Diagnostic avancé de la qualité vocale en présence de discontinuités pour des signaux en bande audio super-élargie

SIBIRI TIEMOUNOU[1,2,3], REGINE LE BOUQUIN JEANNES[2,3], VINCENT BARRIAC[1]

[1]Orange Labs – Lannion, Av. Pierre Marzin, 22307 Lannion Cedex, France
[2]INSERM, U 1099, Rennes, F-35000 France
[3]Université de Rennes 1, LTSI Rennes, F-35000 France
[1,2,3]Sibiri.tiemounou@orange.com, [1]vincent.barriac@orange.com, [2,3]regine.le-bouquin-jeannes@univ-rennes1.fr

Résumé – Ce papier présente un outil de diagnostic des défauts perçus lors des communications téléphoniques, et en particulier des discontinuités présentes sur des signaux audio en bande super-élargie (50-14000 Hz). Cet outil, basé sur la modélisation de trois sous-familles de discontinuités, est doté d'un module de détection qui fournit des informations spécifiques quant à une présence éventuelle de discontinuité et oriente vers les causes possibles. Il rend non seulement une prédiction de la qualité vocale, mais également une quantification de la dégradation relative à chaque famille de discontinuités. L'outil proposé présente une performance globale de détection de discontinuités élevée, supérieure à 80%. En termes de prédiction, il présente une corrélation avec les résultats de tests subjectifs supérieure à 0,72 pour l'ensemble des discontinuités.

Abstract – This paper deals with a diagnostic method for voice quality impairments perceived in telephone communications and particularly discontinuities occurring on Super-wideband (50-14000 Hz) speech signals. This diagnostic method is based on the modeling of three families of discontinuities and integrates a discontinuity detection block providing specific information about the presence of discontinuity and the possible causes. Furthermore, it provides a global predicted Mean Opinion Score relative to the "Continuity" dimension together with additional Mean Opinion Scores for each sub-dimension. The proposed diagnostic method displays a global detection performance higher than 80%. In terms of prediction, it presents a correlation higher than 0.72 with subjective test values.

1 Introduction

L'objet de notre recherche consiste à développer un outil de diagnostic des dégradations de la qualité vocale perçues dans les communications téléphoniques en bande étroite (300-3400 Hz), élargie (50-7000 Hz) et super-élargie (50-14000 Hz), à partir d'une analyse du signal de parole. Cette analyse peut s'effectuer soit en comparant un signal de référence (*i.e.* signal non dégradé) à un signal dégradé, résultant du passage du signal de référence à travers un système de télécommunications, soit directement à partir du signal dégradé. Cet outil sera basé sur la modélisation des quatre dimensions perceptuelles couvrant l'espace perceptuel de la qualité vocale et dont les trois premières sont supposées orthogonales *i.e.* indépendantes les unes des autres. Il s'agit de la « Bruyance » relative au bruit de fond, la « Continuité » regroupant l'ensemble des discontinuités, la « Coloration » relative aux distorsions fréquentielles et la « Sonie » liée au niveau sonore de la parole. La présente étude concerne la dimension « Continuité » dans le cas de signaux audio en bande super-élargie. Cette dimension est relative aux discontinuités perçues dans le signal de parole pouvant être causées par des pertes de paquets (ou de trames) ou par des processus de traitement de signal tels que la réduction de bruit ou l'annulation d'écho. Dans le cas de pertes de paquets, des algorithmes de masquage, appelés PLC (Packet Loss Concealment), utilisés pour minimiser l'effet de ces pertes sur la perception, peuvent plus ou moins impacter la qualité vocale. Notre objectif est de trouver des indicateurs de qualité permettant de mieux caractériser ladite dimension et d'obtenir des informations spécifiques sur l'ensemble des différentes causes de discontinuité. Des études antérieures [1, 2] ont permis de diviser la dimension « Continuité » en trois sous-dimensions respectivement caractérisées par : (i) des *Coupures* perçues dans le signal de parole, (ii) des *Artéfacts Additifs* et (iii) une *Variation de Gain*. Les coupures et les artéfacts sont davantage perçus lorsque les techniques PLC de type « insertion de trames de silence » et « répétition de trames » sont respectivement employées. La *Variation de Gain* regroupe des dégradations impactant le gain du système de transmission, en particulier celles dues aux systèmes de débruitage et de Contrôle Automatique de Gain (CAG). Cette variation de gain se traduit par des atténuations ou des amplifications abruptes du niveau sonore perçues dans le signal de parole.

Si une première modélisation de la dimension « Continuité » a été proposée par Côté [3], celle-ci est uniquement basée sur les deux premières sous-dimensions et ne fournit qu'une prédiction de la qualité globale de la dimension. L'outil que nous proposons prend en compte non seulement la troisième sous-dimension mais fournit aussi, au-delà d'une prédiction globale de la qualité vocale, une prédiction de cette qualité pour chacune des sous-dimensions. Une autre de ses particularités est d'intégrer un module de détection permettant d'obtenir des informations sur la présence

éventuelle d'une discontinuité et sur sa nature. La Section 2 décrit les étapes de la modélisation de la dimension « Continuité », suivie de la validation de l'outil proposé dans la Section 3 avant de conclure dans la Section 4.

2 Modélisation de la dimension « Continuité »

Trois indicateurs sont introduits pour quantifier l'ensemble des discontinuités. Les deux premiers, « r_L » et « r_A » [3], permettent d'estimer respectivement le taux de trames perdues (*Coupures*) et le taux d'artéfacts (*Artéfacts Additifs*) présents dans le signal de parole. Ces indicateurs ont été retenus à l'issue de premières études [4, 5] sur la performance des indicateurs proposés dans l'état de l'art. Quant à la sous-dimension *Variation de Gain*, il n'existe, à notre connaissance, aucun indicateur dans la littérature la caractérisant. Pour pallier ce manque, nous avons proposé un indicateur, noté « V_G ».

Pour le calcul de ces indicateurs, un pré-traitement conforme à celui utilisé dans [3] est appliqué aux signaux de référence et dégradé. Le délai de transmission est d'abord estimé afin de synchroniser le signal dégradé par rapport au signal de référence, suivi d'une égalisation de leur niveau sonore à -26dBoV. Les signaux résultants sont ensuite divisés en trames de 16 ms (correspondant à 768 échantillons par trame, les signaux étant échantillonnés à 48 kHz) avec un recouvrement de 50%, le fenêtrage utilisé étant celui de Hanning. Les Densités Spectrales de Puissance (DSP) de ces signaux, estimées pendant les périodes d'activité vocale, sont obtenues en appliquant une transformée de Fourier à court-terme suivie d'une conversion dans le domaine de Bark suivant l'approche de Zwicker *et al.* [6]. Par ailleurs, l'effet des dégradations liées à la réponse en fréquence du système de transmission est partiellement compensé sur ces signaux [3].

L'indicateur « r_L » est déterminé à partir des variations instantanées dans les enveloppes des signaux de référence et dégradé (notées $d\hat{e}_x(l)$ et $d\hat{e}_y(l)$ respectivement, l représentant une trame donnée). Les coupures sont détectées lorsque la grandeur $\Delta(l) = d\hat{e}_x(l) - d\hat{e}_y(l)$ est supérieure à un seuil noté $\gamma(l)$ dépendant du Rapport Signal à Bruit (RSB) de la trame l. L'expression de « r_L » est la suivante :

$$r_L = \frac{\sum \delta(l)}{L}, \delta(l) = \begin{cases} 1, \text{ si } l \in \{\Delta(l) > \gamma(l)\} \\ 0, \text{ sinon} \end{cases}, \quad (1)$$

où L est le nombre total de trames d'activité vocale. Quant à l'indicateur « r_A », il est défini à partir de la distance de la pente spectrale pondérée (Weighted Spectral Slope) entre les signaux de référence et dégradé, notée $d_{WSS}(l)$. Les artéfacts sont décelés dès que la distance $d_{WSS}(l)$ est supérieure à un seuil, noté $\eta(l)$ dépendant du RSB et de la distribution des valeurs de la distance $d_{WSS}(l)$. Il s'ensuit l'expression de « r_A » :

$$r_A = \frac{\sum \phi(l)}{L}, \phi(l) = \begin{cases} 1, \text{ si } l \in \{d_{WSS}(l) > \eta(l)\} \\ 0, \text{ sinon} \end{cases}. \quad (2)$$

Enfin, l'indicateur « V_G » est déterminé à partir de la DSP du signal dégradé. Une estimation de la variation de gain est effectuée à partir de l'approche proposée dans [3]. En amont, pour atténuer l'impact du bruit, la DSP du bruit est estimée à partir de la DSP du signal dégradé sur les périodes de silence et lui est soustraite. L'expression du gain est alors donnée par :

$$G(l) = 10 \cdot \log_{10}\left(\frac{P_x(l) + \alpha}{P'_y(l) + \alpha}\right) \quad (3)$$

où $\alpha = 10^{-4}$, $P_x(l)$ est la DSP en Bark du signal de référence et $P'_y(l)$ celle du signal dégradé résultant de la compensation du bruit. Les valeurs du gain $G(l)$ sont limitées à l'intervalle [-20dB ; 20dB] afin de compenser l'effet des autres types de discontinuités. On cherche, dans le signal dégradé, les trames l telles que $G(l) < G_s(l) - 6$ (atténuation brusque), ou telles que $G(l) > G_s(l) + 3$ (amplification brusque), où $G_s(l)$ est une version lissée du gain $G(l)$ obtenue en appliquant à ce dernier un filtre passe-bas. Afin de refléter l'effet de ces variations tel qu'il est perçu par le système auditif humain, la DSP en Bark $P'_y(l)$ est transformée en sonie suivant le modèle de Zwicker et Fastl [7] et est notée $L'_y(l)$. L'expression de l'indicateur « V_G » devient alors :

$$V_G = \left(\frac{1}{L^v}\sum_{l^v=1}^{L^v} L'_y\left(l^v\right)^2\right)^{1/2}. \quad (4)$$

où L^v est le nombre total de trames présentant des variations abruptes du niveau sonore de la parole.

2.1 Détection automatique de discontinuités

Pour chaque indicateur, on définit un seuil minimal au-delà duquel la discontinuité correspondante est perçue. Pour ce faire, une base sonore (base 1) a été construite, constituée de quatre sous-ensembles. Le premier sous-ensemble, commun à l'analyse des trois indicateurs, comprend des stimuli impactés par des dégradations autres que des discontinuités : elles correspondent à des conditions de filtrage, de codage, de bruit de fond, et d'atténuation de niveau sonore, appliquées au signal de référence, comme indiqué dans le Tableau 1 (10 conditions au total). Quant au deuxième sous-ensemble, il contient 8 conditions de dégradation correspondant à différents degrés de pertes de paquets/trames associées au codec WB G.722 dont la technique PLC consiste en une « insertion de trames de silence ». Ce sous-ensemble est utilisé pour tester l'indicateur « r_L ». Concernant le troisième sous-ensemble, pour tester l'indicateur « r_A », 18 conditions

de pertes de paquets/trames associées aux codecs SWB G.718 Annexe B et G.729.1 Annexe E, intégrant une PLC par répétition de trames, sont considérées afin de simuler les artéfacts. Enfin, le quatrième sous-ensemble comprend 5 conditions dont 3 niveaux de débruitage (peu agressif, agressif et très agressif) et 2 niveaux de CAG, et est utilisé pour tester l'indicateur « V_G ». Ces conditions ont été appliquées sur 24 doubles phrases pour un total de 240, 192, 432 et 120 stimuli pour les quatre sous-ensembles respectivement.

Pour déterminer le seuil optimal de chaque indicateur, une étape d'apprentissage a été nécessaire durant laquelle 75% des stimuli ont été utilisés (base 1.1). De plus, nous avons adopté comme méthode d'apprentissage, l'arbre binaire de décision [8]. Cet algorithme permet de déterminer un seuil de décision qui classifie les données avec le minimum d'erreurs. Ainsi, pour chaque indicateur, nous avons considéré deux classes de stimuli, la première correspondant aux stimuli ne contenant pas de discontinuités et la deuxième relative aux discontinuités de l'indicateur considéré. Les entrées de cette méthode de décision sont les valeurs de l'indicateur associées aux stimuli et les deux classes correspondantes. Cela étant, pour chaque indicateur, le seuil optimal obtenu par application de l'algorithme de décision est le suivant : le signal est continu (*i.e.* ne contenant pas de discontinuités) si les valeurs de « r_L », « r_A » et « V_G » sont respectivement inférieures à 0,005, 0,0025 et 0,7563, sinon le signal est diagnostiqué comme discontinu. Le Tableau 2 permet d'apprécier les performances de l'outil. Lors de la phase d'apprentissage, il présente un taux de bonne détection de coupures et d'artéfacts supérieur à 90%. Si la performance la plus faible est obtenue par l'indicateur « V_G », elle n'en reste pas moins significative (plus de 80% de bonne détection).

2.2 Prédiction de la qualité vocale en présence de discontinuités

L'outil proposé fournit également une prédiction de la qualité vocale globale relative à la dimension « Continuité » et une prédiction pour chacune des sous-dimensions. La base sonore utilisée pour la prédiction est celle du Tableau 1. Pour la prédiction des différentes grandeurs, des fonctions de mapping ont été calculées à partir des valeurs des indicateurs et les notes moyennes d'opinion ou MOS (Mean Opinion Score) correspondantes. Ces fonctions correspondent à des régressions polynomiales d'ordre 3 avec un intervalle de confiance de 95%. La qualité vocale prédite s'exprime sous la forme :

$$MOS_p = a_0 + a_1 \cdot Ind + a_2 \cdot Ind^2 + a_3 \cdot Ind^3 , \quad (5)$$

où MOS_p correspond à la note MOS prédite, les paramètres a_i (i = 0,…,3) sont les coefficients de la fonction de mapping et *Ind* correspond à la valeur de l'indicateur considéré. Il faut noter que, pour la prédiction de la qualité vocale globale de la dimension « Continuité », *Ind* est défini comme suit :

$$Ind = \alpha \cdot r_L + \beta \cdot r_A + \lambda \cdot V_G , \quad (6)$$

où les coefficients α , β et λ ont été déterminés empiriquement de sorte à optimiser la performance de la prédiction, et valent respectivement 1, 1 et -0,03. L'avantage d'une telle combinaison est qu'elle permet de prédire la qualité en présence de discontinuités multiples. Les stimuli utilisés lors de cet apprentissage ont permis de déterminer les coefficients des fonctions de mapping. Les performances de la prédiction sont évaluées en termes (i) de corrélation (ρ) entre les notes MOS subjectives et les notes MOS prédites, et (ii) d'erreur quadratique moyenne (ε). D'après le Tableau 3, l'outil proposé présente une meilleure prédiction pour la sous-dimension *Coupures* en termes de corrélation (ρ = 0,94). Quant à la prédiction des sous-dimensions *Artéfacts Additifs* et *Variation de Gain*, les performances s'avèrent relativement inférieures (ρ = 0,8 et 0,76 respectivement). Enfin, la performance de prédiction de la dimension « Continuité » de notre outil (ρ = 0,85, ε = 0,33) est supérieure à celle obtenue par le modèle proposé dans [3] (ρ = 0,84, ε = 0,42).

3 Validation de l'outil proposé

La phase de validation a permis d'apprécier le comportement de notre outil sur deux bases de données inconnues et sa performance a été évaluée tant en détection qu'en prédiction. Pour la première base (notée base 1.2), les stimuli sont les 25% de la base 1 n'ayant pas été utilisés lors de la phase d'apprentissage et, pour la seconde base (notée base 2), elle est constituée de stimuli issus de la base sonore élaborée par l'UIT-T pour le développement de la norme P.863 [9].

3.1 Performances de détection

Dans cette phase de validation, la base 2 inclut des conditions réalistes composées de dégradations multiples. Elle comprend 898 stimuli dont 242 présentent des coupures, 226 des artéfacts, 120 des variations de gain dues au CAG et au débruitage, et 310 ne présentent aucune discontinuité (ces dernières relèvent de bruits non stationnaires, de distorsions fréquentielles, …). Il faut noter que, pour un indicateur donné, les types de discontinuité associés aux deux autres indicateurs ne sont pas considérés comme des défauts. Ainsi, les 898 stimuli sont considérés pour chaque indicateur. Le Tableau 2 montre un taux global de bonne détection supérieur à 80% pour les deux bases (bases 1.2 et 2). De plus, l'outil proposé présente une performance de détection de coupures et d'artéfacts supérieure à 89% et 78% respectivement pour ces bases. La plus faible performance est obtenue par l'indicateur « V_G » (79,8% et 73% pour les bases 1.2 et 2 respectivement). Après analyse de nos résultats, il s'est avéré que l'indicateur « r_A » était surtout sensible aux conditions relatives aux distorsions fréquentielles. Quant à l'indicateur « V_G », il était essentiellement impacté par des conditions contenant du bruit non stationnaire, des distorsions fréquentielles et des pertes de paquets/trames.

3.2 Performances de prédiction

Pour la validation de notre modèle sur la prédiction, la base 2 est limitée à 108 stimuli dégradés par des pertes de paquets/trames à des taux de 0, 2 et 20% sans codage (sous-dimension *Coupures*), 84 stimuli dégradés par le codec G722.1C, qui contient une PLC par répétition de trames, associé à des pertes de paquets de 0% et 2% (pertes aléatoires et en rafales) (sous-dimension *Artéfacts Additifs*) et 24 stimuli impactés par un débruitage agressif sans codage (sous-dimension *Variation de Gain*). Le Tableau 3 indique une corrélation supérieure à 0,89 pour la sous-dimension *Coupures* sur les deux bases inconnues et une corrélation supérieure à 0,72 pour les deux autres sous-dimensions. Il révèle aussi une meilleure prédiction de la qualité globale ($\rho \geq 0{,}81$, $\varepsilon \leq 0{,}39$) comparé au modèle proposé dans [3] ($\rho \geq 0{,}80$, $\varepsilon \geq 0{,}45$).

4 Conclusion

Dans cette étude, nous avons proposé un outil de diagnostic avancé des discontinuités perçues dans les contextes téléphoniques en bande super-élargie. En plus de deux indicateurs déjà présents dans la littérature, cet outil intègre un nouvel indicateur permettant de caractériser l'ensemble des causes de discontinuités connues. Le taux de détection des discontinuités s'avère élevé, supérieur à 80%, de même que la performance en matière de prédiction. Notre modèle se révèle un outil efficace de diagnostic pour une application en contexte de supervision et d'optimisation des réseaux de télécommunication tant du point de vue de la détection que de celui de l'identification et de l'impact de la (ou des) discontinuité(s) présente(s). Une ultime étape consistera à optimiser les indicateurs « V_G » et « r_A » en compensant l'effet de dégradations ne relevant pas du domaine des discontinuités.

5 Références

[1] Wältermann M., Scholz K., Raake A., Heute U. and Möller S., "Underlying Quality Dimensions of Modern Telephone Connections". In Proc. *9th International Conference on Spoken Language Processing (ICSLP 2006)*, 2170-2173, USA-Pittsburgh, PA, 2006.

[2] Wältermann M., Möller S., Raake A. and Beerends J. G., "Proposal for benchmarking of the P.OLQA degradation decomposition". ITU-T Study Group 12, Contribution COM 12-C74, September 2007.

[3] Côté N., "Integral and Diagnostic Intrusive Prediction of Speech Quality". Springer, Edition 2011.

[4] Tiémounou S., Le Bouquin Jeannès R. and Barriac V., "Performance evaluation of quality degradation indicators on super wideband speech signals". In Proc. *20th European Signal Processing Conference*, Bucharest, Romania, August 27-31, 2012.

[5] Tiémounou S., Le Bouquin Jeannès R. and Barriac V., "Assessment of speech quality degradation indicators for "continuity" dimension in super wideband telephony context". In Proc. *1st International Conference on Computing, Networking and Communications*, San Diego, USA, Jan. 28-31, 2013.

[6] Zwicker E., Flottorp G., and Stevens S., "Critical Bandwidth in Loudness Summation ". *Journal of the Acoustical Society of America*, 29(5):548–557, 1957.

[7] Zwicker E. and Fastl H., " Psychoacoustics: Facts and models ". Springer, DE-Berlin, 1st edition, 1990.

[8] Breiman L., Frieman J., Olshen R., and Stone C., "Classification and Regression trees". CRC Press, Belmont, California, 1984.

[9] ITU-T Rec. P.863, "Perceptual Objective Listening Quality Assessment (POLQA)". *International Telecommunication Union*, CH Geneva, 2011.

Tab 1 : Synthèse des conditions de dégradation de la base 1. Les cases grises correspondent aux conditions utilisées lors de la prédiction de la qualité vocale. PP/T signifie Pertes de Paquets/Trames. Les niveaux 1, 2 et 3 correspondent respectivement à « peu agressif », « agressif » et « très agressif »

Indicateurs de qualité / Conditions	r_L	r_A	V_G
Conditions (cond.) ne contenant pas de discontinuités (240 stimuli)	Signal de référence (non codé)		
	2 cond. de filtrage passe-bas (7 kHz et 10 kHz)		
	3 cond. de codage (G722, G718B et G729.1E)		
	2 cond. de bruit (voiture et restaurant, RSB = 20 dB)		
	2 cond. d'atténuation du niveau sonore (10 dB et 20 dB)		
Conditions contenant des discontinuités (24 stimuli/condition)	8 cond. : 1 codec (G722) associé à 1, 2, 3, 4, 5, 6, 8 et 10% de PP/T	18 cond. : 2codecs (G718B et G729.1E) associés à 1, 2, 3, 4, 5, 6, 8, 10 et 15% de PP/T	- 3 cond. de débruitage (niveaux 1, 2, et 3) - 2 cond. de CAG (niveaux 1 et 2)

Tab 2 : Performance de détection des discontinuités

Sous-dimensions / Bases de tests	*Coupures*	*Artéfacts Additifs*	*Variation de Gain*	Total
Apprentissage (base 1.1)	99,4%	90,53%	82,14%	91,1%
Validation sur la base 1.2	98,04%	88,3%	79,8%	88,32%
Validation sur la base 2	89,4%	78,62%	73%	80,3%

Tab 3 : Performance de prédiction de la qualité vocale. ρ et ε représentent respectivement la corrélation entre les notes MOS prédites et les notes MOS subjectives et l'erreur quadratique moyenne

Sous-dimensions / Bases de tests	*Coupures*	*Artéfacts Additifs*	*Variation de Gain*	Continuité	Modèle proposé dans [3]
Apprentissage (base 1.1)	ρ = 0,94, ε = 0,33	ρ = 0,80, ε = 0,36	ρ = 0,76, ε = 0,32	ρ = 0,85, ε = 0,33	ρ = 0,84, ε = 0,42
Validation sur la base 1.2	ρ = 0,92, ε = 0,35	ρ = 0,79, ε = 0,37	ρ = 0,75, ε = 0,35	ρ = 0,84, ε = 0,37	ρ = 0,84, ε = 0,45
Validation sur la base 2	ρ = 0,89, ε = 0,36	ρ = 0,77, ε = 0,39	ρ = 0,72, ε = 0,38	ρ = 0,81, ε = 0,39	ρ = 0,80, ε = 0,49