Groupe de travail Réseau
Request for Comments : 3557
Catégorie : En cours de normalisation

Q. Xie, éditeur, Motorola, Inc.
juillet 2003
Traduction Claude Brière de L'Isle

Format de charge utile RTP pour la norme européenne ES 201 108 "Codage réparti de reconnaissance vocale" de l'Institut européen des normes de télécommunications (ETSI)

Statut du présent mémoire
Le présent document spécifie un protocole de l'Internet en cours de normalisation pour la communauté de l'Internet, et appelle à des discussions et suggestions pour son amélioration. Prière de se référer à l'édition en cours des "Protocoles officiels de l'Internet" (STD 1) pour voir l'état de normalisation et le statut de ce protocole. La distribution du présent mémoire n'est soumise à aucune restriction.

Résumé
Le présent document spécifie un format de charge utile RTP pour encapsuler les flux de caractéristiques de signal frontal de la norme européenne 201 108 de l'Institut européen des normes de télécommunications (ETSI, *European Telecommunications Standards Institute*) pour les systèmes répartis de reconnaissance de la parole (DSR, *distributed speech recognition*).

Table des Matières

1. Conventions et acronymes

Dans le présent document, les mots clés "DOIT", "NE DOIT PAS", "EXIGÉ", "DEVRA", "NE DEVRA PAS", "DEVRAIT", "NE DEVRAIT PAS", "RECOMMANDE", "PEUT", et "FACULTATIF" sont à interpréter comme décrit dans le BCP 14, [RFC2119] et indiquent les niveaux d'exigence pour les mises en œuvre conformes.

Les acronymes suivants sont utilisés dans le présent document :

DSR - Reconnaissance répartie de la parole

ETSI - Institut européen des normes de télécommunications

FP - Paire de trames

DTX - Transmission discontinue

2. Introduction

Motivés par les avancées technologiques dans le domaine de la reconnaissance de la parole, des interfaces vocales aux services (comme les systèmes d'information des lignes aériennes, la messagerie unifiée) deviennent de plus en plus présents. En parallèle, la popularité des appareils mobiles a aussi augmenté de façon spectaculaire.

Cependant, les codecs vocaux normalement employés dans les appareils mobiles étaient conçus pour optimiser la qualité vocale audible et non pour la précision de la reconnaissance vocale, et l'utilisation de ces codecs avec la reconnaissance vocale peut avoir pour résultat des performances de reconnaissance assez médiocres. Pour les systèmes auxquels on peut accéder à partir de réseaux hétérogènes qui utilisent plusieurs codecs vocaux, les concepteurs de systèmes de reconnaissance sont de plus confrontés à s'accommoder des caractéristiques de ces différences de façon robuste. Les erreurs de canal et les pertes de paquets de données dans ces réseaux résultent en une dégradation supplémentaire du signal vocal.

Dans les systèmes traditionnels tels que décrits ci-dessus, la totalité de la reconnaissance de la parole repose sur le serveur. Il est forcé d'utiliser la parole entrante dans toutes les condition dans lesquelles elle arrive après que le réseau a décodé la parole codée. Pour traiter ce problème, on utilise une architecture répartie de reconnaissance de la parole (DSR, *distributed speech recognition*). Dans un tel système, l'appareil distant agit comme un client mince, aussi appelé le frontal, en communication avec un serveur de reconnaissance de la parole, aussi appelé un moteur de parole. L'appareil distant traite la parole, compresse les données, et ajoute au flux binaire une protection contre les erreurs d'une façon optimale pour la reconnaissance de la parole. Le moteur de parole utilise alors directement cette représentation, minimisant le traitement de signal nécessaire et tirant parti de l'amélioration de la dissimulation d'erreurs.

Pour réaliser l'interopérabilité avec les différents appareils clients et moteurs de parole, un format commun est nécessaire. Dans le groupe de travail DSR "Aurora" de l'Institut Européen des normes de Télécommunications (ETSI), une charge utile a été définie et a été publiée comme norme européenne [ES201108] en février 2000.

Pour les dialogues vocaux entre un appelant et un service vocal, une faible latence est une forte priorité ainsi qu'une reconnaissance précise de la parole. Bien que la gigue dans l'entrée du reconnaisseur de parole ne soit pas particulièrement importante, de nombreuses questions relatives à l'interaction de la parole sur une connexion fondée sur IP sont encore pertinentes. Donc, il est souhaitable d'utiliser la charge utile DSR dans une session fondée sur RTP.

2.1 Codec frontal DSR de l'ES 201 108 d'ETSI

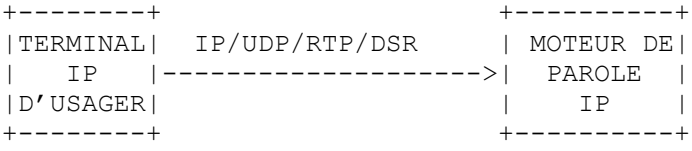
La norme européenne d'ETSI ES 201 108 pour DSR [ES201108] définit un traitement de signal frontal et un schéma de compression pour l'entrée de parole dans un système de reconnaissance de la parole. Certaines caractéristiques pertinentes de ce codec frontal d'ETSI sont résumées ci-dessous.

L'algorithme de codage, une technique mel-cepstral standard commune à de nombreux systèmes de reconnaissance de la parole, prend en charge trois taux d'échantillonnage bruts : 8 kHz, 11 kHz, et 16 kHz. Le calcul de mel-cepstral est un schéma fondé sur la trame qui produit un vecteur de sortie toutes les 10 ms.

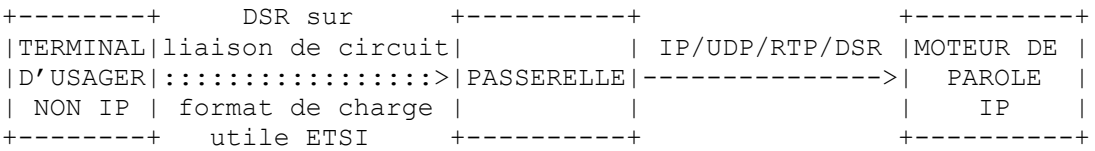
Après le calcul de la représentation mel-cepstral, celle ci est d'abord quantifiée via la quantification de vecteur partagé pour réduire le débit de données du flux codé. Puis, les vecteurs quantifiés provenant de deux trames consécutives sont mis dans une paire de trames (FP), comme décrit plus en détails au paragraphe 4.1.

2.2 Scénarios typiques d'utilisation du format de charge utile DSR

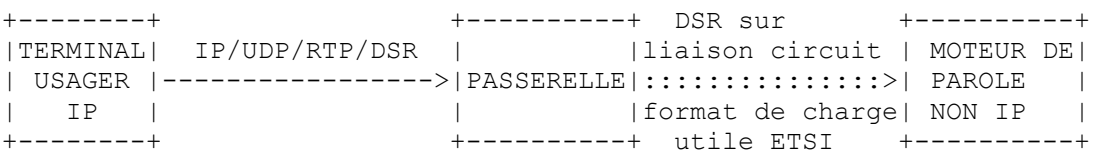
Les diagrammes de la Figure 1 montrent des scénarios typiques d'utilisation du format de charge utile RTP de DSR de l'ES 201 108.



a) De terminal d'utilisateur IP à moteur de parole IP



b) Terminal d'utilisateur non IP à moteur de parole IP via une passerelle



c) Terminal d'usager IP à moteur de parole non IP via une passerelle.

Figure 1 : Scénarios typiques d'utilisation du format de charge utile DSR.

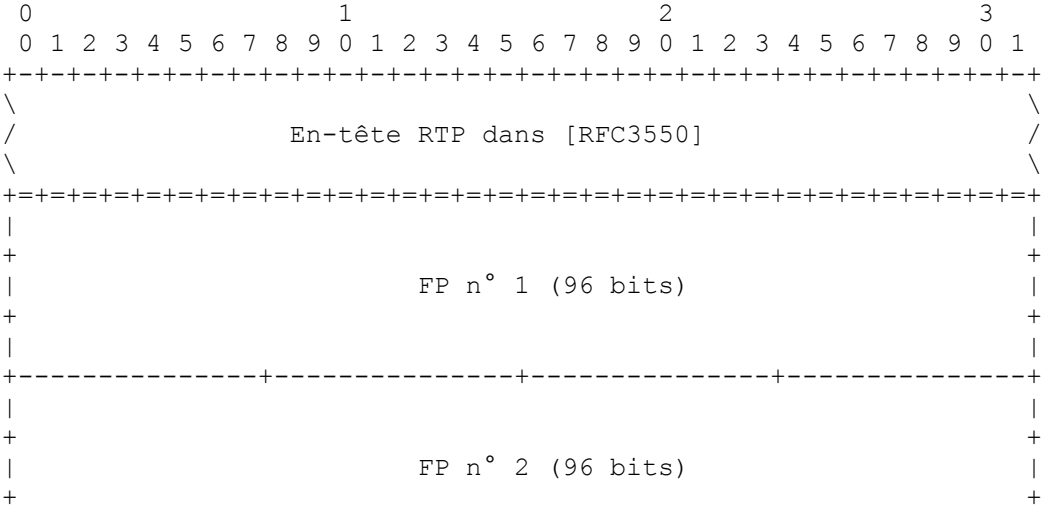
Pour les différents scénarios de la Figure 1, le reconnaisseur de parole réside toujours dans le moteur de parole. Un codeur DSR frontal à l'intérieur du terminal d'utilisateur effectue le traitement frontal de parole et envoie les données résultantes au moteur de parole sous la forme de "paires de trames" (FP). Chaque FP contient deux ensembles de vecteurs de parole codés représentant 20 ms de parole originale.

3. Format de charge utile RTP de DSR de l'ES 201 108

Un datagramme de charge utile RTP de DSR de l'ES 201 108 consiste en un en-tête RTP standard [RFC3550] suivi par une charge utile DSR. La charge utile DSR est elle-même formée en enchaînant une série de FP DSR ES 201 108 (définies à la Section 4).

Les FP sont toujours empaquetées en bits contigus dans les octets de charge utile en commençant par le bit de poids fort. Pour le frontal ES 201 108, la taille de chaque FP est de 96 bits ou 12 octets (voir les paragraphes 4.1 et 4.2). Cela assure qu'une charge utile DSR va toujours se terminer sur une limite d'octet.

L'exemple suivant montre un datagramme RTP de DSR qui porte une charge utile de DSR contenant trois FP longues de 96 bits (le bit 0 est le bit de poids fort *(MSB)*) :



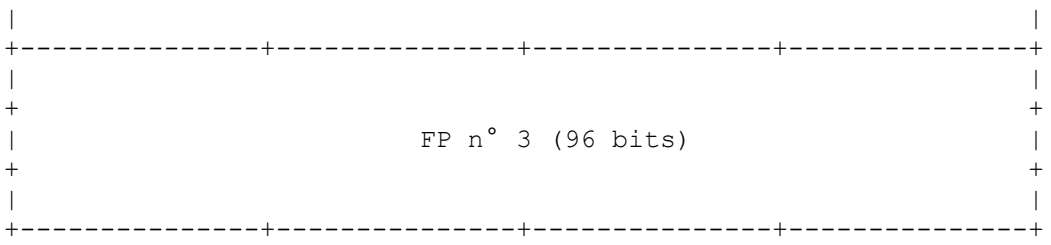


Figure 2 : Exemple de charge utile RTP de DSR ES 201 108

3.1 Considération sur le nombre de FP dans chaque paquet RTP

Le nombre de FP par paquet de charge utile devrait être déterminé par les exigences de latence et de bande passante de l'application DSR qui utilise ce format de charge utile. En particulier, utiliser un plus petit nombre de FP par paquet de charge utile dans une session résultera en une diminution de l'efficacité de la bande passante à cause de la redondance de l'en-tête RTP/UDP/IP, tandis qu'utiliser un plus grand nombre de FP par paquet causera un allongement du délai de bout en bout et donc une latence accrue de reconnaissance. De plus, porter un plus grand nombre de FP par paquet va augmenter la possibilité de pertes catastrophiques de paquets ; la perte d'un grand nombre de FP consécutives est une situation que la plupart des reconnaiseurs de parole ont des difficultés à traiter.

Il est donc RECOMMANDÉ que le nombre de FP par paquet de charge utile DSR soit minimisé, sous réserve de satisfaire les exigences de l'application sur l'efficacité de la bande passante du réseau. Les techniques de compression d'en-tête RTP, comme celles définies dans les [RFC2508] et [RFC3095], devraient être considérées pour améliorer l'efficacité de la bande passante du réseau.

3.2 Prise en charge de la transmission discontinue

Les charges utiles RTP de DSR peuvent être utilisées pour prendre en charge la transmission discontinue (DTX) de parole, ce qui permet que les FP DSR ne soient envoyées que lorsque de la parole a été détectée à l'équipement terminal.

En DTX, un ensemble de trames DSR codant un segment de parole non interrompu transmis du terminal au serveur est appelé un segment de transmission. Une trame DSR à l'intérieur d'un tel segment de transmission peut être soit une trame de parole, soit une trame non de parole, selon la nature de la section de signal de parole qu'elle représente.

La fin d'un segment de transmission est déterminée à l'équipement de l'extrémité d'envoi lorsque le nombre de trames non de parole consécutives excède un seuil pré établi, appelé la période de latence *(hangover time)*. La valeur normale utilisée pour la période de latence est 1,5 secondes.

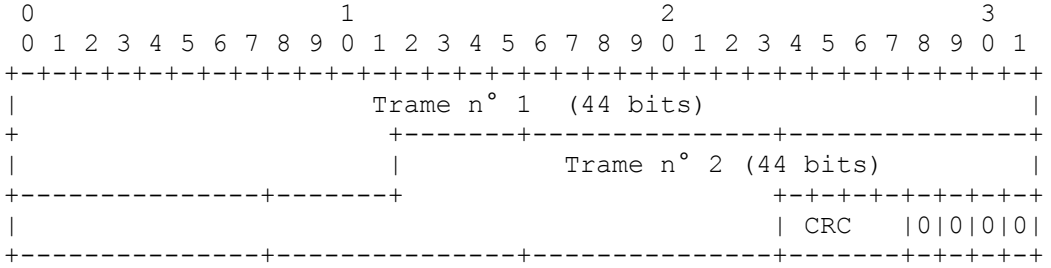
Après l'envoi de toutes les FP dans un segment de transmission, le frontal DEVRAIT indiquer la fin du segment de transmission en cours en envoyant une ou plusieurs FP nulles (définie au paragraphe 4.2).

4. Formats de paire de trame

4.1 Format de FP de parole et de non parole

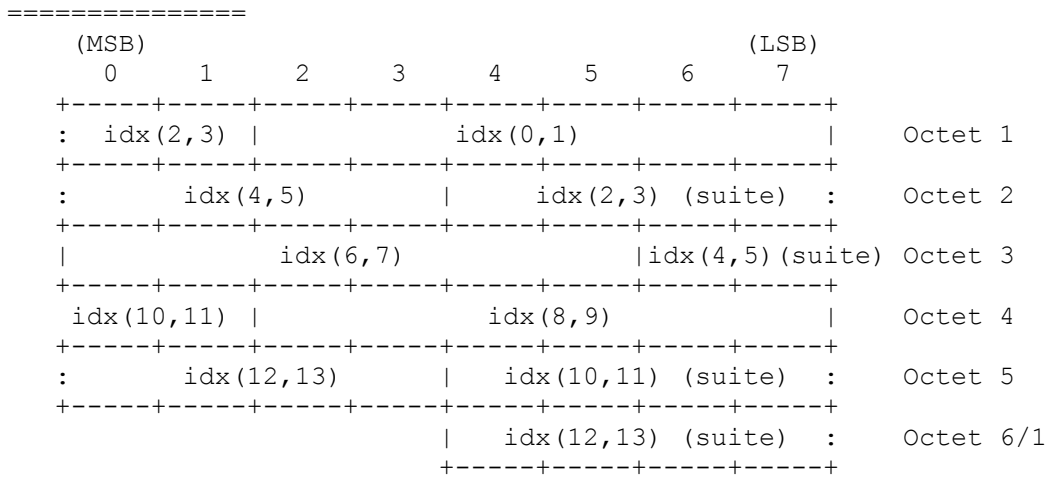
La trame mel-cepstral suivante DOIT être utilisée, comme défini dans [ES201108] :

Comme défini dans [ES201108], les paires de trames mel-cepstral quantifiées de 10 ms DOIVENT être groupées ensemble et protégées avec un CRC de 4 bits, formant une FP longue de 92 bits:

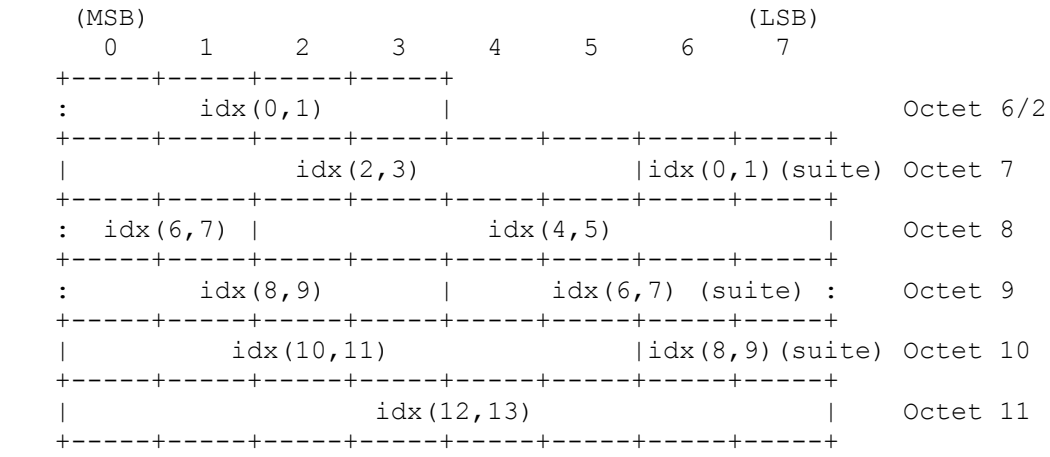


La longueur de chaque trame est 44 bits représentant 10 ms de voix. Les formats de trame mel-cepstral DOIVENT être utilisés pour former une FP :

Trame n° 1 dans la FP :



Trame n° 2 dans la FP :



Donc, chaque FP représente 20 ms de parole originale. Noter que comme on le montre ci-dessus, chaque FP DOIT être bourrée avec 4 zéros à la fin pour réaliser l'alignement sur la limite de mot de 32 bits. Cela donne la taille de 96 bits, ou 12 octets d'une FP. Noter que ce bourrage est distinct du bourrage indiqué par le bit P dans l'en-tête RTP.

Le CRC de 4 bits DOIT être calculé en utilisant la formule définie au paragraphe 6.2.4 de [ES201108]. La définition des indices et la détermination de leur valeur sont aussi décrites dans [ES201108].

4.2 Format de FP nulle

Une FP nulle est définie pour le codec frontal de l'ES 201 108 en réglant le contenu des deux premières trames de la FP à nul (c'est-à-dire, en remplissant les 88 premiers bits de la FP de 0). Le CRC de 4 bits DOIT être calculé de la même façon que décrit au paragraphe 6.2.4 de [ES201108], et 4 zéros DOIVENT être ajoutés à la fin de la FP nulle pour l'aligner sur le mot de 32 bits.

4.3 Utilisation de l'en-tête RTP

Le format de l'en-tête RTP est spécifié dans la [RFC3550]. Ce format de charge utile utilise les champs de l'en-tête d'une manière cohérente avec la présente spécification.

L'horodatage RTP correspond à l'instant d'échantillonnage du premier échantillon codé pour la première FP dans le paquet. La fréquence de l'horloge d'horodatage est la même que la fréquence d'échantillonnage, de sorte que l'unité d'horodatage est l'échantillon.

Comme défini par le codec frontal ES 201 108, la durée d'une FP est 20 ms, correspondant à 160, 220, ou 320 échantillons avec un taux d'échantillonnage respectivement de 8, 11, ou 16 kHz utilisé au frontal. Donc, l'horodatage est augmenté respectivement de 160, 220, ou 320 pour chaque FP consécutive.

La charge utile DSR pour le codec frontal ES 201 108 est toujours un nombre entier d'octets. Si un bourrage supplémentaire est nécessaire pour un autre objet, le bit P dans l'en-tête RTP peut être établi (à 1) et le bourrage ajouté comme spécifié dans la [RFC3550].

Le bit marqueur (M) de l'en-tête RTP devrait être établi en suivant les règles générales définies dans la [RFC3551].

L'allocation d'un type de charge utile RTP pour ce nouveau format de paquet sort du domaine d'application du présent document, et ne sera pas spécifiée ici. Il est prévu que le profil RTP sous lequel sera utilisé ce format de charge utile allouera un type de charge utile pour ce codage ou spécifiera que le type de charge utile sera lié de façon dynamique.

5. Considérations relatives à l'IANA

L'enregistrement d'un nouveau sous-type MIME est demandé pour ce type de charge utile, comme défini ci-dessous.

Cette section définit aussi les paramètres facultatifs qui peuvent être utilisés pour décrire une session DSR. Les paramètres sont définis ici au titre de l'enregistrement de sous-type MIME. Une transposition des paramètres dans le protocole de description de session (SDP, *Session Description Protocol*) [RFC2327] est aussi fournie en 5.1 pour les applications qui utilisent SDP.

Nom du type de support : audio
Nom du sous-type de support : dsr-es201108
Paramètres exigés : aucun
Paramètres facultatifs :

taux : Indique le taux d'échantillonnage de la parole. Les valeurs valides incluent : 8000, 11000, et 16000. Si ce paramètre est absent, on suppose le taux d'échantillonnage de 8000.

maxptime : Quantité maximum de support qui peut être encapsulée dans chaque paquet, exprimée en millisecondes. Le temps devra être calculé comme la somme des temps que représente le support présent dans le paquet. Le temps DEVRAIT être un multiple de la taille de la paire de trame (c'est-à-dire, une FP <-> 20 ms). Si ce paramètre est absent, maxptime est supposé être 80 ms.

Note : comme les performances de la plupart des reconnaiseurs de parole sont extrêmement sensibles aux pertes de FP consécutives, si l'utilisateur du format de charge utile s'attend à un fort ratio de perte de paquet pour la session, il PEUT choisir explicitement pour la session une valeur de maxptime inférieure à la valeur par défaut.

ptime : voir la [RFC2327].

Considérations de codage : Ce type est défini pour le transfert via RTP [RFC3550] comme décrit aux Sections 3 et 4 de la RFC3557.

Considérations de sécurité : voir la Section 6 de la RFC3557.

Adresse personnelle & de messagerie à contacter pour d'autres informations : Qiaobing.Xie@motorola.com

Utilisation prévue : COMMUNE. On prévoit que de nombreuses applications de VoIP (ainsi que d'applications mobiles) vont utiliser ce type.

Auteur/contrôleur de changements : Qiaobing.Xie@motorola.com groupe de travail IETF Transport Audio/Vidéo

5.1 Transposition des paramètres MIME en SDP

Les informations portées dans la spécification de type de support MIME ont une transposition spécifique dans le protocole de description de session (SDP) [RFC2327], qui est couramment utilisée pour décrire les sessions RTP. Lorsque SDP est utilisé pour spécifier les sessions qui emploient le codec DSR ES 201 018, la transposition est la suivante :

- Le type MIME ("audio") passe en SDP "m=" comme le nom du support.
- Le sous-type MIME ("dsr-es201108") passe dans SDP "a=rtpmap" comme le nom du codage.
- Le paramètre facultatif "rate" passe aussi dans "a=rtpmap" comme le débit d'horloge.
- Les paramètres facultatifs "ptime" et "maxptime" passent respectivement dans les attributs SDP "a=ptime" et "a=maxptime".

Exemple d'utilisation de la DSR de l'ES 201 108 :

```
m=audio 49120 RTP/AVP 101
a=rtpmap:101 dsr-es201108/8000
a=maxptime:40
```

6. Considérations pour la sécurité

Les mises en œuvre qui utilisent la charge utile définie dans la présente spécification sont soumises aux considérations pour la sécurité exposées dans la [RFC3550] et le profil RTP [RFC3551]. La présente charge utile ne spécifie aucun service de sécurité différent.

7. Contributeurs

Les individus suivants ont contribué à la conception de ce format de charge utile et à la rédaction du présent document : Q. Xie (Motorola), D. Pearce (Motorola), S. Balasuriya (Motorola), Y. Kim (VerbalTek), S. H. Maes (IBM), et Hari Garudadri (Qualcomm).

8. Remerciements

Le concept présenté ici a largement bénéficié d'un travail antérieur sur le concept de charge utile RTP de DSR de Jeff Meunier et Priscilla Walther. Les auteurs tiennent aussi à remercier Brian Eberman, John Lazzaro, Magnus Westerlund, Rainu Pierce, Priscilla Walther, et les autres de leur relecture et de leurs précieux commentaires sur le présent document.

9. Références

9.1 Références normatives

[ES201108] European Telecommunications Standards Institute (ETSI) Standard ES 201 108, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms", version 1.1.2, 11 avril 2000.

[RFC2026] S. Bradner, "Le processus de normalisation de l'Internet -- Révision 3", (BCP0009) octobre 1996. *(Remplace* RFC1602, RFC1871) (*MàJ par* RFC3667, RFC3668, RFC3932, RFC3979, RFC3978, RFC5378, RFC6410)

[RFC2119] S. Bradner, "Mots clés à utiliser dans les RFC pour indiquer les niveaux d'exigence", BCP 14, mars 1997.

[RFC2327] M. Handley et V. Jacobson, "SDP : Protocole de description de session", avril 1998. *(Obsolète; voir* RFC4566)

[RFC3550] H. Schulzrinne, S. Casner, R. Frederick et V. Jacobson, "RTP : un protocole de transport pour les applications en temps réel", STD 64, juillet 2003. *(MàJ par* RFC7164, RFC7160)

9.2 Références pour information

[RFC2508] S. Casner, V. Jacobson, "Compression d'en-têtes IP/UDP/RTP pour liaisons séries à bas débit", février 1999. *(P.S.)*

[RFC3095] C. Bormann et autres, "Compression d'en-tête robuste (ROHC) : cadre et quatre profils", juillet 2001. (*MàJ par* RFC3759, RFC4815) *(P.S.)*

[RFC3551] H. Schulzrinne et S. Casner, "Profil RTP pour conférences audio et vidéo avec contrôle minimal", STD 65, juillet 2003.

10. Propriété intellectuelle

tentatives faites pour obtenir une licence ou permission générale d'utilisation de tels droits de propriété par ceux qui mettent en œuvre ou utilisent la présente spécification peuvent être obtenues sur répertoire en ligne des IPR de l'IETF à http://www.ietf.org/ipr .

L'IETF invite toute partie intéressée à porter son attention sur tous copyrights, licences ou applications de licence, ou autres droits de propriété qui pourraient couvrir les technologies qui peuvent être nécessaires pour mettre en œuvre la présente norme. Prière d'adresser les informations à l'IETF à ietf-ipr@ietf.org .

11. Adresse des auteurs

David Pearce
Motorola Labs
UK Research Laboratory
Jays Close
Viables Industrial Estate
Basingstoke, HANTS, RG22 4PD
téléphone : +44 (0)1256 484 436
mél : bdp003@motorola.com

Senaka Balasuriya
Motorola, Inc.
600 U.S Highway 45
Libertyville, IL 60048,
USA
téléphone : +1-847-523-0440
mél : Senaka.Balasuriya@motorola.com

Yoon Kim
VerbalTek, Inc.
2921 Copper Rd.
Santa Clara, CA 95051
téléphone : +1-408-768-4974
mél : yoonie@verbaltek.com

Stephane H. Maes, PhD,
Oracle
500 Oracle Parkway, M/S 4op634
Redwood City, CA 94065 USA
téléphone : +1-650-607-6296.
mél : stephane.maes@oracle.com

Hari Garudadri
Qualcomm Inc.
5775, Morehouse Dr.
San Diego, CA 92121-1714, USA
téléphone : +1-858-651-6383
mél : hgarudad@qualcomm.com

12. Adresse de l'éditeur

Qiaobing Xie
Motorola, Inc.
1501 W. Shure Drive, 2-F9
Arlington Heights, IL 60004
USA
téléphone : +1-847-632-3028
mél : Qiaobing.Xie@motorola.com

13. Déclaration complète de droits de reproduction

Remerciement
Le financement de la fonction d'édition des RFC est actuellement assuré par la Internet Society.