Modèle hybride pour l'évaluation de la qualité vocale sans référence, appliqué à la téléphonie

Adrien Leman, Julien Faure, Etienne Parizet

10ème **C**ongrès **F**rançais d'**A**coustique
Lyon, 12-16 Avril 2010

Modèle hybride pour l'évaluation de la qualité vocale sans référence, appliqué à la téléphonie

Adrien Leman[1, 2], Julien Faure[1], Etienne Parizet[2]

[1]Orange Labs - Lannion, Technopole Anticipa 2 Avenue Pierre Marzin 22300 Lannion, adrien.leman@orange-ftgroup.com
[2]Laboratoire de mécanique de l'INSA de Lyon, 20 Avenue Albert Einstein F-69621 Villeurbanne, etienne.parizet@lva.insa-lyon.fr

L'évaluation de la qualité vocale est un phénomène multidimensionnel faisant intervenir certains attributs perceptifs dépendant des dégradations induites par le type de communication utilisé (RTC/RNIS/VoIP/GSM). Le modèle proposé est basé sur un espace perceptif multidimensionnel dont les dimensions se combinent pour fournir un indicateur global de qualité. L'identification des dimensions a permis de révéler trois attributs perceptifs prépondérants: la bruyance, la coloration, et la continuité. Chacune des dimensions est estimée à partir d'indicateurs paramétriques (issues des statistiques réseaux) ou basés sur le signal sonore (indicateurs psychoacoustiques, indicateurs physiques). Cette approche hybride avec un cœur commun et des estimateurs pouvant s'adapter selon l'information accessible permet à la fois d'optimiser les performances et la mise en œuvre du modèle.

1 Introduction

Les services de télécommunication sont de plus en plus nombreux et variés avec l'apparition de nouvelles technologies (RTC, RNIS, GSM, VoIP). Les opérateurs de téléphonie ont ainsi besoin de superviser en temps réel la qualité vocale des services qu'ils proposent. La qualité vocale peut être évaluée par des campagnes de tests subjectifs en demandant directement l'avis aux utilisateurs, cependant les méthodes existantes sont très couteuses et peu adaptées à la supervision. Les modèles objectifs sont ainsi proposés afin de simuler la perception des utilisateurs et d'évaluer la qualité vocale à moindre coût.

Les modèles utilisés actuellement pour la supervision peuvent être classés dans deux familles :

- les modèles intrusifs (avec référence) sont basés sur la comparaison entre un signal de référence injecté dans le réseau et le signal dégradé capté à l'autre extrémité du réseau (e.g. PESQ [1]). Ce sont des modèles fiables mais très difficiles à mettre en œuvre.
- Les modèles non-intrusifs (sans référence) utilisent seulement les informations disponibles au point de mesure. Ils utilisent soit des indicateurs basés sur le signal qui requièrent une grande consommation de CPU, soit des indicateurs paramétriques déterminés par les statistiques du réseau (taux de perte de paquet, gigue) qui rendent uniquement compte de l'état du dernier réseau IP traversé et non de la qualité de bout en bout.

Cette étude présente la construction d'un modèle non-intrusif de l'évaluation de la qualité vocale utilisant des indicateurs hybrides (indicateurs basés sur le signal et/ou paramétriques), appliquée à une transmission téléphonique.

Cette approche hybride permettra d'optimiser la performance du modèle en tirant partie de l'ensemble des indicateurs disponibles au point de mesure.

2 Approche multidimensionnelle

Les utilisateurs sont sensibles aux attributs perceptifs engendrés par les dégradations physiques issues de la transmission de la parole. En présence d'une simple dégradation physique, les attributs perceptifs peuvent être multiples. La structure globale du modèle hybride d'évaluation de la qualité vocale est construite selon une approche multidimensionnelle.

L'approche multidimensionnelle consiste à déterminer l'espace perceptif correspondant à des dégradations physiques (Mc Dermott [2], Bappert [3], Gabrielsson [4], Waltermann [5]). La qualité vocale peut alors être estimée comme une fonction (classiquement, une combinaison linéaire) des dimensions perceptives constituant l'espace. Les principales difficultés consistent alors à identifier les attributs perceptifs liés aux dimensions de l'espace, puis à modéliser ces dimensions.

Deux tests subjectifs ont été réalisés afin de construire le modèle hybride pour l'évaluation de la qualité vocale. Un premier test a été mené afin de mesurer les dissimilarités entre les différents stimuli soumis à diverses dégradations physiques engendrées par une transmission téléphonique. Ces dissimilarités ont ensuite été utilisées afin d'extraire l'espace perceptif. Un deuxième test a permis d'obtenir une note MOS (Mean Opinion Score) globale de qualité vocale pour chacun des signaux dégradés proposés au premier test.

3 Tests subjectifs

3.1 Base sonore

La base sonore est composée de 23 conditions de dégradations de la parole, correspondant aux dégradations rencontrées pour différents types de télécommunications actuelles (ex: VoIP, GSM, RNIS, RTC). Les conditions de dégradations sont composées de pertes de paquets (ppl), d'erreurs de bits (BER), de différents codages à bas débit (G.711, G.726, G.729, GSMEFR) utilisant pour certains l'algorithme "Packet Loss Concealment" (PLC) [6], ainsi que du bruit de fond aléatoire (de type bruit rose) diffusé à 7 niveaux sonores (BDF) (cf. Tableau 1). Chacune des dégradations est représentée par deux phrases prononcées respectivement par un locuteur femme et un locuteur homme pour un total de 46 stimuli. Tous les signaux so-

nores sont échantillonnés à 8 kHz, 16 bit, et filtrés à l'aide d'un filtre passe bande de type IRS (300-3400 Hz). Les stimuli sont restitués par casques aux oreilles des auditeurs en écoute diotique à un niveau de restitution optimal de 79 dB SPL.

1. G.711	13. G.729_ BDF3
2. G.711_ BDF1	14. G.729_G.729
3. G.711_G.726	15. G.729_G.729_ BDF5
4. G.711_ GSMEFR	16. G.729_G.729_ppl4
5. G.711_GSMEFR_ BDF6	17. G.729_G.729_ppl8
6. G.711_GSMEFR_BER2	18. G.729_GSMEFR
7. G.711_GSMEFR_BER6	19. G.729_GSMEFR_BDF7
8. G.711_ppl10_BDF4	20. G.729_GSMEFR_BER2
9. G.711_ppl2noplc_ PBN2	21. G.729_GSMEFR_BER4
10. G.711_ppl4	22. G.729_ppl12
11. G.711_ppl6noplc	23. G.729_ppl6
12. G.729	

Tableau 1: Les 23 conditions de dégradations

3.2 Sujets

48 sujets âgés de 20 à 52 ans participent aux tests subjectifs (24 sujets évaluent la voix d'homme et 24 sujets évaluent la voix de femme). Tous les sujets ont une capacité auditive normale (pas de déficience auditive). Ils n'ont pas été entrainés dans la tâche d'évaluation des caractéristiques sonores et peuvent donc être assimilés à des auditeurs naïfs.

3.3 Evaluation des dissimilarités

Les jugements des dissimilarités ont été obtenus à partir d'un test de comparaison par paire. Il était demandé aux sujets d'évaluer les dissimilarités de chacune des paires sur une échelle linéaire bornée par les adjectifs "identique" et "très différent". Cette échelle était complétée par sept repères afin de guider les sujets dans la tâche du jugement des dissimilarités. Un total de 253 paires de stimuli a été diffusé aux oreilles des auditeurs avec un ordre de présentation aléatoire.

3.4 Evaluation de la qualité vocale globale

L'évaluation de la qualité vocale a été réalisée par une adaptation du test ACR (Absolute Category Rating test). Les stimuli sont présentés un par un et sont évalués de manière indépendante sur une échelle continue à cinq intervalles labélisée par des adjectifs caractérisant la qualité vocale (Excellent, Bon, Moyen, Médiocre, Mauvais). Les sujets ont alors cinq secondes après l'écoute de chaque stimulus pour valider leur réponse.

Les réponses des sujets ont été moyennées afin d'obtenir la note globale de la qualité vocale représentée par la note MOS-LQSN (Mean Opinion Score – Listening Quality Subjective Narrowband).

4 Espace perceptif

4.1 Méthode d'extraction

La méthode INDSCAL introduite par Caroll et Chang (1970) a été utilisée afin d'extraire l'espace perceptif à partir des résultats du test de dissimilarité. Cette méthode statistique permet de définir le meilleur compromis entre le nombre de dimensions, la bonne représentation des dissimilarités, ainsi que la capacité à identifier chaque dimension [7].

4.2 Espace perceptif à trois dimensions

L'espace à trois dimensions est le meilleur compromis pour représenter les caractéristiques sonores des voix de femme et d'homme. La régression linéaire des trois dimensions perceptives explique très bien la note MOS obtenue lors du deuxième test (**r=0.93,** p<0.001, cf. Figure 1.C). Une analyse par "bootstrap" montre que les coordonnées des stimuli le long d'une quatrième dimension ne sont pas significativement différentes.

Afin d'obtenir un espace global commun aux deux types de voix, l'espace perceptif de la voix d'homme a subit une transformation par rotation par rapport à l'espace perceptif de la voix de femme. L'espace global ainsi obtenu est indépendant de la phrase et du locuteur.

Le Tableau 2 expose les coefficients de corrélation entre les deux locuteurs pour les trois dimensions. Les positions des stimuli pour les deux locuteurs sont très similaires pour les deux premières dimensions (respectivement **r=0.98** et **r=0.96**, p<0.001). Cependant, une faible différence est observée entre les deux locuteurs pour la troisième dimension (**r=0.81**, p<0.001). Une discussion est proposée dans la partie (4.3) sur ce sujet.

	Dim1	Dim2	Dim3
Female Vs male	r=0.98	R=0.96	r=0.81

Tableau 2: Coefficient de corrélation entre les positions des stimuli des locuteurs femme et homme pour chacune des trois dimensions

L'espace perceptif tridimensionnel est présenté sur la Figure 1 par les deux graphiques (A et B).

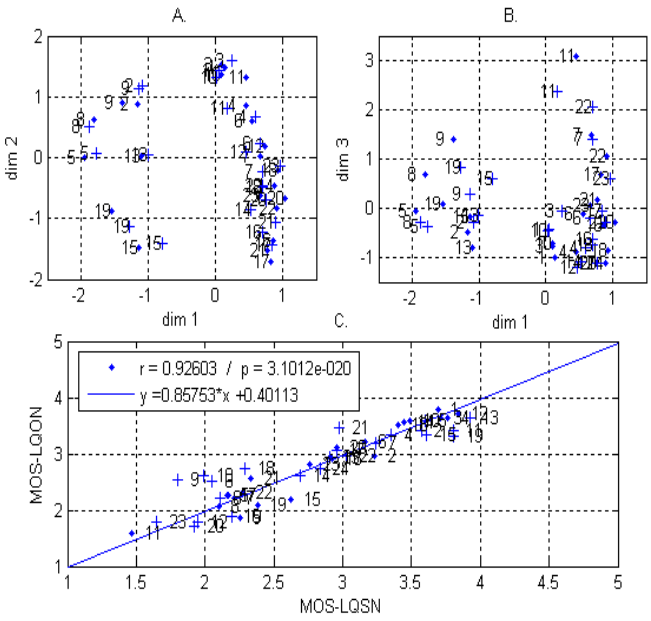

Figure 1: Les figures A et B présentent l'espace tridimensionnel pour les voix d'homme ("+") et de femme (".") . La figure C correspond à la comparaison des notes MOS-LQSN et MOS-LQON. Les stimuli sont numérotés de 1 à 23 correspondants aux conditions de dégradation.

La Figure 1C présente la comparaison entre les notes MOS-LQSN issues du test subjectif et les notes MOS-LQON calculées à partir d'une combinaison linéaire des trois dimensions de l'espace perceptif :

$$MOS.LQON = a + b.\dim 1 + c.\dim 2 + d.\dim 3 \quad (1)$$

Les variables dim1, dim2 et dim3 correspondent aux coordonnées de chaque point dans l'espace. Le coefficient de détermination entre ces deux types de notes MOS est $\mathbf{R^2=0.86}$, p<0.001. Ce résultat montre que l'espace perceptif issu du test de dissimilarité est bien approprié à l'estimation de la qualité vocale.

Les valeurs des coefficients b, c, et d montrent que la troisième dimension (continuité) est la plus influente sur l'impact de l'évaluation de la qualité vocale. Les dimensions 1 et 2 sont ensuite équivalentes.

4.3 Identification des trois dimensions perceptives

La première dimension est caractérisée par le niveau sonore du bruit de fond présent sur le signal de la parole. A l'extrémité négative de l'axe, le niveau sonore du bruit de fond est élevé tandis qu'à l'extrémité positive, les signaux ne comportent pas de bruits de fond (cf. Figure 1.A 1.B). La première dimension est intuitivement interprétée comme **la bruyance** du signal sonore.

La deuxième dimension est étroitement liée au type de codage employé (cf. Figure 1A) L'attribut perceptif correspondant au codage de la parole est principalement la brillance ou la coloration ([8], [3], [9], [5]), mais aussi le sifflement [8], la clarté [3], ou encore le naturel de la voix [10]. La bruyance a aussi été identifiée comme un attribut perceptif correspondant au codage de la parole [8, 10], mais ce dernier est déjà pris en compte dans notre espace perceptif par la dimension "bruyance". Nous nommerons par la suite cette dimension par l'attribut **coloration** du signal sonore.

La troisième dimension est caractérisée par la présence de pertes de paquets et d'erreurs de bits liées à la transmission mobile (cf. Figure 1B). Cette dimension est identifiée comme la **continuité** du signal sonore.
Nous remarquons une faible différence entre les locuteurs femme et homme pour les coordonnées des stimuli le long de la troisième dimension (cf. Tableau 2). Cette différence s'explique par la localisation des pertes sur le signal de la parole qui varie selon les phrases prononcées par les locuteurs femme ou homme. Les patterns d'erreurs sont identiques pour les deux locuteurs, ce qui va générer différentes perceptions de discontinuité selon que les pertes se trouvent sur les zones actives (parole) ou non-actives (bruit de fond) du signal sonore.

5 Construction du modèle

5.1 Structure globale du modèle

La note de qualité vocale est prédite par une combinaison linéaire des trois dimensions de l'espace perceptif. Afin de construire un modèle automatique d'évaluation de la qualité vocale, chaque dimension doit être estimée par des indicateurs physiques. Le modèle hybride utilise les indicateurs paramétriques (issus des statistiques réseaux) et/ou les indicateurs basés sur le signal pour estimer chacune des ces dimensions (cf. Figure 2).

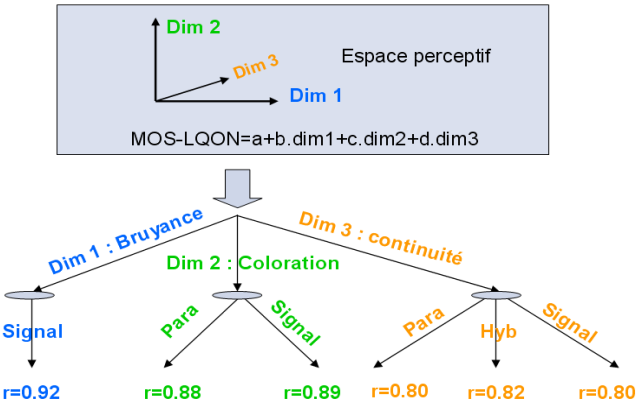

Figure 2: Structure globale du modèle non-intrusif d'évaluation de la qualité vocale, basée sur un espace tridimensionnel, avec les coefficients de corrélation associés à chaque indicateur

La première dimension "bruyance" ne peut pas être approximée par des indicateurs paramétriques car l'information concernant le niveau sonore du bruit de fond n'est pas disponible dans les statistiques réseaux. De ce fait, la première dimension est représentée par un indicateur basé sur le signal. Dans le cas des dimensions 2 et 3, un choix doit être réalisé afin d'utiliser les indicateurs les plus précis et/ou les plus adaptés au contexte de la mesure suivant les informations disponibles au point de mesure.

Les indicateurs utilisés par le modèle proposé sont décrits dans les parties suivantes.

5.2 La bruyance

La dimension bruyance est estimée à l'aide d'une combinaison d'indicateurs basés sur le signal présentée dans l'article [11]. Cette technique prend en compte la sonie du bruit de fond (modèle de Zwicker [12]), ainsi que l'effet du contenu informationnel du bruit de fond grâce à deux indicateurs basés sur le signal. Cette technique permet de représenter cette première dimension "bruyance" avec une précision de r=0.92, p<0.001.

Les conditions bruitées de notre étude sont composées uniquement de bruit rose aléatoire. Si le bruit de fond avait présenté du contenu informationnel (e.g. bruit d'environnement, bruit de musique), ce dernier aurait influencé l'évaluation de la qualité vocale [13]. L'effet du contenu informationnel du bruit de fond est pris en compte par le modèle hybride.

5.3 Coloration

5.3.1 Indicateur paramétrique

La dimension coloration peut être estimée avec les coefficients "Icod" dépendants du type de codage utilisé. Ces coefficients sont disponibles dans G.113 [14]. En présence de transcodage, l'utilisation de plusieurs codecs successifs n'est pas prise en compte dans G.113. Dans ce cas, le modèle E G.107 [15], ainsi que Moller [16] suggèrent d'ajouter les coefficients "Icod" de G.113. La Figure 3 présente la représentation de la dimension coloration par l'indicateur "Icod" en prenant en compte le transcodage.

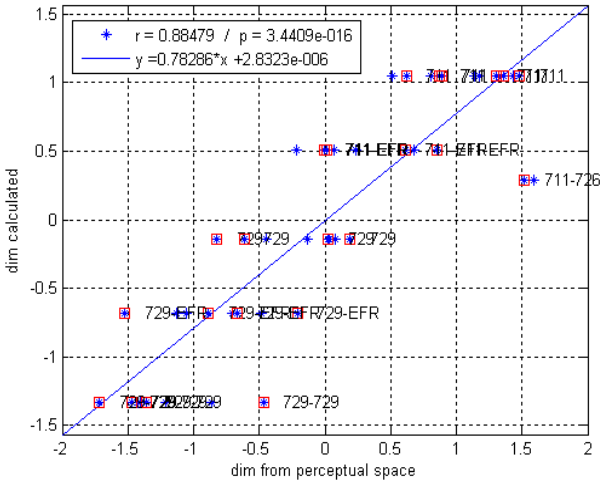

Figure 3: Estimation de la dimension "coloration" avec l'indicateur paramétrique de G.113 pour la voix de femme ("□") et la voix d'homme

Ce résultat (**r=0.88**, p<0.001) peut-être obtenu seulement si les différents codages sont identifiés dans le réseau. Si nous prenons seulement en compte le dernier codec utilisé, la performance de l'indicateur paramétrique chute à **r=0.67**.

5.3.2 Indicateur basé sur le signal

Différents indicateurs ont été calculés afin de quantifier la dimension correspondante à la coloration de la parole. Les indicateurs flux spectral et le centre de gravité spectral (centroïde spectral) représentent bien la coloration pour les deux voix d'homme et de femme séparément. Cependant, dans le cas de l'espace global, les corrélations diminuent (cf. Tableau3). Ces indicateurs ne sont pas adaptés au modèle non-intrusif principalement à cause de la différence de timbre entre les deux voix.

	Femme	Homme	Global
Flux spectral	r=-0.84	r=-0.93	r=-0.68
Centroïde	r=0.63	r=0.84	r=0.35

Tableau3: Coefficients de corrélation de Pearson entre les différents indicateurs et la seconde dimension pour les voix de femme, d'homme, et de l'espace global

Néanmoins, un indicateur basé sur le signal a été développé mais non présenté dans cet article. Cet indicateur ne dépend pas du timbre du signal de parole et représente précisément cette seconde dimension: **r=0.89**, p<0.001.

5.4 La continuité

5.4.1 Indicateur paramétrique

La dimension continuité peut être estimée par le pourcentage de pertes de paquets et le pourcentage d'erreurs de bits déterminés par les statistiques du réseau. Le pourcentage d'erreurs de bits (*pb*) et le pourcentage de pertes de paquets (*ppl*) sont pondérés afin d'obtenir un pourcentage global de discontinuités appelé (*pdg*):

$$(2)\quad \begin{cases} pdg = 12 \times pb & \text{erreurs de bits,} \\ pdg = 3.2 \times ppl & \text{pertes de paquets sans PLC,} \\ pdg = ppl & \text{sinon.} \end{cases}$$

La Figure 4 présente l'estimation de cette troisième dimension à l'aide de l'indicateur paramétrique "*pdg*" (**r=0.80,** p<0.001):

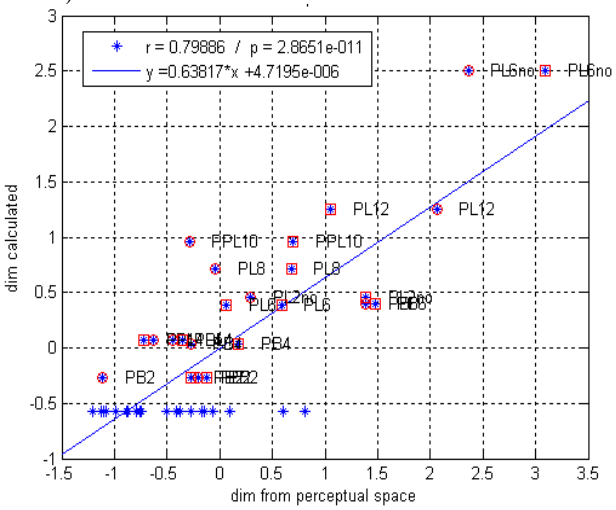

Figure 4: Estimation de la dimension "continuité avec l'indicateur paramétrique pour la voix d'homme ("o") et de femme ("□")

5.4.2 Indicateur hybride

Les discontinuités sont plus gênantes lorsqu'elles sont situées sur les zones actives de la parole plutôt que sur les zones non-actives (cf. 4.3.). Un algorithme de VAD (Voice Activity Detection) [17] basé sur l'analyse du signal de parole est utilisé pour définir les zones de parole active pour ensuite ajuster le paramètre pdg issu des statistiques réseau (cf. 5.4.1). Les résultats de l'estimation de la troisième dimension sont présentés sur la Figure 5.

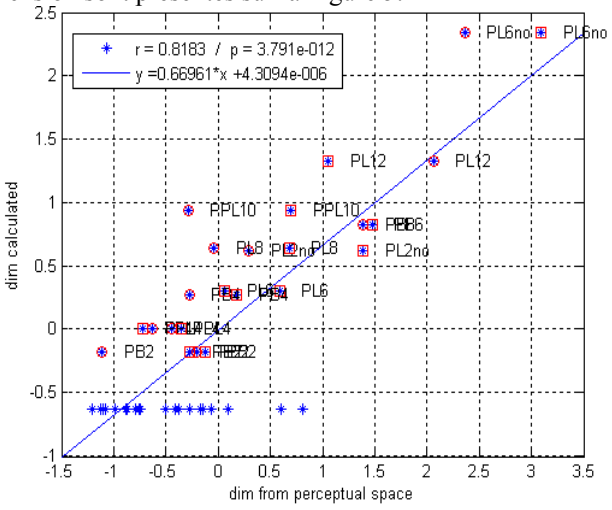

Figure 5: Estimation de la dimension "continuité" avec des indicateurs hybrides pour la voix d'homme ("o") et la voix de femme ("□")

Cette représentation utilisant des indicateurs hybrides obtient de meilleurs résultats qu'avec l'indicateur paramétrique, principalement à cause de la différence de taux d'activité entre les stimuli homme et femme (**r=0.82**, p<0.001).

5.4.3 Indicateur basé sur le signal

Les discontinuités sont difficiles à déterminer lors d'une modélisation non-intrusive à cause des performances des algorithmes de PLC [6]. Cette étude propose une nouvelle

approche afin d'estimer les effets de discontinuité. Elle consiste à analyser seulement la partie active de la parole soumise à un filtre passe bas d'ordre deux à la fréquence de coupure de 80 Hz. Dans ces conditions, il n'y a plus de parole dans le signal résultant (filtre IRS 300-3400 Hz), mais la présence de discontinuités génère des pics dans le domaine temporel. Les spectrogrammes présentés Figure 6 montrent cet effet pour la condition de dégradation 11 correspondant à 6% de pertes de paquets sans utiliser l'algorithme PLC.

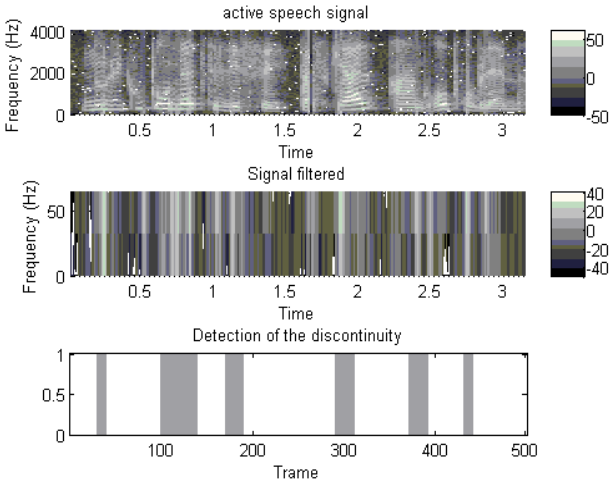

Figure 6: Localisation des zones de discontinuité pour la condition de dégradation 11 prononcée par le locuteur femme. 1er : Spectrogramme du signal de parole actif, 2éme Signal filtré, 3éme Zones de discontinuité.

Des indicateurs sont ensuite déterminés sur le signal filtré, puis une combinaison entre ces indicateurs permet de représenter la dimension continuité avec une corrélation de **r=0.80** (p<0.001).

6 Performances du modèle hybride

Cette partie présente la combinaison entre les indicateurs des trois dimensions afin de modéliser la note globale de la qualité vocale, grâce à l'équation (1). Ensuite, le modèle proposé et comparé avec les modèles existants.

6.1 Modèle adaptatif de l'évaluation de la qualité vocale

Le Tableau 4 expose les résultats des corrélations entre les notes MOS-LQON globales estimées et les notes MOS-LQSN issues du deuxième test subjectif, en considérant que la 1ére dimension "bruyance" ne peut-être estimée que par un indicateur basé sur le signal.

Continuité / Coloration	Param	Hybride	Signal
Param	r=0.83	r=0.86	r=0.76
Signal	r=0.86	r=0.88	r=0.78

Tableau 4: Précisions des différents types de modèles proposés utilisant des indicateurs paramétriques, basés sur le signal ou hybrides, déterminées par le coefficient de corrélation de Pearson entre les notes MOS-LQSN et MOS-LQON (p<0.001)

Le modèle le plus performant est obtenu en estimant la bruyance et la coloration par les indicateurs basés sur le signal, et en estimant la continuité par l'indicateur hybride. La corrélation obtenue est alors de **r=0.88**, p<0.001 (cf. Figure 7).

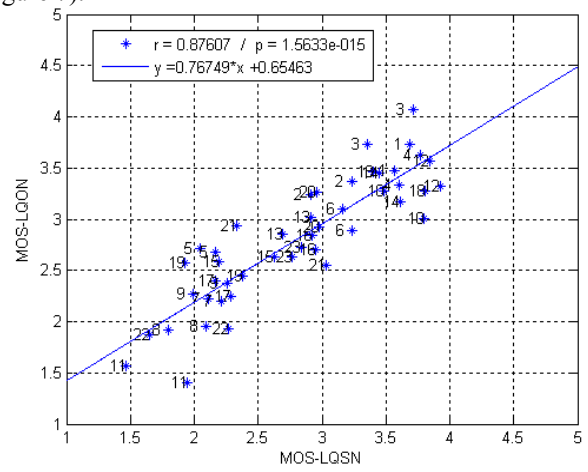

Figure 7: Comparaison des notes MOS-LQSN issues du second test subjectif et les notes MOS-LQON prédites par le modèle le plus performant

Dans le cas de la dimension "coloration", les indicateurs basés sur le signal sont plus précis mais sont plus coûteux en CPU que les indicateurs paramétriques. Cependant, l'indicateur paramétrique nécessite la connaissance préalable de l'ensemble des transcodages depuis la source du flux de parole. Si nous considérons seulement le dernier codec utilisé, les performances de l'indicateur paramétrique de la coloration chutent de **r=0.88** à **r=0.67** (cf. Figure 3).

Les deux types d'indicateurs (paramétrique et basé sur le signal) sont efficaces dans le cas de la dimension "continuité". La précision de l'indicateur hybride est légèrement supérieure à celle de l'indicateur paramétrique. De plus la VAD [17] utilisée par l'indicateur hybride est déjà déterminée pour estimer la première dimension. Les indicateurs paramétrique et hybride sont donc identiques en termes de consommation CPU, lors de la représentation de la dimension continuité.

Néanmoins, l'indicateur basé sur le signal n'est pas assez précis pour l'estimation de la "continuité". Cela cause une baisse générale des performances du modèle.

6.2 Comparaison du modèle proposé avec les modèles existants

Cette partie présente les résultats de la prédiction de la qualité vocale par deux modèles non-intrusifs couramment utilisés. Le modèle non-intrusif P.563 [18] utilise des indicateurs basés sur le signal, tandis que le modèle E (G.107 [15]) utilise des indicateurs paramétriques. Le Tableau 5 présente ces résultats.

Indicateur / Modèle	Param	Signal	Hybride
E-model	r=0.52		
P.563		r=0.51	
Modèle proposé	r=0.61	r=0.78	r=0.88

Tableau 5: Comparaison des différents modèles existants avec le modèle hybride proposé à l'aide des coefficients de corrélation entre les notes MOS-LQSN et MOS-LQON

Les corrélations obtenues avec le modèle non-intrusif proposé sont meilleures que celles obtenues avec les modèles non-intrusifs existants que ce soit pour des indicateurs uniquement paramétriques ou uniquement basés sur le signal. De plus, le modèle hybride proposé améliore encore la précision de la prédiction de la note globale de qualité vocale par rapport aux modèles non hybrides.

Le modèle intrusif normalisé à l'UIT-T d'évaluation de la qualité vocale en condition de bande étroite est le modèle PESQ [1]. Il obtient une précision de $r=0.62$, $p<0.001$.

7 Conclusion

Cet article décrit un modèle non-intrusif utilisant des indicateurs hybrides afin d'évaluer la qualité vocale. La structure globale du modèle repose sur un espace tridimensionnel ("bruyance", "coloration" et "continuité"). Chacune de ces trois dimensions est estimée par différents types d'indicateurs (paramétrique, basé sur le signal et hybride). Cette étude a permis de montrer l'efficacité de l'utilisation d'indicateurs hybrides afin d'augmenter la performance et la précision de l'évaluation de la qualité vocale.

Le choix du type d'indicateur utilisé dépend des informations disponibles au point de mesure sur le réseau. Le choix du type d'indicateur peut-être un compromis entre la précision attendue et les ressources de calcul du modèle global. Quels que soit le choix ou la disponibilité des indicateurs, du fait de la structure multidimensionnelle du modèle, la note globale de la qualité vocale est exprimée sur une échelle unique et comparable.

Le modèle hybride a l'avantage, grâce à cette structure tridimensionnelle, de proposer en plus de la note globale de qualité vocale, des notes de qualité pour chacune des trois dimensions. Suivant les valeurs des trois notes obtenues, le modèle identifie la (ou les) cause principale de la baisse de qualité vocale perçue, correspondant à la présence de bruit de fond (avec ou sans contenu informationnel [11]), ou au codage utilisé, ou encore à la présence de discontinuité. Cette technique est bénéfique à la supervision de la qualité vocale, afin d'améliorer la qualité des services proposés.

Les études futures porteront sur l'amélioration de la modélisation de la dimension "continuité". Ce modèle pourra aussi être amélioré en considérant d'autres dégradations comme le niveau sonore du signal vocal, le time-warping, ou encore la bande élargie.

Références

[1] ITU-T, Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End to End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," Geneva (2002).

[2] B. J. McDermott, "Multidimensional Analyses of Circuit Quality Judgments," *Journal of the Acoustical Society of America*, vol. 45, pp. 774-781 (1969).

[3] V. Bappert and J. Blauert, "Auditory quality evaluation of speech-coding systems," *acta acustica*, vol. 2, pp. 49-58 (1994).

[4] A. Gabrielsson and H. Sjogren, "Perceived sound quality of sound-reproducing systems," *Journal of the Acoustical Society of America*, vol. 65, pp. 1019-1033 (1979).

[5] M. Waltermann, K. Scholz, S. Moller, L. Huo, A. Raake, and U. Heute, "An Instrumental Measure for End-to-end Speech Transmission Quality Based on perceptual Dimensions : Framework and Realization," presented at Interspeech 08, Brisbane, Australia (2008).

[6] UIT-T, G.711 Appendice1, "Algorithme simple de haute qualité pour le masquage des pertes de paquets en codage G.711," Genève (1999).

[7] I. Borg and P. J. F. Groenen, *MODERN MULTIDIMENSIONAL SCALING, Theory and Applications, Second Edition* (2005).

[8] T. E. Etame, "Conception de signaux de référence pour l'évaluation de la qualité perçue des codeurs de la parole et du son" (2008).

[9] V.-V. Mattila, "Ideal point modelling of speech quality in mobile communications based on multidimensional scaling (MDS)," *Audio Engineering Society*, vol. 112 (2002).

[10] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *Acoustical Society of America*, vol. 110, pp. 2167-2182 (2001).

[11] A. Leman, J. Faure, and E. Parizet, "A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises," presented at Interspeech 09, Brighton (2009).

[12] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*: Springer; 2nd updated ed. edition (April 14, 1999).

[13] A. Leman, J. Faure, and E. Parizet, "Influence of informational content of background noise on speech quality evaluation for VoIP application," presented at Acoustics 08, Paris (2008).

[14] ITU-T, Rec. G.113, "Transmission impairments due to speech processing" (2007).

[15] UIT-T, Rec. G.107, "Le modèle E, Modèle de calcul utilisé pour la planification de la transmission" Genève (2003).

[16] S. Moller, *Assessment and Prediction of Speech Quality in Telecommunications*: Kluwer Academic (2000).

[17] UIT-T, Rec. P.56, "Mesure objective du niveau vocal actif," Genève (1993).

[18] L. Malfait, J. Berger, and M. Kastner, "P.563-The ITU-T Standard for Single-Ended Speech Quality Assessment," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14(6), pp. 1924-1934 (2006).