L'acoustique dans les télécommunications

Rozenn Nicol & Al.
France Télécom Recherche & Développement
Technopole Anticipa
2, avenue Pierre Marzin
22307 Lannion CEDEX
E-mail : Rozenn.Nicol@orange-ftgroup.com

Même si le domaine des télécommunications est en constante évolution, et traverse régulièrement des phases de profondes mutations comme ces dernières années, à la fois en termes de technologies, de terminaux, d'usages ou de périmètres, les problèmes fondamentaux dans lesquels l'Acoustique vient mettre son grain de sel restent relativement pérennes. Ils couvrent un large domaine allant de l'acoustique physique, en passant par l'électro-acoustique et le traitement du signal jusqu'à la psychoacoustique. L'article suivant en présente les principaux thèmes.

Production, description et perception de la parole

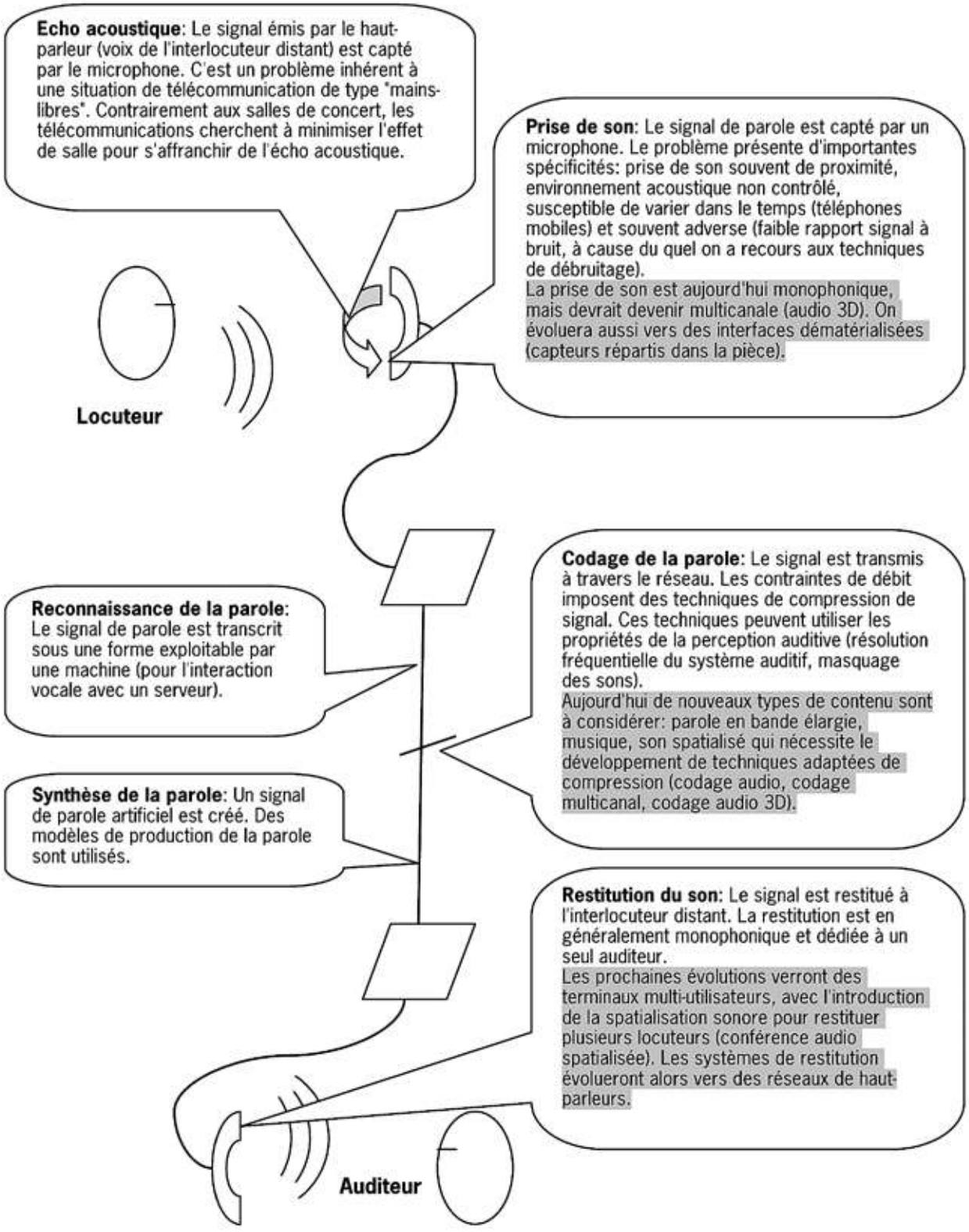
Une bonne compréhension de la production de la parole au niveau physiologique est indispensable pour permettre une description adéquate du signal de la parole en ce qui concerne l'enchaînement et l'anticipation du mouvement des articulateurs. Ces connaissances sont exploitées dans les différents modules des technologies de traitement de la parole, utilisées dans les services vocaux, comme la reconnaissance automatique de parole ou la synthèse de parole. La reconnaissance de la parole intègre ces connaissances au niveau de la modélisation acoustique des phonèmes alors que la synthèse de parole les utilise au moment de la sélection des unités acoustiques les plus appropriées pour une chaîne phonétique donnée. Les paramètres prosodiques sont utilisés par les locuteurs pour structurer linguistiquement la chaîne sonore. Les connaissances sur l'utilisation linguistique de ces paramètres servent pour leur modélisation en vue d'aboutir à une segmentation du signal de parole en unités de sens plus faciles à interpréter syntaxiquement et sémantiquement par un système de traitement automatique de la parole. Quant à la synthèse de la parole, les évolutions des paramètres prosodiques sur une phrase, fournies par le module linguistique, deviennent des paramètres de sélection des unités à concaténer.

Reconnaissance de la parole

La reconnaissance de la parole est la tâche qui consiste à reconnaître dans le signal de parole les phonèmes ou les mots prononcés. Dans un contexte de télécommunications, il s'agit principalement de dialoguer avec un serveur vocal automatique. Dans ce cas, le défi est de pouvoir reconnaître la parole de n'importe quel locuteur (reconnaissance multi-locuteurs) qui s'exprime spontanément : il s'agit ainsi d'être robuste aux divers accents, modes d'élocution, disfluences, dans différents environnements acoustiques. On parle de reconnaissance de parole continue lorsqu'on cherche à reconnaître tous les mots d'un locuteur qui s'exprime de façon naturelle. A ce dessein, on combine une modélisation au niveau acoustique (sur la réalisation des phonèmes) et une modélisation linguistique (sur l'enchaînement plus ou moins probable des mots du vocabulaire). Même si les sorties du module de reconnaissance de parole sont entâchées d'erreurs, elles sont aujourd'hui exploitables et interprétables par un module d'interprétation, pour des tâches applicatives aux contours bien définis, pour permettre un dialogue vocal qui semble «naturel» à l'utilisateur.

Synthèse de la parole

Parmi les technologies de synthèse par concaténation, la synthèse par corpus (SPC) est devenue un standard. Elle repose sur l'exploitation d'un corpus de parole capturant l'univers de production d'un locuteur. Le principe est de sélectionner la séquence de segments acoustiques la plus adaptée au contexte de synthèse. Ce faisant un minimum de traitement est opéré et de ce fait le naturel de la voix originale est préservé. Cette technologie repose néanmoins sur une bonne maîtrise de l'acoustique, d'une part en amont pour caractériser acoustiquement les segments acoustiques à sélectionner et d'autre part en aval, pour



effectuer des traitements correctifs sur le signal produit. Signalons également que les techniques de modification de signaux permettent de modifier le style, le timbre voire l'identité de voix perçue et sont donc d'un grand intérêt pour la personnalisation d'une brique technologique de synthèse vocale.

Codage de la parole et codage audio

Le codage audio vise à établir une représentation du signal adaptée à une transmission efficace ou un stockage à dimension réduite des échantillons audionumériques. Bien que les réseaux de télécommunications et les systèmes de

stockage voient leurs capacités s'accroître, les procédés de codage audio suscitent toujours autant d'intérêt. En effet, les applications à débit contraint (communication mobile par exemple) ou encore les applications en temps réel (streaming audio/vidéo sur internet) nécessitent une transmission du signal à bas débit. Les techniques de codage s'appuient principalement sur deux modèles. Le premier, appelé CELP pour Code Excited Linear Prediction, est basé sur un modèle de production de la parole, et tente de reproduire le souffle, la vibration des cordes vocales et le conduit vocal (bouche, nez). Il obtient de très bonnes performances à bas débits et est très largement utilisés dans les codeurs normalisés pour les applications mobiles et de VoIP (Voice over IP pour voix sur réseau IP). Le second modèle de codage est appelé codage perceptuel ou codage par transformée. Il s'appuie sur une transformation temps-fréquence et un modèle psychoacoustique permettant d'utiliser les effets de masquage présents au niveau du système auditif pour réduire le débit nécessaire à la transmission du signal utile. De nombreux codeurs audio, comme le mp3, l'AAC ou le Dolby Digital, largement utilisés actuellement, sont construits sur ce modèle. De nouvelles techniques de codage bas débits sont récemment apparues. Elles sont basées sur une représentation paramétrique des hautes fréquences du signal audio. Une de ces techniques, appelée SBR (Spectral Band Replication) a été normalisée pour donner le standard HE-AAC, connu aussi sous le nom de Enhanced AAC+.

Electroacoustique : transducteurs de prise et de restitution du son

Le microphone et le haut-parleur (ou plus spécifiquement l'écouteur pour la téléphonie) sont deux maillons essentiels de la chaîne dont les performances jouent de façon déterminante sur la qualité globale. Aucun traitement, aussi sophistiqué soit-il, ne pourra éliminer parfaitement les défauts d'un transducteur. Une contrainte spécifique au contexte des télécommunications est la minimisation de la taille des transducteurs et l'intégration dans des terminaux où l'acoustique entre souvent en conflit avec d'autres impératifs comme l'usage et l'ergonomie. Aujourd'hui les exigences conjointes d'extension de la bande passante et de miniaturisation des terminaux posent de nouveaux défis à la conception des transducteurs.

Fig. 1 : Mesure de caractérisation d'un terminal mobile

Traitements à la prise de son

La prise de son est le point d'entrée d'une succession de traitements qui conduit le signal vers l'interlocuteur distant. S'il paraît alors indispensable de soigner cette interface, l'expérience montre qu'elle fait rarement partie des exigences du cahier des charges d'un terminal et doit souvent s'adapter à des contraintes de design, compacité, etc. La tendance actuelle à la dématérialisation des terminaux, et donc à la généralisation de la communication «mains-libres», fait apparaître des artefacts qui, s'ils étaient négligeables en prise de son de proximité, deviennent problématiques :

- l'**écho acoustique** provoque une forte gêne encore accrue par le délai introduit par les nouveaux modes de transmission de voix sur IP ;
- la **réverbération** et le **bruit ambiant**, particulièrement perceptibles du fait de l'éloignement du locuteur vis-à-vis du microphone, réduisent fortement l'intelligibilité pour le locuteur distant.

Ainsi, différentes solutions sont mises en œuvre pour réduire ces artefacts :

- Des **traitements d'annulation d'écho acoustique** suppriment tout ou partie de l'écho, mais diminuent parfois très fortement l'interactivité de la communication.
- L'emploi **d'algorithmes de débruitage** permet de réduire les bruits ambiants, mais peut dégrader de manière audible et désagréable le signal utile.

Enfin, la prise son multicapteur qui propose une représentation spatiale de la scène (voir la partie sur les «Technologies de spatialisation sonore») peut également tirer parti de la diversité spatiale des sources pour améliorer le rapport signal à bruit et réduire les artefacts (bruit et réverbération), mais elle impose généralement des dimensions peu compatibles avec la compacité des terminaux. En clair, tout est affaire de compromis entre le désir de communication «sans contrainte» et la qualité de la communication.

Évaluation objective et subjective de la qualité audio

Dans le domaine des télécommunications, le terme de qualité audio caractérise généralement l'impact de l'ensemble de la chaîne de transmission sur le signal audio original, du terminal d'émission au terminal de réception, en passant par le codage/décodage et le canal de transmission. On s'intéresse donc principalement à détecter et à quantifier d'un point de vue perceptif les altérations du signal original par le système (bruit, distorsion, réduction de bande

passante, coupures, etc.). Pour évaluer la qualité audio dans le domaine des télécommunications, deux types de méthodes sont considérées : les méthodes dites «objectives» ou instrumentales basées sur l'utilisation d'instruments de mesure qui captent et analysent le signal dans ou en sortie de réseau, ou intègrent des paramètres du réseau, pour prédire une note de qualité. La validation de ces outils est basée sur la mesure dite «subjective» ou perceptive, faisant intervenir des utilisateurs qui évaluent la qualité des séquences audio et/ou vidéo présentées ou des communications effectuées. La rapide évolution des technologies nécessite sans cesse d'adapter voire de renouveler les méthodologies de tests perceptifs et les modèles de prédiction qui en découlent. Une autre piste est d'étudier la qualité audio non plus à travers les jugements de qualité effectués de façon explicite par les sujets mais à travers l'impact de la qualité des médias sur le comportement, l'état émotionnel et la satisfaction des utilisateurs.

Autrefois centrées sur la parole, les études s'étendent à d'autres contenus comme la musique, les films. La spatialisation des sons est une autre nouveauté. Enfin, en termes d'interfaces et des interactions, on évolue vers des technologies transparentes à l'utilisateur qui posent de nouveaux défis notamment au niveau des transducteurs et des traitements à la captation et à la restitution. Ces évolutions ont conduit à de nouveaux thèmes.

Technologies de spatialisation sonore

Dans notre expérience naturelle d'écoute, nous percevons le monde sonore en 3D, c'est-à-dire que nous sommes capables d'identifier la position des sons à la fois en distance, en azimut et en élévation. Les technologies de spatialisation sonore visent à créer l'illusion de cette perception en 3D. En ajoutant une nouvelle dimension à la restitution du son, l'audio 3D permet d'introduire une vraie rupture dans les futurs services de communication (conférence audio spatialisée) en améliorant les interactions (intelligibilité, naturel, immersion) et en enrichissant les contenus. Le monde du grand public (home cinema, jeux sur PC…) est dominé par les formats multicanal (5.1, 6.1, 7.1, 10.2, 22.2) dont la stéréophonie n'est qu'un cas particulier (format 2.0 ou 2.1). Des technologies émergentes, comme le binaural, l'holophonie ou WaveField Synthesis (WFS), Ambisonics ou Higher Order Ambisonics (HOA), proposent une expérience enrichie de la spatialisation sonore. Le binaural vise à imiter les mécanismes de la localisation auditive (reproduction des indices de localisation au niveau des tympans), tandis que l'holophonie et Ambisonics se fondent sur une reconstruction de l'onde acoustique. L'introduction de la spatialisation impose de concevoir de nouveaux systèmes de prise et de restitution du son afin de capter et de reproduire l'information spatiale. Un autre défi est de développer des techniques d'annulation d'écho acoustique adaptées à des dispositifs multicanal.

Fig. 2 : Prototype de microphone HOA pour la captation d'une scène audio 3D

Évaluation subjective des systèmes de spatialisation sonore

L'évaluation des technologies audio 3D nécessite la définition et la mise en œuvre de nouvelles méthodologies pour rendre compte de la dimension spatiale. Pour les contenus multicanal (voir plus haut), des méthodes normalisées existent (MUSHRA). Dans le cas des contenus véritablement 3D, les tests de localisation restent une référence incontournable, mais limitée, car ils ne témoignent que de la précision de la localisation des sons. Des tests basés sur le jugement d'attributs sémantiques (enveloppement, précision, naturel, largeur des sources…) permettent de compléter l'évaluation de la qualité spatiale. Enfin, des évaluations indirectes, pour lesquelles la qualité est évaluée à partir du succès ou de l'échec du sujet à effectuer une tâche, constituent une nouvelle piste prometteuse. C'est notamment une opportunité pour évaluer les technologies dans le contexte de services applicatifs.

Codage audio spatial

Ce domaine concerne les techniques de compression appliquées aux flux audio 3D. Les premières études ont porté sur le codage du format multicanal (5.1) avec d'abord les techniques de matriçage comme le Dolby Surround dans les années 80 et plus récemment les Dolby Pro Logic, puis l'extension des codeurs audio traditionnels (monophoniques) à un nombre de canaux supérieur à deux (normes MPEG, mp3 et AAC). En parallèle, des solutions propriétaires alternatives sont apparues, telles que DTS et Dolby Digital (ou AC-3) pour le format 5.1 dans les applications grand public (par exemple pour le DVD à des débits de 384 kbit/s pour le Dolby Digital et 1.4 Mbit/s pour le DTS). Enfin, des techniques plus évoluées et complètement dédiées à l'audio 3D sont apparues à partir de la fin des années 90 avec le **codage paramétrique** permettant de proposer des schémas de codage bas débit pour des flux stéréophoniques

Il n'est pas facile de chiffrer le nombre d'emplois associés à l'Acoustique dans les Télécommunications. Outre les travaux menés par les opérateurs historiques (Bell Labs, NTT, Orange Labs, Deutsch Telekom, British Telecom…), il faut aussi compter sur de nombreux partenariats avec les laboratoires universitaires ou d'autres industriels (transducteurs par exemple) dont la contribution est difficile à quantifier.

et multicanal de type 5.1. Dans ces codeurs, un *downmix* (composé habituellement de un ou deux canaux) est construit à partir du signal audio original, puis codé par un codeur traditionnel (mp3, AAC, HE-AAC), et enfin transmis parallèlement avec des paramètres d'information spatiale. Le procédé Binaural Cue Coding (BCC) en est un premier exemple qui exploite les propriétés de la perception spatiale du son par l'extraction de paramètres spatiaux liés à la localisation auditive (Inter Channel Time Difference ou ICTD pour les retards entre les canaux, Inter Channel Level Difference ou ICLD pour les différences d'énergie entre canaux, Inter-Channel Coherence ou ICC pour la corrélation entre les canaux). La principale évolution à venir du codage audio spatial portera certainement sur le codage des futurs formats audio 3D (notamment HOA).

Conversion de formats audio 3D

Lorsqu'on écoute un contenu multicanal 5.1 sur un système stéréophonique, on perd l'immersion et les effets introduits par les canaux arrière, ainsi que la stabilité de l'image sonore frontale apportée par le haut-parleur central. De même, quand on écoute un contenu multicanal 5.1 au casque, la scène sonore reste enfermée entre les deux oreilles. Les techniques de conversion de formats visent à réduire ou éliminer ces inconvénients. Face à la multiplicité des formats, un premier objectif est l'**adaptation** des contenus au système de restitution, étant donné qu'il n'est en général pas envisageable pour un utilisateur d'avoir à disposition plusieurs systèmes de restitution. Une seconde finalité est l'**enrichissement** des contenus, afin de rajouter une dimension et de favoriser l'immersion dans la scène sonore (traitements d'élargissement stéréophonique sur enceintes pour donner l'impression que le son sort de deux haut-parleurs virtuels qui seraient plus écartés que les haut-parleurs réels, ou sur casque pour externaliser les sources sonores en simulant des haut-parleurs virtuels par synthèse binaurale). Les outils disponibles aujourd'hui concernent essentiellement la conversion entre les formats stéréophonique et multicanal : techniques de ***upmix*** et techniques de ***downmix***, selon que le système de restitution comprend plus (*upmix*) ou moins (*downmix*) de canaux que les données reçues. ■

ERRATUM

Rozen Nicol de Orange Labs nous demande de vous préciser que les coordonnées que nous avons publiées au début de l'article qu'il a rédigé avec ses collègues du laboratoire d'acoustique dans le numéro 52 Acoustique & Techniques ne sont pas exactes. Si vous souhaitez les contacter, nous vous prions de bien vouloir noter cette adresse :

Orange Labs
TECH/SSTP/TPS
2, avenue Pierre Marzin
22307 Lannion CEDEX
E-mail: rozenn.nicol@orange-ftgroup.com

Résumés de l'article «Le son 3D dans toutes ses dimensions» de Rozen Nicol paru dans le numéro 52 d'Acoustique & Techniques

Résumé
Les technologies de spatialisation sonore ou son 3D proposent une nouvelle expérience de la restitution du son. Nous commençons à les découvrir dans le domaine des jeux ou des systèmes home cinema, avec notamment le son multicanal 5.1. Cet article présente un état des lieux des techniques de spatialisation existantes en faisant la distinction entre ce qui est d'ores et déjà accessible au grand public (stéréophonie, multicanal 5.1) et les technologies du futur qui pourraient bientôt faire vibrer nos haut-parleurs (binaural, holophonie, Ambisonics). Au-delà des aspects liés à la captation et la restitution d'une scène audio 3D, sont abordés les nouveaux enjeux que sont l'émergence et la définition des formats audio 3D, le développement de schémas de codage spécifiques aux flux audio 3D, et la capacité à convertir un format donné en un autre format.

Abstract
Technologies of sound spatialization or 3D audio provide a new listening experience, which we discover in daily life; in particular, through the extensive use of the standardized multi-channel 5.1 format within games or home cinema systems. This article presents an overview of state-of-the-art spatialization techniques used for consumer products (stereophony, multichannel 5.1) and discusses future technologies, which could be available soon (binaural, holophony, Ambisonics). Beyond these aspects related to sound capture and 3D audio rendering, new coding standards are required, i.e. the definition of 3D audio formats, the development of coding schemes specific to 3D audio stream, and the capacity to convert a given format into another format.