UNION INTERNATIONALE DES TÉLÉCOMMUNICATIONS

UIT-T P.862

(02/2001)

SECTEUR DE LA NORMALISATION
DES TÉLÉCOMMUNICATIONS
DE L'UIT

SÉRIE P: QUALITÉ DE TRANSMISSION TÉLÉPHONIQUE, INSTALLATIONS TÉLÉPHONIQUES ET RÉSEAUX LOCAUX

Méthodes d'évaluation objective et subjective de la qualité

Evaluation de la qualité vocale perçue: méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite

Recommandation UIT-T P.862

(Antérieurement Recommandation du CCITT)

RECOMMANDATIONS UIT-T DE LA SÉRIE P

QUALITÉ DE TRANSMISSION TÉLÉPHONIQUE, INSTALLATIONS TÉLÉPHONIQUES ET RÉSEAUX LOCAUX

Vocabulaire et effets des paramètres de transmission sur l'opinion des usagers	Série	P.10
Lignes et postes d'abonnés	Série	P.30
		P.300
Normes de transmission	Série	P.40
Appareils de mesures objectives	Série	P.50
		P.500
Mesures électroacoustiques objectives	Série	P.60
Mesures de la sonie vocale	Série	P.70
Méthodes d'évaluation objective et subjective de la qualité	**Série**	**P.80**
		P.800
Qualité audiovisuelle dans les services multimédias	Série	P.900

Pour plus de détails, voir la Liste des Recommandations de l'UIT-T.

Recommandation UIT-T P.862

Evaluation de la qualité vocale perçue: méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite

Résumé

La présente Recommandation décrit une méthode objective de prévision de la qualité subjective de la téléphonie à 3,1 kHz (bande étroite) avec combiné et des codecs vocaux à bande étroite. La présente Recommandation donne une description approfondie de la méthode, des conseils sur ses modalités d'utilisation et présente une partie des résultats d'un test de performance effectué par la Commission d'études 12 pendant la période 1999-2000. Une implémentation de référence C-ANSI, décrite à l'Annexe A, fait l'objet de fichiers séparés qui font partie intégrante de la présente Recommandation. Est également spécifiée à l'Annexe A une procédure d'essai de conformité qui permet à l'utilisateur de confirmer la validité d'une autre implémentation du modèle. Cette implémentation de référence C-ANSI prévaudra en cas de divergences entre le texte de la description figurant dans la présente Recommandation et la mise en œuvre de référence C-ANSI.

La présente Recommandation comporte un document électronique contenant une implémentation de référence en langage C de l'algorithme PESQ et des données de test de conformité.

Source

La Recommandation P.862 de l'UIT-T, élaborée par la Commission d'études 12 (2001-2004) de l'UIT-T, a été approuvée le 23 février 2001 selon la procédure définie dans la Résolution 1 de l'AMNT.

AVANT-PROPOS

L'UIT (Union internationale des télécommunications) est une institution spécialisée des Nations Unies dans le domaine des télécommunications. L'UIT-T (Secteur de la normalisation des télécommunications) est un organe permanent de l'UIT. Il est chargé de l'étude des questions techniques, d'exploitation et de tarification, et émet à ce sujet des Recommandations en vue de la normalisation des télécommunications à l'échelle mondiale.

L'Assemblée mondiale de normalisation des télécommunications (AMNT), qui se réunit tous les quatre ans, détermine les thèmes d'étude à traiter par les Commissions d'études de l'UIT-T, lesquelles élaborent en retour des Recommandations sur ces thèmes.

L'approbation des Recommandations par les Membres de l'UIT-T s'effectue selon la procédure définie dans la Résolution 1 de l'AMNT.

Dans certains secteurs des technologies de l'information qui correspondent à la sphère de compétence de l'UIT-T, les normes nécessaires se préparent en collaboration avec l'ISO et la CEI.

NOTE

Dans la présente Recommandation, l'expression "Administration" est utilisée pour désigner de façon abrégée aussi bien une administration de télécommunications qu'une exploitation reconnue.

DROITS DE PROPRIÉTÉ INTELLECTUELLE

L'UIT attire l'attention sur la possibilité que l'application ou la mise en œuvre de la présente Recommandation puisse donner lieu à l'utilisation d'un droit de propriété intellectuelle. L'UIT ne prend pas position en ce qui concerne l'existence, la validité ou l'applicabilité des droits de propriété intellectuelle, qu'ils soient revendiqués par un Membre de l'UIT ou par une tierce partie étrangère à la procédure d'élaboration des Recommandations.

A la date d'approbation de la présente Recommandation, l'UIT avait été avisée de l'existence d'une propriété intellectuelle protégée par des brevets à acquérir pour mettre en œuvre la présente Recommandation. Toutefois, comme il ne s'agit peut-être pas de renseignements les plus récents, il est vivement recommandé aux responsables de la mise en œuvre de consulter la base de données des brevets du TSB.

TABLE DES MATIÈRES

Recommandation UIT-T P.862

Evaluation de la qualité vocale perçue: méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite[1]

1 Introduction

La méthode objective décrite dans la présente Recommandation est appelée "Evaluation de la qualité vocale perçue" (PESQ, *perceptual evaluation of speech quality*). Elle constitue l'aboutissement de plusieurs années d'étude et s'applique non seulement aux codecs vocaux mais aussi aux mesures de bout en bout.

Dans la réalité, les systèmes peuvent être soumis à filtrage, présenter un temps de propagation variable et subir des distorsions dues à des erreurs de transmission dans la voie et à des codecs à faible débit binaire. La méthode de mesure de la qualité vocale perçue (PSQM, *perceptual speech quality measure*) décrite dans l'UIT-T P.861 (février 1998), dont l'utilisation n'a été recommandée que pour l'évaluation des codecs vocaux, n'a pas permis de tenir dûment compte du filtrage, du temps de propagation variable et des distorsions localisées de courte durée. La méthode d'évaluation PESQ prend en compte ces effets au moyen de l'égalisation de la fonction de transfert, de l'alignement temporel et d'un nouvel algorithme de calcul des valeurs moyennes de distorsion dans le temps. La validation de la méthode d'évaluation PESQ a été assurée par un certain nombre d'expériences qui ont permis d'en tester expressément la performance compte tenu d'un ensemble de facteurs tel que le filtrage, le temps de propagation variable, les distorsions due au codage et les erreurs de transmission dans la voie.

Il est recommandé d'utiliser la méthode PESQ pour l'évaluation de la qualité de la parole de la téléphonie à 3,1 kHz (bande étroite) avec combiné et des codecs vocaux à bande étroite.

2 Références normatives

La présente Recommandation se réfère à certaines dispositions des Recommandations UIT-T et textes suivants qui, de ce fait, en sont parties intégrantes. Les versions indiquées étaient en vigueur au moment de la publication de la présente Recommandation. Toute Recommandation ou tout texte étant sujet à révision, les utilisateurs de la présente Recommandation sont invités à se reporter, si possible, aux versions les plus récentes des références normatives suivantes. La liste des Recommandations de l'UIT-T en vigueur est régulièrement publiée.

– UIT-T P.800 (1996), *Méthodes d'évaluation subjective de la qualité de transmission*.

– UIT-T P.810 (1996), *Appareil de référence à bruit modulé (MNRU)*.

– UIT-T P.830 (1996), *Evaluation subjective de la qualité des codecs numériques à bande téléphonique et à large bande*.

– UIT-T Supplément 23 de la série P (1998), *Base de données de signaux vocaux codés de l'UIT-T*.

---

[1] La présente Recommandation comporte un document électronique contenant une implémentation de référence en langage C de l'algorithme PESQ et des données de test de conformité.

3 Abréviations

La présente Recommandation utilise les abréviations suivantes:

ACR évaluation par catégories absolues (*absolute category rating*)

CELP prédiction linéaire avec excitation par code (*code excited linear prediction*)

DMOS note moyenne d'appréciation de la dégradation (*degradation mean opinion score*)

HATS simulateur de tête et de torse (*head and torso simulator*)

IRS système de référence intermédiaire (*intermediate reference system*)

LQ qualité d'écoute (*listening quality*)

MIC modulation par impulsions et codage

MOS note moyenne d'opinion (*mean opinion score*)

PESQ évaluation de la qualité vocale perçue (*perceptual evaluation of speech quality*)

PSQM mesure de la qualité vocale perçue (*perceptual speech quality measure*)

4 Domaine d'application

Les Tableaux 1 à 3 donnent un aperçu général des facteurs expérimentaux, des techniques de codage et des applications retenus dans la présente Recommandation, établis d'après les résultats du test de performance effectué par la Commission d'études 12. Le Tableau 1 indique les relations entre les facteurs expérimentaux, les techniques de codage et les applications pour lesquels la présente Recommandation s'est révélée d'une précision acceptable. Le Tableau 2 répertorie les conditions pour lesquelles la méthode d'évaluation PESQ définie dans la présente Recommandation est réputée fournir des prédictions inexactes ou n'est pas censée être utilisée. Enfin, le Tableau 3 énumère les facteurs, techniques et applications pour lesquels la méthode d'évaluation PESQ n'a pas encore été validée à ce jour. Bien que les valeurs de corrélation entre les notes objectives et subjectives obtenues lors du test de performance s'établissaient au voisinage de 0,935 pour les données connues et pour les données inconnues, l'algorithme PESQ ne saurait être utilisé en lieu et place d'essais subjectifs.

Il convient également de noter que l'algorithme PESQ n'offre pas une évaluation complète de la qualité de transmission. Il ne mesure que les effets produits dans un seul sens, par la distorsion du son et le bruit affectant la qualité de la parole. Les effets de l'affaiblissement en sonie, du temps de propagation, de l'effet local, de l'écho et d'autres dégradations liées à l'interaction bidirectionnelle (par exemple, écrêteur de centre) ne sont pas pris en compte dans les notes calculées par l'algorithme PESQ. En conséquence, il est possible d'avoir des notes PESQ élevées malgré une qualité médiocre dans l'ensemble de la connexion.

Tableau 1/P.862 – Facteurs pour lesquels la méthode PESQ s'est révélée d'une précision acceptable

Facteurs expérimentaux
Niveaux d'entrée des signaux de parole dans un codec
Erreur de transmission dans la voie
Perte de paquets et masquage de perte de paquets avec les codecs CELP
Débits lorsque le codec peut fonctionner selon plusieurs modes
Transcodages
Bruit ambiant du côté émission (Voir Note.)

Tableau 1/P.862 – Facteurs pour lesquels la méthode PESQ s'est révélée d'une précision acceptable *(fin)*

Facteurs expérimentaux
Incidence de la variation du temps de propagation dans les essais d'écoute seulement
Prédistorsion à court terme du signal audio
Prédistorsion à long terme du signal audio
Techniques de codage
Codecs temporels (G.711, G.726 et G.727, par exemple)
Codecs CELP et hybrides ≥ 4 kbit/s (G.728, G.729 et G.723.1, par exemple)
Autres codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA
Applications
Evaluation du codec
Sélection du codec
Essais actifs dans le réseau par connexion numérique ou analogique au réseau
Essais de réseaux émulés et prototypes
NOTE – En présence de bruit ambiant, on peut mesurer la qualité en appliquant la méthode PESQ au signal original "propre" (c'est-à-dire exempt de bruit et au signal dégradé avec bruit.

Tableau 2/P.862 – L'évaluation PESQ est réputée fournir des prédictions inexactes lorsqu'elle est utilisée conjointement avec ces variables ou n'est pas censée être utilisée avec ces variables

Facteurs expérimentaux
Niveaux d'écoute (Voir Note.)
Affaiblissement en sonie
Incidence du temps de propagation dans les essais de conversation
Echo pour le locuteur
Effet vocal
Techniques de codage
Remplacement de fractions continues du signal vocal constituant plus de 25% de la conversation active par un silence (écrêtage temporel extrême)
Applications
Dispositifs de mesure sans intrusion en service
Qualité des communications bidirectionnelles
NOTE – L'algorithme PESQ suppose l'existence d'un niveau d'écoute normalisé de 79 dB SPL et compense les niveaux de signaux non optimaux dans les fichiers d'entrée. L'effet subjectif de l'écart par rapport au niveau d'écoute optimale n'est donc pas pris en compte.

Tableau 3/P.862 (Fera l'objet d'un complément d'étude) – Facteurs, techniques et applications pour lesquels la méthode d'évaluation PESQ n'a pas été encore validée à ce jour

Facteurs expérimentaux
Perte de paquets et masquage de perte de paquets avec des codecs de type MIC (Voir Note 1.)
Ecrêtage temporel du signal vocal (Voir Note 1.)
Ecrêtage en amplitude du signal vocal (Voir Note 2.)
Variations selon le locuteur
Locuteurs multiples simultanés
Discordance de débit entre un codeur et un décodeur si un codec possède plusieurs modes de débit
Signaux d'information de couche réseau à l'entrée d'un codec
Signaux vocaux artificiels à l'entrée d'un codec
Signaux de musique à l'entrée d'un codec
Echo pour la personne qui écoute
Effets/artefacts causés par le fonctionnement des annuleurs d'écho
Effets/artefacts causés par les algorithmes de réduction de bruit
Techniques de codage
Codecs CELP et hybrides < 4 kbit/s
HVXC MPEG4
Applications
Essais acoustiques des terminaux/combinés au moyen du simulateur HATS, par exemple
NOTE 1 – La méthode PESQ semble être plus sensible que les sujets à l'écrêtage temporel frontal, notamment dans le cas de mots manquants susceptibles de ne pas être perçus par les sujets. Inversement, la méthode PESQ peut être moins sensible que les sujets à l'écrêtage temporel normal de courte durée (remplacement de courtes fractions du signal vocal par un silence). Dans un cas comme dans l'autre, la corrélation entre la note d'évaluation PESQ et la note MOS subjective pourra s'en trouver limitée. NOTE 2 – Certains éléments de preuve semblent indiquer que l'algorithme PESQ est en mesure de rendre compte de l'écrêtage en amplitude, mais quatre conditions seulement sont censées avoir été incluses (dans deux expériences caractérisées par 50 conditions) dans la base de données de validation décrite au paragraphe 7.

5 Conventions

L'évaluation subjective des réseaux téléphoniques et des codecs vocaux peut être conduite au moyen d'essais d'écoute seulement ou au moyen de méthodes d'essais subjectifs par conversation. Pour des raisons pratiques, les essais d'écoute seulement sont la seule méthode possible d'essais subjectifs au cours de la mise au point de codecs vocaux, lorsqu'une mise en œuvre en temps réel de ces codecs n'est pas réalisable. La présente Recommandation expose une technique de mesure objective qui permet d'estimer la qualité subjective obtenue lors des essais d'écoute seulement, au moyen d'équipements d'écoute conformes aux caractéristiques de réception du système IRS ou du système IRS modifié.

La plupart des informations sur la performance de la méthode PESQ provient des expériences subjectives sur la qualité d'écoute (LQ, *listening quality*) pour la détermination de l'évaluation par catégories absolues (ACR, *absolute category rating*). Par conséquent, il faut considérer que la présente Recommandation se rapporte en premier lieu à l'échelle d'opinion de qualité d'écoute ACR.

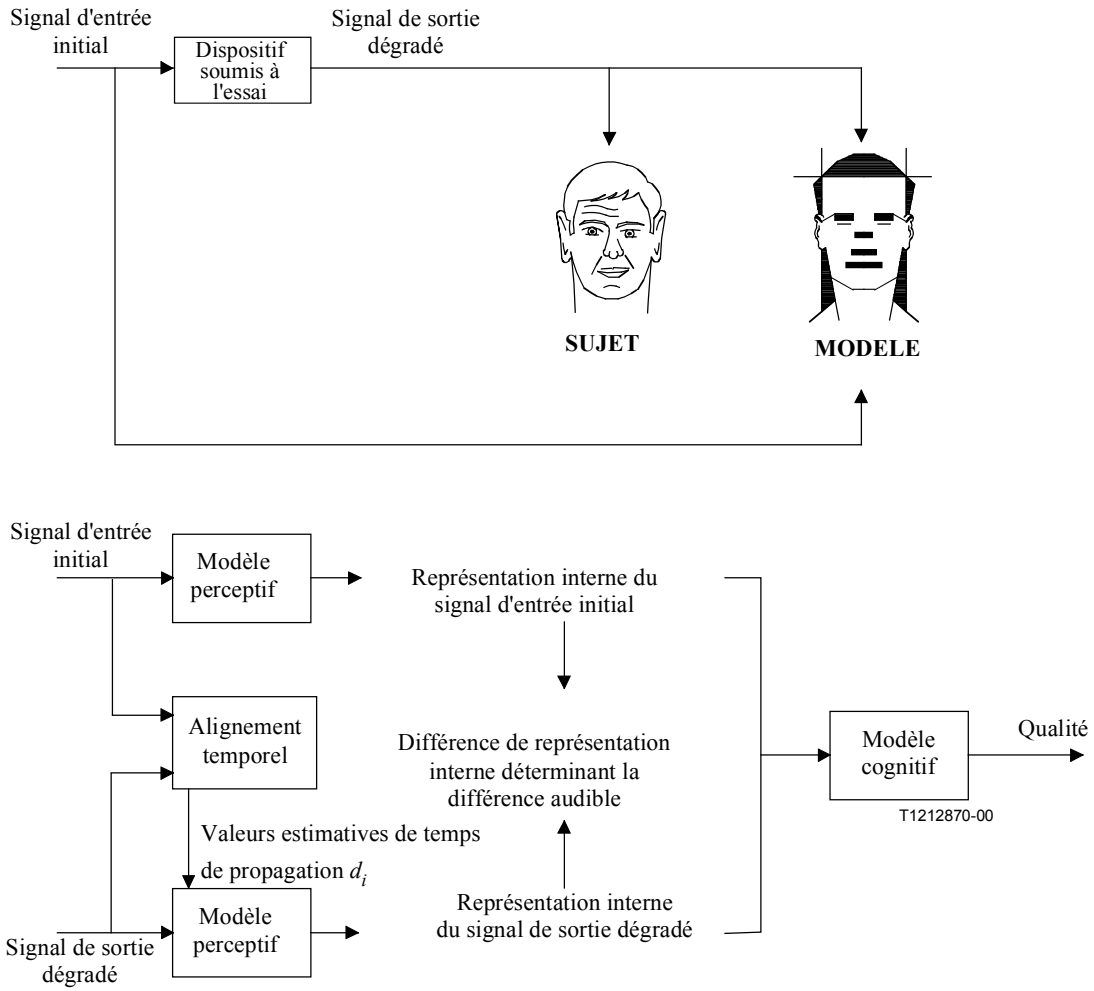
6 Aperçu général de l'évaluation PESQ

L'évaluation PESQ compare un signal initial X(t) à un signal dégradé Y(t) produit par le passage de X(t) à travers un système de communication. Elle permet de prévoir les notes de qualité de perception qu'attribueraient au signal Y(t) les sujets participant à un essai d'écoute subjectif.

Pendant la première phase de l'évaluation PESQ, on calcule une série de temps de propagation entre le signal d'entrée initial et le signal de sortie dégradé, un pour chaque intervalle de temps pour lequel le temps de propagation diffère sensiblement de l'intervalle de temps précédent. Pour chacun de ces intervalles, on calcule un point de départ et d'arrêt correspondant. L'algorithme d'alignement repose sur le principe d'un examen comparatif entre le niveau de confiance obtenu pour deux temps de propagation pendant un intervalle de temps donné et le niveau de confiance obtenu pour un seul temps de propagation pendant ce même intervalle. L'algorithme permet de gérer les variations du temps de propagation pendant les silences et les parties de conversation active.

D'après les divers temps de propagation mesurés, l'évaluation PESQ compare le signal (d'entrée) initial au signal de sortie dégradé et aligné du dispositif soumis à l'essai selon un modèle perceptif, tel que représenté sur la Figure 1. La clé de ce processus réside dans la transformation du signal initial et du signal dégradé en une représentation interne analogue à la représentation psychophysique des signaux audio dans un système auditif humain, compte tenu de la fréquence perçue (en Bark) et de la sonie (en sone). Cette transformation est opérée en plusieurs étapes: alignement temporel, alignement du niveau par rapport à un niveau d'écoute étalonné, correspondance temps-fréquence, prédistorsion des fréquences et échelonnement en sonie par compression.

La représentation interne est traitée pour tenir compte des effets tels que les variations locales de gain ou le filtrage linéaire qui – s'ils ne sont pas trop prononcés – peuvent n'être guère perceptibles. Pour ce faire, on limite l'importance de la compensation en la retardant par rapport à l'effet, ce qui permet de compenser de faibles différences en régime permanent entre le signal initial et le signal dégradé. Les effets plus prononcés, ou les variations rapides, ne sont que partiellement compensés si bien qu'un effet résiduel subsiste qui contribue à renforcer globalement la perturbation perçue. Cela permet d'utiliser un petit nombre d'indicateurs de qualité pour modéliser les différents effets subjectifs. Dans le cadre de l'évaluation PESQ, deux paramètres d'erreur sont calculés dans le modèle cognitif, lesquels sont combinés pour obtenir une note MOS objective concernant la qualité d'écoute. Les principes de base utilisés dans le cadre de l'évaluation PESQ sont définis dans les références bibliographiques [1], ... [5].



NOTE – Un modèle informatique du sujet, composé d'un modèle perceptif et d'un modèle cognitif, est utilisé pour comparer le signal de sortie du dispositif soumis à l'essai à son signal d'entrée, d'après les données d'alignement découlant des signaux temporels du module d'alignement temporel.

Figure 1/P.862 – Aperçu général des principes fondamentaux appliqués pour l'évaluation PESQ

7 Comparaison entre notes objectives et notes subjectives

Les votes subjectifs sont influencés par de nombreux facteurs tels que les préférences des différents sujets et le contexte (et les autres conditions) de l'expérience. Il faut donc prévoir un processus de régression avant de pouvoir procéder à une comparaison directe. La régression doit être monotone pour assurer la préservation des données; on l'utilise normalement pour faire correspondre la note PESQ objective avec la note subjective. Une mesure de qualité objective satisfaisante doit présenter une grande corrélation avec de nombreux essais subjectifs différents si la régression en question est effectuée séparément pour chacun d'entre eux; dans la pratique, s'agissant d'une évaluation PESQ, le mappage assuré par la régression est souvent presque linéaire, selon une échelle de type MOS.

Une des méthodes de régression préférée pour calculer la corrélation entre la note PESQ et la note MOS subjective, qui a été appliquée pour la validation de l'évaluation PESQ, utilise un polynôme de troisième ordre soumis à la contrainte d'être monotone. Ce calcul est effectué pour chaque étude. Dans la plupart des cas, la note MOS pour chaque condition est l'unité de mesure de performance choisie, en sorte que la régression doit être appliquée entre la note MOS pour chaque condition et les notes moyennes correspondantes d'évaluation PESQ. Une condition devrait utiliser

au moins quatre échantillons vocaux différents. La régression débouche sur une note MOS objective dans cet essai. Pour pouvoir comparer les notes objectives et les notes subjectives, les notes MOS subjectives devraient être déterminées à partir d'un essai d'écoute effectué selon l'UIT-T P.830.

7.1 Coefficient de corrélation

Le degré exact de correspondance entre les notes d'évaluation PESQ et les notes subjectives peut être mesuré en calculant le coefficient de corrélation. Pour ce faire, on se fonde en principe sur les notes moyennes pour les différentes conditions, après mise en correspondance des notes objectives et des notes subjectives. Le coefficient de corrélation est calculé selon la formule de Pearson:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

où x_i est la note MOS pour la condition *i*, et $\bar{x}$ est la moyenne des valeurs correspondantes de note MOS x_i. y_i est la note moyenne d'évaluation PESQ appliquée à la condition *i*, et $\bar{y}$ est la moyenne des valeurs prévues correspondantes de note MOS y_i.

Pour les 22 expériences courantes du test de performance de l'UIT, le coefficient de corrélation moyen s'est établi à 0,935. Pour une série agréée de huit expériences utilisées au stade de la validation finale – expériences qui étaient inconnues au stade de la mise au point de la méthode d'évaluation PESQ – le coefficient de corrélation moyen s'est également établi à 0,935.

7.2 Erreurs résiduelles

Le mappage qu'opère la régression élimine tout décalage systématique entre les notes objectives et la note MOS subjective, ce qui réduit la valeur quadratique moyenne des erreurs résiduelles.

$$e_i = x_i - y_i$$

Diverses mesures peuvent être appliquées aux erreurs résiduelles pour donner un autre aperçu de la correspondance étroite entre les notes objectives et la note MOS subjective. Ainsi, l'histogramme des erreurs résiduelles absolues $|e_i|$ donne un bref aperçu de la fréquence à laquelle se produisent des erreurs d'importance différentes

Pour les 22 expériences courantes du test de performance de l'UIT, la distribution moyenne des erreurs résiduelles a montré que l'erreur absolue était inférieure à une note MOS de 0,25 (±0,25 sur une échelle de 5 points) pour 69,2% des conditions et inférieure à une note MOS de 0,5. Pour une série agréée de sept expériences utilisées au stade de la validation finale, expériences qui étaient inconnues au stade de la mise au point de la méthode d'évaluation PESQ, l'erreur absolue était inférieure à une note MOS de 0,25 (±0,25 sur une échelle de 5 points) pour 72,3% des conditions et inférieure à une note MOS de 0,5 (±0,5 sur une échelle de 5 points) pour 91,1% des conditions.

8 Préparation du matériel vocal traité

Il importe que les signaux d'essai destinés à être utilisés pour l'évaluation PESQ soient représentatifs des signaux réels transportés par les réseaux de communication. Les réseaux peuvent traiter la parole et les silences de manière différente. De ce fait, étant souvent optimisés pour la parole, les algorithmes de codage peuvent donner des résultats dénués de sens si on leur applique des signaux d'essai ne contenant pas les propriétés temporelles et spectrales des signaux vocaux. Il est souvent nécessaire de poursuivre le prétraitement pour tenir compte du filtrage sur le trajet d'émission d'un combiné, et pour veiller à ce que les niveaux de puissance soient fixés dans une fourchette appropriée.

8.1 Matériel source

8.1.1 Choix du matériel source

A l'heure actuelle, tous les résultats officiels de performance en matière d'évaluation PESQ se rapportent aux expériences menées en utilisant les mêmes enregistrements de parole naturelle pour l'essai subjectif et l'essai objectif. L'utilisation de signaux vocaux artificiels et de signaux d'essai vocaux réels concaténés n'est recommandée que si ces signaux sont représentatifs de la structure temporelle (intervalles de silence compris) et de la structure phonétique des signaux vocaux réels.

La préparation des signaux d'essai vocaux artificiels peut être assurée de plusieurs manières. On peut construire un signal d'essai vocal réel concaténé en concaténant de courts fragments (d'une seconde, par exemple) de parole réelle tout en conservant une structure représentative de la parole et des silences. On peut aussi adopter une approche phonétique pour produire un signal vocal artificiel très peu redondant qui soit représentatif de la structure tant temporelle que phonétique d'un large éventail de signaux vocaux naturels [6]. Les signaux d'essai doivent être représentatifs des locuteurs et des locutrices. Au cours des essais préliminaires, les signaux vocaux artificiels de haute qualité et les signaux vocaux réels concaténés ont donné de bons résultats dans le cas de l'évaluation PESQ. Au cours de ces essais, les notes objectives pour les signaux d'essai dans chaque condition ont été utilisées pour prévoir les valeurs de note MOS des conditions subjectives. Cette façon de procéder constitue la méthode la plus simple pour déterminer la qualité du système soumis à l'essai. Ce point appelle un complément d'étude.

En cas d'utilisation d'enregistrements de parole naturelle, il convient de suivre les indications données dans le paragraphe 7/P.830; de plus, il est recommandé qu'un minimum de deux locuteurs et de deux locutrices soit mis à contribution pour chaque condition d'essai. Si la variation selon le locuteur doit être évaluée en tant que facteur autonome, il est recommandé de faire appel à un plus grand nombre de locuteurs comme suit: 8 hommes, 8 femmes et 8 enfants. A l'heure actuelle, l'UIT-T P.862 n'est pas validée pour la variation selon le locuteur.

8.1.2 Structure temporelle et durée du matériel source

Pour être représentatifs des pauses naturelles dans la conversation, les signaux d'essai doivent comporter des salves de parole séparées par des périodes de silence. A titre indicatif, la durée d'une salve de parole est généralement de une à trois secondes, encore que cette durée varie considérablement d'une langue à l'autre. Certains types de détecteur d'activité vocale ne sont sensibles qu'aux périodes de silence de plus de 200 ms. La parole devrait être active pendant 40 à 80% du temps, bien que ce pourcentage, là encore, dépende quelque peu de la langue.

La plupart des expériences retenues pour l'étalonnage et la validation de l'évaluation PESQ contenaient des paires de phrases séparées par des silences, d'une durée totale de huit secondes; dans certains cas, trois ou quatre phrases ont été utilisées, avec des enregistrements légèrement plus longs (jusqu'à 12 secondes). Les enregistrements destinés à être utilisés pour l'évaluation PESQ devraient être sensiblement de même longueur et de même structure.

Ainsi, dans le cas où une condition doit être soumise à un essai sur une longue période, il est plus indiqué d'effectuer plusieurs enregistrements distincts de huit à vingt secondes de conversation environ et de traiter chaque fichier séparément aux fins de l'évaluation PESQ. Cette solution présente d'autres avantages: si on utilise dans chaque cas le même enregistrement original, les variations de la qualité de la condition en fonction du temps seront très apparentes; on peut aussi utiliser plusieurs locuteurs et/ou enregistrements sources différents, ce qui permet de mesurer de manière plus précise les variations selon le locuteur ou le matériel dans cette condition.

A noter qu'il découle du processus d'intégration non linéaire dans le cadre de l'évaluation PESQ que la note moyenne pour un ensemble de fichiers ne sera généralement pas égale à la note d'une version concaténée du même ensemble de fichiers, considérée isolément.

8.1.3 Filtrage et étalonnage de niveau

Les signaux doivent traverser un filtre dont les caractéristiques de fréquence permettent de simuler les caractéristiques de fréquence d'émission d'un combiné téléphonique, et doivent être soumis à la même forme d'égalisation de niveau que des voix réelles. L'UIT-T recommande l'emploi de la caractéristique de fréquence à l'émission du système de référence intermédiaire (IRS, *intermediate reference system*) modifié, tel qu'il est défini dans l'Annexe D/P.830. L'égalisation du niveau à une amplitude qui soit représentative du trafic réel doit être effectuée conformément au 7.2.2 de l'UIT-T P.830.

Dans certains cas, le système de mesure utilisé (une interface analogique à 2 fils, par exemple) peut être à l'origine de variations de niveau importantes. Il convient de tenir compte de ces variations pour veiller à ce que le niveau du signal traversant le réseau soit représentatif.

Après filtrage (à l'émission) dans le combiné et égalisation du niveau, le matériel source préparé est normalement utilisé comme signal initial pour l'évaluation PESQ.

8.2 Adjonction de bruit de fond

Il est possible d'utiliser l'évaluation PESQ pour évaluer la qualité de systèmes acheminant des signaux vocaux en présence de bruit de fond ou de bruit ambiant (véhicule, rue, etc.). Les enregistrements de bruit doivent être effectués à travers un filtre approprié reprenant la caractéristique d'émission du système IRS modifié – c'est là une condition particulièrement importante pour les signaux basse fréquence tels que les bruits à l'intérieur d'un véhicule qui sont fortement atténués par le filtre du combiné – puis réglés au niveau souhaité pour l'essai. Pour que l'évaluation PESQ puisse tenir compte de la perturbation subjective dans un contexte ACR, en raison du bruit ainsi que des éventuelles distorsions due au codage, le signal initial utilisé pour évaluation PESQ doit être "propre" (c'est-à-dire exempt de bruit), mais le bruit doit être ajouté avant que les signaux ne traversent le système soumis à l'essai. L'évaluation PESQ est validée pour les signaux d'entrée avec présence de bruit. Le processus d'adjonction de bruit est représenté sur la Figure 2 b).

a) Essai de conversation en l'absence de bruit ambiant

b) Essai de conversation en présence de bruit ambiant

Figure 2/P.862 – Méthodes de mesure de la qualité en présence et en l'absence de bruit ambiant

8.3 Traitement dans le système soumis à l'essai

Le signal source doit être dûment traité dans le système soumis à l'essai. Il est souhaitable d'éviter toute nouvelle distorsion imputable à des opérations inutiles: quantification, écrêtage en amplitude ou rééchantillonnage. Le format préféré pour l'enregistrement du signal initial et du signal dégradé est le codage MIC linéaire à 16 bits et fréquence d'échantillonnage de 8 kHz. L'évaluation PESQ a été validée pour les fréquences d'échantillonnage de 8 kHz et de 16 kHz.

9 Choix des paramètres expérimentaux

Les effets de divers facteurs de qualité sur la performance du codec ou du système peuvent être examinés par une évaluation subjective ou objective. L'UIT-T P.830 donne des indications sur les méthodes d'évaluation subjective des facteurs de qualité suivants:

1) niveaux d'entrée des signaux de parole dans un codec;
2) niveaux d'écoute lors d'expériences subjectives;
3) locuteurs (y compris locuteurs multiples simultanés);
4) erreur de transmission dans la voie entre un codeur et un décodeur;
5) débits lorsque le codec peut fonctionner selon plusieurs modes;
6) transcodages;
7) discordance de débit entre un codeur et un décodeur si un codec possède plusieurs modes de débit;
8) bruit ambiant du côté émission;
9) signaux d'information de couche Réseau à l'entrée d'un codec;
10) signaux de musique à l'entrée d'un codec.

L'évaluation PESQ permet d'évaluer nombre de ces facteurs de qualité (1, 4, 5, 6 et 8).

NOTE 1 – La mesure objective de facteurs de qualité autres que ceux qui sont expressément indiqués dans la présente Recommandation comme étant applicables est encore à l'étude. Ces facteurs ne devront donc être mesurés qu'après vérification de l'exactitude de la mesure objective en liaison avec les essais subjectifs selon l'UIT-T P.830.

Outre les conditions requises pour les codecs, l'UIT-T P.830 préconise l'utilisation de conditions de référence pour les essais subjectifs. Ces conditions sont utiles pour faciliter les comparaisons entre les résultats d'essais subjectifs issus de laboratoires différents ou du même laboratoire à des moments différents. En outre, dans le cas où les résultats des essais objectifs sont exprimés en termes de valeurs Q équivalentes, il convient de soumettre les conditions de référence à des essais en utilisant l'appareil de référence à bruit modulé (MNRU, *modulated noise reference unit*) spécifié dans l'UIT-T P.810.

NOTE 2 – Etendre la mesure objective de la qualité à d'autres codecs normalisés tels que les codecs MIC à 64 kbit/s G.711, MICDA à 32 kbit/s G.726, LD-CELP à 16 kbit/s G.728 et CS-ACELP à 8 kbit/s G.729, ainsi qu'à l'appareil MNRU peut aider à comparer la performance du système soumis à l'essai à celle des codecs normalisés.

Ces paramètres expérimentaux sont expliqués en détail dans l'UIT-T P.830.

10 Description de l'algorithme PESQ

Un grand nombre des phases de l'évaluation PESQ étant relativement difficiles à décrire sous forme d'algorithme, une telle description ne saurait commodément être exprimée par des formules mathématiques. Cette description étant par définition textuelle, le lecteur est invité à se reporter au code source C pour de plus amples précisions. Les Figures 3, 4a et 4b donnent un aperçu général de l'algorithme sous forme d'organigramme. La Figure 3 concerne l'alignement, la Figure 4a l'élément central du modèle perceptif et la Figure 4b la détermination finale de la note d'évaluation PESQ. Chacun des blocs fait l'objet d'une description approfondie.

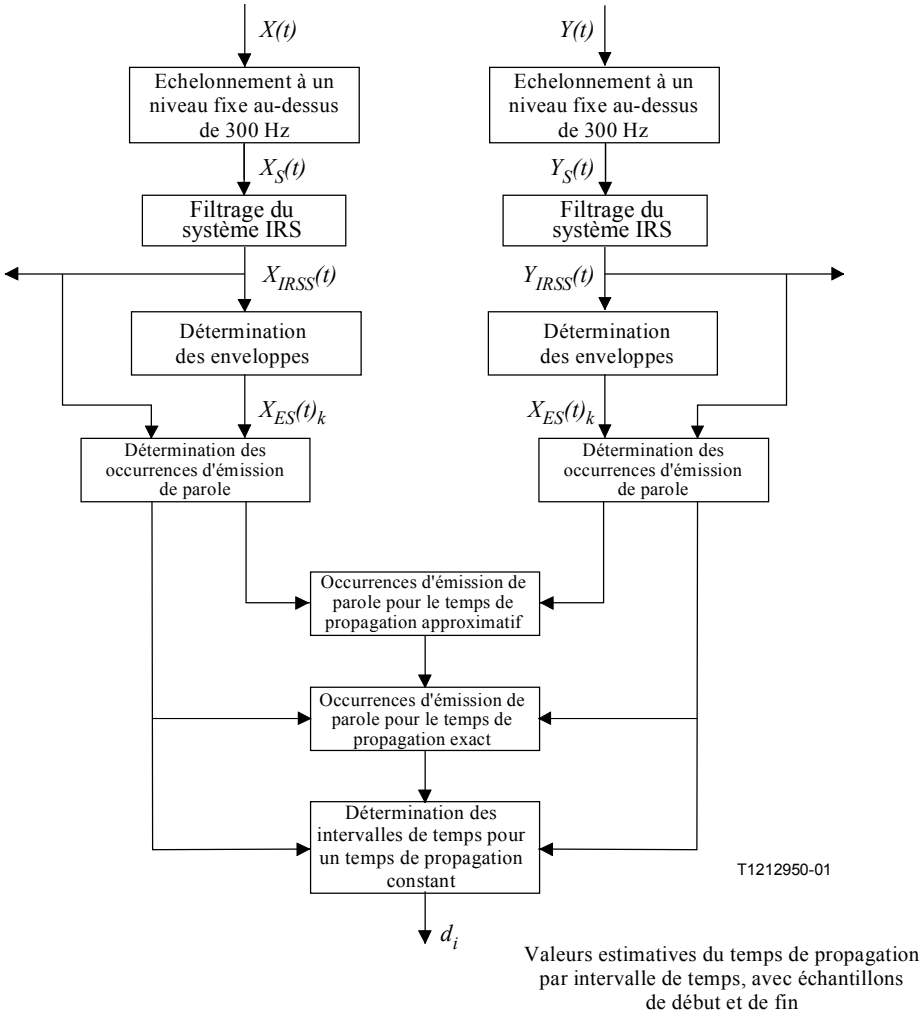
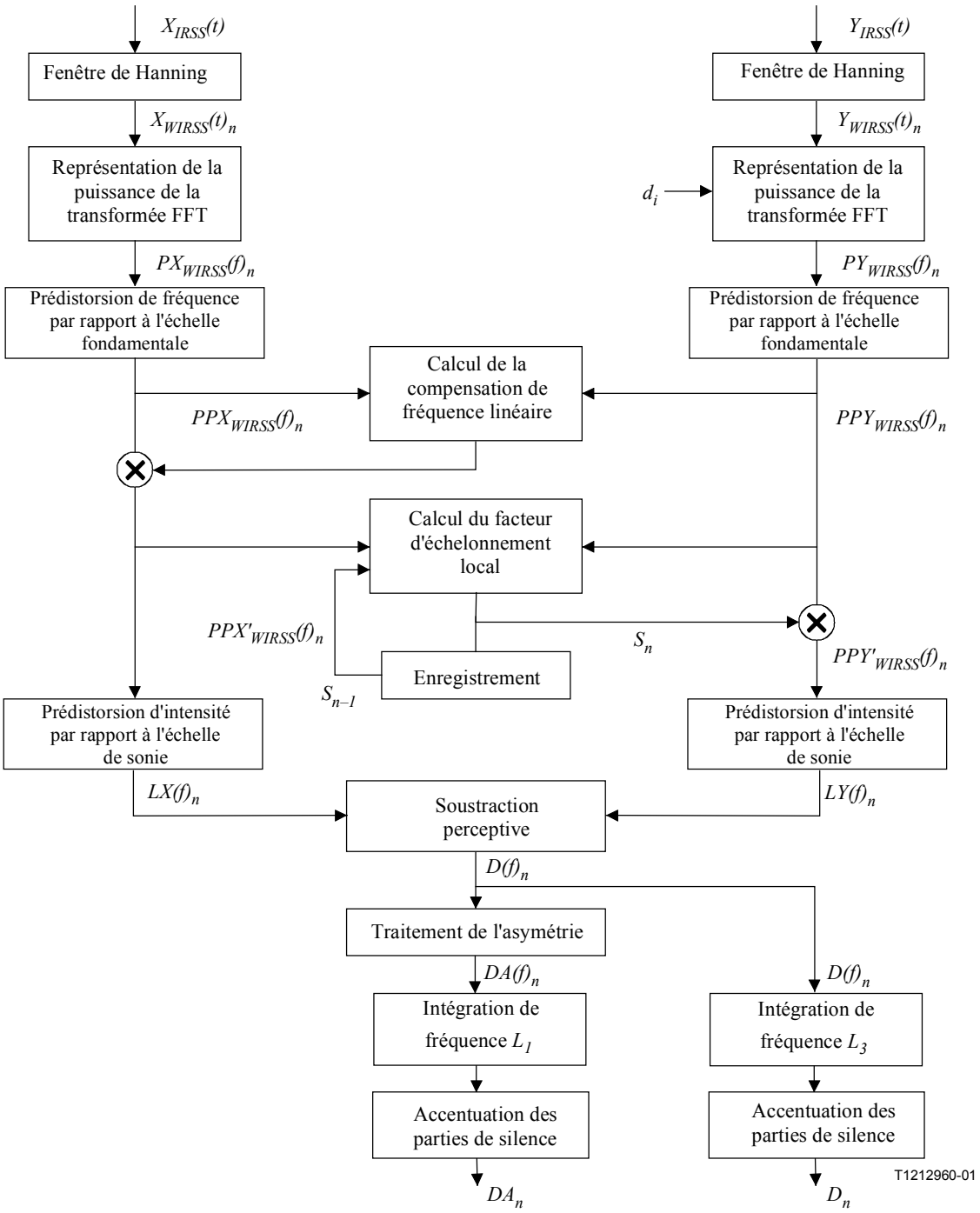
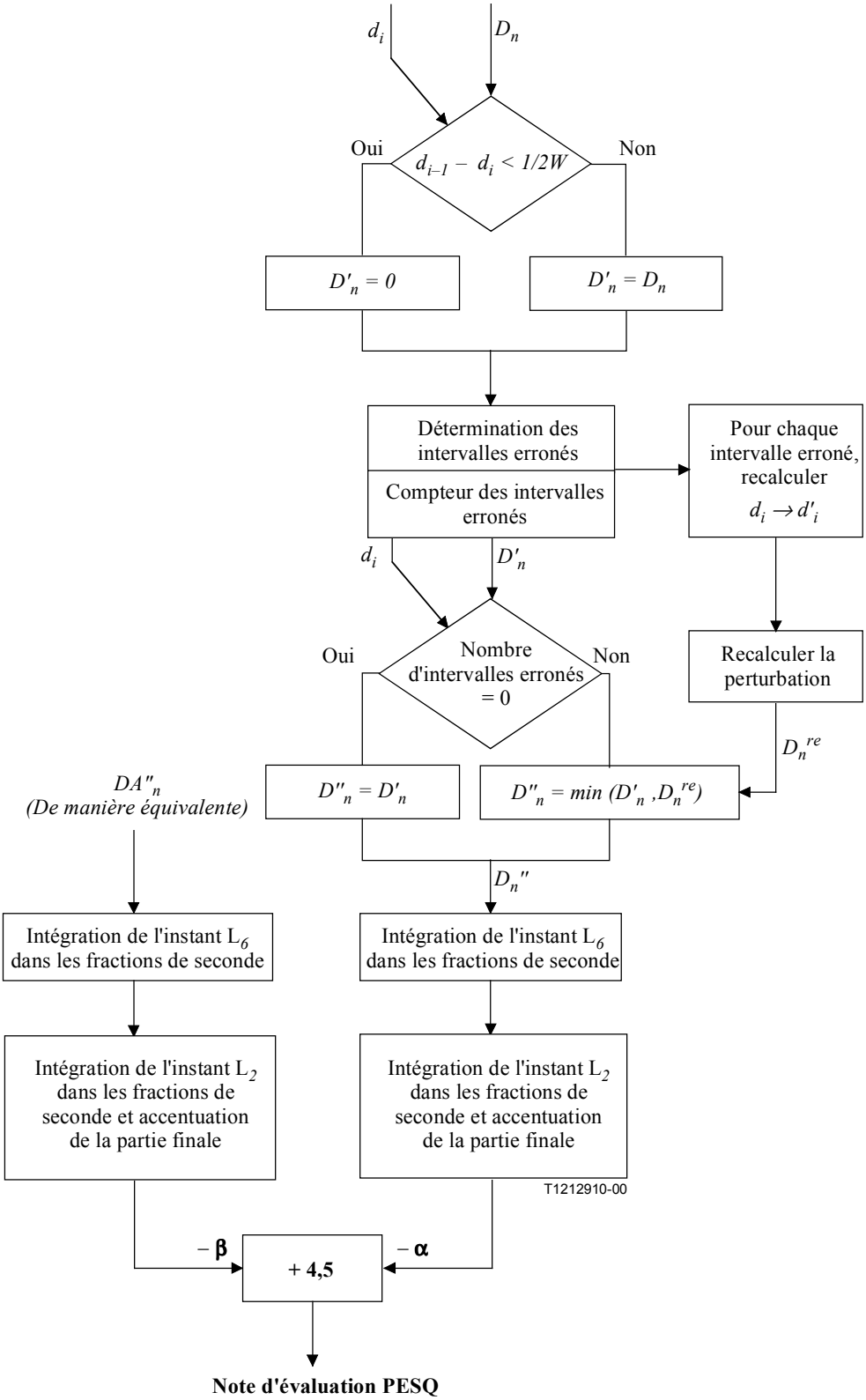


Figure 3/P.862 – Aperçu général de la procédure périodique d'alignement utilisée dans le cadre de l'évaluation PESQ pour déterminer le temps de propagation par intervalle de temps d_i



NOTE – Les distorsions par trame D_n et DA_n ont été cumulées dans le temps (indice$_n$) pour obtenir les perturbations finales (voir Figure 4b qui indique également le réalignement du signal dégradé).

Figure 4a/P.862 – Aperçu général du modèle perceptif



NOTE – Après réalignement des intervalles erronés, les distorsions par trame D''_n et DA''_n sont intégrées par rapport au temps et mappées à la note d'évaluation PESQ. W est la longueur de la fenêtre de la transformée FTT dans les échantillons.

Figure 4b/P.862 – Aperçu général du modèle perceptif

10.1 Prétraitement d'alignement de niveau et d'alignement temporel (Figure 3)

10.1.1 Calcul du gain total du système

Le gain du système soumis à l'essai n'est pas connu a priori et peut varier considérablement selon, par exemple, qu'on a utilisé pour la mesure une connexion RNIS ou une interface analogique à deux fils. En outre, il n'existe pas de niveau d'étalonnage unique auquel le signal initial sera enregistré. Il faut donc aligner le niveau du signal initial *X(t)* et le niveau du signal dégradé *Y(t)* sur le même niveau de puissance constant. L'évaluation PESQ suppose que le niveau d'écoute subjectif soit fixé à une valeur constante de 79 dB SPL au point de référence oreille (voir 8.1.2/P.830). L'algorithme d'alignement de niveau dans le cadre de l'évaluation PESQ fonctionne comme suit:

- les versions filtrées du signal initial et du signal dégradé sont calculées. Le filtre bloque tous les éléments au-dessous de 250 Hz, reste stable jusqu'à 2000 Hz et présente une réponse linéaire qui diminue élément par élément au passage par les points suivants: {2000 Hz, 0 dB}, {2500 Hz, –5 dB }, {3000 Hz, –10 dB}, {3150 Hz, –20 dB}, {3500 Hz, –50 dB}, {4000 Hz et au-dessus, –500 dB}. Ces versions filtrées des signaux sont utilisées uniquement pour l'opération considérée ici, à savoir le calcul du gain total du système;
- on calcule la valeur moyenne des échantillons vocaux du signal initial filtré et des échantillons vocaux du signal dégradé filtré élevés au carré;
- différents gains sont calculés et appliqués pour aligner le signal vocal initial *X(t)* et le signal vocal dégradé *Y(t)* sur un niveau cible constant découlant des versions échelonnées $X_S(t)$ et $Y_S(t)$ de ces signaux.

10.1.2 Filtrage du système IRS

On part du principe que les essais d'écoute ont été effectués en utilisant une caractéristique de réception du système IRS ou du système IRS modifié dans le combiné. Un modèle perceptif d'évaluation de la qualité de la parole par des sujets doit tenir compte de ce principe pour modéliser les signaux que les sujets ont effectivement entendus. On calcule donc les versions filtrées à la réception de type IRS du signal vocal initial et du signal vocal dégradé.

Dans une évaluation PESQ, ce filtrage est mis en œuvre au moyen d'une transformée FFT sur toute la longueur du fichier, dans le domaine fréquentiel, avec une réponse linéaire élément par élément analogue à la caractéristique de réception du système IRS (non modifié) (UIT-T P.830), puis au moyen d'une transformée FFT inverse sur toute la longueur du fichier de signaux vocaux. On obtient ainsi les versions filtrées $X_{IRSS}(t)$ et $Y_{IRSS}(t)$ des signaux d'entrée et de sortie échelonnés $X_S(t)$ et $Y_S(t)$. On utilise un seul et même filtre de réception de type IRS pour l'évaluation PESQ, indépendamment du type de filtrage (IRS ou IRS modifié) utilisé pendant l'expérience subjective réelle. Cette manière de procéder s'explique par le fait que dans la plupart des cas on ignore le type exact de filtrage utilisé et que, même quand on le connaît, le couplage du combiné à l'oreille reste inconnu. Il est donc impératif que la méthode objective soit relativement insensible au type de filtrage utilisé dans le combiné.

Les signaux filtrés par le système IRS sont utilisés à la fois dans la procédure d'alignement temporel et dans le modèle perceptif.

10.1.3 Alignement temporel

La procédure périodique d'alignement temporel fixe des valeurs de temps de propagation applicables au modèle perceptif pour permettre de comparer les parties correspondantes des fichiers du signal initial et du signal dégradé. Ce processus d'alignement se déroule en plusieurs phases:

- estimation du temps de propagation de groupe (enveloppes) d'après l'ensemble signal initial et signal dégradé;

- subdivision du signal initial en un certain nombre de sous-éléments appelés occurrences d'émission de parole (ou paroles émises);
- estimation du temps de propagation de groupe (enveloppes) d'après les occurrences d'émission de parole;
- identification, par corrélation fine ou par histogramme, du temps de propagation par rapport à l'échantillon le plus proche d'après les occurrences d'émission de parole;
- coupure des paroles émises et réalignement des intervalles de temps en vue de repérer les variations du temps de propagation en cours de conversation;
- d'après le modèle perceptif, identification et réalignement de longs passages comportant des erreurs importantes en vue de repérer les erreurs d'alignement.

10.1.3.1 Alignement selon les enveloppes

Les enveloppes $X_{ES}(t)_k$ et $Y_{ES}(t)_k$ sont calculées d'après les signaux initial et dégradé échelonnés $X_S(t)$ et $Y_S(t)$. L'enveloppe est définie par la formule LOG (MAX($E(k)/Ethresh, 1$)), où $E(k)$ est l'énergie dans une trame k de 4 ms et *Ethresh* est le seuil de conversation déterminé par un détecteur d'activité vocale. La corrélation croisée des enveloppes du signal initial et du signal dégradé est utilisée pour estimer le temps de propagation approximatif entre ces signaux, avec une résolution d'environ 4 ms.

10.1.3.2 Alignement temporel fin

Les modèles perceptifs étant sensibles aux décalages temporels, il est nécessaire de calculer une valeur de temps de propagation exacte par échantillon. Cette valeur se calcule comme suit:

- les trames de 64 ms (se chevauchant à 75%) sont soumises à un fenêtrage de Hanning et à une corrélation croisée entre le signal initial et le signal dégradé, après alignement selon les enveloppes;
- le maximum de la corrélation, à la puissance 0,125, est utilisé pour mesurer le niveau de confiance de l'alignement dans chaque trame. L'indice du maximum indique la valeur estimative du temps de propagation pour chaque trame;
- un histogramme de ces valeurs estimatives du temps de propagation, pondérées par le niveau de confiance mesuré, est calculé. L'histogramme est ensuite lissé par convolution avec un noyau triangulaire symétrique d'une largeur de 1 ms;
- l'indice du maximum dans l'histogramme, conjugué à la valeur estimative du temps de propagation précédent, indique la valeur estimative du temps de propagation final;
- le maximum de l'histogramme, divisé par la somme de l'histogramme avant convolution avec le noyau, indique un niveau de confiance compris entre 0 (aucune confiance) et 1 (confiance maximale).

L'alignement temporel fin permet d'obtenir une valeur de temps de propagation et un niveau de confiance correspondant pour chaque occurrence d'émission de parole, compte tenu des variations du temps de propagation pendant les périodes de silence. Connaissant les points où commence et où prend fin chaque occurrence d'émission de parole, l'alignement temporel fin permet d'identifier le temps de propagation de chaque trame dans le modèle perceptif.

10.1.3.3 Coupure des occurrences d'émission de parole

On mesure les variations du temps de propagation pendant la conversation en coupant et en réalignant les intervalles de temps de chaque occurrence d'émission de parole. On procède à l'alignement selon le temps de propagation de groupe (enveloppes) pour calculer la valeur estimative du temps de propagation de chaque élément, puis on procède à un alignement temporel fin pour déterminer le temps de propagation et le niveau de confiance de chaque élément. On procède à de nouvelles coupures en différents points de chaque occurrence d'émission de parole et on identifie la coupure qui produit le niveau de confiance le plus élevé. Si l'on obtient ainsi un niveau de confiance

plus élevé que par l'alignement sans coupures, et que les deux parties ont des temps de propagation sensiblement différents, l'occurrence d'émission de parole est divisée en conséquence. L'essai est appliqué de façon récursive à chaque partie, après coupure, pour mesurer de nouvelles variations du temps de propagation.

Cette manière de procéder permet de tenir compte des variations du temps de propagation tant pendant la conversation que pendant les silences et de calculer le temps de propagation par intervalle de temps (d_i) ainsi que les échantillons adaptés de début et de fin. Le nombre d'intervalles de temps est déterminé par le nombre de variations du temps de propagation.

10.1.3.4 Réalignement perceptif

Après que le modèle perceptif a été appliqué, on identifie les sections qui présentent une perturbation très importante (supérieure à une valeur de seuil donnée) et on les réaligne par corrélation croisée. Cette opération améliore la précision du modèle pour un petit nombre de fichiers difficiles à aligner dans lesquels le processus d'alignement temporel précédent n'a pas permis d'identifier correctement les variations du temps de propagation. Les modalités d'implémentation de cette procédure de réalignement sont décrites au 10.2.13.

10.2 Modèle perceptif (Figures 4a et 4b)

Le modèle perceptif d'évaluation PESQ est utilisé pour calculer une distance entre le signal vocal initial et le signal vocal dégradé (note d'évaluation PESQ). Comme indiqué au paragraphe 7, cette opération doit passer par une fonction monotone pour obtenir une prévision de la note MOS subjective pour un essai subjectif donné. La note d'évaluation PESQ est appliquée à une échelle de type MOS, sous la forme d'un numéro unique compris entre –0,5 et 4,5, bien que dans la plupart des cas le signal de sortie soit compris entre 1,0 et 4,5, ce qui correspond à l'amplitude normale de valeurs MOS observées dans une expérience de qualité d'écoute ACR.

10.2.1 Précalcul de réglages de constantes

Certaines valeurs et fonctions de constantes sont précalculées. Pour celles d'entre elles qui dépendent de la fréquence d'échantillonnage, les versions correspondant aux fréquences d'échantillonnage de 8 kHz et de 16 kHz sont enregistrées dans le programme.

10.2.1.1 Dépendance de la taille de la fenêtre FFT à l'égard de la fréquence d'échantillonnage (8 ou 16 kHz)

Au cours de l'évaluation PESQ, on applique les signaux temporels au domaine des fréquences temporelles en utilisant une transformée rapide de Fourier (FFT) avec une fenêtre Hanning d'une taille de 32 ms. Cela représente, pour la fréquence de 8 kHz, 256 échantillons par fenêtre et, pour la fréquence de 16 kHz, 512 échantillons par fenêtre, avec un taux de chevauchement des trames adjacentes de 50%.

10.2.1.2 Seuil d'audibilité absolu

On procède à l'interpolation du seuil d'audibilité absolu $P_0(f)$ pour obtenir les valeurs au centre des bandes de Bark qui sont utilisées. Ces valeurs sont mises en mémoire et utilisées dans la formule de sonie de Zwicker.

10.2.1.3 Facteur d'échelonnement de la puissance

La transformée FFT produit une constante de gain arbitraire aux fins de l'analyse temps-fréquence. Cette constante est calculée pour une onde sinusoïdale d'une fréquence de 1000 Hz avec une amplitude de 29,54 (40 dB SPL) transposée dans le domaine fréquentiel au moyen d'une transformée FFT fenêtrée sur 32 ms. On assure ensuite la conversion de l'axe des fréquences (discrètes) à une échelle de Bark modifiée par segmentation des bandes FFT. L'amplitude de crête des fréquences spectrales segmentées pour la conversion à l'échelle des fréquences de Bark (appelée la "densité de

puissance fondamentale") peut être de 10 000 (40 dB SPL). Cette valeur est obtenue par une post-multiplication avec une constante, le facteur d'échelonnement de puissance S_p.

10.2.1.4 Facteur d'échelonnement en sonie

La même tonalité de référence de 40 dB SPL est utilisée pour étalonner l'échelle de sonie psycho-acoustique (en sones). Après segmentation des fréquences pour la conversion à l'échelle de Bark modifiée, on soumet l'axe de l'intensité à une prédistorsion pour le convertir à l'échelle de sonie selon la loi de Zwicker, d'après le seuil d'audibilité absolu. La valeur entière de la densité de sonie sur l'échelle de fréquences de Bark, en cas d'utilisation d'une tonalité d'étalonnage de 1000 Hz et de 40 dB SPL, doit être de 1 sone. Cette valeur est obtenue par une post-multiplication avec une constante, le facteur d'échelonnement de la sonie S_l.

10.2.2 Filtrage à la réception du système IRS

Comme indiqué au 10.1.2, on part du principe que les essais d'écoute ont été effectués en utilisant une caractéristique de réception du système IRS ou du système IRS modifié dans le combiné. Le filtrage qu'il est nécessaire d'appliquer aux signaux vocaux est déjà effectué lors du prétraitement.

10.2.3 Calcul de l'intervalle de temps de parole active

Si le fichier des signaux vocaux initial et dégradé commence ou prend fin par des intervalles de silence importants, cela peut avoir une incidence sur le calcul de certaines valeurs de distorsion moyenne dans les fichiers. On procède donc à une estimation des périodes de silence au début et à la fin de ces fichiers. La somme des valeurs de cinq échantillons absolus successifs entre le début et la fin du fichier du signal vocal initial doit être supérieure à 500 pour que cette position soit considérée comme le début ou la fin de l'intervalle actif. L'intervalle entre ces deux instants est défini comme l'intervalle de temps de parole active. Pour réduire les cycles de calcul ou l'encombrement en mémoire, on peut limiter certains calculs à l'intervalle actif.

10.2.4 Transformée rapide de Fourier à court terme

L'oreille humaine opère une transformation temps-fréquence. Dans une évaluation PESQ, cette transformation est mise en œuvre au moyen d'une transformée FFT à court terme fenêtrée à 32 ms. Le chevauchement entre les fenêtres (trames) temporelles successives est de 50%. Les spectres de puissance – la somme des parties réelle et imaginaire élevées au carré des éléments FFT complexes – sont enregistrés séparément dans chacun des ensembles de valeurs réelles pour le signal initial et pour le signal dégradé. L'information de phase figurant dans une fenêtre de Hanning unique est mise au rebut pendant l'évaluation PESQ et tous les calculs sont fondés sur les seules représentations des puissances $PX_{WIRSS}(f)_n$ et $PY_{WIRSS}(f)_n$.

Les points de départ des fenêtres dans le signal dégradé sont décalés par rapport au temps de propagation. L'axe des temps du signal vocal initial est laissé tel quel. Si le temps de propagation augmente, certains éléments du signal dégradé échappent à tout traitement; si le temps de propagation diminue, certains éléments sont traités deux fois.

10.2.5 Calcul des densités de puissance fondamentale

Comme il ressort de l'échelle de Bark, le système auditif humain a une résolution plus fine aux basses fréquences qu'aux fréquences élevées. En application de ce principe, on segmente les bandes FFT et on additionne les puissances correspondantes de ces bandes en normalisant les éléments sommés. La fonction de prédistorsion qui établit une correspondance entre l'échelle des fréquences en Hertz et l'échelle fondamentale en Bark ne suit pas exactement les valeurs indiquées dans la littérature. Les signaux produits par cette fonction sont connus comme étant les densités de puissance fondamentale $PPX_{WIRSS}(f)_n$ et $PPY_{WIRSS}(f)_n$.

10.2.6 Compensation partielle de la densité de puissance fondamentale initiale pour l'égalisation de la fonction de transfert

Pour les besoins du filtrage dans le système soumis à l'essai, on calcule la valeur moyenne dans le temps du spectre de puissance des densités de puissance fondamentale du signal initial et du signal dégradé. Cette valeur moyenne est déterminée pour les trames de parole active n'utilisant que des cellules temps-fréquence dont la puissance est supérieure de plus de 1 000 fois au seuil d'audibilité absolu. Pour chaque segment de Bark modifié, on calcule un facteur de compensation partielle d'après le rapport du spectre du signal dégradé au spectre du signal initial. La compensation maximale n'est jamais supérieure à 20 dB. On multiplie ensuite la densité de puissance fondamentale initiale $PPX_{WIRSS}(f)_n$ de chaque trame n par ce facteur de compensation partielle pour égaliser le signal initial et le signal dégradé. On obtient ainsi la densité de puissance fondamentale initiale $PPY'_{WIRSS}(f)_n$ filtrée en sens inverse.

On recourt à cette compensation partielle du fait qu'un filtrage rigoureux peut être gênant pour la personne qui écoute. La compensation porte sur le signal initial du fait que le signal dégradé est soumis à l'appréciation des sujets dans une expérience ACR.

10.2.7 Compensation partielle de la densité de puissance fondamentale soumise à distorsion pour les variations de gain dans le temps entre le signal dégradé et le signal initial

Les variations de gain de courte durée sont partiellement compensées par le traitement des densités de puissance fondamentale trame par trame. Pour les densités de puissance fondamentale du signal initial et du signal dégradé, on calcule dans chaque trame n la somme de toutes les valeurs supérieures au seuil d'audibilité absolu. On calcule le rapport de la puissance dans les fichiers du signal initial et du signal dégradé, dans les limites de $[3 \cdot 10^{-4}, 5]$. Un filtre passe-bas de premier ordre (sur l'axe des temps) est appliqué à ce rapport. On multiplie ensuite par ce rapport, dans chaque trame n, la densité de puissance fondamentale soumise à distorsion, ce qui permet d'obtenir la densité de puissance fondamentale soumise à distorsion avec compensation partielle de gain $PPY'_{WIRSS}(f)_n$.

10.2.8 Calcul des densités de sonie

A l'issue de la compensation partielle pour le filtrage et les variations de gain de courte durée, on transforme les densités de puissance fondamentale du signal initial et du signal dégradé en une échelle de sonie (en sones) en appliquant la loi de Zwicker [7].

$$LX(f)_n = S_l \cdot \left(\frac{P_0(f)}{0{,}5}\right)^{\gamma} \cdot \left[\left(0{,}5 + 0{,}5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_0(f)}\right)^{\gamma} - 1\right]$$

où $P_0(f)$ est le seuil absolu et S_l est le facteur d'échelonnement en sonie (10.2.1.4).

Au-dessus de 4 Bark, la puissance de Zwicker, γ , est de 0,23, ce qui correspond à la valeur indiquée dans la littérature. Au-dessous de 4 Bark, on accroît légèrement la puissance de Zwicker pour tenir compte de l'effet dit de recrutement. On appelle densités de sonie les ensembles bidimensionnels $LX(f)_n$ et $LY(f)_n$ ainsi obtenus.

10.2.9 Calcul de la densité de perturbation

On calcule la différence de signe entre la densité de sonie du signal dégradé et la densité de sonie du signal initial. Quand cette différence est positive, cela indique que des éléments tels que le bruit ont été ajoutés. Quand cette différence est négative, cela indique que des éléments du signal initial ont été omis. On appelle cet ensemble de différences la densité de perturbation brute.

On calcule la valeur minimale de la densité de sonie du signal initial et du signal dégradé pour chaque cellule de fréquence temporelle et on multiplie par 0,25 les valeurs minimales ainsi obtenues.

On appelle ensemble de masque l'ensemble bidirectionnel correspondant. On applique ensuite à chaque cellule temps-fréquence les règles suivantes:

- Si la densité de perturbation brute est positive et supérieure à la valeur de masque, on soustrait celle-ci de la perturbation brute.
- Si la densité de perturbation brute se maintient entre le plus et le moins de la valeur de masque, on met la densité de perturbation à zéro.
- Si la densité de perturbation brute est inférieure au moins de la valeur de masque, on ajoute celle-ci à la densité de perturbation brute.

Ces règles ont pour effet de ramener vers zéro les densités de perturbation brute, ce qui correspond à une zone de temps mort avant que les cellules temps-fréquence soient effectivement perçues comme étant déformées. Ainsi s'opère la modélisation du traitement des petites différences inaudibles en présence de signaux ayant une valeur élevée de sonie (masquage) dans chaque cellule temps-fréquence. Il en résulte une densité de perturbation en fonction du temps (nombre de fenêtres *n*) et de la fréquence, $D(f)_n$.

10.2.10 Multiplication cellule par cellule avec un facteur d'asymétrie

L'effet d'asymétrie est dû au fait que lorsqu'un codec produit une distorsion du signal d'entrée, il sera généralement très difficile d'introduire une nouvelle composante temps-fréquence qui s'intègre dans le signal d'entrée, ce qui aura pour effet de décomposer le signal de sortie en deux phénomènes perceptifs différents, le signal d'entrée et la distorsion, d'où une distorsion clairement audible [2]. Lorsque le codec écarte une composante temps-fréquence, le signal de sortie produit ne peut être décomposé de la même manière et la distorsion est moins indésirable. On modélise cet effet en calculant la densité de perturbation asymétrique $DA(f)_n$ par trame et en multipliant la densité de perturbation $D(f)_n$ par un facteur d'asymétrie. Ce facteur d'asymétrie est égal au rapport des densités de puissance fondamentale du signal dégradé et du signal initial élevés à la puissance de 1,2. S'il est inférieur à 3, le facteur d'asymétrie est mis à zéro. S'il est supérieur à 12, on l'écrête à cette valeur. Ainsi, seules demeurent les cellules temps-fréquence, sous forme de valeurs non nulles, pour lesquelles la densité de puissance fondamentale du signal dégradé est supérieure à la densité de puissance fondamentale du signal initial.

10.2.11 Cumul des densités de perturbation par rapport à la fréquence et accentuation des parties du signal initial ayant une faible valeur de sonie

La densité de perturbation $D(f)_n$ et la densité de perturbation asymétrique $DA(f)_n$ sont intégrées (sommées) le long de l'axe des fréquences au moyen de deux valeurs d'étalonnage *Lp* différentes et d'une pondération appliquée aux trames ayant une faible valeur de sonie.

$$D_n = M_n \sqrt[3]{\sum_{f=1,\ldots \text{nombre de bandes de Bark}} \left(\left| D(f)_n \right| W_f \right)^3}$$

$$DA_n = M_n \sum_{f=1,\ldots \text{nombre de bandes de Bark}} \left(\left| DA(f)_n \right| W_f \right)$$

où M_n est un facteur de multiplication, *1/(puissance de la trame initiale plus constante)*^{0,04}, représente une accentuation des perturbations survenant pendant les silences dans le fragment vocal du signal initial, et W_f une série de constantes proportionnelles à la largeur des segments de Bark modifiés. Après cette multiplication, les valeurs de perturbation de trame sont limitées à un maximum de 45. Ces valeurs cumulées, D_n et DA_n, sont appelées les perturbations de trame.

10.2.12 Annulation de la perturbation pour les trames durant lesquelles le temps de propagation a subi une diminution importante

Si le signal dégradé subit une diminution du temps de propagation supérieure à 16 ms (une demi-fenêtre) la stratégie de répétition visée au 10.2.4 est modifiée. Il s'est avéré préférable de ne pas tenir compte des perturbations de trame en pareil cas dans le calcul de la qualité objective de la parole. On annule donc les perturbations de trame en pareille situation. Les perturbations de trame obtenues sont appelées D'_n et DA'_n.

10.2.13 Réalignement des intervalles erronés

On appelle intervalles erronés les trames consécutives qui présentent une perturbation supérieure à un certain seuil. Dans une minorité de cas, la mesure objective prévoit de fortes distorsions sur un nombre minimum de trames erronées, imputables à une mauvaise estimation des temps de propagation observés lors du prétraitement. Pour ces intervalles dits erronés, on estime une nouvelle valeur de temps de propagation en maximisant la corrélation croisée entre le signal initial absolu et le signal dégradé absolu, ajustés d'après les temps de propagation observés lors du prétraitement. Lorsque la corrélation croisée maximale est inférieure à un certain seuil, on en conclut que l'intervalle rend bien compte (bruit pour bruit) de la réalité et on en interrompt le traitement du fait qu'il ne saurait plus être considéré comme étant erroné. Dans le cas contraire, on recalcule la perturbation pour les trames durant les intervalles erronés et, si elle est inférieure, on substitue cette perturbation à la perturbation de trame initiale. On obtient ainsi les perturbations de trame finales D''_n et DA''_n qui sont utilisées pour calculer la qualité perçue.

10.2.14 Cumul des valeurs de perturbation dans des intervalles de fraction de seconde

On cumule ensuite les valeurs de perturbation de trames et les valeurs de perturbation de trames asymétriques dans des intervalles de fraction de seconde de 20 trames (compte tenu du chevauchement des trames: 320 ms environ) en utilisant des valeurs d'étalonnage L_6, une valeur p supérieure comme dans le cas du cumul sur toute la longueur du fichier vocal. Ces intervalles présentent aussi un chevauchement de 50% et aucune fonction de fenêtre n'est utilisée.

10.2.15 Cumul des valeurs de perturbation pendant la durée du signal vocal (10 secondes environ), incluant un facteur de récenteté

On cumule à présent les valeurs de perturbation de fraction de seconde et les valeurs de perturbation de fraction de seconde asymétrique pendant l'intervalle actif des fichiers vocaux (les trames correspondantes) en utilisant les valeurs d'étalonnage L_2. La valeur élevée de p pour le cumul dans les intervalles de fraction de seconde par rapport à la valeur inférieure de p pour le cumul dans le fichier vocal est due au fait qu'en cas de distorsion de parties de fractions de seconde, la fraction de seconde concernée perd toute signification, alors qu'en cas de distorsion de la première phrase d'un fichier vocal la qualité des autres phrases reste intacte.

10.2.16 Calcul de la note d'évaluation PESQ

La note d'évaluation PESQ finale est une combinaison linéaire de la valeur de perturbation moyenne et de la valeur de perturbation asymétrique moyenne. La note d'évaluation PESQ s'échelonne de –0,5 à 4,5, bien que dans la plupart des cas la valeur du signal de sortie soit une note de type MOS de la qualité d'écoute comprise entre 1,0 et 4,5, ce qui correspond à l'amplitude normale des valeurs MOS observées dans une expérience ACR.

Bibliographie

[1] BEERENDS (J.G.), STEMERDINK (J.A.): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation, *J. Audio Eng. Soc.,* Vol. 42, No. 3, pp. 115-123, mars 1994.

[2] BEERENDS (J.G.): Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, *Speech Quality Assessment*, Workshop papers, Bochum, pp. 1-9, novembre 1994.

[3] BEERENDS (J.G.): Measuring the quality of speech and music codecs, an integrated psychoacoustic approach, *98th AES Convention*, pre-print No. 3945, 1995.

[4] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.): Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain, *IEE Proceedings – Vision, Image and Signal Processing*, 141 (3), 203-208, juin 1994.

[5] RIX (A.W.), REYNOLDS (R.), HOLLIER (M.P.): Perceptual measurement of end-to-end speech quality over audio and packet-based networks, *106th AES Convention*, pre-print No. 4873, mai 1999.

[6] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.): Characterisation of communications systems using a speech-like test stimulus, *Journal of the AES*, 41 (12), 1008-1021, décembre 1993.

[7] ZWICKER (E), FELDTKELLER (R.): Das Ohr als Nachrichtenempfänger, *S. Hirzel Verlag,* Stuttgart 1967.

ANNEXE A

Implémentation de référence de l'évaluation PESQ et tests de conformité

Liste des fichiers fournis pour la mise en œuvre de référence C-ANSI

La mise en œuvre de référence C-ANSI de l'évaluation PESQ est contenue dans les fichiers de texte suivants, qui figurent dans le sous-répertoire source du CD-ROM distribué:

- dsp.c *Procédures périodiques DSP de base*
- dsp.h *Fichier d'en-tête pour dsp.c*
- pesq.h *Fichier d'en-tête général*
- pesqdsp.c *Procédures périodiques DSP PESQ*
- pesqio.c *Fichiers d'entrée/sortie*
- pesqmain.c *Programme principal*
- pesqmod.c *Modèle PESQ de haut niveau*
- pesqpar.h *Définitions du modèle perceptif PESQ*

La mise en œuvre de référence C-ANSI, qui est fournie dans des fichiers distincts, fait partie intégrante de la présente Recommandation. En cas de divergences entre la description approfondie figurant dans la présente Recommandation et l'implémentation de référence C-ANSI, cette dernière prévaudra.

Liste des fichiers fournis pour la validation de la conformité

Le processus de validation de la conformité décrit ci-après fait référence aux fichiers suivants, qui figurent dans le sous-répertoire "conform" du CD-ROM distribué.

- supp23.txt *Paires de fichiers et notes PESQ pour la validation de la conformité par rapport au Supplément 23*
- supp23.bat *Script batch destiné à faciliter la validation par rapport au Supplément 23*
- voipref.txt *Paires de fichiers et notes PESQ pour la validation de la conformité par rapport à la variation du temps de propagation*
- voipref.bat *Script batch destiné à faciliter la validation par rapport à la variation du temps de propagation*
- or105.wav or109.wav or114.wav or129.wav or134.wav or137.wav
  or145.wav or149.wav or152.wav or154.wav or155.wav or161.wav
  or164.wav or166.wav or170.wav or179.wav or221.wav or229.wav
  or246.wav or272.wav dg105.wav dg109.wav dg114.wav dg129.wav
  dg134.wav dg137.wav dg145.wav dg149.wav dg152.wav dg154.wav
  dg155.wav dg161.wav dg164.wav dg166.wav dg170.wav dg179.wav
  dg221.wav dg229.wav dg246.wav dg272.wav
  u_am1s01.wav u_am1s02.wav u_am1s03.wav
  u_am1s01b1c1.wav u_am1s01b1c7.wav u_am1s02b1c9.wav
  u_am1s01b1c15.wav u_am1s03b1c16.wav u_am1s03b1c18.wav
  u_am1s01b2c1.wav u_am1s02b2c4.wav u_am1s02b2c5.wav
  u_am1s03b2c5.wav u_am1s03b2c6.wav u_am1s03b2c7.wav
  u_am1s01b2c8.wav u_am1s03b2c11.wav u_am1s02b2c14.wav
  u_af1s01b2c16.wav u_af1s03b2c16.wav u_af1s02b2c17.wav
  u_af1s03b2c17.wav u_am1s03b2c18.wav u_af1s01.wav
  u_af1s02.wav u_af1s03.wav

  Fichiers de signaux vocaux fournis pour la validation par rapport à la variation du temps de propagation.

Ces fichiers de signaux vocaux sont de format Wave (modulation MIC linéaire à 16 bits, classement des octets de type Intel, en-tête de 44 octets) avec un taux d'échantillonnage de 8 kHz.

Validation de la conformité

La conformité d'une implémentation de l'évaluation PESQ avec la présente Recommandation doit être validée de la façon suivante. Le processus de validation est fondé sur une comparaison entre l'implémentation de référence (code C-ANSI décrit dans la présente annexe) et l'implémentation par l'utilisateur de l'évaluation PESQ. L'utilisateur peut ainsi procéder à une comparaison détaillée des différentes variables en examinant les différences à tous les niveaux requis.

Pour être considérée comme conforme à la présente Recommandation, une implémentation doit être soumise à deux essais obligatoires consistant à comparer la note PESQ à celle qui est donnée par l'implémentation de référence dans les bases de données suivantes:

- Supplément 23 aux Recommandations UIT-T de la série P (1736 paires de fichiers);
- conditions de variation du temps de propagation (40 paires de fichiers).

Comparaison par rapport au Supplément 23 aux Recommandations UIT-T de la série P

Dans cette comparaison, la totalité des expériences (au nombre de 10) effectuées selon le Supplément 23 aux Recommandations de la série P sont utilisées fichier par fichier. Les noms du fichier initial et du fichier dégradé, ainsi que la note PESQ donnée par l'implémentation de référence, sont indiqués dans le fichier supp23.txt qui est fourni avec le code C-ANSI PESQ de l'UIT.

Lorsqu'une implémentation est soumise à cet essai, le résultat est positif lorsque la différence absolue entre la note PESQ et la note donnée par l'implémentation de référence est inférieure à 0,05 pour **tous** les fichiers.

Cet essai de conformité est obligatoire pour toutes les implémentations.

Le Supplément 23 aux Recommandations de la série P peut être obtenu séparément auprès de l'UIT.

Comparaison par rapport aux fichiers des temps de propagation variables

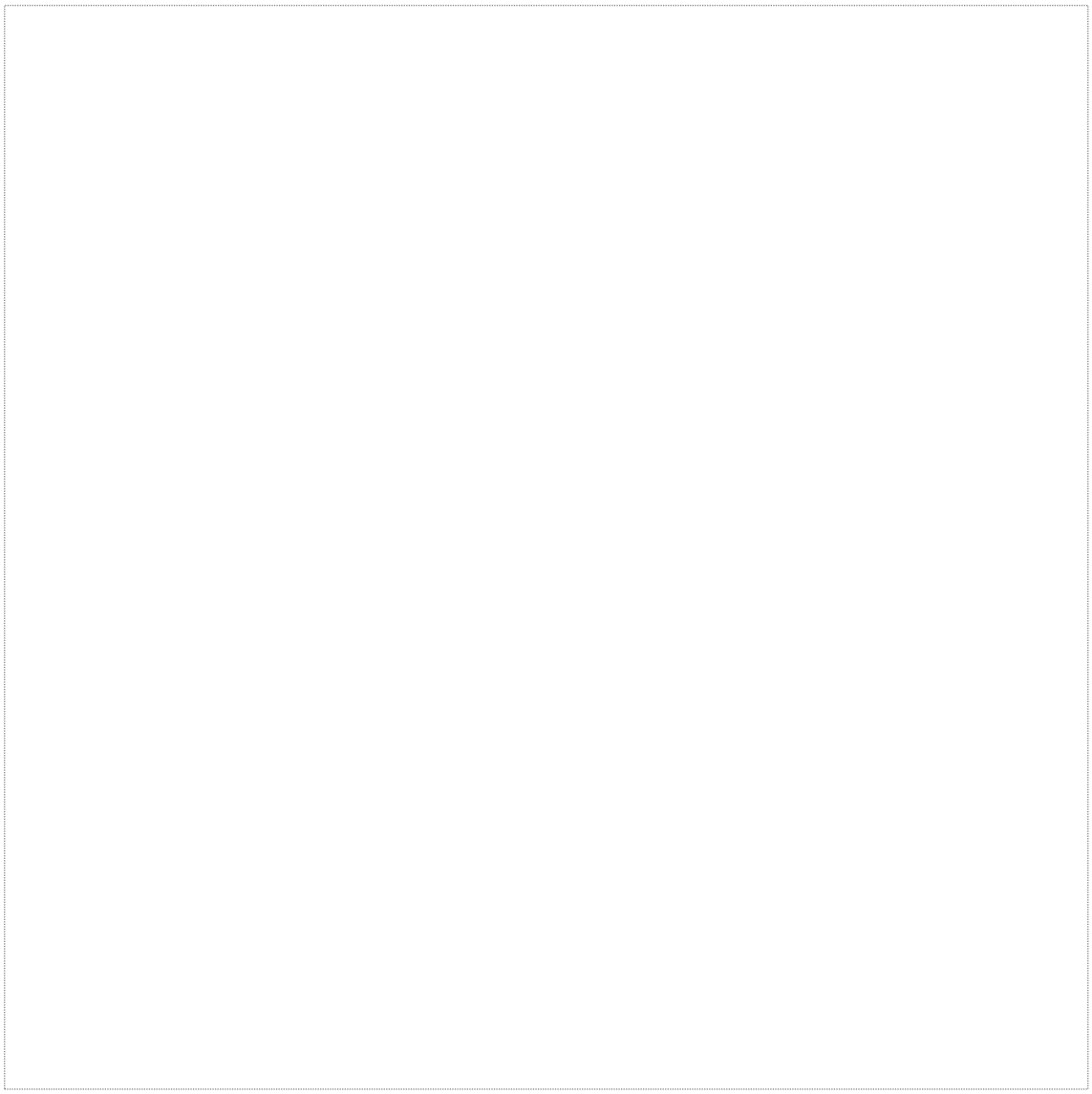
Une base de données variée a été élaborée à partir de 40 conditions (paires de fichiers) provenant de deux essais subjectifs visant des connexions VoIP réelles et simulées, se caractérisant par un temps de propagation variable. Vingt-trois de ces paires de fichiers déclenchent également le processus de réalignement des intervalles erronés. Les noms du fichier initial et du fichier dégradé pour chaque paire de fichiers et la note PESQ donnée par l'implémentation de référence sont indiqués dans le fichier voipref.txt qui est fourni avec le code C-ANSI.

Lorsqu'une implémentation est soumise à cet essai, le résultat est positif lorsque la différence absolue entre la note PESQ et celle qui est donnée par l'implémentation de référence est inférieure à 0,05 pour 39 des 40 paires de fichiers. Une seule paire de fichiers peut présenter une différence absolue en matière de note PESQ inférieure à 0,5. Il peut s'agir de n'importe laquelle des 40 paires de fichiers.

Cet essai de conformité est obligatoire pour toutes les implémentations.

Comparaisons additionnelles

Un autre essai est disponible afin d'empêcher les réalisateurs d'adapter un algorithme pour assurer la conformité avec les prescriptions spécifiées pour les fichiers décrits ci-dessus. Une implémentation de l'évaluation PESQ conforme à la présente Recommandation doit, au moins dans 95% des cas, donner une note dont la différence avec la note PESQ donnée par l'implémentation de référence C-ANSI doit être inférieure à 0,05.



SÉRIES DES RECOMMANDATIONS UIT-T

Série A	Organisation du travail de l'UIT-T
Série B	Moyens d'expression: définitions, symboles, classification
Série C	Statistiques générales des télécommunications
Série D	Principes généraux de tarification
Série E	Exploitation générale du réseau, service téléphonique, exploitation des services et facteurs humains
Série F	Services de télécommunication non téléphoniques
Série G	Systèmes et supports de transmission, systèmes et réseaux numériques
Série H	Systèmes audiovisuels et multimédias
Série I	Réseau numérique à intégration de services
Série J	Réseaux câblés et transmission des signaux radiophoniques, télévisuels et autres signaux multimédias
Série K	Protection contre les perturbations
Série L	Construction, installation et protection des câbles et autres éléments des installations extérieures
Série M	RGT et maintenance des réseaux: systèmes de transmission, circuits téléphoniques, télégraphie, télécopie et circuits loués internationaux
Série N	Maintenance: circuits internationaux de transmission radiophonique et télévisuelle
Série O	Spécifications des appareils de mesure
Série P	**Qualité de transmission téléphonique, installations téléphoniques et réseaux locaux**
Série Q	Commutation et signalisation
Série R	Transmission télégraphique
Série S	Equipements terminaux de télégraphie
Série T	Terminaux des services télématiques
Série U	Commutation télégraphique
Série V	Communications de données sur le réseau téléphonique
Série X	Réseaux de données et communication entre systèmes ouverts
Série Y	Infrastructure mondiale de l'information et protocole Internet
Série Z	Langages et aspects généraux logiciels des systèmes de télécommunication

20783

Imprimé en Suisse
Genève, 2001