EG4 du locuteur V2, qui est jugé différent de ID.

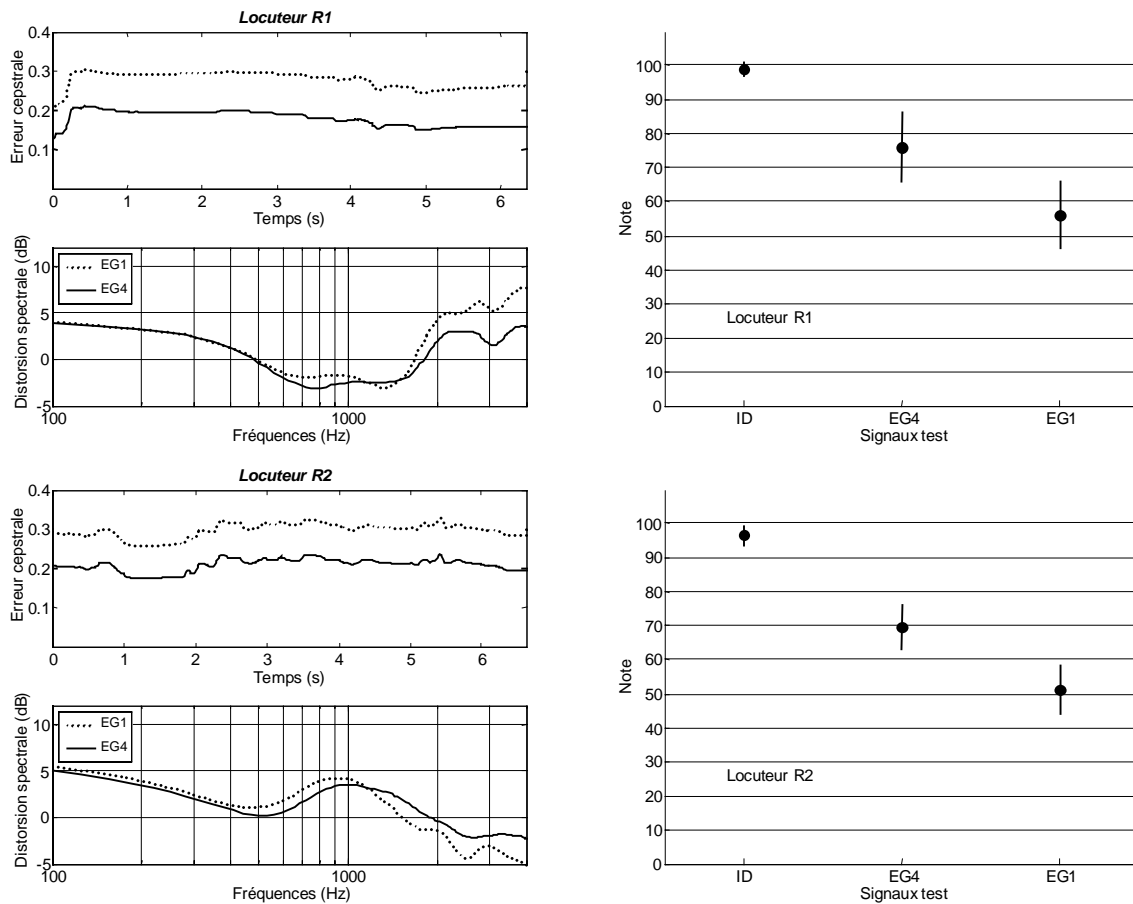


Figure 4.21 : Erreurs cepstrales, distorsions spectrales et notes moyennes pour les deux locuteurs réels

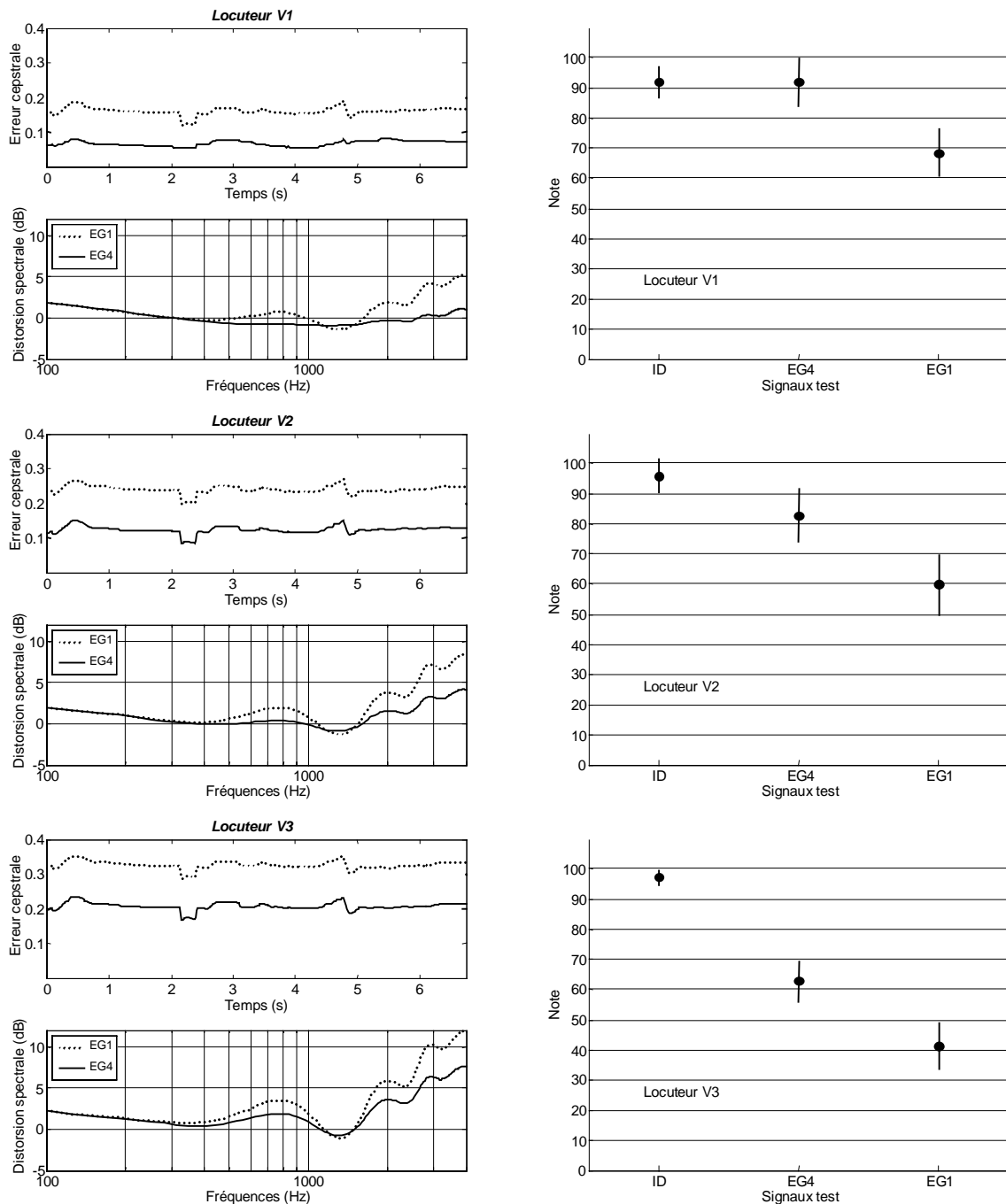


Figure 4.22 : Erreurs cepstrales, distorsions spectrales et notes moyennes pour les trois locuteurs virtuels

La différence d'erreur cepstrale entre EG1 et EG4 est d'environ 0,1 pour tous les locuteurs. A cette constance de la différence d'erreur cepstrale correspond une constance dans l'écart entre les notes moyennes correspondantes, et ce indépendamment du niveau de e_1 . En d'autres termes, l'amélioration de timbre perçue résultant d'une réduction d'erreur cepstrale ne semble dépendre que de l'amplitude de cette réduction. De manière plus générale, nous observons même une relation linéaire entre l'erreur cepstrale et la note moyenne du signal test. Celle-ci apparaît sur la Figure 4.23, où sont représentés pour chaque locuteur ses trois signaux tests ID, EG1 et EG4 dans le plan (erreur cepstrale moyenne, note moyenne).

Compte-tenu de cette dernière observation, nous faisons l'hypothèse que, pour un même type de voix, il existe une relation linéaire entre la note moyenne d'un signal dégradé (EG1 ou

EG4) et son erreur cepstrale. Nous vérifions cette hypothèse par le test de Bravais-Pearson [Guéguen, 1998] de corrélation linéaire entre ces deux variables, pour les signaux EG1 et EG4 des trois locuteurs virtuels (dont les voix sont proches, puisque issues du même locuteur réel). Le coefficient de corrélation entre ces six observations vaut $-0,9891$, pour une valeur seuil de $-0,8116$ (avec un risque d'erreur de 5 %). L'hypothèse est donc bien vérifiée.

Eu égard à l'effet de seuil apparaissant pour le locuteur V1 et à l'éloignement des points représentatifs de ID de la droite correspondant à cette relation linéaire, la relation entre la note et l'erreur cepstrale n'est vraisemblablement pas linéaire sur tout l'espace des erreurs cepstrales, mais est plutôt de type "sigmoïde", avec une partie constante de 0 à une valeur seuil, une partie centrale linéaire décroissante et une partie constante à partir d'une certaine valeur "de saturation". Sous l'hypothèse d'une valeur seuil de 0,07, cette fonction est représentée sur la Figure 4.23, pour ses deux premières parties, par la courbe grise.

Les points correspondant au locuteur R1 respectent la même relation linéaire, tandis que ceux du locuteur R2 suivent la même pente mais sont décalés vers le haut. Il est intéressant de mettre en relation cette observation avec celle des distorsions spectrales des signaux EG1 et EG4 des différents locuteurs. Les distorsions subies par R1 ont en effet une forme très proche de celles des signaux tests des locuteurs virtuels, contrairement à celles de R2. La relation linéaire semble donc être de pente constante et d'ordonnée à l'origine paramétrée par le type de distorsion spectrale, plus que par le type de voix. Ces hypothèses nécessiteraient cependant plus de résultats pour être validées rigoureusement.

Précisons enfin que cette relation linéaire entre l'erreur cepstrale et la perception de dégradation du timbre dépend de la gamme des distorsions présentées aux auditeurs, puisque dans l'expérience présentée au chapitre II, avec des signaux tests présentant des écarts de timbre avec ID nettement plus importants, la note de EG1 était très proche de celle de ID (à comparer avec l'écart de 50 points ici).

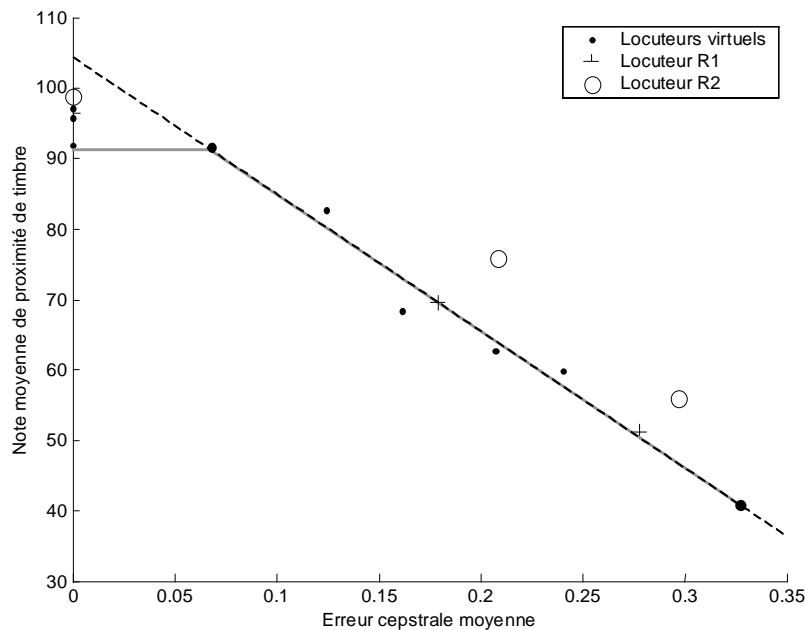


Figure 4.23 : Corrélation entre erreur cepstrale et note moyenne

IV.4. Conclusion

La classification des locuteurs sur la base de leur spectre à long terme restreint à la bande d'égalisation, représenté par le *cepstre partiel à long terme*, fait apparaître l'existence de deux classes de locuteurs correspondant sensiblement à classification en hommes et femmes, séparables chacune en deux sous-classes, pertinentes selon le critère de la variance. Cette classification permet d'envisager une égalisation adaptée utilisant un spectre de référence spécifique à chaque classe de locuteurs plutôt qu'un spectre de référence unique approchant grossièrement le spectre de chaque locuteur.

La difficulté du classement d'un locuteur selon l'observation de son cepstre partiel à l'entrée de l'égaliseur tient à la déviation de celui-ci par le filtrage de la liaison téléphonique. C'est pourquoi nous avons défini des critères de classement robustes aux biais cesptraux introduits par les liaisons. Cette robustesse est obtenue d'une part en combinant au paramètre cepstre partiel le paramètre pitch, peu sensible aux filtrages habituels des liaisons téléphoniques et pertinent du point de vue de la classification hommes / femmes, d'autre part par le choix du corpus d'apprentissage des fonctions discriminantes. Les individus des corpus sont représentatifs à la fois d'une grande variété de locuteurs et d'une diversité de filtrages correspondant à celle des liaisons téléphoniques. Les critères de classement ainsi définis, appliqués à nos corpus de test, aboutissent à des taux de trames mal classées de 11 à 15 % pour un classement homme / femme, de 25 % pour un classement en quatre classes.

L'égalisation différenciée par classe (utilisant un cepstre de référence par classe), permet de réduire l'erreur cespstrale de l'égaliseur pour la plupart des locuteurs. Dans le cas d'une classification en 2 classes, cette réduction de l'erreur cepstrale n'est pas perceptible. Si l'on classe les locuteurs en 4 classes, elle correspond pour certains locuteurs à une légère amélioration de la correction de timbre, d'après les tests subjectifs mis en œuvre. Ces tests mettent par ailleurs en évidence une relation linéaire entre l'erreur cepstrale et la perception de la dégradation du timbre.

Conclusion

Notre objectif était de corriger les distorsions spectrales subies par la parole sur le réseau téléphonique, afin de donner au signal de réception un timbre le plus proche possible de celui de la voix originale du locuteur.

Il n'est naturellement pas envisageable de restaurer le signal original sur toute la bande 0-4000 Hz, eu égard à la forte atténuation de son spectre aux extrémités de cette bande. C'est pourquoi nous avons proposé un algorithme d'égalisation spectrale aveugle, centralisée dans le réseau, qui compense le filtrage de la liaison téléphonique sur une bande de fréquences limitée, typiquement 200-3150 Hz. Cette égalisation consiste à aligner, sur cette bande, le spectre à long terme du signal traité sur un spectre de référence, pris comme une approximation du spectre à long terme original (inconnu) du locuteur. Il en résulte, pour la majorité des locuteurs, une distorsion spectrale très faible, sur la bande d'égalisation, entre le signal de réception et le signal original. Les tests formels subjectifs ont montré que le timbre de la voix perçue en réception de la liaison égalisée reste certes assez éloigné du timbre original, du fait de la restriction de la bande d'égalisation, mais en est significativement plus proche que la voix en réception de la même liaison sans égaliseur.

L'objectif de correction des distorsions spectrales et de leur corollaire que sont les déformations du timbre est donc atteint par ce premier algorithme, validé à la fois par des tests formels sur sa version simulée et par une implantation temps réel sur un autocommutateur expérimental. Cependant, il est apparu au cours des expérimentations que la quantification en loi A des échantillons de sortie de l'égaliseur induit un bruit notable en réception. D'autre part, l'erreur d'approximation du spectre à long terme de tous les locuteurs par un spectre de référence unique conduit, pour certains locuteurs, à des distorsions spectrales importantes, qui se traduisent par une restauration du timbre non optimale. Nous nous sommes donc attachés à dépasser ces deux limites de la méthode d'égalisation spectrale.

Nous avons cherché à réduire le bruit de quantification perçu, par un reformage spectral de celui-ci exploitant les propriétés de masquage fréquentiel de la parole. Deux méthodes originales de reformage du bruit de quantification ont été proposées, l'une fondée sur une réinjection à l'entrée du quantificateur de l'erreur de quantification filtrée, l'autre consistant à chercher une quantification optimale selon un critère probabiliste. Ces deux méthodes permettent de masquer le bruit de quantification sur une large part du signal, mais le masquage est imparfait, laissant apparaître sporadiquement un bruit "rauque". L'évaluation subjective du reformage du bruit met en évidence une préférence pour le bruit de quantification non reformé, mais également une grande tolérance des auditeurs au bruit de quantification, reformé ou non. Les auditeurs préfèrent la voix traitée par l'égaliseur et entachée de bruit de quantification (inhérent à cette égalisation) à la même voix en réception de la même liaison dépourvue d'égaliseur et non bruitée. Ainsi, si l'objectif d'une égalisation non bruyante n'est pas atteint, nous sommes parvenus à un bon compromis entre la restauration du timbre et l'acceptabilité du bruit induit.

La méthode d'égalisation spectrale aveugle a été affinée en cherchant à réduire l'erreur d'approximation du spectre à long terme original de chaque locuteur par le spectre de référence. Nous avons mis en évidence la possibilité de constituer deux ou quatre classes de locuteurs

pertinentes selon le critère de la variance, sur la base du spectre à long terme restreint à la bande d'égalisation. Ainsi, nous disposons non plus d'un spectre de référence unique, mais d'un spectre de référence par classe (le centre de celle-ci), chacun constituant une approximation moins grossière des spectres à long terme des locuteurs de sa classe. Nous avons défini des critères de classement assez robustes aux distorsions spectrales introduites par les liaisons téléphoniques, qui permettent de classer un locuteur selon son spectre à long terme, avec une erreur de 25 % dans le cas d'une classification en quatre classes. L'utilisation d'un spectre de référence par classe dans l'égaliseur permet de réduire la distorsion spectrale entre le signal reçu et le signal émis. Dans le cas d'une classification en quatre classes, cette réduction de distorsion se traduit, d'après les tests subjectifs réalisés, par une amélioration significative de la restauration du timbre pour certains locuteurs, sans que cette restauration ne soit altérée pour les autres.

Dans la définition des objectifs de la correction de timbre, nous avons lié implicitement l'atteinte de l'objectif de restauration du timbre à celui de la réduction de la distorsion spectrale sur la bande d'égalisation (200-3150 Hz). Les tests subjectifs réalisés dans le chapitre IV justifient *a posteriori* partiellement ce postulat, en ce qu'ils font apparaître une corrélation linéaire entre la proximité de timbre et la mesure de la distorsion spectrale par l'erreur cepstrale. Ces tests prennent toutefois comme référence le signal original restreint à la bande 200-3400 Hz. Les résultats des tests du chapitre II permettent de compléter la justification de l'hypothèse initiale, en montrant que de deux signaux issus de l'original restreint à la bande 200-3400 Hz, le plus proche subjectivement de l'original est celui présentant le moins de distorsion spectrale sur cette bande.

Nous disposons donc d'un algorithme capable d'approcher assez finement, en aveugle, le spectre original d'une voix et ainsi de restaurer le timbre de cette voix en réception d'une liaison téléphonique, dans la limite de la bande d'égalisation 200-3150 Hz. L'algorithme proposé peut encore être affiné.

L'amélioration la plus évidente, que ne permettait pas la taille de notre corpus, consisterait à étudier une classification plus fine des locuteurs selon leur spectre à long terme. Ainsi pourrait-on réduire l'erreur cepstrale de l'égaliseur jusqu'au seuil de perceptibilité. Notons cependant que les tests de comparaison de timbre ont été réalisés dans des conditions plus discriminantes que celles d'une communication téléphonique : écoute binaurale sur casque de haute qualité et, pour le test de la classification en quatre classes, évaluation par des auditeurs experts. Pour une écoute sur combiné téléphonique par des non-experts, une plus grande finesse dans l'approximation du spectre original de chaque voix n'est peut-être pas nécessaire.

La tolérance des auditeurs au bruit de quantification induit par l'égalisation n'est naturellement pas une solution pleinement satisfaisante au problème du masquage du bruit de quantification. La simulation de nos algorithmes de reformage spectral du bruit semble indiquer que ce problème est très contraint et que le masquage du bruit est sans doute impossible pour certains spectres de signaux. L'information relative au masque peut également ne pas être totalement fiable (cas des zones non stationnaires). Des solutions sont envisagées pour remédier à ces problèmes : elles reposent sur la détection des zones "à risque" (à la fois dans le temps et dans le domaine fréquentiel) et une modification du traitement (égalisation et reformage du bruit) en conséquence.

Enfin, en vue d'une mise en œuvre sur le réseau des méthodes proposées, l'interaction de celles-ci avec les autres fonctions de traitement de la parole (débruitage et annulation d'écho notamment) doit être étudiée de manière plus précise. D'autre part, les évaluations subjectives réalisées sur des fichiers traités par simulations nécessitent d'être complétées par une validation formelle *in situ*, dans les conditions habituelles de communication, en situation de conversation téléphonique réelle.

Annexe A :

Consignes du test d'évaluation de l'égaliseur

Procédure du test MUSHRA pour l'évaluation de la correction du timbre de la parole téléphonique

Le timbre de la voix en réception d'une liaison téléphonique est dégradé de diverses manières : atténuation des basses, voix étouffée ou au contraire trop claire, effet robotique. Le but du test est d'évaluer des dispositifs de correction du timbre de la parole téléphonique, en notant la proximité du timbre corrigé avec le timbre original.

La session est composée de 10 séquences de test, et dure 50 mn. Il est conseillé de faire une pause au milieu.

1. Mise en route

Si le logiciel n'est pas lancé, double-cliquer sur `c:\cinema2\SEAQ1.36\SeaqSTMms(AES).exe`. Une interface comme celle de la page 2 apparaît.

Ouvrir la session "correc_timbre.ses" (file|open session) du répertoire `c:\Gael`

Créer une feuille de résultats (file|New Score Sheet) du type "Nom_Pénom.sco" à sauver dans le répertoire `c:\Gael`. Le logiciel demande le prénom, le nom et l'âge de l'auditeur.

2. Mode entraînement

On est prêt pour la phase d'entraînement pour laquelle les scores ne sont pas pris en compte dans le fichier *.sco. On remarque que la barre des menus reste visible.

Le test consiste à comparer les fichiers test entre eux et avec la référence, et à noter sur une échelle de 0 à 100 la proximité de timbre entre chaque fichier test et la référence. Par exemple, pour un fichier test A, si le timbre de A est identique à celui de la référence, A sera noté 100. Si le timbre de A est très éloigné de celui de la référence, A sera noté 0. Si deux fichiers test ont le même timbre, ils auront la même note.

Ne pas tenir compte de l'échelle "Excellent ... Bad" de l'interface. Attribuer les notes de la manière suivante :

- **timbre identique : 100**
- **timbre très proche : 80 à 100**
- **timbre assez proche : 60 à 80**
- **timbre moyennement proche : 40 à 60**
- **timbre assez différent : 20 à 40**
- **timbre très différent : 0 à 20**

Le bouton référence REF est à gauche, suivi vers la droite des boutons correspondants aux fichiers à tester (A, B, C, D, E). Au dessus de chaque bouton se trouve un curseur permettant de juger sur une échelle continue la proximité de timbre avec l'original. L'auditeur indique sa note en déplaçant le curseur avec la souris. Le fichier noté est celui couramment sélectionné (en cliquant sur le bouton) et affiché en rouge.

Si on le désire, on peut écouter en continu (symbole loop) et commuter entre les différents fichiers comme on le désire en cliquant sur les boutons REF, A, B, ... On peut également sélectionner des parties du signal par les curseurs du bas et effectuer l'évaluation en ayant isolé ces parties.

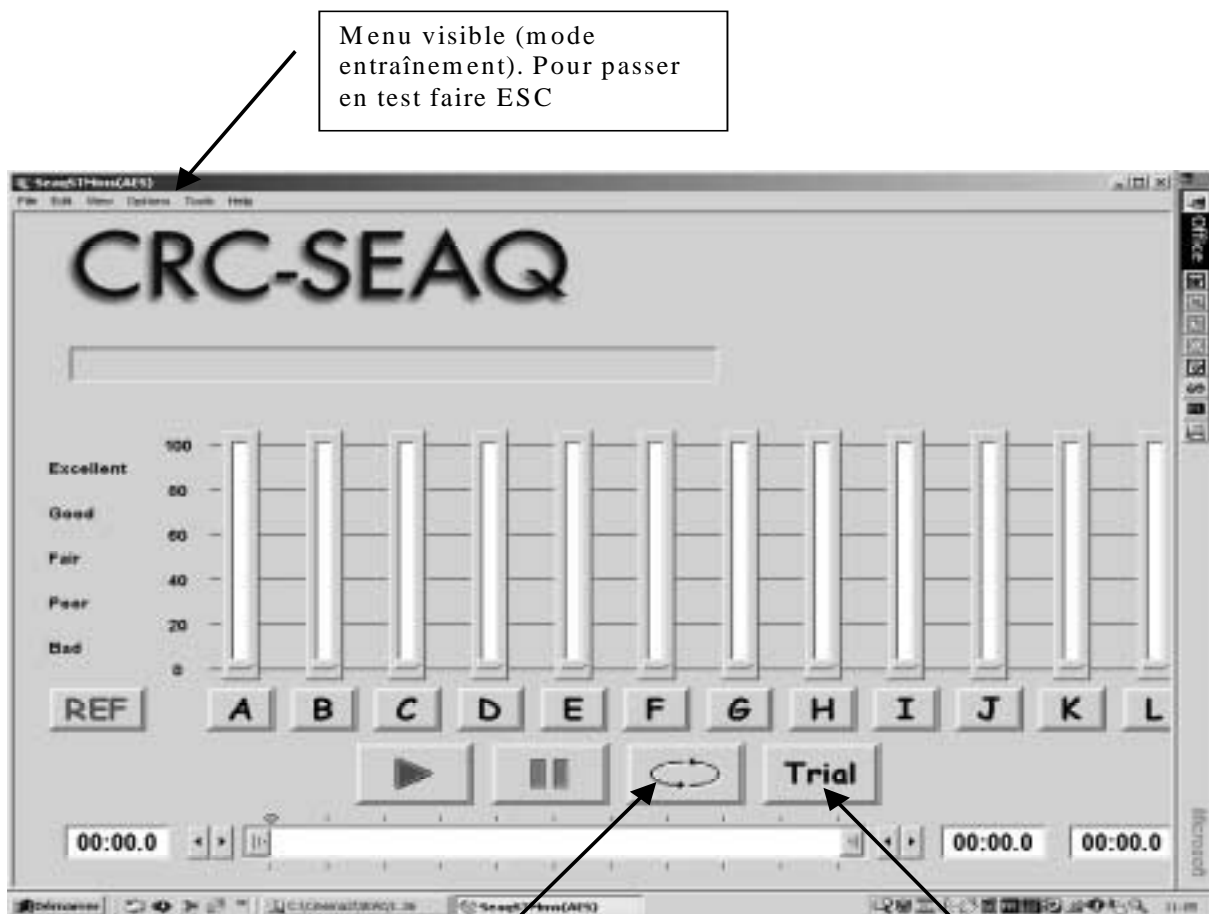
Cette présentation de l'interface permet non seulement au sujet d'évaluer les signaux à tester par rapport à la référence, mais aussi de les comparer entre eux.

3. Phase de test

On appuie sur ESC pour entrer en mode test. On remarque que la barre des menus disparaît. On peut maintenant ajuster les potentiomètres en fonction de la proximité de timbre avec la référence. Pour passer à la séquence suivante, il suffit d'appuyer sur la touche "Trial", le logiciel demande la confirmation des votes et passe à la séquence suivante.

Lorsque toutes les séquences ont été écoutées, le logiciel indique "Session is now over". Il faut alors faire ESC pour laisser la place à un autre sujet.

L'interface du logiciel de Test SEAQ



Menu visible (mode entraînement). Pour passer en test faire ESC

Mode en boucle (joue en continue)

Passer à la séquence sonore suivante

Annexe B :

Résultats des Tukey tests du chapitre II

Dans chaque tableau sont représentés sur la deuxième ligne les notes moyennes, tous locuteurs confondus, des différents signaux test. La valeur inscrite dans une cellule à l'intersection de la ligne X et de la colonne Y est l'indice de significativité de la différence entre la note moyenne de X et celle de Y. La différence est considérée comme significative si cet indice est inférieur à 0,05.

Liaison 1

	ORI	ID	EG	PRE	TRANSP
<i>Note moyenne</i>	96,46875	57,47917	51,14583	41,47917	27,46875
ORI		0,0001176	0,0001176	0,0001176	0,0001176
ID	0,0001176		0,2609154	0,00013024	0,0001176
EG	0,0001176	0,2609154		0,02144521	0,0001176
PRE	0,0001176	0,00013024	0,02144521		0,00030142
TRANSP	0,0001176	0,0001176	0,0001176	0,00030142	

Liaison 2

	ORI	ID	EG	PRE	TRANSP
<i>Note moyenne</i>	96,66666	53,12500	45,30208	51,11458	20,30208
ORI		0,0001176	0,0001176	0,0001176	0,0001176
ID	0,0001176		0,04012495	0,94731653	0,0001176
EG	0,0001176	0,04012495		0,21612108	0,0001176
PRE	0,0001176	0,94731653	0,21612108		0,0001176
TRANSP	0,0001176	0,0001176	0,0001176	0,0001176	

Liaison 3

	ORI	ID	EG	PRE	TRANSP
<i>Note moyenne</i>	98,56250	53,04167	47,31250	31,06250	38,89583
ORI		0,0001176	0,0001176	0,0001176	0,0001176
ID	0,0001176		0,36405712	0,0001176	0,00028324
EG	0,0001176	0,36405712		0,00012726	0,0639798
PRE	0,0001176	0,0001176	0,00012726		0,09973496
TRANSP	0,0001176	0,00028324	0,0639798	0,09973496	

Annexe C

Principes du masquage fréquentiel

Lorsque deux sons de fréquences centrales proches sont présentés simultanément, l'un peut être inaudible. Ce phénomène s'appelle le masquage fréquentiel.

Des expériences de Zwicker [Zwicker, 1981] ont permis de préciser ce masquage dans le cas d'un bruit masquant une tonale. La Figure C.1 présente les courbes d'effet de masque de bruits de bande étroite de fréquence centrale 1 kHz, de largeur de bande 160 Hz, et de niveaux acoustiques L_g 100, 80, 60, 40 et 20 dB. Ces courbes correspondent, pour chaque abscisse f_T , au seuil d'audition d'une tonale (son de fréquence pure) de fréquence f_T en présence du bruit de niveau L_g . La courbe inférieure correspond au seuil d'audition absolu, c'est-à-dire en l'absence de bruit.

Ces courbes mettent en évidence le fait que l'effet de masque du bruit s'étend au-delà de sa bande de fréquences : le seuil d'audition des tonales est modifié sur une large bande de fréquences autour de la fréquence centrale du bruit de bande étroite, avec un maximum en cette fréquence centrale. Notons l'asymétrie du masquage : les courbes d'effet de masque décroissent plus rapidement vers les basses fréquences. Cette asymétrie est d'autant plus marquée que le niveau du bruit est élevé.

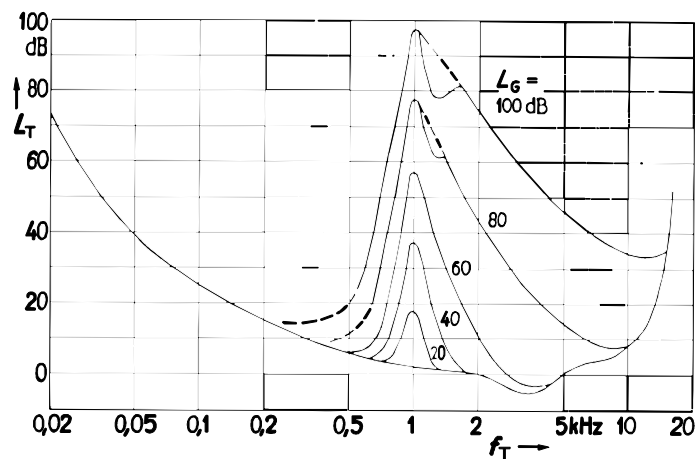


Figure C.1 : Courbes d'effet de masque de bruits à bande étroite

Le cas d'une tonale masquant un bruit a été étudié par Schroeder *et al.* [Schroeder, 1979]. La Figure C.2 représente le seuil d'audition d'un bruit de bande étroite (160 Hz) de fréquence centrale 1 kHz, en présence d'une tonale d'intensité 80 dB, selon la fréquence de la tonale. La décroissance de la courbe est plus forte vers les hautes fréquences. On retrouve donc la même asymétrie du masquage que dans le cas d'un bruit masquant une tonale, sachant que l'axe des abscisses représente ici non plus la fréquence du signal masqué mais celle du signal masquant.

Notons qu'il est plus difficile de masquer un bruit par une tonale que l'inverse : alors que dans l'expérience de Zwicker le maximum du seuil d'audition des tonales est à 4 dB en dessous

du niveau du bruit de bande, la courbe de la Figure C.2 présente un maximum à 56 dB, soit 24 dB en dessous du niveau de la tonale masquante.

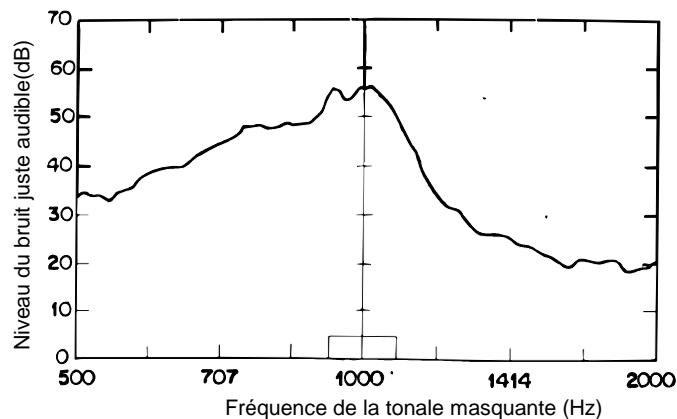


Figure C.2 : Seuil d'audition d'un bruit de bande étroite masqué par une tonale

Le masquage fréquentiel peut être expliqué par les mécanismes de l'audition, décrits dans [Zwicker, 1981].

La Figure C.3 schématise l'anatomie de l'oreille. L'onde sonore captée par le pavillon est transmise jusqu'au tympan, qu'elle fait vibrer. Cette vibration aérienne est transmise par la chaîne des osselets (oreille moyenne) jusqu'à la fenêtre ovale, où elle est transformée en vibration en milieu aqueux, dans la cochlée (ou limaçon), qui est remplie de liquide lymphatique. Dans ce liquide baigne la membrane basilaire : le déplacement du liquide cochléaire fait vibrer celle-ci, ce qui excite le nerf auditif via les cellules ciliées reliées à la membrane basilaire. La cochlée a la forme d'un limaçon, qui est représenté déroulé sur la Figure C.3.

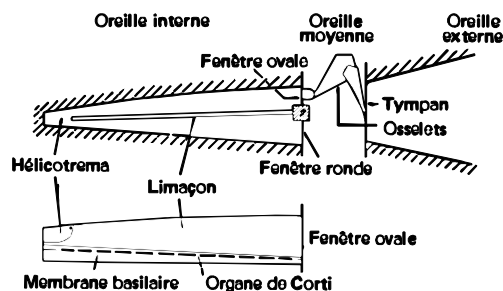


Figure C.3 : Schéma anatomique de l'oreille humaine

Pour un son à une fréquence donnée, la membrane basilaire vibre avec une amplitude maximale en un point dépendant de la fréquence, selon le schéma de la Figure C.4 (l'extrémité extérieure de la spirale correspondant au point de jonction avec la fenêtre ovale). L'oreille réalise donc une transformation fréquence-espace, selon une relation non linéaire. La relation est linéaire entre la tonie (sensation de hauteur) et le lieu de la vibration d'amplitude maximale : une différence de tonie de 1 Bark correspond à un segment de longueur 1,3 mm sur la membrane basilaire.

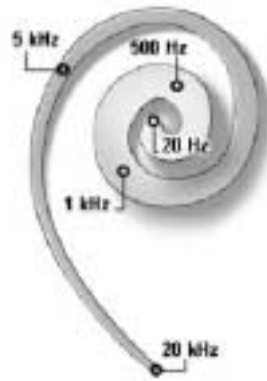


Figure C.4 : Tonotopie de la cochlée [DESS Sc. Co., 2001]

La perception du niveau d'un son est expliquée par la notion de bande critique : on peut considérer que l'oreille divise l'espace des fréquences audibles en 25 bandes adjacentes, réparties selon le tableau C.1, et détermine le niveau subjectif par intégration de l'intensité du son sur chacune de ces bandes, appelées bandes critiques. Ainsi, un bruit est audible si, dans au moins une bande critique, son énergie est supérieure au seuil d'audition. 1 Bark correspond à l'espace entre les centres de deux bandes critiques consécutives.

N° de bande critique	Bornes des bandes (Hz)
1	20 – 100
2	100 – 200
3	200 – 300
4	300 – 400
5	400 – 510
6	510 – 630
7	630 – 770
8	770 – 920
9	920 – 1080
10	1080 – 1270
11	1270 – 1480
12	1480 – 1720
13	1720 – 2000
14	2000 – 2320
15	2320 – 2700
16	2700 – 3150
17	3150 – 3700
18	3700 – 4400
19	4400 – 5300
20	5300 – 6400
21	6400 – 7700
22	7700 – 9500
23	9500 – 12000
24	12000 – 15500
25	15500 – 19500

Tableau C.1 : Liste des bandes critiques [Johnston, 1988]

L'excitation basilaire résultant d'un son au spectre borné par une bande critique s'étend au-delà de cette bande critique. La Figure C.5 représente l'amplitude des oscillations de la membrane basilaire provoquées par des sons purs de fréquences 100, 200 et 300 Hz, en fonction de la distance à la fenêtre ovale. La forme et l'étendue de l'excitation permettent d'expliquer celles des courbes de masquage de la Figure C.2 : le masquage résulte de la comparaison, dans chaque bande critique, de l'excitation du son masquant avec celle du son masqué.

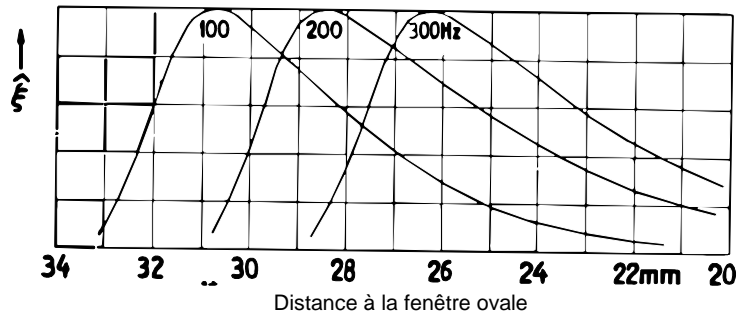


Figure C.5 : Amplitude des oscillations transversales de la membrane basilaire pour 3 tonales

Annexe D :

Consignes du test de comparaison par paires

TEST DE PREFERENCE

Vous allez entendre sur casque des paires d'échantillons de parole.

Chaque paire est constituée de deux échantillons séparés par un silence de 500 ms.

Chaque échantillon est constitué de deux phrases.

Dans la paire d'échantillons, vous devez choisir celui que vous préférez : le premier ou le deuxième.

Pendant l'écoute, le bouton rouge qui est devant vous sera allumé rouge.

Vous voudrez bien écouter chaque paire d'échantillons complètement .
Puis quand le bouton vert qui est devant vous s'allumera vert, indiquez **votre choix en appuyant sur le bouton approprié (chiffre 1 ou 2) selon l'échelle suivante :**

1 : PREFERE LE PREMIER ECHANTILLON .

2 : PREFERE LE DEUXIEME ECHANTILLON .

Vous disposez de 5 secondes pour enregistrer votre réponse (temps pendant lequel le bouton vert reste allumé).

Lorsque vous aurez donné votre opinion , se produira une courte pause avant la paire suivante.

Dans cette expérience nous commencerons par une séance d'apprentissage formée de huit paires d'échantillons. Viendront ensuite les deux mini-séances, comprenant chacune trente six paires d'échantillons.

Annexe E :

Consignes du test de comparaison de dégradations

TEST DE COMPARAISON

Vous allez entendre sur le casque qui est devant vous des séries de 3 paires d'échantillons de parole. Un échantillon est constitué d'une double phrase.

Chaque paire est constituée de l'échantillon de référence suivi d'un échantillon traité, que vous devez comparer à la référence. Les deux échantillons sont séparés par un silence de 600 ms.

Chaque série se présente sous la forme :

paire 1 - paire 2 - paire 1

Les paires sont séparées par des silences de 1,2 s.

Vous devrez choisir la paire dans laquelle la modification de l'échantillon traité par rapport à la référence est la moins gênante : paire 1 ou paire 2. La répétition de la paire 1 doit vous permettre d'être plus sûr de votre jugement.

Pendant l'écoute, le bouton rouge qui est devant vous sera allumé.

Vous voudrez bien écouter chaque série de 3 paires complètement. Puis, quand le bouton vert s'allumera, **indiquez votre choix en appuyant sur le bouton approprié (chiffres 1 ou 2) selon l'échelle suivante :**

1 : MODIFICATION MOINS GÊNANTE DANS LA PAIRE 1.

2 : MODIFICATION MOINS GÊNANTE DANS LA PAIRE 2.

Vous disposez de 5 secondes pour enregistrer votre réponse (temps pendant lequel le bouton vert reste allumé).

Lorsque vous aurez donné votre opinion se produira une courte pause avant la série suivante.

Dans cette expérience nous commencerons par un apprentissage formé de quelques séries de 3 paires. Viendront ensuite les deux mini-séances, comprenant chacune 12 séries.

Annexe F : Evaluation du bruit de quantification

XY	Loc.	Phr.	Amb.	% préférence X	XY + YX	s1 + s2	s1 + s2, XY + YX	% sur tous les locuteurs	% sur tous loc, XY + YX		
A B	M1	s1	cl	20.83	20.83	22.92	20.83	35.94	32.81		
	M1	s2	cl	25.00	20.83						
	M2	s1	cl	29.17	22.92	31.25	25.00				
	M2	s2	cl	33.33	27.08						
	F1	s1	cl	45.83	56.25	54.17	57.29				
	F1	s2	cl	62.50	58.33						
	F2	s1	cl	25.00	18.75	35.42	28.13				
	F2	s2	cl	45.83	37.50						
	M1	s1	br	20.83	18.75	20.83	23.96			29.17	38.54
	M1	s2	br	20.83	29.17						
	F1	s1	br	41.67	54.17	37.50	53.13				
	F1	s2	br	33.33	52.08						
A C	M1	s1	cl	4.17	14.58	18.75	19.79	43.75	42.71		
	M1	s2	cl	33.33	25.00						
	M2	s1	cl	16.67	16.67	18.75	18.75				
	M2	s2	cl	20.83	20.83						
	F1	s1	cl	79.17	77.08	83.33	83.33				
	F1	s2	cl	87.50	89.58						
	F2	s1	cl	45.83	45.83	54.17	48.96				
	F2	s2	cl	62.50	52.08						
	M1	s1	br	25.00	16.67	27.08	21.88			44.79	41.67
	M1	s2	br	29.17	27.08						
	F1	s1	br	70.83	66.67	62.50	61.46				
	F1	s2	br	54.17	56.25						
B C	M1	s1	cl	29.17	43.75	35.42	48.96	56.25	64.32		
	M1	s2	cl	41.67	54.17						
	M2	s1	cl	20.83	39.58	47.92	58.33				
	M2	s2	cl	75.00	77.08						
	F1	s1	cl	62.50	72.92	72.92	78.13				
	F1	s2	cl	83.33	83.33						
	F2	s1	cl	75.00	70.83	68.75	71.88				
	F2	s2	cl	62.50	72.92						
	M1	s1	br	37.50	47.92	50.00	56.25			48.96	59.38
	M1	s2	br	62.50	64.58						
	F1	s1	br	41.67	56.25	47.92	62.50				
	F1	s2	br	54.17	68.75						
Y X	Loc.	Phr.	Amb.	% préférence Y		s1 + s2		% sur tous les locuteurs			
	B A	M1	s1	cl	79.17		81.25		70.31		
		M1	s2	cl	83.33						
		M2	s1	cl	83.33		81.25				
		M2	s2	cl	79.17						
		F1	s1	cl	33.33		39.58				
		F1	s2	cl	45.83						
		F2	s1	cl	87.50		79.17				
		F2	s2	cl	70.83						
		M1	s1	br	83.33		72.92			52.08	
		M1	s2	br	62.50						
		F1	s1	br	33.33		31.25				
F1		s2	br	29.17							
C A	M1	s1	cl	75.00		79.17		58.33			
	M1	s2	cl	83.33							
	M2	s1	cl	83.33		81.25					
	M2	s2	cl	79.17							
	F1	s1	cl	25.00		16.67					
	F1	s2	cl	8.33							
	F2	s1	cl	54.17		56.25					
	F2	s2	cl	58.33							
	M1	s1	br	91.67		83.33			61.46		
	M1	s2	br	75.00							
	F1	s1	br	37.50		39.58					
	F1	s2	br	41.67							
C B	M1	s1	cl	41.67		37.50		27.60			
	M1	s2	cl	33.33							
	M2	s1	cl	41.67		31.25					
	M2	s2	cl	20.83							
	F1	s1	cl	16.67		16.67					
	F1	s2	cl	16.67							
	F2	s1	cl	33.33		25.00					
	F2	s2	cl	16.67							
	M1	s1	br	41.67		37.50			30.21		
	M1	s2	br	33.33							
	F1	s1	br	29.17		22.92					
	F1	s2	br	16.67							

Tableau F.1 : Résultats du test de comparaison par paires de A, B et C

Les tableaux F.1 et F.2 sont à lire de la manière suivante. La moitié supérieure correspond à un ordre de présentation des paires (F.1) ou des séries (F.2), la moitié inférieure à l'ordre inverse. La première colonne indique les paires ou séries testées, les colonnes "Loc", "Phr" et "Amb." indiquent respectivement les locuteurs, phrases et ambiances sonores ("cl" pour silence et "br" pour brouhaha). La cinquième colonne indique les pourcentages de préférence pour chaque comparaison, les colonnes suivantes représentent les résultats cumulés de différentes manières.

	Loc.	Phr.	Amb.	% préférence paire 1	ADA + DAD	s1 + s2	s1 + s2, ADA + DAD	% tous les locuteurs	% tous loc, ADA+DAD
ADA	M1	s1	cl	4.17	14.58	14.58	18.75	27.08	23.96
	M1	s2	cl	25.00	22.92				
	M2	s1	cl	20.83	20.83	27.08	25.00		
	M2	s2	cl	33.33	29.17				
	F1	s1	cl	41.67	29.17	37.50	31.25		
	F1	s2	cl	33.33	33.33				
	F2	s1	cl	25.00	18.75	29.17	20.83		
	F2	s2	cl	33.33	22.92				
	M1	s1	br	16.67	22.92	22.92	23.96	29.17	26.56
	M1	s2	br	29.17	25.00				
	F1	s1	br	33.33	29.17	35.42	29.17		
	F1	s2	br	37.50	29.17				
DAD	M1	s1	cl	75.00		77.08		79.17	
	M1	s2	cl	79.17					
	M2	s1	cl	79.17		77.08			
	M2	s2	cl	75.00					
	F1	s1	cl	83.33		75.00			
	F1	s2	cl	66.67					
	F2	s1	cl	87.50		87.50			
	F2	s2	cl	87.50					
	M1	s1	br	70.83		75.00		76.04	
	M1	s2	br	79.17					
	F1	s1	br	75.00		77.08			
	F1	s2	br	79.17					

Tableau F.2 : Résultats du test de comparaison de dégradations

Annexe G

Significativité de l'écart entre deux pourcentages

La *significativité* de l'écart entre deux proportions observées est déterminée par la méthode suivante [Guéguen, 1998].

On calcule d'abord l'*écart-type commun* aux deux pourcentages p_1 et p_2 , noté σ :

$$\sigma^2 = p(100 - p) \left(\frac{1}{N_1} + \frac{1}{N_2} \right), \quad (5.1)$$

avec N_1 et N_2 les effectifs respectifs du premier et du deuxième échantillons de test et p le *pourcentage commun*, défini par :

$$p = \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2}. \quad (5.2)$$

Cet *écart-type commun* permet de calculer l'*écart-réduit* :

$$z = \frac{p_1 - p_2}{\sigma}. \quad (5.3)$$

Les deux pourcentages sont considérés comme significativement différents au risque de x % si z est supérieur au quantile de la loi normale correspondant à x %, soit 1,96 pour un risque de 5 %.

Annexe H

Construction d'une échelle de Thurstone [Bonnet, 86]

Nous considérons un ensemble de stimuli ayant fait l'objet d'un test de préférences par paires. Disposant pour chaque paire de stimuli XY de la fréquence de préférence de X à Y , nous souhaitons traduire ces résultats par un positionnement des différents stimuli sur une échelle de préférence.

Thurstone postule qu'à chaque grandeur du stimulus S_i correspond un jugement, appelé *processus discriminatif* et noté s_i , considéré comme une variable aléatoire. Un sujet préférera S_i à S_j si $s_i > s_j$. En termes de probabilités,

$$P(S_i > S_j) = P(s_i - s_j > 0). \quad (6.1)$$

On suppose une distribution gaussienne des processus discriminatifs. Chaque processus s_i est alors représenté par sa moyenne $\mu(s_i)$ et sa variance $\sigma^2(s_i)$. Pour un couple de stimuli S_i et S_j , la différence $s_i - s_j$ a aussi une distribution gaussienne, de moyenne $\mu(s_i) - \mu(s_j)$ et de variance $\sigma^2(s_i - s_j)$. Connaissant la fréquence de préférence de S_i à S_j , c'est-à-dire $P(S_i > S_j)$, on peut estimer cette moyenne de $s_i - s_j$ à partir de la valeur de la variable normale réduite $z(S_i > S_j)$ correspondant à $P(S_i > S_j)$, comme l'illustre la Figure H.1.

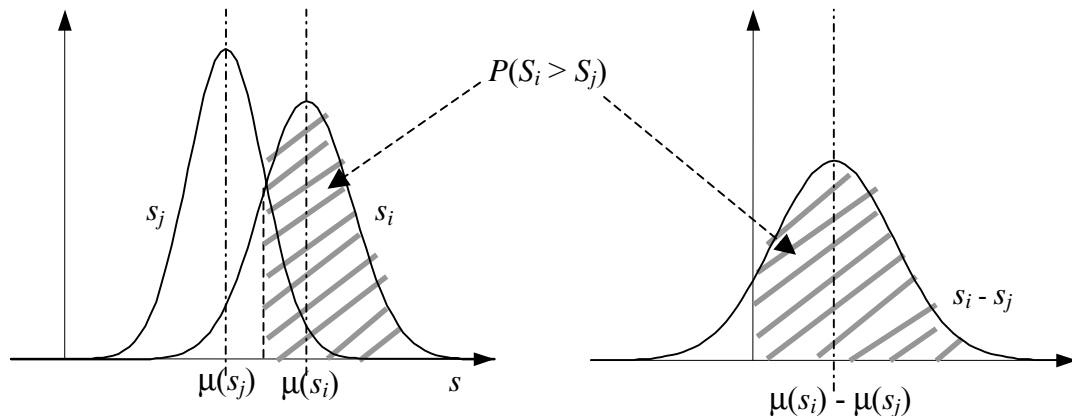


Figure H.1 : Distributions de s_i , s_j et $s_i - s_j$; Probabilité de préférence de S_i à S_j

En effet,

$$z(S_i > S_j) = \frac{\mu(s_i) - \mu(s_j)}{\sigma(s_i - s_j)}, \quad (6.2)$$

avec

$$\sigma^2(s_i - s_j) = \sigma^2(s_i) + \sigma^2(s_j) - \Gamma(s_i, s_j)\sigma(s_i)\sigma(s_j), \quad (6.3)$$

où Γ désigne la corrélation. Sous l'hypothèse que les variances des processus discriminatifs sont toutes égales à l'unité et que les corrélations sont nulles, la variance de $s_i - s_j$ vaut 2. Dans ce cas, la moyenne de $s_i - s_j$ vaut :

$$\mu(s_i) - \mu(s_j) = z(S_i > S_j)\sqrt{2}. \quad (6.4)$$

Cet écart est proportionnel à l'écart entre les points représentatifs de S_i et S_j sur l'échelle de préférences. Chaque stimulus est alors représenté sur l'échelle de préférence par la moyenne de ses écarts avec les autres stimuli.

Annexe J

Calcul des fonctions linéaires discriminantes

[Lebart, 2000b]

Disposant dans un espace à p dimensions de N individus répartis en K classes, nous recherchons les combinaisons linéaires de leurs composantes qui minimisent la variance intra-classes et maximisent la variance inter-classes.

Soit $a(i)$ une combinaison linéaire des composantes x_j^i préalablement centrées d'un individu i représenté par le vecteur x^i :

$$a(i) = \sum_{j=1}^p a_j \cdot (x_j^i - \bar{x}_j), \quad (7.1)$$

où $(\bar{x}_j)_{1 \leq j \leq p} = \bar{x}$ est le centre de la population. La variable a étant centrée, sa variance vaut :

$$\text{var}(a) = \sum_{i=1}^N a(i)^2 = a' T a, \quad (7.2)$$

où T est la matrice $[(t_{jk})_{1 \leq j, k \leq p}]$ de covariance des p variables, de composantes :

$$t_{j,k} = \frac{1}{N} \sum_{i=1}^N (x_j^i - \bar{x}_j)(x_k^i - \bar{x}_k). \quad (7.3)$$

La matrice T se décompose en $T = D + E$, avec D la matrice d'inertie intra-classes et E la matrice d'inertie inter-classes, d'éléments génériques respectifs d_{jk} et e_{jk} tels que :

$$d_{jk} = \frac{1}{N} \sum_{q=1}^K \sum_{i \in q} (x_j^i - \bar{x}_j^q)(x_k^i - \bar{x}_k^q) \quad (7.4)$$

$$e_{jk} = \sum_{q=1}^K \frac{N_q}{N} (\bar{x}_j^q - \bar{x}_j)(\bar{x}_k^q - \bar{x}_k), \quad (7.5)$$

où \bar{x}^q désigne le centre de la $q^{\text{ème}}$ classe et N_q le cardinal de la $q^{\text{ème}}$ classe. Ainsi, la variance de a se décompose en variance interne et variance externe :

$$\text{var}(a) = a' T a = a' D a + a' E a \quad (7.6)$$

Minimiser la variance intra-classes et maximiser la variance inter-classes de a revient à maximiser $f(a)$ tel que :

$$f(a) = \frac{a' Ea}{a' Ta} \quad (7.7)$$

Comme la fonction f est invariante si a est multiplié par un scalaire quelconque, cela revient à maximiser $a' Ea$ sous la contrainte $a' Ta = 1$. La résolution de ce problème conduit à la relation :

$$Ea = \lambda Ta. \quad (7.8)$$

Si T est inversible, on obtient :

$$\begin{cases} T^{-1}Ea = \lambda a \\ a' Ea = \lambda a' Ta = \lambda \end{cases} \quad (7.9)$$

c'est-à-dire que a est le vecteur propre de $T^{-1}E$ relatif à la plus grande valeur propre. Cette valeur propre est appelée le *pouvoir discriminant* de la fonction linéaire a . Pour une partition en K classes, les $K-1$ fonctions linéaires discriminantes correspondront aux vecteurs propres associés aux $K-1$ plus grandes valeurs propres de $T^{-1}E$.

Dans ce cas particulier du classement en deux classes, le terme générique de la matrice E d'inertie inter-classes donné par l'équation (7.5) peut s'exprimer :

$$e_{jk} = \frac{N_1 N_2}{N} (\bar{x}_j^1 - \bar{x}_j^2) (\bar{x}_k^1 - \bar{x}_k^2), \quad (7.10)$$

où N_1 et N_2 sont les cardinaux respectifs des classes 1 et 2, \bar{x}^1 et \bar{x}^2 sont les moyennes de x dans les classes 1 et 2 respectivement. La matrice symétrique E peut être considérée comme le produit d'une matrice colonne c par sa transposée :

$$E = cc', \quad (7.11)$$

avec :

$$c_j = \frac{\sqrt{N_1 N_2}}{N} (x_j^1 - x_j^2). \quad (7.12)$$

La relation (7.8) s'exprime alors :

$$T^{-1}cc'a = \lambda a. \quad (7.13)$$

Pré-multiplions les deux membres par c' :

$$[c'T^{-1}c]c'a = \lambda c'a. \quad (7.14)$$

Comme E est de rang 1 (matrice symétrique), λ , valeur propre de $T^{-1}E$, est unique. Par conséquent, la quantité entre crochets étant un scalaire,

$$\lambda = c'T^{-1}c. \quad (7.15)$$

La fonction discriminante correspond au vecteur propre a , tel que :

$$a = T^{-1}c . \tag{7.16}$$

Références bibliographiques

- [Boite, 1987] René Boite et Murat Kunt, "Traitement de la parole", Presses polytechniques romandes, 1981.
- [Bonnet, 1986] C. Bonnet, "Manuel pratique de psychophysique", Armand Colin, 1986, pp. 136-142.
- [Bowker, 1993] Duane O. Bowker, John T. Ganley, J.H. James, "Telephone network speech signal enhancement", AT&T Bell Laboratories, 1993-1994, brevet US 5333195.
- [Cadoret, 1983] R. Cadoret, Note technique CNET NT/LAA/ELR/289, "Le réseau de lignes d'abonnés", p 23, 1983.
- [Cappé, 1994] O. Cappé, "Elimination of the Musical Noise Phenomenon with Ephraim and Malah Noise Suppressor", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No 2, pp. 345-349, 1994.
- [Com. int., 2002] Communication interne France Télécom R&D, juin 2002.
- [Combesure, 1981] P. Combesure, "Vingt listes de dix phrases françaises phonétiquement équilibrées", *Revue d'Acoustique*, n° 56, pp. 34-38, 1981.
- [Davis, 2002] Gillian M. Davis, "Noise Reduction in Speech Applications", CRC Press, 2002.
- [De Jaco, 1997] Andrew P. De Jaco, John A. Miller, "Adaptative equalizer preprocessor for mobile telephone speech coder to modify non ideal frequency response of acoustic transducer", 1997-1999, brevet US 5915235.
- [DESS Sc. Co., 2001] Cours de perception auditive du DESS de Sciences Cognitives Appliquées de l'Université de Bordeaux 1, décembre 2001
(<http://www.sm.u-bordeaux2.fr/~asco/DESS/cours/polyAudition.pdf>)
- [Ezzaidi, 2001] H. Ezzaidi, J.Rouat et D. O'Shaughnessy, "Towards combining pitch and MFCC for speaker identification systems", *Proc. Eurospeech*, pp. 2825-2828, Aalborg, septembre 1993.
- [Faucon, 1993] G. Faucon, R. Le Bouquin et A. Abkari Azirani, "Mesures objectives de la réduction de bruit", *XIV^{ème} Colloque GRETSI*, Juan-les-Pins, septembre 1993.
- [Gilloire, 1994] A. Gilloire, "Performance evaluation of acoustic echo control : required values and measurement procedures", *Annales des Télécommunications*, T. 49, n° 7-8, pp. 368-372, juillet-août 1994.
- [Glavieux, 1996] Alain Glavieux et Michel Joindot, "Communications numériques – Introduction", Masson, 1996.
- [GRECO-PRC, 1990] GRECO-PRC Communication Homme-Machine, BDSOONS, Base de données des sons du français, avril 1990.

- [Gritton, 1984] C.W.K. Gritton et D.W. Lin, "Echo Cancellation Algorithms", *IEEE ASSP Magazine*, vol. 1, n° 2, pp. 30-38, avril 1984.
- [Guéguen, 1998] N. Guéguen, "Manuel de statistiques pour psychologues", Dunod, 1998.
- [Ho, 1993] Helena S. Ho, Mickael K. Pratt, Pong C. Lim, Thomas T. Oshidari, « *Voice enhancement system and method* », DSC Communications Corporation, 1993-1995, brevet US 5471527.
- [Jambu, 1999] M. Jambu, "Méthodes de base de l'analyse des données", Eyrolles, 1999.
- [Johnston, 1988] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on selected areas in communications*, vol. 6, n° 2, pp. 314-323, février 1988.
- [Lanoë, 1999] R. Lanoë, "Traitements de correction du niveau et du timbre des signaux téléphoniques prenant en compte le bruit", rapport de stage IFSIC / CNET, 1999.
- [Lebart, 2000a] L. Lebart, A. Morineau, M. Piron, "Statistique exploratoire multi-dimensionnelle", Dunod, 2000, pp 145-185
- [Lebart, 2000b] L. Lebart, A. Morineau, M. Piron, "Statistique exploratoire multi-dimensionnelle", Dunod, 2000, pp 251-268.
- [Mahé, 1998] Gaël Mahé, "Etude de traitements centralisés pour la correction du niveau et du timbre de la parole téléphonique", mémoire de DEA STIR, ENST Bretagne, 1998.
- [Mahé, 2002] G. Mahé, A. Gilloire, "Quantization noise spectral shaping in instantaneous coding of spectrally unbalanced speech signals", *Proc. IEEE Workshop on Speech Coding*, Tsukuba, octobre 2002.
- [Makhoul, 1979] J. Makhoul et M. Berouti, "Adaptative noise spectral shaping and entropy coding in predictive coding of speech," *IEEE Transactions on acoustics, speech, and signal processing*, vol. ASSP-27, n° 1, pp. 63-73, février 1979.
- [Mauuary, 1996] L. Mauuary, "Blind equalization for robust telephone based speech recognition", *Proc. Eusipco*, pp 125-128, Trieste, 1996.
- [Mokbel, 1993] C. Mokbel, J. Monné et D. Juvet, "On-line adaptation of a speech recognizer to variations in telephone line conditions", *Proc. Eurospeech*, pp. 1247-1250, Berlin, septembre 1993.
- [Mokbel, 1996] C. Mokbel, D. Juvet et J. Monné, "Deconvolution of telephone line effects for speech recognition", *Speech Communication*, Vol. 19, No. 3, pp. 185-196, septembre 1996.
- [National Semiconductor, 1994] National Semiconductor, documentation technique "TP3054, TP3057 - Enhanced Serial Interface - CODEC/Filter COMBO Family", août 1994.
- [Naylor, 1994] P. Naylor, J. Alcazar, J. Boudy, Y. Grenier, "Enhancement of hands-free telecommunications", *Annales des Télécommunications*, T. 49, n° 7-8, pp. 373-379, juillet-août 1994.
- [Paillard, 1992] B. Paillard, P. Mabilieu, S. Morissette, "PERCEVAL: perceptual evaluation of the quality audio signals", *Journal of the Acoustical Society of America*, Vol. 40, n° 12, pp. 21-30, 1992.

- [Proakis, 1996] John G. Proakis, "Digital signal processing: principles, algorithms, and applications", Prentice Hall PTR, 1996, pp. 838-841.
- [Reynolds, 1995] Douglas A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, vol.17, pp.91-108, 1995.
- [Scalart, 2001] P. Scalart, C. Marro, L. Mauuary, "Procédé et dispositif de réduction de bruit", janvier 2001, brevet FR0101220.
- [Schroeder, 1979] M.R. Schroeder, B.S. Atal, J.L. Hall, "Optimizing digital speech coders by exploring masking properties of the human ear", *J. Acoust. Soc. Am.* 66(6), pp. 1647-1652, décembre 1979.
- [Tukey, 1953] J.W. Tukey, "The problem of multiple comparisons", Ditto, Princeton University, 1953.
- [UIT-R/BS.1534, 1996] UIT-R, Recommandation BS.1534, "Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes de codage", juin 2001.
- [UIT-T/G.121, 1993] UIT-T, Recommandation G.121, "Équivalents pour la sonie des systèmes nationaux", mars 1993.
- [UIT-T/G.168, 1999] UIT-T, Recommandation G.168, "Digital network echo cancellers", juin 2002.
- [UIT-T/G.711, 1988] UIT-T, Recommandation G.711, "Modulation par impulsions et codage (MIC) des fréquences vocales", novembre 1988.
- [UIT-T/G.729, 1996] UIT-T, Recommandation G.729, "Codage de la parole à 8 kbit/s par prédiction linéaire avec excitation par séquences codées à structure algébrique conjuguée", mars 1996.
- [UIT-T/G.VED, 2002] UIT-T, Draft Recommendation G.VED (Voice Enhancement Devices), 2002.
- [UIT-T/P.310, 2000] UIT-T, Recommandation P.310, "Caractéristiques de transmission pour téléphones numériques à bande téléphonique (300-3400 Hz)", mai 2000.
- [UIT-T/P.313, 2000] UIT-T, Recommandation P.313, "Caractéristiques de transmission des terminaux numériques mobiles ou sans cordon", septembre 1999.
- [UIT-T/P.48, 1988] UIT-T, Recommandation P.48, « Spécification d'un système de référence intermédiaire », 1988.
- [UIT-T/P.50/App. I, 1998] UIT-T, Recommandation P.50, "Voix artificielle" – Appendice I: "Signaux d'essai", février 1998.
- [UIT-T/P.800, 1996] UIT-T, Recommandation P.800, "Méthodes d'évaluation subjective de la qualité de transmission", août 1996.
- [UIT-T/P.830, 1996] UIT-T, Recommandation P.830, "Évaluation subjective de la qualité des codecs numériques à bande téléphonique et à bande large", annexe D, février 1996.
- [Vecsys, 1994] Société Vecsys, "Authentification vocale du locuteur à travers le réseau téléphonique", Premier rapport d'avancement pour France Télécom CNET, juillet 1994.

[Zwicker, 1981] E. Zwicker, R. Feldtkeller, "Psychoacoustique – L'oreille récepteur de l'information", Masson, 1981, traduit de l'allemand par C. Sorin.