Conception de signaux de référence pour l'évaluation de la qualité perçue des codecs de la parole et du son

T. ETAME[1], R. LE BOUQUIN JEANNES[2, 3], G. FAUCON[2, 3], L. GROS[1] et C. QUINQUIS[1]

[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

(thierry.etameetame, laetitia.gros, catherine.quinquis)@orange-ftgroup.com
(regine.le-bouquin-jeannes, gerard.faucon)@univ-rennes1.fr

Résumé – Ce travail vise à fournir les signaux de référence représentatifs des nouvelles techniques de codage pour l'évaluation subjective de la qualité de parole codée. La démarche adoptée consiste à caractériser d'un point de vue perceptif les dégradations apportées par les nouveaux codecs, puis à les relier aux techniques de codage afin de pouvoir les recréer artificiellement. La MDS (Multidimensional Scaling ou analyse multidimensionnelle des proximités) non métrique pondérée est utilisée pour générer l'espace perceptif de dégradations de codecs wideband. L'analyse révèle un espace perceptif à quatre dimensions interprétable vis-à-vis des techniques de codage.

Abstract – This work aims to provide reference signals representative of the new techniques of speech coding for subjective assessment tests of speech quality. A non-metric weighted multidimensional scaling (MDS) analysis is used for the perceptual analysis of the defects generated by wideband codecs. The results indicate a four-dimensional space to elucidate the perception of current defects. These dimensions are characterized from a coding point of view.

1. Contexte

La qualité de signaux transmis par les systèmes de télécommunications est évaluée à l'aide de tests d'écoute lors desquels des échantillons de parole, traités par les systèmes à évaluer, sont présentés à des auditeurs. Ceux-ci sont invités à donner leur opinion, après l'écoute de chaque échantillon, sur la qualité vocale de celui-ci sur des échelles de qualité (par exemple : excellente = 5, bonne = 4, moyenne = 3, médiocre = 2, mauvaise = 1) [1]. Afin de faciliter la comparaison des résultats d'un test à l'autre, des points d'ancrage doivent être intégrés dans ces tests. L'appareil MNRU (Modulated Noise Reference Unit ou appareil de référence à bruit modulé) est couramment utilisé comme système de référence lors des tests d'évaluation de la qualité vocale des systèmes de transmission par procédés numériques, fonctionnant tant en bande étroite qu'en bande élargie, pour introduire des dégradations contrôlées dans les signaux [2]. Or, ces dégradations, caractéristiques du bruit de quantification des codecs de type PCM (Pulse Code Modulation ou modulation par impulsions et codage), ne sont plus représentatives, du point de vue perceptif, des défauts apportés par les nouvelles techniques de codage.

L'objectif de notre travail est de développer un système permettant de calibrer les tests subjectifs avec des signaux représentatifs des défauts des codecs actuels. La démarche adoptée consiste à caractériser d'un point de vue perceptif les dégradations apportées par les nouveaux codecs, puis à les relier aux techniques de codage afin de pouvoir les recréer artificiellement.

Partant de l'hypothèse que ces défauts peuvent être décrits sur des continuums perceptifs, nous cherchons à déterminer l'espace perceptif multidimensionnel dans lequel peuvent être représentées les dégradations sonores propres aux codecs actuels.

Nous donnons au paragraphe 2 une brève description de l'analyse multidimensionnelle des proximités (MDS) utilisée dans cette étude pour déterminer l'espace perceptif qui sous-tend la perception des signaux de parole codés. Ensuite, au paragraphe 3, un test de dissimilarité et l'analyse des résultats obtenus par la MDS sont présentés.

2. L'analyse multidimensionnelle des proximités

La qualité vocale est généralement considérée comme un phénomène multidimensionnel [3]. Une des méthodes permettant de prendre en compte la multidimensionnalité de la qualité vocale est la différenciation sémantique utilisant un ensemble d'échelles bipolaires dont les extrémités sont caractérisées par des attributs verbaux opposés (plaisant-déplaisant, sourd-clair, …) [4]. Sur l'ensemble des échelles prédéfinies proposées aux sujets, le résultat obtenu correspond à une représentation multidimensionnelle appelée profil sensoriel. Le désavantage des profils sensoriels est que le vocabulaire est présélectionné par l'expérimentateur.

L'analyse multidimensionnelle des proximités (MDS) est une autre approche de l'étude de la perception multidimensionnelle des sons qui permet d'éviter les biais introduits par les présupposés de l'expérimentateur. II n'y a

aucune présomption sur les dimensions contrairement aux méthodes utilisant des échelles par descripteurs sémantiques ; en revanche, il y a présomption sur la continuité des dimensions.

Dans notre application, cette technique consiste à étudier les structures perceptives qui sous-tendent les jugements de similarité (ou de préférence) entre paires de stimuli en les traduisant en matrice de distance. Celle-ci sert à projeter selon un modèle mathématique l'ensemble des stimuli ou objets sonores dans un espace multidimensionnel. Dans cet espace, des objets similaires se trouvent proches et des objets dissemblables éloignés [5].

L'analyse multidimensionnelle des proximités classique détermine une configuration spatiale des objets dans un espace très souvent euclidien, qui minimise les disparités entre les distances calculées entre les points de cet espace et les dissimilarités contenues dans une matrice de dissimilarités, au sens des moindres carrés. C'est le principe de base de la MDS métrique qui prend en entrée une simple matrice de dissimilarités entre objets étudiés.

Il est difficile pour un sujet de noter uniformément au cours du test d'écoute le degré de différence qu'il perçoit entre les échantillons alors que l'ordre induit par les jugements de dissimilarités semble plus fiable [6]. Ce point de vue est pris en compte dans l'analyse multidimensionnelle non métrique des proximités ([7] [8]) dans laquelle on ne retient que l'ordre des données contenues dans la matrice de dissimilarités.

Par ailleurs, les auditeurs n'attribuent pas les mêmes poids aux dimensions perceptives lors du test d'écoute. Cette variabilité inter-individuelle est prise en compte par l'analyse multidimensionnelle des proximités pondérée INDSCAL (Individual Differences Scaling [9] [10]) qui consiste à appliquer un poids sur chaque dimension en fonction du sujet. La MDS INDSCAL fournit non seulement un espace de stimuli mais également un espace de sujets, espace qui montre le poids donné par chacun des sujets à chacune des dimensions dans l'espace de stimuli.

Dans notre étude, nous avons opté pour une MDS non métrique INDSCAL qui tient compte de la variabilité intra et interindividuelle inhérente à tout test subjectif.

3. Expérimentation

3.1 Sélection des codecs

Pour cette étude, nous avons rassemblé un panel d'une vingtaine de codecs wideband (fréquence d'échantillonnage = 16 kHz), super-wideband (32 kHz) ou fullband (48 kHz) présentant des techniques de codage différentes avec l'objectif de balayer le maximum de défauts possibles.

La technique du "tandeming" ou mise en cascade de codecs est appliquée aux 19 codecs suivants afin d'introduire des degrés de dégradations différentes :

- G722.1 à 24, 32 kbps ; G722.1 C à 24 kbps : codecs par transformée MLT (Modulated Lapped Transform) utilisant une allocation de bits par catégorisation
- HEAAC (High Efficiency Advanced Audio Coding) à 16, 24, 32 kbps ; MP3 à 32, 64 kbps : codecs MDCT (Modified Discrete Cosine Transform) utilisant un modèle psycho-acoustique
- G722.2 à 8,85 ; 12,65 ; 15,85 ; 23,85 kbps : codecs ACELP (Algebraic Code-Excited Linear Predictive)
- G729.1 à 14, 20, 24, 32 kbps : codecs hybrides (CELP – TDBWE, Time-Domain Bandwidth Extension – TDAC, Time-Domain Aliasing Cancellation)
- G722 à 64, 56, 48 kbps : codec SB-ADPCM (Sub-Band Adaptive Differential Pulse Code Modulation).

Dans notre étude, nous considérerons les cas où deux (_x2), ou trois (_x3) mêmes codecs sont mis en cascade, _x1 désigne un codec seul.

Au total 58 tandems/codecs résultants (19 codecs x 3 niveaux de tandem + signal original ou direct) sont considérés. Un test ACR (Absolute Category Rating [1]) a a été conduit afin de sélectionner une vingtaine de tandems/codecs wideband, super-wideband ou fullband de qualité moyenne et voisine afin que les jugements de dissimilarité ne portent pas sur la qualité globale mais bien sur le type de défaut. Des doubles-phrases prononcées par 4 locuteurs (2 hommes et 2 femmes), 2 doubles-phrases par locuteur, sont utilisées pour ce test ACR. Pour chacune des 58 conditions de ce test, le signal de sortie est artificiellement limité en wideband en filtrant la sortie de chaque tandem/codec par un filtre de bande passante 50 Hz – 7 kHz, afin de réduire l'impact du facteur bande. Trente-deux sujets ont participé au test ACR.

Finalement, 18 tandems/codecs ont été retenus pour le test de dissimilarité et sont représentés dans le tableau TAB. 1. Afin de faciliter la perception des dégradations par les sujets, nous avons préféré les signaux résultant de la mise en cascade de 2 ou 3 codecs (_x2, _x3). De même, nous avons considéré des codecs présentant des techniques de codage différentes dans l'objectif d'avoir le maximum de défauts possibles.

3.2 Procédure du test de dissimilarité

Un échantillon de parole de 6 s prononcé par un homme ("La vanille est la reine des arômes. Fragile, il ne résiste pas à l'air glacé."), initialement échantillonné à 48 kHz, est sous-échantillonné et filtré pour être présenté en entrée des codecs super-wideband et wideband. La bande passante du filtre est 50 Hz – 7 kHz pour les codecs wideband et 50 Hz – 14 kHz pour les versions super-wideband. Les stimuli résultants sont ensuite traités par les 18 tandems sélectionnés en limitant artificiellement la sortie de chaque codec en wideband. Les sorties des tandems résultants sont enfin sur-échantillonnées à 48 kHz pour être compatibles avec le matériel de présentation des stimuli du test.

Le test a été réalisé avec un casque audio STAX signature SR-404 (casque ouvert) et son amplificateur SRM-006t. Les stimuli ont été stockés sur un poste de

travail Windows 2000. Le son numérique est envoyé via la carte son PC Digigram VX 222 et converti en 24 bits par le DAC (3Dlab DAC 2000). Les stimuli sont présentés à chaque sujet en écoute diotique à un niveau d'écoute confortable, à l'intérieur d'une cabine insonorisée.

TAB. 1 : les 18 tandems/codecs sélectionnés pour le test de dissimilarité

Codec		Description
+	O1	G722.1C_24kbps_x2
+	O2	G722.1C_24kbps_x3
+	O3	G722.1_24kbps_x2
+	O4	G722.1_24kbps_x3
x	O5	G722.2_12.65kbps_x2
x	O6	G722.2_12.65kbps_x3
x	O7	G722.2_15.85kbps_x2
x	O8	G722.2_8.85kbps_x2
°	O9	G722_48kbps_x2
°	O10	G722_48kbps_x3
°	O11	G722_56kbps_x2
°	O12	G722_56kbps_x3
*	O13	G729.1_14kbps_3
*	O14	G729.1_20kbps_x3
*	O15	G729.1_24kbps_x2
*	O16	G729.1_32kbps_x3
>	O17	MP3_32kbps_x1
>	O18	MP3_32kbps_x2

Le test est individuel. Les stimuli sont présentés à chaque sujet par paires (A-B), dans lesquelles A et B correspondent à deux traitements différents du même échantillon de parole. Au total, 171 (153 + 18 paires nulles) paires de stimuli sont donc considérées. Pour chaque paire, il est demandé au sujet de noter le degré de différence qu'il perçoit entre les stimuli. La dissimilarité entre les deux échantillons de parole codés est notée sur une échelle continue variant de 0 (similaires) à 100 (différents). La consigne de test précise d'écouter au moins une fois intégralement les deux échantillons, de réécouter autant de fois et dans l'ordre qu'on le souhaite les deux échantillons, avant d'exprimer son jugement de dissimilarité. Les jugements subjectifs sont saisis au cours de deux sessions d'une centaine de paires chacune (90 minutes), et ce lors de deux journées distinctes.

3.3 Résultats

L'analyse des matrices de dissimilarités est réalisée à l'aide du logiciel intégré de Statistiques de SPSS (Statistical Package for Social Sciences) qui propose l'algorithme INDSCAL.

L'espace perceptif objet qui sous-tend la perception des 18 échantillons de parole codés est représenté par la figure FIG. 1. Les coordonnées sont données en Annexe A.

Certains critères, tels que la valeur du stress, la proportion de la variance expliquée ou RSQ, et le poids moyen donné par les auditeurs sur chaque dimension, indiquent que le nombre de dimensions approprié est quatre (stress = 0,18, RSQ (squared correlation) = 69%).

On observe un regroupement des codecs par caractéristiques techniques.

La Dimension 1 contribue pour 23,22 % de la variance totale expliquée. En écoutant les stimuli suivant la dimension 1 (par exemple O1-G722.1C_24kbps_x2, O15-G729.1_24kbps_x2, O7-G722.2_15.85kbps_x2), il semble que cette dimension soit caractérisée par l'attribut "clair/sourd". A l'extrémité négative de cette dimension, se trouvent les codecs par transformée MLT (Modulated Lapped Transform) utilisant une allocation de bits par catégorisation [O1-G722.1C_24kbps_x2, O2-G722.1C_24kbps_x3, O3-G722.1C_24kbps_x2, O4-G722.1C_24kbps_x3] qui préservent le "caractère naturel" du signal de parole. L'extrémité positive de la dimension 1 est représentée par les codecs ACELP [O5-G722.2_12.65kbps_x2, O6-G722.2_12.65kbps_x3, O7-G722.2_15.85kbps_x2, O8-G722.2_8.85kbps_x2].

La Dimension 2 contribue pour 23,18 % de la variance totale expliquée. Comme l'indique la figure FIG. 1, l'analyse des codecs suivant la dimension 2 montre que les codecs ADPCM [O9-G722_48kbps_x2, O10-G722_48kbps_x3, O11-G722_56kbps_x2, O12-G722_56kbps_x3] situés à l'extrémité négative, présentent clairement un bruit de fond contrairement aux autres. Ainsi la dimension 2 se définirait par l'attribut "bruit de fond".

La Dimension 3 (12,58 % de la variance totale expliquée) présente à son extrémité négative les codecs hybrides [O13-G729.1_14kbps_x3, O14-G729.1_20kbps_x3, O15-G729.1_24kbps_x2, O16-G729.1_32kbps_x3] et les codecs MP3 [O17-MP3_32kbps_x1, O18-MP3_32kbps_x2] qui contiennent du pré-écho. Ainsi, l'attribut qui caractériserait la dimension 3 serait "pré-écho".

Enfin la dimension 4 (10,49 %) oppose les codecs hybrides [O13-G729.1_14kbps_x3, O14-G729.1_20kbps_x3, O15-G729.1_24kbps_x2, O16-G729.1_32kbps_x3] aux codecs MDCT utilisant un modèle psycho-acoustique [O17-MP3_32kbps_x1, O18-MP3_32kbps_x2] qui laissent percevoir du bruit quand la parole est présente. Ainsi nous qualifierons la dimension 4 de "bruit sur le signal de parole".

4. Conclusion

Le but de cette étude est de proposer un nouveau système de référence pour l'évaluation subjective de la qualité des codecs actuels. Pour cela, il est nécessaire de caractériser d'un point de vue perceptif les dégradations apportées par les nouveaux codecs, puis de les relier aux techniques de codage afin de pouvoir les recréer artificiellement.

Dans une brève présentation de la MDS, nous avons montré comment la MDS non métrique INDSCAL, analyse multidimensionnelle des proximités non métrique pondérée, peut être appliquée à des jugements subjectifs de dissimilarité entre paires de stimuli afin de déterminer une

configuration spatiale sous-tendant la perception de stimuli.

Nous avons ensuite présenté l'analyse des résultats d'un test d'écoute dans lequel les sujets expriment leurs jugements de dissimilarité sur des paires d'échantillons de parole wideband codés. L'analyse des matrices de dissimilarités par INDSCAL a fourni un espace de stimuli où les échantillons de parole codés pour un locuteur homme peuvent être représentés dans un espace perceptif à quatre dimensions (naturel, bruit de fond, pré-écho, bruit sur le signal de parole), et où les codecs se regroupent par caractéristiques techniques.

Ce résultat justifie la suite de notre démarche destinée à faire le lien entre les dégradations sonores identifiées et les caractéristiques techniques des codecs. Cela nous permettra de recréer artificiellement ces dégradations et de les introduire dans les signaux de parole.

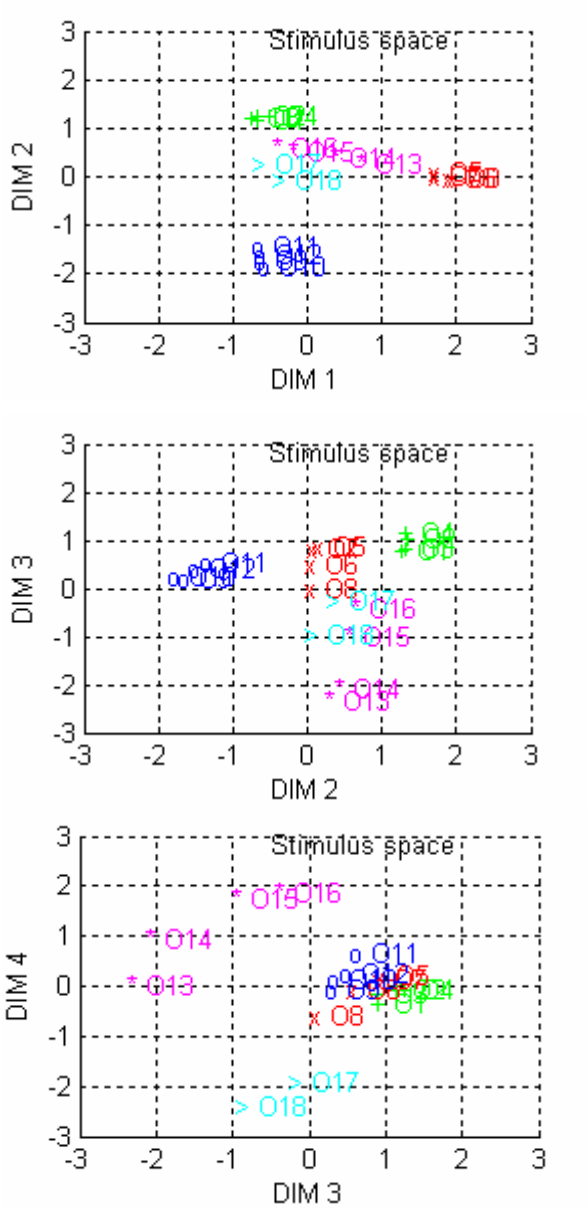

FIG. 1 : espace perceptif objet à quatre dimensions

Références

[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality", International Telecommunications Union, 08/96.

[2] ITU-T Recommendation P.810, "Modulated Noise Reference Unit (MNRU)", International Telecommunications Union, 02/96.

[3] T. Letowski, "Timbre, tone color, and sound quality: concepts and definitions," Archives of Acoustics, 17(1):17–30, 1992.

[4] C. Osgood, "The nature and measurement of meaning," Psychological Bulletin, 49 (197-237), 1952.

[5] 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, March 22-25 2003.

[6] J.L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," J. Acoust. Soc. Am. 110 (4), Oct. 2001.

[7] J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika, Vol. 29, pp. 1-27, 1964.

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

[9] J.D. Carroll, "Individual differences and multidimensional scaling," R.N.Shepard, A.K.Romney & S.B. Nerlove (Eds.) Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Vol.1, pp.105-155, New York and London: Seminar Press, 1972.

[10] J.D. Carroll and J.J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," Psychometrika, 35, pp. 283-319, 1970.

Annexe A

Codec	Dim. 1	Dim. 2	Dim. 3	Dim. 4
G722.1C_24kbps_x2	-0,78	1,17	0,78	-0,34
G722.1C_24kbps_x3	-0,77	1,24	1,03	-0,09
G722.1_24kbps_x2	-0,86	1,21	0,80	-0,18
G722.1_24kbps_x3	-0,61	1,22	1,13	-0,08
G722.2_12.65kbps_x2	1,62	0,06	0,83	0,19
G722.2_12.65kbps_x3	1,80	-0,05	0,46	-0,10
G722.2_15.85kbps_x2	1,62	-0,02	0,82	0,12
G722.2_8.85kbps_x2	1,87	-0,05	-0,01	-0,62
G722_48kbps_x2	-0,74	-1,75	0,20	-0,08
G722_48kbps_x3	-0,67	-1,87	0,22	0,13
G722_56kbps_x2	-0,76	-1,44	0,53	0,63
G722_56kbps_x3	-0,72	-1,61	0,39	0,24
G729.1_14kbps_3	0,66	0,25	-2,37	0,02
G729.1_20kbps_x3	0,31	0,38	-2,12	0,94
G729.1_24kbps_x2	-0,25	0,51	-1,00	1,74
G729.1_32kbps_x3	-0,47	0,58	-0,43	1,84
MP3_32kbps_x1	-0,76	0,23	-0,27	-1,95
MP3_32kbps_x2	-0,49	-0,07	-0,97	-2,42d