Elaboration de nouveaux signaux d'ancrage pour l'évaluation des codeurs

THIERRY ETAME ETAME[1], REGINE LE BOUQUIN JEANNES[2,3], CATHERINE QUINQUIS[1],

LAETITIA GROS[1], GERARD FAUCON[2,3]

[1]France Telecom R&D TECH/SSTP - 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2]INSERM, U 642, Rennes, F-35000, France
[3] Université de Rennes 1, LTSI, F-35000, France
LTSI, Campus de Beaulieu, Université de Rennes 1, 35042 Rennes Cedex, France

[1]th_etame@yahoo.fr, Catherine.Quinquis@orange-ftgroup.com, Laetitia.Gros@orange-ftgroup.com
[2,3]Regine.Le-Bouquin-Jeannes@univ-rennes1.fr, Gerard.Faucon@univ-rennes1.fr

Résumé - La manière la plus fiable d'évaluer la qualité des codecs consiste à réaliser des tests d'écoute. Ceux-ci nécessitent la présence de conditions de référence afin de permettre la comparaison des résultats d'un test à l'autre. De nouvelles techniques de codage apparaissent pour lesquelles le système de référence MNRU (Modulated Noise Reference Unit ou appareil de référence à bruit modulé) n'est plus adapté. L'objectif de ce papier est de proposer un système de référence adapté aux dégradations générées par les nouveaux schémas de compression. Pour 20 codecs de qualité voisine mais présentant des défauts différents et sélectionnés par un test ACR (Absolute Category Rating), nous avons conduit un test de dissimilarité dont les résultats font l'objet d'une analyse multidimensionnelle. A partir des attributs perceptifs caractérisant les quatre dimensions extraites, nous avons généré quatre signaux d'ancrage. Une deuxième analyse sur un ensemble constitué des signaux d'ancrage et de signaux codés a permis de valider trois des attributs perceptifs.

Abstract - Listening tests are the most reliable ways to assess subjective quality of codecs. Reference conditions are needed to allow comparison of results coming from different tests. New coding techniques are developed for which Modulated Noise Reference Unit (MNRU) is no more appropriate. The goal of this paper is to propose a reference system related to the degradations produced by the new coding schemes. An ACR (Absolute Category Rating) test helps in the selection of 20 codecs of similar quality but showing different degradations. The results of dissimilarity tests conducted on signals created using those 20 codecs enter a multidimensional analysis. From the perceptive attributes characterizing the four dimensions of the perceptive space, reference signals have been built. Three of the perceptive attributes have been validated through a second analysis using part of the previous coded signals and the reference signals.

1 Introduction

Si les systèmes de télécommunications modernes de plus en plus complexes permettent un enrichissement du signal transmis, ils ont également un fort impact sur l'évolution des techniques de compression de données, en raison de contrainte de coût et de disponibilité de bande passante. L'environnement fortement concurrentiel dans lequel se développent ces nouveaux services contraint les opérateurs à évaluer et à contrôler la qualité des services qu'ils proposent, notamment des éléments technologiques qui en sont à la base. Nous nous intéressons ici plus spécifiquement à la qualité de codecs de parole et notamment à son évaluation. Actuellement, les tests subjectifs, bien que coûteux et consommateurs de temps, restent la référence pour l'évaluation de la qualité audio perçue par les utilisateurs. Ces tests nécessitent la présence de conditions de référence afin de permettre la comparaison des résultats d'un test à l'autre. Avec le système MNRU (Modulated Noise Reference Unit) [P.810] actuel, seule la dégradation relative au bruit de quantification générée par les codecs de forme d'onde PCM (Pulse Code Modulation) est prise en compte. Il est donc aujourd'hui indispensable d'élaborer un nouveau système de signaux de référence pour prendre en compte les dégradations apportées par les nouveaux codecs. Notre étude consiste à construire un espace d'ancrages en recherchant un ensemble de dégradations perceptibles, indépendantes et les plus représentatives des défauts rencontrés. Les étapes de ce travail sont les suivantes : sélectionner, à partir d'un test ACR (Absolute Category Rating) [P.800], un ensemble de codecs présentant des techniques de codage différentes mais de qualité comparable (afin de juger davantage les défauts que la qualité elle-même du signal), conduire des tests de dissimilarité sur les codecs retenus, analyser les résultats par une analyse multidimensionnelle des proximités (MDS, MultiDimensional Scaling) afin de définir les dimensions perceptives qui sous-tendent la perception des stimuli codés [Mattila 2003, Wältermann 2006], trouver les paramètres physiques ou acoustiques qui expliquent le mieux ces dimensions. Cette caractérisation physique des dimensions nous aidera à élaborer de nouveaux signaux de calibration qui seront testés lors d'une dernière phase de validation. Après un rappel des premières étapes, le présent article traite plus particulièrement des deux derniers points (élaboration des signaux d'ancrage et validation).

2 Sélection des codecs et test de dissimilarité

2.1 Les codecs

Le corpus initial est constitué de dix-neuf codecs WideBand (WB), Super-WideBand (S-WB) ou FullBand (FB) : le codec G.722 aux débits 64, 56 et 48 kbit/s, qui utilise un codec ADPCM (Adaptive Differential PCM), le codec G.722.1 aux débits 24 et 32 kbit/s, qui utilise un algorithme basé sur un codage par transformée MLT (Modulated Lapped Transform), le codec AMR-WB (Adaptive Multi-Rate WB) G.722.2 aux débits 8.85, 12.65, 15.85 et 23.85 kbit/s, dont la bande basse est traitée par un algorithme CELP, la bande haute étant reconstruite simplement par filtrage d'un bruit blanc, le codec G.729.1 aux débits 14, 20, 24 et 32 kbit/s, caractérisé par un codage CELP dans la sous-bande (50 - 4000Hz), une extension de bande par codage paramétrique TDBWE (Time Domain BandWidth Extension) pour la sous-bande (4000 - 7000 Hz) et un codage par transformée TDAC (Time-Domain Aliasing Cancellation) de 50 à 7000 Hz, le codec G.722.1.C à 24 kbit/s qui est une simple extension du G.722.1 pour des signaux en bande Hi-Fi (50 - 14000 Hz), le codec standard MPEG-4 ou HE-AAC (High Efficiency Advanced Audio Coding) aux débits 16, 24 et 32 kbit/s, pour lequel la transformation du signal d'entrée est obtenue par MDCT (Modified Discrete Cosine Transform), et le codec MP3 à 32 et 64 kbit/s, basé sur la technique de codage perceptif dans le domaine fréquentiel et utilisé pour le codage dans les gammes « haute qualité » et « haute fidélité ». Une opération de tandeming (mise en cascade de codecs) a été appliquée à l'ensemble de ces codecs afin d'élargir notre panel, avoir un certain continuum dans la gamme de qualité et constituer finalement une base d'une vingtaine de tandems/codecs de qualité voisine et moyenne.

2.2 Test préliminaire de sélection des codecs

Les 4 échantillons utilisés pour le test préliminaire de sélection des codecs sont extraits de la base de données de France Telecom et prononcés par quatre locuteurs différents (2 hommes et 2 femmes). D'une durée de 8 secondes, ils sont constitués de deux phrases espacées par un silence. Initialement échantillonnés à 48 kHz, ils sont présentés en entrée des codecs fullband. Ils sont sous-échantillonnés à 32 kHz (resp. 16 kHz) et filtrés pour être présentés en entrée des codecs super-wideband (resp. wideband). Les 58 conditions de test (19 codecs x 3 niveaux de tandem + signal original) sont présentées à 32 auditeurs naïfs (test ACR). Les résultats ont permis d'extraire 20 tandems/codecs de qualité moyenne mais recouvrant les différentes techniques de codage. Le Tableau 1 indique les tandems/codecs retenus.

Tableau 1. tandems/codecs retenus
(l'extension x1, x2, x3 indique le nombre de tandems)

	Description		Description
+ 1	G722.1C 24kbps x2	°11	G722 56kbps x2
+ 2	G722.1C 24kbps x3	°12	G722 56kbps x3
+ 3	G722.1 24kbps x2	*13	G729.1 14kbps x3
+ 4	G722.1 24kbps x3	*14	G729.1 20kbps x3
x 5	G722.2 12.65kbps x2	*15	G729.1 24kbps x2
x 6	G722.2 12.65kbps x3	*16	G729.1 32kbps x3
x 7	G722.2 15.85kbps x2	□17	HEAAC 24kbps x2
x 8	G722.2 8.85kbps x2	□18	HEAAC 32kbps x2
° 9	G722 48kbps x2	□19	MP3 32kbps x1
°10	G722 48kbps x3	□20	MP3 32kbps x2

3 Caractérisation perceptive des codecs

Cette étape consiste à établir l'espace perceptif multidimensionnel qui sous-tend la perception de la parole codée.

3.1 Rappels sur la MDS

Une méthode couramment utilisée permettant de prendre en compte la multidimensionnalité de la qualité vocale est l'analyse multidimensionnelle des proximités (MDS, MultiDimensional Scaling). On peut définir la MDS comme une classe de techniques utilisées pour développer une représentation multidimensionnelle des proximités entre stimuli, l'objectif étant de déterminer le nombre de dimensions et la configuration des stimuli dans cet espace multidimensionnel. Chaque dimension de l'espace traduit une caractéristique perceptive commune aux sons étudiés. En d'autres termes, dans cet espace, deux sons proches sont jugés similaires et deux sons éloignés sont jugés dissemblables. L'avantage de cette méthode est qu'elle n'impose aucune présomption sur les dimensions contrairement aux méthodes utilisant des échelles par descripteurs sémantiques [Kruskal 1964, Etame Etame 2008a].

3.2 Test de dissimilarité

Deux échantillons sont extraits de la base de données de France Telecom et durent 6 secondes. Ils sont constitués de deux phrases, espacées par un silence et prononcées par un homme et une femme. Ils sont traités par les 20 tandems/codecs retenus lors du test préliminaire. Deux tests de dissimilarité sont alors réalisés, un avec les échantillons homme et l'autre avec les échantillons femme. Le test sur le locuteur homme (resp. femme) est réalisé avec 29 sujets (resp. 28). La tâche des sujets est d'attribuer une note de dissimilarité comprise entre 0 et 100 pour chacune des 210 paires d'échantillon (*i.e.* toutes les paires possibles constituées avec les 20 conditions, y compris les paires nulles constituées de deux échantillons identiques) présentées aléatoirement. A la fin du test de dissimilarité (réalisé individuellement), les testeurs sont invités à commenter librement, avec leurs propres mots, les dégradations perçues en réécoutant successivement les 20 échantillons. Une procédure basée sur le critère de « paires nulles » a conduit au rejet de 4 auditeurs (resp. 3 auditeurs) pour le locuteur homme (resp. femme). Les jugements de dissimilarité de chaque auditeur « fiable » sont traduits en matrices de dissimilarité symétriques. Lors de la phase suivante, les matrices individuelles de dissimilarité serviront à projeter, selon le modèle de l'algorithme de MDS non métrique INDSCAL, l'ensemble des 20 objets sonores dans un espace multidimensionnel.

3.3 Espace perceptif

Les paramètres en sortie de l'analyse sont la valeur du stress, la proportion de la variance expliquée (RSQ : squared correlation) et le poids moyen apporté par les auditeurs sur chaque dimension et ce pour des configurations allant de 2 à 6 dimensions. La nette amélioration du stress pour une dimensionnalité supérieure à 2 suggère un espace de stimuli à au moins 3 dimensions. D'autre part, le passage de 4 à 5 (voire 6) dimensions n'améliore pas significativement les valeurs de stress et de RSQ : un espace à 4 dimensions apparaît ainsi relativement pertinent pour représenter les vingt objets étudiés. Les verbalisations des auditeurs couplées aux projections des objets dans l'espace à 4 dimensions suggèrent de caractériser ces dimensions par les attributs suivants : « clair/sourd », « bruit de fond », « bruit sur la parole »,

« sifflement ». Les Figures 1 à 3 présentent l'espace perceptif obtenu pour un locuteur masculin.

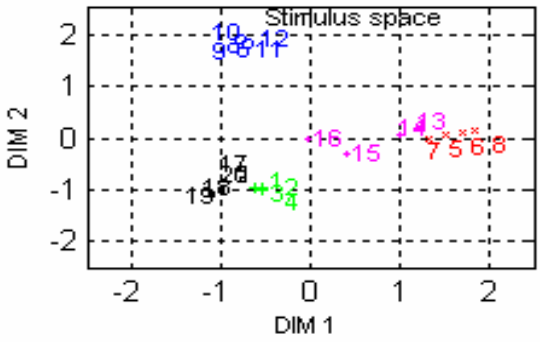

Figure 1 – Espace perceptif (dim. 1 & 2)

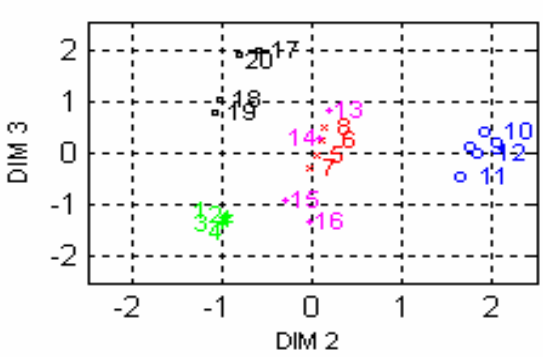

Figure 2 - Espace perceptif (dim. 2 & 3)

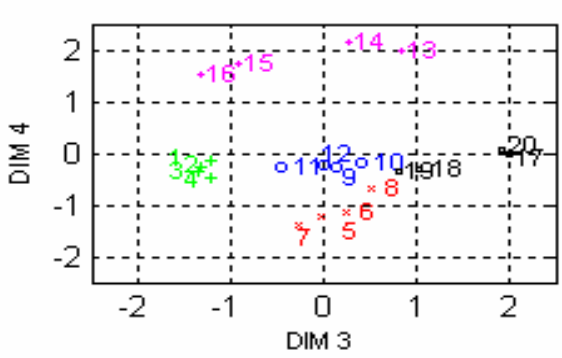

Figure 3 - Espace perceptif (dim. 3 & 4)

4 Corrélats physiques/acoustiques et signaux de référence

L'étape suivante consiste à rechercher des corrélats physiques des dimensions perceptives révélées. Il s'agit de traduire objectivement chacune des dimensions par des mesures physiques pour conduire à l'élaboration de descripteurs du signal. Des coefficients de corrélation élevés entre un descripteur et une dimension suggèrent l'aptitude du descripteur à prédire cette dimension. Cette propriété pourra aider à simuler et calibrer la dégradation perçue et favoriser la construction de signaux de référence.

4.1 Corrélats

4.1.1 Caractérisation physique de l'attribut « clair/sourd »

Les codecs WB CELP se caractérisent principalement par leur défaillance à reproduire les hautes fréquences et génèrent par conséquent une limitation en bande passante en sortie. Pour les codecs G.722.2, la perte d'énergie essentiellement dans les hautes fréquences diminue la sensation de brillance auditive. Or, ceux-ci sont largement projetés sur la première dimension de l'espace perceptif. Le barycentre spectral (SC) étant le paramètre le plus souvent proposé pour décrire cette sensation de brillance [Scholz 2006], il est apparu opportun de le calculer ainsi que son coefficient de corrélation avec les différentes dimensions. Pour cet attribut, nous donnons dans le Tableau 2 les résultats de la corrélation de Pearson entre les projections sur les axes des coordonnées des objets sonores et les descripteurs SC des 20 tandems/codecs (dans tous les tableaux suivants, l'astérisque indique que la mesure est significative).

Tableau 2. Corrélations entre le barycentre et les quatre dimensions

Corrélations	DIM 1	DIM 2	DIM 3	DIM 4
Barycentre (homme)	-0,93*	0,01	-0,07	0,06
Barycentre (femme)	-0,91*	0,17	0,23	-0,35

4.1.2 Caractérisation physique de l'attribut « bruit de fond »

De la même manière que précédemment, à la lumière des verbalisations recueillies auprès des auditeurs, la dimension 2 apparaît caractérisée par l'attribut « bruit de fond », essentiellement perçu dans le silence. Dans un premier temps, nous mesurons la valeur moyenne de l'énergie dans les bandes de fréquences haute (4000 - 7000 Hz) et basse (50 - 4000 Hz) sur les portions de silence contenues dans les signaux. Les corrélations de Pearson entre les énergies moyennes dans la bande haute en l'absence de parole et les quatre dimensions sont données dans le Tableau 3. Elles sont respectivement de 0,85 et 0,84 pour les locuteurs homme et femme pour la dimension 2, ce qui conforte le fait que les codeurs SB-ADPCM G722, qui se détachent des autres suivant cette dimension, présentent clairement un bruit de fond qui se perçoit dans le silence contrairement aux autres codecs.

Tableau 3. Corrélations entre l'énergie en hautes fréquences et les quatre dimensions

Corrélations	DIM 1	DIM 2	DIM 3	DIM 4
Energie moyenne en hautes fréquences (homme)	-0,47	0,85*	-0,05	-0,15
Energie moyenne en hautes fréquences (femme)	-0,50	0,84*	-0,02	-0,14

4.1.3 Caractérisation physique de l'attribut « bruit sur la parole »

Pour caractériser cet attribut et essayer de décrire au mieux une sensation auditive plus ou moins bruitée du signal codé, nous nous sommes intéressés au rapport entre les brillances des composantes déterministes et celles des composantes résiduelles du signal codé. Nous avons ainsi estimé un paramètre (RSC) défini comme le rapport entre le barycentre de la composante déterministe et celui du résidu. Les corrélations de Pearson sont données dans le Tableau 4. Suivant la troisième dimension, elles prennent des valeurs de -0,61 pour les deux locuteurs.

Tableau 4. Corrélations entre les rapports de brillance et les quatre dimensions

Corrélations	DIM 1	DIM 2	DIM 3	DIM 4
RSC (homme)	-0,11	-0,54	-0,61*	-0,27
RSC (femme)	-0,40	-0,54	-0,61*	-0,32

4.1.4 Caractérisation physique de l'attribut « sifflement »

Le coefficient de corrélation entre l'énergie en l'absence de parole dans la bande basse (50 - 4000 Hz) et les coordonnées des objets sonores suivant la dimension 4 s'est avéré important (0,87 pour les 2 locuteurs, homme et femme). Si l'on analyse l'espace perceptif, il s'avère que le codeur G.729.1 se détache des autres codeurs suivant cette dimension, ce qui peut s'expliquer par l'utilisation du « codage descripteur d'insertion de silence » dans la bande basse de ce codeur. Cette interprétation constitue un premier élément de réponse pour expliquer la dégradation « sifflement » associée à cette dimension. Sifflement et réverbération étant liés aux variations dans les spectres des signaux, nous avons calculé les maxima des corrélations (Rmax) entre les spectres de puissance du signal original et de chaque version codée en présence de parole. Les coefficients de corrélation de Pearson entre les projections sur les axes des coordonnées des objets sonores et ces descripteurs sont donnés dans le Tableau 5. Pour la dimension 4, ils sont respectivement de -0,9 pour le locuteur homme et de 0,56 pour le locuteur femme.

Tableau 5. Corrélations entre les valeurs de Rmax et les quatre dimensions

Corrélations	DIM 1	DIM 2	DIM 3	DIM 4
Rmax (homme)	-0,19	0,01	0,08	-0,90*
Rmax (femme)	-0,16	0,11	0,00	0,56*

4.2 Définition des signaux de référence

Même si nous en connaissons les limites, les analyses statistiques précédentes nous ont aidé à générer de nouveaux signaux de référence représentatifs des dégradations perçues. Ainsi, pour simuler le signal d'ancrage représentatif du critère perceptif « clair/sourd » (un codec qui sonne « étouffé » révèle une certaine absence de hautes fréquences ou moins de bande passante en sortie par rapport au signal original d'entrée, et un codec qui sonne « clair » traduit une plus forte présence des hautes fréquences ou autant de bande passante en sortie que pour le signal original), un filtrage passe-bas a été appliqué au signal original wideband d'entrée, le calibrage du signal en sortie du filtre passe-bas se faisant en fonction de la fréquence de coupure de ce filtre, en prenant soin qu'elle soit au moins supérieure à la fréquence de 3400 Hz de la qualité de parole narrowband. Pour l'aspect « bruit de fond », nous avons généré des mélanges de type « signal original wideband » + « bruit blanc gaussien » pour différents rapports signal sur bruit. Les signaux d'ancrage caractéristiques du « bruit sur la parole » ont été modélisés par un bruit modulé par un signal de parole, cette modélisation étant en quelque sorte similaire à celle des signaux produits par l'appareil MNRU. Pour le « sifflement », nous nous sommes fortement inspirés des caractéristiques du codeur G.729.1 qui révélait cet aspect. Dans le cas de ce codeur, le problème de l'inadéquation de phase entre les hautes et basses fréquences survient autour de 4 kHz dans la zone de recouvrement des deux bandes et est dû au filtrage QMF, ce que nous avons reproduit dans le signal de référence correspondant (décomposition en deux sous-bandes, la bande haute étant ensuite filtrée par un filtre de décorrélation avant d'être recombinée avec la bande basse) l'amplification de la dégradation pouvant être obtenue par « tandeming ».

5 Validation et conclusion

La phase de validation utilise les mêmes procédures de tests subjectifs que précédemment. Un test ACR est d'abord réalisé avec 24 auditeurs et conduit à retenir, pour le test de dissimilarité, onze signaux traités par les codecs et neuf signaux de parole dans lesquels les dégradations modélisées précédemment ont été introduites. Au vu des valeurs de stress (0,19) et de RSQ (0,5), une configuration à 5 dimensions apparaît la plus appropriée même si ces grandeurs affichent des valeurs déjà pertinentes pour une dimensionnalité de 4. Toutefois, la projection des objets dans cet espace perceptif à 5 dimensions ne permet pas de les expliquer complètement : si quatre des dimensions peuvent être identifiées avec les attributs définis précédemment, la dimension 5 reste difficilement interprétable. Une meilleure définition des signaux d'ancrage pourrait peut-être améliorer ces résultats [Etame Etame 2008b]. Par exemple, un signal d'ancrage plus représentatif de la dimension « sifflement » pourrait être redéfini à partir d'autres filtres de décorrélation que celui étudié et procurer par là-même une configuration interprétable des objets sonores dans un espace moins perturbé à seulement 4 dimensions.

Références

[Etame Etame 2008a] T. Etame Etame, G. Faucon, R. Le Bouquin Jeannès, L. Gros and C. Quinquis, "Characterization of the multidimensional perceptive space for current speech and sound codecs", AES 124th convention, Amsterdam, The Netherlands, May 17-20, 2008.

[Etame Etame 2008b] T. Etame Etame, "Conception de signaux de référence pour l'évaluation de qualité perçue des codeurs de la parole et du son", Thèse de l'Université de Rennes 1, 2008.

[Kruskal 1964] J.B. Kruskal, "Nonmetric multidimensional scaling: a numerical method", Psychometrika, Vol. 29, pp. 115-129, 1964.

[Mattila 2003] V.V. Mattila, "Ideal point modelling of the quality of noisy speech in mobile communications based on multidimensional scaling", AES 114th convention, Amsterdam, The Netherlands, 2003.

[P.800] ITU-T Recommendation P.810 "Modulated Noise Reference Unit".

[P.810] ITU-T Recommendation P.800 "Methods for subjective determination of transmission quality".

[Scholz 2006] K. Scholz, M. Waltermann, L. Huo, A. Raake, S. Möller, and U. Heute, "Estimation of the quality dimension "directness/frequency content" for the instrumental assessment of speech quality", Interspeech, paper 1219-Wed1A3O.6, 2006.

[Wältermann 2006] M. Wältermann, K. Scholz, A. Raake, U. Heute and S. Möller, "Underlying quality dimensions of modern telephone connections", Interspeech, paper 1089-Wed3FoP.11, 2006.