Enrichissement de la conférence audio en voix sur IP au travers de l'amélioration de la qualité et de la spatialisation sonore

Arnault Nagle

TELECOM ParisTech

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE TELECOM ParisTech

Spécialité : « Signal et Images »

préparée au laboratoire 'Speech & Sound Technologies & Processing' de France Télécom R&D

dans le cadre de **l'Ecole Doctorale** *« INFORMATIQUE, TELECOMMUNICATIONS ET ELECTRONIQUE »*

présentée et soutenue publiquement

par

Arnault NAGLE

le 19 février 2008

ENRICHISSEMENT DE LA CONFERENCE AUDIO EN VOIX SUR IP AU TRAVERS DE L'AMELIORATION DE LA QUALITE ET DE LA SPATIALISATION SONORE

JURY

M. Nicolas MOREAU,	Professeur TELECOM ParisTech	Président
Mme Annie GRAVEY,	Directeur d'Etudes TELECOM Bretagne	Rapporteur
Mme Régine LE BOUQUIN JEANNES,	Professeur Université Rennes 1	Rapporteur
M. Dirk SLOCK,	Eurecom	Directeur de thèse
M. Pascal SCALART,	Professeur ENSSAT	Examinateur
M. Nicolas TSINGOS,	INRIA	Examinateur
M. Olivier HERSENT,	Chief Technology Officer NetCentrex	Invité
M. Aurélien SOLLAUD,	Ingénieur R&D France Télécom	Co-encadrant

RÉSUMÉ

Ce travail de thèse s'intéresse à la conférence audio en Voix sur IP et plus précisément à son enrichissement au travers de l'amélioration de la qualité (bande élargie et traitement d'amélioration) et de la spatialisation sonore. Cette évolution de la conférence audio sera examinée à la lumière des architectures centralisée et distribuée de la conférence sur IP standard. L'objectif est d'étudier des solutions en termes d'architecture intégrant la spatialisation et des extensions pour gérer et contrôler cette spatialisation. Il conviendra aussi d'effectuer les tests montrant les qualités audio et de spatialisation résultantes.

Notre première contribution a été de proposer des architectures permettant d'allier la conférence audio en voix sur IP, les méthodes de spatialisation retenues, les terminaux ou pont de conférence ainsi que les traitements d'améliorations connus (annulation d'écho ou de bruit, etc.). Nous avons montré qu'il semblait difficile par exemple d'utiliser conjointement la spatialisation et la commutation de flux. Par contre la solution utilisant un pont mixeur ne présentait pas d'inconvénient pour inclure la spatialisation tout en conservant l'ensemble des traitements de qualité. Par cette configuration, nous garantissons en plus l'interopérabilité avec les réseaux voix existants. Les solutions distribuées sont tout autant réalisables dans la théorie mais pâtissent actuellement des limites des terminaux. Pour une interopérabilité avec les autres réseaux voix, il est en plus nécessaire d'avoir une entité de mixage pour créer un contenu monophonique. Par la suite, nous avons souligné les avantages et les inconvénients de l'utilisation de pont de conférence de type mixeur et de type répliquant pour proposer une solution de pont mixte. Cette solution fonctionne tantôt en mode répliquant, tantôt en mode mixeur suivant les capacités des terminaux. Par ailleurs, nous avons validé une méthode de réduction de la bande passante d'un pont répliquant vers un terminal, basée sur le masquage auditif.

La seconde contribution de nos travaux consiste en de nouvelles solutions protocolaires adaptées à la gestion et au contrôle de la spatialisation. Nous avons donc défini les extensions nécessaires à la gestion et au transport du son spatialisé. Nous avons tout d'abord défini les spécifications permettant de commander un positionnement de locuteurs dans une conférence audio spatialisée. Nous avons cherché à présenter toutes les solutions possibles pour les gestions automatique ou manuelle. Dans le cas de la spatialisation sur un pont de conférence, nous avons ainsi souligné le fait que cette gestion ne pouvait se faire par l'intermédiaire du protocole SIP, car ce n'est pas le rôle de ce dernier de transporter dans ses messages des informations sur les contenus ou des commandes de spatialisation. Nous avons proposé une solution basée sur ce qui se fait dans les conférences audio standard : une solution de web-pilotage certes propriétaire à chaque fournisseur de services mais en cohérence avec la gestion des protocoles de Voix sur IP.

Pour la conférence avec un pont mixeur, nous avons établi les paramètres du protocole de signalisation SIP nécessaires au transport de flux asymétriques tout en garantissant une interopérabilité avec les terminaux existants. La nécessité de transporter ces flux asymétriques est due à notre hypothèse de départ concernant l'équipement des terminaux : prise de son monophonique et restitution sur casque ou deux haut-parleurs.

Notre troisième contribution s'exprime au travers d'une campagne de tests pour valider nos solutions en termes de qualité audio et de qualité de spatialisation. Ces tests nous ont amené à définir des nouveaux protocoles adaptés à ces architectures audio spatialisées. Nous justifions dans un premier temps nos choix de codeurs et dans un second temps nos choix de tests. Nous avons montré dans un premier temps que les codeurs n'étaient pas perçus de la même façon suivant que l'on écoute en écoute monaurale ou en écoute diotique. Il ressort de ces tests que les codeurs G.711 (PCM) et G.722 (ADPCM) sont les plus adaptés à la conférence audio centralisée avec une qualité jugée nettement supérieure aux codeurs CELP. Ces deux codeurs sont de plus

de faible complexité, robustes au transcodage, à la perte de trames, au transport de contenu binaural et au transport de contenus multi-locuteur. Quant aux codeurs CELP, ils sont à utiliser uniquement lorsque les contraintes de débit sont fortes.

Concernant la conférence audio distribuée wideband, les codeurs AMR-WB à 23.85 kbits/s, G.729.1 à 32 kbits/s et G.722 à 64 kbits/s semblent les plus adaptés quelle que soit la perte de trames. Ils ont une qualité jugée équivalente. En narrowband, les codeurs G.711, AMR à 12.2 kbits/s et G.729.1 à 12 kbits/s obtiennent les meilleures notes de qualité, quelle que soit la perte de trames. Au final, dans tous les cas distribués, le choix du codeur dépendra des contraintes de l'application suivant un compromis complexité/débit.

Mots clés : Conférence audio, Son spatialisé, Architectures, Codeurs, dual-mono, Narrowband, Wideband, SIP, Tests de qualité, Tests de spatialisation, Pont de conférence, Terminal.

ABSTRACT

This thesis deals with audio conferencing over IP and its improvement through high quality and 3D sound. Our goal is to develop solutions enabling the merging of well-known architectures such as the centralized or the loosely coupled ones, techniques which are likely to impact quality (Voice Activity Detection, Noise Reduction …) and 3D sound. We have to define the controls to manage 3D audio conferencing according to each architecture. At last, quality tests and tests about spatialization must be performed to validate our solutions.

The first axis of this thesis is looking further into those current architectures in order to propose solutions integrating 3D sound and improvement techniques such as Noise Reduction, Acoustic Echo Cancellation, etc. Thanks to its interoperability with standard voice networks, we recommend the spatialized centralized architecture for a fast deployment. Concerning the spatialized distributed architecture, its implementation is today limited by the capabilities of the terminals. We also propose a new architecture based on the well-known architectures of VoIP conferencing (forwarding bridge or mixing bridge). This architecture optimizes the use of conferencing bridges and terminals. At last, we validate a method to reduce the bitrate between a forwarding bridge and terminals based on on-the-fly auditory masking.

The second axis of our research relies on the definition of the controls enabling the management of the audio conferencing. We define the necessary extensions to control the positions of each participant in the audio conferencing according to each architecture.

We specify the needs for the protocol SIP in order to manage asymmetric streams and to enable the interoperability with existing terminals.

Our third axis deals with quality tests and tests about spatialization in order to validate the dual-mono coding method and select the most appropriate coders. First, we had to define new protocols to perform our tests and we justify our choice of coders and tests. Next, we prove that the monaural hearing and the diotic hearing are not equivalent. At last, coders G.711 and G.722 are the most suitable for the centralized audio conferencing with a high audio quality and high quality of spatialization compared the CELP coders. They have low-complexity, and are robust to packet losses, multi-talker, 3D sound and tandeming.

For the wideband loosely coupled architecture, AMR-WB at 23.85 kbits/s, G.729.1 at 32 kbits/s, and G.722 at 64 kbits/s seem to be the best coders whatever the packet losses are. In narrowband, G.711, AMR at 12.2 kbits/s, and G.729.1 at 12 kbits/s are the best ones. Coders have to be chosen according to the bitrate and complexity constraints.

Keywords : Audio conferencing, 3D sound, Architectures, Coders, Dual-mono, Narrowband, Wideband, SIP, Quality tests, Tests about spatialization, Conferencing bridge, Terminal.

REMERCIEMENTS

Le travail de thèse, qui est rapporté dans le présent document, a été réalisé au sein de l'Unité de Recherche et Développement *Interfaces Sonores Innovantes* (ISI) du laboratoire *Speech and Sound Technologies & Processing* (SSTP) de France Télécom Recherche et Développement.

Je voudrais remercier tout d'abord Jean-Pierre Petit de m'avoir admis au sein du laboratoire SSTP. Je remercie également sincèrement Yannick Mahieux de m'avoir accueilli au sein de l'équipe ISI et d'avoir régulièrement pris le temps de suivre l'avancée de mes travaux.

Un très grand merci à toi, Aurélien, pour ton investissement au cours de cette thèse, ta disponibilité, ta capacité à simplifier les problèmes et surtout ton soutien. Sans toi, cette thèse n'aurait jamais été au bout. Merci aussi pour tout ton travail de relecture qui a été conséquent au cours de ces 3 années passées ensemble !!

Merci à Dirk Slock d'avoir accepté d'être mon Directeur de Thèse.

Merci à tous les membres du jury d'avoir participé à l'évaluation de ce travail ! Chacun de vous était spécialiste d'un domaine particulier de cette thèse et a dû faire un effort important pour se familiariser avec les autres volets.

Un grand merci à toutes les personnes avec qui j'ai pu travailler au cours de ces 3 années parmi lesquels : Catherine, Greg, Anne, David, Rozenn, Pierre, Marc, Laëtitia, Martine, Manu, Jean-Phi, José, Alex, Jérôme, etc.

Un merci tout particulier à mes voisins de bureau au cours de ces 3 ans : Stouph, Jean-Claude et le strapontiste violoniste ! Grâce à vous, beaucoup d'excellents moments à rigoler dans le bureau, de jeux de mots en tout genre, de pétages de plombs, etc. !! Pour ceux qui liraient cette partie, je confirme que nous arrivions à bosser … de temps en temps !!

Merci beaucoup aussi à la bande de doctorants, post-docs, stagiaires et apprentis du couloir parmi lesquels : Ludo, Titi, Flo, Charly, Zouzou, les Mathieu(s), Adrien, Nico, Manu, Issoufou et Marie. Grâce à eux, les pauses avaient une saveur particulière et permettaient vraiment de penser à autre chose ☺. Merci aussi aux adeptes du squash et du badminton pour le défoulement que cela procure !

Merci aux équipes ISI et TPS pour la bonne ambiance présente tout au long de l'année ! Merci à Noëlle, Annie et Serge pour l'aide qu'ils apportent à longueur de temps !

Merci à tous mes relecteurs et notamment Laëtitia ! Un conseil pour ceux qui voudraient qu'elle relise un document, cela est très facile en l'échange de nounours en chocolat … non périmés …

Je remercie ma famille et mes amis pour leur soutien pendant ces 3 ans.

Pour terminer, comment ne pas remercier celle qui par sa présence et son amour m'a aidé à finir cette thèse… duō xiè bǎo bǎo !

A tous, un grand merci !

Nono.

P.S : Merci à word de m'avoir laissé rédiger ma thèse sans trop de problèmes.

P.S.1 : Merci à Manu pour son template word qui m'a beaucoup aidé !

TABLE DES MATIÈRES

ACRONYMES

AAC	Advanced Audio Coding
AAC-LD	Advanced Audio Coding Low-Delay
ACELP	Algebraic Code Excited Linear Prediction
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
ADSL	Asymmetric Digital Subscriber Line
AEC	Acoustic Echo Cancellation
AGC	Automatic Gain Control
AMR	Adaptive Multi-Rate
AMR-WB	Adaptive Multi-Rate WideBand
ANOVA	ANalysis Of VAriance
APA	Affine Projection Algorithms
AVP	Audio Video Profile
CELP	Code Excited Linear Prediction
CNG	Comfort Noise Generation
CPU	Central Processing Unit
CRM	Coordinate Response Measure
dB	deciBel
dB SPL	deciBel Sound Pressure Level
DCR	Degradation Category Rating
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
DTX	Discontinuous Transmission
DSL	Digital Subscriber Line
ETSI	European Telecommunications Standards Insitute
FEC	Frame Error Correction
FPS	Flower Power Shooter
HP	Haut-Parleur
HRTF	Head-Related Transfer Function
HTTP	Hypertext Transfer Protocol
Hz	Hertz
ICE	Interactive Connectivity Establishment
ILD	Interaural Level Difference
INRIA	Institut National de Recherche en Informatique et en Automatique
IP	Internet Protocol
ITD	Interaural Time Difference
ITU	International Telecommunication Union
kbits/s	kilo-bits par seconde
LSF	Line Spectral Frequencies
MDCT	Modified Discrete Cosine Transform
MIRS	Modified Intermediate Reference System
MPS	MPEG Surround
MSE	Mean Square Error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NAPTR	Naming Authority Pointer
NAT	Network Address Translation
NB	NarrowBand
NLMS	Normalized Least Mean Squares
OPERA	Optimisation PErceptive du Rendu audio

PAR	Positive Acknowledgment Retransmission
PCM	Pulse Code Modulation
PdT	Perte de Trame
PLC	Packet Loss Cancealment
RLS	Recursive Least Squares
RNTL	Réseau National de recherche et d'innovation en Technologies Logicielles
RTC	Réseau Téléphonique Commuté
RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol
RXIRS	Reception Intermediate Reference System
SAC	Spatial Audio Coding
SAOC	Spatial Audio Object Coding
SCTP	Stream Control Transmission Protocol
SIP	Session Initiation Protocol
SPL	Sound Pressure Level
SRV	SeRVice
STUN	Session Traversal Utilities for NAT
SWB	Super WideBand
TCP	Transport Control Protocol
TFSS	Tandem-Free Speaker Selection
TURN	Traversal Using Relays around NAT
UDP	User Datagram Protocol
VAD	Voice Activity Detection
VCET	Voice Communication Effectiveness Test
VoIP	Voice over IP
WB	WideBand
WMOPS	Weighted Millions of Operations Per Second
XML	eXtensible Markup Language
3GPP	3rd Generation Partnership Project

Introduction

Contexte et enjeux

Les services conversationnels de personne à personne ou de groupe connaissent une rupture du fait de l'arrivée de la Voix sur IP (VoIP - Voice over IP en anglais). Son développement rapide est en train de mettre au second plan la téléphonie standard du réseau téléphonique commuté.

Cette rupture s'exprime par l'apparition de nouveaux acteurs sur le marché des télécommunications. La Voix sur IP est en train de bouleverser les positions acquises depuis de nombreuses années par les grands opérateurs historiques de chaque pays. De nombreuses possibilités de communication gratuites sont apparues telles que Skype permettant notamment des communications de personne à personne voire des conférences. Pour tout fournisseur de service, il devient nécessaire de se distinguer de la concurrence.

Parmi les avantages de la Voix sur IP sur la téléphonie standard, citons notamment la convergence du réseau de données et du réseau supportant la voix, aboutissant à l'émergence de nouveaux services voix-données. On peut, par exemple, améliorer l'efficacité des appels en couplant la téléphonie avec un système de gestion de présence et de disponibilité. Par ailleurs, soulignons le développement de la web conférence mélangeant le partage de documents et la conférence audio.

Pour un utilisateur en train de communiquer qu'il soit un particulier ou un professionnel, l'apport de la Voix sur IP se traduit par une réduction des coûts de communication, et aussi une possibilité d'amélioration de la qualité audio. En effet la flexibilité de la VoIP permet une extension de la bande passante du signal audio et du nombre de canaux (stéréophonie voire au-delà).

L'utilisation de cette capacité de transport de contenu multicanal autorise la prise en compte de la dimension spatiale du son. Cette dimension vise à rapprocher encore plus un utilisateur du contexte naturel d'une communication en face à face que l'on peut avoir par exemple autour d'une table de réunion.

La Voix sur IP propose donc des perspectives en termes de nouveaux services et d'enrichissement de l'expérience utilisateur. Un exemple de tel service innovant, qui est l'objet de ce document de thèse, est la conférence audio spatialisée.

Problématique, orientations et plan de la thèse

Le sujet de la thèse porte sur la conférence audio en Voix sur IP et plus précisément sur son enrichissement au travers de l'amélioration de la qualité (bande élargie et traitement d'amélioration) et de la spatialisation sonore.

Tout système de télécommunication se caractérise par une partie média et une partie signalisation. A la lumière des différentes architectures possibles de la conférence sur IP, il convient d'établir de quelle manière le traitement de spatialisation impacte ces deux parties.

Au niveau média, il est nécessaire de s'interroger sur la possibilité d'intégrer les traitements d'amélioration de la qualité et sur l'emplacement de la spatialisation dans la chaine de traitement audio. La VoIP intègre classiquement des codeurs normalisés monophoniques, dans le but de réduire le débit réseau. Se pose alors la question de l'impact sur la qualité audio de la juxtaposition de la spatialisation et du codage. En effet, le signal spatialisé étant par essence multicanal, l'utilisation de codeurs monophoniques appliqués indépendamment sur chacun des canaux peut avoir un effet néfaste sur la qualité.

Au niveau signalisation, les protocoles actuels de VoIP ne prennent pas en compte la dimension spatiale. Il faut d'une part pouvoir exprimer les capacités des terminaux en termes de rendu spatial et d'autre part contrôler la spatialisation principalement via un pilotage de la position virtuelle de chaque participant.

Les objectifs sont d'établir les différentes options possibles en termes d'architecture et de déterminer ce qu'il est nécessaire de développer pour proposer un service optimal de conférence audio, notamment pour la gestion de la spatialisation. Enfin il convient de ne jamais perdre de vue qu'il est nécessaire de garantir une qualité audio et une qualité de spatialisation les meilleures possibles ainsi qu'une interopérabilité avec les terminaux existants.

Le travail de cette thèse a tout d'abord consisté à analyser l'état de l'art d'un vaste domaine couvrant à la fois la prise de son jusqu'à sa restitution en passant par l'architecture réseau choisie, le codage de compression audio, la spatialisation, et d'éventuels traitements d'amélioration, etc. Le chapitre 1 introduit la conférence audio spatialisée en Voix sur IP en se basant sur les exemples de la conférence audio standard.

Dans ce contexte, la première orientation de nos travaux de recherche, décrite au chapitre 2, est de proposer des architectures permettant d'allier la conférence audio en voix sur IP, les méthodes de spatialisation retenues, les terminaux ou pont de conférence ainsi que les traitements d'améliorations connus (annulation d'écho ou de bruit, etc.). Des travaux complémentaires seront présentés concernant une technique d'optimisation de l'utilisation des ponts et des terminaux, ainsi qu'une technique de réduction de bande passante réseau entre un pont de conférence et un terminal.

La seconde orientation de nos travaux, décrite au chapitre 3, est d'établir au niveau protocolaire des solutions adaptées à la gestion et au contrôle de la spatialisation.

A partir des architectures présentées au chapitre 2, nous décrivons, au chapitre 4, l'ensemble de notre campagne de tests pour valider nos solutions en termes de qualité audio et de qualité de spatialisation. Nous justifierons dans un premier temps nos choix de codeurs et dans un second temps nos choix de tests. Nous conclurons en recommandant les codeurs les plus adéquats aux configurations retenues (centralisées ou distribuées), tout en mettant l'accent sur les besoins à terme.

1. Etat de l'art sur la conférence audio spatialisée en Voix sur IP

Ce chapitre introduit la conférence audio spatialisée en Voix sur IP en se basant sur les exemples de la conférence audio standard (sans spatialisation) des réseaux téléphonique commuté et à commutation de paquets. Son domaine d'application est vaste et ses nombreuses composantes s'étendent de la prise de son jusqu'à sa restitution en passant par l'architecture réseau choisie, le codage de compression audio, la spatialisation, voire d'éventuels traitements d'amélioration, etc.

Tout d'abord, nous expliciterons les termes *conférence audio* et nous présenterons par un exemple les parties signalisation et média d'un appel téléphonique en général. Nous exposerons ensuite les différentes composantes de la conférence audio en VoIP et nous soulignerons la flexibilité mais aussi les inconvénients qu'offre le monde de l'IP. Un exemple concret d'appel VoIP sera de même explicité, dans le cadre d'une conférence audio standard.

Par la suite, la conférence audio spatialisée sera introduite en insistant sur ses contextes d'utilisation et l'apport du son spatialisé. Les différentes briques de traitement de signal disponibles et choisies pour réaliser la spatialisation seront enfin décrites.

Pour conclure, l'étude de toutes ces composantes a permis d'établir les différents axes de recherche étudiés pendant cette thèse.

1.1 La conférence audio standard

Commençons par définir ce qu'est une conférence audio standard ainsi que les moyens de l'évaluer.

1.1.1 Définition de la conférence audio standard

La conférence audio standard, à comparer avec la conversation point à point où seuls deux participants sont présents, est un service basé sur un système de télécommunication permettant d'échanger de l'information au moyen de la voix entre plusieurs participants. Il est utile de rappeler que :

- La télécommunication est l'ensemble des transmissions, émission ou réception de signaux représentant des signes, des écrits, images, sons ou renseignements de toute nature, par fil, radioélectricité, optique ou autres systèmes électromagnétiques [45].

- Un système de télécommunication est généralement composé de terminaux émetteurs et récepteurs, d'un réseau de télécommunication et d'un équipement de routage au sein de ce réseau.
- Un service est défini par l'utilisation du susdit système pour réaliser une tâche spécifique [33].

Dans notre cas, nous transmettons la parole des différents participants à la conférence. Cette parole a pour objectif la communication entre différentes personnes, dont, en schématisant, au moins une a pour objectif de se faire comprendre et l'autre (ou les autres) de comprendre.

1.1.2 Evaluation de la conférence audio standard

Afin de déterminer l'acceptabilité du service proposé, les termes *qualité de service* sont employés. La qualité de service est l'effet global produit par la qualité de fonctionnement d'un service, celle-ci influençant le degré de satisfaction de l'usager du service [43]. Cette qualité de service s'évalue historiquement à travers quatre critères [43] :

- La logistique de service, qui est l'aptitude d'une organisation à fournir un service et à faciliter son utilisation.
- La facilité d'utilisation du service, qui est l'aptitude d'un service à être utilisé de façon satisfaisante et aisée par un usager.
- La servibilité du service, qui est l'aptitude d'un service à être obtenu à la demande d'un usager et à continuer d'être fourni, sans dégradations excessives, pendant la durée voulue, avec des tolérances spécifiées et d'autres conditions données.
- La sécurité du service, qui est la protection contre la surveillance clandestine, l'utilisation frauduleuse, les dégradations malveillantes, l'usage abusif, l'erreur humaine et les catastrophes naturelles.

Cependant, nous nous sommes basés sur une nouvelle définition de la qualité de service, qui semble plus appropriée à la VoIP. Elle a été proposée dans [62] et est illustrée Figure 1.1. Elle permet une intégration plus aisée des différentes contraintes liées à la VoIP comme la gigue (section 1.4.2.1), la perte de paquets (section 1.4.2.2), etc.

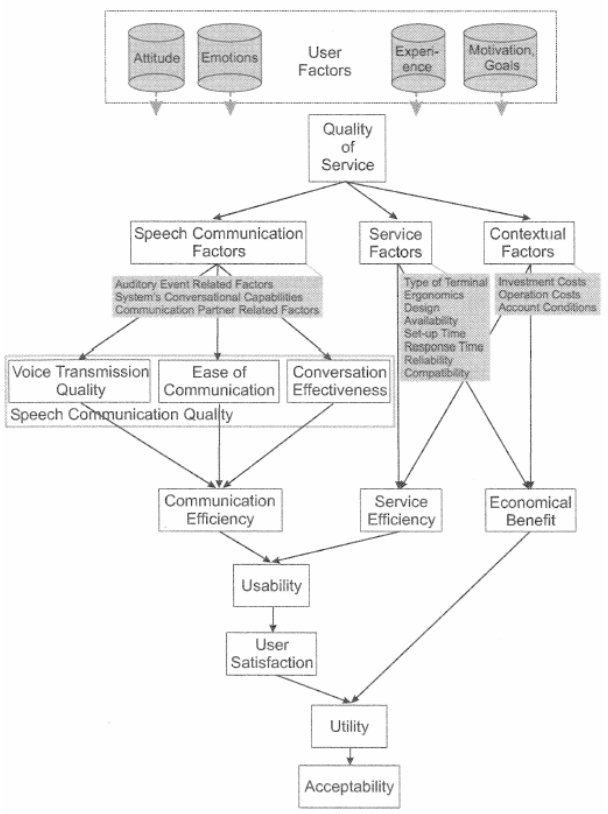

Figure 1.1 Qualité de service proposée dans [62]

La qualité de service se décompose cette fois en 3 éléments majeurs :

- Les facteurs relatifs à la communication de la parole (*Speech Communication Factors*) que sont la qualité de la transmission vocale (incluant l'effet de la gigue sur la qualité perçue, etc.), la facilité de la communication et l'efficacité de la conversation.
- Les facteurs de service (*Service Factors*) qui couvrent notamment la sécurité et la maintenance.
- Les facteurs contextuels (*Contextual Factors*) comme les coûts, les conditions de contrats, etc.

1.1.3 Présentation de la conférence audio du réseau téléphonique commuté

L'exemple de la conférence audio du réseau téléphonique commuté (RTC) permet d'introduire les notions de *partie signalisation* et de *partie média* existantes dans tout système de télécommunication

1.1.3.1 Introduction aux notions de "partie signalisation" et "partie média"

La partie signalisation permet la gestion de l'appel de sa création jusqu'à sa conclusion. Pour illustrer cette définition, un exemple est pris avec un appel standard effectué sur un téléphone du réseau RTC. Pour un appel de ce type, la partie signalisation correspond [26] :

- au décrochage du combiné téléphonique par l'appelant,
- à la génération de la tonalité invitant l'appelant à composer le numéro,
- à la composition du numéro permettant de trouver le correspondant désiré,
- à l'analyse par les commutateurs intermédiaires du numéro composé, afin de déterminer le chemin devant être parcouru par le contenu audio grâce à la signalisation SS7 [97] sur un réseau sémaphore (indépendant des données média),
- à faire sonner l'appareil du distant,

- à la mobilisation des ressources pour transporter les données audio, et
- à la gestion de la fin de l'appel par un des deux participants.

La partie média concerne, quant à elle, le transport de la voix d'un locuteur à un auditeur [38]. Concrètement, cela correspond généralement :

- dans un premier temps, à un transport de la voix entre le poste de l'abonné et le 1er commutateur (ou commutateur local) sur la boucle locale, grâce à une modulation de l'amplitude du courant dans la bande de fréquences 300 Hz à 3.4 kHz,
- à une transformation analogique vers numérique au niveau de ce commutateur local,
- à une compression du contenu numérisé par le codeur G.711 [42] loi A ou loi μ,
- à un transport de la voix, numérisée et compressée, entre différents commutateurs (le chemin ayant été établi par la signalisation) et en utilisant un multiplexage temporel (TDM, Time Division Multiplexing), grâce à des conducteurs métalliques (paires torsadées, câbles coaxiaux), à des faisceaux hertziens ou à des fibres optiques,
- et enfin, au niveau du commutateur final, à une décompression du contenu numérique puis à une transformation numérique vers analogique pour un envoi sur la boucle locale vers le destinataire final.

En guise de complément, le multiplexage temporel divise le flux d'information de chaque canal de transmission en blocs (généralement d'un octet) et juxtapose, les uns après les autres, les blocs de chaque conversation et cela de manière cyclique [38].

1.1.3.2 Exemple de conférence RTC

La Figure 1.2 illustre le principe de la conférence RTC classique. Les participants ont souhaité entrer en conférence. Le pont est l'entité qui va centraliser leurs appels et leur permettre de communiquer.

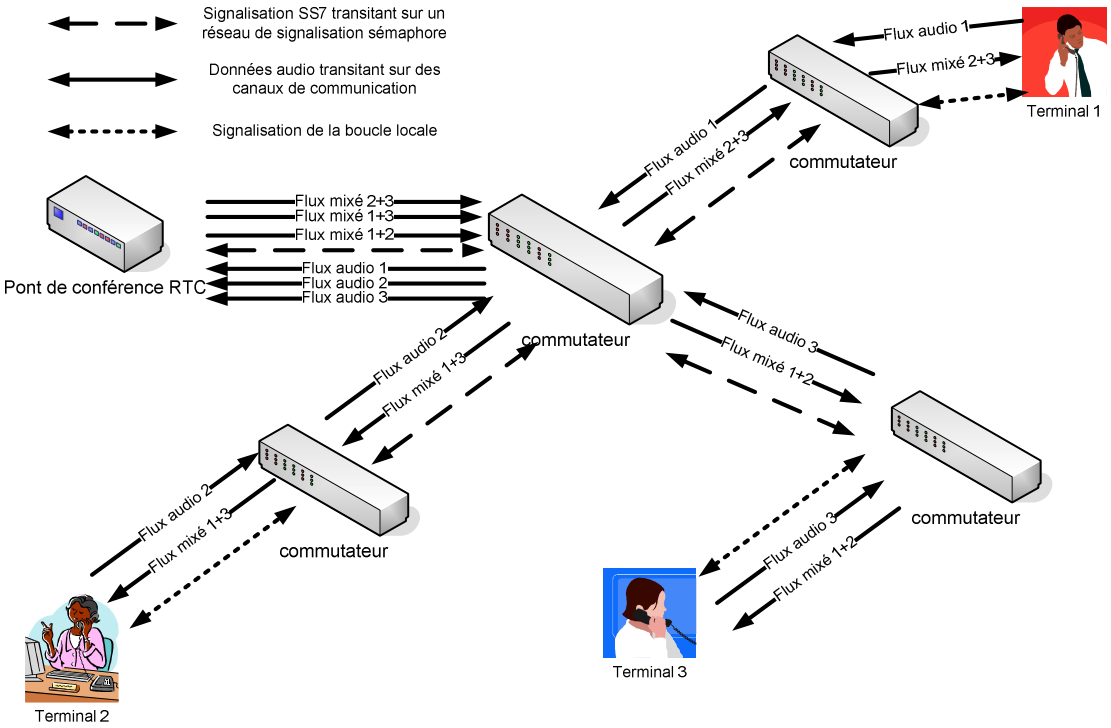

Figure 1.2 Principe de la conférence RTC

Pour chaque terminal appelant le pont, la première étape a été l'échange des informations de décrochage du combiné et de composition du numéro du pont (défini à l'avance) à travers la signalisation sur la boucle locale.

Chaque commutateur local a ensuite pris en charge l'appel et a dialogué grâce au langage de signalisation SS7 avec d'autres commutateurs par le réseau de signalisation sémaphore, et cela jusqu'au pont de conférence. Ce dernier est une entité pouvant recevoir plusieurs appels et pouvant mixer des flux média.

Ainsi chaque participant est en communication avec le pont de conférence. Ce dernier reçoit, à travers les canaux de communication réservés par la signalisation, un flux audio pour chacun d'eux et leur envoie un flux mixé des autres participants.

1.2 La conférence audio sur IP

Avant d'intégrer la spatialisation dans la conférence audio, il est nécessaire de savoir comment la conférence audio sur IP standard est faite aujourd'hui. La voix sur IP possède également une partie signalisation et une partie média, qui vont être détaillées ci-dessous.

Concrètement, en voix sur IP, la partie signalisation permet de rechercher et d'authentifier les différents participants, de négocier les codeurs utilisés pour compresser les données lors de la communication, de spécifier le format des données échangées (protocoles de transport, options, …), de préciser les ports et les adresses IP utilisées pour la communication, de préciser tout ou partie des capacités des terminaux, etc. Les différents échanges de signalisation entre deux terminaux ou un terminal et un pont de conférence seront appelés ici dialogue en référence à SIP [82].

Le choix du protocole de signalisation s'est porté sur le protocole SIP (Session Initiation Protocol), normalisé à l'IETF (*Internet Engineering Task Force*). Il est adopté par tous les grands acteurs de l'industrie en tant que protocole retenu pour l'évolution des réseaux. Un appel SIP standard sera explicité dans la section 1.6. En quelques mots, cet appel SIP réunit les parties signalisation et média, et permet de mettre en communication deux entités en négociant au préalable des flux audio, la façon dont ils sont éventuellement codés, comment et où ils doivent être envoyés, etc. Il convient de préciser que les exemples ci-dessous en matière d'architecture réseau sont valables quel que soit le protocole de signalisation (H.323 [2], MGCP [8], etc.), même si des références à des exemples SIP sont faites. Il est supposé que l'établissement des dialogues et les négociations des médias se sont déroulés correctement. Pour la suite de cette section, aucune connaissance particulière en SIP n'est requise.

La partie média de la VoIP correspond à toute opération (dans un sens large) susceptible d'être appliquée sur les flux média (mélange des flux audio, prétraitement, encodage pour compresser les données, transport sur les réseaux IP, décodage pour décompresser les données, etc.), suivant les architectures choisies (voir ci-dessous pour leur énumération). On retrouvera ainsi les notions de pont et de mixage présentés dans la conférence audio du réseau RTC.

Grâce à l'utilisation d'un réseau de paquets pour faire de la conférence audio, il est possible de définir plusieurs architectures rendant beaucoup plus libres les parties signalisation et média. Il existe ainsi trois types d'architectures de conférence utilisables pour la partie signalisation : la conférence centralisée, la conférence distribuée multi-unicast et la conférence distribuée multicast. De nombreuses variantes peuvent en découler. L'objectif de la section suivante sera de présenter ces différentes possibilités ainsi que leurs avantages et inconvénients.

Dans les parties ci-dessous, il n'est pas présenté la manière dont les terminaux sont informés de la création des conférences ou la façon dont ils les ont rejoint. En guise d'information, les terminaux sont avertis de la création d'une conférence par des messages SIP ou des liens web par exemple. Cela sera illustré dans la section 1.6.3.

1.2.1 La conférence centralisée

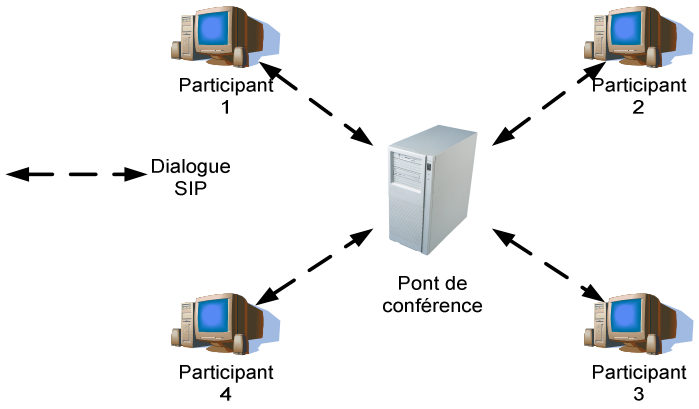

Figure 1.3 Illustration de la conférence audio centralisée

Le modèle de la conférence centralisée correspond au cas où tous les flux de signalisation passent par la même entité appelée pont de conférence. Plus simplement, chaque terminal de chaque participant est en dialogue ou communication point à point avec le pont. Sans tenir compte des flux média, on peut résumer ce modèle par la Figure 1.3 aussi appelé topologie en étoile. Les figures suivantes correspondent aux différentes configurations de la conférence centralisée, suivant la manière dont les flux média sont gérés [75].

1.2.1.1 Configuration les plus courantes de la conférence audio centralisée avec une entité de mixage

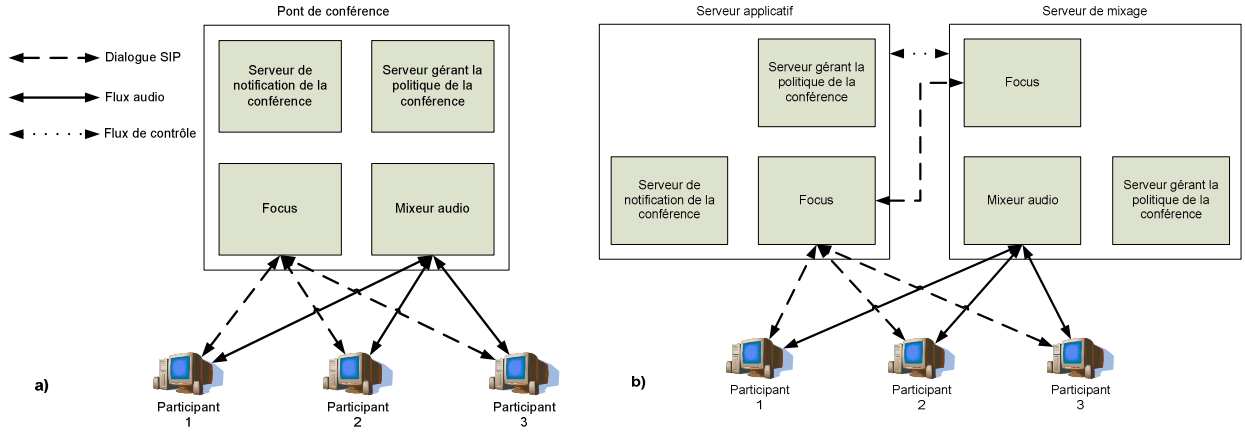

Figure 1.4 a) Pont de conférence mixeur signalisation/média b) Serveur applicatif + serveur de mixage formant le pont de conférence mixeur

Le modèle Figure **1.4** a) présente le cas le plus courant du pont de conférence mixeur pour centraliser les flux de signalisation et de média. Il est à la fois le *focus* (entité qui réceptionne les flux de signalisation), le mixeur audio, les serveurs de notification (pour avertir les participants d'évènements dans la conférence) et de politique de conférence (pour gérer les règles de la conférence).

La négociation des flux média se fait comme pour un dialogue standard entre chaque participant et le pont de conférence. C'est à ce dernier de gérer, de décoder et de mélanger les flux venant des participants ainsi que d'envoyer les mixages encodés vers chacun d'eux.

Le modèle Figure 1.4 b) présente le cas où le serveur utilisé, pour centraliser les flux de signalisation, est dissocié de celui qui régit les flux de média. Dans cette alternative, les deux entités sont des serveurs centralisés qui forment le pont de conférence mixeur.

Le serveur utilisé pour les flux de signalisation est appelé serveur applicatif. Il administre la politique média, les dialogues avec chaque participant ainsi que les autorisations et l'adhésion de ceux-ci. Il est le *focus* vu par les participants, le serveur de notification de la conférence et le serveur de politique de conférence.

Le serveur utilisé pour les flux média est appelé le serveur de mixage. Il inclut un *focus*, un serveur de politique de conférence et un mixeur audio. La politique de la conférence indique au *focus* qu'il doit accepter toutes les invitations du *focus* du serveur applicatif.

Le *focus* du serveur applicatif se comporte en fait comme un tiers d'appel (ou un intermédiaire) pour contrôler les flux média entre les participants et le serveur de mixage, permettant à ce dernier de remplir son rôle. Il faut bien comprendre que les flux de signalisation restent bien d'une part entre le *focus* du serveur applicatif et les participants et d'autre part entre le serveur applicatif et le serveur de mixage. Il est à noter que le serveur applicatif peut aussi dialoguer avec le serveur de mixage par d'autres protocoles que SIP.

La négociation des flux média se fait virtuellement comme pour un dialogue standard entre les participants et le serveur de mixage. Cela reste virtuel car, comme on l'a vu, chaque extrémité (participant ou le serveur de mixage) ne voit que le serveur applicatif. C'est au serveur de mixage de gérer, de décoder et de mixer les flux venant des participants et de renvoyer codés les résultats vers chacun d'eux.

1.2.1.2 Première variante de la conférence audio centralisée avec un terminal mixeur

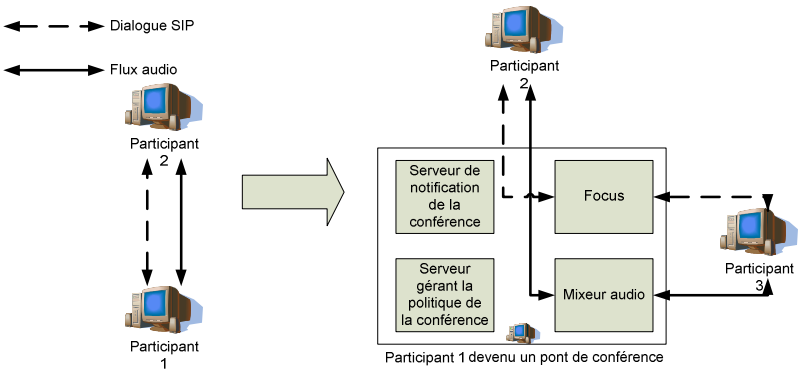

Figure 1.5 Terminal devenant un pont mixeur

Le modèle Figure 1.5 présente le cas où un terminal devient le pont de conférence mixeur pour les flux de signalisation/média. Dans cet exemple, le dialogue à deux, suite à l'arrivée du participant 3 (P3), s'est transformé en conférence. Le participant 1 (P1) doit donc faire office de *focus* pour les flux de signalisation et de mixeur audio pour les flux média. Il continue évidemment de participer à la conférence. Il envoie au participant 2 (P2), les flux mixés de P3 et de lui-même, et à P3, ceux de P2 et de lui-même.

La négociation des flux média se fait comme pour un dialogue standard entre chaque participant (participants 2 et 3) et celui faisant office de point central (participant 1). C'est à ce dernier de gérer, de décoder et de mixer les flux venant des autres participants et de renvoyer codés les résultats vers chacun d'eux. Ce genre de conférence est évidemment très limité en nombre de participants du fait de l'impossibilité matérielle pour un terminal utilisateur de gérer de nombreux flux média et les débits associés.

Il est à noter que cette configuration de conférence audio centralisée présente un inconvénient majeur. Elle ne permet pas de garantir une qualité de service car elle dépend des capacités, potentiellement limitées et non contrôlables, d'un terminal.

1.2.1.3 Autres variantes de la conférence audio centralisée avec une entité de mixage

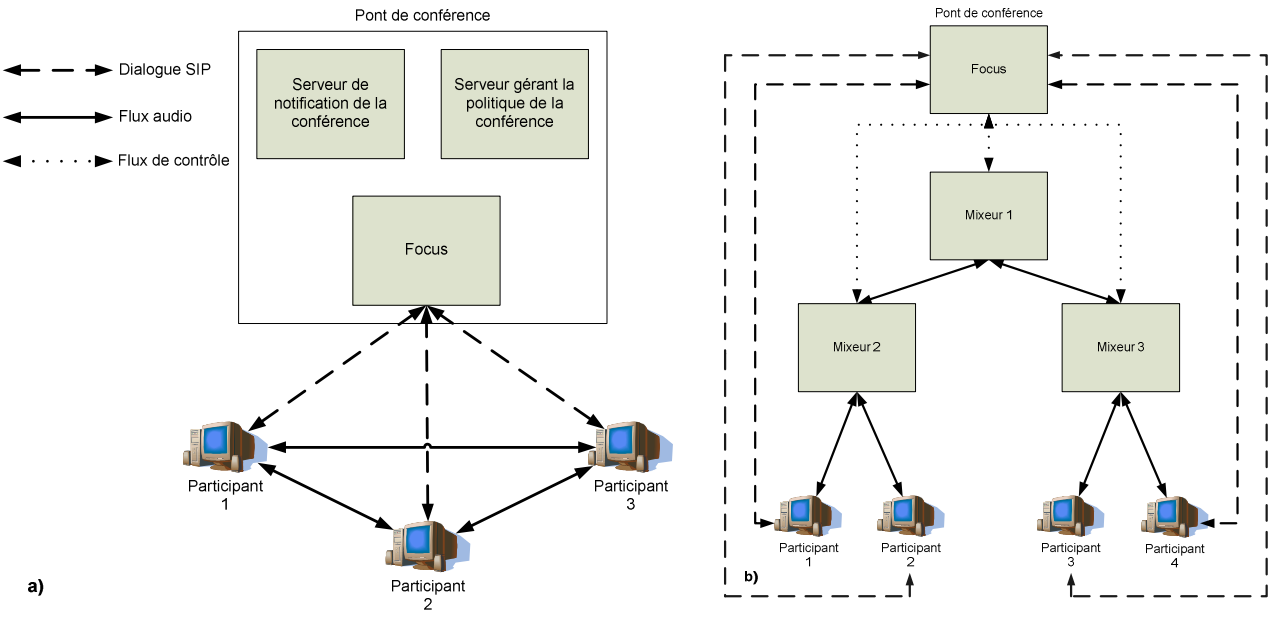

Figure 1.6 a) Pont de conférence + terminaux en mixeurs b) Pont de conférence + mixeurs cascadés

Le modèle Figure 1.6 a) présente le cas d'un pont de conférence utilisé pour centraliser les flux de signalisation. La négociation des flux média se fait comme pour un dialogue standard entre chaque participant et le pont de conférence. Cependant, dans cette configuration, les flux média sont mixés au niveau des terminaux. Le *focus* dirige le média de chaque participant, en se comportant comme un tiers d'appel entre chacun d'eux. Les flux peuvent être distribués de 3 manières :

- Chaque participant envoie ses flux à tous les autres (cas de la Figure 1.6 a).
- Chaque participant rejoint un groupe multicast pour recevoir les flux des autres participants (voir la section 1.2.3 pour une introduction au multicast) et envoie ses flux audio vers cette adresse de groupe.
- Dans un modèle SSM (Source Specific Multicast), chaque participant envoie ses flux en unicast vers un point central, qui utilisera le multicast afin de le diffuser aux autres acteurs de la conférence.

Le modèle Figure **1.6** b), présente le cas d'un pont de conférence jouant le rôle de *focus* pour centraliser les flux de signalisation et de multiples mixeurs audio utilisés pour mixer les contributions des différents participants. En effet dans le cas de conférence avec un grand nombre de participants, il est impossible à un serveur de gérer tous les flux média. En conséquence, cette architecture, réunissant plusieurs serveurs en cascade, permet de résoudre ce problème. Le *focus* gère les mixeurs audio par des flux de contrôle SIP ou autre.

La négociation des flux média se fait comme pour un appel SIP point à point standard entre chaque participant et le *focus*. C'est au mixeur audio spécifié dans la négociation de gérer, de décoder et de mixer les flux venant des participants qu'il gère et de renvoyer les résultats vers chacun d'eux. En guise de remarque, chaque mixeur reçoit un flux audio monophonique (issu du mixage des flux audio de plusieurs participants) des mixeurs avec qui il est en relation.

1.2.1.4 Variantes de la conférence audio centralisée avec une entité répliquante

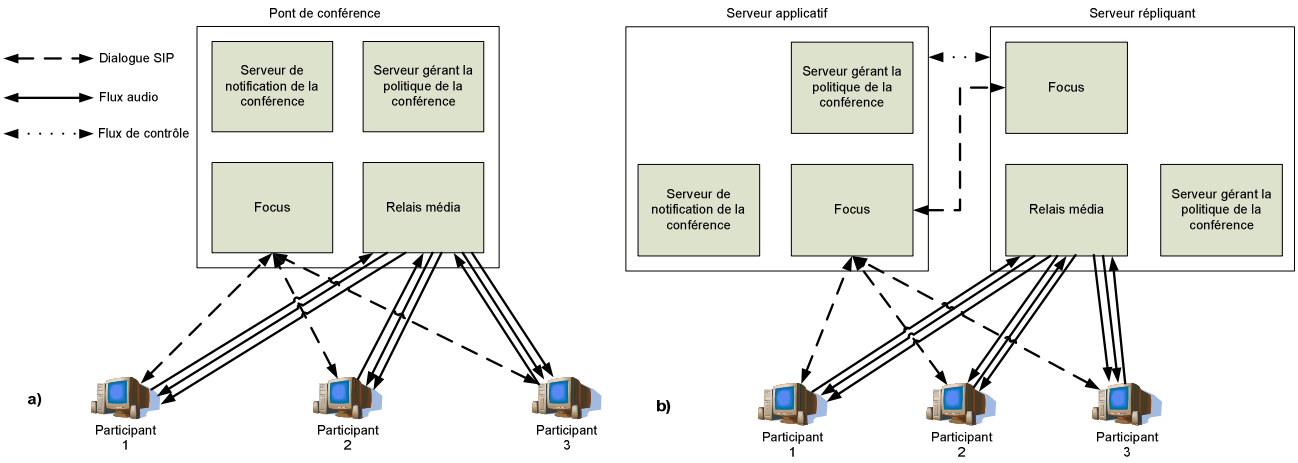

Figure 1.7 a) Pont de conférence répliquant + terminaux en mixeurs b) Serveur applicatif + serveur répliquant + terminaux en mixeurs

A titre indicatif, il existe un modèle illustré Figure 1.7 pour la gestion de conférence, qui consiste en fait à utiliser un pont de conférence répliquant. Ce modèle n'est pas présent dans les documents SIP étudiés, mais son principe est très simple puisque ce pont répliquant est tout simplement un relais média. Du point de vue du terminal, ce dernier enverra un flux audio vers le pont de conférence et il en recevra autant qu'il y a de participants distants. Pour le pont de conférence en mode répliquant, la gestion des flux média est d'une certaine manière distribuée, car il enverra un flux média issu d'un participant à tous les autres participants, sans faire de traitement de mixage ou autre… Le mixage se fera sur les terminaux des participants à la conférence.

Tout comme les exemples illustrés Figure 1.7 dans le cas d'un serveur média, il existe deux cas d'utilisation :

- Un pont de conférence répliquant, qui gère à la fois la signalisation et le relais des flux média (Figure 1.7 a)). La négociation des flux média se fait comme pour un dialogue standard entre chaque participant et le pont de conférence (Figure 1.4 a)). Par contre, elle se différencie dans le sens où pour un terminal, il faut spécifier un flux en émission et en accepter N-1 en réception (dans le cas où il y a N participants à la conférence), au lieu d'un flux bidirectionnel dans le cas d'un serveur média avec mixeur (cf. Figure 1.4 a)). Pour le pont de conférence, il négocie l'opposé, c'est-à-dire un flux en réception et N-1 en émission pour chaque participant.
- Deux serveurs formant le pont de conférence répliquant : un qui gère la signalisation et l'autre la réplication des flux média Figure 1.7 b). La négociation des flux média se fait virtuellement comme pour un dialogue standard entre les participants et le pont répliquant. Cela reste virtuel car en fait chaque extrémité (participant ou le pont répliquant) est inconnue pour l'autre. En effet, chaque participant ou le pont répliquant ne dialogue qu'avec le serveur applicatif. Tout comme le cas précédent, pour un terminal, il faut spécifier un flux en émission et N-1 en réception. Pour le serveur applicatif, il négocie un flux en réception et N-1 en émission pour chaque participant.

1.2.1.5 Avantages de la conférence centralisée en général

Nous reviendrons plus en détail sur les avantages de la conférence centralisée en termes de médias dans le chapitre suivant. Concernant les avantages pour la signalisation, citons :

- La réduction du trafic des messages de signalisation SIP pour les participants : équivalent à un dialogue standard. Idem pour les flux média en présence d'un serveur spécifique de mixage dans le réseau.

- La gestion facilitée de la conférence pour un administrateur : invitation ou exclusion d'un participant, création ou suppression d'une conférence, diffusion des informations spécifiques à chaque participant, ainsi que le contrôle et une administration facilités.
- La facturation aisée pour un fournisseur de services.

1.2.1.6 Inconvénients de la conférence centralisée en général

Nous reviendrons plus en détail sur les inconvénients de la conférence centralisée en termes de médias dans le chapitre suivant. Le principal désavantage de la conférence centralisée au niveau média est qu'elle dépend d'un point central de signalisation. En cas de panne de celui-ci, la conférence est terminée.

1.2.2 La conférence distribuée multi-unicast

Ce type de conférence est très peu explicité dans les documents IETF, que cela soit dans les *Internet-Drafts* (qui sont des documents de travail) ou les RFCs (*Request For Comments* qui sont des documents validés par l'IETF). Il en est question à plusieurs reprises, mais aucune publication n'a été effectuée récemment.

Ces conférences distribuées multi-unicast sont en fait très coûteuses en termes de ressources de calcul et réseau, ce qui est sûrement la raison du manque de succès des recherches dans ce domaine, par rapport aux conférences centralisées.

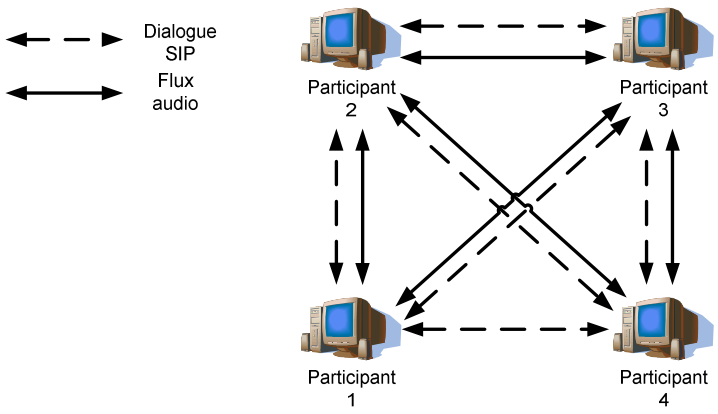

Figure 1.8 Illustration de la conférence audio distribuée multi-unicast

Dans ce modèle de conférence (voir Figure 1.8), la gestion des flux de signalisation et média est très simple, puisque chaque entité SIP établit un dialogue point à point avec toutes les autres entités. Le nombre de dialogues établis est susceptible d'augmenter très vite : pour N participants à une conférence, il y a C_N^2 dialogues. Le principe de ce type de conférence est très simple, puisqu'il se base sur un appel SIP standard. La difficulté réside dans le fait d'informer un nouveau participant de l'existence des autres participants à la conférence. Cela peut néanmoins se faire grâce à de nouvelles méthodes proposées dans les Internet-Drafts.

Dans cette architecture, chaque terminal envoie son flux audio vers les autres participants.

1.2.2.1 Avantages de la conférence distribuée multi-unicast

- Toutes les entités sont confrontées à la même charge.
- Ce type de conférence est robuste aux pannes car la panne d'une entité n'entraîne pas la fin de la conférence.

- Le départ d'une entité n'entraîne pas la fin de la conférence, même si elle est à l'origine de celle-ci.
- Chaque terminal n'encode qu'un seul flux audio.
- Ce type de conférence est intéressant pour un faible nombre de participants.

1.2.2.2 Inconvénients de la conférence distribuée multi-unicast

- Pour chaque participant, il convient de gérer N-1 dialogues, ce qui peut être important pour un nombre N élevé de participants.
- De plus, il faut pouvoir envoyer N-1 flux audio vers les participants à la conférence. Dans ce cas, une entité limitée en bande passante aura de grandes difficultés à gérer l'envoi de flux.
- En réception, il faut pouvoir gérer N-1 flux audio.
- Il n'y a pas de contrôle sur la création ou la suppression de conférence dans ce type de conférence. Idem pour l'ajout ou la suppression de participants.

1.2.3 La conférence distribuée multicast

Tout d'abord, on rappelle que l'unicast est un moyen de communication sur un réseau d'une source vers un destinataire et que le broadcast (ou diffusion) permet l'envoi de messages d'une source vers tous les destinataires. Pour sa part, le multicast permet à une source l'envoi de messages vers N destinataires, tous inscrits à un groupe multicast pour les recevoir, avec 1<N<tous. La Figure 1.9 illustre les différents modes de communication qui viennent d'être cités.

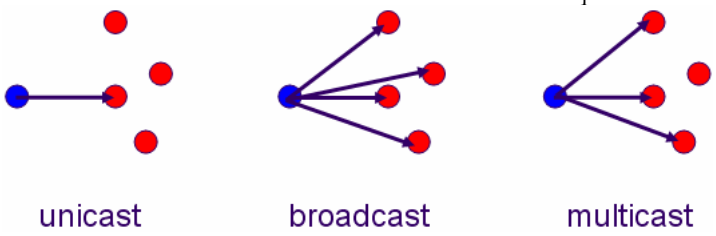

Figure 1.9 Principe du multicast

Le multicast est une technologie, qui conserve la bande passante, tout en réduisant le trafic puisque chaque paquet IP ne passe qu'une fois par un nœud. La Figure 1.10 illustre ce principe et précise le nombre de paquets IP arrivant ou partant de chaque terminal ou routeur.

La conférence distribuée multicast, quant à elle, est un modèle de conférence à grande échelle (cf. [75], [59]), qui ne possède pas de réelle limite d'implémentation (sauf son faible déploiement dans les réseaux), puisqu'elle ne nécessite pas d'entité centrale susceptible d'avoir trop de données à gérer.

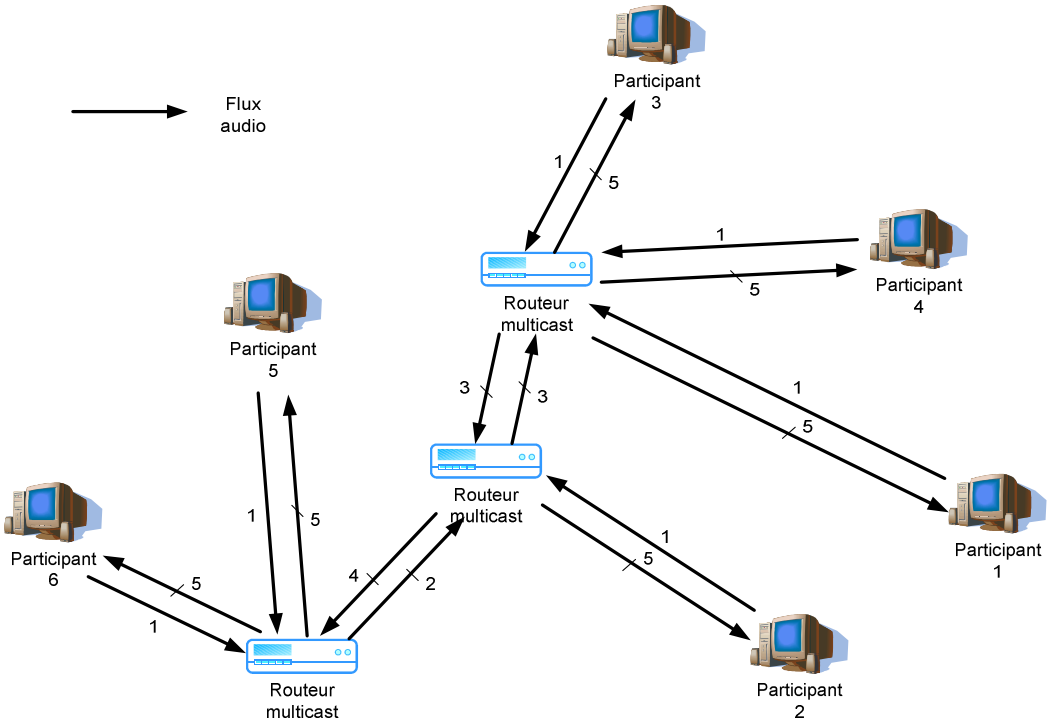

Figure 1.10 Illustration de la conférence audio distribuée multicast

Tout d'abord, il convient d'allouer une adresse multicast pour cette conférence. Chaque participant devra s'inscrire au groupe associé pour récupérer les flux audio que les autres participants auront envoyé à cette adresse multicast.

Dans le cas de ce type de conférence, la difficulté réside évidemment dans le fait de connaître l'adresse du groupe multicast : par des pages web (sorte d'annuaire des conférences en cours) ou des mails d'invitation généralement, ou plus rarement par des messages d'un protocole de signalisation.

En utilisant des pages web, il est possible de n'utiliser aucun message de signalisation. En effet, il suffit de s'inscrire au groupe concerné par la conférence, tenir compte des données spécifiées (codeurs, adresse IP multicast et numéro de port), et de gérer en conséquence les différentes données que l'on reçoit.

Les participants envoient un flux audio vers une adresse multicast et en reçoivent N-1 au maximum (voir Figure 1.10) si N est le nombre de participants. En effet, dans le cas de l'utilisation d'une suppression de silence pour un flux audio, ce type de gestion des flux média permet de ne recevoir que les flux des personnes parlant à un moment donné. Cela limite donc considérablement le nombre de flux à mixer si la détection d'activité vocale est utilisée, sachant que dans une conférence, généralement deux ou trois personnes, au maximum, parlent en même temps.

1.2.3.1 Avantages de la conférence distribuée multicast

- Cette configuration est robuste car il n'y a aucun point central de signalisation donc il n'y a pas de point de rupture pour la conférence. En cas de panne d'un des routeurs, le protocole IP fera en sorte qu'un autre routeur prenne le relais.
- Il n'y a aucune difficulté pour un participant de quitter ou rejoindre la conférence. Rejoindre une conférence peut être fait grâce uniquement à des contacts déjà présents ou des pages web. Il n'y a pas de gestion particulière centralisée de celle-ci.
- Le départ d'un participant ou une panne de sa machine n'entraîne pas la fin de la conférence.

- Le réseau s'occupe seul de la distribution des flux média aux utilisateurs concernés. Il n'y a donc besoin d'aucune entité supplémentaire à ajouter et maintenir. Cela est peu coûteux.
- Cette configuration est utile pour les conférences à grande échelle.
- La signalisation SIP peut être inexistante.

1.2.3.2 Inconvénients de la conférence distribuée multicast

- Ce sont ceux du multicast en général : la difficulté d'obtenir les adresses multicast disponibles, les différents routeurs doivent supporter le multicast de bout en bout (ce qui est loin d'être courant aujourd'hui), etc.
- Tout le monde est susceptible de pouvoir écouter un flux audio multicast, ce qui pose des problèmes de confidentialité et de cryptage des données.
- Il est difficile de pouvoir gérer les participants (ajout ou suppression) ainsi que les création et suppression de conférence.
- Il y a peu d'information sur les participants (manière dont ils sont entrés dans la conférence, etc.).

1.2.4 Quelques exemples de logiciels de conférence audio sur IP

Plusieurs logiciels de VoIP permettent aujourd'hui de créer et de gérer des conférences audio entre des participants distants. Peu d'informations sont cependant disponibles sur les protocoles de signalisation utilisés et l'architecture qu'ils utilisent. En voici quelques-uns parmi les plus connus :

- Skype (http://www.skype.com/intl/fr/), jusqu'à 25 participants par conférence. Skype utilise un protocole de signalisation propriétaire et un terminal effectue le mixage pour tous.
- Teamspeak (http://www.goteamspeak.com/). Pas d'informations sur le protocole de signalisation utilisé mais le mixage se fait sur un serveur contrôlé par un utilisateur.
- …

Des solutions hardware existent aussi de type pont de conférence mixeur, parmi lesquels :

- Les solutions de Polycom qui sont des entités dédiées à la conférence. http://www.polycom.com/usa/en/products/network/conferencing_platform/conferencing_platform.html.
- Une des solutions de Avaya http://www.avaya.com/gcm/master-usa/en-us/products/offers/meeting_exchange_express_edition.htm&View=ProdOverview.
- Une des solutions de la société Compunetix présentée à l'adresse suivante : http://www.compunetix.com/summit/index.html.
- …

Il existe des logiciels de VoIP permettant de faire des dialogues point à point, c'est-à-dire entre uniquement deux terminaux. En autres, nous pouvons citer :

- Windows Live Messenger (http://get.live.com/messenger/overview), qui permet uniquement des communications d'un terminal à un autre terminal.
- Yahoo Messenger (http://fr.messenger.yahoo.com/), qui permet uniquement des conférences d'un terminal à un autre terminal.

Des clients SIP sont également disponibles parmi lesquels :

- Linphone : http://www.linphone.org/.
- X-Lite : http://www.counterpath.com/.
- OpenWengo : http://openwengo.com.

- MiniSIP : http://www.minisip.org/)
- …

1.3 Les différents éléments de la conférence audio sur IP

Ce paragraphe vise à présenter les différentes éléments de la voix sur IP afin de mieux appréhender les contraintes associées à la partie média de la conférence. Nous nous intéresserons principalement aux aspects réseaux et matériels. Il convient au lecteur de bien lier les architectures présentées dans la section 1.2 et les différents éléments présentés ci-dessous et illustrés Figure 1.11.

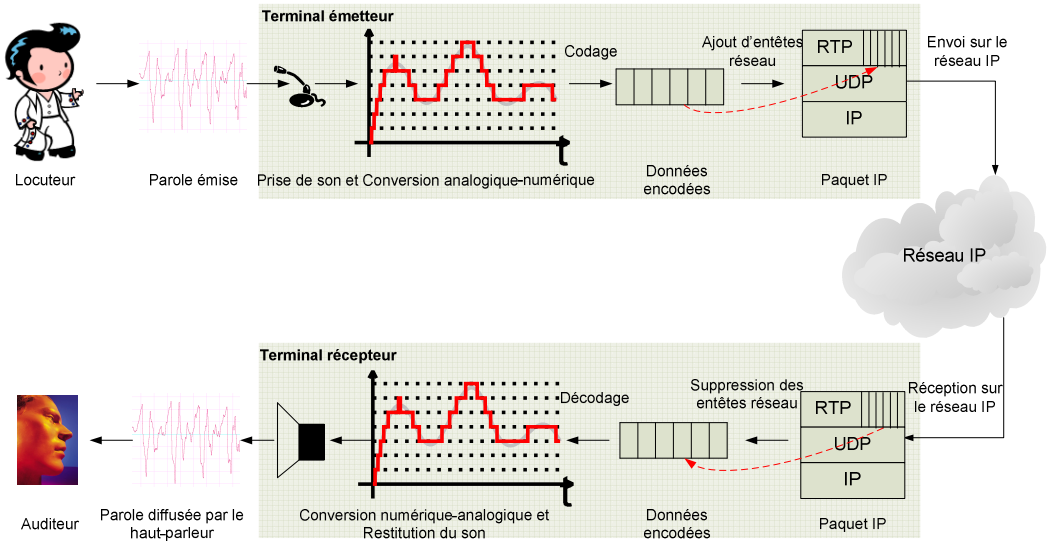

Figure 1.11 Principe de la VoIP

1.3.1 La prise et la restitution du son

La prise de son s'effectue, généralement, par un simple microphone monophonique du côté émetteur. Le contenu capturé est ensuite envoyé vers la carte de traitement sonore ou carte son.

C'est cette dernière qui convertit le signal analogique en signal numérique ou séquence d'échantillons. Cette opération, appelée échantillonnage, dépend de deux critères spécifiques aux cartes son : la résolution et la fréquence d'échantillonnage.

- La résolution est le nombre de bits sur lesquels sont codés les échantillons. Généralement, ce nombre est 16 ou 24 et est lié aux valeurs de la quantification.
- La fréquence d'échantillonnage est le nombre d'échantillons qui seront produits à la seconde. Son unité est le Hertz ou plus généralement le kiloHertz. Généralement, sur les cartes son standard, elle varie de 8 kHz à 192 kHz.

Dans le cas qui nous intéresse, une carte son, fonctionnant sur 16 bits et à une fréquence d'échantillonnage dans une plage de 8 à 48 kHz, est suffisante.

En effet, nous nous sommes intéressés seulement à deux bandes de qualité audio. La bande classique dite bande étroite ou narrowband, ainsi que la bande dite élargie ou wideband.

- Le narrowband : Bande de qualité de la téléphonie RTC, entre 300 et 3400 Hz, qui sert de référence basse. Cette limitation historique est due au matériel de l'époque utilisé et au système de transmission qui en a découlé (TDM).

- Le wideband : Le transfert des données sur des réseaux IP a permis l'utilisation de cette nouvelle bande de qualité. Elle correspond à une bande de 50 à 7000 Hz. Le document [23] montre l'importance de l'apport de la bande passante dans la sensation de qualité.

Au niveau du récepteur, les données décodées sont ensuite envoyées vers la carte son pour être transformées en signaux analogiques joués par la suite par un système de restitution.

Suivant la logique de nos hypothèses (explicitées dans la section 1.7.3.2), nos systèmes de restitution privilégiés seront le casque à deux oreillettes ainsi qu'un système de deux haut-parleurs.

1.3.2 Les codeurs audio

A partir du moment où les données sont disponibles sous forme numérique, se pose, pour l'émetteur (terminal ou pont), le problème de la compression audio avant de les envoyer sur le réseau. Il est évident, qu'avec la croissance des bandes passantes disponibles, il est légitime de se demander si la compression est encore utile. Il est assez simple d'y répondre en soulignant que de plus en plus d'applications ne font pas seulement de la conférence audio mais aussi du partage de documents, de vidéos, etc., d'où l'intérêt de préserver la bande passante. Il est de plus toujours intéressant de compresser les données audio ne serait-ce que pour pouvoir augmenter les capacités d'accueil, en termes de participants ou de conférence, d'un pont mixeur de conférence par exemple. Enfin cette bande passante peut être limitée par le fournisseur de service.

De plus, le débit monophonique non compressé est tout de même de 128 kbits/s pour un contenu narrowband et de 256 kbits/s pour un contenu wideband. Dans le cadre d'une configuration distribuée, le débit arrivant sur un terminal est susceptible d'être important.

Les données audio numériques fournies à la plupart des codeurs récents sont découpées en trames de 10 ou 20 ms. Il en ressort des trames compressées qui sont envoyées sur le réseau IP (voir la section 1.3.3), sous forme de paquets. Cette compression est une raison de la dégradation en termes de qualité globale du signal audio. Elle introduit de plus du retard et l'utilisation de ressources de calcul (*Central Processing Unit* ou CPU).

Au niveau d'un récepteur (terminal ou pont), les données sont décodées par le décodeur adapté à l'encodeur. Cela souligne la nécessité que les deux entités se soient mises d'accord par l'intermédiaire de la partie signalisation.

1.3.3 Le réseau IP

1.3.3.1 Réseau de données à commutation de paquets

Une fois les trames compressées sur un pont ou un terminal, elles sont envoyées sur le réseau IP à l'intérieur de paquets (classiquement l'équivalent de 20 ms de signal compressé dans un paquet) et cela comme n'importe quel paquet IP. Le réseau IP est un réseau de données à commutation de paquets, à comparer au réseau RTC qui est à commutation de circuits.

Il est rappelé qu'un réseau à commutation de circuits est une méthode de transfert de données consistant à établir un circuit dédié au sein d'un réseau [1]. Une sorte de tunnel de données est ainsi réservé entre les deux abonnés et les nœuds intermédiaires (des commutateurs en RTC).

Pour sa part, un réseau à commutation de paquets consiste à envoyer l'information sous forme de paquets, dont chacun peut suivre un chemin différent d'un autre. Cela dépend en effet des nœuds (des routeurs pour les réseaux IP) de ce réseau et notamment de leur table de routage. De même, rien ne garantit que l'ordre de départ des paquets soit conservé à l'arrivée. La perte de paquets, suite à une congestion du réseau, est également possible.

1.3.3.2 Comment transporter des données isochrones?

Les applications courantes sur Internet utilisent la suite de protocoles TCP/IP. Les données généralement transportées ne sont cependant généralement pas isochrones. On rappelle que les données isochrones sont des données qui doivent être restituées dans le temps avec un décalage fixe par rapport au moment où elles ont été capturées. Cela sous-entend une absence de déséquencement. La Voix sur IP (VoIP), quant à elle, emploie l'ensemble RTP/UDP/IP adapté au transport de ces données.

Le protocole UDP fournit un service de remise sans connexion et non fiable utilisant IP pour transporter les messages point à point. Le protocole RTP (*Real-time Transport Protocol*), utilisé avec UDP, fournit le moyen de transport des données temps réel comme l'audio ou la vidéo. RTP est souvent utilisé conjointement avec RTCP (*RTP Control Protocol*), ce dernier permettant de transporter des informations concernant la qualité effective de la transmission ainsi que des informations concernant l'identité des participants, et de synchroniser audio et vidéo, RTP et RTCP sont deux protocoles de l'IETF utilisés pour le transport de flux médias sur le réseau IP.

Il est à noter que les protocoles RTP et RTCP, définis dans la RFC 3550 [85], n'ont aucun contrôle sur la qualité de service contrairement à RSVP (*Resource Reservation Protocol*). RTP et RTCP permettent simplement aux récepteurs de s'adapter au mieux et de corriger les imperfections du réseau en gérant des tampons et des mécanismes de remise en ordre de paquets et de masquage des paquets perdus.

Ces protocoles sont sommairement introduits ci-dessous.

1.3.3.2.1 Le protocole IP

Le protocole IP (*Internet Protocol*) [87] assure un service sans connexion. Il est de plus non fiable car il n'assure aucune garantie quant au fait que les paquets IP arrivent à destination. Sur ce fait, il se repose sur les protocoles des couches supérieures pour garantir la fiabilité, comme TCP et même des utilisations particulières de UDP (en SIP, par exemple avec des protocoles au dessus de UDP qui vont assurer la fiabilité de certaines données). On parle alors de remise au mieux.

La couche IP dans la suite de protocoles TCP/IP est une couche abstraite où chaque paquet est routé indépendamment, c'est-à-dire qu'un paquet *i* peut prendre un chemin totalement différent du paquet précédent *i-1*. Un paquet peut donc arriver après celui qu'il précède. Simplement, le protocole IP permet aux paquets de se déplacer sur le réseau Internet indépendamment les uns des autres.

Grâce au routage, le protocole IP permet aux paquets de trouver le meilleur chemin vers une destination. Il est à noter que les tables de routage évoluent lentement ce qui permet d'assurer une stabilité des chemins entre deux points.

1.3.3.2.2 Le protocole UDP

Le protocole UDP (*User Datagram Protocol*) [87] est un protocole de transport orienté paquets, sans connexion et non fiable. Il permet donc aux applications d'échanger des paquets sans accusé de réception ni remise ou ordre garantis. Voici les points forts d'UDP :

- Il est efficace pour les diffusions car il ne nécessite pas autant de connexions (contrairement à TCP) que de personnes.
- Il est plus rapide, plus simple, plus efficace que TCP au détriment de la fiabilité de la transmission (possibilité de pertes de trame ou de déséquencement).
- Il possède un temps d'exécution court qui permet de tenir compte des contraintes temps réel ou de limitation de mémoire sur un processeur. En conséquence, il est très utilisé pour le transport des données temps réel.
- Il peut éventuellement contrôler l'intégrité des données.
- UDP peut être utilisé par plusieurs processus (multiplexage). Ils auront la même adresse IP mais des numéros de ports différents.

1.3.3.2.3 Le protocole TCP

Le protocole TCP (*Transmission Control Protocol*) [87], qui garantit une fiabilité point à point entre deux processus d'application, est un protocole fiable orienté connexion. Il assure les fonctions suivantes :

- Transferts de données de base.
- Fiabilité : Utilisation du PAR (Positive Acknowledgment Retransmission) basé sur l'envoi de données et l'attente de leur validation par un acquittement ACK sous peine de leur re-transmission.
- Contrôle de flux : Utilisation de la fenêtre de contrôle de flux pour réguler les transmissions.
- Connexions : Ensemble composé de "l'adresse IP et du numéro de port de l'émetteur, de l'adresse IP et du numéro de port du récepteur".
- Remise en ordre des paquets.
- Multiplexage : TCP peut être utilisé par plusieurs processus. Ils auront la même adresse IP mais des numéros de ports différents.

TCP est utilisé pour transporter toute donnée dont l'intégrité et le séquencement sont primordiaux.

1.3.3.2.4 Le protocole SCTP

Le protocole SCTP (*Stream Control Transmission Protocol*), défini dans la RFC 2960 [90], est un récent protocole de transport sur IP qui joue le même rôle que les protocoles plus standard TCP et UDP. Comme TCP, il fournit un service de transport fiable orienté session assurant que la donnée est transportée sans erreur, à un débit adapté et sans déséquencement. Contrairement à TCP, il rajoute des fonctions caractéristiques pour le transport de la signalisation téléphonique.

Le nom SCTP provient de multi-streaming function qui autorise la partition des données en plusieurs flux qui ont la propriété de remise indépendante en séquence. La perte d'un flux entraînera du retard uniquement sur ce flux et non sur les autres. Cela permettra un gain de temps par rapport au TCP qui n'a qu'un seul flux et qui doit retarder la remise de l'ensemble en cas de pertes. En signalisation téléphonique par exemple, il est important que seuls les messages d'un même appel ou canal soient conservés dans l'ordre. Un ordre strict par rapport aux messages des autres appels n'est pas primordial. D'où, on affectera dans ce protocole un flux pour un appel ou un canal, mais il n'y aura qu'une session ouverte.

1.3.3.2.5 Le protocole RTP

Le protocole RTP (*Real-time Transport Protocol*) permet le transfert de données isochrones (données devant être restituées dans le temps avec un décalage relatif fixe par rapport au moment où elles ont été capturées) à travers un réseau de paquets. Il a été conçu pour permettre aux logiciels de réception de compenser la gigue (section 1.4.2.1) due au multiplexage statistique et les éventuels pertes et déséquencements de paquets introduits par le réseau de transport IP. RTP est utile pour tout type de données temps réel.

RTP, généralement utilisé au dessus de UDP, est non fiable et fournit des services de remise de données temps réel avec notamment un identifiant de type de données, un numéro de séquence, un marquage en temps et un identifiant de source. RTP ne fournit pas de garanties de qualité de service (voir la section 1.1.2) ni de remise en temps.

1.3.3.2.6 Le protocole RTCP

Le protocole RTCP (*RTP Control Protocol*), généralement utilisé au dessus de UDP, offre une vue sur la performance et est basé sur la transmission périodique de paquets de contrôle à tous les participants de la session contenant diverses statistiques sur :

- la gigue (section 1.4.2.1) mesurée,

- les taux moyens de perte de paquets,
- le nombre de paquets reçus ou transmis,
- des informations sur les participants de la session.

Il fournit un retour sur la qualité de la distribution des données et transporte l'information de session minimum. Les paquets RTCP ne transportent évidemment pas de données temps réel mais seulement des informations relatives aux flots de données RTP. Le protocole RTCP propose 5 types de paquets pour transporter des informations de contrôle :

- *SR* : Sender Report qui contient à la fois des informations sur les flux émis et les flux reçus.
- *RR* : Receiver Report qui contient des informations sur les flux reçus quand le récepteur n'est pas un émetteur.
- *SDES* : Source Description items (*CNAME* nom d'utilisateur unique de la forme usager@serveur, *NAME* : nom de la source, *EMAIL*, *PHONE*,...)
- *BYE* : Message envoyé par un participant quand il quitte la session.
- *APP* : fonctions spécifiques à l'application.

1.3.4 Le mixage audio

Dans le cadre de la conférence audio à plusieurs participants, il existe deux possibilités pour effectuer le mixage audio suivant l'architecture de la conférence (cf. la section 1.2). Soit les données à mixer sont au niveau d'un pont, soit au niveau d'un récepteur. Le mixage s'effectue sur les données décodées. Il convient de préciser que le pont n'ajoute évidemment pas au contenu mixé à destination d'un participant la contribution audio de ce dernier.

1.4 Les problématiques liées aux communications en général et au transport de la Voix sur IP

Dans ce paragraphe, les principaux inconvénients de la VoIP et des communications en général vont être présentés. Pour un utilisateur, ceux-ci jouent grandement dans la perception de la qualité de service et de la qualité vocale.

1.4.1 Les problématiques liées aux communications en général

1.4.1.1 L'écho acoustique côté locuteur

Ce phénomène est très connu notamment dans le cas de l'utilisation de terminaux en mains-libres, c'est-à-dire lors d'une restitution sur des haut-parleurs.

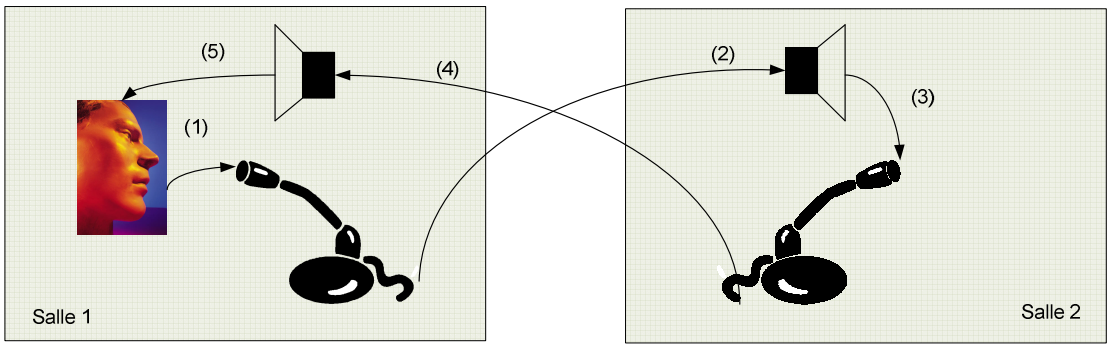

Figure 1.12 Illustration de l'écho acoustique

Le phénomène est illustré Figure 1.12 pour le locuteur de la salle 1. Sa voix est captée tout d'abord par son microphone, puis restituée sur le haut-parleur de la salle 2. C'est à ce moment que l'écho est généré, puisque le microphone de la salle 2 capte ce qui sort du haut-parleur et le renvoie vers le haut-parleur (ou un casque par exemple) de la salle 1. Dans ce cas, le locuteur de la salle 1 entend sa propre voix avec un décalage temporel dû à la transmission, ce qui provoque un phénomène de gêne si ce décalage est supérieur à 30 ms [9].

Des techniques de réduction d'écho seront vues dans le chapitre suivant.

1.4.1.2 Les codeurs

Comme cela a été vu dans la section 1.3.2, le codage de compression sert à réduire la bande passante au niveau réseau. Outre le fait qu'ils contribuent au retard (cf. la section 1.4.2.1), les codeurs de parole sont source de dégradation de la qualité du signal audio.

Suivant les configurations testées (architecture centralisée dans le cadre de la conférence par exemple – section 1.2.1) ou lors d'un appel IP vers un poste RTC classique, des transcodages peuvent être effectués. Cela correspond en fait à une succession d'encodage-décodage avec différents codeurs (potentiellement les mêmes). Ces transcodages dégradent encore plus la qualité, puisque cela correspond à l'encodage de parole ayant déjà subi une dégradation par un précédent encodage-décodage.

1.4.2 Les problématiques liées à la VoIP

1.4.2.1 Le délai

Le délai ou retard lors d'une communication en Voix sur IP est dû à 5 types de problèmes [38] :

- Le retard dû au couple codeur-décodeur. Celui-ci a généralement besoin d'un délai d'une trame pour fonctionner, mais ce retard peut coïncider avec celui du système d'exploitation. Cependant, certains codeurs ont besoin de connaître quelques échantillons en avance (look-kahead) pour éventuellement anticiper une perte de trames. Enfin, certains traitements plus complexes au sein des codeurs-décodeurs comme une transformée peut rajouter l'équivalent d'une trame de retard.
- La gigue ou jitter : du fait que les files d'attente des routeurs sont plus ou moins remplies de manière aléatoire, les paquets IP sont susceptibles d'arriver à destination avec des délais différents. Ils ne peuvent donc être joués directement sous peine de laisser des silences entre les trames audio. Afin d'éviter de trop dégrader la qualité audio (voir la section 1.7.2), des buffers de gigue sont mis en place afin de resynchroniser les trames. Cela introduit cependant du délai mais qui est jugé moins gênant que la perte de la trame [71]. Ces buffers permettent aussi de remettre les paquets dans le bon ordre afin d'éviter le déséquencement si les paquets suivent des chemins différents (cf. la section 1.3.3.1).
- Le délai de transmission sur le réseau et le passage entre les différents réseaux.
- L'influence du système d'exploitation et de sa capacité à aller récupérer les données mises dans le buffer par la carte son lorsque celui-ci est plein. Suivant le système d'exploitation, les données mémorisées dans le buffer de la carte son seront disponibles plus ou moins rapidement, entrainant éventuellement un retard.
- Le nombre de trames audio contenues dans un paquet IP. Plus ce nombre est important, plus le délai est important puisque l'on attend un certain temps avant d'envoyer les données. Ceci a néanmoins l'avantage de diminuer le débit du fait de l'économie d'entêtes IP.

Pour des soucis d'interactivité entre les participants, le délai de transmission total doit être inférieur à 150 ms [44] pour une communication optimale.

1.4.2.2 La perte de paquets

La perte de paquets est l'une des principales différences par rapport au réseau classique RTC où la perte de signal est impossible. Elle est due à la congestion du réseau à un instant donné qui aboutit au rejet des paquets au niveau d'un routeur.

Il est à noter qu'il existe, intégrés à certains codeurs, des algorithmes de correction de perte de trames. Ceux-ci seront utilisés dans le cadre de la thèse. Il convient de distinguer deux méthodes importantes [71] souvent retenues dans ce domaine : la correction de la perte de paquets (PLC - *Packet Loss Concealment*) et la correction d'erreur par anticipation (FEC - *Forward Error Correction*).

- Pour la PLC, aucune donnée supplémentaire n'est envoyée en anticipation au récepteur qui doit donc faire au mieux. La solution par défaut est l'insertion d'un silence de la longueur de la trame perdue. Une autre solution, meilleure en termes de qualité audio (voir section 1.7.2), consiste à remplacer le phonème perdu par un bruit, une tonale ou par lui-même suivant que le son est respectivement voisé, non voisé, ou seulement une partie de ce phonème. La dernière possibilité est l'interpolation en fonction de la trame précédente. L'avantage majeur de la PLC est qu'elle n'introduit pas de retard supplémentaire.
- Pour la FEC, des données supplémentaires redondantes sont transmises au sein d'une trame par l'émetteur, concernant la trame précédente ou la suivante. Celles-ci sont utilisées par le récepteur en cas de perte de trames. Cela aboutit généralement à la restitution d'une bonne qualité audio par rapport à la PLC, mais cela introduit encore un délai supplémentaire et une augmentation de débit. Son utilisation principale est généralement le streaming qui a moins de contraintes de délai par rapport à la VoIP.

1.4.2.3 Les erreurs binaires

Elles se produisent notamment sur des réseaux de données sans fil ou à technologie d'accès filaires tel que le DSL (Digital Subscriber Line).

Elles peuvent aboutir à une suppression du paquet IP si une erreur se produit (vérification possible grâce à la somme de contrôle de l'en-tête, ou en anglais header checksum d'un paquet).

1.4.2.4 La traduction d'adresses réseau

Le principe de base du traducteur d'adresses réseau *(Network Address Translator*) est de savoir traduire, à la volée, une adresse privée en une adresse publique ou l'inverse, comme l'illustre la Figure 1.13. Il modifie également les valeurs des différents checksums des entêtes IP, TCP et UDP, afin que les contenus des paquets ne soient pas considérés comme erronés [89]. Il est d'ailleurs à noter que le traducteur d'adresses peut être évidemment à l'interface de deux réseaux privés.

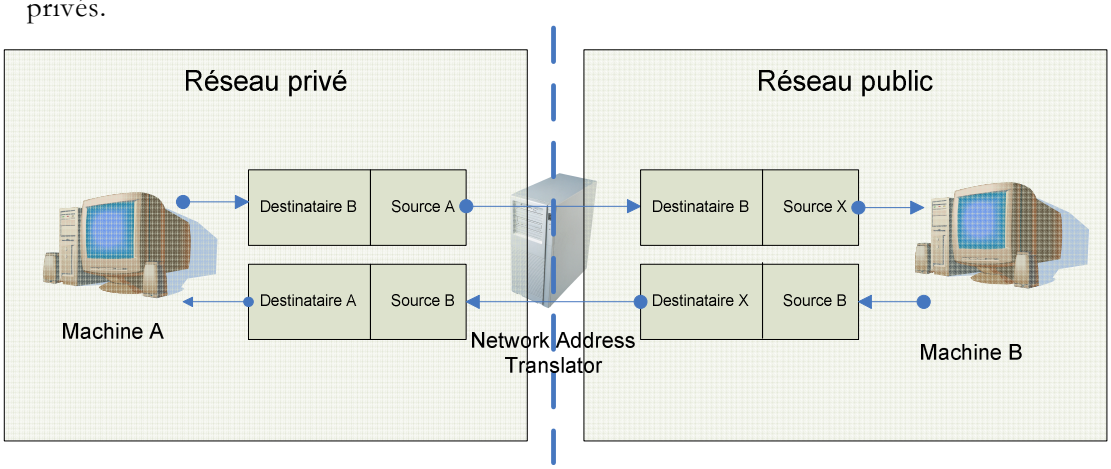

Figure 1.13 Illustration de la NAT

La traduction d'adresses réseau (NAT - *Network Address Translation* en anglais) effectuée par le *Network Address Translator* est un des problèmes totalement liés au réseau IP et à sa croissance. En effet, le nombre de plus en plus important de machines nécessitant une adresse IP, additionné au nombre limité de ces dernières (adresse IPv4 codée sur seulement quatre octets)

et à leur gaspillage au début d'Internet, ont contribué à rechercher une solution permettant de résoudre ce problème d'allocation. Publié en mai 1994, le but premier de la NAT était de pallier le manque croissant des adresses IPv4 disponibles. La NAT ne devait être que provisoire, mais elle semble à présent "condamnée" à durer.

La traduction d'adresses a aussi été utilisée pour protéger les réseaux privés des intrusions venant d'Internet. En effet, un PC, avec une adresse privée, ne peut pas être contacté de l'extérieur, tant qu'il n'y a pas de lien le concernant dans la table de correspondance du traducteur d'adresses réseau. De plus, la présence d'un lien n'implique pas toujours un moyen de contact possible [10] (par exemple *Address-Dependent Filtering*, *Address and Port-Dependent Filtering*,…).

Néanmoins, les traducteurs d'adresses réseau n'apportent pas que des avantages. En effet, en Voix sur IP en SIP notamment, les problèmes sont nombreux :

- Adresses IP privées locales, non routables, échangées par l'intermédiaire de protocoles de haut niveau comme SIP, ne peuvent aboutir à un appel entre deux entités sur deux réseaux différents.
- De même, il existe des problèmes liés aux réponses à des requêtes au niveau signalisation SIP et à la génération de nouvelles requêtes pour l'entité ne se trouvant pas derrière le traducteur d'adresses réseau.
- De même, se présentent des problèmes liés à l'annonce des canaux RTP et RTCP dans les messages SDP.
- Problèmes dus à la suppression des liens dans la table de correspondance des traducteurs d'adresses réseau, lors de l'absence de trafic de paquets RTP pendant un temps trop long, lors d'une mise en *mute* d'un participant par exemple.

L'IETF propose des solutions encore en cours de normalisation pour résoudre les problèmes générés par l'utilisation de traducteurs d'adresses réseau en SIP. Nous pouvons citer les 3 plus connues : ICE (Interactive Connectivity Establishment) [76], TURN (Traversal Using Relays around NAT) [79] et STUN (Session Traversal Utilities for NAT) [78].

1.4.3 La rentabilité pour un opérateur

Il en ressort qu'il est possible de faire de la conférence audio gratuitement sur le réseau public qu'est Internet notamment dans le cas d'une architecture distribuée ou lorsqu'un terminal devient mixeur pour les autres (cf. Figure 1.5). Il est clair que cela entraîne un chamboulement dans le monde des télécommunications dans lequel les revenus des opérateurs historiques obtenus grâce à la téléphonie classique étaient importants. Les clients ont ainsi un autre moyen de communiquer et ils en profitent. Skype en est le parfait exemple.

La solution des opérateurs historiques, pour contrecarrer ces solutions peu rentables pour eux, est de proposer généralement des services de qualité supérieure notamment pour l'audio ou des fonctionnalités supplémentaires notamment grâce à l'avantage d'une solution centralisée pour laquelle il contrôle facilement les conférences et les coûts des communications.

1.5 Avantages de la VoIP

Malgré les inconvénients vus ci-dessus, les avantages de la VoIP sont nombreux par rapport au réseau classique commuté.

1.5.1 La qualité audio

Tout d'abord, les réseaux IP permettent plus facilement une évolution vers une qualité audio (cf. la section 1.7.2) bien supérieure à la qualité historique du RTC (voir introduction à la section 1.2). La qualité des communications n'est pas limitée à la qualité téléphonique audio standard (300-3400 Hz). Actuellement des codeurs permettent de transporter à des débits raisonnables des contenus wideband (50-7000 Hz, voir section 1.3.1) voire super wideband (SWB de bande 50-15000 Hz) avec une très bonne qualité audio.

1.5.2 Le couplage audio/vidéo/données

Un des avantages clés, outre la qualité audio, est la possibilité de coupler facilement des conférences audio/vidéo/données, facilitant ainsi la gestion de ces types de conférences (cf. section 1.7.3.1) en plein essor.

1.5.3 Les choix d'architectures

Comme cela a été vu dans la section 1.2, il existe de nombreuses architectures possibles contrairement au réseau téléphonique classique. Cela permet de s'adapter aux besoins définis pour le service de conférence audio proposé, et aux limites du matériel.

1.6 Un appel VoIP

A présent que les différentes entités physiques et les architectures sont définies, nous allons nous intéresser à deux exemples d'appel en VoIP, afin de montrer comment le protocole de signalisation SIP permet de gérer un appel. En préambule, le protocole SIP sera introduit.

1.6.1 Les principales bases de SIP

Initialement, SIP est un protocole de l'IETF visant à établir, modifier et terminer des sessions multimédia. Il a été défini une première fois dans la RFC 2543 [77] en mars 1999, mais ce document restait trop vague. En juin 2002, une nouvelle version décrite dans la RFC 3261 [82] a été spécifiée. SIP est un protocole de signalisation appartenant à la couche application du modèle OSI, comme illustré Figure 1.14. SIP possède l'avantage de ne pas être associé à un médium particulier et est censé être indépendant du protocole de transport (UDP, TCP, …).

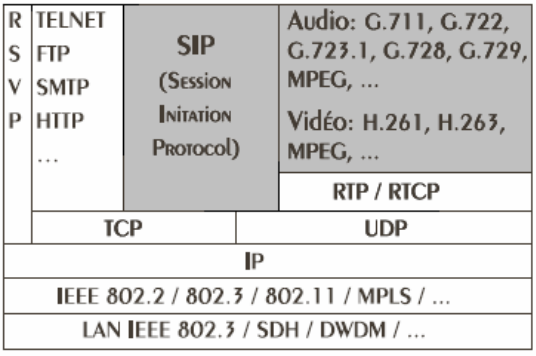

Figure 1.14 Architecture des protocoles selon SIP

Les fonctions principales de SIP sont les suivantes :

- *User Location* : qui permet de localiser un terminal utilisé pour communiquer.
- *User Capabilities* : qui détermine quels médias vont être échangés ainsi que les paramètres associés.
- *User Availability* : qui détermine si le poste appelé souhaite communiquer.
- *Call Setup* ou *Ringing* : qui avertit les parties appelante et appelée de la demande d'ouverture de session et mise en place des paramètres d'appel.
- *Call Handling* : qui gère le transfert et la fermeture des appels.

1.6.1.1 Quelques définitions

Tout d'abord, tous les messages échangés en SIP se font sous le modèle client-serveur, un terminal pouvant tantôt être la partie serveur d'une transaction, tantôt la partie client.

Les points source et destination sont appelés des adresses SIP ou URI (*Uniform Resource Identifier*), ces dernières sont de la forme :

sip :infos-utilisateur@domaine

Avec :

- *infos-utilisateur* de la forme "nom_utilisateur :mot_de_passe".
- *domaine* de la forme "nom_de_domaine :port" ou "adresse_IP :port".

Il y a deux types particuliers d'adresses en SIP : *address-of-record* et *contact address*. Le premier spécifie un nom unique, connu de tous qui permet d'être contacté (qui existe toujours et qui n'a pas de réalité physique) et l'autre correspond à une adresse physique où l'on est réellement à un instant *t* (adresse qui varie si l'on change de terminal par exemple).

1.6.1.2 Les messages SIP

Ils se décomposent en deux catégories : les requêtes et les réponses qui ont la forme générique suivante, Figure 1.15 :

Figure 1.15 Format du message SIP

La Figure 1.16 montre un exemple de message SIP typique.

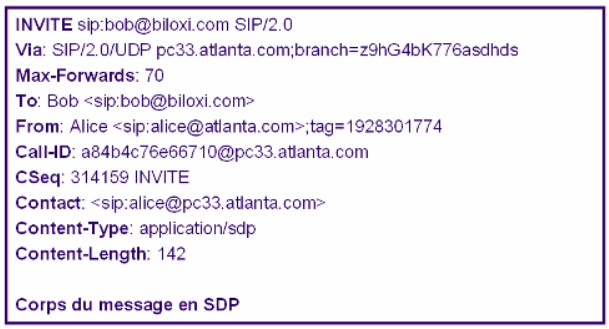

Figure 1.16 Exemple de message SIP

La *Start Line* dépend du type de message comme on le verra dans les deux sections suivantes.

Les messages SIP ont un entête commun (*general header*) composé des principaux paramètres suivants :

- *Via* : qui permet d'identifier les étapes que doit suivre la réponse, ainsi que la version de SIP. Chaque entité, par laquelle passe le message, y rajoute en tête une ligne en spécifiant son adresse SIP de contact. Il est à noter que l'entité suivante rajoute généralement sur la ligne *Via* de l'entité la précédant un paramètre *received* qui indique l'adresse IP de ce dernier. Cela permet de simplifier le traitement pour la réponse. Ainsi cette dernière pourra suivre exactement le même chemin que la requête. Chaque entité, au retour, supprime l'information lui correspondant dans *Via*. Il est aussi ajouté un paramètre *branch* qui identifie la transaction. Toute réponse de la partie serveur utilise cette valeur de *branch*. Toute requête initiée par ce client gardera, de même, cette valeur. Ce paramètre permet aussi de déterminer la version de SIP (RFC 2543 [77] ou RFC 3261 [82]). En effet, un paramètre *branch*, commençant par le *magic cookie* de valeur "z9hG4bK", identifie la version RFC 3261 [82] de SIP.
- *To* : qui est l'identifiant de l'appelé. Il est conçu comme le *From*. Il n'a pas de *tag* initialement mais il sera mis dans les requêtes suivantes dès que l'entité serveur aura répondu une fois. Le champ *To* d'une réponse est le même que celui du champ *To* de la requête associée. Il ne désigne donc pas la cible de la réponse mais celle de la requête.
- *From* : qui est l'identifiant de l'appelant. Il contient un nom d'usage (ou identité logique), entouré de "" en cas d'ambiguïté syntaxique, et une URI (ou *address-of-record* qui est une URI particulière) entourée de <>. Exemple : *"John" <john@g_un_bo_domaine>*. Le paramètre *From* est souvent associé à un *tag* qui sert à identifier un dialogue. Le champ *From* d'une réponse est le même que celui du champ *From* de la requête associée. Il ne désigne donc pas l'émetteur de la réponse mais celui de la requête.
- *Call-ID :* qui est un identifiant globalement unique pour un appel. Il est composé généralement d'une partie localement unique ainsi que d'un identifiant global comme une adresse IP ou un nom de type Domain Name System. (*12345@mon_bo_domaine.com*)
- *Cseq* : qui est un identifiant qui sert à rapprocher les requêtes des réponses correspondantes au sein d'une même transaction. Il est composé d'un numéro et du nom de la méthode. Les réponses à une requête doivent avoir le même numéro de *Cseq*. Ce numéro s'incrémente pour chaque nouvelle requête envoyée dans un dialogue sauf pour un *ACK* ou un *CANCEL* car ils ont trait à la requête précédente.
- *Contact :* adresse physique ou réelle (*contact address*) de l'expéditeur du message. Cela permet à l'utilisateur distant de le contacter directement sans passer par une entité intermédiaire de routage de signalisation SIP appelée proxy.

Le CRLF est un retour à la ligne.

Le corps du message, selon la méthode, indique des informations sur la progression de la requête, la description de la session, la description des destinations ou des services intermédiaires et des causes d'erreurs. Pour l'établissement ou la modification des sessions par exemple, c'est la syntaxe SDP [35] qui est utilisée.

1.6.1.2.1 Les requêtes SIP

La *Start Line* d'une requête correspond au nom de la requête suivi du champ *Request URI* et de la version de SIP. Le champ *Request URI* indique généralement l'adresse d'enregistrement (*address-of-record*) puis ensuite quand l'adresse précise est connue au niveau du proxy du domaine concerné, l'adresse de contact (*contact address*).

Voici les différentes requêtes SIP : *ACK* (voir ci-dessous), *BYE* (relâche un appel), *CANCEL* (annule la requête précédente), *INFO* (transporte une information comme un DTMF), *INVITE* (établit un appel), *MESSAGE* (pour l'envoi de messages instantanés), *NOTIFY* (notifie un évènement), *OPTION* (message pour s'enquérir des méthodes supportées), *PRACK* (confirme la réception d'une réponse provisoire), *REFER* (redirection d'appel), *REGISTER* (permet d'enregistrer leur localisation sous forme d'une ou plusieurs adresses particulières : *contact addresses*), *SUBSCRIBE* (demande de notification d'évènements) et *UPDATE* (met à jour les paramètres média).

La requête *ACK* confirme une réponse finale uniquement pour la réponse *INVITE* et n'est jamais acquittée (3ème volet du processus en 3 coups pour un message *INVITE*). Les réponses de type 2xx (voir paragraphe suivant) sont acquittées par les entités initiatrices du message *INVITE* et les autres types de réponses définitives (3xx, 4xx, 5xx ou 6xx) peuvent être acquittés par des proxys intermédiaires.

1.6.1.2.2 Les réponses SIP

La *Start Line* d'une réponse SIP correspond à la version de SIP, au code d'état ainsi que sa signification (*Reason-Phrase*).

Une réponse à une requête est caractérisée par un code appelé code d'état. Les codes d'état du type *1xx* sont ceux correspondant aux réponses provisoires, qui évitent la retransmission des requêtes *INVITE*. Les autres sont des réponses finales. Quand un terminal ne reconnaît pas un code *Cxx*, (ex : 385) il le remplace par *C00* (ex : 300).

La Figure 1.17 montre les différentes catégories de codes d'états.

Code d'état	Signification
1xx	Information
2xx	Succès
3xx	Réacheminement
4xx	Erreur requêtes
5xx	Erreur serveur
6xx	Erreur globale

Figure 1.17 Récapitulatif des codes d'état

1.6.1.3 Négociation des flux média

La négociation des flux média se fait par le protocole SDP (Session Description Protocol). Le but de SDP est de fournir des informations sur une session multimédia, elle-même étant définie comme un ensemble de flux média pendant une durée donnée. A la base, SDP était plutôt fait pour décrire des sessions multicast. Du fait de la redondance de certains champs présents dans les messages SIP, certains champs spécifiques de SDP ne sont pas utilisés. Il faut savoir que les messages *INVITE*, *2xx*, *UPDATE*, les réponses provisoires et leurs validations peuvent contenir ces descriptifs de session.

SDP inclut :

- Le nom et le but de la session.
- Le temps d'activité de la session : SDP peut fournir une liste pour le temps de début et de fin ou de périodicité de la session.
- Les médias de la session.
- Le type de média.
- Le protocole de transfert.
- Les informations pour recevoir ces médias (adresses, port, format)

Les spécifications SDP sont en fait encapsulées dans des messages SIP. SDP est pratique car il correspond à un encodage texte qui est beaucoup plus facile à utiliser qu'un encodage de type binaire de type ASN.1 (utilisé en H.323). Il est cependant moins efficace en termes de fiabilité lors de la génération des codes de sérialisation.

1.6.1.3.1 Le modèle offre/réponse

Le modèle proposé par la RFC 3264 [84] est celui de offre/réponse comme l'illustrent les exemples des Figure 1.18 et Figure 1.19. Ce modèle sera détaillé de manière complète car il est important de comprendre son fonctionnement.

Dans un premier temps, on ne s'intéresse qu'à la première figure. Voici la signification des différents champs :

- Le champ "*v*=" donne la version du protocole SDP.
- Le champ "*o*=<*username*><*session id*><*session version*><*network type*><*address type*><*address*>" donne l'initiateur de la session ainsi que son *username* et les adresses du terminal de l'utilisateur.
- Le champ "*s*=<*session name*>" fournit le nom de la session et il est généralement vide dans le cadre de SIP pour l'unicast.
- Le champ "*c*=<*network type*><*address type*><*connection address*>" contient les informations sur les adresses et leurs types.
- "*t*= " fournit le temps de la session, généralement à 0 0 pour les sessions unicast sous SIP.

Une offre peut contenir plusieurs propositions de flux média, chacune incluse dans une ligne commençant par "m" et suivie de lignes d'attributs optionnels. Une ligne "m" spécifie le type de média (audio, vidéo…), le port utilisé en réception pour ce type de média ainsi que le profil utilisé (RTP/AVP pour Audio Video Profile) et les types de format RTP (voir 1.3.3.2.5). Un numéro de port à zéro signifie (même si cela peut paraître surprenant dans une offre !) que le flux offert ne doit pas être utilisé. Cette mise à zéro du numéro de port sert à informer le distant que ce média est supporté.

Des lignes "a" sont associées aux lignes "m". Elles sont utilisées pour spécifier non seulement le type de flux (RTP dans le cas de la voix sur IP donc *rtpmap*), le type de chaque codeur (PCMU/8000, PCMA/8000, H261/90000…) et son type de format RTP associé, mais aussi pour spécifier si le flux (quel que soit le codec qui sera ensuite utilisé) est bidirectionnel (*sendrecv*), unidirectionnel (*recvonly* en réception ou *sendonly* en émission) ou inactif (*inactive*). Le paramètre par défaut est "*a=sendrecv*".

Revenons en aux lignes "m". Pour celles-ci, il est à noter que les flux RTCP (qui existent toujours sauf dans le cas où le numéro de port est à 0) sont envoyés à la même adresse IP mais à un numéro de port incrémenté de un par rapport au numéro de port spécifié. Par convention, les numéros des ports RTP sont pairs, et ceux de RTCP impairs.

Il y a deux cas qui se produisent concernant ces lignes "m" :

- Lorsqu'il y a plusieurs lignes "m", cela indique les flux que le terminal peut envoyer et/ou recevoir simultanément. Il est prêt à recevoir immédiatement et en parallèle les flux RTP correspondants, dans le cas " *a=recvonly* " ou " *a= sendrecv* " sur le port spécifié. Dans le cas " *a=sendonly* " ou " *a= sendrecv* ", il ne peut évidemment rien envoyer puisqu'il n'a pas encore de réponse du terminal distant mais cela indique aussi qu'il peut envoyer, en parallèle des flux reçus, des flux RTP.
- Lorsqu'il y a une ligne "m" avec plusieurs choix possibles pour un type précis de média (comme pour le choix d'un codec audio dans l'exemple), cela signifie que l'appelant a une préférence. Dans ces cas là, seul un codec doit être choisi parmi l'ensemble proposé mais il peut être changé en cours de session. Dans l'exemple Figure 1.18, pour l'audio, le terminal préfère le codec 0 (qui correspond au G.711 loi μ) au codec 8 (qui correspond au G.711 loi A).

Tout terminal, ayant envoyé une offre, doit donc être prêt à recevoir immédiatement des flux sur les ports qu'il a spécifiés. Il pourra envoyer des flux seulement lorsqu'il aura reçu l'adresse du destinataire incluse dans la réponse de ce dernier.

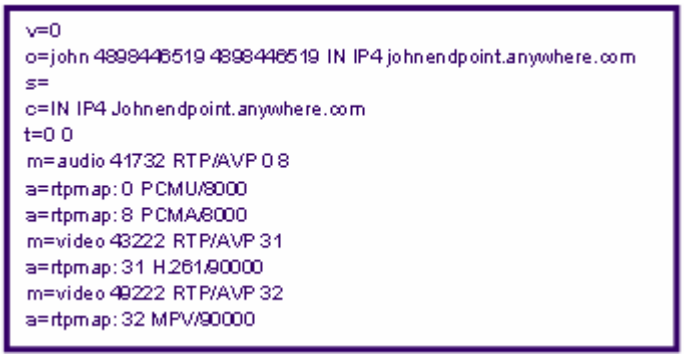

Figure 1.18 Négociation en Emission ⇔ offre

Concrètement, Figure 1.18, le terminal "john" propose d'établir des flux bidirectionnels (car il n'y a pas de ligne "*a=xxxx*" qui indique le contraire) pour tous les flux "m" proposés. Il peut établir un canal audio avec de préférence comme codec le G.711 loi μ mais il acceptera aussi le G.711 loi A. De même, il propose deux canaux vidéo avec les codecs H.261 et MPV qu'il peut supporter en parallèle avec le canal audio. Mathématiquement et sans la notion de préférence, cela correspondrait à : **(**G.711 loi μ "OU EXCLUSIF" G.711 loi A) "ET" H.261 "ET" MPV.

Les numéros de port associés aux lignes "m" sont ceux des ports RTP en réception (soit 41732, 43222 et 49222). Les numéros des ports RTCP de réception sont donc les numéros de ports précédents incrémentés de un.

Intéressons-nous à la Figure 1.19 qui montre un exemple de réponse du terminal "mark" à l'offre du terminal "john".

Figure 1.19 Négociation en Réception ⇔ réponse

Quelques contraintes sont à respecter. Sur la ligne "o", on peut remarquer qu'il y a un nouveau numéro de session car la réponse provient d'une autre entité. La ligne "*t*= " doit être identique à l'offre.

De plus, dans une réponse, il doit y avoir le même nombre de lignes "m" que pour l'offre. Un flux (bidirectionnel ou unidirectionnel) rejeté aura un port à 0 comme le montre le cas du flux vidéo bidirectionnel avec le numéro de codec 32. Il est à noter que les flux RTCP sont envoyés à la même adresse IP mais à un numéro de port incrémenté de un par rapport au numéro de port spécifié.

Dans le cas où le flux est accepté, si dans l'offre, la ligne "a" spécifiait :

- "a=sendrecv" : la réponse doit spécifier "a=inactive" ou "a= sendonly" ou "a=recvonly" ou "a=sendrecv".
- "a=recvonly" : la réponse doit spécifier "a=inactive" ou "a= sendonly".
- "a=sendonly" : la réponse doit spécifier "a=inactive" ou "a=recvonly".
- "a=inactive" : la réponse doit spécifier "a=inactive".

Chaque type de réponse doit au moins contenir un des choix proposés dans une offre "m" (sinon le flux est rejeté et le port est mis à 0). Elle peut en inclure cependant d'autres mais dans le cas "*sendonly*" ou "*sendrecv*" pour le terminal *mark* ayant reçu l'offre, aucun flux ne pourra être

envoyé sans l'accord du terminal offrant *john*. La réponse précise aussi le port choisi en réception pour les flux média RTP (soit 51762 et 53222). Bien que le terminal *mark* qui répond, puisse lister dans une ligne "m" son ordre de préférence, il est recommandé qu'il suive le même ordre relatif que le terminal *john* afin de pouvoir établir le même codec dans les deux sens. Dans le cas de l'exemple, le terminal *mark* appelé a choisi le codec audio G.711 loi A au lieu de celui en loi μ qu'il ne peut manifestement pas supporter. Le terminal *mark* a refusé d'avoir deux flux vidéo en parallèle.

Une fois la réponse envoyée, le terminal *mark* doit s'attendre à recevoir ou à envoyer des flux média de n'importe quel format listé dans celle-ci. Contrairement au terminal *john* qui a fait l'offre, il peut envoyer directement du média s'il a accepté un des codecs proposés.

Lorsque le terminal *john* reçoit la réponse, il ne peut envoyer des flux médias que vers ceux validés dans cette réponse. Il doit choisir la première proposition dans la réponse dans le cas où plusieurs codecs sont proposés dans une même ligne "m".

1.6.1.4 Les différentes entités fonctionnelles d'un dialogue SIP

SIP utilise plusieurs entités fonctionnelles utilisées lors d'une communication :

- Un *User Agent* (UA) qui est le terminal de l'utilisateur.
- Un *Registrar* qui permet l'enregistrement des terminaux.
- Un *Location Server* qui est capable de fournir la localisation courante d'un utilisateur identifié par son adresse.
- Un *Proxy Server* qui peut être *stateless* (sans état), *stateful* (il peut contrôler l'état de l'appel) ou *forking* (il peut dupliquer un message *INVITE* vers plusieurs terminaux). Cette entité permet de router les messages SIP.
- Un *Back-to-Back User Agent* pour les entités qui pourraient être des UA mais qui sont utilisées en tant que serveurs par exemple entre deux réseaux.

Un même *proxy* peut être plusieurs des entités précédemment citées. Un exemple, Figure 1.20, permet d'illustrer l'utilisation de ceux-ci.

1.6.1.4.1 Le User Agent

Les points d'extrémité, qu'utilise SIP pour négocier les caractéristiques de session, sont appelés les UA. Les UA sont en fait généralement des applications qui tournent sur le PC de l'utilisateur. Cela peut être aussi des téléphones cellulaires, des IP phones, des passerelles IP/RTC… Les UA se décomposent en deux sous entités comme toute entité SIP : UAS (*User Agent Server)* et UAC (*User Agent Client*) puisque tous les dialogues SIP se font sous le modèle client-serveur. L'UA est tantôt l'un, tantôt l'autre. L'UAC envoie des requêtes et reçoit des réponses appropriées de la part d'un serveur distant. L'UAS reçoit des requêtes de la part d'un client distant et envoie des réponses appropriées.

1.6.1.4.2 Localisation et Enregistrement avec le *Registrar* et le *Location Server*

Le protocole de signalisation SIP doit pouvoir "retrouver" le destinataire d'un appel quel que soit l'endroit où il se trouve à un moment donné. Les Serveurs proxys et les serveurs de redirection doivent en effet pouvoir déterminer où envoyer les messages. Cela est possible grâce à l'envoi de la part d'un UA de ses adresses de contact (*contact addresses*) à un *Registrar.*

Un UA a trois moyens de trouver son *Registrar :* par configuration, par l'utilisation de l'*address-of-record* ou par multicast à l'adresse 224.0.1.75.

Cet UA lui envoie une requête *REGISTER* qui contient son *address-of-record* (dans le champ *To*), éventuellement des *contact addresses* (de la forme d'une URI SIP, d'un numéro de téléphone ou une adresse IP dans le champ *Contact*), la personne qui fait l'enregistrement (dans le champ *From*) ainsi que le domaine du *location service*. Un message *REGISTER* permet d'ajouter de

nouveaux liens, d'en supprimer ou de savoir lesquels sont actifs. Un terminal doit s'enregistrer périodiquement.

Le *Registrar* prévient ensuite le *location service* de la mise à jour des liens entre une *address-of-record* et une ou plusieurs *contact addresses*.

Dans l'exemple Figure 1.20, un UA s'enregistre auprès du *Registrar*, ce dernier spécifiant ensuite au *Location Service* le lien entre sip:Carole@chicago.com (*address-of-record*) et sip:Carole@cube2214a.chicago.com (*contact address*). Ensuite lorsqu'une requête *INVITE* avec l'*address-of-record* de carole arrive au proxy du domaine chicago.com, celui-ci fait une demande au *Location Service*, afin de savoir quelle est la *contact address* de Carole. Une fois qu'il a obtenu la réponse, il lui envoie le message. Il est à noter que le *Registrar* et le Proxy sont des entités logicielles qui peuvent se trouver sur la même machine.

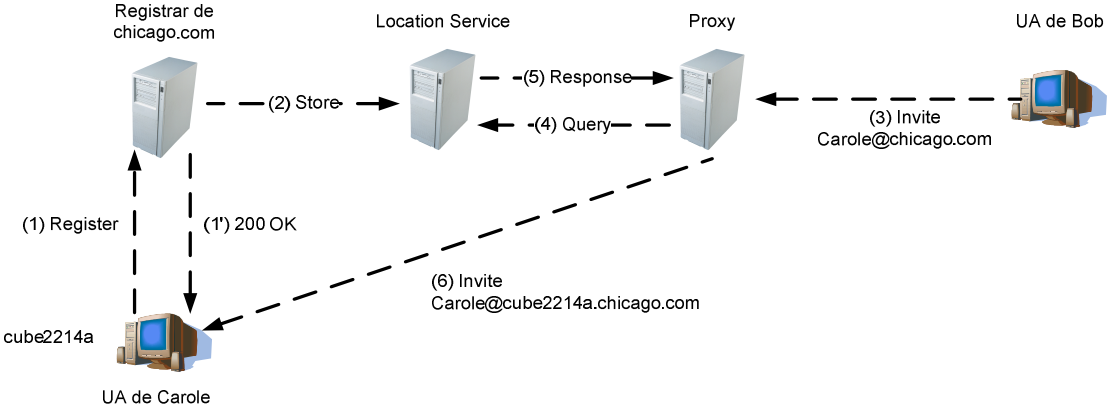

Figure 1.20 Exemple d'enregistrement en SIP

Pour savoir où envoyer un message, les RFC 2543 [77] puis 3263 [81] ont spécifié de manière précise comment localiser les serveurs SIP correspondant à une adresse URI et quelle couche de transport utiliser. Dans la RFC 3263 [81], il est spécifié que si l'URI de destination contient une adresse IP (avec ou sans port) et sans spécifier de protocole alors on utilise UDP. Idem si la cible n'est pas une adresse IP et qu'un port est spécifié. Dans tous les autres cas, si l'URI ne contient ni une adresse IP ni un port ni une spécification de protocole, il faut utiliser les mécanismes des enregistrements DNS (Domain Name System) de type NAPTR (Naming Authority Pointer), définis dans la RFC 3403 [60], pour résoudre en adresse de transport, le lieu où on doit envoyer le message *INVITE*.

1.6.1.4.3 Le serveur de redirection (*Redirect Server*)

Les serveurs de redirection aident à localiser les UA en fournissant une adresse de substitution à laquelle l'utilisateur peut être joint. Il réalise une association d'adresse vers une ou plusieurs nouvelles adresses dans le champ *Contact*. Il répond concrètement à une requête *INVITE* par une réponse de type *3xx*.

1.6.1.4.4 Les Proxy Server

Ils ont un rôle très important en SIP et ils agissent comme un serveur d'un côté (réception des requêtes) et un client de l'autre (retransmission des requêtes). Un proxy peut relayer la requête sans aucun changement, décider d'en vérifier la validité, dupliquer des requêtes, résoudre des adresses… Mais la tâche la plus importante des proxys reste de router l'information au plus proche de la partie appelée. Le plus souvent une invitation aura d'ailleurs traversé plusieurs proxys avant d'atteindre celui qui sait précisément où est l'appelé.

1.6.1.4.5 Exemple de proxy : L'agent d'appel

Un agent d'appel est une fonction qui gère les appels entrants et/ou sortants d'un usager. Il peut réaliser les tâches suivantes :

- Tenter de localiser un utilisateur en redirigeant des messages d'établissement d'appel vers le ou les terminaux appropriés.
- Rediriger l'appel.
- Gérer des règles de filtrage.
- Enregistrer les listes d'appels infructueux.

Un proxy peut choisir de relayer toute la signalisation ce qui est utile dans le cadre de la facturation. Les appels entrants sont contrôlables en forçant le DNS de manière appropriée à passer par un proxy donné contrôlant la facturation. Les appels sortants quant à eux sont envoyés au proxy pour la résolution d'adresse et la transmission du message au proxy suivant.

1.6.1.5 La couche transport de SIP

Les messages sur SIP peuvent être envoyés indépendamment sur UDP ou TCP (voir 1.3.3.2). Si aucun port spécifique n'est précisé dans l'URI SIP (vue ci-dessus) alors celui-ci sera, par défaut, 5060. La RFC 3261 [82] spécifie que notamment dans les cas où la taille de paquets est proche du MTU (*Maximum Transfer Unit*), TCP doit être utilisé.

SIP permet aussi d'assurer une fiabilité sur les couches de transport non fiable comme UDP en se basant sur des mécanismes de retransmission lorsque des réponses provisoires ou définitives (équivalent à une bonne transmission de la requête) n'apparaissent pas dans un laps de temps donné. Il faut noter que seule la dernière réponse définitive est retransmise.

Seul le type de message *INVITE* nécessite un type d'acquittement particulier dit en 3 coups : le serveur envoie une réponse provisoire ou éventuellement une réponse finale, dès réception de ce message pour éviter sa retransmission. En cas de réponse finale, celle-ci est validée par un message *ACK*, concluant ainsi ce processus d'acquittement particulier. En effet, le temps de réponse à un message *INVITE* peut être très variable (il dépend du moment où l'appelant décroche) et il est donc dans ce cas difficile de fixer un laps de temps avant la retransmission (comme pour les autres requêtes), tout en ne sachant pas si le message est bien arrivé.

Le problème des réponses provisoires est qu'elles ne sont pas transmises de manière fiable sous SIP car par exemple : si plusieurs réponses provisoires sont transmises et toutes perdues sauf la dernière, et que cela est fait avant que le laps de temps de non validation d'une requête soit écoulé (avant nouvelle retransmission), alors toutes ces réponses provisoires sont réellement perdues. La méthode PRACK (RFC 3262 [80]) a été mise en place pour solutionner ce problème. Elle stipule que toute réponse provisoire doit être acquittée par un message *PRACK*. Ce message doit être considéré comme une requête SIP normale acquittée par une réponse (*200 OK* par exemple). L'émetteur d'un message *INVITE* doit stipuler qu'il supporte cette méthode, dans le cas où il souhaite l'utiliser.

1.6.2 Un appel VoIP par l'exemple

Maintenant que les bases sont posées, nous allons illustrer cela par un appel VoIP en SIP.

Ce cas suivra l'exemple de la Figure 1.21. Les réponses provisoires, qui ont lieu dans cet exemple, ne seront pas représentées pour simplifier le scénario. Il y en aurait par exemple dès qu'une entité reçoit un message *INVITE*.

Joe, qui appartient à la zone A, veut joindre Bob de la zone B à l'aide de son identité SIP (*address-of-record*) : sip:*Bob@b.com* où *b.com* est le domaine du fournisseur de service SIP de Bob. Comme le terminal de Joe ne connaît pas la localisation de Bob (*contact address*) ou du serveur SIP de son domaine, il envoie le message *INVITE* avec son adresse de contact à son serveur *Proxy.a.com*. L'adresse de ce serveur est obtenue par configuration du softphone (logiciel qui fait l'appel) ou par DHCP (Dynamic Host Configuration Protocol).

Ce serveur proxy, avec l'aide d'un serveur DNS, détermine l'adresse et le type de transport souhaité du proxy de Bob en analysant le nom de domaine associé à l'URI de ce dernier (utilisation des mécanismes des enregistrements DNS de type SRV), en l'occurrence *b.com*.

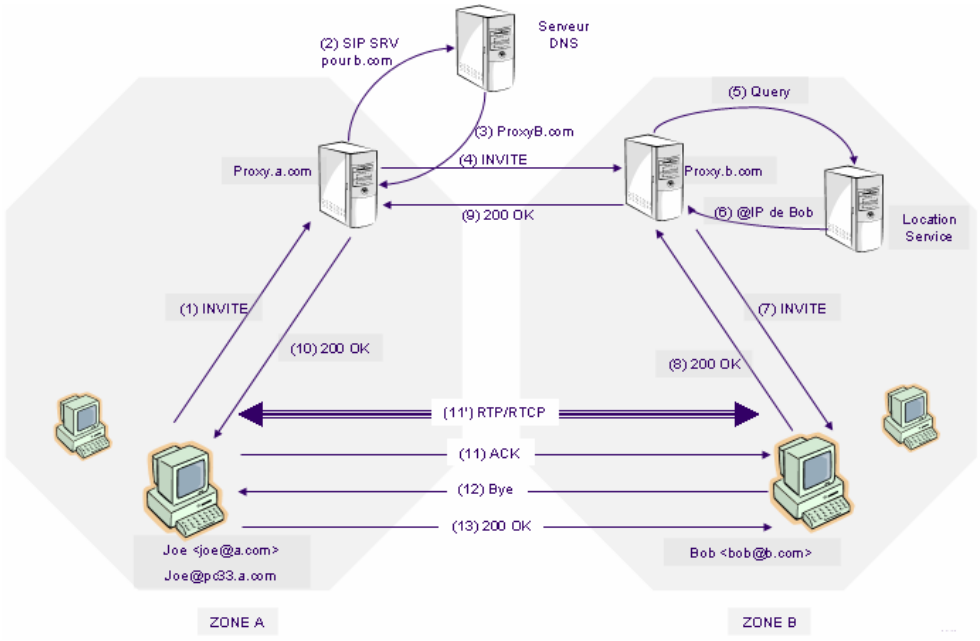

Figure 1.21 Exemple d'appel VoIP SIP

Une fois cette réponse obtenue, il peut faire suivre le message *INVITE* tout en rajoutant dans le champ *Via* (comme l'a fait auparavant le terminal de Joe) son adresse IP et son adresse de contact. Quand le serveur *Proxy.b.com* reçoit le message, il modifie de la même façon le champ *Via* et cherche, dans le *location Service*, l'adresse de contact de Bob. S'il y est enregistré (par la méthode vue dans la section 1.6.1.4.2), cela veut dire que l'on a une adresse où le joindre. Le message *INVITE* avec cette fois le champ Request URI, contenant une *contact address* au lieu d'une *address-of-record*, y est envoyé.

Toutes les réponses provisoires et finale envoyées par le terminal de Bob passeront par le même chemin grâce aux éléments contenus dans le champ *Via* à l'aller. Ces éléments (composés d'adresses IP et d'adresses de contact) sont en fait ajoutés dans l'entête *Via* quand l'UA de Bob envoie un message et chaque proxy intermédiaire, en recevant ce message, va enlever l'élément qu'il a apporté à l'aller dans cet entête et en déduire où envoyer ensuite le message. Dans le cas de l'exemple, on suppose que Bob décroche directement et le message *200 OK*, issu du décrochage, arrive à l'UA de Joe par l'intermédiaire des deux proxys.

Lorsque l'UA de Joe reçoit le *200 OK*, il lui renvoie un *ACK* directement grâce au champ *Contact* du message *200 OK* reçu. Si l'adresse est une adresse de contact URI, il peut aussi utiliser la résolution DNS pour obtenir une adresse IP. Les flux média RTP peuvent à présent passer directement entre les deux UAs, suite à la négociation des flux média contenue dans ces messages (cf. 1.6.1.3). Dans l'exemple, Bob a raccroché et son terminal a envoyé un message *BYE* directement à celui de Joe.

Il est à noter que les proxys pourraient demander à être sur le chemin de toutes les requêtes échangées entre les deux terminaux, en s'inscrivant dans le champ *Record Route* de la première requête *INVITE* envoyée.

Durant toute la session, Bob ou Joe auraient pu changer les paramètres média de la session avec des messages *INVITE* ou *UPDATE*.

1.6.3 Un exemple dans le cadre de la conférence

Prenons un exemple simple correspondant au cas de la conférence centralisée illustré Figure 1.4 a).

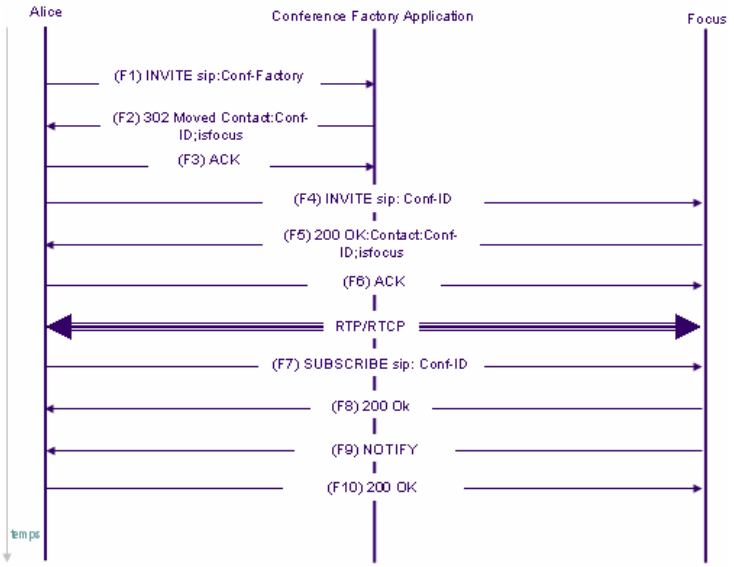

Figure 1.22 Création puis entrée en conférence

Dans cet exemple, Figure 1.22, Alice souhaite créer une conférence. Pour cela, son UA envoie un message *INVITE* (F1) à l'URI de la *Conference Factory* (connu par défaut et dédié à la création de conférence). En réponse, il reçoit un message de redirection (F2), spécifiant dans le champ Contact l'URI du *focus*. La *Conference Factory* aura en fait, par des moyens non-SIP, créé une conférence, dont il renverra l'URI au demandeur par cette redirection. L'UA d'Alice n'a plus qu'à rediriger son appel (on garde donc le même Call-ID) vers l'URI du *focus* de la conférence (F4). Le *focus* accepte l'appel et renvoie une réponse, en spécifiant qu'il est bien le *focus* (champ Contact de F5). Lorsque d'autres participants sont entrés en conférence, on se trouve bien dans le cas où le *focus* a dû effectuer le mixage audio lui-même et envoyer au terminal d'Alice le contenu mixé. On perçoit bien pourquoi cette architecture est similaire à plusieurs dialogues SIP entre chaque participant et le *focus*.

Les messages F7 à F10 correspondent à la souscription d'Alice aux évènements de la conférence. Nous y reviendrons dans le chapitre 3.

Il est à noter qu'Alice peut spécifier la liste des participants [20] qu'elle souhaite inviter dans la conférence.

1.7 La conférence audio spatialisée

Les bases de la conférence audio standard posées, il est maintenant possible d'introduire la conférence audio spatialisée et de définir dans la partie suivante les axes de recherche. Tout d'abord, il est primordial d'être convaincu de l'apport du son spatialisé.

1.7.1 Les apports du son spatialisé

Le son spatialisé permet de placer des sources virtuelles dans l'espace, comme illustré Figure 1.23. Concrètement dans le cadre de la conférence audio spatialisée, des sources virtuelles sont créées et placées à des positions virtuelles autour de l'utilisateur. Chaque participant est associé à

une source virtuelle et toutes les phrases qu'il prononce sont spatialisées pour sembler provenir de sa position virtuelle associée. Les différentes techniques retenues dans le cadre de la conférence audio spatialisée seront décrites dans la section 1.7.4.

Il convient de souligner que nous ne traitons pas de la prise de son naturelle spatialisée. Celle-ci est généralement réalisée avec des microphones adaptés (stéréo, ambisonique, etc.) alors que nous n'utilisons que des microphones monophoniques.

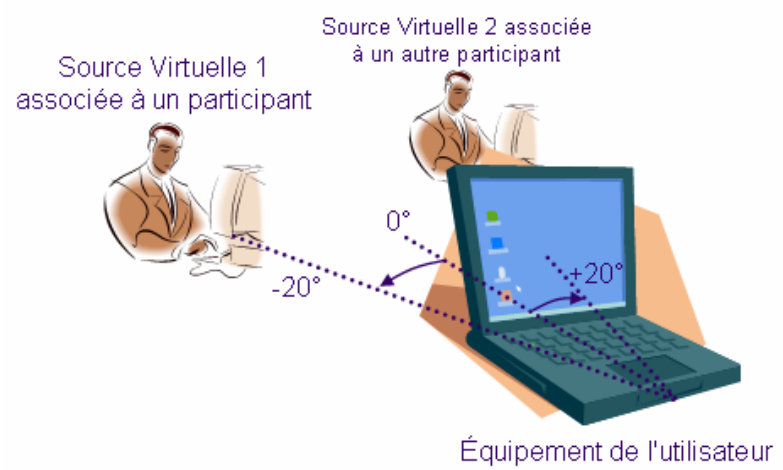

Figure 1.23 Illustration du son spatialisé

Différents articles montrent l'intérêt du son spatialisé par rapport au son monophonique ou stéréophonique. Nous présentons ci-dessous un panel représentatif de ceux-ci.

Le but de [11] a notamment été de comparer différents systèmes de conférences audio à quatre participants sur des critères de mémoire, d'intelligibilité, de compréhension, de focalisation et de préférence.

Des remarques intéressantes sur le fonctionnement de la mémoire court terme sont à noter. Elles sont basées sur le fait que notre cerveau analyse beaucoup mieux qui parle lorsqu'il a une information de type sonore couplée avec une information de type spatiale.

L'importance d'appartenance à un groupe est soulignée avec les notions de : retour des autres participants, savoir qui est présent et surtout qui interrompt ou qui parle. Ce dernier point est notamment renforcé par le son spatialisé.

Les configurations testées dans [11] simulaient une conférence "monophonique" (un seul haut-parleur placé devant les auditeurs), une conférence audio spatialisée avec des participants devant l'auditeur et faiblement espacés (+/-5°, +/- 15°), et une configuration de conférence audio spatialisée pour laquelle les écarts angulaires étaient plus élevés (+/- 20°, +/- 60°). Pour les configurations spatialisées, des haut-parleurs (HP) ont été utilisés pour simuler chaque participant (un HP utilisé pour simuler un locuteur).

Les résultats obtenus pour la reconnaissance des locuteurs montrent que les solutions spatiales sont bien meilleures. Idem pour la confiance en ces résultats et la confiance dans la capacité de donner le point de vue de chaque participant.

Concernant la facilité d'établir qui parlait, les personnes ont de même préféré les solutions spatialisées. L'attention des participants concernant l'identification des personnes a donné le même genre de résultat.

La compréhension est très supérieure pour les solutions spatialisées par rapport aux autres. 89.5 % des participants du test considèrent que le son spatialisé procure un apport non négligeable.

L'ordre de préférence des configurations est le suivant :

- 1) La conférence audio spatialisée avec des participants bien espacés,
- 2) La conférence audio spatialisée avec des participants légèrement séparés,
- 3) La conférence audio "monophonique".

Pour sa part, l'article [67] a notamment montré que, plus le nombre de locuteurs est important, plus le son spatialisé aide les sujets à comprendre ce qui se dit. Cependant à partir de 6 locuteurs concurrents, l'intelligibilité est très faible que le son soit spatialisé ou non.

Des tests réalisés dans l'article [96] montre l'impact du son spatialisé sur la perception en profondeur, sur la performance qu'il apporte par exemple dans la recherche de personnes et sur les notions de présence, de collaboration et d'immersion qu'il entraîne dans un contexte de réalité augmentée (des objets virtuels ajoutés à une scène spatialisée réelle). Les auteurs ont montré que, dans un contexte de jeu, le son spatialisé apporte un plus, intéressant pour l'immersion du joueur.

A titre d'information, [50] montre l'apport du son spatialisé pour des tâches de recherche et de détection dans des environnements virtuels. De plus, il est montré que les sujets, utilisant le son spatialisé, effectuent un apprentissage, ce qui leur permet d'obtenir de meilleurs résultats en termes de rapidité au fil du temps.

En conclusion, le son spatialisé engendre ainsi un apport notable sur les plans de :

- L'intelligibilité, notamment en présence de bruit et plusieurs locuteurs simultanés,
- La compréhension,
- Les effets sur la mémoire et la reconnaissance des locuteurs,
- Le confort d'écoute, qui est un mélange difficilement évaluable des caractéristiques ci-dessus,
- La sensation d'immersion et de réalisme.

1.7.2 Evaluation de la conférence audio spatialisée

Comme il a été vu dans le paragraphe 1.1.2, l'évaluation du service de la conférence audio spatialisée est complexe. Nous nous restreindrons pour notre part à la qualité de la transmission vocale ainsi qu'à une sous-partie de l'efficacité de la conversation qu'est la qualité de la spatialisation (voir chapitre 4).

Dans ce document, par qualité audio, nous entendons la qualité audio subjective globale de la parole (voir la section 1.7.3), entendue par un auditeur à un bout du système, lors de tests d'opinion d'écoute [47] (Test ACR ou DCR par exemple explicité dans le chapitre 4). Chaque auditeur doit noter en tenant compte de toutes les caractéristiques du signal et en conséquence sans en privilégier aucune. Autrement dit la spatialisation est incluse dans le jugement de la qualité audio, mais comme n'importe quel facteur comme la perte de trame, le codage, etc. On cherche ainsi à évaluer le ressenti global qu'aurait un auditeur standard sans notion particulière des facteurs précédemment cités.

Par qualité de la spatialisation, nous entendons la qualité de discernement des personnes, par rapport à une scène de référence n'ayant pas subi de dégradation comme du codage ou de la perte de trames. Cela peut se traduire par exemple par une conservation d'une bonne précision des positions des locuteurs suite à un traitement.

1.7.3 La conférence audio spatialisée pour deux contextes d'utilisation privilégiés

Deux contextes d'utilisation ont été choisis dans le cadre de la thèse : la conférence audio en entreprise d'une part, et la réalité virtuelle ou réalité augmentée d'autre part.

1.7.3.1 La conférence audio en entreprise

Le premier contexte retenu est celui de la conférence audio en entreprise. Historiquement, la conférence audio en entreprise (sur le réseau téléphonique commuté - section 1.3.3.1) était traditionnellement génératrice de revenus considérables pour les opérateurs téléphoniques et ce jusqu'à ce jour. On peut citer par exemple les revenus de la conférence audio en 2005 en Europe occidentale de l'ordre de 512 millions de dollars pour 119 millions de minutes consommées ([39],[27]). Néanmoins du fait de l'avènement des réseaux IP, celle-ci semble en perte de vitesse lorsqu'elle est fournie seule, puisque son revenu prévisionnel en 2010 serait de 425 millions de dollars pour 98 millions de minutes ([39]).

En effet, l'extension des réseaux sur IP au sein des entreprises et de leur utilisation pour transporter des données temps-réel ont permis le développement de deux types de conférences intégrant l'audio : les conférences vidéos et les conférences dites intégrées. Ces conférences dites intégrées mélangent à la fois la conférence audio et le partage de documents (web conferencing). Ces solutions rendant plus naturelles les réunions attirent de plus en plus de clients.

En 2005, toujours en Europe occidentale et selon [39], les revenus de la conférence vidéo (respectivement intégrée) étaient de 76 (respectivement 3) millions de dollars et sont estimés à 133 (respectivement 230) millions de dollars en 2010.

Les analyses effectuées dans [39] et [27] s'accordent pour dire que les revenus de la conférence audio augmenteront lorsque l'on enrichira l'audio avec de la vidéo ou du partage de documents. Globalement, en Europe occidentale, la conférence audio en général (seule, avec du partage de documents et/ou de la vidéo) rapporterait entre 715 millions de dollars selon [27] et 788 (425+133+230) millions de dollars selon [39] en 2010.

Comme ces chiffres l'ont montré, la conférence audio possède un bel avenir devant elle. Outre le développement de ces technologies et leur intégration logique au sein des entreprises dans le temps, leur essor est accéléré par le contexte actuel et le souhait des entreprises de réduire les coûts en termes de déplacements et de se donner une meilleure image en investissant sur le développement durable.

Pour notre part, comme annoncé ci-dessus, nous traiterons uniquement de la conférence audio, sachant notamment que son couplage avec un logiciel de partage de document ne change rien à notre problématique finale. Ce type de conférence a pour but la discussion, la négociation en tout genre, la communication de renseignements ou la présentation d'idées ou de rapports. Il privilégie donc l'échange d'informations et il est nécessaire que celui-ci soit le plus possible facilité par la meilleure qualité audio possible.

Il semble évident a priori que le son spatialisé sera un avantage en termes d'intelligibilité, de reconnaissance des locuteurs, de confort d'écoute, etc. par rapport à une conférence audio standard (cf. section 1.7.1). Il conviendra par la suite de montrer que l'intégration du son spatialisé au sein des architectures de conférence permet de garantir une qualité audio et une qualité de spatialisation.

1.7.3.2 Les contextes de réalité virtuelle et de réalité augmentée

Le second contexte d'application est le contexte de la réalité virtuelle et de la réalité augmentée, utilisé notamment avec les jeux en réseau. Il est évident que la conférence audio dans les jeux permet de renforcer l'immersion, le réalisme, l'appartenance à un groupe et aussi une plus grande simplicité comparée à un chat texte. Ce type de conférence s'est développé avec le développement de la voix sur IP. Cependant, peu de jeux à notre connaissance intègrent des possibi-

lités de conférence audio, mis à part Unreal Tournament 2004 (http://www.unrealtournament.com/). Ce jeu permet même l'utilisation du son spatialisé à condition que les cartes sons puissent le gérer.

En fait pour la plupart des jeux pour lesquels la conférence audio serait utile, ce sont le plus souvent les joueurs eux-mêmes qui la créent. Elle se trouve être donc complètement indépendante du jeu avec lequel elle est utilisée. Un logiciel de conférence sur IP est généralement utilisé comme teamspeak (cf. section 1.2.4).

Il est à noter que la console de jeux Xbox (http://www.xbox.com) intègre un chat audio nativement permettant de créer des conférences audio en VoIP. Selon les jeux, il est possible d'effectuer la conférence pendant leur déroulement.

L'apport du son spatialisé se situe au niveau du réalisme et de l'immersion (cf. section 1.7.1). Dans ce cas, la spatialisation est associée à la position d'un locuteur dans le monde virtuel, ce qui est complètement différent du premier contexte d'utilisation. En effet, dans le premier contexte, la scène sonore créée est totalement détachée d'une représentation physique des locuteurs et la spatialisation peut être qualifiée de libre, c'est-à-dire libérée des contraintes de cohérence. Dans le second contexte, le son doit être spatialisé dynamiquement car il doit suivre la position de son personnage. Il est facile d'imaginer l'apport de cette information pour des jeux de recherche d'individus, de sources sonores, etc.

Afin d'augmenter le réalisme, il est ici possible d'ajuster le son par rapport à la distance entre l'auditeur et le locuteur, afin de simuler encore plus la réalité de tous les jours. Cela a par contre des désavantages en termes d'intelligibilité et de netteté des discussions. Il est aussi possible d'inclure les sons locaux issus du jeu (comme des bruits de pas, etc.) et de les spatialiser en tenant aussi compte de la distance.

Ces choix de contexte permettent de nous restreindre à certaines technologies de son spatialisé. En effet dans ces deux cas, l'utilisateur lambda se trouve être seul devant son terminal (PC, téléphone, mobile, etc.), équipé d'un casque ou de deux haut-parleurs, et d'un microphone monophonique. Cela nous réduit ainsi le choix des méthodes de spatialisation au binaural (spatialisation sur casque) et au stéréo-dipôle (spatialisation sur 2 haut-parleurs), qui seront explicités dans la section 1.7.4. Le contenu sera exclusivement de la parole, ce qui nous permettra de faire des choix de codeurs dédiés.

1.7.4 Les méthodes de spatialisation

La restitution sonore spatialisée permet, à un auditeur, de reproduire une source sonore virtuelle positionnée dans l'espace et de la suivre dans ses éventuels déplacements. Cette localisation se fait en 3 dimensions (3D) ou 2 dimensions (2D).

Il existe de nombreuses méthodes de reproduction 3D qui n'utilisent pas les mêmes types de signaux ou le même matériel de restitution. Voici la liste de celles qui nous intéressent :

- La stéréophonie.
- Le binaural
- Le stéréo-dipôle

Après un petit rappel sur la perception d'un champ sonore 3D, ces méthodes vont être présentées, dans les paragraphes suivants.

1.7.4.1 La perception 3D d'un champ sonore

1.7.4.1.1 Localisation d'une source sonore

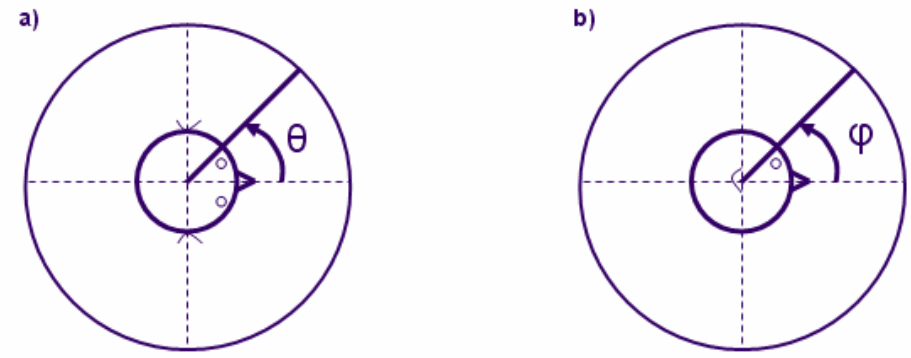

Figure 1.24 Perception 3D d'un champ sonore : a) notion d'azimut, b) notion d'élévation

Les mécanismes, permettant à un auditeur de localiser une source sonore, sont au nombre de trois [68] :

La localisation dans le plan horizontal avec la notion d'azimut θ (cf. Figure 1.24 a)). Celle-ci se base principalement sur les différences, appelées différences interaurales, que perçoit le système auditif entre les signaux captés par les oreilles. Ces différences sont de deux sortes :

- Différences interaurales de temps (Interaural Time Difference ou ITD) ou de phase.
- Différences interaurales d'intensité (Interaural Intensity Difference ou IID) issues du problème de diffraction provoqué par la présence de la tête.

La localisation dans le plan médian avec la notion d'élévation φ (cf. Figure 1.24 b)). Celle-ci se base principalement sur les critères monauraux. Ces derniers caractérisent le fait que l'onde acoustique subit des réflexions sur le pavillon de l'oreille, les épaules et le torse, lors de son trajet. Toutes ces réflexions, dépendant de l'incidence de l'onde, modifient le timbre du son perçu. Il est ainsi possible d'évaluer l'élévation de la source.

La localisation en distance entre la source du son et l'auditeur. Celle-ci se base sur trois indices :

- Le niveau sonore : plus il est fort, plus la source est proche.
- Le rapport entre l'énergie du champ direct et l'énergie du champ réverbéré : un rapport faible implique une source distante.
- Le contenu spectral : l'absence de hautes fréquences donne une impression de distance à la source.

1.7.4.2 La stéréophonie

Un système stéréophonique est composé de deux haut-parleurs disposés afin que ceux-ci et la tête de l'auditeur forment les trois sommets d'un triangle isocèle (cf. Figure 1.25). L'auditeur perçoit en fait une source sonore, dite source fantôme (source virtuelle), qui se situe à une position entre les deux haut-parleurs (HPs). Si les deux signaux dans les HPs sont identiques, alors la source virtuelle semblera être exactement au centre des deux HPs. En faisant évoluer en amplitude ou en phase l'un des signaux (ΔI ou ΔT), cette source virtuelle se déplace entre les haut-parleurs. Cette méthode n'est pas à proprement parler une méthode de restitution spatialisée 3D, car elle permet de localiser une source uniquement entre les deux HPs. Il est à noter que si l'auditeur s'écarte de la ligne de symétrie des HPs, la source virtuelle coïncidera en fait avec le HP le plus proche.

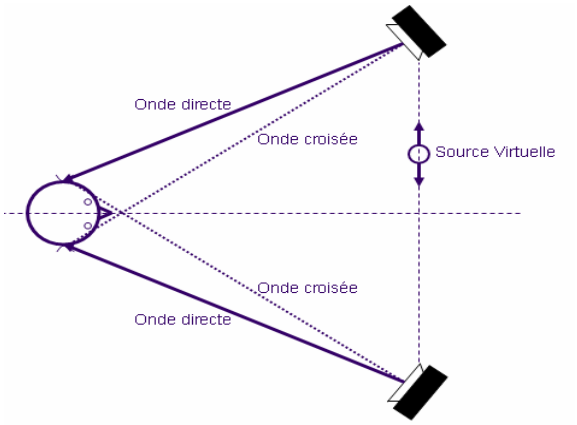

Figure 1.25 Exemple de restitution stéréophonique

Dans le cas d'utilisation de signaux monophoniques, les signaux stéréo sont obtenus de façon synthétique par contrôle de gain ou de retard.

1.7.4.2.1 Avantages de la stéréophonie

- Mise en œuvre simple
- Simplement deux HPs pour la restitution.

1.7.4.2.2 Inconvénients de la stéréophonie

- Zone d'écoute limitée.
- Modèle de spatialisation sonore "grossier" : Performances à la fois imprécises et limitées en termes de spatialisation sonore.
- L'apport en termes d'intelligibilité est moindre par rapport aux techniques binaurales [6,25].

1.7.4.3 Les méthodes binaurales

Le principe des méthodes binaurales est de reproduire le champ sonore induit au niveau des oreilles de l'auditeur. Le champ restitué contient donc les effets de la propagation entre la source et l'auditeur (retard, effet de salle, atténuation…) ainsi que l'ensemble des phénomènes engendrés par le corps de l'auditeur, les réflexions sur le haut de son corps et sur son oreille externe (pavillon). Les techniques de synthèse binaurale restituent donc les indices ou HRTFs (*Head Related Transfer Function*) qu'utilise le système auditif pour interpréter le champ sonore. Ce dernier peut être dans ce cas spatialisé dans les trois dimensions.

1.7.4.3.1 Principe des HRTFs

Les HRTFs sont des fonctions de transfert relatives à la tête ($h_g(t)$ et $h_d(t)$) représentées par des filtres, qui modélisent la propagation acoustique entre la source (r, θ, φ) et les oreilles de l'auditeur (cf. Figure 1.26 [22]). Elles contiennent donc les informations sur les ITD, les IID ainsi que sur les indices monauraux. L'utilisation des indices monauraux les rendent particulières à chaque personne.

Il existe deux possibilités pour une restitution spatialisée du son : la synthèse binaurale et la synthèse stéréo-dipôle.

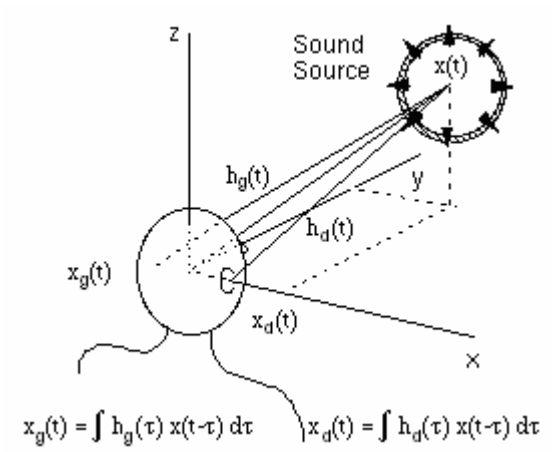

Figure 1.26 Illustration des HRTFs

1.7.4.3.2 La synthèse binaurale

La synthèse binaurale permet de restituer, par l'intermédiaire des HRTFs, un environnement spatialisé, en partant de sources monophoniques, écoutable au casque [61]. Il est possible de placer les sources suivant des directions particulières en azimut θ et en élévation φ. Le système doit évidemment disposer d'un jeu de paires de HRTFs pour chaque direction. Dans l'exemple, Figure 1.27, 3 sources monophoniques sont spatialisées à 3 endroits différents grâce à 3 couples gauche-droite de HRTFs. Il en ressort deux canaux (un pour chaque oreille) B_d et B_g.

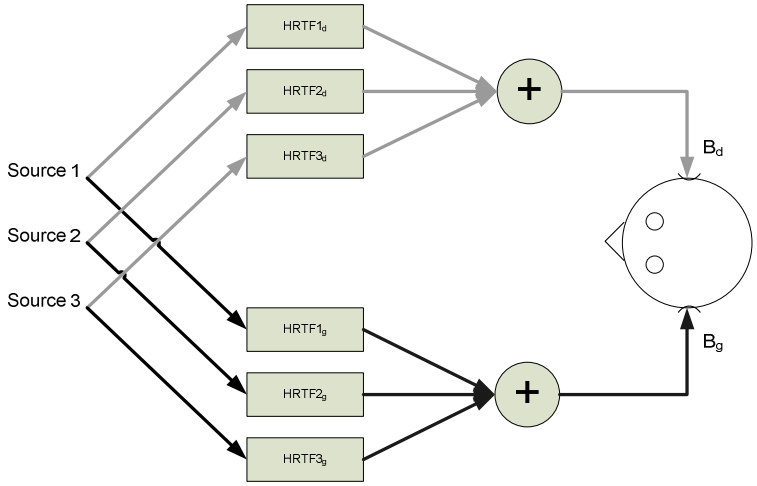

Figure 1.27 Synthèse d'une scène sonore pour une restitution binaurale

1.7.4.3.3 La synthèse stéréo-dipôle

La synthèse stéréo-dipôle permet de générer un contenu spatialisé sur deux haut-parleurs en partant d'un contenu binaural (B_d et B_g). Le problème qui survient est que, sans traitement, le contenu destiné à l'oreille gauche est aussi entendu par l'oreille droite et réciproquement. On souhaite donc que le contenu entendu à l'oreille droite Y_d (respectivement l'oreille gauche Y_g) soit équivalent au contenu du canal droit (respectivement gauche) du binaural. Il est alors nécessaire de générer des signaux X_g et X_d adaptés, correspondant aux signaux issus des HPs gauche et droit.

Le principe est de compenser les trajets croisés afin que le contenu soit adapté à une restitution sur haut-parleurs. La solution proposée dans [31] est illustrée Figure 1.28.

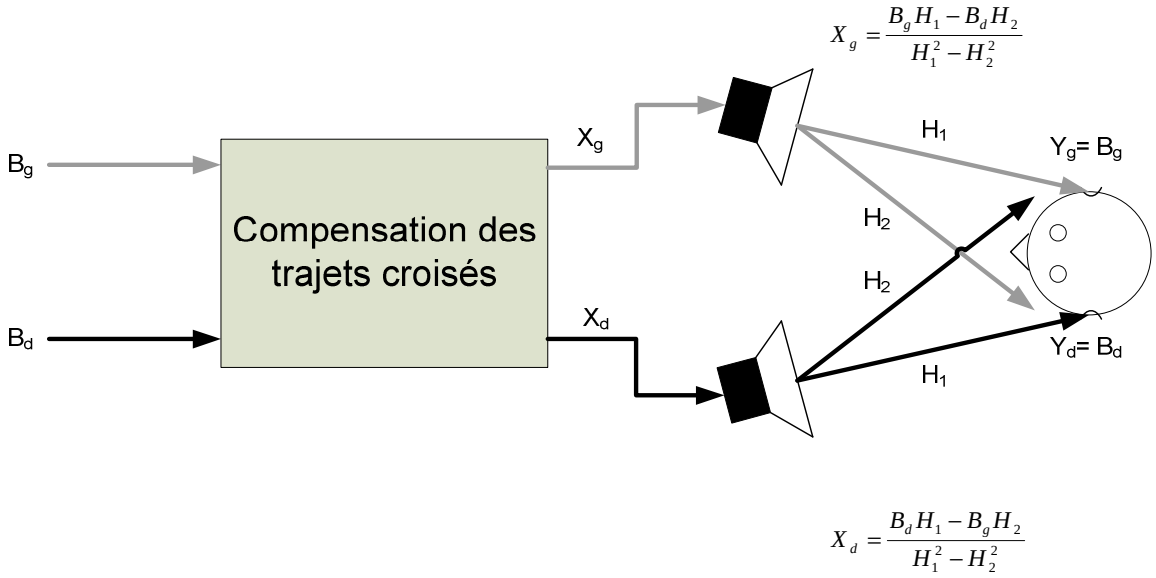

Figure 1.28 Synthèse d'une scène sonore pour une restitution stéréo-dipôle

Afin de compenser les trajets croisés, il convient de résoudre le système suivant [31] :

$$\begin{bmatrix} Y_g \\ Y_d \end{bmatrix} = \begin{bmatrix} B_g \\ B_d \end{bmatrix} \tag{1-1}$$

Cela revient à résoudre :

$$\begin{bmatrix} H_1 & H_2 \\ H_2 & H_1 \end{bmatrix} \begin{bmatrix} X_g \\ X_d \end{bmatrix} = \begin{bmatrix} B_g \\ B_d \end{bmatrix} \tag{1-2}$$

Avec H_1 et H_2 :

- H_1 : la fonction de transfert entre le haut-parleur gauche et l'oreille gauche (qui est aussi égale à la fonction de transfert entre le haut-parleur droit et l'oreille droite, car le système est supposé symétrique). Cela correspond à la fonction de transfert du trajet direct entre le haut-parleur et l'oreille la plus proche.
- H_2 : fonction de transfert entre le HP gauche et l'oreille droite (qui est aussi égale à la fonction de transfert entre le HP droit et l'oreille gauche). Cela correspond à la fonction de transfert du trajet de contournement de la tête.

En inversant le premier élément de l'équation, on obtient :

$$\begin{bmatrix} X_g \\ X_d \end{bmatrix} = \begin{bmatrix} H_1 & H_2 \\ H_2 & H_1 \end{bmatrix}^{-1} \begin{bmatrix} B_g \\ B_d \end{bmatrix} \tag{1-3}$$

Ce qui revient à :

$$\begin{bmatrix} X_g \\ X_d \end{bmatrix} = \frac{1}{H_1^{\,2} - H_2^{\,2}} \begin{bmatrix} H_1 & -H_2 \\ -H_2 & H_1 \end{bmatrix} \begin{bmatrix} B_g \\ B_d \end{bmatrix} \tag{1-4}$$

Finalement, on obtient X_g et X_d les signaux à envoyer aux deux HPs :

$$X_g = \frac{B_g H_1 - B_d H_2}{H_1^{\,2} - H_2^{\,2}} \tag{1-5}$$

$$X_d = \frac{B_d H_1 - B_g H_2}{H_1^{\,2} - H_2^{\,2}} \qquad (1\text{-}6)$$

1.7.4.3.4 Avantages des méthodes binaurales

- Ce sont des outils de reproduction sonore 3D performante puisqu'ils permettent une reproduction spatialisée sur deux canaux seulement, lorsque les HRTFs sont celles de l'auditeur.
- Ils sont simples à implémenter à l'aide de filtres FIR.

1.7.4.3.5 Inconvénients des méthodes binaurales

- Généralement, les HRTFs ne sont pas individualisées, ce qui peut amener un défaut dans le ressenti de la spatialisation avec notamment des inversions avant-arrière [13]. La plupart des sujets utilisant des HRTFs non-individualisées [13] ont l'impression que les sources sont au dessus du plan horizontal. Cela s'expliquerait par le fait que les indices spectraux de localisation en élévation se situent au dessus de 7 kHz, où l'énergie des signaux de parole est faible. Une solution serait donc d'avoir des HRTFs généralisées pour tous ou de simplifier la mesure des HRTFs, actuellement complexe, en sélectionnant des directions particulières avec lesquelles il est possible de déterminer l'ensemble des autres directions. Cela reste encore à l'état de recherche.
- La synthèse binaurale ne permet pas de jouer sur la distance radiale entre une source et l'auditeur car les HRTFs sont mesurées sur une sphère donc à une distance fixe du centre de la tête.
- La synthèse binaurale ne permet pas une excellente externalisation des sons en dehors de la tête mais un ajout de réverbération permet souvent de pallier ce problème.
- Dans le cas de la méthode stéréo-dipôle, les filtres de corrections H_1et H_2 sont calculés pour une position d'écoute. En conséquence, un système stéréo-dipôle ne possède qu'un point d'écoute optimal et uniquement dans le cas où l'auditeur est immobile.

1.7.4.4 Impact sur les terminaux

Du fait des méthodes de spatialisation choisies, les terminaux choisis pour la conférence audio spatialisée auront nécessairement deux haut-parleurs ou un casque stéréo. Ils pourront être des terminaux mobiles connectés par un accès WIFI par exemple, ou des terminaux fixes connectés par des accès filaires.

L'utilisation de terminaux possédant uniquement un haut-parleur ne permet pas de faire de la spatialisation. Cependant ces terminaux sont susceptibles de pouvoir participer à une conférence audio spatialisée, même s'ils ne profitent pas de ces avantages. Il est donc nécessaire de pouvoir leur proposer un contenu mixé monophonique issu par exemple d'un simple mixage de flux monophoniques.

1.8 Axes de recherche

Tout d'abord, le domaine de la conférence audio sur IP n'est certes pas innovant comme le prouve l'état de l'art, mais l'intégration en son sein du son spatialisé l'est. Le premier objectif de la thèse est de déterminer la manière dont il est possible d'intégrer le son spatialisé tout en préservant ses atouts et la qualité audio globale. Il est également primordial de garder une interopérabilité avec le monde monophonique. Le second objectif de la thèse est d'établir les différentes possibilités pour enrichir la conférence audio en VoIP (spatialisée ou standard) en termes de traitements d'amélioration.

Il est ainsi nécessaire de déterminer comment implémenter le son spatialisé au sein des réseaux IP, ce qui revient à savoir comment l'intégrer dans les parties signalisation et média. Il conviendra de définir les différentes solutions possibles en termes de spatialisation et de traitements d'amélioration dans la partie média suivant les architectures que l'on souhaite. De même, il conviendra de définir les besoins protocolaires pour pouvoir gérer la spatialisation.

Une fois les architectures média fixées, il conviendra de tester la qualité audio et la qualité de la spatialisation résultante. Cela signifie qu'il faudra définir des tests de qualité audio adaptés à la spatialisation en prenant soin de pouvoir extraire l'impact de chaque élément entrant en jeu dans les architectures.

2. Définition d'architectures pour un enrichissement de la conférence audio sur IP

Le chapitre 1 a montré la flexibilité que permettent les réseaux IP en termes d'architectures pour la conférence audio. La spatialisation, comprenant une étape de mixage de flux spatialisés par synthèse binaurale ou synthèse stéréo-dipôle (voir section 1.7.4), est une opération similaire à l'opération de mixage monophonique de la conférence audio standard. Notre point de départ sera en conséquence les architectures vues dans la section 1.2.

L'objectif de ce chapitre est de proposer des architectures permettant d'allier la conférence audio en voix sur IP, la spatialisation et les traitements d'améliorations connus. Un choix sera effectué parmi elles afin d'effectuer des tests qualitatifs dans le chapitre 4.

D'un point de vue strictement voix sur IP, nous nous intéresserons dans ce chapitre à la partie média. Les différents cas, qui vont être étudiés, sont basés sur :

- les différents types de conférence.
- les deux types de pont existants (pont mixeur ou pont répliquant).
- les terminaux standards avec au minimum une prise de son monophonique et une restitution sur casque ou deux haut-parleurs et ayant éventuellement des capacités à effectuer eux-mêmes la spatialisation.
- les différentes briques de spatialisation sélectionnées pour la thèse (binaural et stéréo-dipôle) qui ne peuvent se situer qu'à deux endroits : soit sur un pont de conférence IP (section 2.2 pour la conférence centralisée avec pont mixeur), soit sur un terminal (section 2.3 pour les conférences centralisée avec pont répliquant, distribuées multi-unicast et multicast).
- les différents traitements d'amélioration possibles pour la qualité des signaux audio : la détection de l'activité vocale (Voice Activity Detection ou VAD), la commutation de flux, le contrôle automatique de gain (Automatic Gain Control ou AGC) et l'annulation d'écho acoustique (Acoustic Echo Cancellation ou AEC) et de réduction de bruit.

Les sections 2.2 et 2.3 ont pour objectif de présenter les différentes architectures de conférence audio spatialisée que nous proposons, à la vue des techniques disponibles et de leurs contraintes. La section 2.2 considérera les architectures utilisant un pont mixeur spatialiseur. La section 2.3 traitera des autres possibilités pour lesquelles la spatialisation se fait sur un terminal. En préambule, les blocs pouvant jouer sur la qualité suivant leur présence ou leur absence, seront explicités (section 2.1).

Il est à noter que les blocs d'encodage ou de décodage utilisés sont des blocs de compression que l'on suppose adaptés au transport des flux monophoniques ou spatialisés.

Des travaux complémentaires (section 2.4) ont été menés pour créer une architecture innovante utilisant un pont mixte afin de tirer profit des avantages des deux entités de base que sont les ponts mixeur et répliquant. De plus, des travaux effectués en collaboration avec l'INRIA (section 2.5) seront présentés avec pour but de diminuer la bande passante entre un pont répliquant et un terminal en ne transmettant du premier vers le second que les trames audibles et en rejetant les autres, grâce à un critère de masquage auditif.

2.1 Les blocs jouant sur la qualité

La voix sur IP, comme toutes les technologies de transmission de la voix, possède des limites : bande passante limitée, bruit, écho, …. Des solutions existent pour repousser ces limites mais elles ont aussi un coût qu'il faut prendre en compte, notamment en termes de retard, de ressources CPU voire de qualité. Cette section a pour objectif de présenter ces traitements d'amélioration. Il est à noter qu'il n'est pas donné d'ordre de grandeur en termes de retard ou de ressources CPU, car ces caractéristiques dépendent grandement de la plateforme d'implémentation (PC, DSP, …).

2.1.1 La détection d'activité vocale

2.1.1.1 Introduction à la détection d'activité vocale

La détection d'activité vocale (*Voice Activity Detection* – VAD) consiste à segmenter sur un signal audio les zones de parole active et les zones de silence (parole inactive). Pour cela, de nombreux algorithmes existent avec des comportements très différents. On peut cependant résumer le fonctionnement global d'un bloc de VAD en deux étapes [70] :

- Calcul de paramètres tels que l'énergie pleine bande, la table des coefficients de prédiction linéaire LSF (*Line Spectral Frequencies*), …
- Application d'une règle de classification sur les paramètres précédemment cités.

Chaque application étant différente, il faut trouver le bon compromis entre la complexité de l'algorithme, ses performances objectives (en termes de taux de fausses alarmes et de taux de non détection, de gain en bande passante si un système de transmission discontinue est mis en place – voir ci-dessous, etc.) et subjective (qualité perçue suite à l'utilisation de la VAD). On rappelle que la fausse alarme est la détection de parole alors qu'il n'y en a pas, et que la non-détection est le fait de ne pas détecter la parole alors qu'elle est présente.

En outre, selon les algorithmes, on obtient soit un drapeau de présence/absence de parole (booléen), soit une probabilité de présence de parole ou encore d'autres grandeurs discrètes ou continues.

Les algorithmes de VAD prennent généralement en entrée le signal décodé (format linéaire sur seize bits PCM ou Pulse Code Modulation) et renvoient en sortie les informations citées précédemment. Il existe des VAD codec-dépendants. Dans ce cas, la VAD est intégrée dans le processus de codage et utilise des paramètres calculés par l'algorithme de codage lui-même. Leurs sorties sont des informations du même type que pour les VAD de signaux décodés.

Dans le cas de la conférence audio, quelques contraintes doivent être posées. D'une part, il est préférable que le bloc de VAD génère une grandeur continue plutôt qu'un booléen afin de pouvoir comparer les voix entre elles. De plus, la qualité de détection de la VAD est importante dans le cas de la commutation (voir section 2.1.6) si bien qu'il doit présenter en priorité un taux de non-détection proche de 0 afin de ne pas perdre d'information utile. Un taux de fausses alarmes le plus bas possible est également intéressant mais moins critique. Pour des raisons d'homogénéisation des algorithmes, les VAD codec-dépendants ne conviennent pas pour un

pont de conférence sauf si ce dernier est dédié à un type de codec (exemple pont exclusivement G.729). Enfin, la complexité n'est pas le critère le plus décisif, bien qu'elle ait son importance.

2.1.1.2 Introduction au système de transmission discontinue

On rappelle que la transmission discontinue (*Discontinuous Transmission* – DTX) consiste à envoyer ou non des paquets audio suivant la détection ou non de parole dans les flux à transmettre. Il est intéressant d'observer que dans le cas d'un système de DTX, l'information de VAD calculée sur le terminal est intrinsèquement incluse dans la présence ou l'absence d'un paquet, après l'envoi d'une trame de silence signifiant le début de la "non parole". La DTX permet donc une réduction de la bande passante montante.

Cette trame de silence aboutit généralement à diffuser du côté d'un auditeur, soit un silence, soit le bruit de fond capté côté émission, soit un bruit blanc… [88]. Le bruit généré pour simuler la présence du distant est appelé bruit de confort. On parle aussi de *Comfort Noise Generation* en anglais (CNG) pour la génération de ce bruit.

En guise de remarque, effectuer une VAD sur un pont mixeur n'est pas suivi d'une réduction de la bande passante descendante entre un pont et un terminal, sauf si personne ne parle. Cela peut présenter un intérêt sur un pont répliquant si on utilise une VAD codec-dépendante, c'est-à-dire sans décodage incluant un retard supplémentaire correspondant à la durée de l'analyse. On peut dans ce cas diminuer la bande passante descendante d'un pont répliquant vers les terminaux.

Une option possible est l'insertion de ces informations de VAD (de type probabilité) calculées sur les terminaux dans les trames transmises vers le pont ce qui évite à ce dernier de les recalculer pour faire de la sélection de flux (voir section 2.1.5). Cependant, un problème se pose quant à l'homogénéité des différentes VADs faites sur les terminaux.

En guise d'information, la VAD peut être aussi utilisée pour réaliser une commutation de vidéo lors d'une visioconférence. En effet, si la VAD d'une salle détecte de la parole, la vidéo correspondant à celle-ci est affichée.

2.1.1.3 Un exemple de bloc de VAD idéalement adapté à la conférence audio

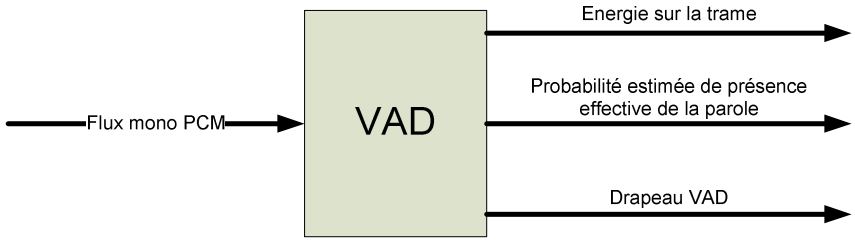

Figure 2.1 Illustration de la VAD

Le bloc illustré Figure 2.1 est un exemple de ce qu'il serait intéressant d'implémenter pour notamment pouvoir profiter de la sélection de flux ou d'un système DTX. Placé indifféremment sur le terminal émetteur ou sur un pont de conférence, il permet de déterminer sur une trame si le flux contient un signal de parole. L'entrée du bloc est le flux d'un conférencier au format décodé PCM. Les sorties du bloc sont :

- Un drapeau VAD : valeur binaire indiquant lorsqu'elle est positionnée à 1 que le signal en entrée contient de la voix. La valeur du drapeau est une fonction des deux autres données.
- L'énergie sur la trame correspond au résultat de la formule suivante, pouvant servir de référence de comparaison avec d'autres trames :

$$E = \sum_{n=0}^{N-1} x_n^2 \text{ où } (x_n) \text{ représentent les N échantillons de la trame.}$$

- La probabilité estimée de présence effective de parole donne un nombre entre 0 et 1, calculé en fonction de différents critères et notamment en fonction du rapport signal à bruit.

2.1.1.4 Avantages et inconvénients de la détection d'activité vocale

L'utilisation d'un bloc VAD permet de réduire la bande passante montante d'un terminal et de contrôler une éventuelle sélection de flux.

L'inconvénient majeur lié à l'utilisation d'une VAD réside dans les erreurs de non-détection de la parole utile dont les effets sont particulièrement audibles. En effet, ces erreurs sont susceptibles de supprimer de l'information (la parole) et donc de dégrader la qualité, en générant un phénomène similaire à la perte de paquets ou des coupures (traduit de l'anglais *clipping*) [34].

2.1.2 Le contrôle automatique de gain ou AGC

2.1.2.1 Principe de fonctionnement

Le rôle du bloc de contrôle automatique de gain (*Automatic Gain Control* – AGC) est d'égaliser, selon un critère énergétique (par exemple l'énergie pleine bande), les différents signaux issus des locuteurs au niveau d'un pont, d'un terminal récepteur ou du signal local au niveau d'un terminal émetteur. Il s'agit d'un bloc central dans l'application de conférence audio, puisque son objectif est d'éviter que des interlocuteurs soient masqués par d'autres du fait de différences d'énergie entre les signaux. En effet, des différences de niveau interviennent fréquemment en raison des disparités dans les équipements et dans les réglages audio (réglage du microphone).

Le bloc d'AGC peut être localisé à différents endroits dans la chaîne de transmission. Situé sur le terminal émetteur, il permet de soulager le serveur de conférence ou le terminal récepteur mais cela implique que tous les terminaux soient équipés du même type d'AGC, ce qui n'est pas forcément garanti. Si tel n'est pas le cas, l'intérêt de l'utilisation du bloc AGC sur seulement quelques terminaux est limité. Situé dans le pont ou dans le terminal récepteur (en conséquence avant le mixage suivant la configuration de la conférence), il permet de garantir une cohérence entre tous les terminaux puisque tous les signaux seront traités par le même type de bloc AGC pour arriver à un niveau d'énergie équivalent.

Un bloc AGC est susceptible d'être couplé avec le bloc de VAD afin de n'égaliser que le niveau de parole. Cela peut par contre produire des contrastes de bruit entre les phases de silence et de parole.

Figure 2.2 Bloc de contrôle automatique de gain

Ce bloc (Figure 2.2) permet de calculer le gain à appliquer sur le signal afin que tous les signaux audio aient une énergie globale équivalente. Le bloc prend en entrée le flux d'un conférencier : format PCM. La sortie du bloc est un gain à appliquer le cas échéant sur le signal.

2.1.2.2 Avantages et inconvénient du système AGC

Comme cela vient d'être décrit, le bloc d'AGC permet d'homogénéiser les niveaux énergétiques des différents intervenants d'une conférence audio.

La principale contrainte du bloc AGC est que la détermination du gain se fait généralement sur des signaux décodés. Une procédure de décodage est donc nécessaire qui augmente naturellement la complexité. Du fait du fonctionnement de l'AGC (temps d'adaptation de

l'algorithme), il faut s'attendre à une gêne le temps que le niveau sonore se stabilise lors de la prise de parole d'un participant. Cependant, aucun test de qualité ne semble avoir été fait pour mesurer cet impact en termes de qualité perçue par un auditeur au cours d'une conversation.

2.1.3 Le bloc de débruitage

2.1.3.1 Principe de fonctionnement

La prise de son capte généralement le locuteur et son bruit environnant. Ce bruit peut être gênant pour les autres interlocuteurs de la conversation, surtout si le rapport signal utile à bruit ambiant est défavorable. Dans ce cas, l'utilisation d'un bloc de débruitage permet de supprimer (ou du moins réduire) toutes les informations inutiles (bruits) contenues dans le signal. Le bloc prend en entrée des trames de signal décodé. Le signal de sortie est également un signal PCM de même fréquence d'échantillonnage que le flux d'entrée.

La technique généralement utilisée repose sur le traitement du signal par un filtre, calculé dans le domaine fréquentiel grâce à un estimateur de la densité spectrale de puissance du bruit lors des phases de silence établies par un bloc de VAD [95]. Il est à noter qu'un bloc de VAD couplé à un système de transmission discontinue peut être considéré comme un bloc de débruitage, puisque le bruit est surtout audible en période de silence lorsqu'il n'est pas masqué par la parole.

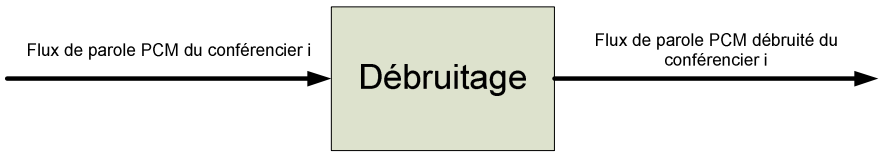

Figure 2.3 Bloc de débruitage

Ce bloc (Figure 2.3) permet de filtrer le signal de parole d'un conférencier et de sortir un flux où le bruit sera atténué. Il sera généralement présent après un bloc de contrôle automatique de gain sur un pont, un terminal émetteur ou récepteur.

2.1.3.2 Avantages et inconvénients d'un bloc de débruitage

Il est à noter que les choix de réglage du bloc de débruitage peuvent altérer la qualité du signal de parole tout en laissant un bruit résiduel [71]. En effet, un débruitage agressif introduit des distorsions. Aussi, l'introduction d'un algorithme de débruitage s'accompagne du choix d'un compromis entre le niveau de réduction de bruit que l'on souhaite atteindre et la distorsion que l'on peut accepter d'introduire sur le signal utile. Il conviendrait de tester l'impact sur la qualité audio d'un débruitage entre un décodage et un encodage au niveau d'un pont afin de voir si la qualité perçue est meilleure pour un auditeur. Le débruitage pourrait introduire de la distorsion sur la parole plus gênante que le bruit de fond qu'il est censé enlever.

2.1.4 L'annulation d'écho acoustique

2.1.4.1 Principe de fonctionnement

Ce bloc a pour but d'éviter le problème présenté dans la section 1.4. On rappelle le schéma général :

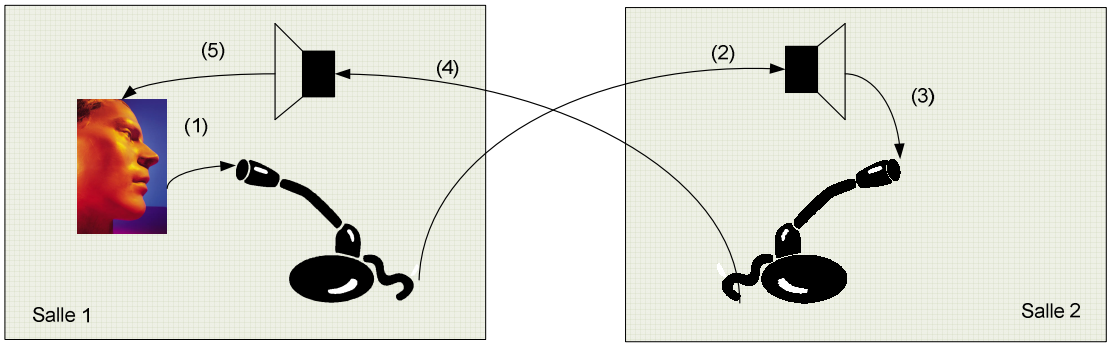

Figure 2.4 Illustration de l'écho acoustique

Dans le cadre de la conférence audio spatialisée avec les hypothèses que nous avons retenues, nous sommes susceptibles de nous trouver dans un cas générateur d'écho acoustique lors de la diffusion sur deux haut-parleurs (HPs) et d'une prise de son monophonique.

Les solutions disponibles sont :

- Un mode half-duplex qui, lorsque la personne de la salle 1 parle, applique un gain (plus précisément une atténuation) sur le microphone de la salle 2 pour éviter que l'écho capté soit entendu dans la salle 1 et réciproquement. Ce mode pénalise l'interactivité de la conversation puisque quand un participant parle, l'autre peut difficilement se faire entendre. Il n'est évidemment pas adapté à la double parole, c'est-à-dire lorsque les deux participants parlent en même temps.
- Les algorithmes d'annulation d'écho reposant sur un filtrage adaptatif visent à soustraire au signal microphonique y(n) une estimée de l'écho $\hat{z}(n)$, obtenue à partir de la référence haut-parleur x(n), comme illustré Figure 2.5. Ainsi le filtre adaptatif $\hat{H}_L$ modélise le chemin acoustique réel H entre le haut-parleur et le microphone (en réalité, H ne modélise que la partie linéaire du canal acoustique et ignore toutes les non-linéarités) et cherche à estimer l'écho z(n). Il existe de nombreux algorithmes permettant d'adapter les coefficients du filtre dont les plus connus [36] sont le *Normalized Least Mean Squares* (NLMS), *Affine Projection Algorithms* (APA), *Recursive Least Squares* (RLS). Le NLMS est le plus répandu pour ses propriétés de complexité et de stabilité numérique. Dans le cas multi-haut-parleur et mono-microphone, il faut estimer les contributions liées au couplage entre chaque haut-parleur et le microphone, ce qui nécessite l'utilisation (et l'adaptation) de 2 filtres.

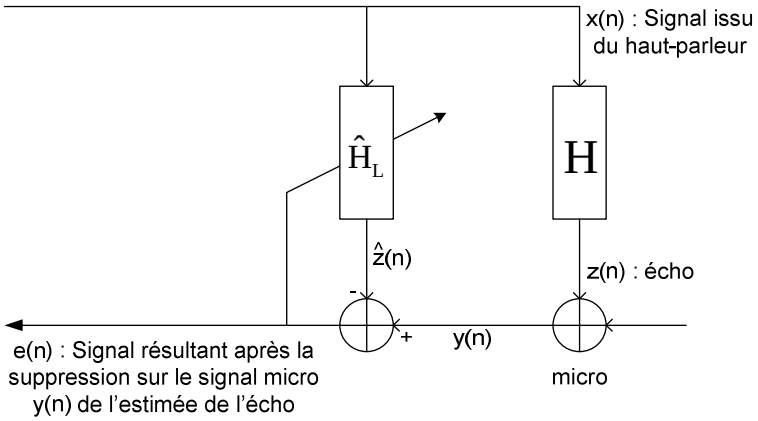

Figure 2.5 Illustration d'un algorithme d'annulation d'écho reposant sur un filtrage adaptatif

- Il existe des solutions combinées d'annulation d'écho et de réduction de bruit [57] reposant sur du filtrage de Wiener, notamment dans le domaine fréquentiel. Le filtre de Wiener consiste à estimer le signal utile que l'on souhaite obtenir comme une fonction linéaire de l'observation, puis à déterminer cette fonction linéaire optimale (au sens des moindres car-

rés). Ces techniques souffrent de difficultés de réglages qui sont souvent à l'origine de distorsions audibles.

Toutes ces méthodes permettent d'obtenir une réduction d'écho efficace en période de simple parole, mais leurs performances se dégradent généralement en période de double parole.

Ces méthodes ont été généralement déployées sur des signaux en bande étroite (narrow-band), mais on peut les étendre aux signaux large bande (wide-band, super wide-band…), en réalisant par exemple des traitements en sous bande. Ainsi, pour traiter un écho en mode wide-band (50 - 7000 Hz), on pourra appliquer un filtrage adaptatif sur la bande basse (jusqu'à 3400 Hz) et une simple atténuation (application d'un gain) dans la bande haute. On applique rarement un traitement d'annulation d'écho sur une pleine large bande (50 - 7000 Hz) afin d'éviter l'adaptation de filtres longs, souvent délicats à converger.

2.1.4.2 Avantages et Inconvénients de l'annulation d'écho acoustique

En termes de qualité [71], les traitements d'annulation d'écho acoustique peuvent laisser un écho résiduel audible, générer de la distorsion sur la parole, mais aussi introduire du délai ainsi que des coupures audio. Un effet pervers du délai de transmission est la diminution de l'interactivité, ceci engendrant l'augmentation de la parole simultanée [71].

2.1.5 La sélection de flux

2.1.5.1 Principe de base de la sélection de flux

La sélection de flux consiste simplement, comme son nom l'indique, à sélectionner M flux audio (encodés ou non) parmi N avec $M \leq N$ [21]. Cette sélection de flux située généralement sur un pont est généralement basée dans le cadre de la conférence audio standard sur la présence ou l'absence de parole donc sur des informations de VAD ([72,86]).

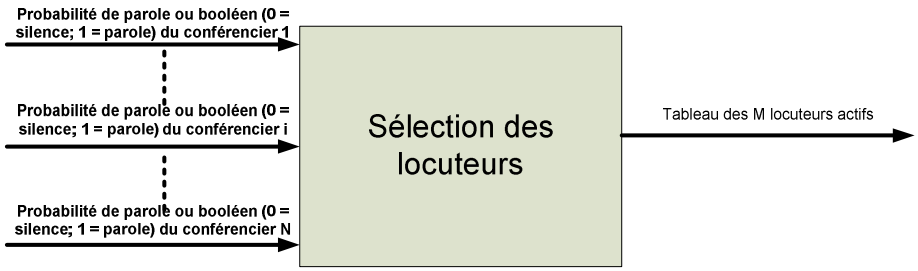

Figure 2.6 Sélection des locuteurs

Comme l'illustre la Figure 2.6 suivant le type de VAD, on peut ainsi sélectionner :

- soit les M participants actifs à un moment donné pour une VAD tout ou rien (1 parole et 0 silence ou autre), avec M non constant.
- soit les M participants parmi les P actifs avec $M \leq P \leq N$ pour une VAD renvoyant une probabilité de parole dans la trame, avec M constant. En guise de remarque, il est inutile de sélectionner M participants si P<M. On en sélectionne à ce moment seulement P.

Ensuite ce tableau des M locuteurs actifs permet de commander le bloc M parmi N décrit Figure 2.7 et de sélectionner les M flux de parole (codés ou non) :

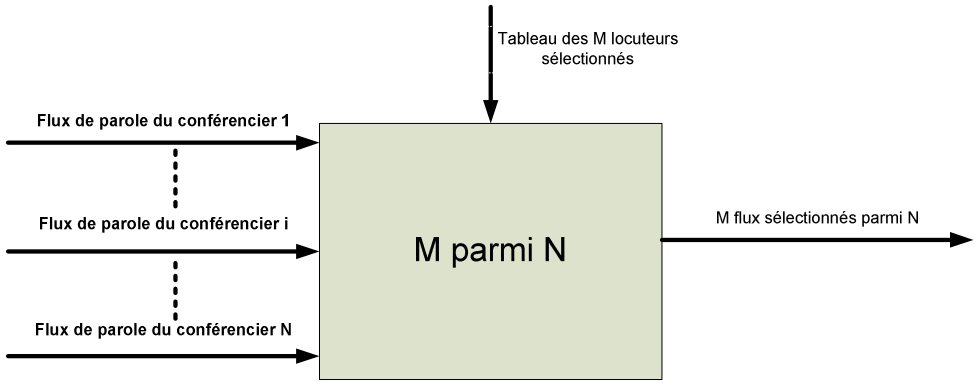

Figure 2.7 Bloc M parmi N

Il convient de bien différencier la sélection de flux de la commutation de flux (voir section 2.1.6), cette dernière pouvant utiliser la sélection de flux pour remplir sa fonction. On effectue ici seulement une sélection à partir d'informations de VAD, que celle-ci ait été calculée sur le terminal en émission ou sur un pont après un décodage total ou partiel.

2.1.5.2 Exemple d'implémentation et résultats en termes de qualité

Dans le contexte de [88], un pont fonctionnant en mode répliquant (sans transcodage) est utilisé. Des informations de VAD (absence ou présence de parole) et d'énergie (issue du calcul de la VAD) sont incluses dans les trames de chaque terminal, avant d'être transmises vers le pont. La sélection de flux est basée sur ces paramètres et l'algorithme TFSS (Tandem-Free Speaker Selection). Cet algorithme se base sur une machine à 6 états qui sert à détecter, classifier le début, le milieu et la fin du dialogue :

- Idle : le participant est silencieux.
- Entry : le participant a commencé à parler après une période de silence.
- Short Hangover : après avoir commencé à parler, le participant s'est interrompu.
- Bridged : le participant a parlé pendant au moins 0.86 seconde.
- Long Hangover : le participant a fait une pause pendant qu'il parlait ou il a été interrompu.
- Short Entry : le participant s'est remis à parler après une pause inférieure à 1.56 secondes.

Les transitions entre deux états dépendent des différents paramètres de VAD et d'énergie pour chaque terminal. Pour les transitions entre états, se référer directement aux pages 68 et 69 du document [88]. La sélection de flux se fait grâce à la combinaison de l'énergie et de l'ordre d'activité.

Cette sélection de flux basée sur l'énergie et la VAD de chaque trame issue de chaque participant a obtenu de bons résultats en termes de qualité et de réduction de bande passante dans [88], en sélectionnant M égal à deux quel que soit le nombre de participants à la conférence. L'auteur de [88] justifie le choix de M égal à 2 en citant l'article [91]. Ce dernier spécifie que toute conférence à N participants peut être ramenée à une conférence à 2 participants avec un pourcentage de parole simultanée entre 5 et 11% du temps total. L'auteur de [88] précise cependant que dans ses tests de conférence à 4 participants, il a détecté une période de double parole d'environ 40% du temps total. Par contre, la période de triple ou quadruple parole a été estimée à moins de 10% du temps total rendant tout de même la valeur M égal à deux la plus pertinente.

L'inconvénient majeur est que les terminaux doivent posséder la même VAD pour rendre équitable, en termes de sélection de flux, le traitement. Cette méthode requiert des terminaux plus complexes et augmente de même le débit du flux montant de ceux-ci vers le pont.

2.1.5.3 Remarques

Nous proposons une solution d'implémentation basée sur le masquage auditif dans la section 2.5. Comme nous avons pu le voir, l'algorithme TFSS de [88] s'applique surtout dans un contexte de conférence audio en entreprise où il est rare que plus de deux locuteurs s'expriment en même temps. Notre solution, ne nécessitant pas que l'on spécifie une valeur de M contrairement à la solution de [88], s'applique évidemment dans ce contexte mais aussi dans un contexte de jeu virtuel où plus de 2 participants sont susceptibles de s'exprimer en même temps.

Pour un pont mixeur (conférence centralisée) ou un terminal (conférence distribuée multi-unicast), ce bloc peut permettre d'éviter certains décodages et mixages inutiles, sauf si les décodeurs ont besoin d'informations sur la trame précédente.

2.1.6 La commutation de flux

2.1.6.1 Le pont mixeur à commutation de flux

On rappelle tout d'abord, Figure 2.8, le cœur d'un pont mixeur gérant des flux monophoniques. Pour chaque conférencier, tous les signaux, préalablement décodés, des autres participants sont additionnés les uns aux autres. Enfin, on encode chaque flux mixé pour le transmettre aux récepteurs.

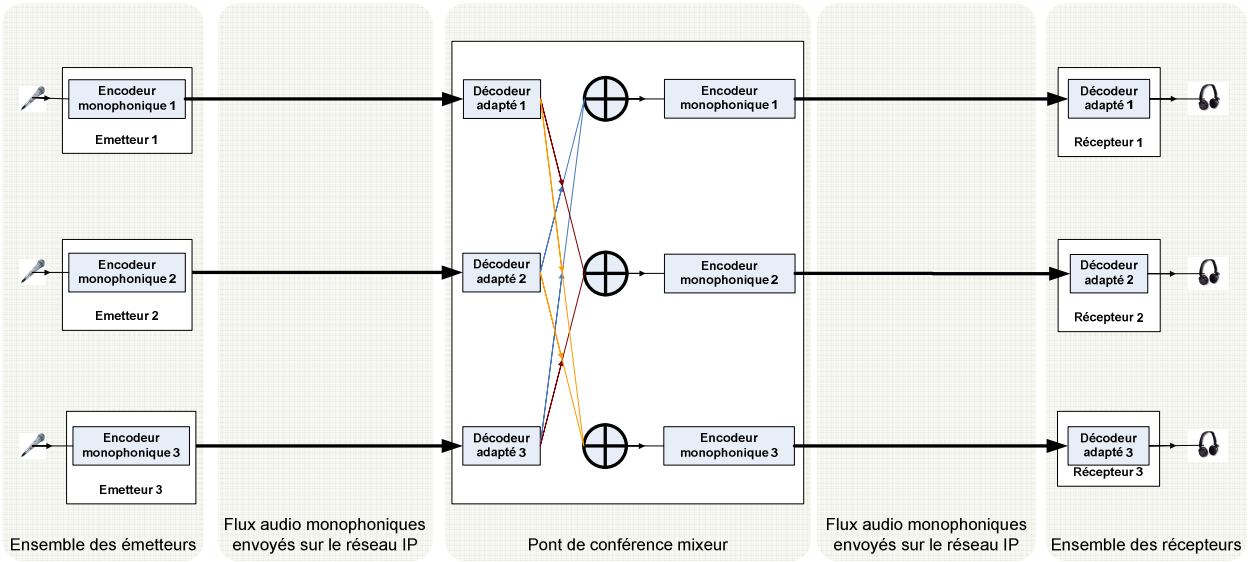

Figure 2.8 Schéma d'un pont mixeur standard

Comme on l'a vu dans le chapitre précédent, une dégradation de qualité apparaît suite à l'étape de décodage puis encodage au niveau du pont. L'utilisation de la commutation de flux va permettre d'éviter une perte de qualité en présence d'un seul locuteur. Cette méthode introduite dans [66] consiste à ne laisser que le flux de la seule personne qui parle (si elle existe) ou bien le mixage des voix des différents locuteurs simultanés.

En revanche, étant donné que certains traitements ne fonctionnent que sur du signal décodé, il faut prévoir, en plus du détecteur d'activité vocale, des signaux de contrôle de la commutation.

Ainsi, le schéma de la Figure 2.9 est la base du pont mixeur à commutation de flux monophoniques. Chaque bloc sera explicité dans la partie suivante.

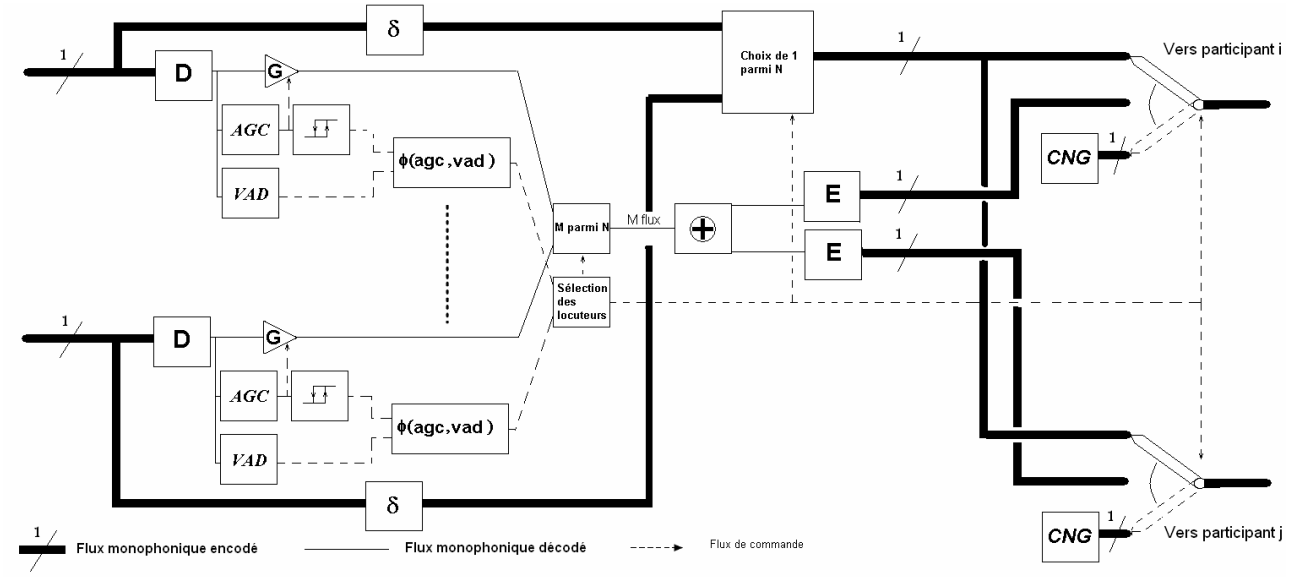

Figure 2.9 Schéma d'un pont mixeur à commutation de flux

2.1.6.2 Les différents composants

Tout d'abord, le signal de chaque participant est décodé (bloc "D") afin de pouvoir être traité. Un bloc de contrôle automatique de gain (bloc "AGC" explicité dans la section 2.1.2) calcule ensuite l'éventuel coefficient correcteur de niveau sonore G à apporter au signal.

En parallèle, une VAD est calculée sur chaque signal décodé et la combinaison de tous ces résultats (un par conférencier) permet de savoir combien de personnes sont actives sur une durée de trame.

2.1.6.2.1 Premier cas : cas du décodage

Si au moins deux personnes sont déclarées actives, on corrige les signaux avec leurs valeurs de gain de l'AGC (bloc "G"), et le mixage va être effectué. Optionnellement, une sélection des M locuteurs les "plus actifs" pourra être effectuée selon la valeur donnée par la VAD pour chaque locuteur actif (bloc "M parmi N").

S'il n'y a qu'un seul locuteur actif dont le niveau est trop faible ou trop fort entrainant une valeur d'AGC hors d'une fourchette (bloc "hystérésis"), on doit corriger le signal de ce locuteur. On n'envoie donc pas le flux non décodé mais celui qui a été égalisé après décodage. Cela montre la nécessité du bloc Φ(agc,vad), fonction de la VAD et de l'AGC, afin de piloter le bloc "sélection des locuteurs".

Le mixage se fait sur les M locuteurs (M étant strictement supérieur à 1 et pouvant être égal à N). On effectue ainsi M-2 additions pour chacun des M locuteurs et M-1 additions pour chacun des N-M auditeurs. Ces signaux mixés sont transmis à un encodeur avant d'être envoyés au participant correspondant.

2.1.6.2.2 Second cas : cas du flux commuté

Si au plus un locuteur est actif et si sa valeur d'AGC est dans la fourchette autorisée, on transmet sa trame ou la trame du dernier locuteur actif, retardée d'un temps δ pour compenser le retard pris par la ligne de mixage, à savoir le décodage, le traitement et l'encodage. La sélection de locuteur choisit le flux du locuteur unique et le transmet à tous les autres.

Enfin, la dernière étape consiste à choisir, via un commutateur piloté par la sélection des locuteurs, la source pour le destinataire. Le flux entrant retardé du locuteur actif est transmis directement à tous les autres participants. Ce locuteur reçoit lui soit un bruit de confort (CNG), soit rien (non représenté sur la Figure 2.9).

2.1.6.3 Avantages et inconvénients de la commutation de flux

L'avantage de cette méthode est la préservation d'une bonne qualité en présence d'un seul locuteur. Néanmoins, des tests sont à mener quant à la variation de la qualité perçue par un auditeur selon qu'il y ait un ou plusieurs locuteurs. Nous verrons dans les tests de qualité du chapitre 4 que peu de codeurs conversationnels peuvent transporter un contenu multi-locuteur simultanés avec une qualité jugée bonne.

La principale contrainte de ce genre de méthode est qu'il faut que tous les terminaux utilisent le même codec. En pratique, le pont peut décider de rejeter un participant qui ne possède pas le codec requis pour effectuer les traitements ci-dessus.

Par ailleurs, le retard sur la ligne de commutation est codec dépendant puisqu'il dépend du décodage de la trame codée ainsi que du traitement de celle-ci. En guise de remarque, le cas échéant, il peut être utile que le mixage et l'encodage soient réalisés afin de ne pas perdre la continuité dans les variables d'état de l'encodeur de sortie.

2.2 La spatialisation sur un pont de conférence IP

Dans cette partie, nous chercherons à définir les différentes configurations possibles pour fournir un service de conférence audio spatialisée de haute qualité. Nous nous intéresserons aux configurations ayant une entité centrale dans le réseau pour gérer les flux médias telle qu'un pont mixeur ou un pont répliquant. A partir de ces configurations, nous chercherons à établir les différentes contraintes qu'elles peuvent générer notamment sur les terminaux.

Seules les configurations jugées les plus complètes seront présentées. Il conviendra de laisser ou non certains des blocs de qualité vus dans la section précédente suivant le niveau de qualité, de complexité et de latence que l'on désire.

2.2.1 Pont spatialiseur mixeur sans commutation

Le pont spatialiseur mixeur sans commutation, que nous proposons, est dans sa conception un pont relativement standard, remplaçant un mixage de flux monophoniques par une spatialisation de ces mêmes flux.

2.2.1.1 Schéma général du pont

Tout d'abord, Figure 2.10, le signal issu de chaque participant est décodé (bloc "D") afin de pouvoir être traité par les blocs suivants. Un bloc de contrôle automatique de gain (bloc "AGC") calcule le coefficient correcteur à apporter au signal pour le niveler en énergie (bloc "G") par rapport aux autres. Le signal est ensuite débruité par le bloc du même nom. En guise de remarque, faire l'enchaînement inverse (débruitage puis égalisation en niveau) aurait pu faire réapparaître en sortie du bloc AGC le bruit que l'on avait atténué en sortie du débruiteur.

En parallèle, une VAD est calculée sur chaque signal décodé, afin de savoir au final combien de personnes sont actives en même temps. Il est à noter que cette VAD pourrait contrôler l'égalisation en niveau effectuée par le bloc "G", afin que seules les séquences contenant de la parole soient traitées.

Une sélection des M locuteurs les "plus actifs" est effectuée selon la valeur donnée par la VAD (valeur binaire, énergie, …) pour chaque locuteur actif (blocs "Sélection des locuteurs" et "M parmi N"). On peut noter que, pour économiser du temps processeur, un déplacement des blocs AGC et du débruitage après la sélection de flux est possible.

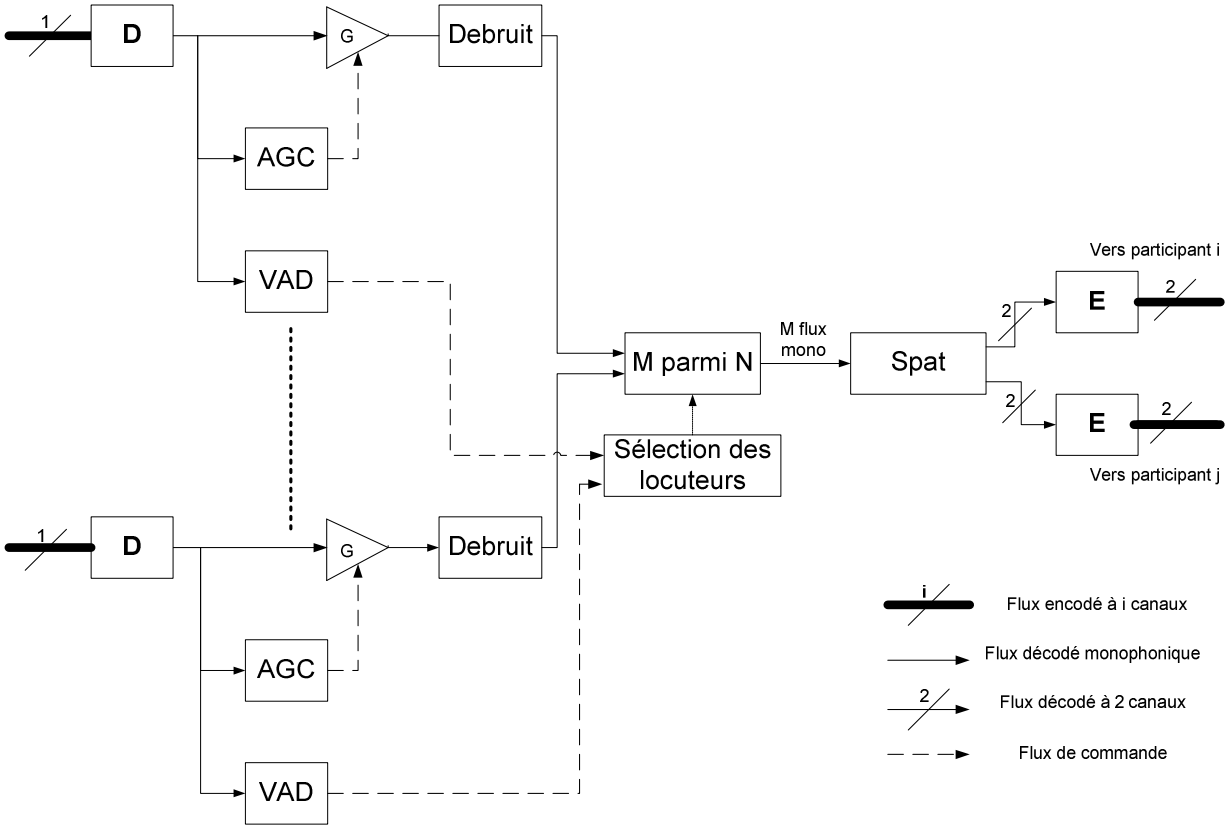

Figure 2.10 Schéma du pont spatialiseur mixeur sans commutation

La spatialisation se fait sur les M locuteurs (M pouvant être égal à N). On effectue ainsi :

- M-1 spatialisations et M-2 additions pour chacun des M locuteurs,
- Ainsi que M spatialisations et M-1 additions pour chacun des N-M auditeurs.

Ces N signaux spatialisés sur 2 canaux sont ensuite transmis à N encodeurs avant d'être envoyés aux conférenciers.

Les blocs AGC, VAD et débruitage pourraient être situés sur un terminal émetteur et sont ceux explicités dans la section 2.1. Dans le cas du bloc de VAD sur chaque terminal émetteur, ces informations devraient être transmises avec les trames audio pour être utilisées dans la sélection de flux. Les modules de sélection de flux et M parmi N sont ceux explicités en 2.1.

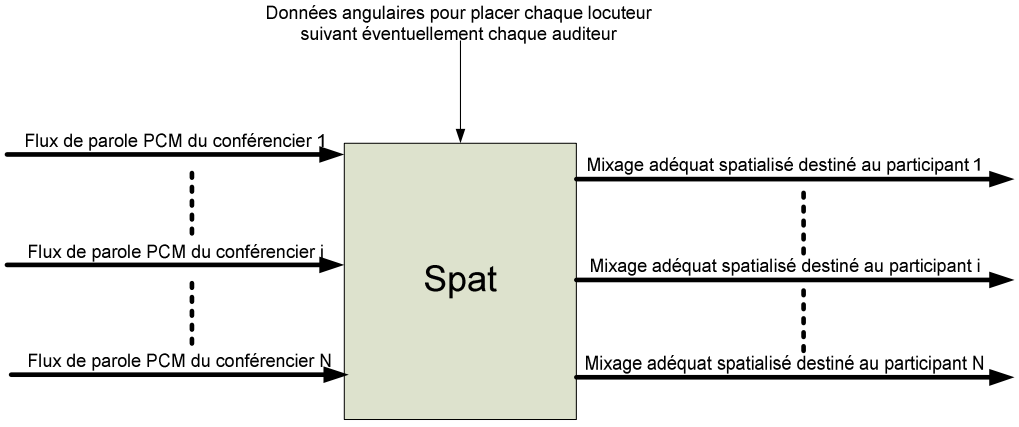

Figure 2.11 Schéma du bloc de spatialisation

Le bloc, illustré Figure 2.11, permet la spatialisation des M (pouvant être égal à N) flux audio monophoniques PCM pour N participants. En sortie de ce bloc, on obtient donc N flux audio spatialisés sur deux canaux.

2.2.1.2 Les terminaux

Avec cette configuration de pont, notre terminal de réception est très simple car il ne possède aucune fonction de haut niveau telle que la VAD, AGC… Le schéma du terminal est présenté en Figure 2.12.

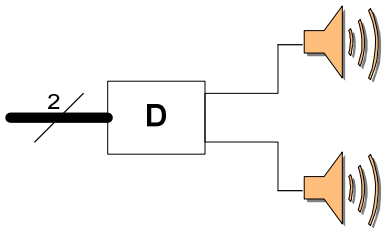

Figure 2.12 Schéma d'un terminal utilisé avec un pont spatialiseur mixeur sans commutation

Le signal sur deux canaux est décodé et émis vers les organes de restitutions : haut-parleurs (cas de l'exemple) ou casque.

2.2.1.3 Contraintes et problèmes anticipés

- Cette configuration n'évite pas le transcodage et donc ni la perte de qualité ni le retard associés.
- Le pont spatialiseur mixeur est susceptible de devoir supporter une charge CPU importante suivant le nombre de conférences et de participants à chacune de ces conférences.
- Il convient de sélectionner des codeurs adaptés aux transports de flux spatialisés.

2.2.1.4 Avantages

- Aucune difficulté particulière pour mettre en œuvre ce type de pont.
- Les terminaux sont très simples. Ils ont comme seule contrainte de pouvoir permettre la stéréophonie ce qui n'est pas un problème pour un ordinateur.
- La sélection de flux permet la réduction d'une partie de la charge CPU.
- Il est à noter que suite à une spatialisation binaurale sur le pont, il est possible de convertir sur le terminal les flux binauraux en flux stéréo-dipôle par la formule vue dans la section 1.7.4.3. Cela peut être utile pour éviter au pont ce traitement et pour simplifier le dialogue entre celui-ci et un terminal souhaitant changer de mode de restitution.
- Il est à noter que cette configuration est celle qui s'adapte aux terminaux. En effet, il est aussi possible d'envoyer un contenu monophonique si le terminal le souhaite en remplaçant le bloc de spatialisation par un simple mixage. L'interopérabilité avec le RTC est aussi possible. En effet, communément, un appel issu du réseau RTC transite par une passerelle RTC/IP pour aboutir au pont de conférence mixeur. Celui-ci ne voit en fait que la passerelle comme destinataire et la considère comme un terminal standard. Il peut en retour lui envoyer un contenu mixé monophonique adapté au moyen de restitution caché derrière cette passerelle.

2.2.2 Pont spatialiseur mixeur avec commutation

Nous proposons une structure avec pont spatialiseur mixeur et commutation Figure 2.13. Cette entité est un pont plus évolué que le précédent, et est censé améliorer la parole en présence d'un seul locuteur.

2.2.2.1 Schéma général du pont

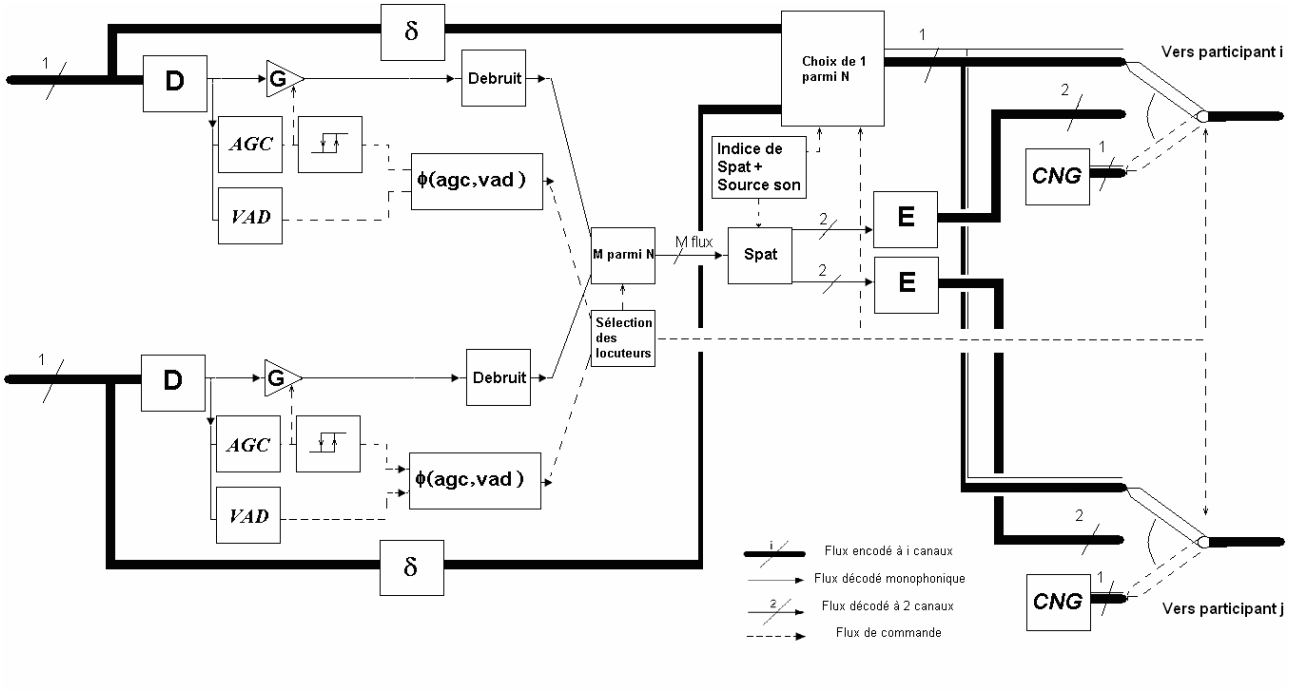

Figure 2.13 Schéma du pont spatialiseur mixeur avec commutation

Voici les différences par rapport au cas de la section 2.1.6 :

2.2.2.1.1 Premier cas : cas du décodage

La partie mixage vue en 2.1.6 est remplacée par un bloc de spatialisation. Cette dernière se fait sur les M (respectivement M-1) locuteurs pour chaque auditeur (respectivement locuteur). Ces signaux sur deux canaux sont transmis à un encodeur pour chaque conférencier.

2.2.2.1.2 Second cas : cas du flux commuté

Le flux émis n'étant pas spatialisé, il faut que la spatialisation soit réalisée par le terminal récepteur. Il convient donc de lui fournir l'identité du terminal dont le flux est commuté. Ainsi deux informations sont issues du bloc de choix : le flux encodé monophonique et l'identité du terminal associé. Il est à noter que l'information d'identité de l'émetteur du flux est contenue dans l'entête RTP du paquet IP et ne nécessite donc pas de flux complémentaire.

2.2.2.1.3 Les blocs de qualité

Les blocs AGC et débruitage sont ceux explicités dans la section 2.1. Le bloc de sélection des locuteurs est celui explicité en 2.1, avec en entrée par exemple les probabilités de parole continues de chaque participant. Le bloc de spatialisation est le même que celui de la section 2.2.1.1

Un filtre par hystérésis permettant le seuillage de la valeur d'AGC, couplé avec un bloc de VAD, permet de commander la commutation à travers une fonction Φ (VAD, AGC).

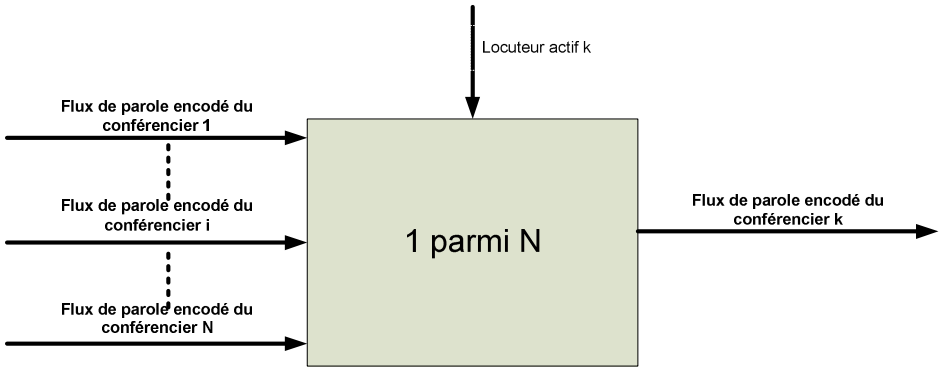

Figure 2.14 Bloc 1 parmi N

Ce bloc, illustré Figure 2.14, permet de choisir le flux encodé qui sera transmis à tous les autres conférenciers en by-pass. Ce flux a déjà été retardé par le bloc de retard δ qui permet de compenser l'avance du flux en by-pass par rapport au flux mixé (temps de décodage additionné avec le temps du traitement et le temps d'encodage).

Au final, un commutateur sélectionne, soit le flux encodé en by-pass pour tous les conférenciers sauf pour le $k^{ème}$ conférencier qui aura un bruit de confort généré par le bloc CNG, soit les flux encodés mixés.

2.2.2.2 Les terminaux

Voici Figure 2.15, ci-dessous, l'architecture de notre terminal utilisable dans ce type de conférence :

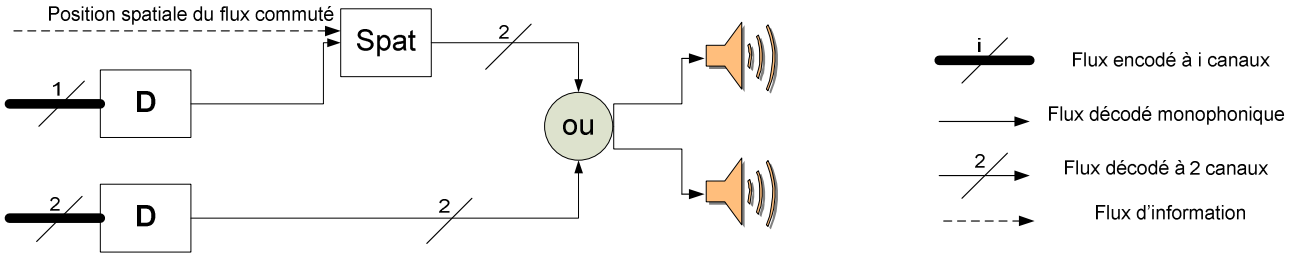

Figure 2.15 Schéma d'un terminal utilisé avec un pont spatialiseur mixeur avec commutation

Un problème important à résoudre avec ce pont est qu'il envoie à chaque participant tantôt un flux monophonique (cas de la commutation), tantôt un flux spatialisé. Cela ne pose normalement pas de problème dans le sens où SIP permet aux données codées différemment de transiter par le même canal de transmission si cela est spécifié dans la négociation. Cependant, les terminaux ont tendance à négocier généralement qu'un seul codec pour un canal. Dans ce cas, cela complique évidemment la gestion des flux et implique dans le cas du protocole de signalisation SIP une négociation de 2 canaux de transmission : un pour le flux mono et l'autre pour le flux stéréo. Le terminal est soit dans un mode soit dans l'autre suivant le port de réception sur lequel le paquet audio est reçu.

Etant donné que le flux du locuteur unique n'est pas spatialisé sur le pont (il est en by-pass), il faut faire la spatialisation sur le terminal. Pour cela, un dialogue d'une nature à définir doit être instauré entre le pont et le terminal pour que ce dernier sache où le spatialiser suivant sa provenance. Il faut cependant tenir compte de la difficulté de synchroniser au cours du temps, les éventuels changements de position des locuteurs dans l'espace de restitution et gérer la synchronisation des basculements entre les flux mixés spatialisés sur pont et ceux non mixés spatialisés sur le terminal. La façon dont doit être restitué le bruit de confort (diffus, localisé, …) est à étudier.

Dans le cas ou le flux audio est déjà spatialisé, on peut le jouer directement.

2.2.2.3 Contraintes et Problèmes anticipés

- Ce pont nécessite que le terminal possède des fonctionnalités de haut niveau avec un bloc de spatialisation. De plus, ce terminal devrait posséder des blocs de contrôle automatique de gain et/ou débruitage, car ces opérations n'auront pas été faites dans le pont pour le signal commuté.
- Il convient de savoir qui parle au niveau d'un terminal récepteur pour bien effectuer la spatialisation en cohérence avec la scène sonore au niveau du pont spatialiseur mixeur. Il est donc nécessaire de développer ou modifier un protocole de communication afin de permettre le transfert de cette information.
- Le traitement sur pont est susceptible de mobiliser beaucoup de ressources (CPU).
- Les transcodages seront certes diminués mais pas supprimés. Il faut noter que des variations de qualité peuvent aussi perturber des auditeurs.
- Il est nécessaire que tous les participants puissent utiliser le même codeur pour le flux commuté sinon il faut effectuer un transcodage et la commutation est impossible.
- Cette solution paraît difficilement développable ...

2.2.2.4 Avantages

- Ce pont possède l'avantage d'utiliser la spatialisation et la commutation de flux sur une même architecture. Ainsi, le problème dû au transcodage est atténué.
- Gestion dynamique des résultats de la VAD.
- L'apport en termes de qualité est à évaluer.
- Réduction de la charge CPU grâce à la sélection de flux au niveau du pont.
- Il serait ici possible de déporter à nouveau la partie stéréo-dipôle de la spatialisation sur un terminal (voir section 2.2.1.4).
- L'interopérabilité avec les terminaux monophoniques est aussi possible (RTC par exemple).

2.2.3 Conclusion

Cette section a montré les différentes solutions pour effectuer au niveau média de la conférence audio spatialisée avec un pont spatialiseur mixeur. La première solution sans commutation est sans contestation la plus simple à développer et à tester. Elle ne nécessite au niveau média que des terminaux capables de jouer des flux sur deux canaux.

2.3 La spatialisation sur un terminal

Comme cela a été vu, seule la conférence audio centralisée avec un pont spatialiseur mixeur possède un bloc de spatialisation sur un pont. Pour les 3 autres types de conférence (les conférences centralisée avec pont répliquant, distribuées multi-unicast et multicast), la spatialisation a lieu sur le terminal récepteur. Nous chercherons dans cette partie à établir l'ensemble des choix possibles et les contraintes associées à la vue des blocs de qualité disponibles établis dans la section 2.1.

2.3.1 La configuration centralisée avec un pont répliquant

2.3.1.1 Version avec sélection de flux sur le pont sans décodage

2.3.1.1.1 Schéma général du pont répliquant sans décodage

Le pont proposé ici est un pont répliquant avec les caractéristiques décrites dans la section 1.2.1.4. Son principe est de diffuser les paquets audio reçus à tous sauf à l'émetteur. Un pont répliquant standard est donc considéré comme un pont non intelligent dans le sens où il ne fait aucun traitement. Ce cas ne sera pas traité car étant un cas particulier du cas décrit ici.

La première variante proposée à ce pont de base est illustrée Figure 2.16. Dans ce module, les flux audio arrivent codés et on extrait sans décoder des paquets RTP des informations utiles calculées sur les terminaux émetteurs telles que la VAD par exemple. Ces informations alimentent un bloc de sélection des locuteurs qui va permettre de déterminer les paquets contenant a priori de la parole. Ces informations vont permettre de commander le bloc M parmi N qui va ou non laisser passer les flux codés pour être envoyés vers les autres participants. Une solution de sélection de flux plus élaborée est proposée dans la section 2.5 se basant globalement sur cette architecture avec extraction de paramètres dans les paquets provenant des participants et analyse de ceux-ci au niveau du pont répliquant.

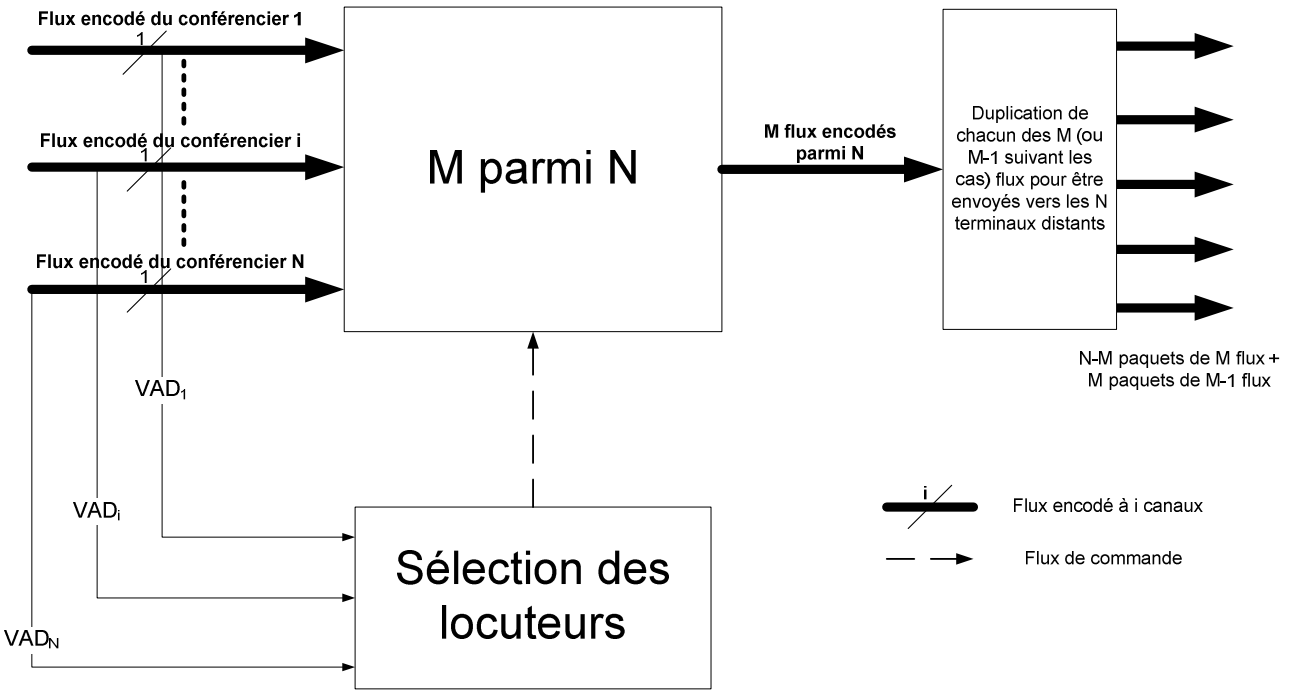

Figure 2.16 Schéma d'un pont répliquant intelligent

Sans décodage, on ne peut ici effectuer du contrôle automatique de gain, de la détection vocale, de la spatialisation ou du débruitage. Il est à noter que la VAD sans décodage est possible mais cela nécessite une VAD spécifique à chaque codeur.

Les blocs M parmi N et Sélection des locuteurs sont évidemment potentiellement supprimables et déplaçables sur le terminal en réception. Ils pourraient ainsi personnaliser leur sélection de flux pour éviter des décodages inutiles. Le bloc de sélection des locuteurs est celui explicité en 2.1, avec en entrée les probabilités de parole continues de chaque participant.

2.3.1.1.2 Schéma des terminaux

Comme on l'a vu dans la partie précédente, le pont se base sur des informations utiles contenues dans les trames pour effectuer sa sélection. Il est donc nécessaire que nos terminaux déterminent ces informations, comme la VAD, pour les inclure dans les trames en émission. Elles seraient incluses dans les paquets audio pour permettre la sélection des flux codés au niveau du pont. A noter que ceci n'est pas standard au niveau de RTP et que si l'on veut que cela soit indépendant des codecs, il faudrait établir (et normaliser…) un format générique, voire utiliser les mécanismes d'extension d'entête.

La Figure 2.17 illustre le côté réception du terminal.

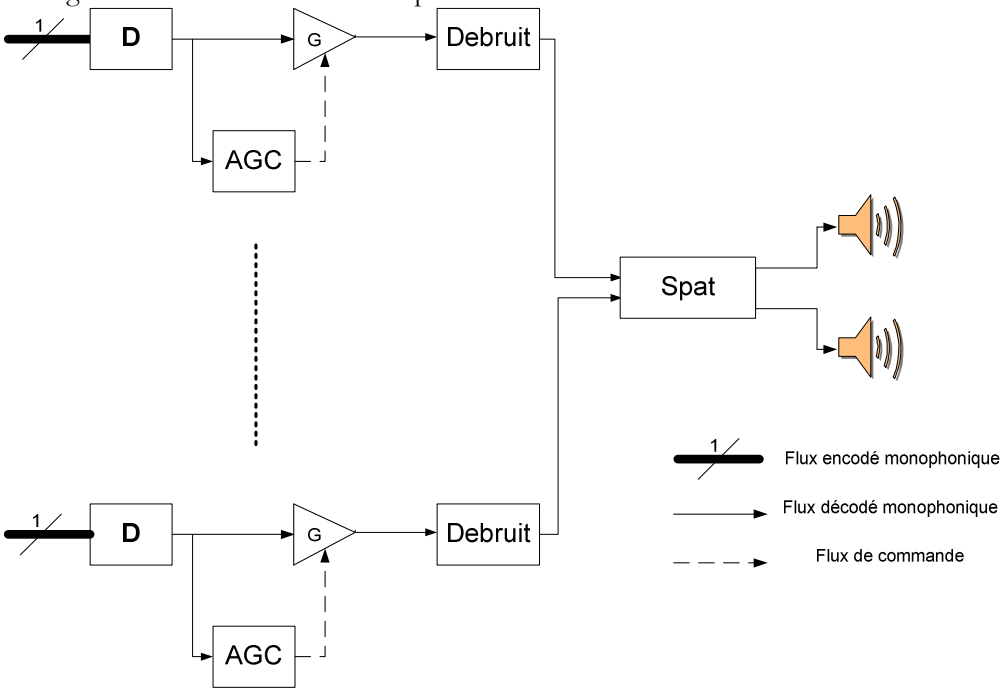

Figure 2.17 Schéma d'un terminal utilisé avec un pont répliquant intelligent

Les flux codés sont décodés par un premier bloc, puis sont traités par les blocs AGC et de débruitage avant d'être envoyés vers le module de spatialisation. A noter que le débruitage pourrait être fait par chaque terminal en émission ce qui aboutirait à un traitement de débruitage par terminal émetteur au lieu de N-1 pour chaque terminal récepteur. Par contre, ce traitement serait moins homogène du fait d'algorithmes différents voire inexistants sur certains terminaux.

Les blocs AGC et débruitage sont ceux explicités dans la section 2.1. Le bloc de spatialisation est le même que celui de la section 2.2.1.1

2.3.1.1.3 Contraintes et Problèmes anticipés

- Il y a beaucoup de contraintes en termes de charges de calcul sur le terminal.
- Les terminaux ne peuvent concrètement ouvrir que 1 ou 2 canaux en réception actuellement.
- Le nombre de participants à la conférence est limité par les capacités des terminaux.
- Tous les terminaux doivent inclure la VAD calculée avec le même algorithme et utilisée avec le même format de transport dans leurs flux audio. Cela implique une homogénéisation des terminaux peu crédible.
- Il n'existe pas de pont répliquant disponible sur le marché.
- L'interopérabilité avec le RTC est compliquée puisqu'elle nécessite une entité dans le réseau effectuant le mixage pour obtenir un flux monophonique.

- Au niveau des terminaux, le fait que certains paquets soient supprimés peut poser des problèmes, selon les codeurs, dans les variables d'état des décodeurs.

2.3.1.1.4 Avantages

- Le pont est très léger logiciellement. Il peut donc gérer de nombreuses conférences.
- Il n'y a pas de transcodage au niveau du pont, donc la qualité est préservée.
- Une réduction de débit importante est possible grâce à la sélection de flux au niveau du pont.
- La spatialisation sur les terminaux permet de spatialiser au mieux dans le sens où elle est adaptée à leurs capacités. Un terminal peut, s'il le souhaite, ne pas effectuer de spatialisation.

2.3.1.2 Version avec calcul de paramètres de détection d'activité vocale après décodage

2.3.1.2.1 Schéma général du pont répliquant avec calcul de paramètres de détection d'activité vocale

La seconde variante proposée ici pour ce pont répliquant est illustrée Figure 2.18. Dans ce module, les flux audio arrivent codés et on décode totalement ou partiellement (suivant le type de codeurs) les paquets audio afin de calculer des paramètres de VAD qui seront envoyés vers un bloc de sélection des locuteurs. Ce bloc commandera le bloc M parmi N qui prendra en entrée les flux codés (retardés pour prendre en compte le temps de décodage et la sélection des locuteurs) issus des participants, et renverra les flux sélectionnés vers le réseau. Cette solution reste intéressante car c'est généralement l'encodage qui a un coût important et non le décodage. Elle inclut cependant un retard δ du fait de ce traitement de décodage et de VAD.

Il est à noter que dans le cas de décodage partiel, cette solution devient codec dépendante puisque la VAD sera spécifique à un codeur. Le bloc VAD est celui explicité dans la section 2.1.

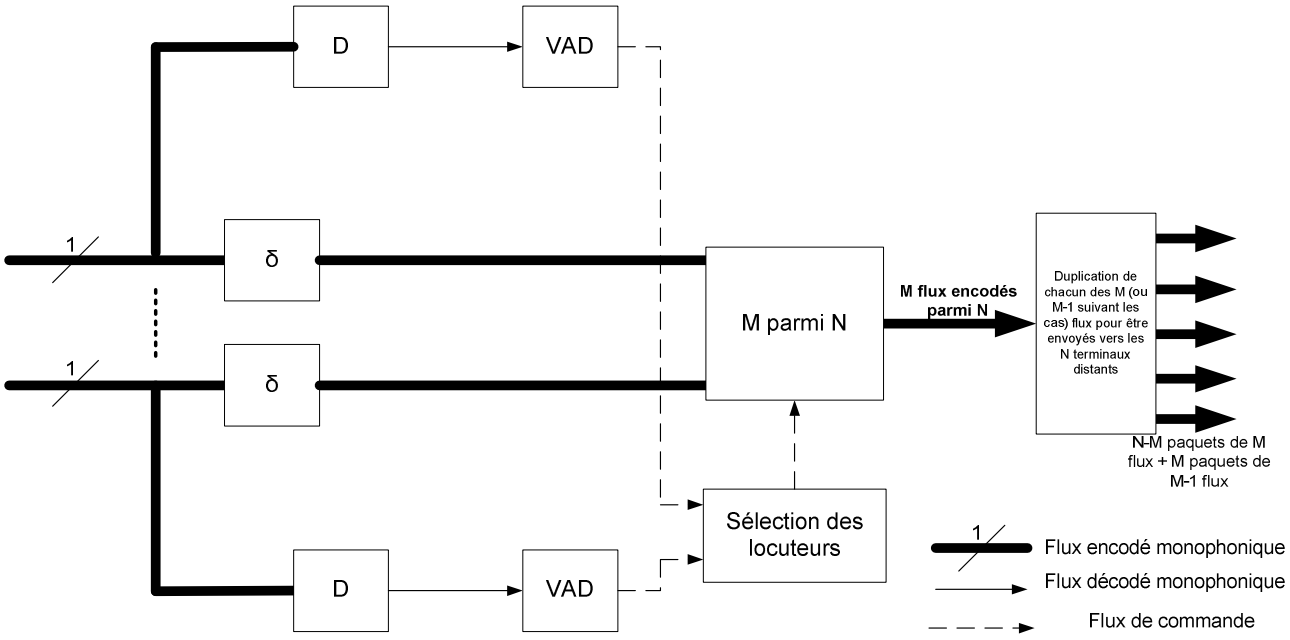

Figure 2.18 Schéma d'un pont répliquant intelligent avec calcul de VAD

2.3.1.2.2 Schéma des terminaux

La seule différence avec 2.3.1.1.2 étant qu'en émission nos terminaux n'effectuent pas la VAD sauf éventuellement pour l'utiliser en local avec un bloc DTX. Ces informations ne sont donc évidemment pas transmises dans les trames montantes vers le pont.

2.3.1.2.3 Contraintes et Problèmes anticipés

- Il y a beaucoup de contraintes sur le terminal ce qui nécessite des terminaux capables de supporter des charges de calcul plus ou moins importantes suivant les options choisies.
- Le nombre de participants à la conférence est limité par les capacités des terminaux.
- Les terminaux ne peuvent concrètement ouvrir que 1 ou 2 canaux.
- Il n'existe pas de pont répliquant disponible sur le marché.
- L'interopérabilité avec le RTC est compliquée puisqu'elle nécessite une entité dans le réseau effectuant le mixage pour obtenir un flux monophonique.

2.3.1.2.4 Avantages

- Le pont est un peu moins léger logiciellement par rapport au précédent mais il peut toujours gérer de nombreuses conférences.
- Du fait de l'absence de transcodage, il n'y a donc pas de perte de qualité au niveau du pont.
- La VAD est sur le pont donc cela entraine moins de contraintes pour le terminal et cela garantit une homogénéisation de son calcul.
- Il y a une réduction de la charge réseau grâce à la sélection de flux au niveau du pont.
- La spatialisation sur les terminaux permet de spatialiser au mieux dans le sens où elle est adaptée à leurs capacités. Un terminal peut, s'il le souhaite, ne pas effectuer de spatialisation.

2.3.2 Les configurations distribuées multi-unicast et multicast

Pour ces deux configurations, les traitements sont exactement les mêmes. Ils se rapprochent de ceux que nous avons présentés pour le pont répliquant dans la section 2.3.1.1.

En émission un terminal peut inclure une VAD pour effectuer de la DTX, ou pour l'inclure comme information complémentaire dans les trames. Il peut de même effectuer du contrôle automatique de gain et du débruitage.

En réception, tout comme sur la Figure 2.17, le terminal peut aussi :

- Effectuer du contrôle automatique de gain et du débruitage.
- Sélectionner les trames avec de la parole si des informations de VAD sont incluses.
- Eventuellement calculer les informations de VAD en sortie du décodage pour effectuer lui-même une sélection de flux (non représenté sur le schéma). Cela est peu intéressant du fait des coûts de décodage et de VAD qui risquent d'être plus importants que ceux économisés pour la spatialisation.
- Spatialiser les différents contenus sélectionnés.

2.3.2.1 Contraintes et Problèmes anticipés

- Beaucoup de contraintes sur le terminal ce qui nécessite des terminaux capables de supporter des charges de calcul plus ou moins importantes suivant les options choisies.
- Les terminaux ne peuvent concrètement ouvrir que 1 ou 2 canaux actuellement, ce qui limite le développement de la conférence audio distribuée multi-unicast.
- Nombre de participants à la conférence limité par les capacités des terminaux.
- Tous les terminaux doivent inclure la VAD dans leurs flux audio si l'on souhaite faire une sélection de flux utile. Cela implique une homogénéisation des terminaux non crédible.
- L'interopérabilité avec le RTC est compliquée puisqu'elle nécessite une entité dans le réseau effectuant le mixage pour obtenir un flux monophonique.

2.3.2.2 Avantages

- L'absence de transcodage évite une perte de qualité.
- La spatialisation sur les terminaux permet de spatialiser au mieux dans le sens où elle est adaptée à leurs capacités. Un terminal peut, s'il le souhaite, ne pas effectuer de spatialisation.
- Il n'y a pas d'entité centrale comme un pont de conférence à gérer.

2.3.3 Conclusion

Les architectures présentes dans cette section montrent les différentes possibilités au niveau média pour faire de la conférence audio spatialisée sans entité de mixage dans le réseau à condition que les terminaux aient les capacités en termes de ressources et de bande passante en réception. Il faut de plus que les terminaux aient en commun des codeurs pour pouvoir communiquer.

2.4 Proposition d'une architecture de pont de conférence mixte pour optimiser l'utilisation des terminaux et des ponts de conférences

Nous avons vu dans les paragraphes précédents que les traitements dépendaient des capacités des différents acteurs physiques de la conférence : terminaux, pont mixeur ou répliquant. L'idée dans cette section est de proposer une architecture dite mixte pouvant utiliser au mieux les capacités de chacun.

Nos points de départ sont les deux architectures centralisées vues auparavant. Dans un premier temps, une comparaison sera effectuée entre ces dernières pour lister les avantages et inconvénients de chacune, avant de les illustrer par un exemple.

Cette architecture mixte a fait l'objet d'un dépôt de brevet de numéro Fr 06 55908.

2.4.1 Comparaison des deux architectures centralisées

2.4.1.1 Avantages et inconvénients d'un pont répliquant

On rappelle les avantages du pont répliquant :

- Ce pont est très léger logiciellement ce qui permet de supporter de nombreuses conférences.
- Les conférences standards sont généralement composées de moins de cinq personnes donc réserver une entité telle qu'un pont mixeur est inutile.
- Du fait de l'absence de transcodage au niveau de ce type de pont, il n'y a pas de perte de qualité audio et pas de retard supplémentaire, contrairement à un pont mixeur.
- Les contraintes en termes de flux sont un flux montant et N-1 descendants pour chaque terminal, si N est le nombre de participants à la conférence et si tous les participants envoient chacun un flux audio. Pour le terminal, cela est mieux dans le sens où c'est souvent le flux montant vers un serveur qui est limitant (par exemple avec un accès asymétrique comme l'ADSL).
- Pour une éventuelle spatialisation, les terminaux peuvent spatialiser de la manière qu'ils souhaitent suivant leurs capacités locales.

Les inconvénients du pont répliquant sont les suivants :

- Les terminaux ne peuvent aujourd'hui ouvrir que 1 ou 2 canaux simultanément ce qui explique une absence de pont répliquant sur le marché.

- Le nombre de participants à la conférence est limité par les capacités des terminaux.
- La bande passante descendante au niveau d'un terminal est importante.
- Les ressources CPU nécessaires au décodage et au mixage des différents flux descendants au niveau d'un terminal sont proportionnelles au nombre de participants.
- Les terminaux doivent avoir des codeurs communs pour communiquer. Il existe donc une codec-dépendance de leur part.
- L'interopérabilité avec les réseaux voix existants est complexe (RTC par exemple). Il faut une entité dans le réseau pour effectuer un mixage comme une passerelle (IP/RTC par exemple).
- La négociation des codecs peut entraîner de nombreux messages de signalisation entre le pont et les terminaux, le pont servant d'intermédiaire entre les participants.
- Toujours pour les terminaux, dans l'hypothèse d'une spatialisation, ceux-ci doivent gérer N-1 flux à spatialiser.

2.4.1.2 Avantages et inconvénients d'un pont mixeur

On rappelle les avantages du pont mixeur :

- Les terminaux légers ne gèrent qu'un flux montant et un flux descendant et les décodage/codage associés.
- La gestion de la conférence est possible avec de nombreux participants du fait d'une entité centrale dans le réseau.
- Le bloc de spatialisation n'est pas au niveau des terminaux mais sur le pont, ce qui permet en plus de gérer facilement une scène sonore commune entre tous les participants.
- Il y a un net gain en bande passante lorsqu'il y a de nombreux participants par rapport à un pont répliquant même doté d'une sélection de flux.
- L'interopérabilité avec les réseaux voix existants est aisé (RTC par exemple).
- La puissance des ponts croît de manière régulière.
- Suivant la "richesse" du pont en terme de codecs, la conférence audio est codec-indépendante pour les terminaux.

Les inconvénients du pont mixeur sont les suivants :

- Le traitement nécessite beaucoup de ressources sur le pont. Il convient de trouver un compromis entre bande passante et CPU. Le pont doit effectivement décoder N flux, effectuer N mixages de N-1 flux, encoder N flux mixés et envoyer ces derniers vers les N participants de la conférence. Des serveurs intermédiaires effectuant une partie de ce travail peuvent soulager un serveur central de mixage pour les très grandes conférences.
- Il est dommage de réserver un pont et ses moyens pour des conférences à 3 car les terminaux peuvent avoir les ressources suffisantes pour faire du décodage/mixage pour 2 participants distants.
- Il est difficile de spécifier la spatialisation souhaitée.
- La qualité audio est moins bonne du fait du transcodage et du retard plus important.

2.4.1.3 Comparaison des deux architectures en termes de débit et coût CPU pour les ponts et pour les terminaux

Les courbes, illustrant les contraintes en termes de bande passante et de coût CPU des différentes architectures, en fonction du nombre de participants, ont été tracées. Les hypothèses faites sont les suivantes : un encodage ou un décodage a un coût unitaire, tout comme le traitement CPU. Cela ne correspond évidemment pas à un cas réel où les codeurs-décodeurs n'ont pas le même coût, mais cela permet d'évaluer grossièrement les architectures.

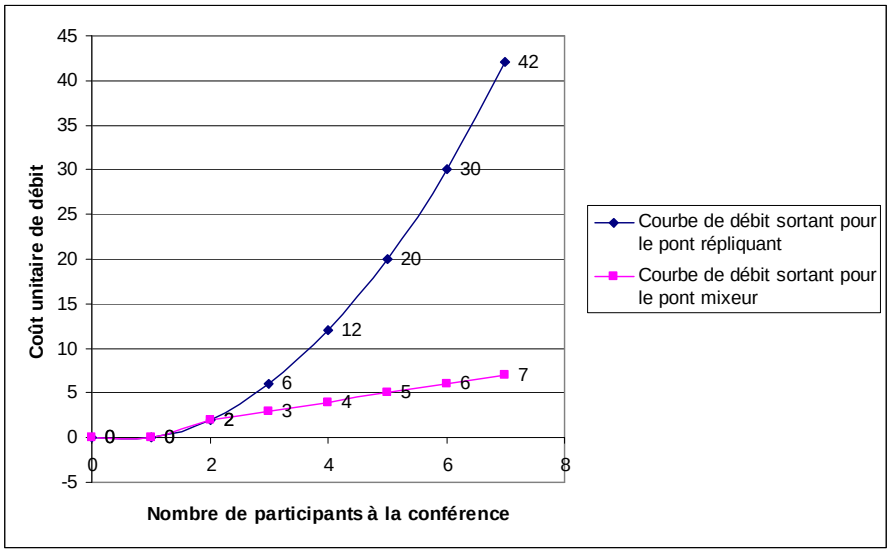

Figure 2.19 Comparaison en termes de bande passante sortante pour les deux ponts

Comme illustré Figure 2.19, les débits sortants des deux types de pont, pour N le nombre de participants, sont de la forme :

- y = N pour le pont mixeur avec N ≥ 2, sinon 0.
- y = N.(N-1) pour le pont répliquant avec N ≥ 2, sinon 0.

Pour les débits entrants au niveau des ponts, ils sont du type y = N.

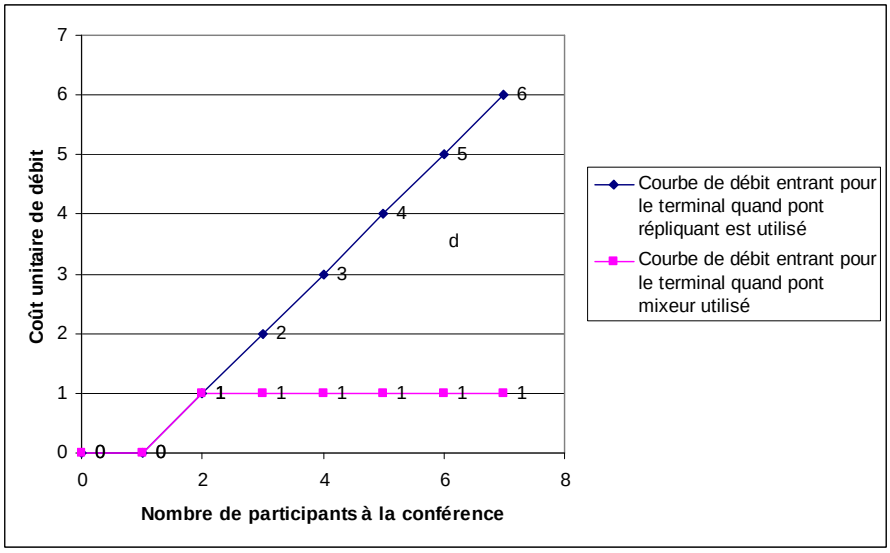

Figure 2.20 Comparaison en termes de bande passante entrante pour un terminal de chaque architecture

Comme illustré Figure 2.20, les débits entrants des terminaux sont de la forme :

- y = 1 avec le pont mixeur avec N ≥ 2, sinon 0.
- y = N-1. avec le pont répliquant avec N ≥ 2, sinon 0.

Pour les débits sortants au niveau des terminaux, ils sont du type y = 1.

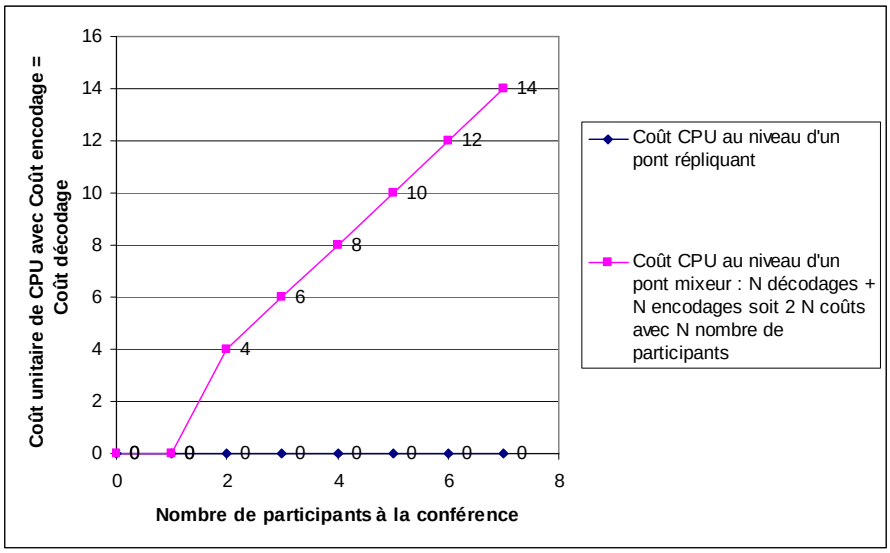

Figure 2.21 Comparaison en termes de CPU pour les deux ponts

Comme illustré Figure 2.21, les coûts CPU au niveau des ponts sont de la forme :

- y = 2.N pour le pont mixeur avec N ≥ 2, sinon 0. Il y a en effet N décodages puis N encodages au niveau du pont. Il n'a pas été tenu compte des coûts de mixage et autre traitement de qualité qui impliquerait une translation de la courbe suivant l'axe des ordonnées.
- y = 0 pour le pont répliquant.

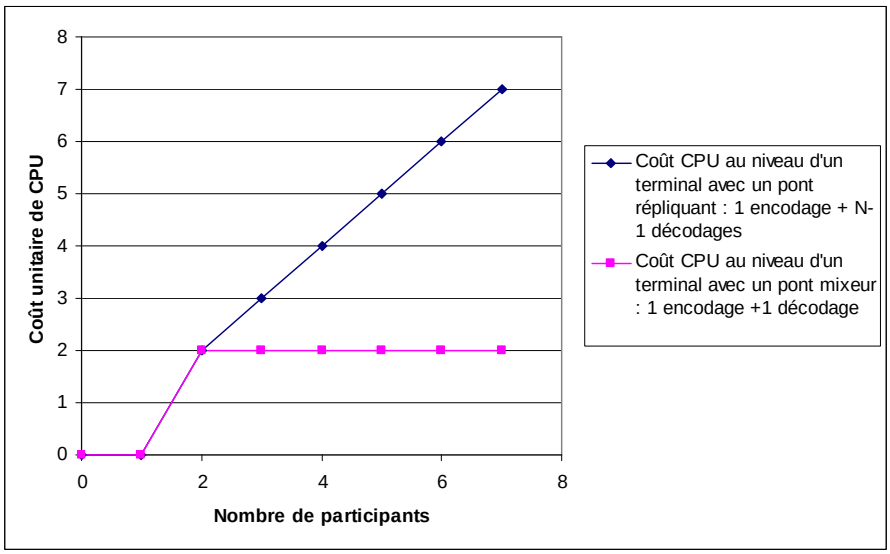

Figure 2.22 Comparaison en termes de CPU pour un terminal de chaque architecture

Comme illustré Figure 2.22, les coûts CPU des terminaux sont de la forme :

- y = 2 avec le pont mixeur avec N ≥ 2 sinon 0. Il y a en effet 1 encodage et 1 décodage au niveau du terminal.
- y = N avec le pont répliquant avec N ≥ 2. Il y a en effet 1 encodage et N-1 décodages au niveau du terminal.

2.4.1.4 Avantages des ponts de conférence en général

Malgré les défauts exposés ci-dessus des différents types de ponts, l'utilisation de pont dans le cadre de la conférence est importante car :

- Réduction du trafic de messages de signalisation en VoIP.

- Gestion facilitée de la conférence comme cela a été vu dans le chapitre précédent : Invitation, Suppression, Diffusion des informations, Contrôle et administration facilités
- Pour les opérateurs, cela permet de facturer facilement les services.

2.4.1.5 *Illustration des problèmes de ces deux architectures*

Comme le laissent deviner les paragraphes précédents, il est peu intéressant d'utiliser des ponts répliquants pour un nombre important de participants et un pont mixeur pour un nombre faible de participants. Cela est illustré sur les Figure 2.23 et Figure 2.24.

Par exemple, comme le montre la Figure 2.23, un pont mixeur est utilisé pour gérer la conférence audio de 3 participants. Il mixe donc les flux 1 et 2 pour le participant 3, les flux 1 et 3 pour le participant 2 et les flux 2 et 3 pour le participant 1.

Cependant dans ce cas, il n'est peut-être pas tenu compte des capacités des terminaux qui pourraient supporter la bande passante descendante ainsi que la charge de calculs nécessaires au mixage. En faisant cette supposition qui est acceptable à la vue des possibilités des terminaux actuels, cela permettrait d'éviter le transcodage au niveau du pont.

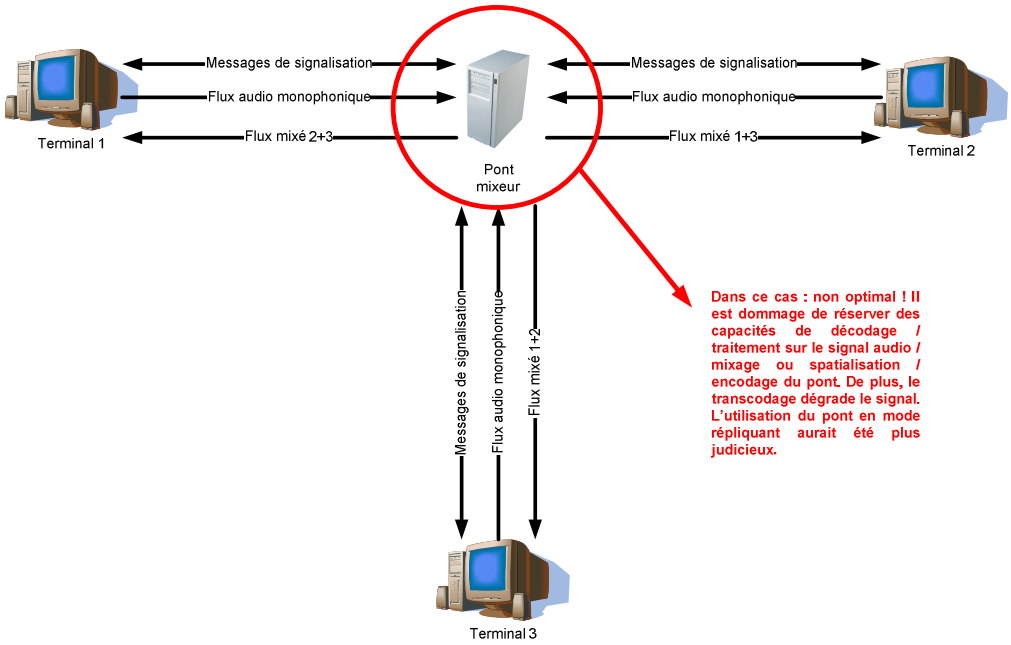

Figure 2.23 Utilisation peu intéressante du pont mixeur

Par exemple, comme le montre la Figure 2.24, un pont répliquant est utilisé pour gérer la conférence audio de 5 participants. Il transmet par exemple au terminal 3 les flux des participants 1, 2, 4 et 5.

Cependant dans ce cas, le terminal 3 est susceptible de ne pas pouvoir accepter autant de flux et il est probable que les flux ne puissent même pas être envoyés vers lui puisqu'ils seront refusés au moment de la négociation. En supposant cela réalisable, au niveau du traitement, il faut de plus qu'il dispose des codecs appropriés et surtout qu'il effectue le traitement de décodage et de mixage ce qui reste hypothétique sur un terminal peu puissant, comme un mobile par exemple.

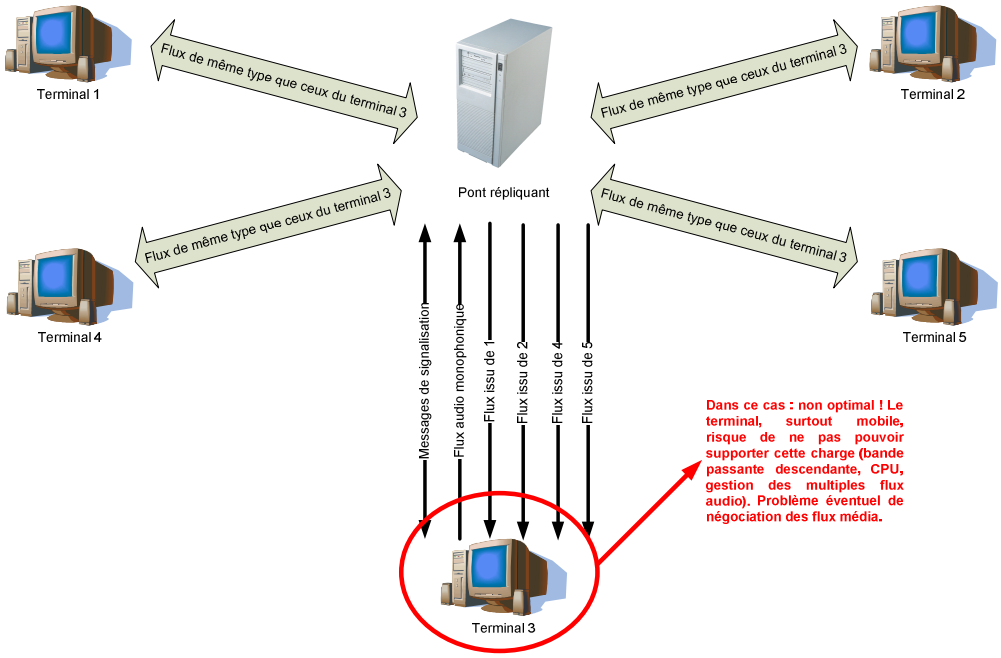

Figure 2.24 Utilisation peu intéressante du pont répliquant

L'objectif de notre architecture est d'optimiser la gestion d'un pont de conférence et des terminaux associés en utilisant un pont mixte répliquant ou mixeur suivant les cas de figure lors d'une conférence audio. Il est de même proposé un moyen de déterminer les limites pour lesquelles il est conseillé d'utiliser un pont répliquant ou un pont mixeur.

2.4.2 Solution proposée

La solution s'inscrit dans le cas d'une utilisation point à point (2 terminaux en discussion) ou multipoint (plusieurs terminaux en discussion), et évidemment en mode centralisé.

La solution est de déterminer un mode de fonctionnement d'un pont mixte afin d'optimiser l'utilisation d'un pont et des terminaux lors d'une communication.

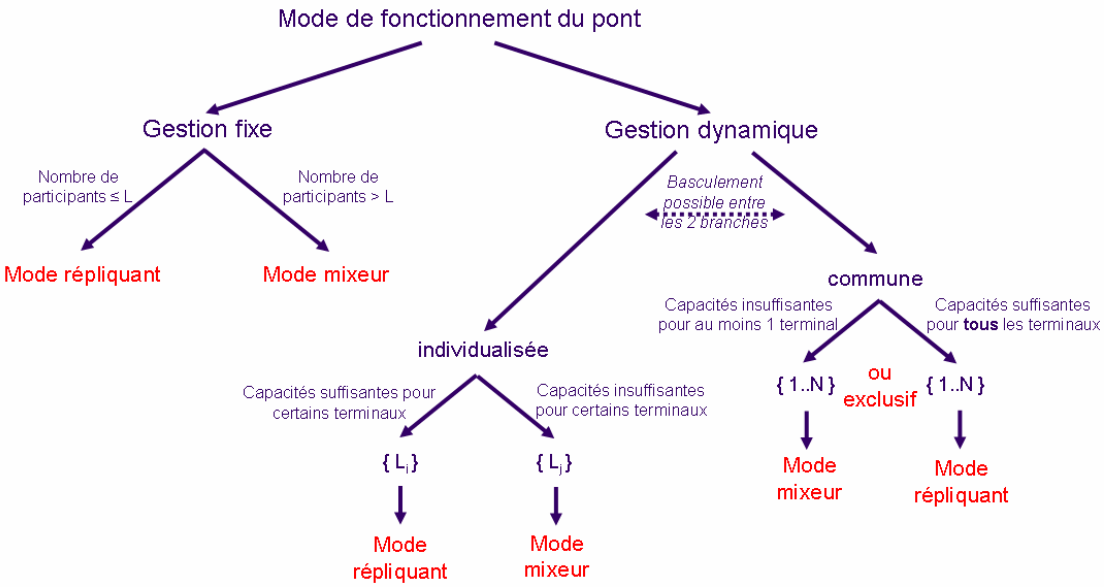

Figure 2.25 Illustration du fonctionnement d'un pont mixte

Le mode de fonctionnement du pont mixte est établi Figure 2.25. Ce schéma illustre comment le pont se comporte vis-à-vis des terminaux. Deux gestions sont possibles : gestion dynamique ou gestion fixe.

2.4.2.1 Gestion dynamique

Initialement, le pont sera par défaut en mode répliquant. Ce choix s'explique par le fait que les participants arrivent les uns après les autres en conférence. Il est donc logique de choisir le mode qui fonctionne avec le plus petit nombre de participants. Il cherchera à négocier de cette manière le plus longtemps possible une conférence (par l'intermédiaire d'un protocole de signalisation comme SIP) entre les participants.

Autrement dit, il cherchera à gérer à la fois la signalisation et le relais des flux média. En SIP par exemple, la négociation des flux média se fera comme pour un appel SIP point à point standard entre chaque participant et le serveur de conférence. Par contre elle se différencie dans le sens où, pour un terminal, il faudra spécifier un flux en émission et N-1 en réception (dans le cas où il y a N participants à la conférence), au lieu d'un flux bidirectionnel dans le cas d'un pont mixeur. Le pont de conférence négociera l'opposé, c'est-à-dire un flux en réception et N-1 en émission pour chaque participant.

Si ces négociations se passent bien pour chaque participant, c'est-à-dire que les canaux audio sont bien créés avec les codecs adéquats, cela veut dire que tous les terminaux ont la capacité de supporter cette bande passante descendante et que la charge CPU limite (pour les traitements de décodage/mixage) n'est pas atteinte. Tous les terminaux pourront ainsi participer de manière optimale à la conférence.

Si au moins un des participants ne peut supporter cette charge, il est alors possible :

- De lui/leur mixer les flux audio pour n'en former plus qu'un, après renégociation d'un flux média (voir paragraphe ci-dessous pour des détails en SIP). On aura donc d'un côté des terminaux qui recevront un flux audio mixé, car ils n'ont pas les capacités suffisantes et de l'autre des terminaux qui recevront plusieurs flux audio qu'ils pourront traiter à leur guise. On parlera alors de gestion dynamique individualisée. Une bascule totale peut être faite vers le mode mixeur si, par exemple, trop peu de terminaux restent avec le mode répliquant.
- De renégocier pour tous le passage en un flux mixé. On parlera alors de gestion dynamique commune. En SIP, par exemple, la négociation des flux média se fait comme pour un appel SIP point à point standard entre chaque participant et le pont de conférence. C'est à ce dernier de gérer, de décoder et de mixer les flux venant des participants et de renvoyer les résultats vers chacun d'eux.

Le fonctionnement du pont mixte est dans ce cas en gestion dynamique, c'est-à-dire que l'entrée ou le départ d'un nouveau participant va entraîner des négociations entre les terminaux et les ponts, afin de déterminer comment le pont va se comporter envers certains participants ou tous les participants. Dans ce cas, aucune hypothèse n'est faite sur les terminaux et leurs capacités ce qui est évidemment le cas le plus intéressant, car le plus général.

2.4.2.2 Gestion fixe

Le pont peut aussi fonctionner en gestion fixe (par opposition à la gestion dynamique). Autrement dit, on peut spécifier que toutes les conférences avec un nombre de participants inférieur ou égal à une limite L doivent être gérées par un pont répliquant et les autres avec un pont mixeur. Cela est intéressant dans le cas où on se trouve par exemple sur un LAN (Local Area Network) sur lequel on connaît les machines, leurs activités (si elles sont dédiées ou non) et leurs limites. Lors de l'entrée ou la sortie d'un nouveau participant, le pont peut ainsi en fonction du nombre de conférenciers savoir comment négocier les flux. La contrainte forte est la connaissance du matériel disponible ce qui en cas de problème peut aboutir à ce que des terminaux ne puissent suivre la conversation.

Les négociations en SIP sont comme celles spécifiées dans la section 2.4.2.1.

2.4.2.3 *Variantes possibles*

On peut décider de fixer l'utilisation du pont mixte en mode répliquant ou mixeur dès le départ afin de bénéficier des avantages du pont répliquant ou du pont mixeur.

Il est à noter que lorsque l'on parle de mixage qui comprend tout ce qui va du mixage simple standard aux options incluant différents traitements audio (contrôle automatique de gain, …) ainsi que de la spatialisation. Pour cette dernière, le changement de mode nécessite de tenir compte des positions dans les espaces de restitution. En effet pour un pont mixeur, la spatialisation se fait sur le pont alors que sur un pont répliquant, elle se fait sur le terminal. Cela implique lors d'un changement de mode que la cohérence des localisations des sources soit respectée. Il est donc nécessaire d'établir un dialogue entre le terminal et le pont à travers le protocole de signalisation par exemple afin que ces informations soient transmises.

Ce traitement n'empêche donc pas des traitements plus complexes comme la spatialisation et permet à l'inverse à des terminaux disposant d'un seul haut-parleur et de capacités en bande passante ou CPU très limitées de participer à la conférence, sans dégrader la qualité de celles des autres.

Il est à noter de plus que l'on parle dans ce document de pont mixte, mais il faut bien comprendre que cela peut être deux entités physiques différentes, avec des adresses IP différentes, mais contrôlées par la même entité "intelligente" contenant les capacités de dialogue avec les terminaux à travers un protocole de signalisation comme SIP.

2.4.2.4 *Description d'un mode particulier de la solution proposée*

Dans cette partie, un mode particulier de réalisation est décrit avec un schéma fonctionnel du pont ainsi que les descriptions de fonctionnement du pont en mixte en gestion dynamique et fixe.

2.4.2.4.1 Schéma fonctionnel du pont

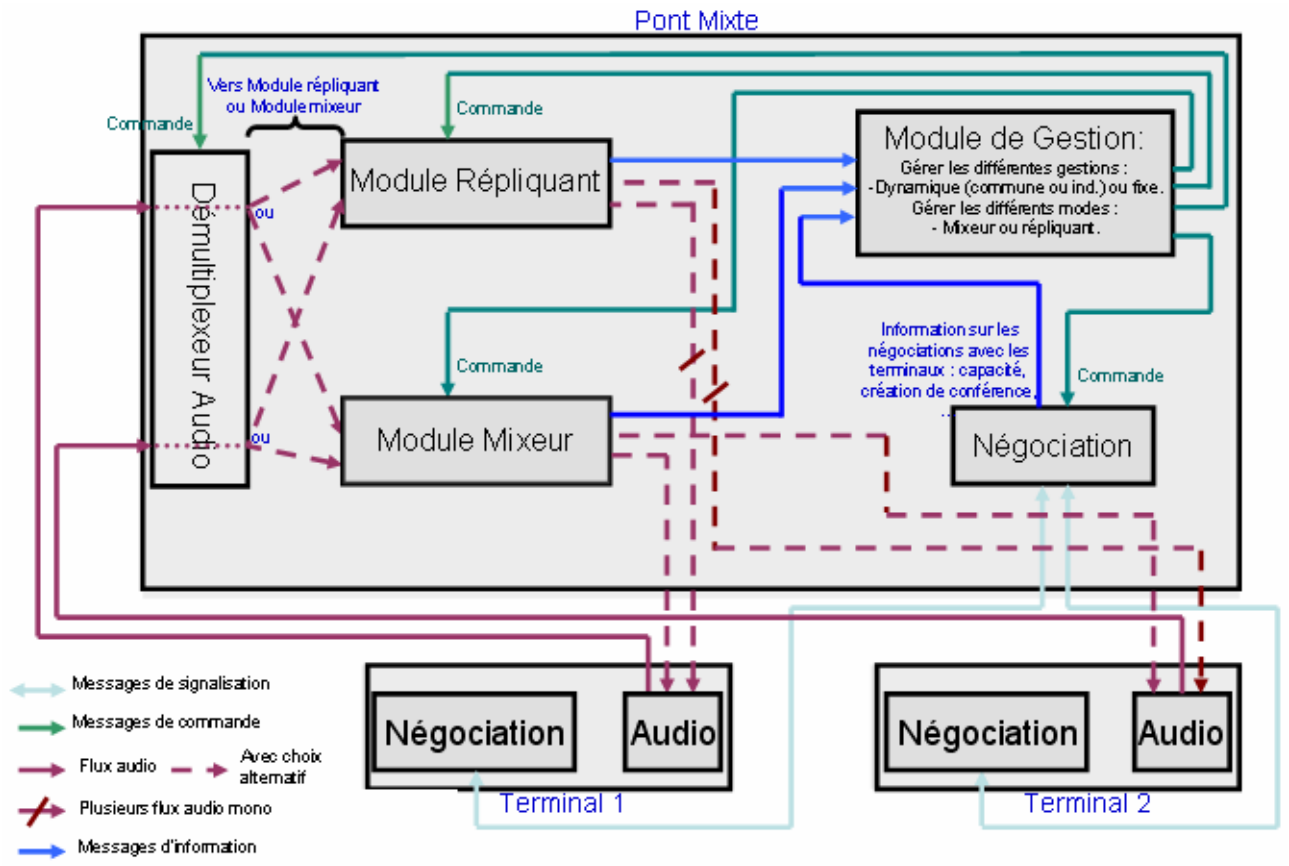

Figure 2.26 Schéma fonctionnel du pont mixte

Comme l'illustre la Figure 2.26, le pont mixte peut être vu comme une entité composée de 5 modules :

- Un module négociation, dont le rôle est d'échanger les messages de signalisation avec les terminaux. Il permet de créer des conférences, négocier avec les terminaux et en déduire leurs capacités. Ce module transmet ces informations vers le module de gestion et en reçoit les ordres. Ces ordres permettent de répondre ou de proposer des alternatives aux propositions des terminaux, suivant que le pont mixte fonctionne en mode mixeur et/ou répliquant, pour une conférence donnée.
- Un module démultiplexeur audio, qui va gérer les flux audio entrants issus des terminaux et les envoyer vers les modules Répliquant et/ou Mixeur, suivant les consignes du module de gestion.
- Un module pont répliquant, dont le rôle va être de transférer les flux audio issus de chaque participant vers tous les autres participants. Il est commandé par le module de gestion et lui envoie des informations concernant sa charge CPU ou les limites pour le débit sortant ou entrant. Il reçoit des flux audio par l'intermédiaire du démultiplexeur audio, les duplique et les renvoie vers les différents terminaux. (Les modules, comme l'encodage ou le décodage, pour les interfaces audio ne sont pas représentés)
- Un module pont mixeur, dont le rôle va être de mixer les différents flux audio en fonction de chaque participant, avant de lui envoyer le flux résultant. Il est commandé par le module de gestion et lui envoie des informations concernant sa charge CPU ou les limites pour le débit sortant ou entrant. Il reçoit des flux audio par l'intermédiaire du démultiplexeur audio, les mixe et les renvoie vers les différents terminaux. (Les modules, comme l'encodage ou le décodage, pour les interfaces audio ne sont pas représentés)
- Un module de gestion, qui contrôle le pont mixte. Il analyse les renseignements reçus de la part des différents modules et leur envoie des commandes. En fonction des renseignements reçus ou de la façon dont il est programmé, il fonctionne sous une gestion fixe ou dynamique (individualisée ou commune), et ordonne pour chaque conférence qu'il héberge d'utiliser les modules mixeurs et/ou répliquant.

2.4.2.4.2 Fonctionnement du pont mixte en gestion dynamique

Cette partie présente succinctement le fonctionnement des modules du pont et leurs dialogues, lors de la création d'une seule conférence, dans le cas de la gestion dynamique explicitée dans la section 2.4.2.1. Il est bien évidemment possible de gérer plusieurs conférences à plusieurs participants en parallèle.

On se met dans le cas où une conférence est créée avec plusieurs participants en discussion et le pont est en attente de l'arrivée ou du départ d'un participant. Lorsque cela se produit, le module de négociation avertit le module de gestion qu'un participant vient de rejoindre ou quitter la conférence. Le module de gestion avertit à son tour le démultiplexeur audio qu'il va recevoir ou non des paquets audio pour ce participant. Il y a ensuite deux types de gestion dynamique : individualisée ou commune.

Pour la première citée, à partir du moment où le pont fonctionne dans les deux modes (répliquant et mixeur) pour une même conférence, le démultiplexeur audio envoie les flux audio reçus vers le module répliquant et le module mixeur.

Lors de l'arrivée d'un nouveau participant, le module négociation négocie avec le terminal en tenant compte d'un pont répliquant.

- Si le terminal supporte le mode répliquant, le module de gestion avertit le pont répliquant de tenir compte du nouveau participant.
- Dans le cas contraire, le module de négociation négocie avec le terminal en tenant compte d'un pont mixeur. Si le terminal supporte le mode mixeur, le module de gestion avertit le pont mixeur de tenir compte du nouveau participant.

- Le module de négociation doit renégocier avec tous les terminaux utilisant le pont en mode répliquant pour ouvrir des canaux supplémentaires et éventuellement les faire utiliser le pont en mode mixeur, si la négociation échoue.

Lors du départ d'un participant, le module de gestion, informé par le module négociation, avertit le pont mixeur et/ou répliquant qu'il doit en tenir compte. Il peut éventuellement demander au module de négociation de renégocier le passage en mode répliquant pour les terminaux utilisant le pont mixeur. Pour les terminaux utilisant le mode répliquant du pont, les canaux relatifs au participant partant sont fermés après renégociation.

Si trop peu de participants sont en mode répliquant, le module de gestion peut demander au module de négociation de négocier en tenant compte d'un pont mixeur pour une conférence donnée.

Pour la gestion dynamique commune, dans l'hypothèse où le pont est en mode répliquant et lors de l'arrivée d'un participant, le module de gestion commande le module de négociation de négocier avec tous les participants en tenant compte d'un pont répliquant et avec le flux audio supplémentaire du nouveau participant :

- Si tous les terminaux supportent le mode répliquant, rien ne change. Le module de gestion avertit le pont répliquant qu'il doit tenir compte de l'arrivée d'un participant.
- Dans le cas contraire, le module de gestion ordonne au pont mixeur de traiter les flux audio de cette conférence et au pont répliquant d'arrêter cette même tâche. Le module démultiplexeur audio reçoit des ordres en conséquence. Le module de négociation doit renégocier avec tous les terminaux pour passer en mode mixeur.

Si le pont est en mode mixeur et lors du départ d'un participant :

- Le module de gestion avertit le pont mixeur qu'il doit tenir compte du départ d'un participant.
- Le module de gestion peut demander au module de négociation de renégocier avec tous les terminaux restants pour repasser éventuellement en mode répliquant. Cela ne se fera que si tous les terminaux restants supportent le mode répliquant.

Si le pont est en mode mixeur lors de l'arrivée d'un nouveau participant ou si le pont est en mode répliquant lors du départ d'un participant, les modes ne changent pas. Le module de gestion avertit :

- Dans le premier cas, le pont mixeur qu'il doit tenir compte de l'arrivée d'un participant et le module de négociation qu'il doit négocier avec le participant en tenant compte d'un pont mixeur.
- Dans le second cas, le pont répliquant qu'il doit tenir compte du départ d'un participant, et le module de négociation qu'il doit négocier avec les terminaux pour fermer les canaux du participant partant.

2.4.2.4.3 Fonctionnement du pont mixte en gestion fixe

Cette partie présente succinctement le fonctionnement des modules du pont et leurs dialogues, lors de la création d'une seule conférence, dans le cas de la gestion fixe explicitée dans la section 2.4.2.2. Il est bien évidemment possible de gérer plusieurs conférences à plusieurs participants en parallèle.

On se met à nouveau dans le cas où une conférence est créée avec plusieurs participants en discussion et le pont est en attente de l'arrivée ou du départ d'un participant. Lorsque cela se produit, Le module de négociation avertit le module de gestion qu'un participant vient de rejoindre ou de quitter la conférence. Le module de gestion avertit le démultiplexeur audio qu'il va recevoir ou non des paquets audio pour ce participant.

On rappelle que la gestion fixe dépend du nombre de participants à la conférence. Si ce nombre est inférieur ou égal à une limite L, la conférence est gérée par un pont répliquant et dans le cas contraire par un pont mixeur.

Dans l'hypothèse où le nombre de participants N est inférieur ou égal à L :

- Si on vient de passer de N=L+1 à N=L suite au départ d'un participant, le module de gestion ordonne au pont répliquant de traiter les flux audio de cette conférence et au pont mixeur d'arrêter. Le module démultiplexeur audio reçoit des ordres en conséquence. Le module de négociation doit renégocier avec tous les autres terminaux pour le passage en mode répliquant.
- En cas d'arrivée d'un participant, le module de gestion ordonne au module de négociation de négocier avec le participant en tenant compte d'un pont répliquant. Le module de négociation doit aussi renégocier avec les autres terminaux pour l'ouverture de canaux supplémentaires pour ce nouveau participant.
- En cas de départ d'un participant, le module de négociation doit renégocier avec les terminaux restants la fermeture des canaux relatifs à ce participant.
- Le module de gestion avertit le pont répliquant qu'il doit tenir compte de l'arrivée ou du départ d'un participant.

Dans l'hypothèse où N est strictement supérieur à L :

- Si N = L+1 suite à l'arrivée d'un nouveau participant, le module de gestion ordonne au pont mixeur de traiter les flux audio de cette conférence et au pont répliquant d'arrêter cette même tâche. Le module démultiplexeur audio reçoit des ordres en conséquence. Le module de négociation doit négocier avec tous les autres terminaux pour le passage en mode mixeur.
- Le module de gestion avertit le pont mixeur qu'il doit tenir compte de l'arrivée ou du départ d'un participant.
- En cas d'arrivée d'un participant, le module de gestion ordonne au module de négociation de négocier avec le participant en tenant compte d'un pont mixeur.

2.4.2.4.4 Illustration de la solution proposée

Un exemple d'application est la conférence audio spatialisée en VoIP, avec le protocole de signalisation SIP et plusieurs terminaux en conférence dont les flux média sont centralisés sur un pont de conférence.

Le pont fonctionne en gestion dynamique individualisée et par défaut en mode répliquant.

Initialement, trois terminaux (T1, T2 et T3) sont en communication et ils sont capables de recevoir et de traiter les flux audio des autres. Les négociations de flux se sont bien déroulées et le pont reste donc en mode répliquant comme illustré Figure 2.27.

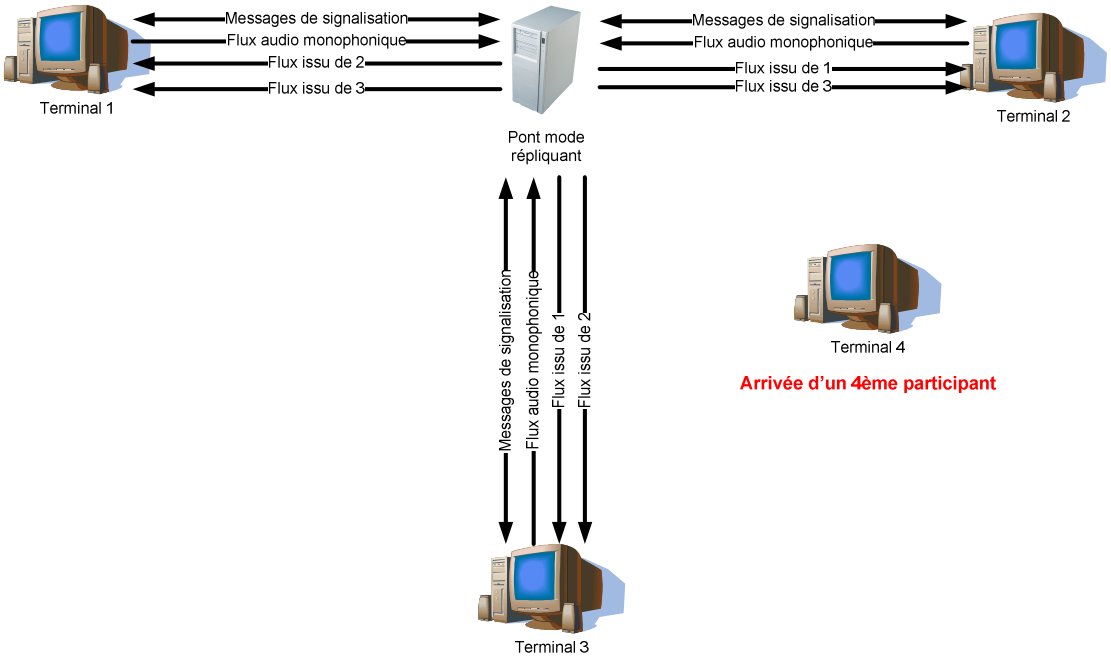

Figure 2.27 Fonctionnement du pont mixte avant l'arrivée du participant 4

Un 4ème participant (T4) rejoint la conférence et des renégociations SIP s'effectuent. Les participants T1 et T3 ne peuvent ouvrir de canaux supplémentaires. Le pont renégocie donc pour eux un seul flux mixé en retour. Pour les participants T2 et T4 qui ont suffisamment de capacités, trois canaux descendants sont ouverts avec le pont et ils peuvent traiter les flux audio de la manière qu'ils souhaitent avec par exemple une spatialisation au lieu d'un simple mixage, cela sans influence pour T1 et T3. Le résultat est illustré Figure 2.28.

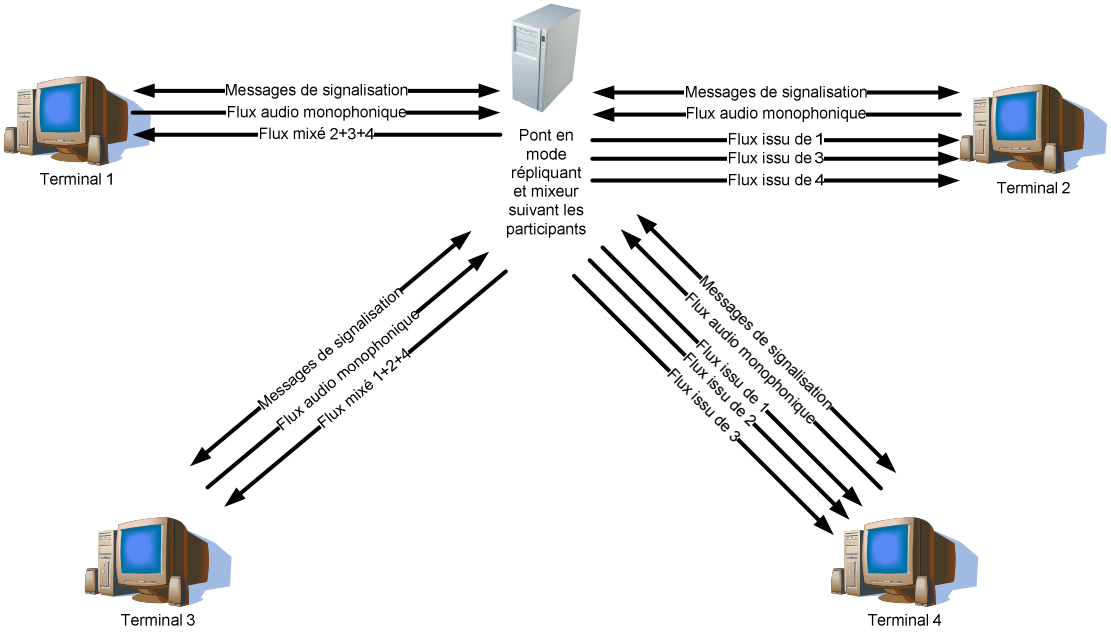

Figure 2.28 Fonctionnement du pont mixte après l'arrivée du participant 4

2.4.2.5 Avantages de cette solution

- Optimisation de l'utilisation des ponts de conférence et des terminaux.
- Traitement adapté à la dynamique des conférences.
- Traitement adapté à des terminaux hétérogènes en CPU, accès et bande passante.

- Traitement adapté dans le cas de l'utilisation de traitement audio comme la spatialisation à condition de gérer la cohérence des images spatiales entre les modes répliquant et mixeur.

2.5 Masquage auditif temps réel pour un pont répliquant en VoIP

Comme cela a été vu dans la partie 2.4.1.1, les avantages d'un pont répliquant sont notamment l'absence de transcodage par rapport à une configuration centralisée avec pont mixeur, et l'envoi d'un seul flux montant d'un participant vers un pont par rapport à une configuration distribuée multi-unicast. Cependant, l'un des inconvénients majeurs est la consommation excessive de bande passante du pont vers les terminaux dès que le nombre de participants augmente. Pour diminuer cette bande passante, nous présentons dans cette section une solution basée sur un algorithme de masquage pour ne transmettre que les trames audibles à un instant donné et pour rejeter les autres.

Outre son principe de masquage, cette solution se distingue de celle présentée dans [88] par le fait qu'elle ne fait pas de supposition sur le nombre de locuteurs à un instant donné et qu'elle ne supprimera pas d'information si plus de deux personnes parlent en même temps. Cette solution vise plutôt le contexte de jeu virtuel (beaucoup de locuteurs simultanés) même si son application est évidemment possible pour la conférence audio en entreprise.

L'algorithme sera tout d'abord présenté, ainsi que par la suite ses performances et son intégration dans un pont répliquant VoIP de France Télécom. Cette section est le résultat d'un travail réalisé avec l'INRIA (Institut National de Recherche en Informatique et Automatique) dans le cadre d'un projet collaboratif RNTL de nom OPERA (Optimisation PErceptuelle du Rendu Audio) [94]. Elle a abouti à une publication pour l'AES 30 [65].

2.5.1 Principe et évaluation de l'algorithme de masquage

Nous nous intéressons dans cette partie à la validation de l'algorithme dans un contexte non temps-réel. Le but de l'algorithme développé par l'INRIA est de comparer les trames de différents flux audio à un instant donné et de rejeter avant mixage celles qui seront masquées par les autres. L'intérêt comme on le verra par la suite est son implémentation temps-réel au sein d'un pont répliquant pour que ces trames ne soient tout simplement pas envoyées du pont vers les terminaux. Nous présentons ici brièvement son principe. Plus de détails sont disponibles dans [30,92,93].

La première étape consiste à séparer les différents flux audio (échantillonnés à 44.1 kHz) en trames. Ensuite, sur chacune de ces trames, est effectuée une transformée de Fourier à court terme sur 1024 échantillons. Puis, l'algorithme calcule la densité spectrale de puissance sur chacune des 30 premières bandes de Bark. On rappelle que les bandes de Bark ou bandes critiques sont des bandes fréquentielles dans lesquelles l'ouïe humaine regroupe des excitations sonores ayant des fréquences voisines.

Une fois cela effectué, et après utilisation d'une fonction d'étalement (*spreading function*) [69,98], on somme les 30 valeurs obtenues suivant 4 bandes de fréquences contigües. Pour chaque trame audio de chaque flux audio, nous disposons donc de 4 valeurs d'énergie, une par bande de fréquence. Pour chacune de ces mêmes bandes de fréquence, une valeur de tonalité [69] est calculée entre 0 et 1 : la valeur sera proche de 0 pour un signal de type bruit ou de 1 pour un signal de type tonal.

La seconde étape de notre algorithme (voir Figure 2.29) consiste à sélectionner dans la scène sonore les trames audibles à un instant donné en fonction de ces critères d'énergie.

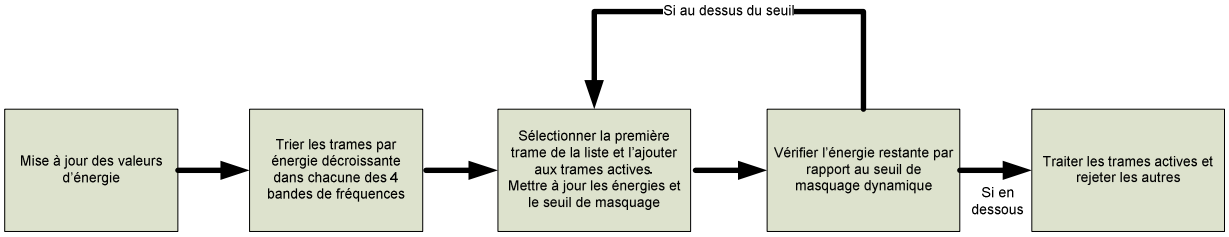

Figure 2.29 Vue schématique de la sélection des trames audibles à effectuer pour chaque bande de fréquences

Sur chaque bande de fréquences, les trames sont triées par énergie décroissante. A chaque fois, la première trame de la liste est choisie et un estimateur de l'énergie totale dans la scène sonore est décrémenté de l'énergie de cette trame. Symétriquement, un estimateur de l'énergie des sources audibles, dépendant de la tonalité, est incrémenté de la même énergie. Cette trame est ensuite enlevée de la liste. Le processus s'arrête lorsque l'énergie des sources audibles est supérieure à l'énergie totale restante. On dispose donc pour chaque bande de fréquences d'un ensemble de trames audibles et de trames non audibles. L'ensemble de cette procédure sera appelée par la suite *CullFrames*. On mixe finalement l'ensemble des trames qui sont audibles dans au moins une des 4 bandes de fréquences.

Par la suite un test a été mené pour valider cet algorithme. Un panel de 21 sujets a écouté 18 stimuli consistant en des mixtures de différents signaux sonores choisis dans quatre catégories : musique (différentes pistes instrumentales et vocales de deux morceaux de musique pop), parole (hommes et femmes parlant anglais, français, grec, allemand et polonais), bruits environnementaux et fragments de sons réverbérés. A partir de ces éléments, 18 mixtures ont été construites et l'estimation de masquage effectuée.

Le test était basé sur un protocole de type 2AFC (two-alternative forced-choice) [63] avec une référence cachée. Les sujets pouvaient écouter 3 signaux : la référence et deux stimuli tests A et B, ces derniers étant soit le signal "dégradé" (le résultat de l'algorithme), soit la référence cachée. Les sujets ayant été informés que l'un des deux signaux n'était pas identique à la référence devaient indiquer lequel.

Les résultats ont montré qu'il n'était pas possible de rejeter l'hypothèse "sur l'ensemble des sons, le taux d'identification est 50%", autrement dit, les sujets n'étaient pas capables de repérer la référence, autrement que par hasard.

Ces résultats ont montré que les résultats de l'algorithme étaient transparents à l'écoute. Cet algorithme va donc être implémenté dans le pont répliquant VoIP de France Télécom et testé dans le jeu *Flower Power Shooter* de "Virtools of Dassault Systèmes" disponible sur leur site internet pour test à l'URL suivante : http://www.virtools.com/applications/games-fps.asp.

2.5.2 Implémentation dans un pont répliquant VoIP

Dans cette section, nous décrivons l'intégration de l'algorithme dans une architecture avec pont répliquant nommé ComIP.

La Figure 2.30 décrit le fonctionnement de la chaîne de traitement audio, depuis la capture de la voix d'un participant P1 jusqu'à sa restitution chez un autre participant P4. Deux autres participants P2 et P3 sont aussi en conférence. Tous les clients fonctionnent de la même façon que P1 et P4.

Dans notre application finale (voir section 2.5.3), nous prenons en considération la position des joueurs dans le cadre d'un jeu. C'est la raison pour laquelle le client Virtools apparaît sur le schéma. Il est susceptible de fournir à l'algorithme de masquage une normalisation de l'énergie par les positions des joueurs (voir section suivante). Il faut bien comprendre que l'algorithme fonctionne aussi pour une restitution non spatialisée. Sans le client Virtools, nous aurions une conférence audio standard.

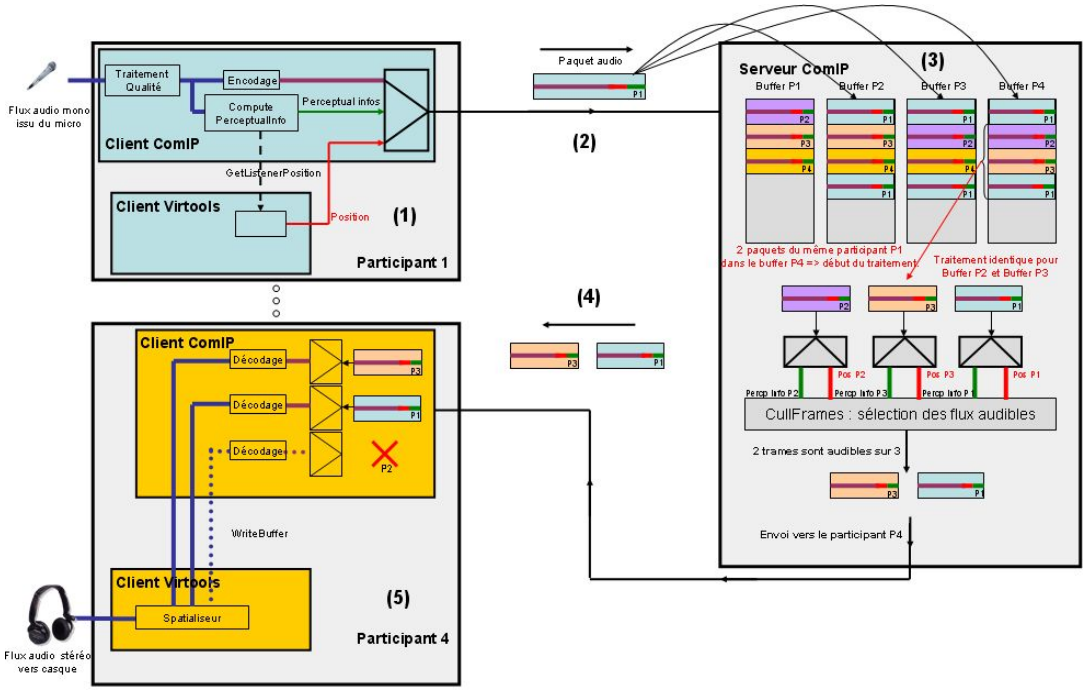

Figure 2.30 Schéma général de fonctionnement de la chaîne audio ComIP/Virtools

2.5.2.1 *Etapes (1) et (2)*

Au niveau du participant P1, Le son est enregistré par le microphone puis numérisé et découpé en trames de 60 ms (960 échantillons à 16 kHz). Pour chaque trame, la première étape de l'algorithme est effectuée (voir section 2.5.1) pour déterminer les 4 valeurs d'énergie/tonalité pour les bandes en Hz suivantes : 0-500, 500-2000, 2000-5000 et 5000-8000. Ensuite, les données audio sont encodées par un codeur propriétaire wideband France Télécom à 32 kbits/s et insérées dans un paquet IP avec les huit informations d'énergie/tonalité et avec les éventuelles informations de position du participant dans le jeu.

2.5.2.2 *Etape (3)*

Chaque paquet audio issu du terminal d'un participant (P1 par exemple) arrive au niveau du pont répliquant ComIP et est dupliqué, puis placé dans les buffers de sortie des autres participants. Si deux paquets audio issus d'un même participant (par ex P1) se trouvent dans un buffer d'un autre participant (par ex P4) au niveau du serveur, alors le traitement *CullFrames* (seconde étape de l'algorithme de la section 2.5.1) est appliqué. Ce traitement prend en entrée les données perceptives (énergie/tonalité) ainsi que les éventuelles positions disponibles des participants contenues dans les paquets audio disponibles dans le buffer de P4, sauf le dernier arrivé (en l'occurrence celui de P1). *CullFrames* permet de comparer relativement l'importance de signaux audio issus de différents participants et de ne sélectionner que ceux qui seront audibles.

En l'occurrence, dans l'exemple, seuls deux paquets sur trois sont audibles et seront envoyés au participant 4. Afin d'éviter des alternances trop fréquentes d'envoi ou non de paquets issus d'un même client pouvant créer des hachages, une fonction de lissage des résultats a été développée. Il faut en effet plusieurs résultats du même type ("envoi" / "non envoi") pour changer de statut : passage de "non envoi" à "envoi" / passage de "envoi" à "non envoi".

Après plusieurs essais, il a été choisi de basculer au bout de 3 décisions consécutives identiques.

En guise de remarque, un même paquet de P1 peut être envoyé vers P4 mais pas vers P3, par exemple. La fonction *CullFrames* travaille dans le buffer de chaque participant P_x indépendamment des buffers des autres clients.

2.5.2.3 Etape (4)

Les paquets audio sélectionnés par la fonction *CullFrames* sont envoyés au participant (ici P4).

2.5.2.4 Etape (5)

Au niveau du terminal du participant 4, les flux audio sont extraits des paquets puis décodés. Ils sont ensuite envoyés vers le client Virtools par la fonction WriteBuffer puis spatialisés par la couche openAL implémentée dans Virtools. Comme cela a déjà été précisé, le client Virtools pourrait ne pas être utilisé et le client audio ComIP ferait lui-même le mixage et éventuellement la spatialisation.

2.5.2.5 Les problèmes d'implémentation

L'ensemble a été implémenté avec succès, mais a montré quelques imperfections. En effet, lorsque certains participants parlent, on remarque bien au niveau du serveur audio qu'il ne transmet que les trames relatives à ces participants. Cependant lorsque personne ne parle, le serveur audio transmettait toutes les trames des participants puisque celles-ci avaient à peu près la même importance ou alors uniquement celles dont le bruit de fond était prédominant. Dans ce cas, il convenait évidemment de ne rien transmettre puisque les trames ne contenaient aucune information. Il est à noter que les tests ont été effectués avec du matériel audio standard (carte son, microphone, casque).

La première solution proposée a été de placer un seuil basé sur l'énergie au niveau du serveur pour supprimer ces trames de bruit mais cela aboutissait à placer un seuil fixe qui finalement supprimait aussi bien la parole que le bruit lorsque le niveau sonore de celle-ci était faible. De plus ce seuil fixe pouvait convenir dans certains cas, mais pas dans d'autres. Un choix a été fait, en concertation avec l'INRIA, de laisser ce seuil fixe assez bas mais de rehausser la qualité sonore des trames audio afin d'aider le serveur à ne pas envoyer les trames de silence/bruit et d'avoir un confort d'écoute accru.

Ces remarques ont donc abouti, dans un second temps, à l'implémentation d'une chaîne de qualité (visible au niveau du bloc émetteur Figure 2.30) composée :

- D'un filtre passe-haut de fréquence de coupure à 50 Hz, afin de supprimer d'éventuels bruits continus provenant par exemple d'un microphone. Le choix de 50 Hz comme fréquence de coupure se justifie par le fait qu'il n'y a pas vraiment d'information de parole sous cette fréquence. Par ailleurs, le codage audio effectué en bout de chaîne ignore les fréquences inférieures à 50 Hz.
- D'un réducteur de bruit.
- D'un contrôle automatique de gain permettant de niveler les signaux audio au niveau de chaque émetteur.

Ces résultats sont meilleurs, même si selon le matériel et le bruit ambiant, ces problèmes peuvent ressurgir. Différents réglages ont été effectués afin de rendre la spatialisation plus réaliste directement dans le jeu Virtools lui-même.

2.5.3 Evaluation de l'algorithme dans le cadre du jeu de Virtools

2.5.3.1 Intégration de l'architecture ComIP dans l'architecture Virtools

Afin d'évaluer l'algorithme CullFrames, cette architecture avec pont répliquant a été intégrée dans le jeu *Flower Power Shooter* (FPS) pour permettre la conférence audio spatialisée en tenant compte de la position des joueurs pour spatialiser le son. On se trouve dans le cadre de réalité virtuelle explicitée dans la partie 1.7.3.2. Ce jeu est un jeu multi-joueurs dont l'objectif est de toucher ses adversaires avec des boules de peinture (voir illustration Figure 2.31).

Figure 2.31 Illustration du jeu *Flower Power Shooter* de Virtools of Dassault Systèmes

FPS et ComIP sont tous deux des logiciels client-serveur. La solution développée relie les deux briques clients, les serveurs restant indépendants l'un de l'autre. La Figure **2.32** ci-dessous décrit schématiquement l'interaction entre les deux logiciels.

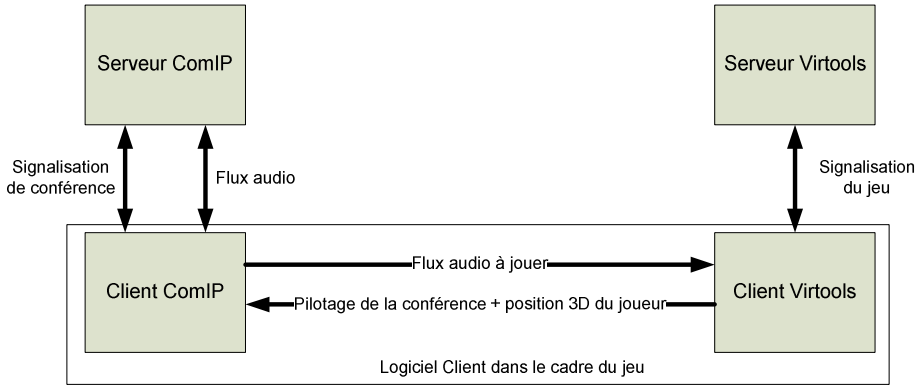

Figure 2.32 Schéma simplifié de l'interaction entre les composants ComIP et Virtools

La logique du jeu repose sur le client Virtools, qui pilote le client ComIP au niveau notamment de la création de la conférence audio associée à la partie de FPS, et des entrées/sorties des joueurs. Ainsi, sans synchronisation directe entre les serveurs, il est possible de garder une cohérence entre le jeu et la conférence audio.

Si l'on prend l'exemple de l'arrivée d'un nouveau joueur dans la partie FPS, elle se fait au niveau du serveur de jeu Virtools, qui avertit les autres clients Virtools de cet événement. Alors chaque client Virtools informe son client ComIP associé de l'arrivée d'un nouveau participant et donc d'un nouveau flux audio à gérer.

Une première étape de l'intégration de ComIP dans l'environnement Virtools a consisté à développer une couche d'interface permettant au client Virtools de piloter la conférence audio. Chaque fonction de l'API ComIP a été encapsulée dans un formalisme Virtools appelé Building Block. Ainsi on retrouve l'intégralité des fonctionnalités de ComIP accessible dans l'environnement de développement Virtools.

Au terme de cette première étape, les fonctions de conférence audio, sans rendu 3D, étaient pilotables depuis Virtools. On avait donc à ce stade le jeu multi-joueurs FPS intégrant un chat audio monophonique.

Le logiciel ComIP standard prend entièrement en charge la gestion de la carte son, c'est-à-dire la capture de la voix des participants et le rendu de la voix des distants. Dans le cadre du démonstrateur, il a été décidé d'utiliser la brique rendu audio 3D présente dans le client Virtools pour le jeu. Ainsi il a été nécessaire de désactiver la fonction de rendu audio de ComIP et d'établir une liaison entre ComIP et Virtools pour la transmission des trames audio reçues des autres participants. Le client Virtools utilise la brique openAL pour effectuer la spatialisation.

En ce qui concerne l'interface nécessaire au bon déroulement d'une partie de jeu intégrant les principes décrits ci-dessus, le travail s'est porté en grande partie sur la gestion de la connexion ComIP (adresse du serveur de jeu, etc.) et l'intégration de ComIP au sein du jeu.

2.5.3.2 Tests effectués

Des tests ont été effectués au sein de France Télécom de 3 à 5 participants, tous équipés d'un microcasque (Sennheiser HMD 280 pro). Un ordinateur était utilisé en tant que serveur ComIP et un autre en tant que serveur de jeu. Les testeurs ont joué à *Flower Power Shooter*, tout en bénéficiant d'une conférence audio spatialisée grâce à des connexions ethernet ou Wi-Fi. Deux modes ont été testés : un où personne ne parle et un autre où les joueurs jouent normalement en se parlant les uns les autres.

Pour expliciter les résultats, il faut savoir que pour une trame reçue au niveau du pont répliquant, elle est, sans optimisation, envoyée N-1 fois avec N le nombre de participants à la conférence. Dans les Figure 2.33 et Figure 2.34 ci-dessous, elles sont appelées "trames qui devraient être envoyées". Dans une conférence standard avec pont répliquant, une trame reçue par un participant est systématiquement envoyée à tous les autres. Avec le démonstrateur intégrant les optimisations de simplification de la scène sonore, une trame peut être envoyée ou non suivant les trames avec lesquelles elle est en concurrence. La différence entre les "trames acceptées par l'algorithme" et "les trames réellement envoyées" est due à l'impact de la fonction de lissage explicitée dans les paragraphes précédents.

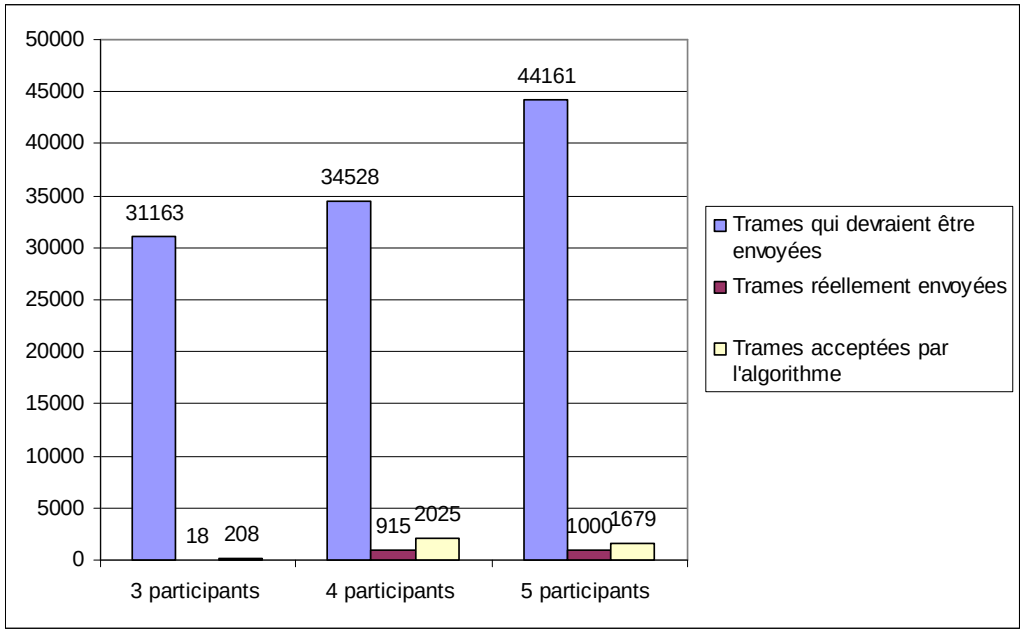

Figure 2.33 Statistiques obtenues en période de silence pour différents nombres de participants

En période de silence, Figure 2.33, l'algorithme supprime quasiment toutes les trames et aboutit à une baisse très significative de la bande passante. Le pont répliquant se comporte comme un système DTX.

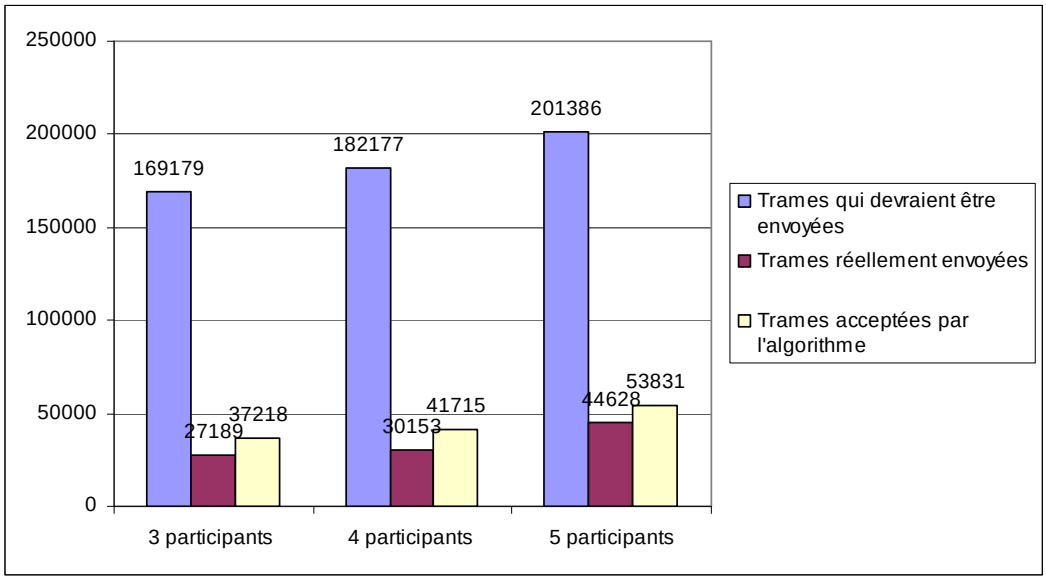

Figure 2.34 Statistiques obtenues en période de silence/parole/paroles simultanées pour différents nombres de participants

En période de silence, parole, et paroles simultanées, Figure 2.34, la réduction de la bande passante est encore significative. Nous pouvons noter d'une part l'influence de la fonction de lissage et d'autre part que le pourcentage de trames acceptées augmente du fait de l'augmentation des possibilités de dialogue. Ces résultats dépendent bien sûr de la volonté des joueurs à parler, du réseau, de la qualité du matériel, de la fonction de lissage et de la position des joueurs dans le jeu. L'impact sur le temps de calcul côté client est difficile à chiffrer sur une petite conférence mais on économise un traitement de décodage et de spatialisation pour chaque trame non envoyée. La qualité de la communication reste bonne malgré la forte diminution des données envoyées par le pont.

2.5.4 Discussion et perspectives

L'utilisation de cet algorithme dépend de nombreux facteurs dans un contexte temps-réel : le matériel audio, la bande passante disponible dans le réseau, les techniques de suppression de bruit, la fonction de lissage au niveau du pont répliquant, l'algorithme de masquage, le niveau de prise de son de chaque locuteur, la localisation du joueur dans le contexte de réalité virtuelle, … Tous ces facteurs peuvent dégrader la qualité audio, mais les résultats en gain de bande passante sont très intéressants.

Dans le contexte du jeu, du fait que l'on tienne compte des positions des locuteurs, l'intelligibilité peut être compromise, car on a tenu compte de l'atténuation du son avec la distance croissante. Cela rend le jeu plus réaliste mais peut gêner certains joueurs. Il serait aussi intéressant de pouvoir tenir compte des obstacles entre deux joueurs pouvant aussi jouer sur la perception sonore.

Comme améliorations techniques, nous avions testé un bloc de VAD pilotant un bloc DTX sur les clients VoIP mais cela a généré des soucis de niveau sonore dans le client Virtools. Cette solution aurait été intéressante pour aider le pont répliquant à sélectionner les trames audibles et pour diminuer les problèmes liés au bruit de fond.

Le coût en termes de débit pour inclure les positions et les informations d'énergie et de tonalité est de 44 octets (4 octets pour une donnée de type float multipliés par 11 valeurs : 3 positions + 8 informations d'énergie/tonalité) à comparer avec les 240 octets compressés du contenu audio. Cela est évidemment important, mais nous n'avons pas cherché à compresser ces informations, ni même à effectuer les calculs des informations d'énergie/tonalité dans le pont

En supposant une bande passante fixe du pont répliquant vers les terminaux, nous pourrions utiliser un critère d'importance calculé à partir des informations d'énergie pour attribuer dynamiquement plus ou moins de bande passante à certains terminaux [52,53,92]. Cela permettrait d'utiliser des codeurs scalables dont on peut tronquer une partie du contenu compressé pour faire baisser le débit mais inévitablement la qualité suivant l'importance donnée au signal. Un codeur scalable comme le G.729.1 [46] se différencie des codeurs standards par le fait que les données codées se présentent sous forme de couches empilées. En incluant plus ou moins de couches suivant les contraintes dues aux réseaux, on fait ainsi varier le débit et par conséquence la qualité selon l'importance de chaque source. On attribue aux sources les plus importantes le maximum de débit pour une qualité supérieure.

Dans notre application, nous n'avons pas traité les sons locaux du jeu qui pourraient masquer ou être masqués par les paroles des différents locuteurs. Des tests effectués dans [93] ont montré que l'algorithme restait performant dans ces conditions.

Actuellement, la décision de masquage dépend uniquement des critères d'énergie et de tonalité. Une amélioration pourrait être d'inclure des critères plus perceptifs introduits dans [51]. Une version prenant en compte les effets audio 3D à la manière de [93] pourrait être aussi testée.

Cette solution de pont répliquant avec algorithme de masquage intégré est à rapprocher des solutions présentées en 2.3.1.1 avec une solution plus élaborée. Les résultats obtenus sont globalement bons et le principe est à retenir.

2.6 Conclusions et perspectives

Tout d'abord, il est rappelé que les hypothèses de départ sont que les terminaux ont une prise de son monophonique et possèdent des dispositifs d'écoute stéréophoniques adaptés à la spatialisation. De plus, les blocs d'encodage ou de décodage sont des blocs de compression que l'on suppose adaptés au transport des flux monophoniques ou spatialisés.

Ce chapitre a montré la faisabilité au niveau média du transport de flux spatialisés sur des réseaux IP standard. Différentes solutions sont possibles, qui ne nécessitent pas la même contribution en termes de recherche, de développement et de mise en application. Il est difficile de faire un dimensionnement réel des coûts de chaque solution, du fait de la difficulté de pondérer ceux-ci en termes de charges CPU ou latence. Il convient de laisser ou non certains blocs suivant la qualité, le niveau de complexité et de latence que l'on désire.

Cependant il nous semble réaliste de penser que des méthodes telles que l'utilisation d'un pont mixeur spatialiseur sans commutation (Voir la section 2.2.1) est implémentable à court terme. Nous baserons donc nos tests du chapitre **4** concernant la spatialisation sur pont spatialiseur mixeur sur cette configuration. Nous testerons la configuration simplifiée suivante, Figure 2.35, sans les blocs de traitements pouvant jouer sur la qualité, cela afin de bien identifier l'impact du codage audio et des pertes de trames :

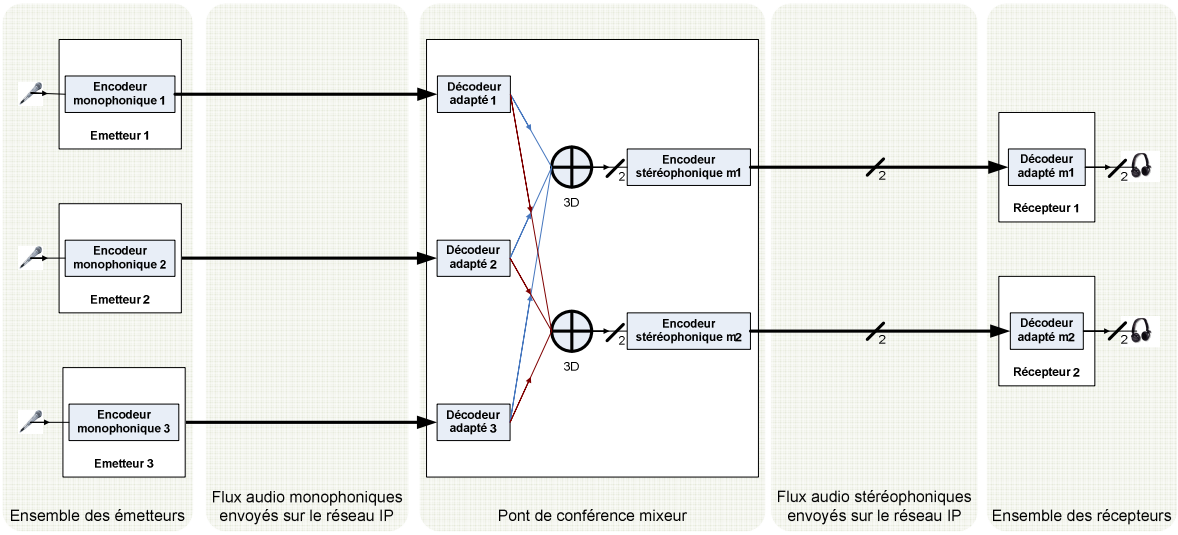

Figure 2.35 Illustration de la conférence audio centralisée avec l'utilisation d'un pont spatialiseur mixeur (les émetteurs et les récepteurs sont évidemment généralement confondus dans une conférence)

Il semble difficile d'inclure conjointement la spatialisation et la commutation (Voir la section 2.2.2) étant donné que cela implique à la fois une gestion de la signalisation des appels VoIP et une coordination de la spatialisation entre un terminal et le pont. Cela peut de plus poser des soucis d'écoute dans le sens où il est nécessaire d'assurer une qualité sans coupure apparente pour un auditeur. Il semble plus sage à court terme de choisir entre spatialisation et commutation, plutôt que de chercher à associer les deux. Cette association serait intéressante à tester dans l'avenir.

L'utilisation de pont répliquant pose le problème de la charge supportable par les terminaux suivant les traitements qu'on souhaite réaliser pour obtenir une qualité intéressante. De plus, il n'existe pas pour l'instant de pont répliquant sur le marché et les terminaux ne peuvent actuellement ouvrir plus de un ou deux canaux à la fois. Cette solution reste cependant très efficace dans le cas d'un faible nombre de participants lorsque les terminaux de ces derniers sont capables de supporter les traitements audio nécessaires. Une solution pour réduire la bande passante, problème majeur des ponts répliquants, a été testée et a montré son intérêt. Elle nécessite cependant des précautions d'emploi du fait des nombreux paramètres entrant en jeu.

Ces réflexions sur les ponts mixeur et répliquant ont amené à un dépôt de brevet concernant un pont mixte : à la fois pont répliquant et pont mixeur, pour bénéficier des avantages de chacun. Une mise en application de ce dernier serait intéressante à tester.

Tout comme la configuration centralisée avec pont répliquant, les configurations distribuées sont également limitées par les capacités des terminaux mais restent implémentables.

Pour nos tests du chapitre 4 avec spatialisation sur terminal, nous nous sommes basés sur une configuration simplifiée illustrée Figure 2.36. Nous retrouvons bien toutes les possibilités des configurations distribuées et centralisée avec pont répliquant.

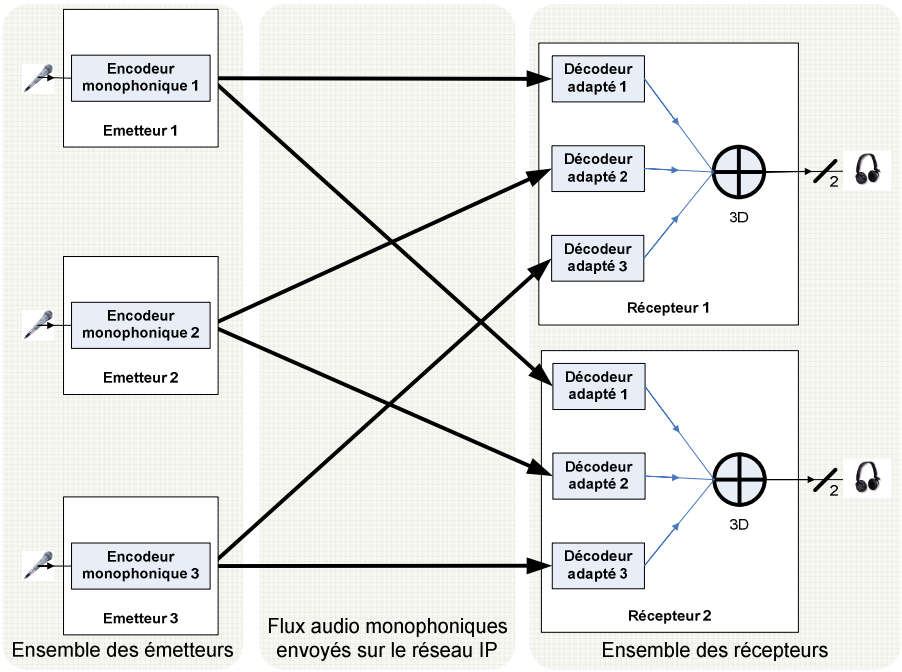

Figure 2.36 Illustration de la conférence audio distribuée multi-unicast (les émetteurs et les récepteurs sont évidemment généralement confondus dans une conférence)

3. Extensions nécessaires pour la gestion de la conférence spatialisée sur IP

Le chapitre 2 a proposé différentes architectures permettant d'allier la conférence audio en voix sur IP, la spatialisation et les traitements d'améliorations connus. Il convient à présent de déterminer les extensions nécessaires à la conférence audio standard en VoIP pour gérer et commander cette spatialisation.

L'objectif de ce chapitre est d'établir au niveau protocolaire des solutions adaptées à la gestion et au contrôle de la spatialisation. Ces solutions devront tenir compte de l'existant en SIP pour permettre l'interopérabilité avec d'autres terminaux. Une autre contrainte est de proposer des réponses qui tiennent compte de l'esprit d'utilisation du protocole SIP, c'est-à-dire en respectant sa logique de conception.

Nous chercherons dans une première partie (voir section 3.1) à verbaliser les différents besoins en termes de positionnement des locuteurs dans la conférence audio spatialisée. Nous déterminerons les différentes options possibles et nous proposerons des solutions adaptées aux besoins.

Dans une seconde partie (voir section 3.2), nous verrons les extensions spécifiques au protocole SIP pour intégrer la conférence audio spatialisée.

3.1 Positionnement des locuteurs de la conférence audio

Nous présentons dans cette section une réflexion concernant ce que l'on entend par positionnement. Nous nous restreindrons dans la majorité de nos exemples à la spatialisation 2D, c'est-à-dire dans le plan horizontal pour simplifier les représentations.

3.1.1 Qu'est-ce que le positionnement en conférence audio spatialisée ?

Le positionnement des locuteurs consiste à placer des locuteurs dans l'espace autour d'un utilisateur. Dans un premier temps, cela consiste à suivre une gestion de positionnement pour attribuer à chaque locuteur une position exprimée par exemple en coordonnées cartésiennes, sphériques ou polaires. Dans un second temps, on utilise le couple d'HRTFs correspondant à cette position pour spatialiser le locuteur.

Deux gestions sont possibles pour le positionnement des locuteurs :

- La gestion de positionnement dite automatique de la spatialisation pour laquelle l'utilisateur n'entre pas en jeu pour la disposition des différentes sources sonores.
- La gestion de positionnement dite manuelle de la spatialisation pour laquelle l'utilisateur commande lui-même la disposition des différentes sources sonores.

3.1.2 La gestion de positionnement dite automatique

3.1.2.1 Définition des spécifications de la gestion de positionnement dite automatique

La gestion de positionnement automatique est la configuration la plus simple à gérer dans notre cas, puisqu'elle est uniquement intégrée dans le logiciel gérant le pont de conférence, dans un softphone IP, dans un IP Phone, etc. Qu'elle soit appliquée sur un terminal ou sur un pont de conférence mixeur, l'utilisateur ne commande rien, ce qui implique qu'il n'y a besoin d'aucune interface de contrôle de la spatialisation, ni d'éventuel dialogue entre le pont mixeur et le terminal pour contrôler la spatialisation. Par contre, pour aider à l'identification et à mieux ressentir la position des sources sonores, l'utilisateur peut avoir un retour visuel représentant dans l'espace les positions des locuteurs ([55,56]). Cela sera traité dans la section 3.1.3.1.1.

Pour un utilisateur donné, la gestion de positionnement automatique doit obligatoirement tenir compte de données telles que :

- L'arrivée ou le départ d'un autre participant dans la conférence. Les participants d'une conférence n'arrivent ou ne partent jamais tous en même temps. Par exemple, leur arrivée dépend généralement du moment où ils appellent le pont de conférence ou lorsqu'ils répondent à l'invitation de ce dernier. Le procédé doit donc savoir gérer au fur et à mesure les arrivées et les départs.
- Le moment d'arrivée des participants (avant ou après l'utilisateur) dans la conférence. Un utilisateur donné peut être le premier participant à entrer en conférence et dans ce cas, le procédé doit gérer les arrivées progressives des autres participants. A l'inverse pour le dernier utilisateur entrant, le procédé doit placer en une seule fois tous les participants déjà présents. Dans les deux cas, le procédé suit la stratégie de placement.
- La stratégie de placement choisie pour l'arrivée ou le départ des locuteurs. Elle dépend du contexte (conférence audio en entreprise, jeu vidéo, etc.), de la dimension de la spatialisation (2D ou 3D), de la façon dont on souhaite placer les locuteurs (dans le cas d'une conférence audio d'entreprise : cercle, demi-cercle, ligne, etc.). Par exemple en tenant compte de l'arrivée d'un nouveau locuteur E dans une conférence d'entreprise avec une représentation des locuteurs en demi-cercle en face d'un utilisateur, différents modes sont possibles et sont illustrés sur la Figure 3.1, ci-dessous. Ces modes sont adaptables aussi pour le départ d'un participant. Nous en profitons pour introduire la notion d'ordre qui nous permet de savoir quel locuteur est à placer à côté de quels autres dans l'espace. Cet ordre est par exemple un ordre de gauche à droite comme dans l'exemple de la Figure 3.1.
    - Mode statique : les positions angulaires sont réajustées sans changer l'ordre suite à l'arrivée d'un participant.
    - Mode position figée : L'ordre et les positions des autres participants ne changent pas suite à l'arrivée d'un participant.
    - Mode dynamique : on regénère la scène sonore (mise à jour de l'ordre et des positions) en tenant compte de l'arrivée d'un participant. Il faut préciser que ce mode ne nous semble pas adapté à une conférence audio d'entreprise, car un utilisateur risquerait d'être perturbé par les déplacements des locuteurs.

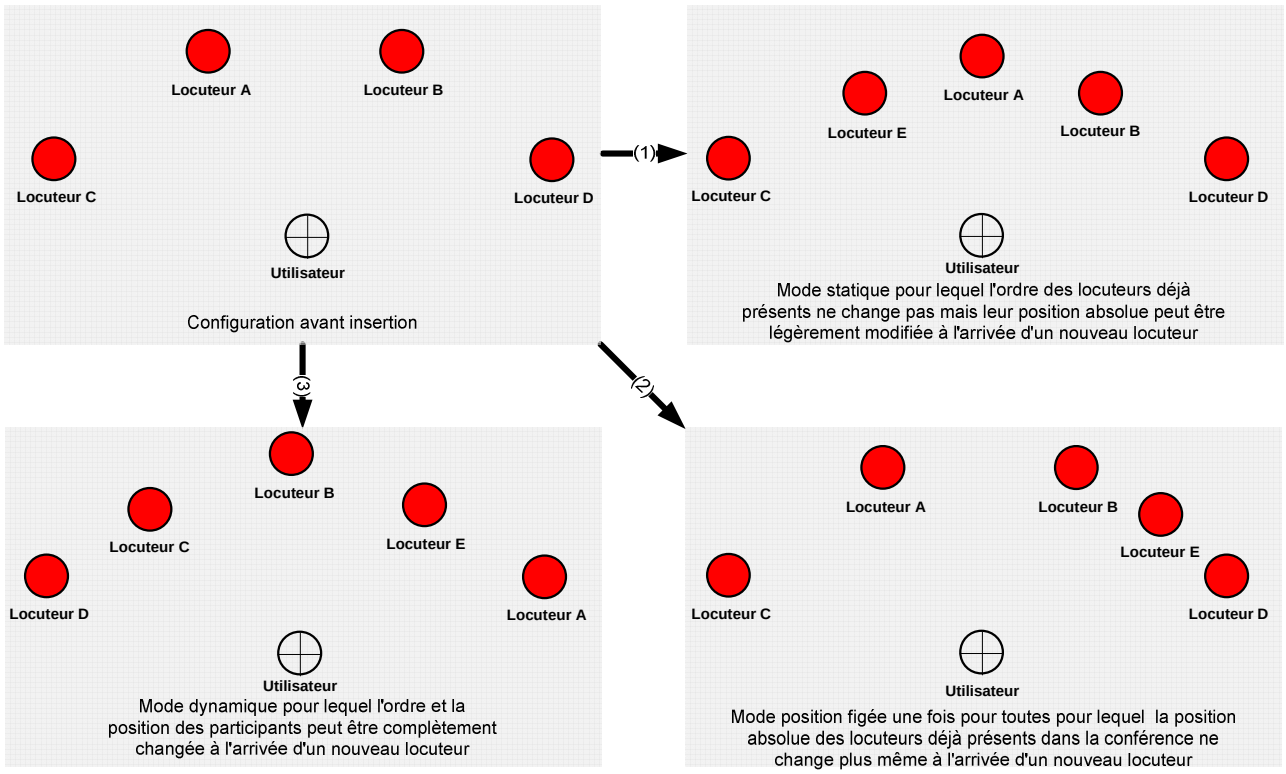

Figure 3.1 : Exemple de différentes stratégies d'insertion d'un nouveau participant E

Toute gestion de positionnement automatique doit avoir une politique de placement. Voici quelques exemples de politique de placement devant prendre en compte les données présentées ci-dessus et basées sur :

- Une politique de placement aléatoire. Les participants déjà présents à l'arrivée de l'utilisateur sont placés aléatoirement. Pour les participants arrivant par la suite, ils peuvent être aussi placés aléatoirement ou par défaut placés par exemple toujours à la gauche du participant le plus à la gauche.
- La cohérence par rapport à une table virtuelle. Les participants sont virtuellement placés autour d'une table. Le procédé fait donc en sorte que pour toute combinaison de deux participants chacun derrière un terminal, si l'un entend l'autre à gauche (respectivement droite), alors l'autre l'entendra à droite (respectivement gauche).
- Des paramètres issus des signaux des participants. Nous présenterons dans la section suivante une méthode innovante pour placer les locuteurs.
- La position dans un univers virtuel, par exemple dans le contexte d'un jeu. Dans ce cas, la politique de placement des locuteurs est simple puisqu'elle tient compte de la position relative de ces locuteurs vis-à-vis de l'utilisateur.
- Le statut des participants. Certains participants (organisateur, présentateur principal, etc.) peuvent être placés de préférence en face de l'utilisateur par exemple.
- etc.

3.1.2.2 Proposition d'une politique de placement basée sur des paramètres issus des signaux des participants

Le point de départ de nos travaux est l'article [54]. Ce document montre que la facilité d'identifier un locuteur est augmentée par un positionnement des voix par le sujet comparé à un positionnement aléatoire. Il montre également que les erreurs d'identification entre deux voix similaires sont moindres lorsque le sujet a eu la possibilité de bien les séparer.

Le positionnement des locuteurs dans une conférence audio en entreprise est un problème critique qui trouve donc une solution lorsque l'auditeur place lui-même les voix dans l'espace

sonore, plutôt que de se contenter d'un positionnement aléatoire dans lequel des voix jugées similaires peuvent se retrouver proches, et mener ainsi à une dégradation de l'intelligibilité, du confort d'écoute, de la compréhension ou de la reconnaissance des locuteurs.

Cette solution a un inconvénient en pratique puisque cela implique la présence d'une interface de positionnement et des manipulations de la part de l'auditeur. De plus, pour le cas de la conférence audio centralisée, la possibilité laissée à un utilisateur de pouvoir placer virtuellement les auditeurs distants implique un dialogue entre le pont spatialiseur et le terminal. Cela amène une complexification des dialogues entre les différentes entités puisqu'une décision de l'utilisateur d'un terminal concernant la position d'un de ses locuteurs distants doit déclencher un traitement sur le pont pour adapter la spatialisation.

Notre étude propose un système de positionnement automatique des voix dans l'espace audio virtuel, reposant sur un système d'optimisation de distances entre les voix. Ces travaux réalisés avec Gregory Pallone de France Télécom ont fait l'objet d'un dépôt de brevet de numéro Fr 07 04712.

Notre positionnement automatique des voix dans l'espace virtuel audio 3D dans le contexte d'une audioconférence repose sur 3 points :

- L'établissement d'une matrice de distances entre les voix.
- L'établissement d'une méthode de positionnement par maximisation de distances afin de positionner de manière optimale chacune des voix sous des contraintes précises.
- L'intégration de ces deux premiers points dans un terminal ou un pont de conférence audio spatialisée.

Nous détaillons dans la suite chacun de ces points successivement.

3.1.2.2.1 L'établissement d'une matrice de distances entre les voix

L'établissement d'une matrice de distances D entre les voix a pour objectif de fournir des valeurs représentatives de la distance (au sens euclidien) qu'il existe entre chacune des voix (deux voix qui se ressemblent auront une distance faible, et deux voix très dissemblantes auront une distance élevée). Il est donc nécessaire de se baser sur des critères perceptifs afin d'établir cette matrice.

Nous proposons ici une façon d'établir cette matrice de distances entre voix. D'autres possibilités sont présentées dans le brevet.

La matrice de distances peut être obtenue de la manière suivante. On réalise un banc de K filtres (par exemple K=24) nommés TimeFilters(n), dont la largeur de bande fréquentielle se calque sur l'échelle des Bark [28], afin de s'approcher au mieux de la perception de l'oreille. On passe d'une échelle linéaire des fréquences à une échelle en Bark grâce à la relation suivante :

$$Freq2bark = \frac{26.81 * Freqscale}{Freqscale + 1960} - 0.53 \qquad (3\text{-}1)$$

Pour information, la largeur des filtres est relativement constante à basses fréquences, et proportionnelle à la fréquence à hautes fréquences. Les fréquences centrales sont par exemple données par : 1.0e+004 * (0.0020 0.0122 0.0212 0.0309 0.0416 0.0534 0.0663 0.0807 0.0968 0.1148 0.1353 0.1585 0.1853 0.2166 0.2533 0.2972 0.3506 0.4170 0.5018 0.6139 0.7687 0.9968 1.3662 1.9996).

On calcule, sur une portion du signal vocal échantillonné sig de chacun des locuteurs x, les sous-bandes fréquentielles subband(k) en filtrant le signal vocal par les réponses impulsionnelles obtenues :

$$subband(k) = (sig * TimeFilters)(k) \quad \text{où * désigne le produit de convolution} \qquad (3\text{-}2)$$

Dans chacun des signaux en sous-bandes, on calcule l'énergie à court terme STE de la manière suivante : on applique au signal en sous-bandes la fenêtre win (de type Hanning par exemple) centrée en h et on calcule l'énergie de ce signal fenêtré, puis on réitère l'opération, en déplaçant la fenêtre win (de 4 ms par exemple), jusqu'à épuiser la portion de signal vocal.

$$\mathrm{STE_h(k)} = \sum (\mathrm{subband(k).win_h})^2 \qquad (3\text{-}3)$$

L'énergie à court terme est ensuite moyennée afin d'obtenir un coefficient énergétique par sous-bande et par locuteur, puis normalisée par l'énergie totale (énergie de la portion vocale complète toutes bandes confondues) afin de s'affranchir des différences de niveaux des différents locuteurs.

$$\mathrm{coeffEnergy(k)} = \frac{\overline{STE_h(k)}}{\sum sig^2} \qquad (3\text{-}4)$$

On établit enfin la matrice de distances en calculant la distance quadratique des coefficients énergétiques entre chaque locuteur :

$$D(x,y) = \sqrt{\sum_{k=1}^{K} \left(\mathrm{coeffEnergy}_x(k) - \mathrm{coeffEnergy}_y(k)\right)^2} \qquad (3\text{-}5)$$

Cette matrice est symétrique avec des valeurs nulles sur la diagonale. On interprète cette matrice de la manière suivante : plus la valeur D(x,y) est élevée, plus la dissemblance entre les voix x et y est grande.

Nous nous trouvons à présent à l'entrée du second bloc sur la Figure 3.2 :

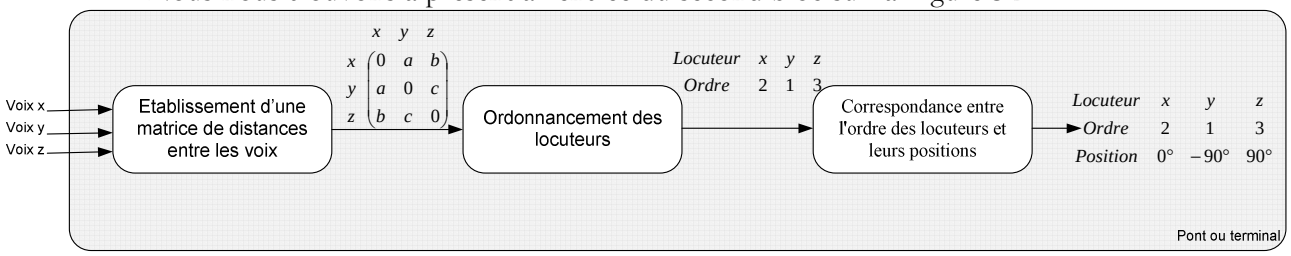

Figure 3.2 : Enchainement des étapes de notre politique de placement

3.1.2.2.2 L'établissement d'une méthode de positionnement par maximisation de distances

L'établissement d'une méthode de positionnement par maximisation de distances vocales a pour objectif :

- D'ordonner les locuteurs, par exemple de gauche à droite (voir la Figure 3.2).
- De réaliser une correspondance entre l'ordre des locuteurs et leurs positions dans l'espace virtuel audio 3D, par exemple d'associer l'ordre 1 à la position -90° dans le cas d'une représentation en demi-cercle (voir la Figure 3.2).

Tout d'abord, l'ordre des locuteurs est à déterminer à partir des distances vocales contenues dans la matrice. Il est possible de modéliser le problème d'ordonnancement de la manière suivante en posant :

- N, le nombre de locuteurs déjà présents dans la conférence, et L = {L_1, L_2, …, L_N, L_{N+1}} l'ensemble ordonné des N locuteurs auquel on ajoute le nouveau locuteur L_{N+1}.
- *T*, l'ensemble des permutations possibles de L, et τ_i une permutation donnée de L tel que :

$$T = \{\tau_i\}_{i \in C}$$

- *C*, l'ensemble des choix de permutation possibles. En se basant sur les modes vus pour la stratégie de placement dans la section 3.1.2.1, il y en a :

- (N+1)! en mode dynamique.
- N+1 en mode statique ou position figée.

Pour chaque permutation, on utilise une variable de décision M_i nous permettant d'évaluer la pertinence de l'insertion du nouveau participant à l'emplacement décrit par i. Cette variable M_i dépend évidemment de la matrice de distances vocales établie précédemment. On détermine ensuite l'indice J tel que $M_J=\max(M_i)_{i\in C}$, ce qui nous donne l'indice i aboutissant au meilleur résultat d'ordonnancement. On en déduit τ_J la permutation associée et donc l'emplacement du nouveau locuteur.

Exemple dans un mode dynamique (chaque locuteur peut changer de position) pour N=4 : soit L = {L_1, L_2, L_3, L_4, L_5}. Si on détermine que τ_J est la permutation la plus adéquate et qu'elle est définie comme suit :

$$\tau_J(L_1)=L_2,\ \tau_J(L_2)=L_4,\ \tau_J(L_3)=L_1,\ \tau_J(L_4)=L_5,\ \tau_J(L_5)=L_3$$

Alors on obtient l'ordre suivant : (L_2, L_4, L_1, L_5, L_3) qui est optimal au sens de notre variable de décision.

Donnons à présent un exemple de variable de décision dans le cas du mode statique avec une disposition en cercle des locuteurs autour de l'utilisateur. Un exemple d'ordre est illustré ci-dessous et on cherche à introduire un nouveau participant.

Figure 3.3 : Illustration d'un ordre avec des locuteurs (x, y, z, ..) se trouvant à des positions quelconques de 1 à N

Un critère de maximisation de la voix doit être utilisé pour tester quelle position d'insertion est la meilleure. On essaie donc toutes les possibilités et on garde celle qui permet de placer de manière optimale un locuteur lorsque l'on veut conserver l'ordre précédemment établi.

Pour mettre à jour l'ordre des participants, on travaille dans l'espace des distances vocales caractérisé par la matrice D des distances vocales établie précédemment. Dans le cas illustré ci-dessous, on cherche à estimer une valeur caractéristique de l'insertion d'un participant entre deux autres.

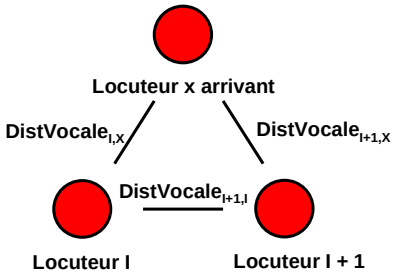

Figure 3.4 : Configuration de test

Cette valeur nous est donnée par la variable de décision M_i représentative de la pertinence d'insérer le nouveau locuteur à un endroit. M est donné par le rapport entre 2 valeurs :

$$M_i=\frac{\mu}{\sigma} \tag{3-6}$$

avec μ la moyenne des distances entre deux locuteurs I et I+1 :

$$\mu=\frac{D_{I+1,X}+D_{I,X}}{2.D_{I,I+1}} \tag{3-7}$$

et σ l'écart-type donné par :

$$\sigma = \sqrt{\left(\frac{D_{I,X}}{D_{I,I+1}} - \mu\right)^2 + \left(\frac{D_{I+1,X}}{D_{I,I+1}} - \mu\right)^2} \qquad (3\text{-}8)$$

Le choix de normaliser μ par $D_{I,I+1}$ permet de privilégier l'insertion du nouveau locuteur entre 2 locuteurs relativement proches au sens de la distance vocale et, de fait dans notre exemple, voisins dans l'espace audio spatialisé.

La normalisation de la moyenne par un écart-type standard consiste à privilégier l'insertion d'un nouveau locuteur distant de manière équitable des deux locuteurs déjà placés.

On obtient donc N valeurs de M, avec N le nombre de participants déjà présents. On choisit le maximum parmi ces valeurs et l'indice résultant nous donne ainsi entre quels locuteurs déjà présents il convient d'insérer le nouveau locuteur. Un nouvel ordre est donc établi et une correspondance entre l'ordre des locuteurs et leurs positions dans l'espace virtuel audio 3D est à réaliser.

Une fois l'ordre établi, la méthode effectue une correspondance entre l'ordre obtenu et une représentation audio spatialisée en fournissant des positions spatiales utilisables par un système de spatialisation. On peut imaginer un positionnement automatique par exemple dans un plan tout autour de l'auditeur avec la formule définie ci-dessous qui met à jour la position angulaire de chacun des locuteurs suite à l'arrivée du nouveau participant :

$$P(\tau_J(L_i)) = \left((i-1) * \frac{2\pi}{N+1}\right), i \in [1, N+1] \qquad (3\text{-}9)$$

Il est aussi possible de définir des fonctions plus subtiles qui prennent la résolution spatiale de l'oreille humaine (plus de locuteurs en frontal car meilleure discrimination [14]). De plus, on s'est ici restreint à l'espace 2D, mais il est évidemment possible de généraliser à l'espace 3D. On peut aussi tenir compte des notions comme l'éloignement à l'auditeur.

Pour le départ d'un participant, il est possible suivant la stratégie de placement choisie d'appliquer n'importe lequel des 3 modes :

- Mode statique : les positions angulaires sont réajustées sans changer l'ordre.
- Mode dynamique : on régénère la scène sonore (mise à jour de l'ordre et des positions) en tenant compte du départ d'un participant.
- Mode position figée : L'ordre et les positions des autres participants ne changent pas suite au départ d'un participant.

3.1.2.2.3 L'intégration de ces deux premiers principes dans un terminal ou un pont de conférence audio spatialisée de type mixeur

Les deux principes énoncés ci-dessus sont à intégrer dans un pont ou dans un terminal. Concernant le fonctionnement avec un pont, les éléments suivants établissent son fonctionnement.

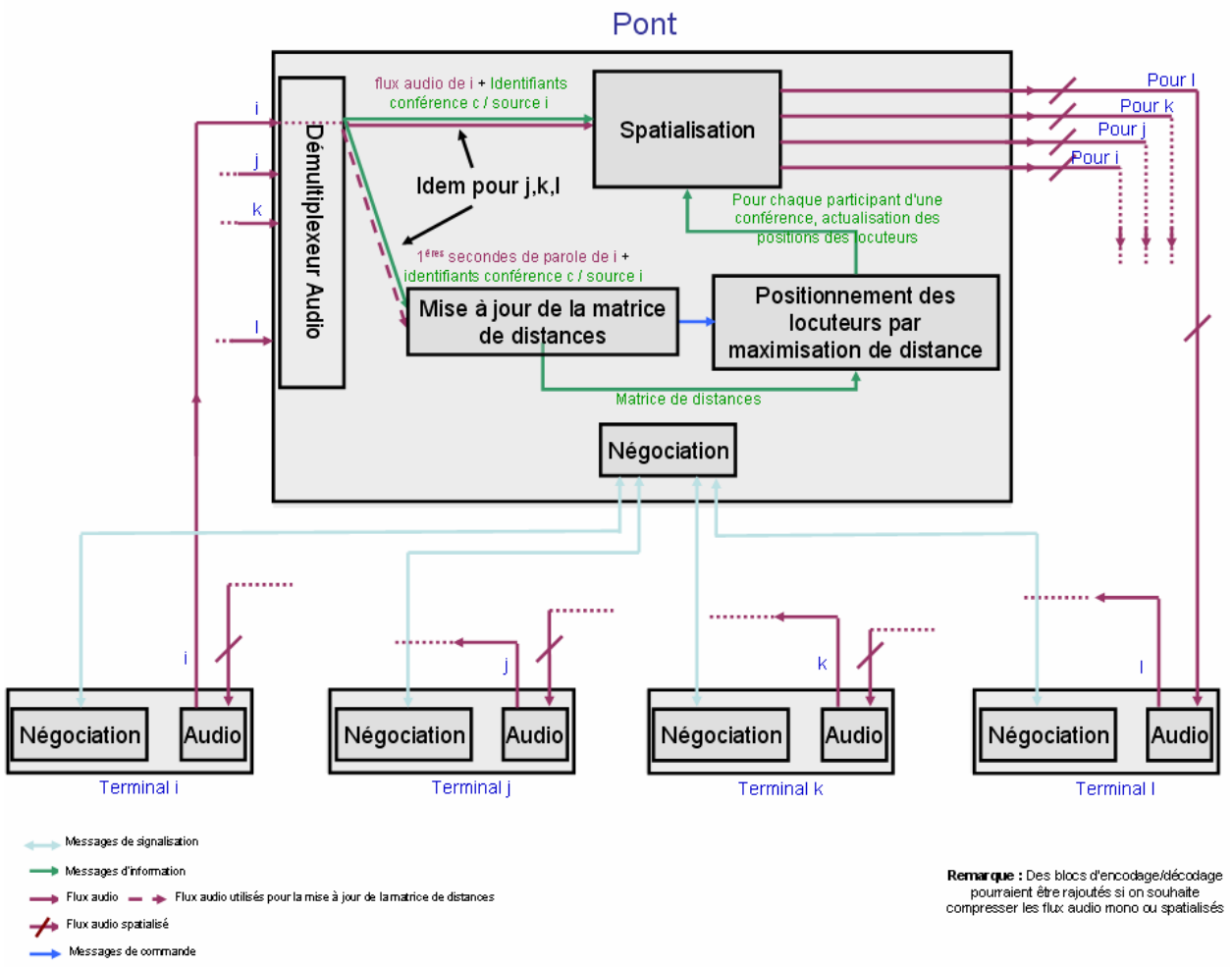

Figure 3.5 : Schéma du pont intégrant le positionnement automatique des locuteurs

Comme l'illustre la Figure 3.5, le pont peut être vu comme une entité composée de plusieurs modules :

- Un module *Négociation*, dont le rôle est d'échanger les messages de signalisation avec les terminaux. Il permet de créer des conférences, de négocier avec les terminaux et de leur donner leur identifiant.
- Un bloc *Démultiplexeur Audio* qui envoie les premières secondes de parole d'un participant vers le bloc de mise à jour de la matrice de distances et éventuellement en parallèle vers le bloc de spatialisation.
- Un bloc de *Mise à jour de la matrice de distances* qui mesure l'écart entre deux voix et donne un indice de dissemblance entre celles-ci. Il prend en entrée quelques secondes de voix d'un participant, en extrait des paramètres d'identification vocale et met à jour une matrice de distance vocale.
- Un bloc de *Positionnement des locuteurs par maximisation de distances* qui fournit au bloc de Spatialisation les positions de tous les locuteurs distants et cela pour chaque auditeur. Il prend en entrée la matrice de distance vocale pour établir ces positionnements propres à chaque participant puisque deux auditeurs n'ont pas la même scène sonore, n'ayant pas les mêmes locuteurs distants.
- Un bloc de *Spatialisation* qui spatialise les flux audio issus de chaque participant pour générer autant de scènes sonores qu'il y a de participants. La spatialisation se base sur les informations qu'elle reçoit du bloc Positionnement des locuteurs par maximisation de distances.

Du côté terminal, tout ce traitement est transparent. Il est à noter que le fait de prononcer quelques mots hors conférence permet notamment de ne pas avoir à attendre que le participant

parle pendant la conférence pour le positionner. Cependant s'il est impossible de déterminer la position du locuteur avant ses premiers mots en communication (si la personne n'a pas parlé, ou si l'on décide de ne pas demander de parler avant l'entrée en conférence), on pourrait par exemple entendre le locuteur à une position arbitraire (au centre) jusqu'à ce que le système ait effectué le choix de la position, position que le locuteur rejoindrait directement ou progressivement.

Concernant le fonctionnement de ce même algorithme sur un terminal, la Figure 3.5 peut être réutilisée avec une petite modification : le flux spatialisé en sortie du spatialiseur est destiné à la carte son et au système de rendu audio 3D de l'utilisateur (casque, haut-parleurs, …) et non plus à d'autres terminaux.

Un terminal entrant en conférence n'a aucune information sur les voix des autres participants. Contrairement au cas avec pont où les paramètres d'identification des autres participants sont immédiatement disponibles à l'arrivée d'un nouveau participant, il doit ici au contraire utiliser les premières secondes de parole de chaque participant (et cela même plusieurs minutes après son entrée en conférence) pour les établir. Il doit donc en plus jouer en parallèle ces flux sans les avoir placés de manière optimisée. On peut ainsi imaginer que chaque personne qui prend la parole une première fois depuis l'arrivée du nouveau participant soit placée par défaut au centre. Ensuite elle bouge de manière fluide ou instantanée au bout de quelques secondes vers une autre position déterminée grâce à l'algorithme.

On peut aussi envisager une solution dans laquelle chaque participant communique aux autres ses paramètres d'identification vocale par exemple dans le corps d'un message de protocole de signalisation. Cela permet d'éviter à chaque participant tout le processus de calcul des paramètres d'identification et ainsi diminuer le temps de calcul et le temps d'attente.

3.1.2.2.4 Discussion sur la solution proposée

En résumé, le but de cette solution est de coupler 3 blocs (détermination d'une matrice de distances vocales, ordonnancement et correspondance ordre/position) pour offrir un positionnement optimisé au sens du critère choisi (ici la distance vocale) des locuteurs.

Cette solution est susceptible de fonctionner quelle que soit la largeur de bande (narrowband, wideband, …) ou la dégradation de la qualité (due au bruit, à une communication mobile, …). On peut d'ailleurs imaginer qu'une personne avec un bruit trop important puisse être isolée des autres spatialement. Cela impliquerait d'autres calculs qui permettraient d'extraire l'énergie du bruit ambiant en période d'absence de parole de la part du locuteur et ainsi d'adapter la position du locuteur. On pourrait ainsi coupler (ou choisir uniquement un des deux critères) les notions de distances vocales inter-participants avec un critère "niveau de bruit ambiant" pour disposer les participants.

Bien que son apport semble évident par rapport à une politique de placement aléatoire, cette politique de placement doit être encore validée subjectivement. La difficulté que nous rencontrons est de fixer la méthode de test adaptée (identification des locuteurs, détermination du nombre de locuteurs, test d'intelligibilité, etc.), ainsi que les paramètres que sont : le nombre de locuteurs à partir duquel la méthode est intéressante, l'impact des écarts angulaires choisis, l'effet mémoire ou d'apprentissage que la méthode de test peut engendrer sur les résultats, etc.

En termes d'inconvénients, la contrainte la plus forte est le moment d'extraction de ces paramètres vocaux. Nous avons en effet souligné le fait que ce n'est pas toujours évident de récupérer ces informations suivant la configuration de la conférence.

Une fois cette méthode validée, une publication est envisagée.

3.1.2.2.5 Avantages attendus de la solution

- La solution devrait aboutir à une meilleure compréhension, à une meilleure identification des différents locuteurs et à un confort d'écoute accru.
- Aucune manipulation de la position des locuteurs de la part de l'auditeur quel que soit l'endroit où a lieu la spatialisation.

- Aucune modification pour définir un dialogue entre un pont et un terminal pour placer des locuteurs.
- Quel que soit le nombre de participants, la solution les place de la manière optimale au sens du critère utilisé, ce qui s'avère difficile pour un utilisateur au-delà de quelques participants.
- L'adaptation du système à la prise en compte d'un participant qui entre/sort.
- La distance vocale peut être basée sur des critères orientés "signal" autres que la ressemblance entre les voix, comme par exemple la puissance vocale.
- La distance vocale peut aussi être basée sur des critères orientés "identification du locuteur" tels que l'importance hiérarchique du locuteur, pour lesquels l'identification peut être réalisée grâce au signal vocal ou de toute autre manière (transmission de paramètres d'identification).
- La solution peut s'adapter à de multiples autres méthodes de positionnement, comme par exemple maximiser simplement la somme des distances vocales selon toutes les combinaisons possibles.

3.1.3 La gestion de positionnement dite manuelle

Le positionnement manuel ne suit aucune règle particulière si ce n'est celle de l'utilisateur. Cet utilisateur a donc besoin d'une interface pour placer tel qu'il le souhaite les différentes sources sonores.

Un exemple d'interface pour commander la spatialisation est donnée Figure 3.6. En cliquant avec le bouton droit sur la souris sur le nom du participant *arnault*, le participant *aurélien* peut régler la position à laquelle il entend *arnault*. En l'occurrence suite à son réglage, *aurélien* entendra *arnault* à une position angulaire de -50° par rapport à une position à 0° se trouvant en face de lui.

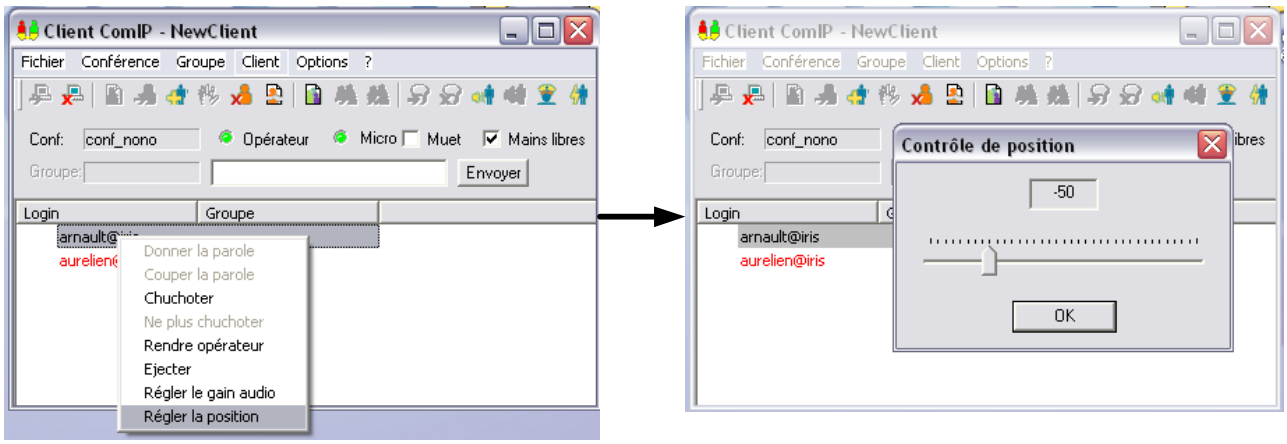

Figure 3.6 : Exemple d'interface du démonstrateur *ComIP 3D* développée par Aurélien Sollaud de France Télécom

Tout comme pour la gestion de positionnement automatique, la gestion de positionnement manuelle doit obligatoirement tenir compte de données telles que :

- L'arrivée ou le départ des autres participants dans la conférence.
- Le moment d'arrivée des participants (avant ou après l'auditeur) dans la conférence. Tous les participants arrivés avant l'utilisateur sont placés par défaut avant que l'utilisateur modifie éventuellement leur position suivant son souhait. Tout participant arrivant après l'utilisateur sera placé par défaut à une position, avant une éventuelle modification par l'utilisateur.

Par contre, suivant que la spatialisation est effectuée sur un pont de conférence mixeur ou un terminal, le traitement est complètement différent.

- Spatialisation manuelle sur un terminal : elle ne nécessite pas de traitement complexe puisque le softphone IP situé sur le terminal et gérant les flux médias peut facilement récupérer les positions souhaitées par un utilisateur par l'intermédiaire d'une interface telle que celle présentée ci-dessus (Figure 3.6) pour placer les sources sonores. En conséquence, nous ne la détaillerons pas de manière plus complète.
- Spatialisation manuelle sur un pont mixeur : cette gestion de la spatialisation est la plus complexe car elle nécessite un dialogue entre le pont de conférence mixeur et le terminal pour spécifier la position des locuteurs. Il faut de plus être informé de l'arrivée ou du départ de tout participant. Nous présentons une réflexion sur ce domaine dans la section 3.1.3.1.

3.1.3.1 Données nécessaires pour contrôler la spatialisation effectuée sur un pont mixeur à partir d'un terminal

La dissociation de l'entité effectuant la spatialisation de l'entité la contrôlant nécessite un dialogue entre ces deux parties. Nous chercherons dans cette section à spécifier les informations nécessaires à ce dialogue.

La liste des informations que tout terminal doit recevoir est donnée ci-dessous. Nous y reviendrons dans la section 3.1.3.1.1. Il est à noter que cet échange peut aussi avoir lieu pour une gestion de positionnement automatique avec spatialisation sur pont, à la différence près qu'il n'y aura pas de retour sur les commandes effectuées du fait du mode de gestion.

- Arrivée/départ d'un participant à la conférence pour pouvoir éventuellement demander une mise à jour des positions.
- Mise à jour des informations concernant les participants.
- Retour sur les commandes effectuées pour connaître leurs résultats.

La liste des commandes à pouvoir effectuer sur le pont est donnée ci-dessous. Nous y reviendrons dans la section 3.1.3.1.2.

- Modifier la position d'un locuteur en accord avec la demande faite par l'utilisateur. Pour cette méthode, l'identifiant et la position du locuteur doivent être spécifiés.
- Inverser la position d'un participant avec celle d'un autre pour par exemple un meilleur confort d'écoute. Pour cette méthode, les identifiants des deux locuteurs doivent être spécifiés.
- Grouper plusieurs participants lorsque le nombre de participants devient important et la discrimination difficile. Pour cette méthode, les identifiants des locuteurs à regrouper doivent être fournis ainsi que la position à laquelle ils doivent être placés. La commande inverse "Dissocier plusieurs participants" doit pouvoir être effectuée.
- Demander un positionnement automatique lorsque le nombre de participants devient important, cela afin d'aider l'utilisateur.

L'identification du contenu spatialisé est envoyée du pont vers le terminal. Nous y reviendrons dans la section 3.1.3.1.3.

3.1.3.1.1 Recevoir les informations des autres participants par le pont mixeur

Tout terminal de la conférence audio spatialisée doit être informé de l'arrivée ou du départ d'un autre participant pour pouvoir le placer à la position qu'il souhaite. Il doit aussi être informé si ses ordres ont été réalisés. Nous allons pouvoir pour cela utiliser le protocole de signalisation SIP.

Le protocole SIP permet d'être informé de certains évènements d'une conférence audio standard centralisée grâce à des mécanismes que nous allons présenter : la souscription à des évènements ainsi qu'une structure permettant de décrire une conférence audio centralisée.

Le protocole SIP supporte un mécanisme autorisant la souscription à des évènements [74]. Ce mécanisme sert notamment pour des applications de présence. Un *User Agent*, Figure 3.7, qui veut être notifié de certains évènements, envoie un message *SUBSCRIBE* à un serveur (qui génère des évènements ou les reçoit). Le serveur répond par un message *200 OK* et envoie une requête *NOTIFY* automatiquement au *User Agent* pour préciser la durée de la souscription. Cette dernière reçoit évidemment une réponse *200 OK* de la part de l'UA. Ensuite tout changement en rapport avec la souscription sera notifié à l'UA. Le message SUBSCRIBE doit être renvoyé périodiquement.

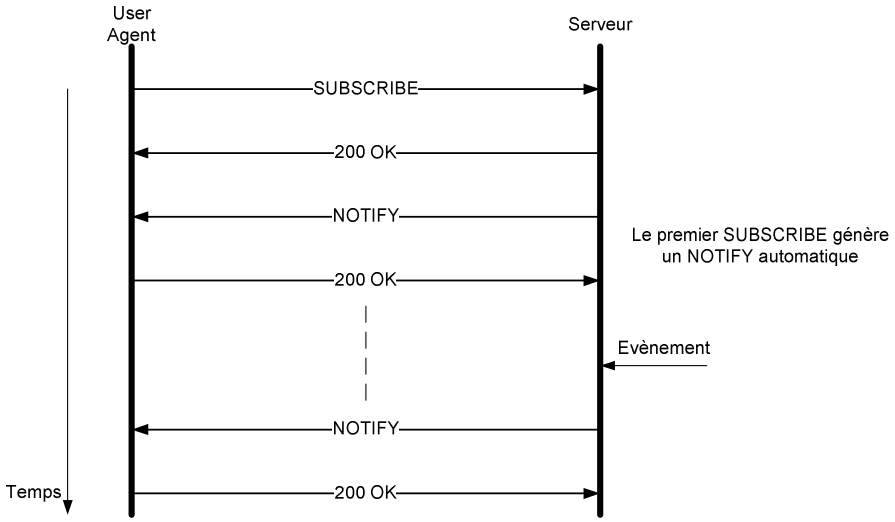

Figure 3.7 : Principe des messages SUBSCRIBE/NOTIFY

Concrètement, le *User Agent* est un terminal, et le serveur un pont de conférence de type mixeur. Les données exprimées en XML, auxquelles le terminal souscrit, sont mises à jour par des évènements relatifs à la conférence audio (arrivée ou départ d'un participant, etc.), suivant le *conference event package* définis dans la RFC 4575 [83]. Cela sous-entend que le notifieur (le pont) et le souscripteur (le terminal) doivent supporter ce package. Les conditions courantes d'envoi de *NOTIFY* sont l'arrivée ou le départ d'un participant, le changement d'un type de média ou de l'URI de la conférence, ainsi que toute modification dans l'état de cette dernière ou des participants.

Le corps du message *NOTIFY* est donc en XML et est composé d'éléments imbriqués. Ces derniers peuvent transporter trois états de l'information : *full*, *partial* ou *deleted*.

Un élément reçu par le souscripteur avec la mention *full* signifie que les informations, correspondant à celui-ci et reçues précédemment, doivent être entièrement remplacées. Cela est effectué évidemment si les informations reçues sont plus récentes que les précédentes (dépend des aléas de la transmission de données sur un réseau à commutation de paquets), ce qui est identifiable par un numéro de version.

Un élément reçu par le souscripteur avec la mention *partial* signifie que les informations, correspondant à cet élément et reçues précédemment, doivent être comparées et mises à jour.

Un élément reçu par le souscripteur avec la mention *deleted* signifie que les informations, correspondant à cet élément et reçues précédemment, doivent être supprimées.

Il est à noter que ce type de structure peut être imbriqué. Ainsi les sous-éléments d'éléments avec un état *full* (resp. *deleted*), sont eux-mêmes *full* (resp. *deleted*). Par contre, les sous-éléments d'éléments avec un état *partial* peuvent être *full*, *partial* ou *deleted*.

Voici brièvement les principales informations apportées par ce package :

- Concernant le type de la conférence :
  - l'URI.
  - Le numéro de version afin de savoir ordonner temporellement les *NOTIFYs*.

- Concernant le type de la conférence :
  - Le sujet de la conférence.
  - La description de la conférence.
  - Les URIs pour avoir des services supplémentaires.
  - Le nombre de participants maximum.
  - Le type de média disponible, etc.
- Concernant l'état de la conférence :
  - Le nombre de participants.
  - Le type de média actif et leurs identifiants, etc.
- Concernant les utilisateurs :
  - Leur URI.
  - Leur statut (connecté ou non).
  - Les raisons de leur déconnexion.
  - La façon dont ils ont rejoint la conférence (par qui, à quel moment…).
  - Le type de média qu'ils utilisent actuellement, etc.

Ces informations sont pour nous très importantes puisqu'elles nous permettent notamment d'être informés de l'arrivée ou du départ d'un participant. Les informations à transmettre du pont vers le terminal peuvent être incluses dans ce package. En effet dans la partie concernant les utilisateurs, nous pouvons introduire une extension XML normalisable insérant pour chaque utilisateur sa position. Elle nous permet de contrôler la position actuelle des participants, de mettre à jour l'interface graphique et de vérifier le résultat de nos actions (changement de position, inversion de deux participants, etc.).

3.1.3.1.2 Transmettre des commandes et des informations d'un terminal vers un pont

Selon nous, la liste des commandes à effectuer sur le pont n'est pas intégrable dans le protocole SIP. En effet, le but du protocole SIP est d'établir et de gérer des sessions multimédias entre 2 terminaux en négociant les codeurs, le nombre de canaux, les paramètres de transport etc. Son objectif n'est pas de commander des modifications sur le contenu transporté. De plus, il nous semble difficile de normaliser les commandes décrites ci-dessus (en introduction de la section 3.1.3.1) du fait des nombreuses possibilités de représenter ces informations.

Selon nous, le plus simple est d'utiliser par exemple une interface web comme cela se fait actuellement pour les ponts de conférence standard (voir Figure 3.8). On parle alors de web-pilotage. Il est à souligner que d'autres méthodes propriétaires peuvent être utilisées.

Dans ce cas, chaque utilisateur se connecte au site web de la conférence avec son nom d'utilisateur et le mot de passe spécifique à la conférence (donné par l'organisateur ou par le système). Par cette interface, nous transmettons grâce au protocole HTTP (*Hypertext Transfer Protocol*) les informations concernant la commande choisie (déplacement d'un locuteur, inversion, etc.) ainsi que ses paramètres (identifiant, etc.).

Nous aurons donc en plus des actions standards illustrées Figure 3.8 (fermer le microphone, déconnecter le participant, etc., du participant *aurélien*) nos propres commandes spécifiques à la gestion du son spatialisé. Une fois les ordres reçus, le pont de conférence mixeur gère la scène sonore telle qu'on lui a demandée et renvoie les informations par les messages de notification vus dans la section précédente et par l'interface web. Un terminal peut ainsi avoir deux sources d'informations : les messages SIP et, s'il le peut, l'interface web.

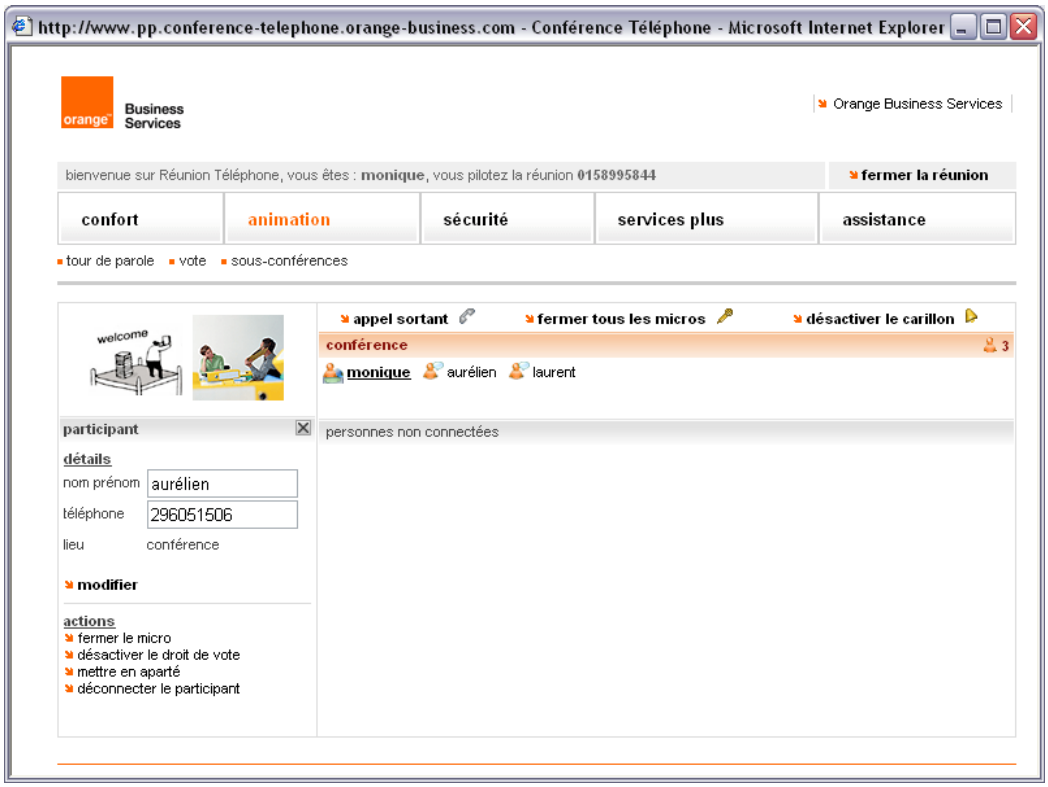

Figure 3.8 : Exemple de web-pilotage d'une conférence classique

3.1.3.1.3 L'importance de l'identification du contenu transporté

Comme on l'a vu, le protocole SIP permet notamment de négocier un codeur et un nombre de canaux. Tant que les conférences restaient monophoniques, stéréophoniques ou de type 5.1, le protocole SIP était suffisamment robuste pour que les deux entités en train de communiquer puissent bien gérer les contenus.

Néanmoins avec l'apparition de nouveaux types de contenus, binaural et stéréo-dipôle notamment, se pose la question de l'utilisation de ceux-ci. En effet, un contenu stéréo-dipôle n'est nativement pas idéal pour une restitution au casque et réciproquement. Le problème est même vrai au-delà de ce contexte. Quel que soit le format de la prise de son ou du dispositif de restitution du son (nombre de haut-parleurs etc.), il est en effet toujours nécessaire de séparer le contenant du contenu pour notamment pouvoir effectuer des traitements complémentaires.

Ainsi bien qu'il garantisse un décodage adapté au contenu, le protocole SIP ne permet donc pas d'utiliser de manière optimale les contenus car il ne les identifie pas. Nous avons déposé un brevet (n° FR 05 54106) pour spécifier le type de contenu transporté (binaural, stéréo-dipôle, etc.) dans toute communication, notamment par l'intermédiaire d'un protocole de signalisation comme le protocole SIP. Néanmoins, comme nous l'avons souligné, le rôle de SIP est de mettre en relation deux entités, et non d'identifier un contenu qui peut être sous des formes multiples.

Il se pose donc le problème pour un terminal récepteur de savoir comment exploiter le contenu reçu. Il convient de définir le type de format échangé et ainsi la méthode de spatialisation souhaitée par un autre moyen que le protocole SIP. L'interface web illustrée ci-dessus serait un exemple de méthode de spécification du type de format échange (par exemple binaural).

Néanmoins, un terminal ne pouvant pas se connecter à cette interface ne pourra pas connaître le type de flux échange sauf si un type par défaut est connu. Ce terminal est donc susceptible de ne pas restituer le contenu spatialisé de manière optimale (contenu binaural sur un casque par exemple) mais cela ne devrait pas être trop gênant pour l'intelligibilité.

3.2 L'asymétrie du contenu dans le cas de la conférence audio centralisée avec pont mixeur

Dans notre hypothèse de départ, chaque terminal envoie un flux audio monophonique codé vers le pont de conférence et reçoit de celui-ci un contenu binaural codé sur deux canaux. Le premier besoin exprimé concerne donc la nécessité de pouvoir négocier un contenu asymétrique entre deux entités (pont ou terminal) grâce au protocole SIP. En effet, dans les conférences audio standards, les flux échangés sont de type monophoniques que cela soit d'un terminal vers un autre terminal, d'un terminal vers un pont ou réciproquement. Cette asymétrie ne peut pas être exprimée facilement par le protocole SIP/SDP à ce jour. Nous proposons ici une évolution du protocole pour adresser ce besoin dans la section 3.2.

Cette partie a pour objectif de présenter notre proposition SIP pour permettre un échange de contenu asymétrique entre deux entités dans un réseau. Cette proposition fera l'objet d'une action en normalisation à l'IETF courant 2008.

3.2.1 Exemple d'un échange SDP classique

Rappelons brièvement un exemple de SDP (issu du logiciel eConf, client VoIP de France Télécom) basé sur le modèle offre/réponse défini dans la RFC 3264 [84] pour négocier les codeurs et envoyé lors d'un établissement d'appel classique (voir chapitre 1 pour plus de détails).

```
v=0
o=offerer 1184762939 1184762939 IN IP4 192.0.2.102
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.102
b=AS:384
t=0 0
m=audio 6000 RTP/AVP 103 8 106
a=rtpmap:103 AMR/8000/1
a=rtpmap:8 PCMA/8000
a=rtpmap:106 AMR/8000/2
a=sendrecv
```

Cette offre SDP propose un flux audio avec 3 possibilités de codeurs pour le compresser, dans l'ordre de préférence : AMR, G.711, AMR stéréo. L'attribut `a=sendrecv` indique le caractère bidirectionnel de la communication souhaitée, mais ne permet pas de choisir un codeur par direction.

Une réponse possible à l'offre précédente est donnée ci-dessous :

```
v=0
o=answerer 1184759768 1184759768 IN IP4 192.0.2.25
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.25
b=AS:384
t=0 0
m=audio 8000 RTP/AVP 103
a=rtpmap:103 AMR/8000/1
a=sendrecv
```

Le terminal appelé a choisi le premier codeur dans la liste, AMR. La communication va s'établir de manière symétrique et monophonique.

3.2.2 Etat de l'art

L'état de l'art de la négociation de media en SDP a été parcouru, avec notamment les documents IETF suivants :

- La RFC 4566 de nom SDP : Session Description Protocol [35]
- La RFC 3264 de nom An Offer/Answer Model with the Session Description Protocol [84]
- La RFC 3388 de nom Grouping of Media Lines in the Session Description Protocol [19]
- La RFC 3407 de nom Session Description Protocol Simple Capability Declaration [7]
- Le draft draft-ietf-mmusic-sdp-capability-negotiation-06 de nom SDP Capability Negotiation
- Le draft draft-ietf-mmusic-sdp-media-capabilities-01 de nom SDP media capabilities Negotiation

Le moyen de négocier une communication non symétrique est d'utiliser 2 lignes media (`m=`) en utilisant pour l'une l'attribut `a=sendonly` et pour l'autre l'attribut `a=recvonly`. L'inconvénient de cette technique est qu'on est obligé d'utiliser des numéros de ports différents, soit 2 ports RTP et 2 ports RTCP, ce qui ne facilite pas le passage de NAT et le déploiement à grande échelle. De plus, elle sépare artificiellement les deux sens de communications alors qu'il s'agit bien d'un même appel.

De plus, la plupart des terminaux ne sait pas gérer plus d'une ligne d'un média donné (audio, vidéo…). On risquerait alors de se retrouver avec une communication monodirectionnelle.

3.2.3 Solution proposée

La méthode la plus simple et la plus interopérable avec l'existant est de définir un nouvel attribut SDP pour préciser les codeurs souhaités dans chaque direction de la communication.

3.2.3.1 Syntaxe du nouvel attribut média Asymmetric Send Receive

On définit un attribut `a=asr`, valable uniquement au niveau média :

```
a=asr s=<liste des codeurs en émission> r=<liste des codeurs en réception>
```

Les listes sont constituées à partir des numéros de codeurs (payload types) séparés par des virgules.

On définit 2 cas d'utilisation :

- La définition des codeurs souhaités dans chaque sens, dans ce cas les listes `s=` et `r=` sont obligatoirement présentes.
- L'indication de la capacité : dans ce cas on indique seulement l'attribut `a=asr` et rien derrière.

3.2.3.2 Utilisation dans le cadre de l'offre/réponse SDP

On se restreint au cas unicast, seul cas pertinent pour la conférence centralisée.

3.2.3.2.1 Génération de l'offre

L'offre SDP est construite de manière conforme à la RFC 3264.

Pour les média ayant un attribut `a=sendrecv` ou `a=inactive`, l'émetteur de l'offre peut indiquer sa volonté d'établir une communication asymétrique en insérant l'attribut `a=asr`. La liste de codeurs `s=` est classée selon l'ordre de préférence des codeurs pour l'émission, indépendamment de l'ordre indiqué sur la ligne `m=`. Il en est de même pour la liste `r=` pour les codeurs en réception. Les listes `s=` et `r=` sont obligatoirement des sous-ensembles de la liste de formats sur la ligne `m=`.

Conformément à la RFC 3264, l'émetteur de l'offre doit être prêt à recevoir du média pour toutes les lignes m=, sans se restreindre à la liste r=.

3.2.3.2.2 Génération de la réponse

Si le récepteur de l'offre ne supporte pas l'attribut a=asr, il ignore la ligne correspondante et génère une réponse selon la RFC 3264.

Si le récepteur supporte a=asr, il doit ignorer l'ordre de préférence des codeurs indiqué sur la ligne m=, et tenir compte de l'ordre des listes s= et r=.

Le récepteur doit obligatoirement insérer une ligne a=asr dans sa réponse, et construire ses propres listes s= et r= en fonction de l'offre. La liste s= de la réponse est construite à partir de la liste r= de l'offre conformément à ce qui est décrit dans la RFC 3264 pour les listes de codeurs des lignes m=. De même, la liste r= de la réponse est construite à partir de la liste s= de l'offre.

La liste des codeurs sur la ligne m= est la réunion des listes s= et r=, et son ordre n'a pas d'importance. De même les attributs a=rtpmap correspondant à cette liste de codeurs doivent être présents.

Si le récepteur supporte a=asr mais que l'offre ne contient pas cet attribut, il est possible de l'insérer dans la réponse pour indiquer le support de cette fonctionnalité.

3.2.3.2.3 Interprétation de la réponse

Si la réponse ne contient pas l'attribut a=asr, elle est traitée par l'émetteur de l'offre initiale de manière conforme à la RFC 3264.

Si la réponse contient a=asr, les listes s= et r= sont traitées séparément, d'une manière équivalente au traitement dans la RFC 3264 des listes des codeurs des lignes m=. L'ordre des codeurs de la ligne m= est ignoré.

3.2.4 Exemples

3.2.4.1 Exemple 1 : Un terminal stéréo appelle un pont de conférence spatialisé

Le terminal va offrir un flux montant mono et un flux descendant stéréo :

```
v=0
o=endpoint 1184762939 1184762939 IN IP4 192.0.2.102
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.102
b=AS:384
t=0 0
m=audio 6000 RTP/AVP 103 8 106
a=rtpmap:103 AMR/8000/1
a=rtpmap:8 PCMA/8000
a=rtpmap:106 AMR/8000/2
a=sendrecv
a=asr s=103,8 r=106,103,8
```

Le pont de conférence va choisir le codeur préféré pour chaque sens de la communication :

```
v=0
o=pont de conférence 1184759768 1184759768 IN IP4 192.0.2.25
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.25
b=AS:384
t=0 0
m=audio8000 RTP/AVP 103 106
a=rtpmap:103 AMR/8000/1
```

```
a=rtpmap:106 AMR /8000/2
a=sendrecv
a=asr s=106 r=103
```

On établit alors un canal AMR mono montant vers le pont et un canal AMR stéréo descendant vers le terminal.

L'exemple d'un pont de conférence spatialisé qui appelle un terminal stéréo est équivalent, en permutant les `s=` et `r=`.

3.2.4.2 Exemple 2 : Un terminal stéréo appelle un pont de conférence classique

L'offre est la même que pour 3.2.4.1 ci-dessus, le terminal ne sait pas à l'avance quelles sont les capacités du pont de conférence.

```
v=0
o=endpoint 1184762939 1184762939 IN IP4 192.0.2.102
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.102
b=AS:384
t=0 0
m=audio 6000 RTP/AVP 103 8 106
a=rtpmap:103 AMR/8000/1
a=rtpmap:8 PCMA/8000
a=rtpmap:106 AMR/8000/2
a=sendrecv
a=asr s=103,8 r=106,103,8
```

Le pont de conférence ignore l'attribut `a=asr` qu'il ne connait pas, et répond classiquement :

```
v=0
o=pont de conférence 1184759768 1184759768 IN IP4 192.0.2.25
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.25
b=AS:384
t=0 0
m=audio 8000 RTP/AVP 8
a=rtpmap:8 PCMA/8000
a=sendrecv
```

La communication est alors symétrique en G.711 mono.

L'exemple d'un pont de conférence spatialisé qui appelle un terminal classique est équivalent, en permutant les `s=` et `r=`.

3.2.4.3 Exemple 3 : Un terminal classique appelle un pont de conférence spatialisé

Le terminal émet l'offre suivante :

```
v=0
o=endpoint 1184762939 1184762939 IN IP4 192.0.2.102
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.102
b=AS:384
t=0 0
m=audio 6000 RTP/AVP 103 8
a=rtpmap:103 AMR/8000/1
a=rtpmap:8 PCMA/8000
a=sendrecv
```

Le pont de conférence répond à cette offre de manière classique mais indique en plus son support du mode asymétrique. Le terminal émetteur pourrait, s'il le souhaite, effectuer une nouvelle offre en tenant compte de cette possibilité.

```
v=0
o=pont de conférence 1184759768 1184759768 IN IP4 192.0.2.25
s=-
i=eConf4.2.36
c=IN IP4 192.0.2.25
b=AS:384
t=0 0
m=audio 8000 RTP/AVP 103
a=rtpmap:103 AMR/8000/1
a=sendrecv
a=asr
```

3.3 Conclusion et perspectives

Dans ce chapitre, nous avons posé les bases de la gestion du positionnement des locuteurs en conférence audio spatialisée. Dans un premier temps, nous avons défini les termes à utiliser (gestion de positionnement automatique, manuelle, stratégie de placement, politique de placement, etc.). Dans un second temps, nous avons proposé des solutions adaptées à chaque contexte (gestion de positionnement automatique ou manuelle, et, spatialisation sur pont ou sur terminal).

Dans le cas de la spatialisation sur un pont de conférence, nous avons ainsi souligné le fait que cette gestion ne pouvait se faire par l'intermédiaire du protocole SIP, car ce n'est pas le rôle de ce dernier de transporter des informations sur le contenu. De plus, les configurations de restitution 3D sont multiples. Il y a certes le binaural, mais nous rappelons qu'il existe de nombreuses possibilités pour la position d'écoute par exemple pour le stéréo-dipôle. Nous avons proposé une solution basée sur ce qui se fait dans les conférences audio standard : une solution de web-pilotage certes propriétaire à chaque fournisseur de services mais en cohérence avec la gestion des protocoles de Voix sur IP.

Pour la conférence avec un pont mixeur, nous avons de même établi les paramètres du protocole de signalisation SIP nécessaires au transport de flux asymétriques tout en garantissant une interopérabilité avec les terminaux. La nécessité de transporter ces flux asymétriques est due à notre hypothèse de départ concernant l'équipement des terminaux : prise de son monophonique et restitution sur casque stéréo ou deux haut-parleurs. Nous avons enfin souligné l'importance de bien dissocier le contenu du contenant. Il est en effet important de pouvoir identifier le contenu pour pouvoir par la suite le restituer de manière adéquate ou pour pouvoir le modifier à la guise de l'utilisateur.

4. Evaluation de la qualité audio suivant les différentes architectures de conférence

Quel que soit l'objectif d'une recherche, d'une industrialisation ou d'un service que l'on souhaite mettre en place, une exigence de qualité est requise. Cela s'applique évidemment pour la conférence audio spatialisée. Ce chapitre a pour intention de tester la qualité audio et la qualité de spatialisation (voir section 1.7.2 pour la définition des termes employés) des deux architectures retenues dans le chapitre 2.

4.1 Présentation du contexte des études effectuées

Cette section vise notamment à justifier d'une part nos choix de codeurs et d'autre part la sélection des tests retenus dans les sections suivantes.

4.1.1 Justification des différents choix effectués pour les tests

4.1.1.1 Rappel des architectures retenues

Les deux schémas d'architecture de conférence audio spatialisée retenues sont rappelés ci-dessous, et vont servir de base pour nos tests de qualité audio et de spatialisation. Ils sont présentés de manière décomposée afin de bien séparer les émetteurs des récepteurs.

Le premier schéma, illustré Figure 4.1, correspond à la configuration centralisée avec pont mixeur. On retrouve bien :

- L'encodage monophonique au niveau de chaque terminal émetteur,
- Le décodage de chaque flux compressé monophonique au niveau du pont par le décodeur adapté,
- Le mixage des flux préalablement spatialisés sur le pont de conférence mixeur,
- L'encodage de ces flux spatialisés par un codeur stéréophonique adapté (dans le sens deux canaux dans ce chapitre),
- Le décodage de ces flux stéréophoniques compressés par le décodeur adapté au niveau du terminal récepteur.

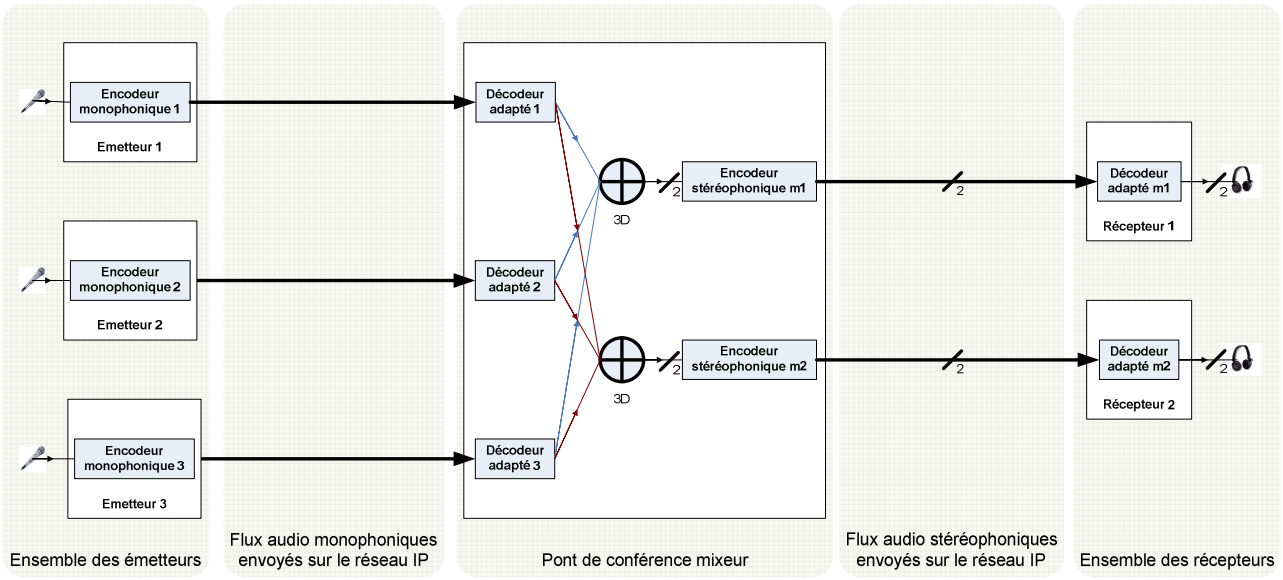

Figure 4.1 Illustration de la conférence audio centralisée avec l'utilisation d'un pont mixeur (les émetteurs et les récepteurs sont évidemment généralement confondus dans une conférence)

Le second schéma, illustré Figure 4.2, regroupe l'ensemble des autres configurations que nous avons appelées : centralisée avec pont répliquant, distribuée multi-unicast et distribuée multicast. Pour simplifier, nous les regrouperons dans ce chapitre sous les termes *architecture distribuée* dans le sens où un seul encodage/décodage est effectué. On y retrouve bien :

- L'encodage monophonique au niveau de chaque terminal émetteur,
- Le décodage de chaque flux compressé monophonique au niveau de chaque terminal récepteur par le décodeur adapté,
- Le mixage des flux préalablement spatialisés sur le récepteur.

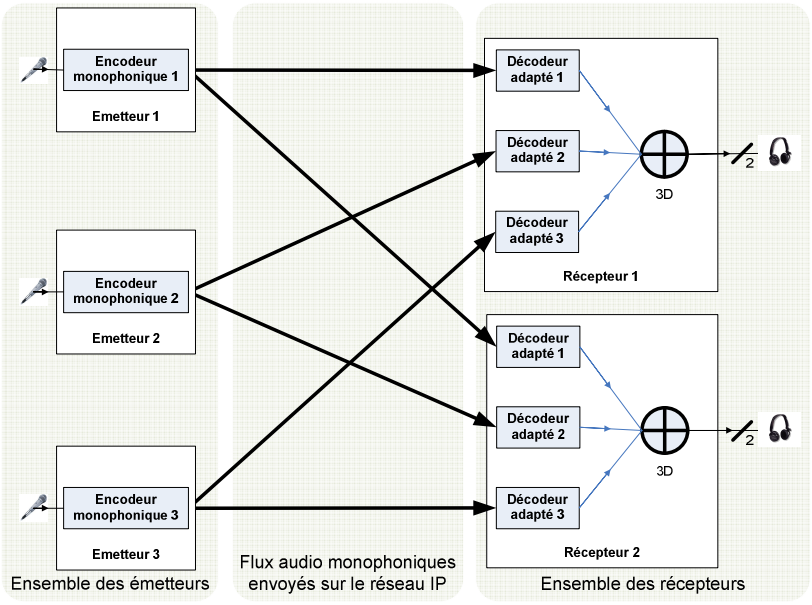

Figure 4.2 Illustration de la conférence audio distribuée (les émetteurs et les récepteurs sont évidemment généralement confondus dans une conférence)

4.1.1.2 Justification de la méthode retenue pour la spatialisation

Il convient tout d'abord de sélectionner une méthode de spatialisation pour ces tests. Le choix s'est porté sur la méthode binaurale pour plusieurs raisons. Tout d'abord d'un point de vue technique, notre méthode de spatialisation stéréo-dipôle découle de la méthode binaurale (cf. la

section 1.7.4) et peut être effectuée par un terminal quelle que soit l'architecture sélectionnée comme souligné dans le chapitre **2**. Ensuite on fait l'hypothèse raisonnable que la qualité audio n'évolue pas perceptivement lors du passage du binaural vers le stéréo-dipôle. Enfin, d'un point de vue pratique notamment pour les tests, l'écoute au casque est beaucoup plus intéressante car elle peut être réalisée par plusieurs sujets en parallèle.

4.1.1.3 Justification des codeurs retenus pour le transport du contenu binaural

Pour le transport des données monophoniques, les codeurs de parole monophoniques standards narrowband ou wideband sont logiquement adaptés et par conséquence utilisés. Par contre, plusieurs solutions sont disponibles pour transporter le contenu binaural entre le pont mixeur et le terminal récepteur pour la conférence audio illustrée Figure 4.1 :

- L'utilisation de codeurs stéréophoniques adaptés au transport de contenus sur deux canaux comme le binaural,
- L'utilisation de techniques de compression de contenus multicanal et les codeurs associés,
- L'utilisation de codeurs monophoniques pour compresser indépendamment chaque canal du contenu binaural appelé aussi codage dual-mono.

4.1.1.3.1 Les codeurs stéréophoniques

Les codeurs stéréophoniques sont généralement dédiés au transport de la musique plutôt que de la parole. De ce fait, leurs délais algorithmiques ne sont pas prévus pour être compatibles avec les contraintes d'interactivité de la conférence audio (le délai de bout en bout devant être inférieur à 150 ms [44]).

Le seul codeur stéréophonique existant respectant les contraintes conversationnelles est le codeur MPEG 4 *Advanced Audio Coding Low Delay* (AAC-LD) [40]. Ce codeur est une norme MPEG pour laquelle seul le décodeur est normatif. Ainsi différentes implémentations de l'encodeur sont possibles suivant les contraintes que l'on se fixe en termes de qualité et de complexité. Le codeur AAC-LD permet de transporter un contenu monophonique, stéréophonique ou multicanal à un débit entre 32 et 80 kbits/s par canal pour une fréquence d'échantillonnage entre 22.05 et 48 kHz. Le délai algorithmique de ce codeur (encodeur et décodeur) peut varier entre 20 et 42 ms selon le débit et la fréquence d'échantillonnage utilisés. Ce délai est inférieur au délai du codeur MPEG 4 AAC [40,58] (au minimum 55 ms à 48 kHz) dont il est issu.

Nous n'avons pas sélectionné ce codeur pour nos tests par un souci de qualité et d'interopérabilité avec l'existant. En effet, aucune garantie n'est apportée sur la qualité d'une implémentation d'encodeur AAC-LD (puisqu'il n'est pas normalisé), ce qui n'est pas acceptable pour un service. De plus, ce codeur n'est pas déployé sur les plateformes conversationnelles actuelles, ni sur les terminaux. Pour ces derniers, les codeurs stéréophoniques généralement utilisés sont des codeurs de diffusion sans contraintes temps-réel tels que le MP3, l'AAC et le HE-AAC [40].

Il est à noter que des travaux ont été lancés à MPEG pour développer un nouveau codeur : le codeur MPEG 4 AAC *Enhanced Low Delay*. Une fois encore, seul le décodeur sera normalisé (date prévue : fin 2007), ce qui, de nouveau, ne garantit pas la qualité de l'ensemble encodeur-décodeur. Ce codeur permettra de transporter un contenu monophonique, stéréophonique ou multicanal. Les débits seront compris entre 24 et 64 kbits/s par canal pour une fréquence d'échantillonnage de 48 kHz. Le délai algorithmique sera compris entre 15 et 32 ms.

Des extensions stéréophonique et super wideband de codeurs conversationnels (G.729.1, EV-VBR) sont en cours de normalisation à l'ITU-T/SG16, mais le calendrier de normalisation prévoit une sélection des candidats seulement en juillet 2008. Il conviendra de surveiller ces codeurs dans un proche avenir pour étudier leur capacité à transporter un contenu binaural.

4.1.1.3.2 Méthodes de compression de contenu multicanal

Dans le cadre de la compression de contenu multicanal, outre les technologies de stéréo jointe, des méthodes reposant sur la modélisation paramétrique de l'information spatiale sont apparues dans les années 2000 [18].

Les méthodes de base sont le BCC (*Binaural Cue Coding*) [12] et le *Parametric Stereo* [17], visant toutes deux à réduire tout type de contenu sur deux canaux en un contenu monophonique d'une part, et en informations spatiales (différence de temps, d'intensité et de cohérence entre les deux canaux) d'autre part. Le contenu monophonique résulte de la somme par portion temps-fréquence des deux canaux du contenu de départ. Le débit total à transmettre correspond donc au débit d'un flux monophonique codé (dépendant du codeur monophonique retenu), additionné du débit nécessaire pour transporter les informations spatiales (de l'ordre de 2 kbits/s). Cela permet une réduction de débit intéressante par rapport à un codage dual-mono. Cependant les codeurs comme le codeur MPEG 4 HE-AAC v2 [40], utilisant cette technologie utile pour transporter un contenu binaural, ne sont pas adaptés au transport de contenu conversationnel du fait de leur délai algorithmique trop important (de l'ordre de 130 ms [58]).

Par la suite, des travaux [32,37] ont permis la généralisation de la méthode de compression vue dans le paragraphe précédent mais pour un nombre quelconque de canaux. Le principe appelé aussi *Spatial Audio Coding* (SAC) part d'un contenu multicanal (5.1, 22.2, …) et le réduit à un contenu sur un nombre de canaux quelconques (inférieur au nombre de canaux du signal original) couplé avec des informations spatiales additionnelles (différences de niveaux et cohérence entre les différents canaux). Le contenu réduit est généralement codé par un codeur adapté, par exemple le codeur MPEG 4 HE-AAC. L'architecture retenue (réduction du nombre de canaux couplé à des informations spatiales) permet une rétrocompatibilité avec les équipements ne comprenant pas les informations spatiales incluses dans le codage. Ainsi, un terminal pourra tout de même jouer le contenu réduit même s'il ne sait pas comment interpréter les informations spatiales. Début 2007, ces travaux ont abouti à la norme MPEG Surround (MPS) [5]. Des recherches basées sur une extension de cette approche sont actuellement en cours.

Cette extension appelée *Spatial Audio Object Coding* (SAOC) [37] rejoint la philosophie orientée objet de MPEG 4. Elle est beaucoup plus intéressante pour nous que MPEG Surround car elle part d'une hypothèse de non corrélation des canaux de départ. Autrement dit, chaque canal est vu comme un objet et peut être un flux audio monophonique complètement indépendant des autres, par exemple un flux audio issu d'un participant à une conférence… L'intérêt est l'utilisation de ce genre de technique dans le cadre de la conférence centralisée avec un pont mixeur pour une transmission à un débit optimal entre le pont et chaque terminal.

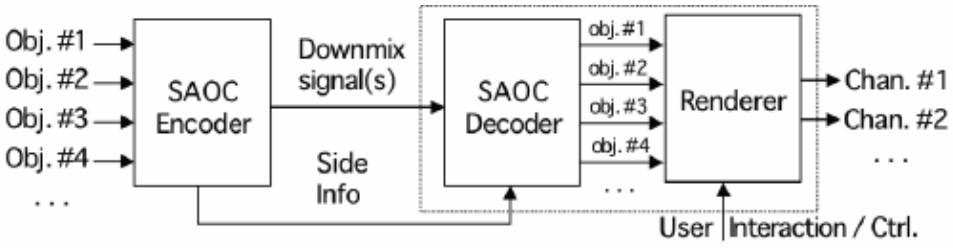

Figure 4.3 Illustration de la méthode SAOC (issu de [37])

Plus précisément cette méthode SAOC, illustrée Figure 4.3, utilise un encodeur SAOC afin de créer un signal monophonique ou stéréophonique (*Downmix signals*) plus des informations complémentaires (*Side Info*) à partir des signaux (ou objets) initiaux. Le principe de cet encodeur est de générer une scène virtuelle pour créer une corrélation entre les objets et se rapprocher du principe vu pour l'encodeur MPEG Surround. Côté réception, un décodeur SAOC utilise le signal reçu plus les informations complémentaires pour créer à nouveau les objets. Ces derniers peuvent être par la suite spatialisés par une entité complémentaire (appelée *Renderer* sur le schéma) afin d'être restitués aux positions voulues par l'utilisateur et sur un nombre quelconque de canaux (deux pour le binaural par exemple).

Dans [37], il est aussi question de réutiliser les briques créées pour le décodeur MPEG Surround, comme illustré Figure 4.4. On remarque que le bloc Renderer défini ci-dessus s'est transformé en *SAOC-to-MPS Transcoder* et le flux Side Info en *SAOC bitstream*. Ce décodeur SAOC inclut et commande ainsi un décodeur MPEG Surround.

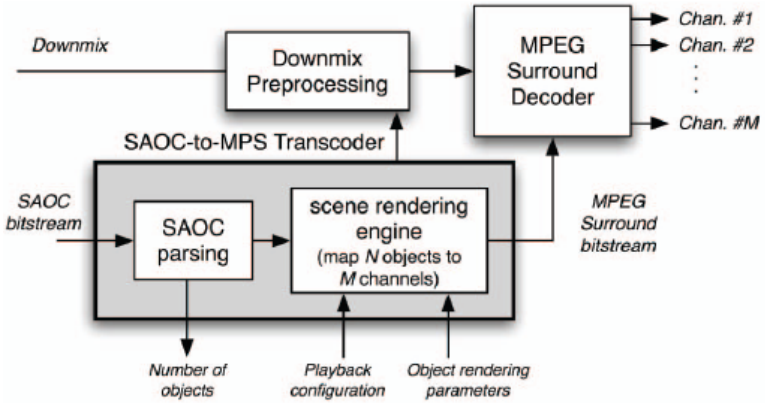

Figure 4.4 Exemple de décodeur simplifié SAOC (issu de [37])

Cette méthode est prometteuse et méritait donc d'être présentée cependant elle en est encore au stade de la conception. En conséquence, nous ne l'utiliserons pas pour nos tests. A priori, elle n'est pas adaptée pour respecter les contraintes conversationnelles, mais il semble possible d'utiliser un codeur tel que l'AAC LD. Néanmoins, le problème de l'encodeur non normalisé restera toujours d'actualité...

4.1.1.3.3 Le codage dual-mono

Le codage dual-mono code chaque canal d'un contenu stéréo indépendamment de l'autre par un codeur monophonique, qui est dans notre cas dédié au transport de la parole et normalisé. Ce type de codeur possède donc des délais algorithmiques compatibles avec les contraintes d'interactivités de la voix sur IP. Cette méthode possède des inconvénients immédiats par rapport à l'utilisation d'un codeur stéréophonique qui sont : l'absence d'économie de bande passante si les deux canaux sont fortement corrélés et le risque d'une perte de qualité de spatialisation en modifiant, par le codage dual-mono, des paramètres tels que l'ILD ou l'ITD (cf. section 1.7.4).

Cependant cette solution présente un grand avantage en termes d'interopérabilité. Dans le cadre d'un service qui vise à connecter tous les participants possédant un terminal ayant des fonctionnalités monophonique ou stéréophonique, il est difficilement possible de se passer de ces codeurs et donc du codage dual-mono. Notre choix de partir de ces codeurs normalisés s'explique ainsi par le fait :

- Qu'il est important que différents opérateurs ou clients puissent communiquer entre eux qu'ils soient sur les réseaux mobile (codeurs utilisés : AMR, AMR-WB du 3GPP), ou fixe (codeurs utilisés : G.711, G.729.1, G.729 ou G.722 de l'ITU-T).
- Que nous avons une première évaluation de leur qualité perçue grâce aux différents tests réalisés par les grands organismes de normalisation tels que l'ITU, le 3GPP ou l'ETSI.
- Que dans un premier temps, cette solution sera la plus facile à mettre en œuvre avant l'arrivée de nouvelles techniques ou de nouveaux codeurs.

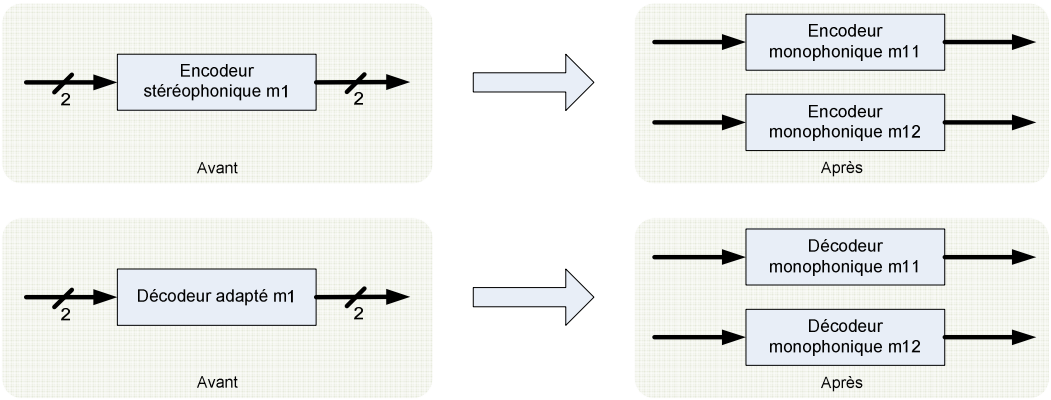

Figure 4.5 Illustration de la transformation pour les blocs encodeur et décodeur de la conférence audio centralisée avec pont mixeur

Cette méthode est celle que nous avons choisie pour transporter du contenu binaural d'un pont mixeur vers un terminal. Ce choix entraîne donc des modifications illustrées Figure 4.5 pour la conférence audio centralisée avec pont mixeur.

Le but de nos tests est de déterminer dans quelle mesure il est possible de transporter et de restituer un contenu binaural avec les codeurs de parole actuels en termes de qualités (audio et spatialisation). La section suivante va présenter les tests choisis pour ces évaluations.

4.1.1.4 Justifications des différents tests retenus dans la cadre de la validation de la conférence audio spatialisée

Nous avons décidé de séparer nos tests suivant que le contenu est mono-locuteur ou multi-locuteur. Il nous a semblé en effet important de bien considérer séparément les deux types de contenu sachant qu'ils pouvaient avoir un impact sur les résultats de qualité obtenus par des codeurs monophoniques conversationnels. Nous pouvons ainsi bien séparer les effets issus de la synthèse binaurale et ceux provenant du nombre de locuteurs.

Nos trois premiers tests ont pour objectif d'évaluer la qualité audio des différents codeurs sélectionnés suivant les différentes configurations de conférence audio et en présence d'un seul locuteur (condition la plus fréquente en conférence audio).

Notre quatrième test a pour but d'évaluer la qualité audio et la qualité de la spatialisation suite à un codage dual-mono sur un contenu binaural multi-locuteur. Nous avons décidé d'effectuer un test de qualité de spatialisation uniquement sur un contenu multi-locuteur afin que les sujets puissent comparer les différences entre les positions des locuteurs d'une condition de spatialisation à une autre.

4.1.1.4.1 Tests de qualité audio pour contenu mono-locuteur

Nous cherchons à évaluer l'impact global de l'ensemble de la chaîne de traitement pour chacune des deux configurations vues ci-dessus. En conséquence, le test à effectuer consisterait à se placer dans la position de l'auditeur en bout de chaîne qui juge la qualité audio des signaux spatialisés. Cependant en effectuant uniquement ce test et en analysant la Figure 4.1, il en ressort que deux étages peuvent dégrader la qualité sans que nous puissions dire si l'un des deux est plus impactant que l'autre :

- Le premier étage, partant de la prise de son au niveau des émetteurs et allant jusqu'au mixage spatialisé inclus, est similaire à la configuration distribuée de la conférence. Il sera donc étudié pendant le test des configurations globales (section 4.4).
- Le second étage partant de la sortie du mixage spatialisé et allant jusqu'au récepteur impacte *a priori* la qualité audio du fait de l'encodage dual-mono. En effet, les deux canaux étant corrélés, le codage dual-mono est susceptible de dégrader la qualité du signal. Il convient donc d'étudier cet étage en y injectant non pas un signal ayant déjà subi des altérations de qualité

dues à un premier encodage-décodage mais un signal spatialisé sans défaut. Il sera ainsi possible d'évaluer qualitativement l'impact de cet étage.

De même, en analysant la Figure 4.2 et sachant que les tests réalisés dans les grands organismes de normalisation sont réalisés en écoute monaurale (sur une seule oreille), il semble important de mesurer l'impact d'une écoute diotique (même contenu sur les deux oreilles) sur des fichiers ayant subi un simple traitement avec un encodage et un décodage. C'est cette étude que nous avons décidé de mener en premier, afin de nous donner une référence avec une écoute sur deux oreilles avant les tests de spatialisation.

La liste des études à mener est la suivante :

- L'impact d'une écoute diotique comparée à une écoute standard monaurale sur la qualité de contenus monauraux (donc non spatialisés) encodés puis décodés. Voir section 4.2.
- L'impact du codage dual-mono sur la qualité de contenus binauraux. Voir section 4.3.
- Plus globalement, l'impact des deux configurations de conférence audio sur la qualité. Voir section 4.4.

4.1.1.4.2 Test de qualité audio et qualité de spatialisation pour contenu multi-locuteur

Ce test (voir section 4.5) vise à évaluer la qualité audio et la qualité de spatialisation ressenties par les sujets lors de la diffusion d'un contenu binaural multi-locuteur. Nous testerons uniquement le second étage de la conférence centralisée avec pont mixeur, avec un contenu binaural initial qui n'aura pas subi de dégradation.

Nous n'évaluerons pas les autres configurations en multi-locuteur car la spatialisation étant la dernière étape du traitement de la conférence audio distribuée et du premier étage de la conférence audio centralisée, le contenu final restitué à un auditeur est la somme de contenus mono-locuteur dont nous aurons déjà testé la qualité dans le test de la section 4.4. La qualité audio finale de la conférence audio distribuée ne devrait donc pas être impactée par le nombre de locuteurs.

De plus, la qualité de la spatialisation sera optimale puisqu'elle n'aura pas subi de dégradation.

4.1.2 Liste des codeurs choisis et présentation succincte de leur fonctionnement

Les études sont menées en narrowband et en wideband avec des codeurs conversationnels normalisés. Les codeurs, que nous avons choisis pour ces études, sont issus de différentes catégories :

- Les codeurs références du narrowband et du wideband que sont respectivement le codeur ITU G.711 [42] et le codeur ITU G.722 [42].
  - Le codeur ITU G.711 est un codeur narrowband fonctionnant à une fréquence d'échantillonnage de 8 kHz pour un débit fixe de 64 kbits/s. Sa technique de codage repose sur le principe du Pulse Code Modulation (PCM), qui est un processus selon lequel le signal est échantillonné, puis chaque échantillon quantifié indépendamment des autres échantillons et converti par codage en une valeur numérique. Dans le cadre de ce codeur, l'échelle de quantification est non linéaire, permettant de diminuer le rapport signal-sur-bruit de l'erreur de quantification pour les sons de faible amplitude. Sa complexité est évaluée à 0.3 WMOPS (Weighted Millions of Opérations per Second), la place nécessaire en RAM à 0.002 kwords (un word est un mot de 16 bits) et la place nécessaire en ROM à 0.05 kwords. Le calcul en WMOPS est basé sur les opérations requises pour effectuer l'encodage et le décodage, en assignant à chaque opération une pondération basée sur une mesure du temps relatif pris pour effectuer chaque type d'opération (e.g. addition, multiplication etc.).
  - Le codeur ITU G.722 est un codeur wideband fonctionnant à une fréquence d'échantillonnage de 16 kHz pour un débit fixe de 64 kbits/s. Sa technique de codage repose sur l'algorithme nommé Adaptive Differential Pulse Code Modulation (ADPCM) appliqué

sur les bandes de fréquence 0 – 4 kHz et 4 kHz – 8 kHz. Cet algorithme permet de réduire le débit binaire par prédiction adaptative et quantification adaptative. Sa complexité est évaluée à 9 WMOPS, la place nécessaire en RAM à 0.256 kwords et la place nécessaire en ROM à 0.05 kwords.

- Des codeurs dédiés à un environnement mobile que sont le codeur narrowband 3GPP AMR [3] et le codeur wideband 3GPP AMR-WB [4].
  - Le codeur 3GPP AMR (Adaptive Multi-Rate) est un codeur narrowband fonctionnant à différents débits fixes. Il utilise différentes techniques telles que l'ACELP (Algebraic Code Excited Linear Prediction), le DTX (Discontinuous Transmission), la VAD (Voice Activity Detection) et le CNG (Comfort Noise Generation). Sa complexité est évaluée entre 14.2 WMOPS (au débit 4.75 kbits/s) et 16.3 WMOPS (au débit 12.2 kbits/s), la place nécessaire en RAM à 4.8 kwords et la place nécessaire en ROM à 15 kwords.
  - Le codeur 3GPP AMR-WB (Adaptive Multi-Rate-WideBand) est un codeur wideband fonctionnant à différents débits fixes. Il s'agit d'un encodage AMR auquel on a élargi la bande des fréquences. Dans les fréquences supérieures à 3.5 kHz, le signal est encodé avec très peu de bits, et à la reconstruction, une méthode d'extension de bande est utilisée. Sa complexité est évaluée entre 28.29 WMOPS (au débit 12.65 kbits/s) et 35.97 WMOPS (au débit 23.85 kbits/s), la place nécessaire en RAM à 6.5 kwords et la place nécessaire en ROM à 16 kwords.
- Un codeur scalable, ou à débit variable (et non plus fixe), narrowband et wideband qu'est le codeur ITU-T G.729.1 [46]. Sa complexité est évaluée entre 18.86 WMOPS (au débit 8 kbits/s) et 35.79 WMOPS (au débit 32 kbits/s), la place nécessaire en RAM à 8.7 kwords et la place nécessaire en ROM à 40.5 kwords. Ce codeur se différencie des autres par le fait que les données codées se présentent sous forme de couches empilées. En incluant plus ou moins de couches suivant les contraintes dues aux réseaux, on fait ainsi varier le débit et par conséquence la qualité. Il peut fonctionner ainsi à différents débits au cours du temps (entre 8 kbits/s et 32 kbits/s) suivant le choix effectué par l'application, et utilise quatre briques technologiques :
  - Le CELP (Code Excited Linear Prediction) pour un contenu narrowband et un débit à 8 et 12 kbits/s. Sa complexité est évaluée à 18.86 WMOPS à 8 kbits/s et 21.70 WMOPS à 12 kbits/s.
  - L'extension de bande qui, additionnée à la technique précédente, permet d'obtenir un contenu wideband à 14 kbits/s,
  - L'algorithme MDCT (Modified Discrete Cosine Transform) pour un contenu wideband à des débits entre 16 et 32 kbits/s.
  - La transmission et utilisation d'information supplémentaire FEC (Forward Error Correction) pour corriger la perte de trames et donc limiter la perte de qualité.

4.2 Etude de l'impact sur la qualité perçue d'une écoute diotique par rapport à une écoute monaurale

Le but de ce test, dont les résultats ont été publiés à l'AES 123 à New-York [64], est d'étudier l'impact du type d'écoute, monaurale ou diotique, sur les jugements de qualité audio pour des signaux de parole traités par des codeurs VoIP.

Ce problème est loin d'être anodin, car lors d'une conférence audio classique sur IP, il est fréquent que les écoutes se fassent au casque (écoute diotique). Or à l'ITU, la plupart des tests sont effectués au combiné téléphonique (écoute monaurale).

Lors de notre étude, deux tests ont été réalisés, l'un avec une qualité narrowband et l'autre avec une qualité wideband. Les codeurs utilisés sont ceux présentés ci-dessus.

4.2.1 Stimuli

Les 8 échantillons de test sont extraits de la base de données de France Télécom et durent 8 secondes. Ils sont constitués de deux phrases espacées par un silence, et sont échantillonnés à 16 kHz, pour une bande de fréquences de 8 kHz. Ils sont énoncés par quatre locuteurs différents (2 femmes et 2 hommes), prononçant chacun deux échantillons (tous différents deux à deux). Chaque échantillon est traité par l'ensemble des conditions des tests.

4.2.2 Conditions de test narrowband et wideband

Les traitements narrowband et wideband sont explicités dans les deux sections suivantes. Leur rôle est de simuler une transmission monophonique de parole sur un réseau en utilisant un des codeurs conversationnels normalisés présentés ci-dessus.

Une procédure aléatoire a été mise en place pour générer, pour chaque échantillon initial, une séquence de trames effacées. Pour chaque traitement, des pertes de trames (PdT) à 0%, 3% et 6% ont été introduites, les algorithmes de correction de perte de trame de chaque codeur étant utilisés. Tous les codeurs subiront au même endroit les pertes de trames pour un même échantillon de test.

Il est à noter que la correction de perte de trames du G.711 est la seule à fonctionner sur des trames de 10 ms (contre 20 ms pour les autres). Pour que les pertes de trames soient de la même longueur temporelle, deux trames sont effacées au lieu d'une pour le G.711.

4.2.2.1 Conditions narrowband

Le traitement appliqué aux fichiers est illustré Figure 4.6. Les briques utilisées (*Filtrage MIRS16, Filtrage RXIRS16, Egalisation à -26 dB$_{ov}$* et *Sur/Sous-échantillonnage*) sont normalisées à l'ITU, et permettent de pré-traiter les fichiers. Le bloc principal a pour nom *Traitement avec ou sans perte de trames*.

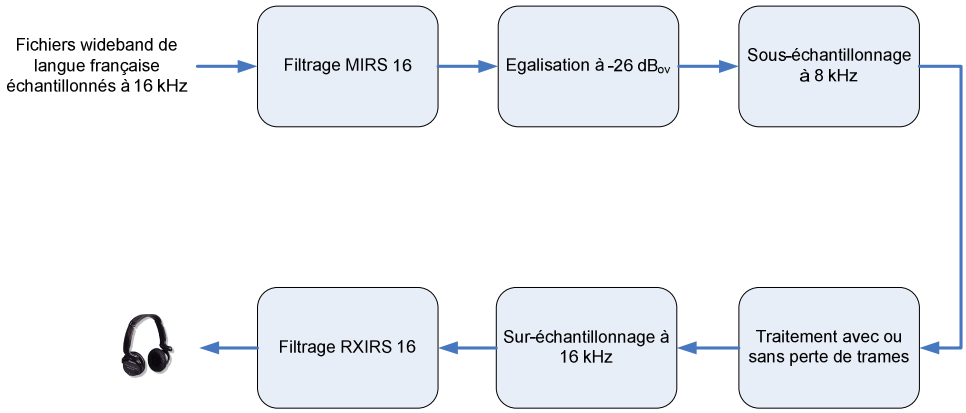

Figure 4.6 Traitement appliqué aux fichiers pour le test monaural/diotique narrowband

Les filtrages MIRS 16 (*Modified Intermediate Reference System*) [42] et RXIRS 16 (*Receive-Side IRS*) [42] permettent de simuler des appareils de prise de son et de restitution numériques narrowband pour des signaux de fréquence d'échantillonnage initiale de 16 kHz.

Les blocs de sous-échantillonnage et sur-échantillonnage permettent de ré-échantillonner des contenus à des fréquences d'échantillonnage respectivement de 8 et 16 kHz, afin de pouvoir être traités par les blocs suivants. L'égalisation à -26 dB$_{ov}$ permet, comme son nom l'indique, d'égaliser en énergie les fichiers de départ pour que les sujets aient la même perception sonore en niveau quel que soit le fichier. Ces trois traitements sont effectués grâce aux filtres définis dans [42].

Le bloc *Traitement avec ou sans perte de trames* effectue les encodage-décodage ainsi que les pertes de trame sur les fichiers pré-traités (par filtrage, égalisation et sous-échantillonnage), suivant les conditions listées dans le Tableau 4.1. La condition 1 est notre condition de référence sans dégradation, aussi appelée *Direct* par la suite.

Numéro de la condition	Codeur	Pourcentage de perte de trame
1	Aucun	0
2	G.711	0
3		3
4		6
5	G.729.1 à 8 kbits/s	0
6		3
7		6
8	G.729.1 à 12 kbits/s	0
9		3
10		6
11	AMR à 4.75 kbits/s	0
12		3
13		6
14	AMR à 12.2 kbits/s	0
15		3
16		6

Tableau 4.1 Liste des conditions du test monaural-diotique narrowband

Les fichiers sont ensuite sur-échantillonnés et filtrés avant diffusion sur casque de manière monaurale ou diotique.

4.2.2.2 Conditions wideband

Le traitement appliqué aux fichiers est illustré Figure 4.7. Les briques *Filtrage P341* et *Egalisation à -26 dB*$_{ov}$ sont normalisées à l'ITU et permettent de pré-traiter les fichiers avant le bloc *Traitement avec ou sans perte de trames*.

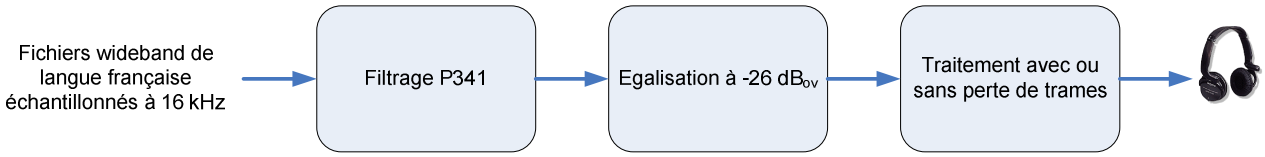

Figure 4.7 Traitement appliqué aux fichiers pour le test monaural/diotique wideband

Le filtrage *P.341* [42] permet de simuler l'appareil de prise de son et de restitution numériques. L'égalisation à -26 dB$_{ov}$ est effectuée grâce à un filtre défini dans [42].

Le bloc *Traitement avec ou sans perte de trames* effectue les encodage-décodage ainsi que les pertes de trame sur les fichiers pré-traités (par filtrage et égalisation), suivant les conditions listées Tableau 4.2. La condition 1 est notre condition de référence sans dégradation, aussi appelée *Direct* par la suite.

Numéro de la condition	Codeur	Pourcentage de perte de trame
1	Aucun	0
2	G.729.1 à 16 kbits/s	0
3		3
4		6
5	G.729.1 à 32 kbits/s	0
6		3
7		6
8	G.722	0
9		3
10		6
11	AMR-WB à 12.65 kbits/s	0
12		3

13		6
14	AMR-WB à 23.85 kbits/s	0
15		3
16		6

Tableau 4.2 Liste des conditions du test monaural-diotique wideband

Les fichiers sont ensuite disponibles pour être diffusés sur casque de manière monaurale ou diotique.

4.2.3 Niveau d'écoute

A un niveau sonore physique donné (en dB Sound Pressure Level - SPL) par canal, l'écoute sur deux oreilles amplifie l'intensité subjective par rapport à l'écoute sur une oreille.

Pour des sons purs, il est montré qu'un son présenté à une oreille a une sonie globalement deux fois inférieure au même son présenté aux deux oreilles [29]. On rappelle que l'échelle des sones est une échelle d'intensité subjective établie en demandant aux sujets de choisir des nombres proportionnels à la sonie des sons qui leur sont présentés [16]. Ainsi, un son, qui a une sonie de 1 sone (une sinusoïde de 40 dB SPL à 1000 Hz), donne l'impression d'être deux fois moins fort qu'un son qui a une sonie de 2 sones.

Afin de comparer des contenus comparables, les niveaux d'écoute doivent donc être ajustés de façon à ce que les sujets aient la même impression de niveau quel que soit le type d'écoute. Deux choix sont possibles :

- Le premier choix est une atténuation ou une amplification en dB SPL par bande d'octave [16].
- Le second choix est une atténuation ou une amplification en dB SPL, de manière uniforme sur l'ensemble de la bande des fréquences.

C'est cette dernière solution qui a été retenue. Ce choix se justifie par l'application finale que l'on souhaite. En conférence audio en VoIP, la seule possibilité laissée à l'utilisateur est le contrôle du volume global. Libre à lui d'écouter le contenu sur une ou deux oreilles, et s'il change de l'un pour l'autre, il aura tendance à ajuster le volume en conséquence.

Une fois la méthode d'ajustement choisie, sachant que le contenu monaural est joué à 79 dB SPL (en niveau de pression acoustique [16]) à l'ITU-T, il convient de déterminer l'atténuation (en dB SPL) à appliquer sur chaque canal du contenu diotique pour que les sujets aient la même impression de niveau d'écoute.

Il est trouvé dans [73] que pour des bruits blancs de bande 100 Hz -500 Hz diffusés à 80 dB, une atténuation de 10 dB SPL doit être appliquée à l'écoute diotique (cf. Figure 4.8) pour avoir la même sensation de niveau. Cette valeur se retrouve dans [16] et a été validée par des pré-écoutes.

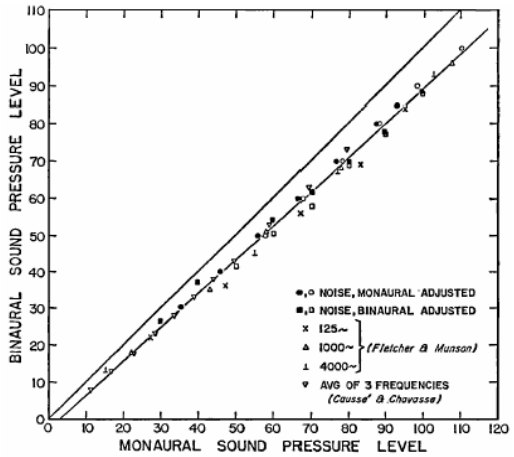

Figure 4.8 Correspondance entre les niveaux d'écoute en dB SPL d'une écoute monaurale par rapport à une écoute binaurale pour un bruit [73]

Notre choix se porte donc sur une atténuation de 10 dB SPL. Les niveaux d'écoute sont établis à 79 dB SPL pour l'écoute monaurale et 69 dB SPL par canal pour l'écoute diotique.

4.2.4 Les auditeurs

Les auditeurs, au nombre de 32 par test, sont recrutés au hasard parmi la population selon la recommandation de la méthodologie de test. Aucune procédure de rejet n'a été instaurée.

4.2.5 Procédure expérimentale

La méthode *Absolute Category Rating* (ACR) [47] a été choisie pour les deux tests car elle permet d'évaluer la qualité audio perçue des échantillons écoutés. Les sujets donnent une note de qualité pour chaque échantillon sur l'échelle absolue illustrée par le Tableau 4.3. Les notes moyennes obtenues sont appelées *Mean Opinion Score* (MOS).

Qualité de la parole	Note
Excellente	5
Bonne	4
Passable	3
Médiocre	2
Mauvaise	1

Tableau 4.3 Echelle de note ACR

Les deux tests se sont déroulés dans une salle d'écoute acoustiquement isolée, suivant les recommandations de la norme P.800 [47].

Pour chaque test, les sujets ont passé deux sessions avec une session par type d'écoute (monaurale / diotique) et avec un apprentissage en début de chaque session. La moitié des sujets a passé tout d'abord la session monaurale, puis la session diotique. L'autre moitié a effectué l'inverse, cela afin d'annuler un éventuel effet d'ordre de passation des deux sessions.

Les 32 sujets ont été séparés en 4 groupes de 8 (G1-G4). L'ordre de présentation des conditions était aléatoire pour chaque groupe. Les 128 fichiers (8 échantillons x 16 conditions), écoutés en monaural et en diotique selon la session, sont présentés de manière aléatoire aux sujets, suivant le groupe dans lequel ils se trouvent. Seul le type d'écoute et le volume associé changent d'une session à l'autre. Le contenu n'est donc pas spatialisé puisque les deux écoutes partent du même contenu qui est monophonique.

Il aurait été évidemment souhaitable de pouvoir comparer directement dans une même session les deux types d'écoute. Cependant, cela n'est évidemment pas réalisable car cela signifierait qu'un sujet doive continuellement rendre libre ou non sa deuxième oreille, suivant l'échantillon testé. Le sujet perdrait de sa concentration à réaliser ce genre de manipulation. Une solution alternative consistant à rester en écoute diotique tout le temps et à inclure un bruit de confort sur l'oreille supposée laissée libre a été envisagée. Cependant, cela a été rejeté du fait de la difficulté à choisir un bruit de confort adapté.

Il a été nécessaire de choisir un casque particulier permettant de laisser libre une oreille pour l'écoute monaurale. Le choix s'est porté sur le casque Sennheiser HD25, qui possède une réponse plate dans la zone de fréquence qui nous intéresse [50 Hz - 7000 Hz].

4.2.6 Résultats narrowband

Les notes MOS obtenues par chaque codeur, aux différents taux de PdT et selon les deux écoutes, sont illustrées sur les Figure 4.9, Figure 4.10 et Figure 4.11. Sur chaque figure, l'histogramme de gauche et l'histogramme de droite correspondent aux notes de qualité obtenues respectivement en écoute monaurale et en écoute diotique.

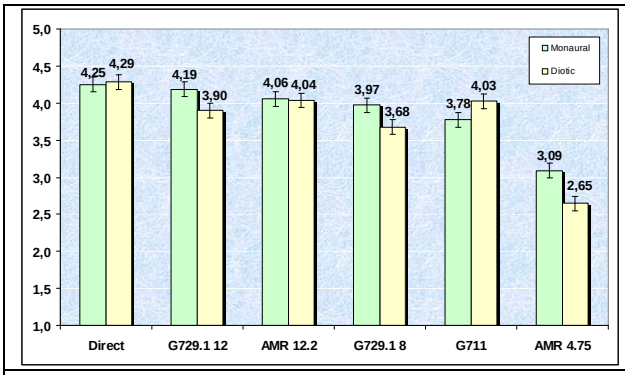

Figure 4.9 MOS pour chaque codeur narrowband en écoute monaurale et en écoute diotique à 0% de PdT

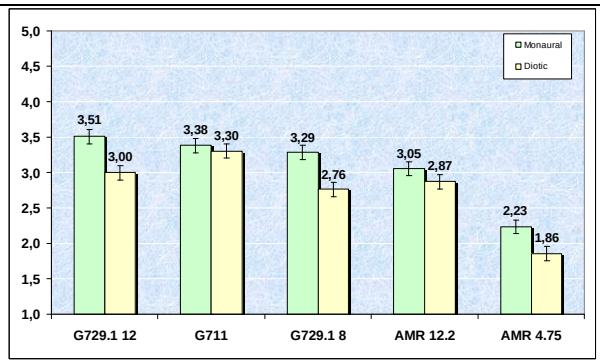

Figure 4.10 MOS pour chaque codeur narrowband en écoute monaurale et en écoute diotique à 3% de PdT

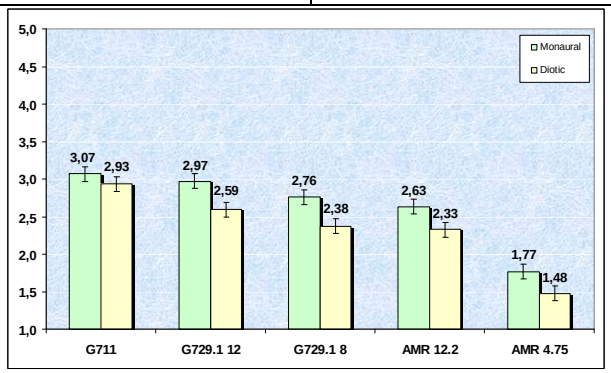

Figure 4.11 MOS pour chaque codeur narrowband en écoute monaurale et en écoute diotique à 6% de PdT

Sur ces figures, il apparaît que le type d'écoute a généré des différences de notation. Les jugements obtenus en écoute diotique semblent globalement inférieurs à ceux obtenus en écoute monaurale. On remarque des différences assez nettes pour de nombreux codeurs entre les deux types d'écoute (par exemple à 3 ou 6 % de PdT). Il est à noter que seuls la référence et le G.711 sans perte de trame semblent avoir de meilleurs résultats en écoute diotique qu'en écoute monaurale.

La condition Direct et le codeur AMR 4.75 sont sans surprise les bornes haute et basse des résultats. On rappelle que le direct n'est pas codé et que l'AMR 4.75 présente le plus bas débit du test. On remarque aussi la dégradation de la qualité perçue avec la diminution des débits.

Enfin, la perte de trame a eu de même un impact sur la qualité perçue, ce qui n'est pas surprenant au vu des pourcentages utilisés.

Ces effets sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles en prenant en compte les facteurs du test (Type d'écoute, Codeur, Perte de trames, Locuteur, Echantillon, Ordre d'écoute). Les résultats de l'ANOVA (effets principaux) sont donnés Tableau 4.4 :

Facteur	Degrés de liberté	F-ratio	p = Significativité
Ecoute	1	157.94	0.00
Codage	5	509.24	0.00
Perte de trames	2	1564.14	0.00
Locuteur	3	40.39	0.00
Echantillon	1	12.99	0.00
Ordre d'écoute	1	0.06	0.81

Tableau 4.4 Effets des différents facteurs du test monaural/diotique narrowband

L'ANOVA a aussi montré la non-significativité de l'ordre d'écoute des sujets sur la qualité perçue (session monaurale puis diotique, ou, session diotique puis monaurale). Nous pouvons aussi relever que les locuteurs et les échantillons ont eu un effet. Nous avons pu en effet constater qu'un échantillon d'un des locuteurs avait des résultats significativement différents des autres échantillons tous locuteurs confondus.

Afin de savoir pour quels codeurs l'effet de l'écoute a été significatif, nous avons comparé les deux types d'écoute grâce à des t-tests [48]. En prenant N = 8192 et MSE (Mean Square Error) comme étant l'erreur quadratique moyenne calculée dans l'ANOVA, on peut définir l'intervalle de confiance à 95 % comme suit :

$$CI = 1.96 \times \sqrt{\frac{MSE}{N}}$$

Le rapport MSE sur N est le meilleur estimateur de la variance. Cet intervalle de confiance est le même quel que soit le codeur, l'écoute ou la PdT. MSE étant égal à 0.65, l'intervalle de confiance vaut 0.0991. Les résultats obtenus sont les suivants :

- A 0% de PdT :
  - Le Direct et l'AMR 12.2 sont équivalents en écoute monaurale et diotique.
  - Les autres codeurs sont significativement différents en écoute monaurale et diotique. Nous notons que les résultats obtenus par le codeur G.711 en écoute diotique sont significativement meilleurs que ceux obtenus en écoute monaurale, contrairement aux autres codeurs.
- A 3% de PdT :
  - Le G.711 est équivalent en écoute monaurale et diotique.
  - Les autres codeurs sont significativement différents en écoute monaurale et diotique. Pour tous ces codeurs, les résultats obtenus en écoute diotique sont moins bons qu'en écoute monaurale.
- A 6% de PdT, tous les codeurs sont significativement différents en écoute monaurale et diotique. Pour tous ces codeurs, les résultats obtenus en écoute diotique sont moins bons qu'en écoute monaurale.

Les différentes équivalences, issues des t-tests, en écoute monaurale et diotique pour une PdT donnée sont données Tableau 4.5, Tableau 4.6 et Tableau 4.7 ci-dessous. Ces tableaux donnent à la fois les différentes classes d'équivalence, mais aussi l'ordonnancement entre les codeurs. Par exemple, l'ordonnancement des codeurs dans le Tableau 4.5 est : Direct, G.729.1 à 12 kbits/s, AMR à 12.2 kbits/s, G.729.1 à 8 kbits/s, G.711 et AMR à 4.75 kbits/s. Néanmoins la condition Direct est équivalente au codeur G.729.1 à 12 kbits/s, etc.

N°	Ecoute monaurale	Ecoute diotique
1	Direct ⇔ G.729.1 12	Direct
2	G.729.1 12 ⇔ AMR 12.2	AMR 12.2 ⇔ G.711
3	AMR 12.2 ⇔ G.729.1 8	G.711 ⇔ G.729.1 12
4	G.711	G.729.1 8
5	AMR 4.75	AMR 4.75

Tableau 4.5 Comparaison des classements entre les écoutes monaurale et diotique des codeurs narrowband à 0% de perte de trames

N°	Ecoute monaurale	Ecoute diotique
1	G.729.1 12 ⇔ G.711	G.711
2	G.711 ⇔ G.729.1 8	G.729.1 12 ⇔ AMR 12.2
3	AMR 12.2	AMR 12.2 ⇔ G.729.1 8
4	AMR 4.75	AMR 4.75

Tableau 4.6 Comparaison des classements entre les écoutes monaurale et diotique des codeurs narrowband à 3% de perte de trames

N°	Ecoute monaurale	Ecoute diotique
1	G.711 ⇔ G.729.1 12	G.711
2	G.729.1 8 ⇔ AMR 12.2	G.729.1 12
3	AMR 4.75	G.729.1 8 ⇔ AMR 12.2
4		AMR 4.75

Tableau 4.7 Comparaison des classements entre les écoutes monaurale et diotique des codeurs narrowband à 6% de perte de trames

Nous pouvons noter que les différences notables obtenues entre les deux types d'écoute pour certains codeurs ont abouti à des ordonnancements et des équivalences différents, notamment pour les pertes de paquets à 0% et 3%. L'impact est moindre à 6% de PdT, mais il semble que les sujets aient eu plus de facilité à différencier, avec l'écoute diotique, la qualité des codeurs (G711 significativement différent du G.729.1 à 12 kbits/s).

4.2.7 Résultats wideband

Les notes MOS obtenues par chaque codeur, aux différents taux de PdT et selon les deux écoutes, sont illustrées sur les Figure 4.12, Figure 4.13 et Figure 4.14. Sur chaque figure, l'histogramme de gauche et l'histogramme de droite correspondent aux notes obtenues respectivement en écoute monaurale et en écoute diotique.

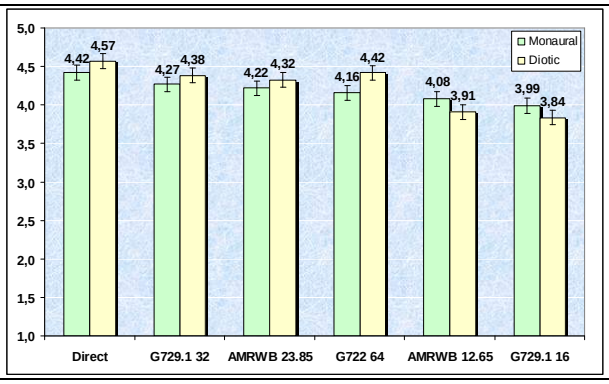

Figure 4.12 MOS pour chaque codeur wideband en écoute monaurale et en écoute diotique à 0% de PdT

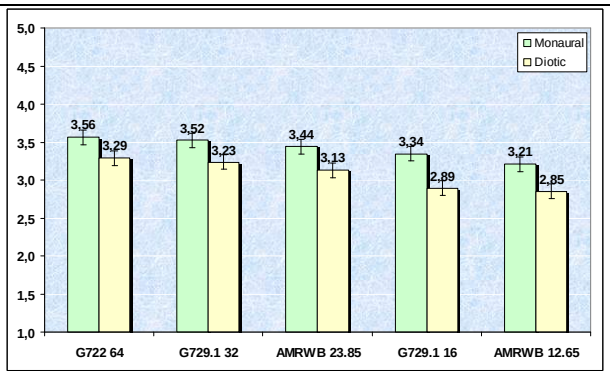

Figure 4.13 MOS pour chaque codeur wideband en écoute monaurale et en écoute diotique à 3% de PdT

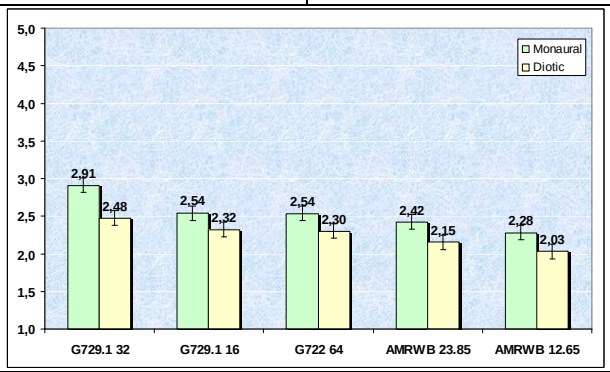

Figure 4.14 MOS pour chaque codeur wideband en écoute monaurale et en écoute diotique à 6% de PdT

Tout comme en narrowband, il apparaît que le type d'écoute a généré des différences de notation. Il est intéressant de remarquer que sans perte de trame, les codeurs bas-débit (G.729.1 à 16 kbits/s ou AMR-WB à 12.65 kbits/s) sont moins bien notés en écoute monaurale qu'en écoute binaurale, contrairement aux codeurs haut-débit (G.729.1 à 32 kbits/s ou AMR-WB à

23.85 kbits/s). En présence de PdT, les résultats en écoute diotique sont toujours plus mauvais qu'en écoute monaurale.

On retrouve sans surprise la condition Direct avec la meilleure moyenne ainsi qu'une dégradation de la qualité perçue avec la diminution des débits. De même, l'introduction de pertes de trames fait logiquement baisser la qualité perçue.

Ces effets sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles en prenant en compte les facteurs du test (Type d'écoute, Codeur, Perte de trames, Locuteur, Echantillon, Ordre d'écoute). Les résultats de l'ANOVA sont donnés dans le Tableau 4.8 :

Facteur	Degrés de liberté	F-ratio	p = Significativité
Ecoute	1	94.82	0.00
Codage	5	69.75	0.00
Perte de trames	2	3244.08	0.00
Locuteur	3	206.36	0.00
Echantillon	1	41.08	0.00
Ordre d'écoute	1	45.57	0.00

Tableau 4.8 Effet des différents facteurs du test monaural/diotique wideband

Cette analyse a montré un effet de l'ordre d'écoute. Cet effet (session monaurale puis session diotique, ou, session diotique puis session monaurale) s'explique simplement. Un des deux groupes a utilisé différemment l'échelle de notation, notant plus sévèrement les mauvaises conditions et moins durement les bonnes conditions en écoute diotique qu'en écoute monaurale.

Afin de savoir pour quels codeurs l'effet de l'écoute a été significatif, nous avons comparé les deux types d'écoute grâce à des t-tests [48]. Les résultats obtenus sont les suivants :

- A 0% de PdT :
  - Le G.729.1 32 et l'AMR-WB 23.85 sont équivalents en écoute monaurale et diotique.
  - Les autres codeurs et le Direct sont significativement différents en écoute monaurale et diotique. Seuls le Direct et le G.722 sont jugés significativement meilleurs en écoute diotique.
- A 3% de PdT, tous les codeurs sont significativement différents en écoute monaurale et diotique. Pour tous ces codeurs, les résultats obtenus en écoute diotique sont moins bons qu'en écoute monaurale.
- A 6% de PdT, tous les codeurs sont significativement différents en écoute monaurale et diotique. Pour tous ces codeurs, les résultats obtenus en écoute diotique sont moins bons qu'en écoute monaurale.

Afin de classer les différents codeurs pour un type d'écoute et une PdT donnés, des t-tests [48] ont été effectués. La procédure est la même que celle explicitée en 4.2.6 pour le t-test narrowband. Pour ce test, les valeurs obtenues sont 0.61281 pour le MSE et en conséquence 0.096 pour l'intervalle de confiance.

Les différentes équivalences en écoute monaurale et diotique pour une PdT donnée sont données Tableau 4.9, Tableau 4.10 et Tableau 4.11 ci-dessous.

N°	Ecoute monaurale	Ecoute diotique
1	Direct	Direct
2	G.729.1 32 ⇔ AMR-WB 23.85 ⇔ G.722	G.722 ⇔ G.729.1 32 ⇔ AMR-WB 23,85
3	G.722 ⇔ AMR-WB 12.65	AMR-WB 12.65 ⇔ G.729.1 16
4	AMR-WB 12.65 ⇔ G.729.1 16	

Tableau 4.9 Comparaison des classements entre les écoutes monaurale et diotique des codeurs wideband à 0% de perte de trames

N°	Ecoute monaurale	Ecoute diotique
1	G.722 ⇔ G.729.1 32 ⇔ AMR-WB 23.85	G.722 ⇔ G.729.1 32
2	AMR-WB 23.85 ⇔ G.729.1 16	G.729.1 32 ⇔ AMR-WB 23.85
3	G.729.1 16 ⇔ AMR-WB 12.65	G.729.1 16 ⇔ AMR-WB 12.65

Tableau 4.10 Comparaison des classements entre les écoutes monaurale et diotique des codeurs wideband à 3% de perte de trames

N°	Ecoute monaurale	Ecoute diotique
1	G.729.1 32	G.729.1 32
2	G.729.1 16 ⇔ G.722 ⇔ AMR-WB 23.85	G.729.1 16 ⇔ G.722
3	AMR-WB 12.65	AMR-WB 23.85 ⇔ AMR-WB 12.65

Tableau 4.11 Comparaison des classements entre les écoutes monaurale et diotique des codeurs wideband à 6% de perte de trames

Nous pouvons remarquer que les différences pourtant significatives obtenues pour la plupart des codeurs entre l'écoute diotique et l'écoute monaurale ont finalement abouti à des ordonnancements et équivalences quasi-identiques.

Il est à noter que nous retrouvons les résultats connus en matière de correction de paquets qui stipule que la correction de paquets du G.729.1 à 32 kbits/s est supérieure ou égale à celle du G.722 [24].

4.2.8 Discussion et perspectives

Tout d'abord, notre étude montre que le type d'écoute a eu une influence sur la perception de la qualité des codeurs. De façon générale, en écoute diotique, la qualité est jugée plus sévèrement, notamment lorsque celle-ci est dégradée (soit par la PdT ou la diminution des débits). Une explication possible est que l'écoute diotique permet de mieux percevoir les dégradations. Selon le codeur, la perte de trames, le débit, l'impact est plus ou moins important aboutissant à des ordonnancements différents entre codeurs selon le type d'écoute.

En revanche, une amélioration de qualité perçue en écoute diotique par rapport à l'écoute monaurale est observée pour les conditions de très bonne qualité (les Directs dans le test narrowband et wideband, ou les codeurs de très bonne qualité dans le test wideband). Dans ces contextes, l'écoute diotique met en valeur la bonne qualité de ces codeurs comparée à une écoute monaurale car le confort d'écoute est accru en écoute diotique.

Une exception est à noter en narrowband avec le codeur G.711. Il est plus faible que la plupart des autres codeurs en écoute monaurale mais il obtient de meilleurs résultats en écoute diotique. Une des pistes est que la méthode d'ajustement des niveaux (atténuation uniforme sur l'ensemble des fréquences) a pu jouer un rôle notamment sur le codeur G.711 et la perception de son bruit blanc de quantification. Nous avons mesuré en conséquence le niveau sonore de la parole et du silence en narrowband dans les mêmes conditions que le test. Les mesures ont été effectuées grâce à un mannequin HATS et leurs résultats sont donnés dans le Tableau 4.12.

Pour la parole, on retrouve bien globalement les valeurs de 79 dB SPL et de 69 dB SPL. On obtient les mêmes résultats en utilisant les dBA, ceux-ci prenant en compte la sensibilité de l'oreille par rapport aux fréquences.

Pour le bruit du G.711, il a été mesuré au niveau du bruit ambiant, c'est-à-dire à 50 dB SPL en écoute monaurale et sur chaque oreille en écoute diotique. Par contre en dBA, on remarque une différence de 6.5 dBA entre les deux types d'écoute. La valeur de 31 dBA pour l'écoute diotique se retrouve par la somme logarithmique des deux bruits indépendants présents : celui du G.711 de l'ordre de 27.5 dBA (après l'égalisation) et celui du bruit ambiant de l'ordre de 28.5 dBA. Ainsi en écoute diotique, le bruit du G.711 est proche du bruit ambiant de la salle donc peu audible. Cela pourrait expliquer le meilleur classement du G.711 en écoute diotique.

	Ecoute monaurale	Ecoute diotique (pour chaque oreille)
Parole seule	~80 dB SPL ~79.5 dBA	~70 dB SPL ~69 dBA
Bruit du G.711 seul	~50 dB SPL ~37.5 dBA	~50 dB SPL ~31 dBA
Bruit ambiant	~49 dB SPL, ~28.5 dBA	

Tableau 4.12 Niveau sonore pour chaque système d'écoute en dB SPL et dBA

Ces tests soulignent donc l'importance du niveau d'écoute. La différence de 10 dB SPL a été justifiée dans notre protocole, mais il est vrai qu'elle est susceptible d'avoir eu un impact notamment sur le codeur G.711 comme on vient de le voir. Il serait intéressant de faire un test en écoute monaurale en effectuant une session à 79 dB SPL, puis le même test avec les mêmes sujets mais à d'autres niveaux d'écoute. Cela permettrait de voir si des changements sont susceptibles d'intervenir dans le classement des codeurs.

Les résultats obtenus dans ces deux tests ne montrent pas une meilleure discrimination de l'écoute diotique par rapport à une écoute monaurale. L'écoute diotique permet en effet de séparer certains codeurs jugés équivalents en écoute monaurale, mais à l'inverse elle permet aussi de rendre équivalents des codeurs qui ne l'étaient pas en écoute monaurale. Cela va à l'encontre des résultats de [49] obtenus pour un même contenu diffusé sur une ou deux oreilles au même niveau (sonie doublée d'un contenu sur l'autre).

Ces résultats nous ont aussi permis d'avoir une information pour les tests de spatialisation à l'angle 0° d'un contenu monophonique encodé-décodé par la suite en dual-mono. En effet pour cet angle, le contenu binaural sera identique sur les deux oreilles et sera donc similaire au cas diotique.

4.3 Evaluation de la qualité de signaux de parole binauraux mono-locuteur encodés-décodés en dual-mono par des codeurs monophoniques de parole

Le but de notre étude est d'évaluer la possibilité de transporter un contenu binaural encodé-décodé en dual-mono, grâce à des codeurs de parole monophoniques normalisés. Nous testons ainsi l'impact sur la qualité audio perçue du second étage de la conférence centralisée illustrée par la Figure 4.1. Nous chercherons aussi à vérifier s'il existe une perception symétrique de la qualité audio entre les demi-plans frontaux gauche et droit.

De même qu'en 4.2, deux tests ont été menés, l'un avec une bande de qualité narrowband et l'autre avec une bande de qualité wideband. Les codeurs utilisés sont ceux présentés dans la section 4.1.2.

4.3.1 Stimuli

Les 4 échantillons de test sont extraits de la base de données de France Télécom et durent 8 secondes. Ils sont constitués de deux phrases espacées par un silence, et sont échantillonnés à 16 kHz, pour une bande de fréquences de 8 kHz. Ils sont énoncés par quatre locuteurs différents (2 femmes et 2 hommes), prononçant chacun un échantillon (tous différents deux à deux). Chaque échantillon est traité par l'ensemble des conditions des tests.

4.3.2 Conditions de test narrowband et wideband

Les traitements narrowband et wideband sont explicités dans les deux sections suivantes. Leur rôle est de simuler une transmission d'un contenu binaural sur un réseau en utilisant un codage dual-mono avec des codeurs conversationnels normalisés. Il est à noter que les HRTFs utilisés

sont symétriques. En conséquence, pour un angle donné, en inversant les deux canaux, on obtient un contenu semblant venir de l'angle opposé.

Pour chaque traitement, des pertes de trames (PdT) à 0%, 3% et 6% ont été à nouveau introduites. Elles ont lieu sur les deux canaux simultanément d'un contenu binaural pour simuler ce qui se passe lors de la transmission d'un contenu stéréo en Voix sur IP.

4.3.2.1 *Conditions narrowband*

Le traitement appliqué aux fichiers est illustré Figure 4.15. Les deux éléments de gestion non présents dans la section 4.2.2.1 (*Séparation/Fusion*) sont deux outils de gestion de fichiers permettant de réunir ou de séparer les deux canaux d'un fichier.

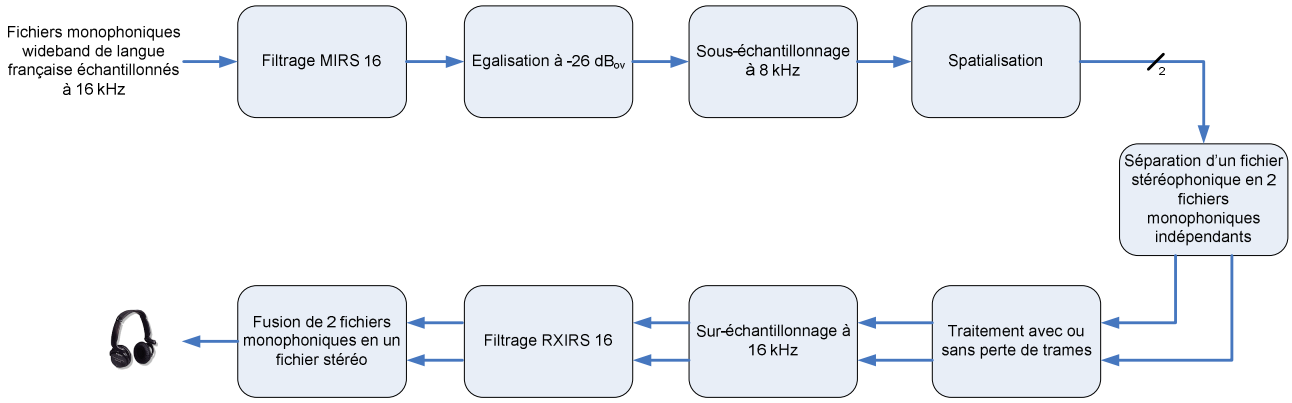

Figure 4.15 Traitement appliqué aux fichiers pour le test narrowband de transport de contenu binaural mono-locuteur

Les filtrages MIRS 16 (*Modified Intermediate Reference System*) [42] et RXIRS 16 (*Receive-Side IRS*) [42] permettent de simuler des appareils de prise de son et de restitution numériques narrowband pour des signaux de fréquence d'échantillonnage initiale de 16 kHz.

Les blocs de sous-échantillonnage et sur-échantillonnage permettent de ré-échantillonner des contenus à des fréquences d'échantillonnage respectivement de 8 et 16 kHz, afin de pouvoir être traités par les blocs suivants. L'égalisation à -26 dB_{ov} permet, comme son nom l'indique, d'égaliser en énergie les fichiers de départ pour que les sujets aient la même perception sonore en niveau quel que soit le fichier. Ces trois traitements sont effectués grâce aux filtres définis dans [42].

Le bloc *Spatialisation* effectue une spatialisation binaurale, selon les 9 angles suivants : 0°, ±20°, ±45°, ±60° et ±90°, d'un contenu monophonique pré-traité (après filtrage, égalisation et sous-échantillonnage) de fréquence d'échantillonnage 8 kHz. Nous avons choisi des valeurs d'angles rapprochées les unes des autres en face de l'auditeur (0°, ±20°, ±45°), car la capacité humaine de localisation et discernement des sources est plus performante en frontal que sur les côtés [13].

Le bloc *Traitement avec ou sans perte de trames* effectue les encodage-décodage dual-mono ainsi que les pertes de trames des contenus binauraux, suivant les conditions listées dans le Tableau 4.13 et la valeur de l'angle choisie dans le bloc de Spatialisation. Les conditions de 1 à 9 sont nos conditions de référence spatialisées, sans encodage-décodage (aussi appelé codeur Ori pour Original par la suite), ni perte de trame.

Numéros des conditions	Codeur	Pourcentage de perte de trame	Angles
1-9	Aucun	0	0°, ±20°, ±45°, ±60° et ±90°
10-18	G.711	0	
19-27		3	
28-36		6	
37-45	G.729.1 à 8 kbits/s	0	
46-54		3	
55-63		6	

64-72	G.729.1 à 12 kbits/s	0	
73-81		3	
82-90		6	
91-99	AMR à 4.75 kbits/s	0	
100-108		3	
109-117		6	
118-126	AMR à 12.2 kbits/s	0	
127-135		3	
136-144		6	

Tableau 4.13 Liste des conditions du test de transport de contenu binaural narrowband

Les 576 fichiers (144 conditions x 4 échantillons) sont ensuite sur-échantillonnés et filtrés avant diffusion sur casque.

4.3.2.2 Conditions wideband

Le traitement appliqué aux fichiers est illustré Figure 4.16.

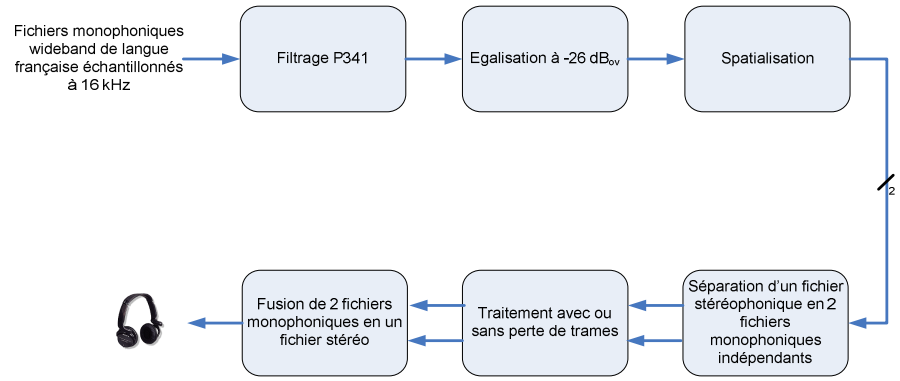

Figure 4.16 Traitement appliqué aux fichiers pour le test wideband de transport de contenu binaural mono-locuteur

Le filtrage *P.341* [42] permet de simuler l'appareil de prise de son et de restitution numériques. L'égalisation à -26 dB_{ov} est effectuée grâce à un filtre défini dans [42].

Le bloc *Spatialisation* effectue une spatialisation binaurale, selon les 9 angles suivants : 0°, ±20°, ±45°, ±60° et ±90°, d'un contenu monophonique de fréquence d'échantillonnage 16 kHz pré-traité (par filtrage et égalisation).

Le bloc *Traitement avec ou sans perte de trames* effectue les encodage-décodage dual-mono ainsi que les pertes de trames des contenus binauraux, suivant les conditions listées Tableau 4.14 et la valeur de l'angle choisie dans le bloc de Spatialisation. Les conditions de 1 à 9 sont nos conditions de référence spatialisées, sans encodage-décodage (aussi appelé codeur Ori pour Original par la suite), ni perte de trame.

Numéros des conditions	Codeur	Pourcentage de perte de trame	Angles
1-9	Aucun	0	0°, ±20°, ±45°, ±60° et ±90°
10-18	G.722	0	
19-27		3	
28-36		6	
37-45	G.729.1 à 16 kbits/s	0	
46-54		3	
55-63		6	
64-72	G.729.1 à 32 kbits/s	0	
73-81		3	
82-90		6	
91-99	AMR-WB à 12.65 kbits/s	0	
100-108		3	
109-117		6	
118-126	AMR-WB à 23.85 kbits/s	0	

127-135		3	
136-144		6	

Tableau 4.14 Liste des conditions du test de transport de contenu binaural wideband

Les 576 fichiers (144 conditions x 4 échantillons) sont ensuite disponibles pour être diffusés sur casque.

4.3.3 Les auditeurs

Les auditeurs, au nombre de 32 par test, sont choisis au hasard parmi la population normale selon la recommandation de la méthodologie de test. Aucune procédure de rejet n'a été instaurée.

4.3.4 Procédure expérimentale

La spatialisation pouvant être perturbante pour un sujet non averti, notamment avec la sensation d'un son "au milieu de la tête", et interprétée comme un défaut par rapport à une condition stéréophonique standard, il nous a semblé judicieux de demander aux sujets de comparer une condition de test avec sa référence au même angle (la condition 1, 2,… ou 9). Cette référence est évidemment la même phrase prononcée par le même locuteur et spatialisée au même angle que la condition testée, mais sans la dégradation apportée par le codage et la perte de trame.

La méthode *Degradation Category Rating* (DCR) [47] a en conséquence été choisie et appliquée pour les deux tests, car elle permet d'évaluer la qualité globale d'un échantillon testé par rapport à une référence. Les sujets doivent donner une note de dégradation de qualité en fonction d'une échelle relative d'évaluation.

L'échelle d'évaluation de la procédure DCR est donnée Tableau 4.15. Les notes moyennes obtenues sont appelées *Degradation Mean Opinion Score* (DMOS).

Qualité de la parole	Note
Dégradation imperceptible ou même parfois amélioration	5
Dégradation perceptible audible mais pas gênante	4
Dégradation perceptible légèrement gênante	3
Dégradation perceptible gênante	2
Dégradation perceptible très gênante	1

Tableau 4.15 Echelle de note DCR

Les deux tests se sont déroulés dans une salle acoustiquement isolée, suivant les recommandations de la norme P.800 [47].

Du fait du grand nombre de conditions, chaque test a été divisé en 3 sous-tests T1, T2 et T3, avec 3 angles par sous-test :

- T1 : les angles testés sont ±60° et 0°.
- T2 : les angles testés sont -90°, -20° et 45°.
- T3 : les angles testés sont -45°, 20° et 90°.

Pour chaque test, seize personnes ont passé T1 en deux sous-groupes de 8 (G1-1 et G1-2). Seize autres personnes ont passé T2 et T3 à 2 jours d'intervalles, en deux sous-groupes de 8 (G2-1 et G2-2). Au sein d'un même sous-test, l'ordre de présentation des conditions était aléatoire pour chaque sous-groupe. Chaque sous-test commençait par un apprentissage.

Le casque utilisé est le Sennheiser HD25, qui possède une réponse plate dans la zone de fréquence qui nous intéresse [50 Hz - 7000 Hz]. Le niveau d'écoute par canal est 69 dB SPL.

4.3.5 Résultats narrowband

Nous cherchons tout d'abord à vérifier la symétrie de perception de la qualité de parole, pour un contenu symétrique narrowband (présenté avec des HRTF symétriques gauche/droite).

En conséquence, une analyse de variance ANOVA, suivie d'un test HSD de tukey [48] est utilisée pour les groupes G1 et G2 afin de comparer les notes obtenues pour les angles gauche et droit. Quel que soit l'angle, aucune dissymétrie n'a été trouvée dans les résultats des sujets. Cela est illustré pour le groupe G1 sur la Figure 4.17. Ces résultats montrent que la perception de la qualité de la parole est symétrique (gauche/droit) en narrowband, quels que soient le codage ou la perte de paquets. En conséquence, les résultats pour chaque angle sont regroupés en valeur absolue.

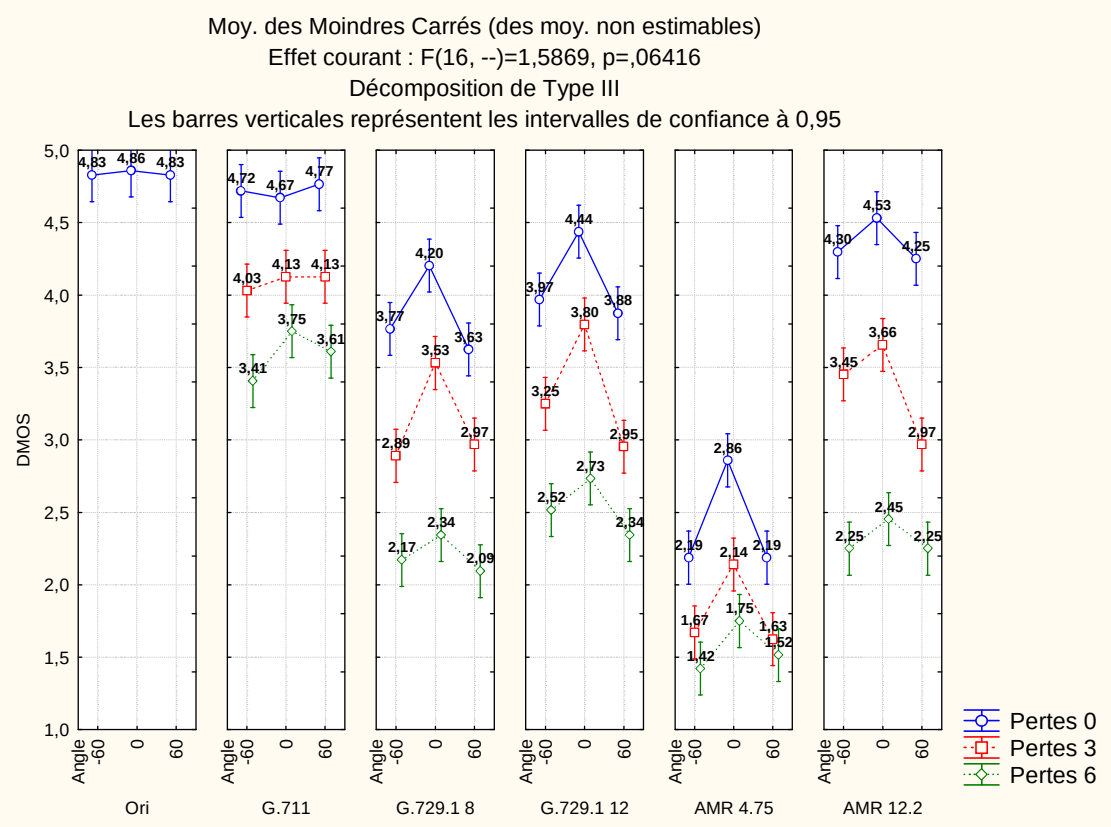

Figure 4.17 DMOS pour les angles ±60° et 0° suite au codage dual-mono narrowband

Suite à ce regroupement, les notes DMOS obtenues par chaque codeur, aux différents taux de PdT et selon les différents angles, sont illustrées sur la Figure 4.18.

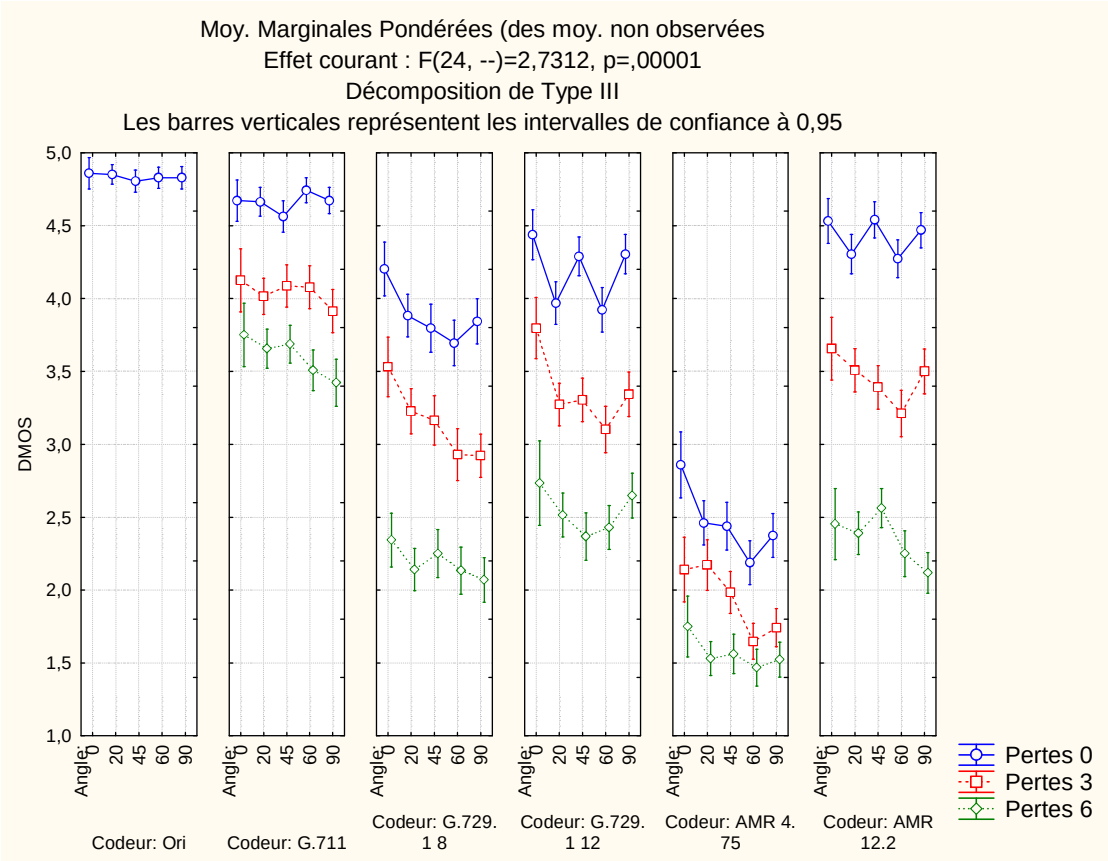

Figure 4.18 Illustration de l'impact de l'angle sur la qualité perçue des codeurs narrowband

Il apparaît que la valeur de l'angle a généralement généré des différences de notation assez importantes. On peut remarquer que l'angle 0° est généralement aussi bien voire mieux noté que les autres angles, même si ce n'est pas souvent significatif comme le montre le test HSD de Tukey réalisé ci-dessous. L'angle 60° semble être la borne basse pour la plupart des configurations testées même si, à nouveau, cela n'est pas forcément significatif. Le test HSD de Tukey a en effet montré les différences significatives suivantes:

- Pour le codeur G.729.1 à 8 kbits/s :
  - entre les angles 0° et 60°, à 0% et 3% de PdT.
  - entre les angles 0° et 90° à 3% de PdT
- Pour le codeur G.729.1 à 12 kbits/s :
  - entre les angles 0° et 60°, à 0% et 3% de PdT.
  - entre les angles 0° et 20° d'une part et les angles 0°et 45° d'autre part à 3% de PdT.
- Pour le codeur AMR à 4.75 kbits/s :
  - entre les angles 0° et 60° à 0% et 3% de PdT.
  - entre les angles 20°et 60° d'une part et les angles 20° et 90° d'autre part à 3% de PdT.
- Pour le codeur AMR à 12.2 kbits/s :
  - entre les angles 0°et 20° d'une part et les angles 0° et 45° d'autre part à 3% de PdT.
  - entre les angles 45° et 90° à 6% de PdT

Les notes DMOS obtenues pour chaque codeur, tous angles confondus et à une PdT donnée, sont présentées sur la Figure 4.19. Le regroupement se justifie par le fait qu'un codeur est choisi pour l'ensemble d'une communication et non selon un angle donné.

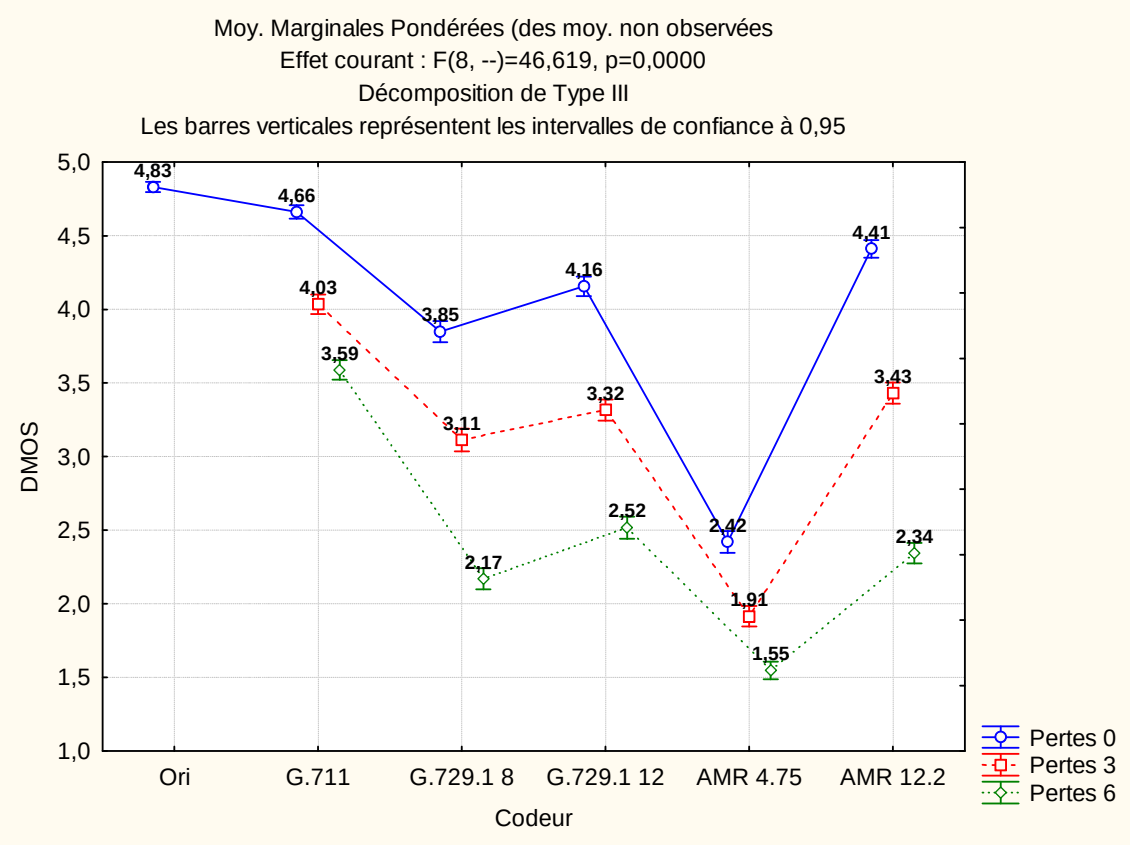

Figure 4.19 Résultats des codeurs narrowband tous angles confondus

Sur la Figure 4.19**,** il apparaît que la référence est la condition la mieux notée à 0% de PdT. On note aussi que, pour une même famille de codeurs, les codeurs haut-débit sont mieux notés que les bas-débit.

Le codeur PCM G.711 supporte le mieux l'ensemble des conditions testées avec ou sans PdT. Il est intéressant de souligner que ce codeur semble être le codeur de bonne qualité le moins impacté par la perte de trames.

L'ensemble composé par les codeurs CELP (G.729.1 et AMR) subit quant à lui de manière plus prononcée la dégradation de qualité due à la PdT. Plus la dégradation est importante, plus l'écart entre les codeurs CELP et le codeur G.711 se creuse.

Les effets des facteurs (Codeur, Perte de trames et Angle) et de leurs interactions sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles. Les résultats de cette analyse sont donnés par le Tableau 4.16.

Facteur	Degrés de liberté	F-ratio	p = Significativité
Codeur	5	427,918	0,00
Perte de trames	2	917,536	0,00
Angle	3	36,905	0,00
Codeur*Perte de trames	8	46,619	0,00
Codeur*Angle	15	5,473	0,00
Angle*Perte de trames	6	4.945	0.00
Angle*Perte de trames * Codeur	24	2.73	0.00

Tableau 4.16 Effets des différents facteurs du test narrowband de transport de contenu binaural

Les différentes équivalences selon les différentes pertes de trames sont données dans le Tableau 4.17 et confirment de manière significative les tendances énoncées ci-dessus. Elles ont été obtenues grâce au test HSD de Tukey.

N°	0%	3%	6%
1	Ori	G.711	G.711
2	G.711	AMR 12.2 ⇔ G.729.1 12	G.729.1 12
3	AMR 12.2	G.729.1 8	AMR 12.2
4	G.729.1 12	AMR 4.75	G.729.1 8
5	G.729.1 8		AMR 4.75
6	AMR 4.75		

Tableau 4.17 Classement des codeurs narrowband à différentes pertes de trames pour le transport de contenu binaural

Il est intéressant de souligner que le codeur G.729.1 à 12 kbits/s résiste mieux à l'augmentation du pourcentage de PdT que le codeur AMR à 12.2 kbits/s, et est même significativement mieux classé à 6% de PdT. De même à 6% de PdT, le G.729.1 à 8 kbits/s a mieux résisté à la PdT en termes de qualité audio que l'AMR à 12.2 kbits/s mais reste cependant encore moins bien classé. Le codeur G.729.1 semble être plus robuste à la PdT que le codeur AMR. Ces ordres se retrouvent globalement quel que soit l'angle.

4.3.6 Résultats wideband

De même qu'en narrowband, nous avons cherché tout d'abord à vérifier l'éventuelle symétrie (gauche/droite) de la perception de la qualité de parole. En suivant le même protocole qu'en 4.3.5, les résultats montrent à nouveau que la perception de la qualité de la parole est symétrique (gauche/droite) en wideband (cf. Figure 4.20 pour différentes PdT). En conséquence, les résultats pour chaque angle sont regroupés en valeur absolue.

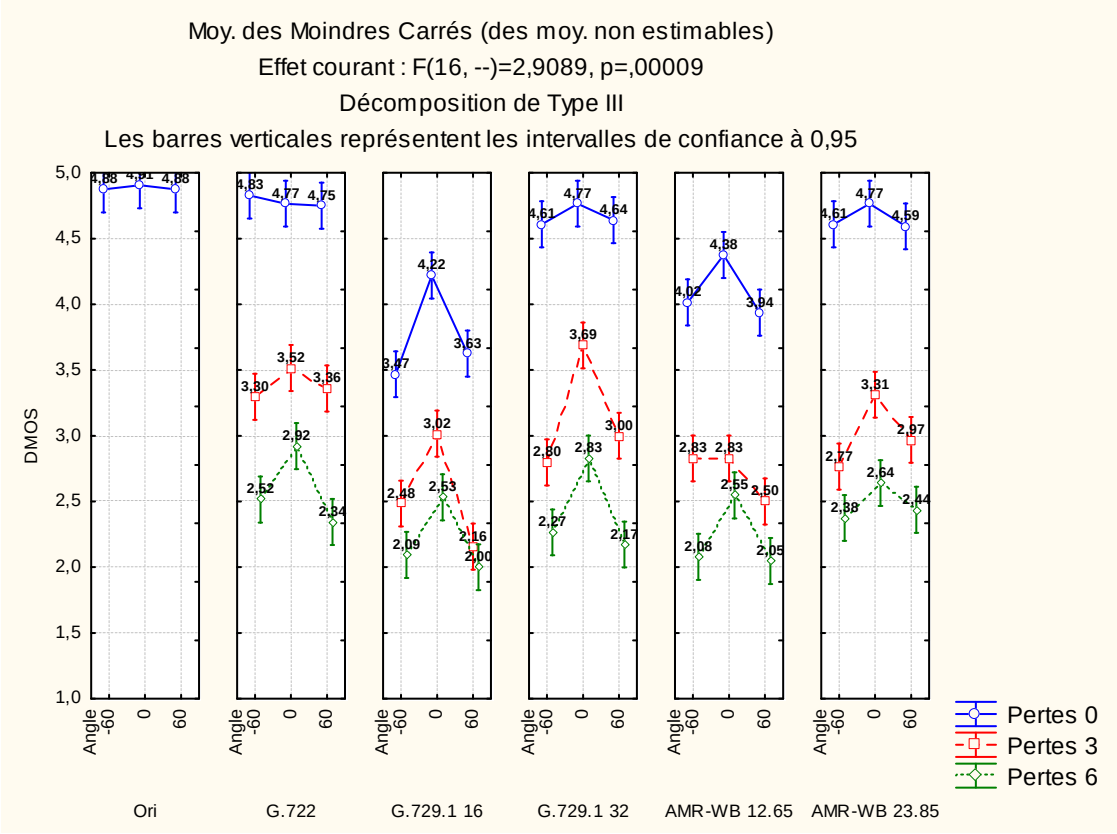

Figure 4.20 DMOS pour les angles ±60° et 0° suite au codage dual-mono wideband

Suite à ce regroupement, les notes DMOS obtenues par chaque codeur, aux différents taux de PdT et selon les différents angles, sont illustrées sur la Figure 4.21.

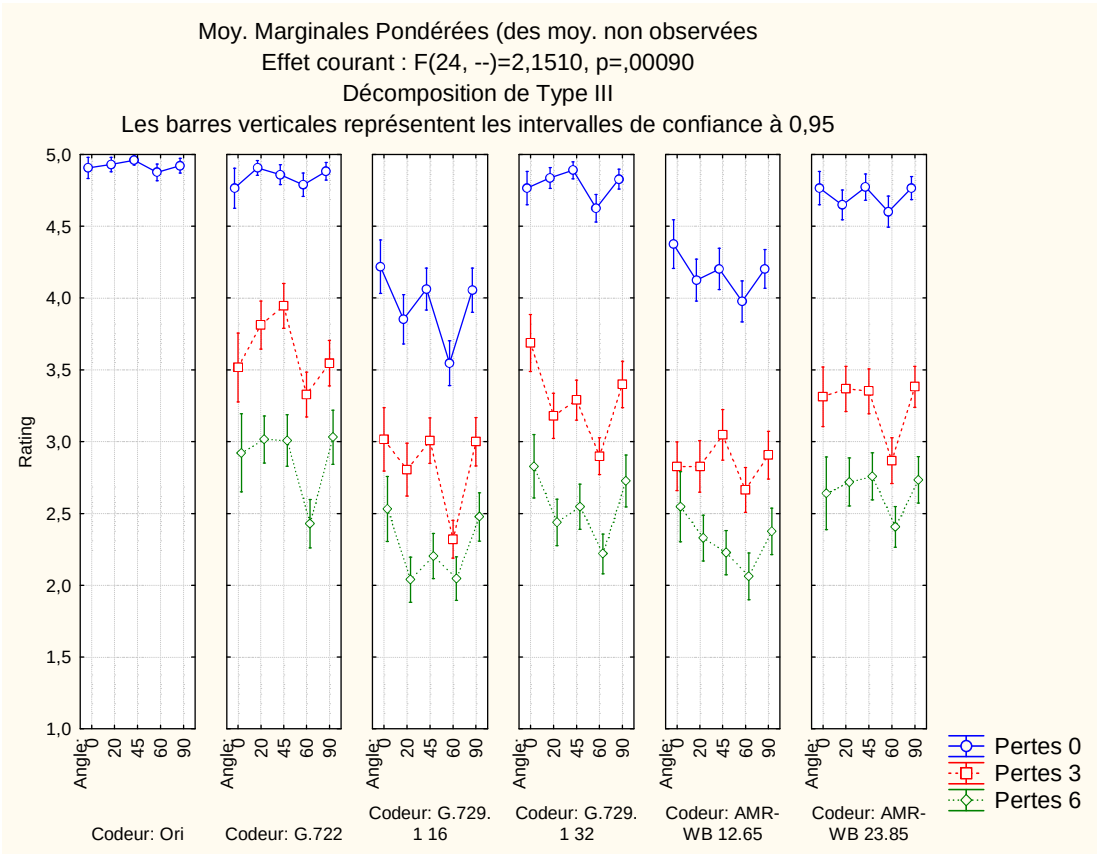

Figure 4.21 Illustration de l'impact de l'angle sur la qualité perçue des codeurs wideband

Il apparaît à nouveau que la valeur de l'angle a généralement généré des différences de notation assez importantes. On peut noter que l'angle 0° a des résultats généralement meilleurs que les autres angles même si cela est moins marqué qu'en narrowband. A l'inverse l'angle 60° semble être une borne basse assez fréquente et de manière significative parmi les configurations testées. Cette dernière caractéristique apparaissait moins en narrowband. Ces tendances sont confirmées par le test HSD de Tukey donnant les différences significatives suivantes :

- Pour le codeur G.729.1 à 16 kbits/s :
  - entre l'angle 60° et les angles 0°, 45° et 90° à 0% de PdT
  - entre l'angle 60° et les angles 0°, 20°, 45° et 90° à 3% de PdT
  - entre les angles 0° et 20°, les angles 0° et 60°, les angles 20° et 90°, et les angles 60° et 90°, à 6% de PdT.
- Pour le codeur G.729.1 à 32 kbits/s :
  - entre les angles 0° et 20°, les angles 0° et 60°, les angles 60° et 90° à 3% de PdT.
  - entre l'angle 60° et les angles 0°et 90°, à 6% de PdT
- Pour le codeur AMR-WB à 12.65 kbits/s :
  - entre les angles 0° et 60° à 6% de PdT.
- Pour le codeur AMR-WB à 23.85 kbits/s :
  - entre l'angle 60° et les angles 20°, 45° et 90° à 3% de PdT.
- Pour le codeur G.722 :
  - entre les angles 20° et 60°, les angles 45° et 60°, et les angles 45° et 90°, à 3% de PdT.
  - entre l'angle 60° et les angles 0°, 20°, 45° et 90° à 6% de PdT

Les notes DMOS par chaque codeur tous angles confondus à une PdT donnée sont présentées sur la Figure 4.22, ci-dessous.

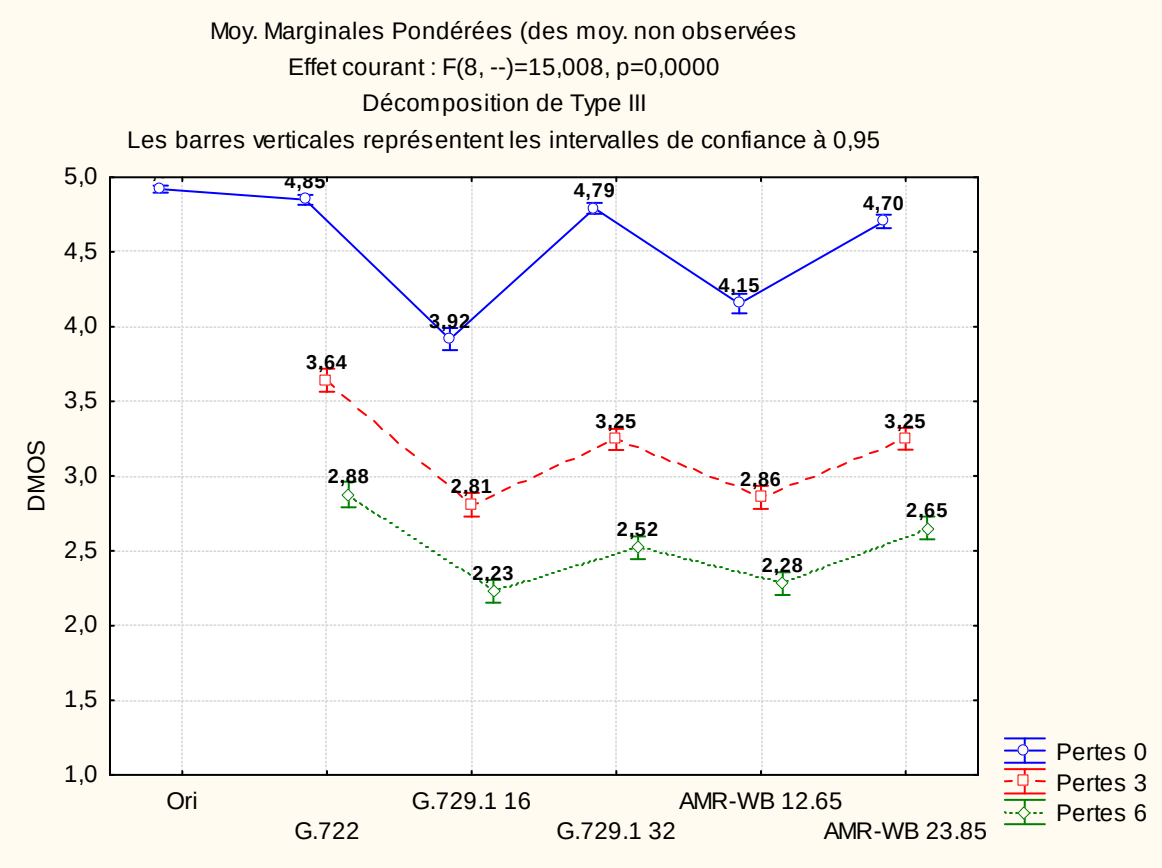

Figure 4.22 Résultats des codeurs wideband tous angles confondus pour le transport de contenu binaural

Sur la Figure 4.22, il apparaît d'une part que la référence est la condition la mieux notée à 0% de PdT et d'autre part que pour une même famille de codeurs, les codeurs haut-débit sont mieux notés que les bas-débit.

On note aussi que l'augmentation de la PdT entraîne une diminution des notes de qualité. Tout comme en narrowband, les codeurs CELP supportent moins bien la PdT que le codeur ADPCM. Plus la perte est importante, plus l'écart se creuse entre eux.

Les effets des facteurs (Codeur, Perte de trames et Angle) et de leurs interactions sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles prenant en compte les facteurs du test. Les résultats de cette analyse sont donnés par le Tableau 4.18.

Facteur	Degrés de liberté	F-ratio	P = Significativité
Codeur	5	138,891	0,00
Perte de trames	2	945,310	0,00
Angle	3	32,332	0,00
Codeur*Perte de trames	8	15,008	0,00
Codeur*Angle	15	6,154	0,00
Perte de trames*Angle	6	4,138	0,00
Angle*Perte de trames*Codeur	24	2.151	0.00

Tableau 4.18 Effets des différents facteurs du test wideband de transport de contenu binaural

Les différentes équivalences selon les différentes pertes de trames sont données dans le Tableau 4.19 et confirment de manière significative les tendances énoncées ci-dessus. Elles ont été obtenues grâce au test HSD de Tukey.

N°	0%	3%	6%
1	Ori ⇔ G.722 ⇔ G.729.1 32	G.722	G.722
2	G.722 ⇔ G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85
3	AMR-WB 12.65	AMR-WB 12.65 ⇔ G.729.1 16	AMR-WB 12.65 ⇔ G.729.1 16
4	G.729.1 16		

Tableau 4.19 Classement des codeurs wideband à différentes pertes de trames

Grâce au Tableau 4.19, on remarque, qu'en présence de PdT, les codeurs G.729.1 à 32 kbits/s et l'AMR-WB à 23.85 kbits/s restent similaires même si à 6% un écart semblait se dessiner sur la figure précédente. A plus bas-débit, le codeur G.729.1 à 16 kbits/s est devenu équivalent au codeur AMR-WB à 12.65 kbits/s. Ces ordres se retrouvent globalement quel que soit l'angle.

4.3.7 Discussion et perspectives

Nous cherchons tout d'abord à retrouver pour l'angle 0°, les résultats obtenus en diotique dans la section 4.2. On rappelle les équivalences dans les Tableau 4.20 et Tableau 4.21 :

N°	0%	3%	6%
1	Direct	G.711	G.711
2	AMR 12.2 ⇔ G.711	G.729.1 12 ⇔ AMR 12.2	G.729.1 12
3	G.711 ⇔ G.729.1 12	AMR 12.2 ⇔G.729.1 8	G.729.1 8 ⇔ AMR 12.2
4	G.729.1 8	AMR 4.75	AMR 4.75
5	AMR 4.75		

Tableau 4.20 Classement issu de 4.2. des codeurs narrowband en écoute diotique à différentes pertes de trames

N°	0%	3%	6%
1	Direct	G.722 ⇔ G.729.1 32	G.729.1 32
2	G.722 ⇔ G.729.1 32 ⇔ AMR-WB 23,85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 16 ⇔ G.722
3	AMR-WB 12.65 ⇔ G.729.1 16	G.729.1 16 ⇔ AMR-WB 12.65	AMR-WB 23.85 ⇔ AMR-WB 12.65

Tableau 4.21 Classement issu de 4.2. des codeurs wideband en écoute diotique à différentes pertes de trames

En utilisant les Figure 4.18 et Figure 4.21, à 0° et quelles que soient la perte de paquets et la largeur de bande, nous retrouvons globalement le même ordonnancement en narrowband. Les différences d'ordonnancement sont plus importantes en wideband. Les différences entre les codeurs ne sont pas significatives à l'angle 0° dans ce test pour une perte de trames donnée. Nous ne pouvons donc pas conclure. Cela est notamment dû au nombre de sujets utilisés (16 dans ce test contre 32 dans le précédent) et au nombre d'échantillons issus de chaque locuteur (1 dans ce test contre 2 dans le précédent) qui a augmenté les intervalles de confiance et donc diminué la précision.

Globalement, ce test a montré que les codeurs G.711 et G.722 étaient plus adaptés pour transporter un contenu binaural. Néanmoins, ils ont un coût très important en termes de débits (64 kbits/s par voie) par rapport aux codeurs AMR, AMR-WB ou G.729.1, qu'ils compensent il est vrai par une complexité moindre. Cette complexité est aussi un facteur important lors d'un déploiement à grande échelle.

Il a été aussi montré que la perception de la qualité de la parole est symétrique, ceci permettant de faire des simplifications supplémentaires dans le test suivant.

Le fait que les sujets donnent de meilleures notes à l'angle 0° peut se justifier par une gêne moindre lorsque les dégradations sont symétriques (le contenu étant identique sur chaque canal).

4.4 Evaluation des deux configurations de conférence audio avec des codeurs monophoniques de parole

L'étude précédente a montré qu'il était possible de transporter du contenu binaural en dual-mono grâce à des codeurs monophoniques avec une qualité proche de la référence. Il convient à présent de tester les deux configurations (centralisée et distribuée) dans leur intégralité en termes de qualité audio.

De même qu'en 4.2 et 4.3, deux tests ont été menés, l'un avec une bande de qualité narrowband et l'autre avec une bande de qualité wideband. Les codeurs utilisés sont ceux présentés dans la section 4.1.2.

4.4.1 Stimuli

Les 4 échantillons de test sont extraits de la base de données de France Télécom et durent 8 secondes. Ils sont constitués de deux phrases espacées par un silence, et sont échantillonnés à 16 kHz, pour une bande de fréquences de 8 kHz. Ils sont énoncés par quatre locuteurs différents (2 femmes et 2 hommes), prononçant chacun un échantillon (tous différents deux à deux). Chaque échantillon est traité par l'ensemble des conditions des tests.

4.4.2 Conditions de test narrowband et wideband

Les traitements narrowband et wideband sont explicités dans les deux sections suivantes. Leur rôle est de simuler les deux configurations de conférence en utilisant les codeurs conversationnels normalisés.

Pour chaque configuration de chaque traitement, des pertes de trames à 0% et 3% ont été utilisées. Pour la configuration centralisée de chaque traitement, 2 PdT différentes (chacune à 0% et 3%) ont été introduites pour simuler des pertes entre un terminal émetteur et le pont de conférence, mais aussi entre ce dernier et le terminal récepteur (simultanément sur les deux canaux).

Pour la configuration distribuée de chaque traitement, la perte de trame a lieu sur le contenu monophonique.

L'étude précédente ayant montré que les résultats de qualité étaient symétriques en termes d'angles, dans chaque test, les signaux ont donc été spatialisés uniquement à 5 angles : 0°, -20°, 45°, -60° et 90°.

4.4.2.1 Conditions narrowband

Le traitement appliqué aux fichiers pour la configuration centralisée est illustré Figure 4.23. On remarque bien les deux étages de la configuration avec chacune une étape d'encodage-décodage (bloc *Traitement avec ou sans perte de trames*). Le 1er étage comprend la mise en forme des fichiers (filtrage, égalisation et sous échantillonnage), l'encodage-décodage monophonique et la spatialisation. Le 2nd étage effectue le codage dual-mono du contenu binaural avant les traitements (Sur-échantillonnage, filtrage) pour la diffusion sur casque. Afin de diminuer les combinaisons à tester, le codeur utilisé pour le codage monophonique est le même que celui utilisé pour le codage dual-mono.

Les filtrages MIRS 16 (*Modified Intermediate Reference System*) [42] et RXIRS 16 (*Receive-Side IRS*) [42] permettent de simuler des appareils de prise de son et de restitution numériques narrowband pour des signaux de fréquence d'échantillonnage initiale de 16 kHz.

Les blocs de sous-échantillonnage et sur-échantillonnage permettent de ré-échantillonner des contenus à des fréquences d'échantillonnage respectivement de 8 et 16 kHz, afin de pouvoir être traités par les blocs suivants. L'égalisation à -26 dB_{ov} permet, comme son nom l'indique,

d'égaliser en énergie les fichiers de départ pour que les sujets aient la même perception sonore en niveau quel que soit le fichier. Ces trois traitements sont effectués grâce aux filtres définis dans [42].

Le bloc *Spatialisation* effectue une spatialisation binaurale, selon les 5 angles suivants : 0°, -20°, 45°, -60° et 90°, d'un contenu monophonique pré-traité (après filtrage, égalisation et sous-échantillonnage) et encodé-décodé, de fréquence d'échantillonnage 8 kHz.

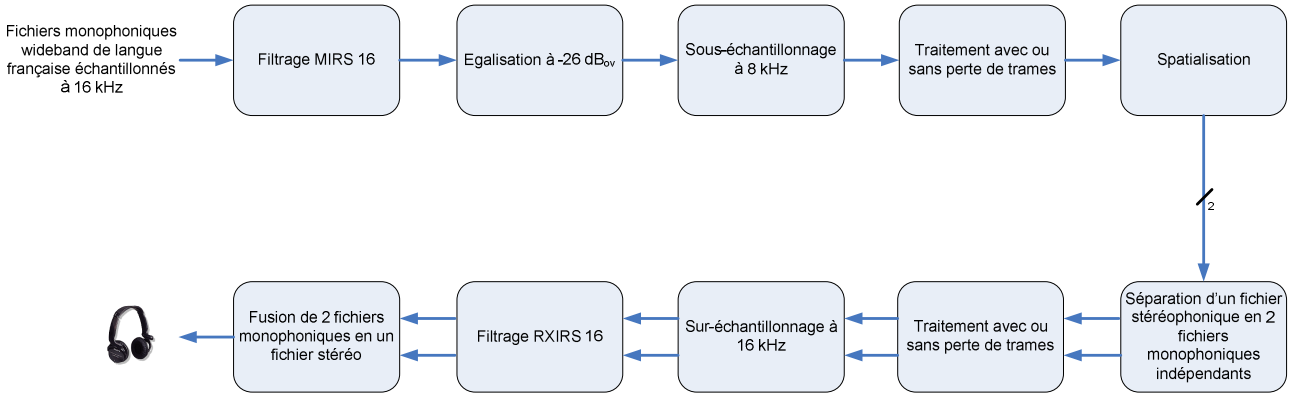

Figure 4.23 Traitement appliqué aux fichiers pour le test narrowband pour la conférence centralisée

Le traitement appliqué aux fichiers narrowband pour la conférence distribuée est illustré Figure 4.24. Cette configuration est similaire au 1er étage de la conférence audio centralisée vu ci-dessus mis à part les traitements (sur-échantillonnage et filtrage) de mise en forme des fichiers avant diffusion sur casque.

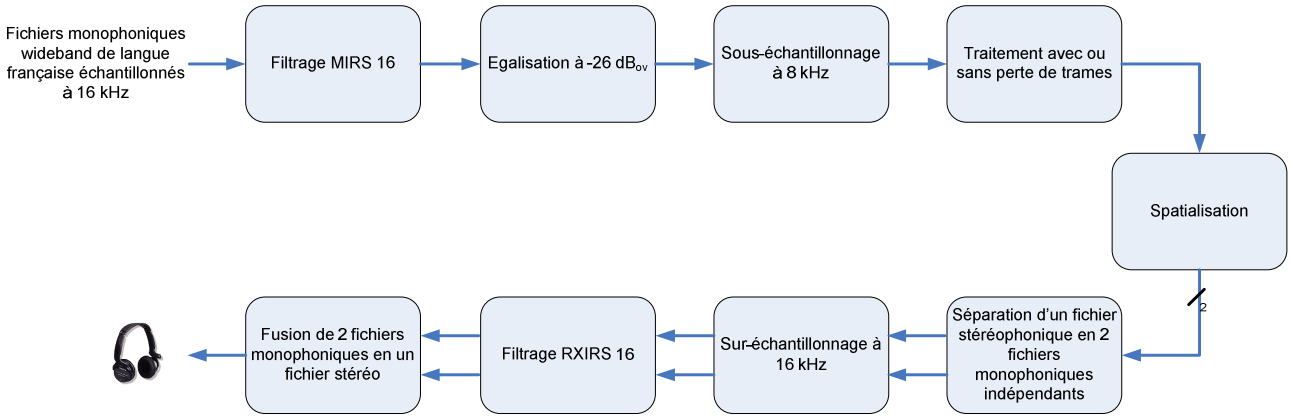

Figure 4.24 Traitement appliqué aux fichiers pour le test narrowband pour la conférence distribuée

Les différentes conditions du test narrowband sont données Tableau 4.22. Les conditions de 1 à 5 sont nos conditions de référence spatialisées sans encodage-décodage (aussi appelé codeur Ori pour Original par la suite), ni perte de trame.

Il convient de préciser que par exemple la perte 0 correspond à la conférence distribuée sans PdT, que les pertes 3&3 correspond à la conférence centralisée avec 3% de PdT de chaque côté du pont de conférence, etc.

Numéro de la condition	Codeur	Type de configuration	Pourcentage de perte de trame	Angles
1-5	Aucun	Aucun	0	0°, -20°, 45°, -60° et 90°
6-10	G.711	Centralisé	0 & 0	
11-15			0 & 3	
16-20			3 & 0	
21-25			3 & 3	
26-30		Distribué	0	
31-35			3	

36-40	AMR à 4.75 kbits/s	Centralisé	0 & 0	
41-45			0 & 3	
46-50			3 & 0	
51-55			3 & 3	
56-60		Distribué	0	
61-65			3	
66-70	AMR à 12.2 kbits/s	Centralisé	0 & 0	
71-75			0 & 3	
76-80			3 & 0	
81-85			3 & 3	
86-90		Distribué	0	
91-95			3	
96-100	G.729.1 à 8 kbits/s	Centralisé	0 & 0	
101-105			0 & 3	
106-110			3 & 0	
111-115			3 & 3	
116-120		Distribué	0	
121-125			3	
126-130	G.729.1 à 12 kbits/s	Centralisé	0 & 0	
131-135			0 & 3	
136-140			3 & 0	
141-145			3 & 3	
146-150		Distribué	0	
151-155			3	

Tableau 4.22 Liste des conditions du test de conférence narrowband

Les auditeurs écouteront ainsi 620 fichiers (155 conditions x 4 échantillons).

4.4.2.2 Conditions wideband

Le traitement appliqué aux fichiers pour la configuration centralisée est illustré Figure 4.25. On retrouve bien les deux étages de la configuration centralisée avec le premier étage de mise en forme des fichiers (par filtrage et égalisation), l'encodage-décodage monophonique et la spatialisation. Le second étage est très simple avec uniquement l'encodage dual-mono.

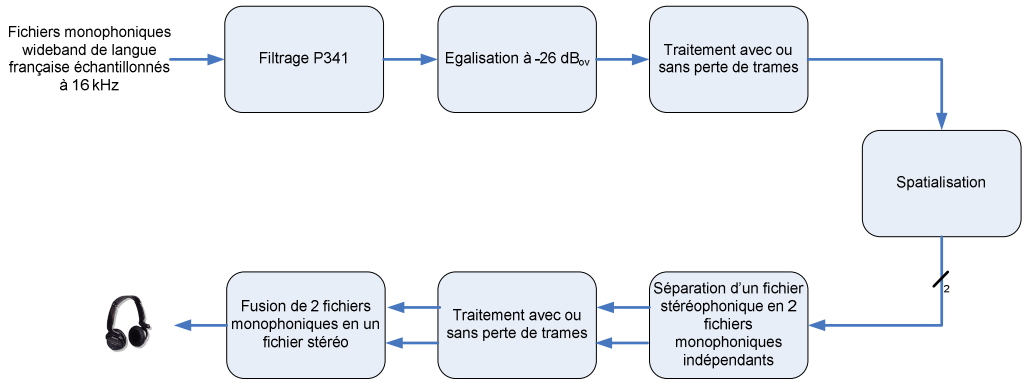

Figure 4.25 Traitement appliqué aux fichiers pour le test wideband pour la conférence centralisée

Le traitement appliqué aux fichiers wideband pour la conférence distribuée est illustré Figure 4.26. Il comprend un filtrage, une égalisation en niveau, un encodage-décodage monophonique ainsi que la spatialisation.

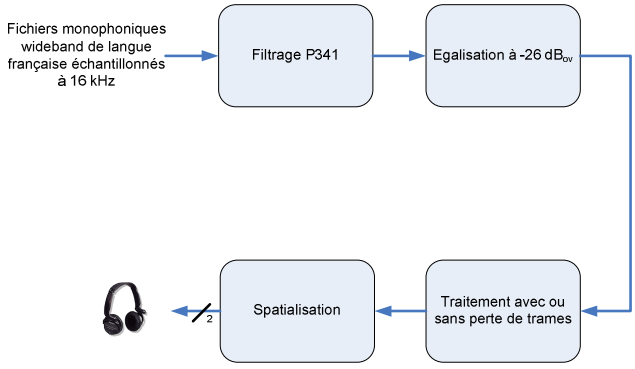

Figure 4.26 Traitement appliqué aux fichiers pour le test wideband pour la conférence distribuée

Le filtrage P.341 [42] permet de simuler l'appareil de prise de son et de restitution numériques. L'égalisation à -26 dB_{ov} est effectuée grâce à un filtre défini dans [42].

Les différentes conditions du test narrowband sont données Tableau 4.23. Les conditions de 1 à 5 sont nos conditions de référence spatialisées sans encodage-décodage (aussi appelé codeur Ori pour Original par la suite), ni perte de trame.

Numéro de la condition	Codeur	Type de configuration	Pourcentage de perte de trame	Angles
1-5	Aucun	Aucun	0	0°, -20°, 45°, -60° et 90°
6-10	G.722	Centralisé	0 & 0	
11-15			0 & 3	
16-20			3 & 0	
21-25			3 & 3	
26-30		Distribué	0	
31-35			3	
36-40	AMR-WB à 12.65 kbits/s	Centralisé	0 & 0	
41-45			0 & 3	
46-50			3 & 0	
51-55			3 & 3	
56-60		Distribué	0	
61-65			3	
66-70	AMR-WB à 23.85 kbits/s	Centralisé	0 & 0	
71-75			0 & 3	
76-80			3 & 0	
81-85			3 & 3	
86-90		Distribué	0	
91-95			3	
96-100	G.729.1 à 16 kbits/s	Centralisé	0 & 0	
101-105			0 & 3	
106-110			3 & 0	
111-115			3 & 3	
116-120		Distribué	0	
121-125			3	
126-130	G.729.1 à 32 kbits/s	Centralisé	0 & 0	
131-135			0 & 3	
136-140			3 & 0	
141-145			3 & 3	
146-150		Distribué	0	
151-155			3	

Tableau 4.23 Liste des conditions du test de conférence wideband

Les auditeurs écouteront ainsi 620 fichiers (155 conditions x 4 échantillons). Pour simplifier, nous parlerons par la suite des angles uniquement en valeur absolue.

4.4.3 Les auditeurs

Les auditeurs, au nombre de 24 par test, sont choisis au hasard parmi la population normale selon la recommandation de la méthodologie de test. Aucune procédure de rejet n'a été instaurée.

4.4.4 Procédure expérimentale

Pour les mêmes raisons que dans la section 4.3.1, la méthode de test DCR a été utilisée. A nouveau, les références (conditions 1 à 5) sont évidemment les mêmes phrases prononcées par le même locuteur et spatialisée au même angle que la condition testée, mais sans la dégradation apportée par le codage et la perte de trame.

L'échelle d'évaluation de la procédure DCR est donnée Tableau 4.24. Les notes moyennes obtenues sont appelées *Degradation Mean Opinion Score* (DMOS).

Qualité de la parole	Note
Dégradation imperceptible ou même parfois amélioration	5
Dégradation perceptible audible mais pas gênante	4
Dégradation perceptible légèrement gênante	3
Dégradation perceptible gênante	2
Dégradation perceptible très gênante	1

Tableau 4.24 Echelle de note DCR

Les deux tests se sont déroulés dans une salle d'écoute acoustiquement isolée, suivant les recommandations de la norme P.800 [47].

Pour chaque test, les 24 sujets ont été séparés en 3 groupes de 8 (G1-G3). Dû au grand nombre de conditions, chaque test a été divisé en 4 sessions de 1h30 avec à chaque fois un apprentissage. Chacun des 3 groupes a ainsi passé les 4 sessions avec un jour d'intervalle entre chaque session. L'ordre de présentation des conditions était aléatoire pour chaque groupe.

Le casque utilisé est le Sennheiser HD25, qui possède une réponse plate dans la zone de fréquence qui nous intéresse [50 Hz - 7000 Hz]. Le niveau d'écoute par canal est 69 dB SPL.

4.4.5 Résultats narrowband

Les notes DMOS obtenues par chaque codeur, aux différents taux de PdT et selon les différents angles, sont illustrées sur la Figure 2.24.

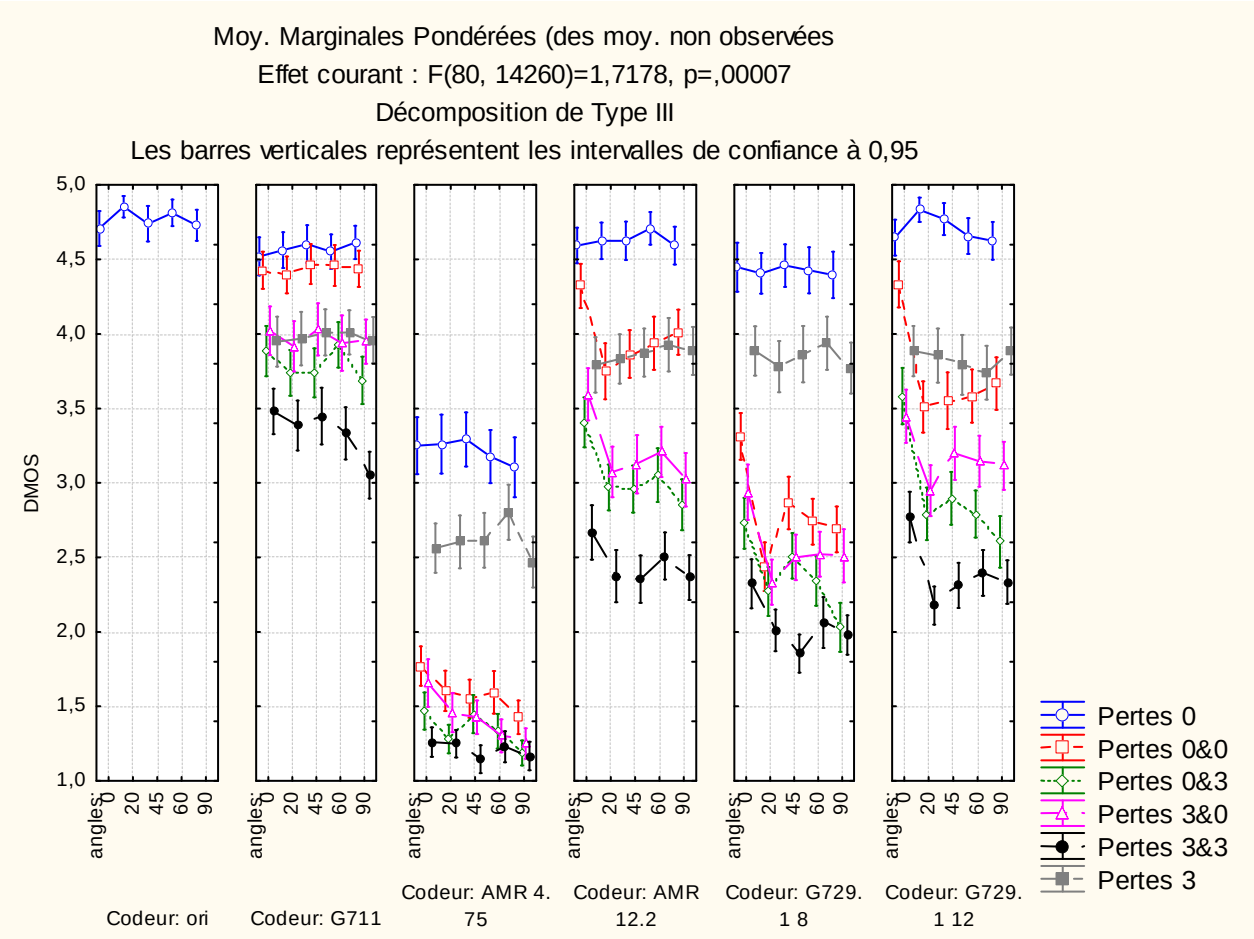

Figure 4.27 Illustration de l'impact de l'angle sur la qualité perçue des codeurs narrowband pour les configurations de conférence

Il apparaît que l'impact de l'angle est important notamment sur les codeurs AMR à 12.2 kbits/s, G.729.1 à 8 et 12 kbits/s. De plus, l'angle 0° se distingue de nouveau sur les configurations centralisées tout comme pour le test de la section 4.3. Sinon pour la plupart des autres angles et quel que soit le codeur, les résultats sont globalement uniformes. Le test HSD de Tukey a en effet confirmé les différences significatives suivantes:

- Pour le codeur AMR à 12.2 kbits/s pour les configurations :
  - 0&0 entre les angles 0° et 20°,
  - 0&3 entre les angles 0° et 90°,
  - 3&0 entre l'angle 0° d'une part et les angles 20°, 45° et 90° d'autre part.
- Idem pour le G.729.1 à 8 kbits/s pour les configurations :
  - 0&0 entre les angles 0° d'une part et 20°, 60° et 90° d'autre part,
  - 0&3 entre l'angle 90° et les angles 0° et 45°,
  - 3&0 entre l'angle 0° et 20°,
  - 3&3 entre les angles 0° et 45°.
- Idem pour le G.729.1 à 12 kbits/s pour les configurations :
  - 0&0 et 0&3 entre l'angle 0° et les autres,
  - 3&0 et 3&3 entre les angles 0° et 20°.

La Figure 4.28 présente les résultats des codeurs narrowband tous angles confondus pour chaque configuration.

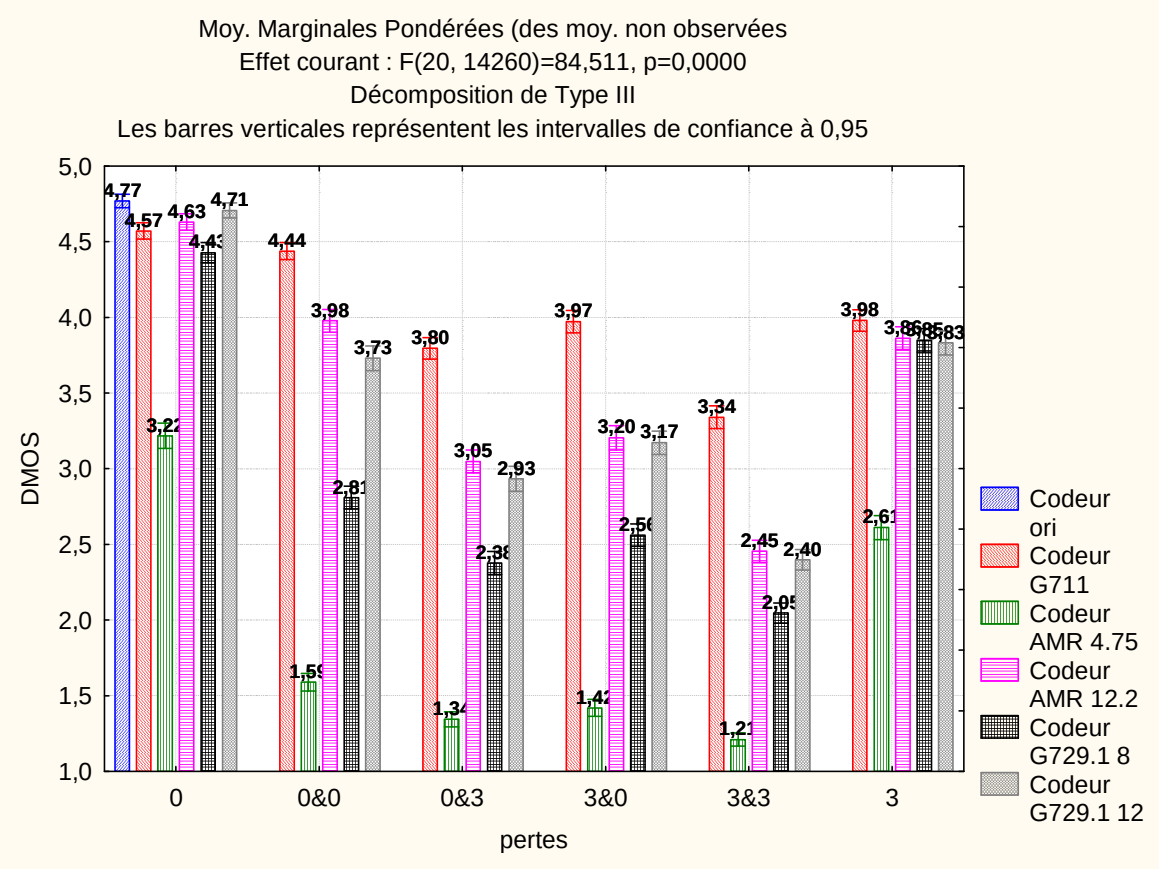

Figure 4.28 Résultats des codeurs narrowband tous angles confondus par configuration de conférence

Il apparaît que la référence est sans surprise la condition la mieux notée à 0% de PdT. Un autre résultat logique et attendu est que l'augmentation de la PdT entraine une diminution des notes de qualité. Pour une même famille de codeurs, les codeurs haut-débit sont mieux notés que les bas-débit (de façon nette sauf pour le G.729.1 pour la conférence distribuée à 3% de PdT).

Le codeur G.711 est le meilleur codeur pour la conférence audio centralisée quelle que soit la PdT. Il est à noter que les sujets font nettement plus la différence entre les codeurs dans la configuration de conférence audio centralisée, quelle que soit la perte de trames.

L'ensemble des effets des facteurs (Codeur, Perte de trames et Angle) du test et leurs interactions sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles. Les résultats de cette analyse sont donnés par le Tableau 4.25.

Facteur	Degrés de liberté	F-ratio	p = Significativité
Codeurs	5	1497,6	0,00
pertes	5	1255,4	0,00
angles	4	27,4	0,00
Codeur*pertes	20	43,7	0,00
Codeur*pertes*angles	20	2,5	0,00
Codeur*angles	80	0,9	0,00

Tableau 4.25 Effets des différents facteurs pour le test de conférence narrowband

Les différentes équivalences selon les différentes pertes de trames sont données dans le Tableau 4.26 et confirment de manière significative les tendances énoncées ci-dessus. Elles ont été obtenues grâce au test HSD de Tukey.

N°	0	3	0&0	0&3	3&0	3&3
1	Ori ⇔ G.729.1 12 ⇔ AMR 12.2	G.711 ⇔ AMR 12.2 ⇔ G.729.1 8 ⇔ G.729.1 12	G.711	G.711	G.711	G.711
2	AMR 12.2 ⇔ G.729.1 12 ⇔ G.711	AMR 4.75	AMR 12.2	AMR 12.2 ⇔ G.729.1 12	AMR 12.2 ⇔ G.729.1 12	AMR 12.2 ⇔ G.729.1 12
3	G.711 ⇔ G.729.1 8		G.729.1 12	G.729.1 8	G.729.1 8	G.729.1 8
4	AMR 4.75		G.729.1 8	AMR 4.75	AMR 4.75	AMR 4.75
5			AMR 4.75			

Tableau 4.26 Classement des codeurs narrowband tous angles confondus par configuration de conférence

Pour la conférence distribuée à 0% de PdT, on retrouve bien l'équivalence entre le G.711, l'AMR 12.2 et le G.729.1 établi pour l'écoute diotique (voir le Tableau 4.5). Néanmoins, les résultats sont plus resserrés notamment entre le G.711 et le G.729.1 à 8 kbits/s d'une part, et la référence et les deux codeurs CELP haut-débit d'autre part. Pour la conférence distribuée à 3% de PdT, les résultats obtenus ici (G.711 ⇔ AMR 12.2 ⇔ G.729.1 8 ⇔ G.729.1 12) sont compatibles avec les résultats obtenus dans le Tableau 4.6. Globalement à 0% de PdT ou 3% de PdT, au vu des notes obtenues, la spatialisation ne semble logiquement pas avoir impacté la qualité audio des signaux.

Pour les configurations centralisées, on remarque que tout d'abord sans PdT, le classement 0&0 est exactement identique au classement établi dans le test précédent sans PdT pour le transport d'un contenu binaural (voir Tableau 4.17).

De la même façon pour le classement de la configuration centralisée 0&3, on retrouve le classement établi dans le test précédent à 3% de PdT pour le transport d'un contenu binaural (voir Tableau 4.17). Par rapport aux configurations distribuées, on se rend bien compte de l'impact du second étage de la conférence centralisée qui impacte nettement la qualité.

Le G.711 se classe 1er dans les configurations centralisées et subit moins les pertes ainsi que les deux encodages-décodages que les codeurs CELP. On peut noter que le codeur G.729.1 à 12 kbits/s réagit mieux à la PdT que l'AMR à 12.2 kbits/s. En effet, moins bien noté pour la configuration 0&0, il est jugé équivalent à l'AMR à 12.2 kbits/s pour les configurations 0&3, 3&0 et 3&3.

Les classements restent identiques pour les configurations 0&3, 3&0 et 3&3.

Le codeur G.711 est incontestablement le meilleur dans le cas de l'utilisation d'un pont de conférence. Les codeurs G.729.1 à 12 kbits/s, G.711 et l'AMR à 12.2 kbits/s semblent être les meilleurs codeurs pour la conférence distribuée. Il est à noter que globalement les codeurs les mieux notés sont ceux disposant de plus de débit sauf pour le cas de la conférence distribuée à 0% de PdT.

La Figure 4.29 présente les résultats des codeurs narrowband tous angles confondus pour chaque configuration.

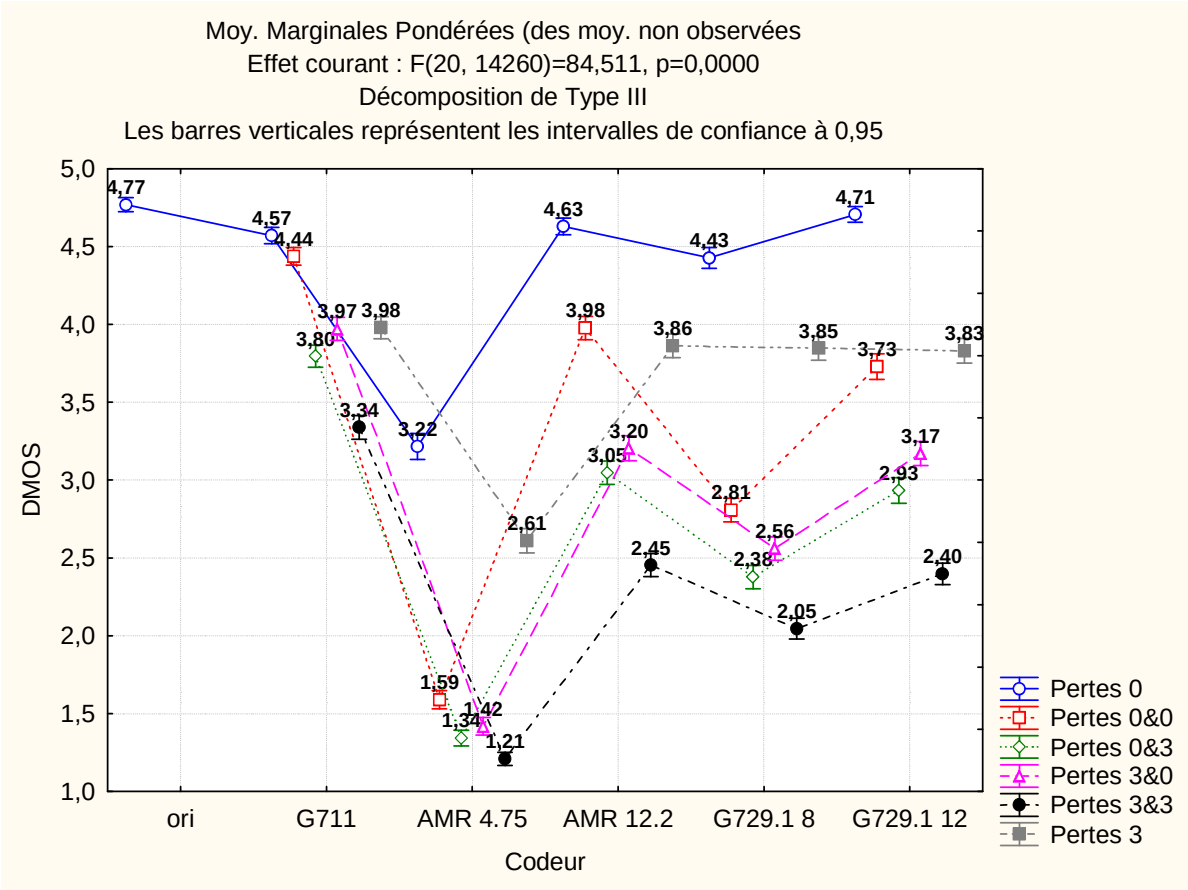

Figure 4.29 Résultats des différentes configurations de conférence pour chaque codeur narrowband

Il apparaît que les notes obtenues pour chaque codeur en conférence distribuée et sans perte de trames sont sans surprise les meilleures.

Ce résultat est confirmé suite à un test HSD de Tukey effectué sur les données. Les équivalences sont présentées dans le Tableau 4.27.

N°	G.711	AMR 4.75	AMR 12.2	G.729.1 8	G.729.1 12
1	0 ⇔ 0&0	0	0	0	0
2	3⇔ 3&0	3	0&0 ⇔ 3	3	3 ⇔ 0&0
3	0&3	0&0 ⇔ 3&0	3&0 ⇔ 0&3	0&0	3&0
4	3&3	3&0 ⇔ 0&3	3&3	3&0	0&3
5		0&3 ⇔ 3&3		0&3	3&3
6				3&3	

Tableau 4.27 Classement des différentes configurations de conférence pour chaque codeur

La configuration distribuée sans PdT se classe première pour chaque codeur. On remarque que pour le G.711 que la configuration centralisée sans PdT lui est équivalente. Cela montre la bonne résistance du G.711 au transcodage.

Globalement pour les codeurs CELP, la configuration distribuée à 3% de PdT se classe second, avec même une égalité avec la configuration centralisée sans PdT pour les plus hauts débits. Pour le G.711, la configuration distribuée à 3% de PdT se classe en 2nde position.

La tendance concernant le classement des configurations centralisées avec PdT est : 0&0, 3&0, 0&3 et 3&3. Sans surprise pour les positions extrêmes, le classement de la configuration centralisée 3&0 devant celle à 0&3 peut s'expliquer par le fait que la tentative de reconstruction par le codeur de chaque canal sur la PdT est plus perceptible sur le contenu binaural que celle effectuée par le même codeur sur le contenu monophonique.

4.4.6 Résultats wideband

Les notes DMOS obtenues par chaque codeur, aux différents taux de PdT et selon les différents angles, sont illustrées sur la Figure 4.30.

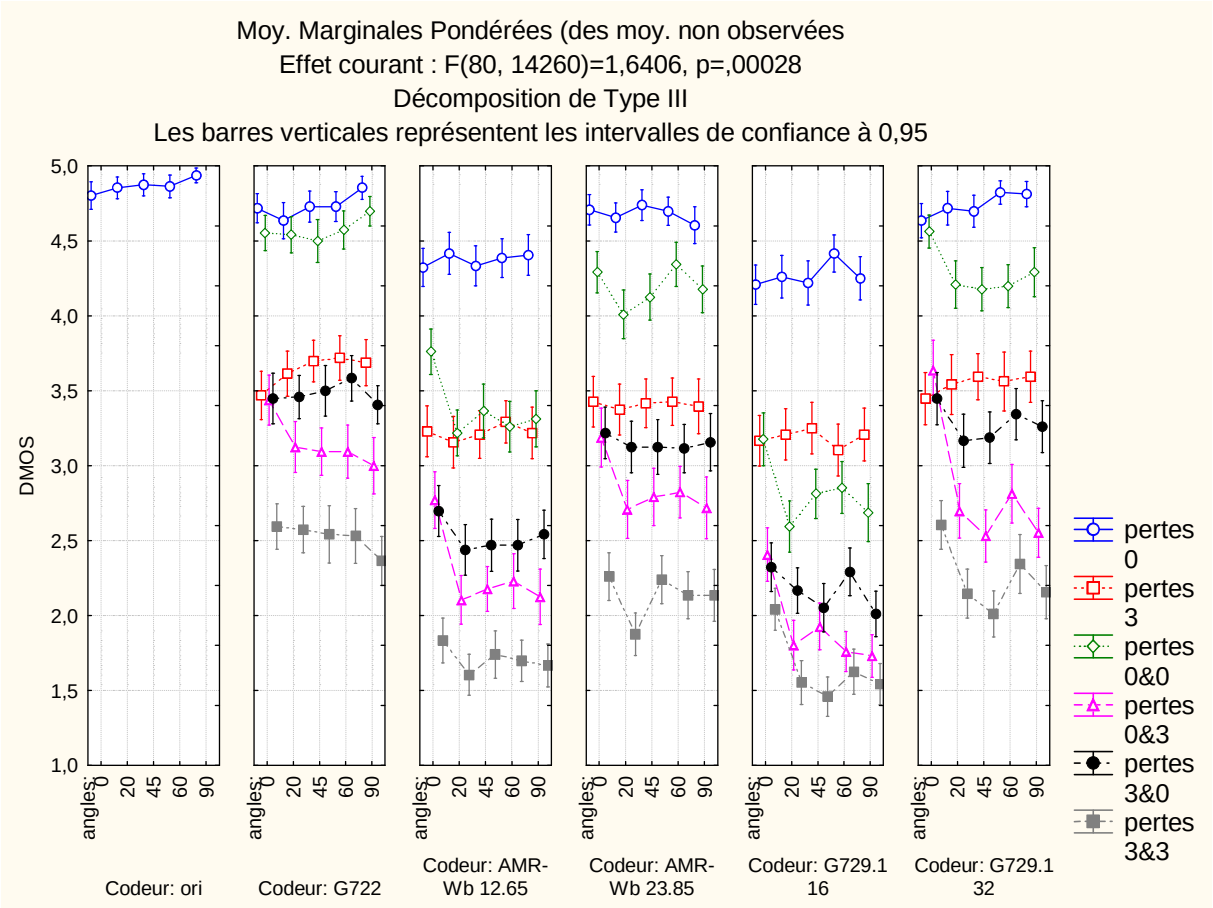

Figure 4.30 Illustration de l'impact de l'angle sur la qualité perçue des codeurs wideband pour les configurations de conférence

Il apparaît que globalement l'angle 0° se distingue de nouveau sur les configurations centralisées tout comme pour le test précédent. Sinon pour la plupart des autres angles, les résultats sont globalement uniformes. Plus précisément, le test HSD de Tukey confirme cette tendance avec les différences significatives suivantes :

- Pour le codeur AMR-WB à 12.65 kbits/s pour les configurations :
  - 0&0 entre l'angle 0° d'une part et les angles 20° et 60° d'autre part,
  - 0&3 entre l'angle 0° et les autres angles.
- Idem pour le codeur AMR-WB à 23.85 kbits/s pour les configurations :
  - 0&3 entre les angles 0° d'une part et 20° et 90° d'autre part,
- Idem pour le G.729.1 à 16 kbits/s pour les configurations :
  - 0&3 entre l'angle 0° et les autres,
  - 0&0 entre l'angle 0° et les angles 20° et 90°.
  - 3&3 entre l'angle 0° et les angles 20°, 45° et 90°.
- Idem pour le G.729.1 à 32 kbits/s pour les configurations :

- 0&3 entre l'angle 0° et les autres,
- 3&3 entre l'angle 0° et les angles 20° et 45°.

La Figure 4.31 présente les résultats des codeurs wideband tous angles confondus pour chaque configuration.

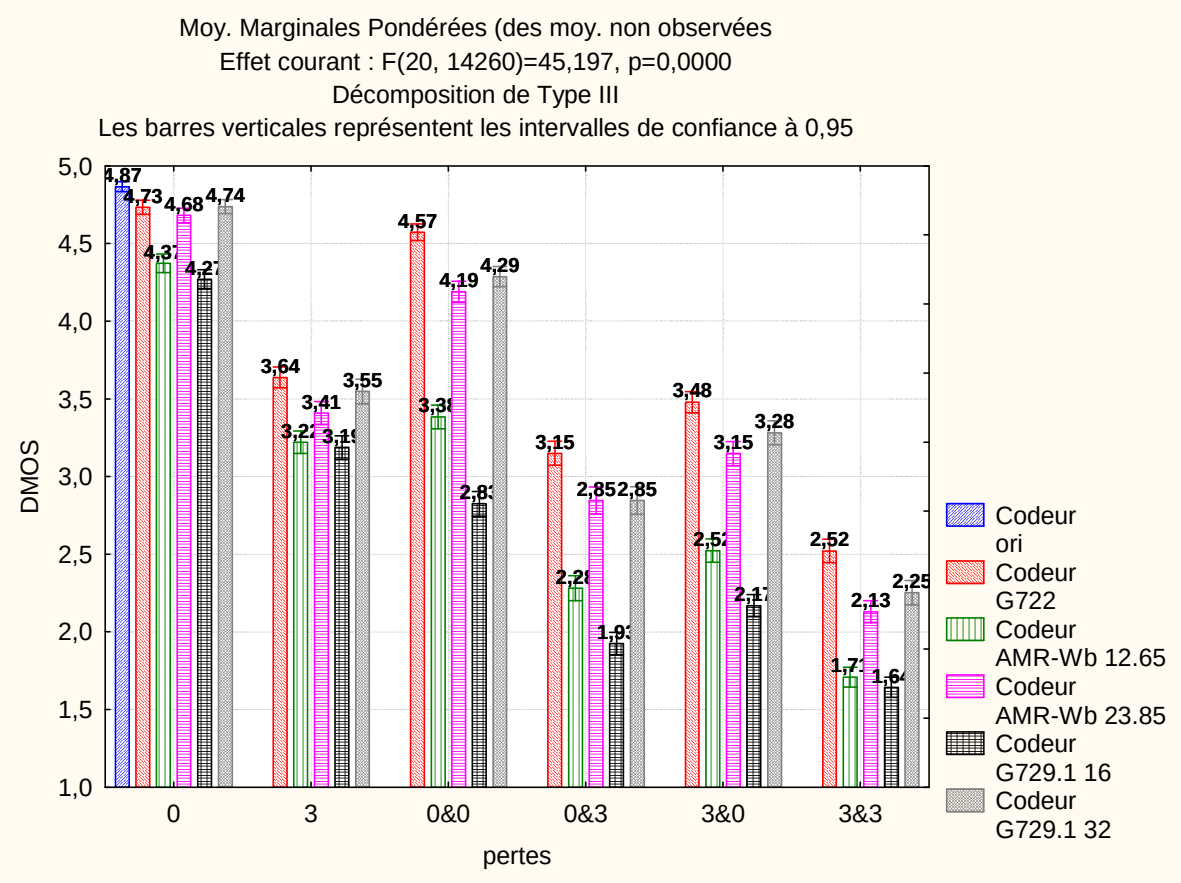

Figure 4.31 Résultats des codeurs wideband tous angles confondus par configuration de conférence

Il apparaît que la référence est la condition la mieux notée à 0% de PdT. L'augmentation de la PdT entraîne logiquement une diminution des notes de qualité. De même, pour une même famille de codeurs, les codeurs haut-débit sont mieux notés que les bas-débit. Enfin il semble que le codeur G.722 soit le meilleur en configuration centralisée. Il est difficile d'extraire une tendance pour la configuration distribuée.

Les effets des facteurs du test (Codeur, Perte de trames et Angle) et leurs interactions sont confirmés ($p < 0.05$) par une analyse de variance ANOVA conduite sur les notes individuelles. Les résultats de l'analyse sont donnés par le Tableau 4.28.

Facteur	Degrés de liberté	F-ratio	p = Significativité
Codeur	5	871,2	0,00
pertes	5	4042,3	0,00
angles	4	40,4	0,00
Codeur*pertes	20	45,2	0,00
Codeur*angles	20	3,1	0,00
Codeur*pertes*angles	80	1,6	0,00

Tableau 4.28 Effets des différents facteurs pour le test de conférence narrowband

Les différentes équivalences selon les différentes pertes de trames sont données dans le Tableau 4.29 et confirment de manière significative les tendances énoncées ci-dessus. Elles ont été obtenues grâce au test HSD de Tukey.

N°	0	3	0&0	0&3	3&0	3&3
1	Ori ⇔ G.729.1 32 ⇔ G.722	G.722 ⇔ G.729.1 32	G.722	G.722	G.722	G.722
2	G.729.1 32 ⇔ G.722 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85	G.729.1 32 ⇔ AMR-WB 23.85
3	AMR-WB 12.65 ⇔ G.729.1 16	AMR-WB 12.65 ⇔ G.729.1 16	AMR-WB 12.65	AMR-WB 12.65	AMR-WB 12.65	AMR-WB 12.65 ⇔ G.729.1 16
4			G.729.1 16	G.729.1 16	G.729.1 16	

Tableau 4.29 Classement des codeurs wideband tous angles confondus par configuration de conférence

Pour la conférence distribuée à 0% de PdT, on retrouve globalement les équivalences établies pour l'écoute diotique à 0% de PdT, Tableau 4.9. Le même type d'équivalence se retrouve entre la conférence distribuée à 3% de PdT et l'écoute diotique à 3% de PdT, Tableau 4.10. On remarque donc que la spatialisation ne semble logiquement pas avoir eu d'effet sur la qualité audio ressentie par les sujets au vu des notes obtenues.

Pour le classement de la configuration centralisée 0&0, le G.722 s'est détaché des deux codeurs CELP haut-débit par rapport aux résultats en binaural à 0% de PdT, Tableau 4.19.

Pour le classement de la configuration centralisée 0&3, on retrouve le classement établi en binaural à 3% de PDT, Tableau 4.19. Les classements restent identiques pour les configurations 0&0, 0&3 et 3&0. Pour la configuration centralisée 3&3, le G.729.1 à 16 kbits/s est au niveau de l'AMR-WB à 12.65 kbits/s.

Dans le cadre de la conférence distribuée, les codeurs, semblant les plus adaptés, sont les G.729.1 à 32 kbits/s, G.722 et AMR-WB à 23.85 kbits/s. Par contre pour la conférence centralisée avec pont mixeur, le codeur G.722 est incontestablement le meilleur, étant performant sur le double encodage/décodage et en présence de PdT. Il est à noter que globalement les codeurs les mieux notés sont ceux disposant du plus de débit.

La Figure 4.32 présente les résultats des codeurs wideband tous angles confondus pour chaque configuration.

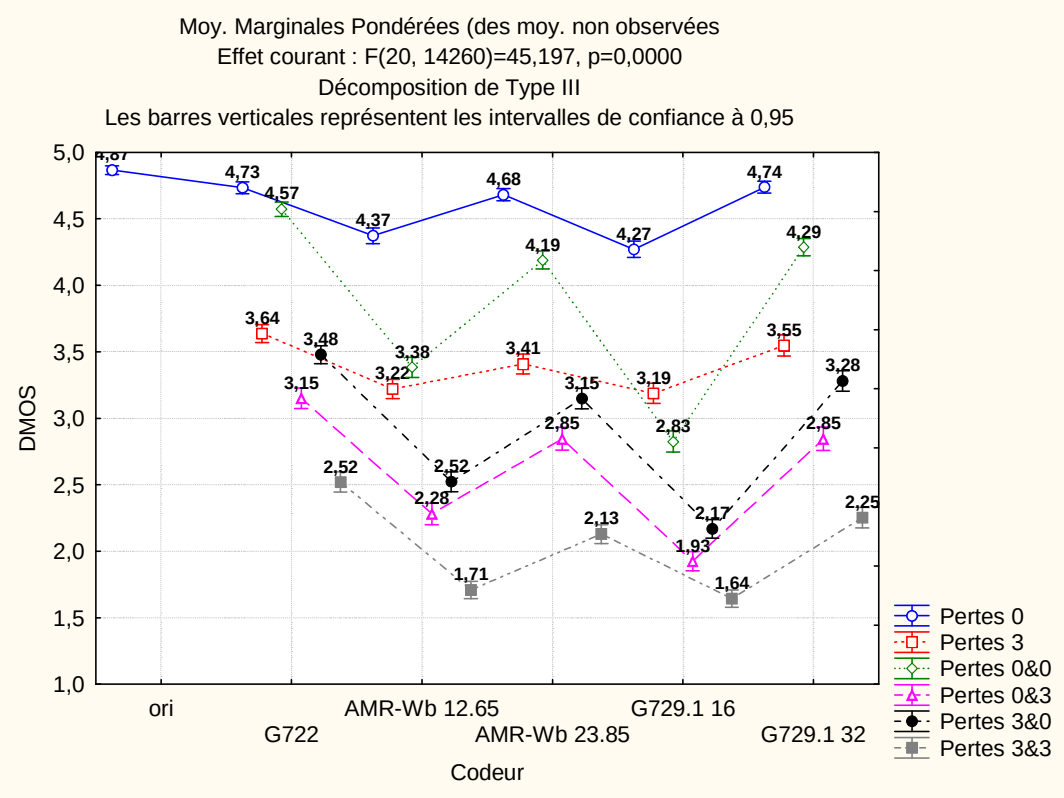

Figure 4.32 Résultats des différentes configurations de conférence pour chaque codeur wideband

On remarque la proximité des notes obtenues sans PdT pour le G.722 des deux configurations centralisée et distribuée. Cela montre sa bonne résistance au transcodage. La seule particularité notable par rapport aux résultats obtenus en narrowband est que pour les codeurs CELP, à haut-débit, la configuration centralisée sans PdT arrive devant la configuration distribuée à 3% de PdT.

Un test HSD de Tukey, dont les résultats d'équivalence sont donnés dans le Tableau 4.30, nous donne les différences significatives suivantes.

N°	G.722	AMR-WB 12.65	AMR-WB 23.85	G.729.1 16	G.729.1 32
1	0 ⇔ 0&0	0	0	0	0
2	3 ⇔ 3&0	0&0 ⇔ 3	0&0	3	0&0
3	0&3	3&0	3	0&0	3
4	3&3	0&3	3&0	3&0	3&0
5		3&3	0&3	0&3	0&3
6			3&3	3&3	3&3

Tableau 4.30 Classement des différentes configurations de conférence pour chaque codeur

4.4.7 Discussion et perspectives

Nous avons vu par ce test qu'il est possible de proposer un service de conférence audio spatialisée avec des codeurs monophoniques standard et un contenu mono-locuteur. Pour la configuration distribuée, nous obtenons logiquement une qualité proche de la référence, la spatialisation en tant qu'étape finale ne dégradant pas plus le signal transmis. Concernant la configuration centralisée, la dégradation est sans surprise notable par rapport à la configuration distribuée. Comme cela avait été vu pour le test précédent sur le transport des flux binauraux, l'impact dégradant du second étage de la conférence centralisée est important et confirmé.

Les codeurs G.711 et G.722 semblent bien adaptés pour la configuration de conférence centralisée en termes de complexité et de qualité. La qualité est nettement supérieure à celles de codeurs CELP et est jugée équivalente à celle de la configuration distribuée. Leur coût important en débit est cependant à prendre en compte.

Les codeurs AMR-WB à 23.85 kbits/s, G.729.1 à 32 kbits/s et G.722 à 64 kbits/s semblent quant à eux les plus adaptés à la conférence audio distribuée wideband quelle que soit la perte de trames. Ils ont une qualité jugée équivalente mais possèdent des atouts différents : les codeurs CELP ont globalement un débit deux fois plus faible que le G.722 mais une complexité quatre fois plus importante (~35 WMOPS contre 9 WMOPS). En narrowband, les codeurs G.711, AMR à 12.2 kbits/s et G.729.1 à 12 kbits/s semblent les plus intéressants quelle que soit la perte de trames. Le codeur PCM présente l'avantage d'une complexité très faible par rapport aux codeurs CELP (0.3 WMOPS contre ~16-21 WMOPS) mais un débit cinq fois plus important (64 kbits/s contre ~12 kbits/s).

Néanmoins pour la conférence centralisée et en période de parole simultanée, la qualité audio d'un contenu spatialisé avec plusieurs locuteurs risque d'être dégradée par le codage dual-mono des codeurs conversationnels monophoniques. Le test de la section suivante va évaluer la possibilité d'utiliser nos codeurs dans ce contexte.

4.5 Evaluation de la qualité de signaux de parole binauraux multi-locuteur encodés-décodés en dual-mono par des codeurs monophoniques de parole

Les tests précédents ont montré qu'il était possible de transporter un contenu binaural mono-locuteur en préservant une bonne qualité audio quelle que soit la configuration testée. L'objectif de ce test est maintenant d'évaluer la qualité audio et la qualité de spatialisation ressenties par un auditeur, suite à un codage dual-mono sur un contenu binaural multi-locuteur. Cela revient à tester le second étage de la conférence audio centralisée avec pont mixeur. Nous reprenons le test 4.3 mais cette fois en multi-locuteur.

De même que dans les sections précédentes, lors de cette étude, deux tests ont été menés, l'un avec des codeurs narrowband et l'autre avec des codeurs wideband.

4.5.1 Stimuli

Les 2 échantillons de test sont extraits de la base de données de France Télécom et durent 8 secondes. Ils sont constitués de deux phrases espacées par un silence, et sont échantillonnés à 16 kHz, pour une bande de fréquences de 8 kHz. Ils sont énoncés par deux hommes, prononçant chacun un échantillon différent. Suite à des pré-tests, ces échantillons nous semblaient les plus pertinents pour juger de la qualité audio et de la qualité de la spatialisation.

4.5.2 Conditions de test narrowband et wideband

Les traitements narrowband et wideband sont explicités dans les deux sections suivantes. Leur rôle est de simuler une transmission d'un contenu binaural multi-locuteur sur un réseau en utilisant un codage dual-mono avec des codeurs conversationnels normalisés. Aucune perte de trame n'a été introduite.

4.5.2.1 Conditions narrowband

Le traitement appliqué aux fichiers est illustré Figure 4.33. L'élément *Somme* permet d'effectuer une somme canal par canal de deux contenus binauraux.

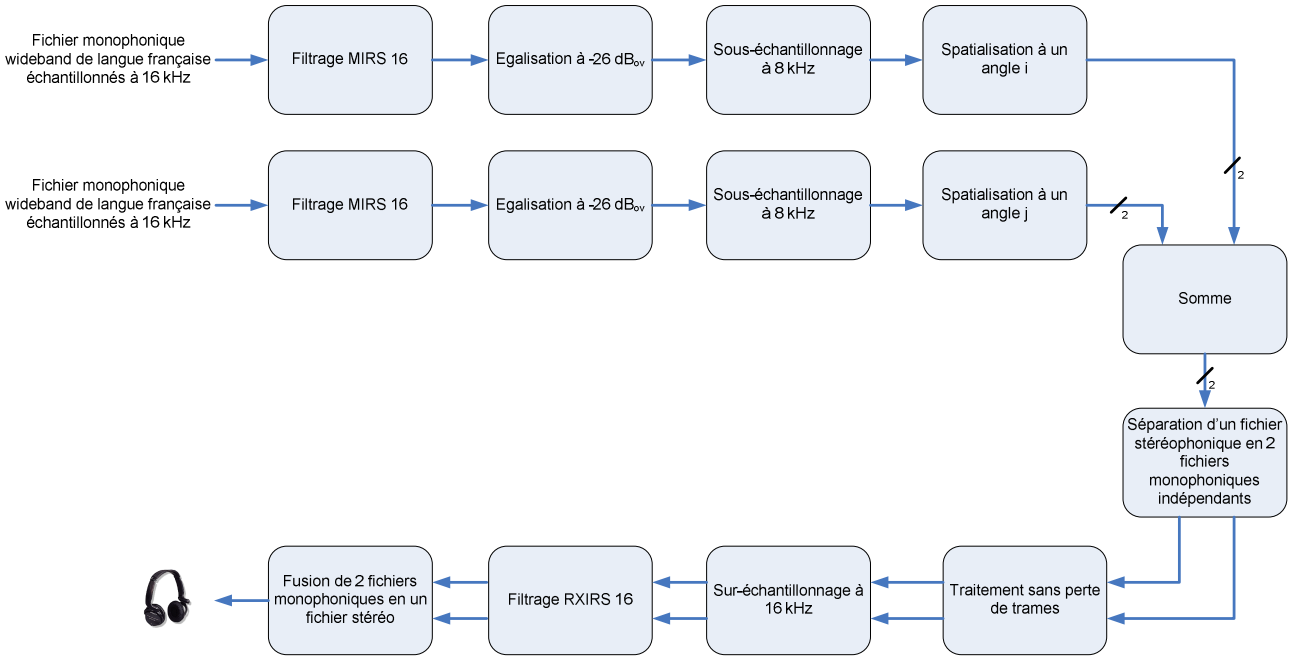

Figure 4.33 Traitement appliqué aux fichiers pour le test narrowband de transport de contenu binaural multi-locuteur

Le principe retenu est le même que pour le test narrowband de la section 4.3. Nous avons simplement rajouté une branche correspondant à un participant supplémentaire avant l'envoi sur le réseau.

Les filtrages MIRS 16 (*Modified Intermediate Reference System*) [42] et RXIRS 16 (*Receive-Side IRS*) [42] permettent de simuler des appareils de prise de son et de restitution numériques narrowband pour des signaux de fréquence d'échantillonnage initiale de 16 kHz.

Les blocs de sous-échantillonnage et sur-échantillonnage permettent de ré-échantillonner des contenus à des fréquences d'échantillonnage respectivement de 8 et 16 kHz, afin de pouvoir être

traités par les blocs suivants. L'égalisation à -26 dB_{ov} permet, comme son nom l'indique, d'égaliser en énergie les fichiers de départ pour que les sujets aient la même perception sonore en niveau quel que soit le fichier. Ces trois traitements sont effectués grâce aux filtres définis dans [42].

Les deux participants sont spatialisés à des angles différents (sauf pour la condition n°7) afin de se mettre dans le cas concret d'utilisation de la spatialisation pour lequel on évite de placer virtuellement deux locuteurs à la même place. Il est à noter que nous avons souhaité mélanger la simple et la double parole dans nos échantillons de test. Ainsi le 1er locuteur parle seul pendant les 4 premières secondes, puis les deux locuteurs parlent ensemble durant les 4 secondes suivantes et enfin le second locuteur conclut en parlant seul lui aussi pendant 4 secondes. Les conditions traitées par le bloc *Traitement sans perte de trames* du test sont données ci-dessous :

Numéros des conditions	Codeur	Pourcentage de perte de trame	Angles
1	Aucun	0	Voir section Processus expérimental pour le choix des combinaisons de deux angles retenues
2	G.711	0	
3	G.729.1 à 8 kbits/s	0	
4	G.729.1 à 12 kbits/s	0	
5	AMR 4.75 kbits/s	0	
6	AMR 12.2 kbits/s	0	
7	AMR 4.75 kbits/s	0	Les deux locuteurs sont spatialisés à 0°

Tableau 4.31 Liste des conditions du test de transport de contenu binaural narrowband multi-locuteur

La condition 1 est notre condition de référence spatialisée sans dégradation, aussi appelée *Ori_Spat ou Ori_Qua* par la suite, suivant que l'on étudie les résultats de spatialisation ou de qualité.

Il est à noter qu'une condition (la n° 7, appelée par la suite *Ref AMR 4.75*) est ajoutée par rapport aux tests narrowband précédents. Pour avoir une référence basse en termes de spatialisation, nous avons introduit cette condition où les deux locuteurs sont spatialisés au même angle : 0°.

4.5.2.2 Conditions wideband

Le traitement appliqué aux fichiers est illustré Figure 4.34.

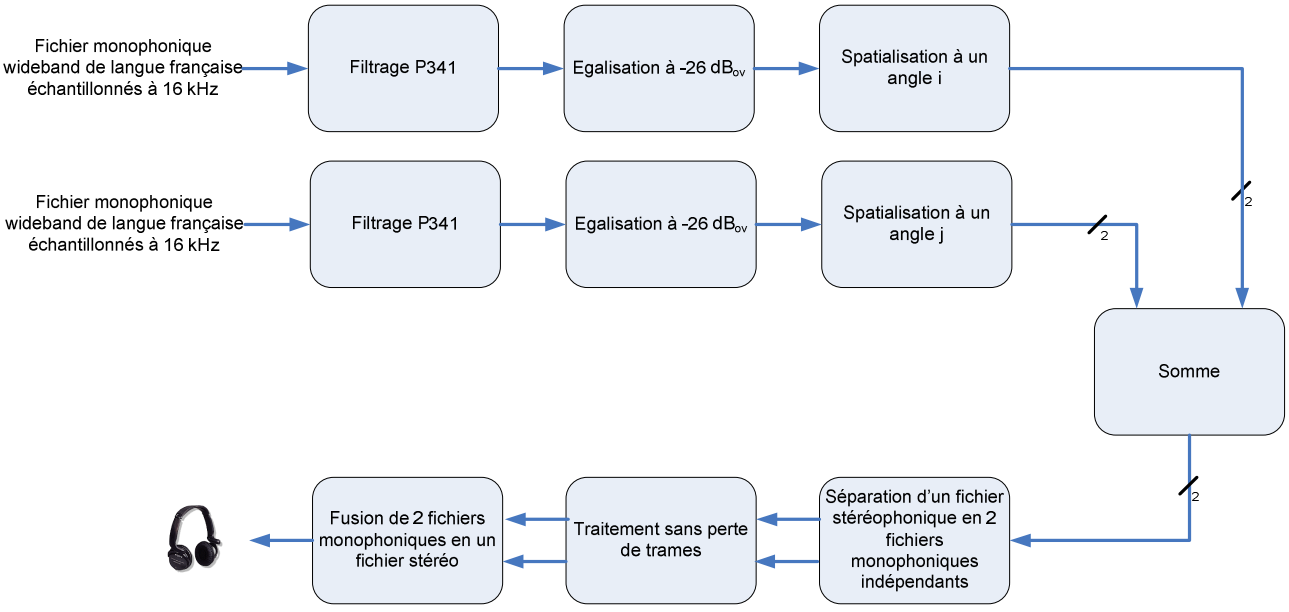

Figure 4.34 Traitement appliqué aux fichiers pour le test wideband de transport de contenu binaural multi-locuteur

Le filtrage *P.341* [42] permet de simuler l'appareil de prise de son et de restitution numériques. L'égalisation à -26 dB_{ov} est effectuée grâce à un filtre défini dans [42].

Tout comme pour le test narrowband, le principe retenu est le même que pour le test wideband de la section 4.3 avec simplement le rajout d'une branche correspondant à un participant supplémentaire avant l'envoi sur le réseau. Le mélange de simple et double parole est également utilisé. Les conditions traitées par le bloc *Traitement sans perte de trames* du test sont données ci-dessous :

Numéros des conditions	Codeur	Pourcentage de perte de trame	Angles
1	Aucun	0	Voir section Processus expérimental pour le choix des combinaisons de deux angles retenues
2	G.722	0	
3	G.729.1 à 16 kbits/s	0	
4	G.729.1 à 32 kbits/s	0	
5	AMR-WB 12.65 kbits/s	0	
6	AMR-WB 23.85 kbits/s	0	
7	AMR-WB 6.60 kbits/s	0	Les deux locuteurs sont spatialisés à 0°

Tableau 4.32 Liste des conditions du test de transport de contenu binaural wideband multi-locuteur

La condition 1 est notre condition de référence spatialisée sans dégradation, aussi appelée *Ori_Spat* ou *Ori_Qua* par la suite suivant que l'on étudie les résultats de spatialisation ou de qualité.

Tout comme le test narrowband, il est à noter qu'une référence basse en spatialisation est ajoutée, appelée par la suite *Ref AMR-WB 6.60*.

4.5.3 Procédure expérimentale

Notre objectif dans ce test est d'évaluer la qualité audio ainsi que la qualité de la spatialisation. Nous nous sommes posé plusieurs contraintes fortes :

- Pour juger de la qualité de la spatialisation, il est nécessaire :
  - d'avoir une référence.
  - d'avoir des experts qui savent évaluer les dégradations. Du fait du faible nombre d'experts à notre disposition, il en découle que nous devons choisir une méthode de test qui nous permette d'avoir des résultats significatifs avec très peu de sujets.
- Quel que soit l'élément testé (qualité de spatialisation ou audio), nous ne pouvons pas multiplier les conditions de test (notamment en faisant varier les angles testés ou les locuteurs) du fait de la durée et de la difficulté des tests.
- Il serait intéressant d'avoir une comparaison directe des différents codeurs.

La méthode, qui nous a semblé la plus appropriée et répondant à tous ces critères, est la méthode MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*) [41].

Nous avions dans un premier temps décidé de comparer dans une même session à la fois la qualité de la spatialisation et la qualité audio mais des pré-tests réalisés sur deux sujets ont montré qu'il était difficile de noter les deux composantes en même temps. Nous avons donc décidé de séparer chaque test en deux sessions MUSHRA : l'une évaluant la qualité audio, l'autre la qualité de la spatialisation.

Concernant la session de qualité de spatialisation, nous avons demandé aux sujets de noter les différences par rapport à la scène de référence en termes :

- De flou (notion introduite par [14]) : si la largeur de la source semble plus grande par rapport à la référence.

- De précision des sources (notion introduite par [14]) : si l'emplacement de la source est différent de celui de la référence.
- Et d'externalisation ressentie : son en dehors ou dans la tête par rapport à la référence.

En guise de remarque, nous avons rejeté les méthodes de mesure d'intelligibilité de type CRM (Coordinate Response Measure) [15] et VCET (Voice Communication Effectiveness Test) [15]. Ces méthodes permettent de manière indirecte de mesurer la qualité de la spatialisation mais ne permettent pas d'évaluer par exemple si la scène sonore a bougé. Elles permettent surtout de montrer l'apport du son spatialisé par rapport à un son monophonique ou stéréophonique. De plus, nous avons remarqué que les dégradations apportées par le codage ou la perte de trame ne sont pas suffisantes pour dégrader l'intelligibilité.

L'interface utilisée pour les tests est donnée Figure 4.35. La référence explicite est marquée "Ref". Elle correspond à la condition n°1. L'auditeur doit noter le long de l'échelle continue de qualité audio ou de qualité de spatialisation (suivant la session) chacune des différentes étiquettes 1, 2, … 7 par rapport à la référence. Pour cela, il utilise les curseurs au-dessus de chacune des étiquettes notées de 1 à 7. La référence est aussi cachée parmi les versions ainsi que l'ancre basse de spatialisation (la condition n°7). Pour chaque séquence audio, les 7 conditions à noter ont été aléatoirement distribuées pour chaque auditeur derrière les étiquettes de 1 à 7. De plus, l'ordre de présentation des séquences audio est différent d'un auditeur à un autre. On s'attend à ce que l'étiquette correspondant à la condition n°1 ait la meilleure note dans les deux sessions, puisque cette condition n'a pas subi de dégradation. De même, la condition n°7 devrait être l'ancre basse de la session de spatialisation.

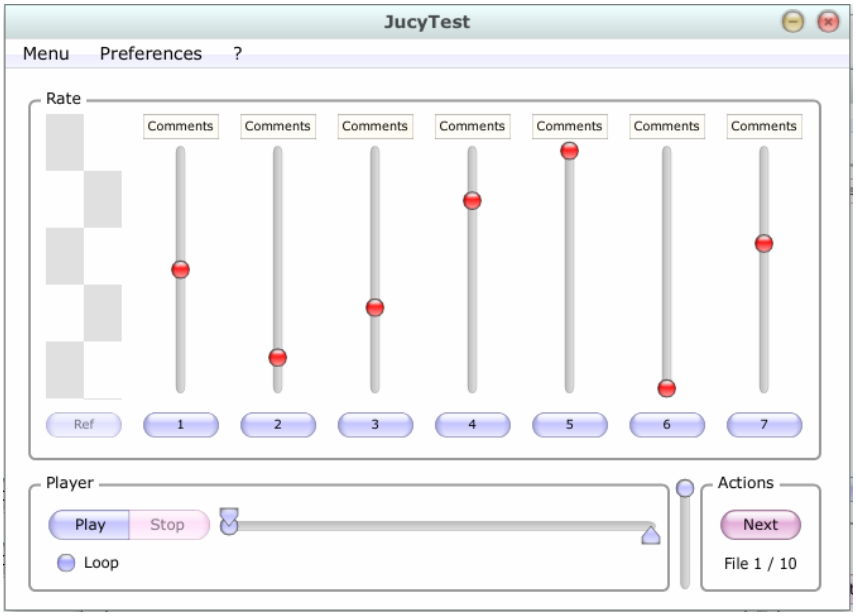

Figure 4.35 Interface logicielle du test MUSHRA

L'échelle d'évaluation de la procédure MUSHRA est donnée Tableau 4.15. Cette échelle continue sera utilisée pour les deux sessions (qualité de la parole et qualité de la spatialisation) des tests.

Qualité de la parole/de la spatialisation	Note
Qualité excellente	80-100
Qualité bonne	60-79
Qualité passable (dans le sens, ça peut passer)	40-59
Qualité médiocre	20-39
Qualité mauvaise	0-19

Tableau 4.33 Echelle de note MUSHRA

Les deux tests se sont déroulés dans une salle d'écoute acoustiquement isolée, suivant les recommandations de la norme P.800 [47].

Pour la restitution sonore, un casque STAX Signature SR-404 ouvert et son amplificateur SRM-006t ont été utilisés. Le son numérique en sortie de la carte son interne (Digigram VX 222) est converti par un convertisseur analogique-numérique 24 bits (3Dlab DAC 2000) avant d'être envoyé vers l'amplificateur du casque.

Chaque auditeur est libre d'ajuster le niveau sonore au début de l'écoute. Le niveau est ensuite figé pendant la session de test.

Les sujets effectueront un MUSHRA avec un apprentissage d'une séquence audio et avec 10 séquences audio de test. On rappelle que pour chaque séquence audio, ils auront 7 notes à donner. Voici les caractéristiques angulaires des séquences audio :

Combinaison d'angles	Angle du 1er locuteur	Angle du second locuteur	Catégorie des angles des deux locuteurs
1	0°	20°	Même côté
2	0°	45°	Même côté
3	-20°	20°	Côtés opposés
4	-20°	60°	Côtés opposés
5	-45°	-20°	Même côté
6	-45°	45°	Côtés opposés
7	45°	60°	Même côté
8	-60°	0°	Même côté
9	-60°	-20°	Même côté
10	-60°	60°	Côtés opposés

Tableau 4.34 Liste des combinaisons d'angles

Ces combinaisons ont été choisies pour essayer de couvrir globalement l'ensemble des possibilités : 2 participants du même côté ou un participant de chaque côté de l'auditeur. Le but est de voir si cela a un impact sur les résultats de qualités.

4.5.4 Les auditeurs

Les auditeurs, au nombre de 10 pour le test wideband et 9 pour le test narrowband, sont choisis parmi les experts du codage et de la spatialisation, selon la recommandation de la méthodologie de test.

Une procédure de rejet a été mise en place pour ce test. Tout sujet mettant moins de 90 à la référence haute que cela soit pour la session de qualité de spatialisation ou de qualité audio est exclu des résultats. Un sujet pour chaque test a été rejeté. Le nombre final de sujets retenus est donc 9 pour le test wideband et 8 pour le test narrowband.

4.5.5 Résultats narrowband

4.5.5.1 Résultats narrowband de la session de qualité audio

Nous comparons, Figure 4.36, les deux types de configuration angulaire sur les résultats de la session de qualité audio du test narrowband :

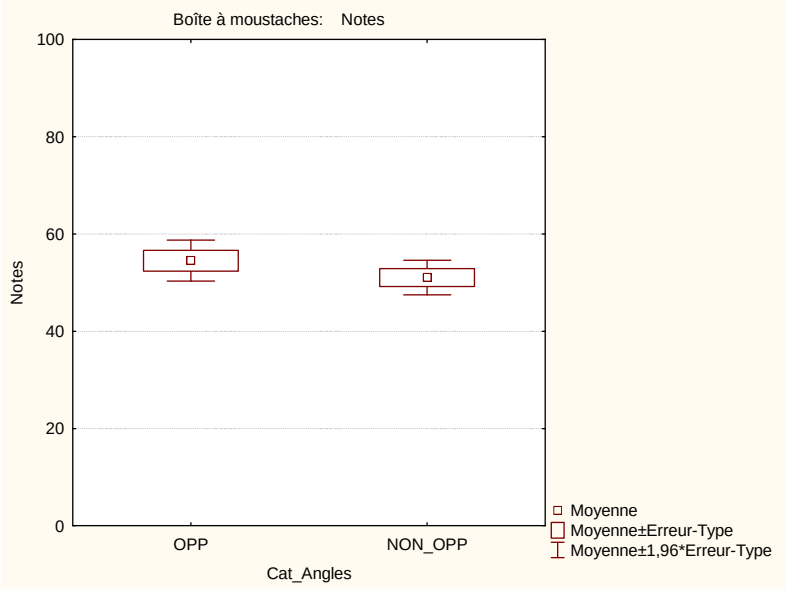

Figure 4.36 Comparaison des deux types de configuration angulaire sur les résultats de la session de qualité audio du test narrowband

Nous pouvons noter que les notes obtenues pour le type d'angle "NON_OPP" (même côté) sont moins bonnes que les notes obtenues pour le type d'angle "OPP" (côtés opposés). Cependant un t-test a montré que la différence entre les deux types d'angles n'est pas significative ($t(558)=1.23$, $p= 0,22$). Les sujets ont donc jugé que la qualité perçue ne variait pas suivant les deux types de configurations angulaires. Cette tendance se retrouvera dans les paragraphes suivants pour la spatialisation ainsi que pour le test wideband. Nous reviendrons sur ce point dans la discussion.

Les résultats obtenus pour les différents codeurs sont présentés sur la Figure 4.37, quelle que soit la configuration angulaire.

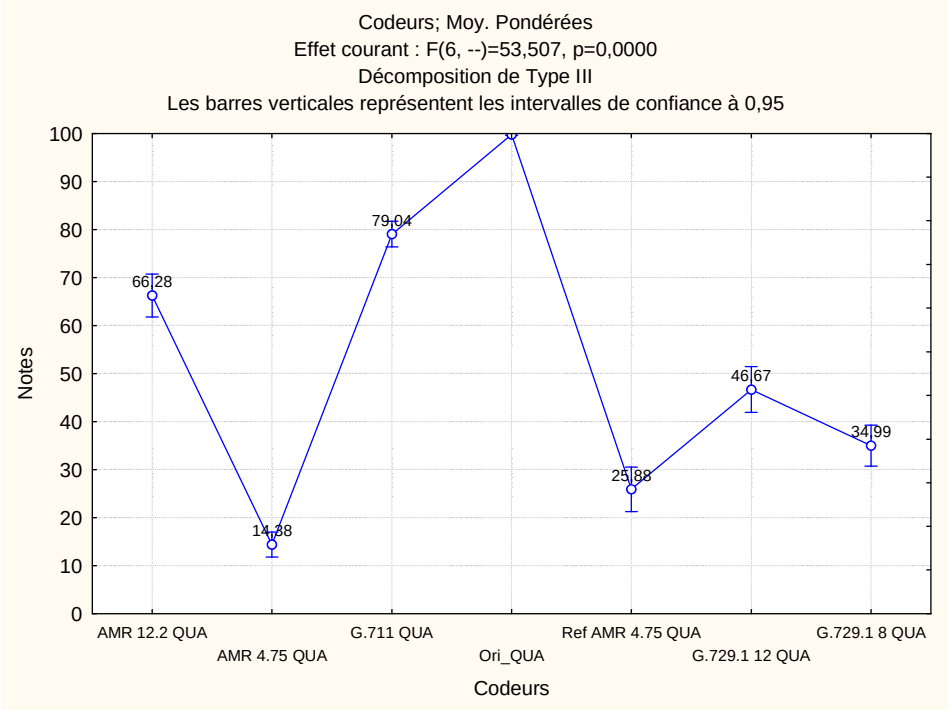

Figure 4.37 Résultats des codeurs narrowband tous angles confondus pour la session de qualité audio

Il apparaît que la référence est la condition la mieux notée et que pour une même famille de codeurs, les codeurs haut-débit sont logiquement mieux notés que les bas-débit. Seuls deux codeurs ont une qualité audio jugée bonne : le codeur G.711 (proche de la qualité excellente) et le codeur AMR à 12.2 kbits/s. Tous les autres codeurs ont une qualité audio jugée au mieux passable et nettement moins bonne que les codeurs G.711 et AMR à 12.2 kbits/s. Il est à noter que la version de l'AMR 4.75 avec spatialisation à des positions différentes a de moins bons résultats en termes de qualité audio que la version avec spatialisation à 0°. Nous y reviendrons dans la discussion et dans le test suivant.

Une analyse de variance ANOVA a confirmé la significativité de l'impact du codage ($p < 0.05$). Les résultats de l'ANOVA sont donnés par le Tableau 4.35 :

Facteurs	Degrés de liberté	F-ratio	P =Significativité
Codeurs	6	53.5071	0.00
Sujets	7	2.48	0.00
Combinaison d'angles	9	1.8491	0.08
Codeurs*Angles	54	1.0291	0.42

Tableau 4.35 Effets des différents facteurs pour le test de conférence narrowband pour la session de qualité audio du test narrowband

Nous n'avons par contre pas d'effet de la combinaison d'angles, ce qui sous-entend que les notes obtenues sont indépendantes de la combinaison d'angles.

Le test HSD de Tukey effectué sur ces données a montré que tous les codeurs étaient différents deux à deux. Le classement des codeurs est donc :

N°	Codeurs
1	Ori
2	G.711
3	AMR 12.2
4	G.729.1 12
5	G.729.1 8
6	Ref AMR 4.75
7	AMR 4.75

Tableau 4.36 Classement des codeurs narrowband tous angles confondus pour la session de qualité audio

4.5.5.2 Résultats narrowband de la session de spatialisation

Nous comparons, Figure 4.38, les deux types de configuration angulaire sur les résultats de la session de qualité de la spatialisation du test narrowband :

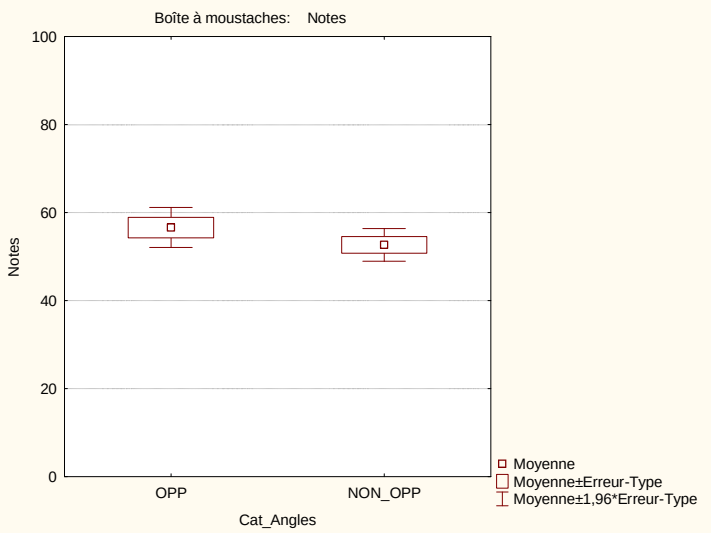

Figure 4.38 Comparaison des deux types de configuration angulaire sur les résultats de la session qualité de spatialisation narrowband

Un t-test a de nouveau montré que les deux types d'angles ont obtenu des résultats similaires (t(558)=1.31, p= 0,19) pour la session narrowband de qualité de spatialisation.

Nous présentons sur la Figure 4.39 les résultats obtenus pour les différents codeurs.

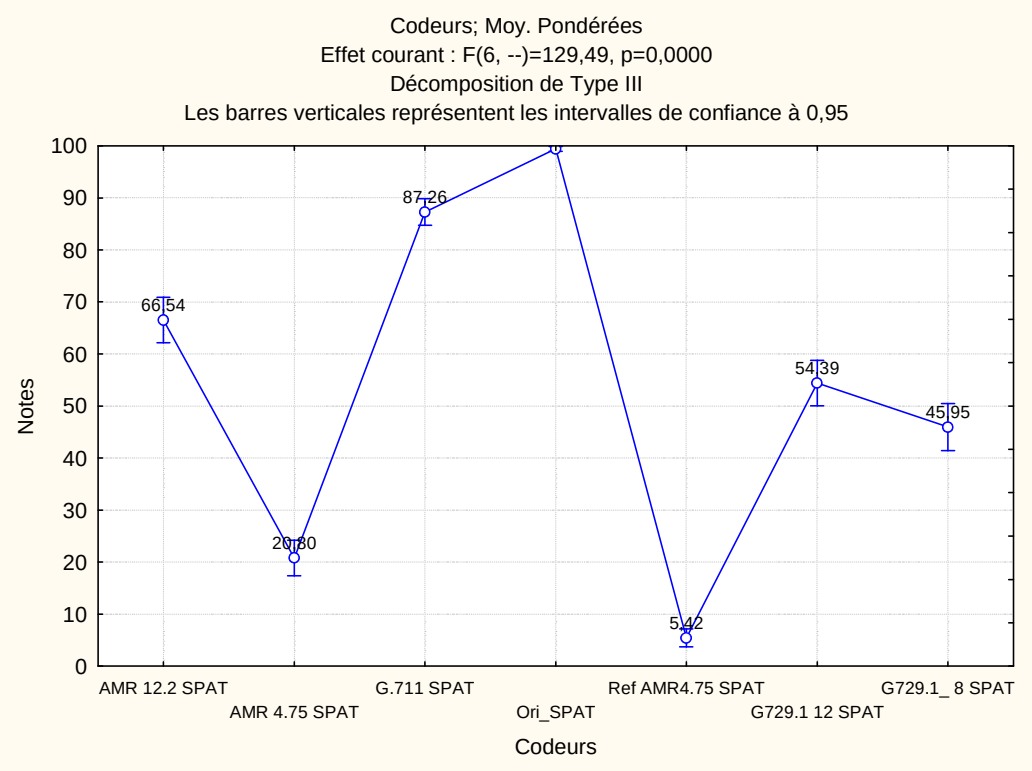

Figure 4.39 Résultats des codeurs narrowband tous angles confondus pour la session de qualité de spatialisation

On peut remarquer grâce à la Figure 4.39 que les références haute (la condition n°1) et basse (la condition n°7) sont bien retrouvées par les sujets. Pour une même famille de codeurs, les codeurs haut-débit sont logiquement mieux notés que les bas-débit.

On peut remarquer qu'un seul codeur a une spatialisation jugée excellente : le codeur G.711. Le codeur AMR à 12.2 kbits/s a quant à lui une spatialisation jugée bonne. Tous les autres codeurs ont une qualité jugée au mieux passable.

Une analyse de variance ANOVA a confirmé la significativité de l'impact du codage ($p < 0.05$). Les résultats de l'ANOVA sont donnés par le Tableau 4.37 :

Facteurs	Degrés de liberté	F-ratio	p =Significativité
Codeurs	6	129.4860	0.00
Sujets	7	3.69	0.00
Combinaison d'angles	9	1.8393	0.08
Codeurs*Angles	54	2.1035	0.00

Tableau 4.37 Effets des différents facteurs pour le test de conférence narrowband pour la session de qualité de spatialisation du test narrowband

Nous avons de nouveau un effet du codage mais pas d'effet de la combinaison d'angles.

Le test HSD de Tukey effectué sur ces données a montré que tous les codeurs étaient différents deux à deux. Le classement des codeurs est donc :

N°	Codeurs
1	Ori
2	G.711
3	AMR 12.2

4	G.729.1 12
5	G.729.1 8
6	AMR 4.75
7	Ref AMR 4.75

Tableau 4.38 Classement des codeurs narrowband tous angles confondus pour la session de qualité de spatialisation

4.5.6 Résultats wideband

4.5.6.1 Résultats wideband de la session de qualité

Nous comparons les deux types de configuration angulaire sur les résultats de la session de qualité audio du test wideband, Figure 4.40 :

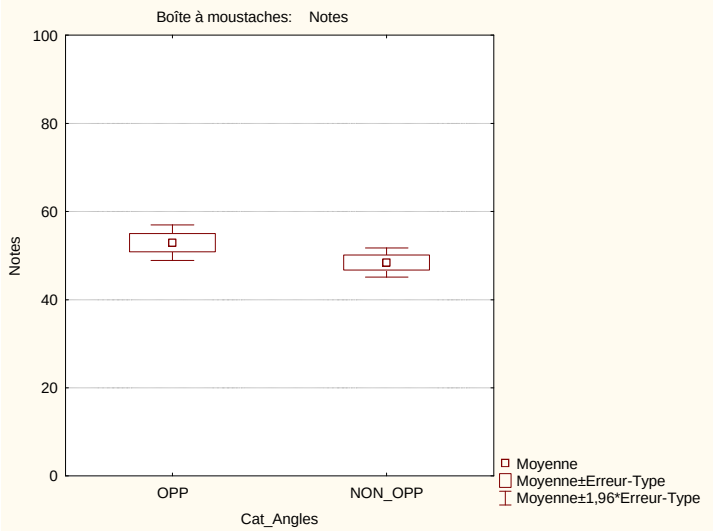

Figure 4.40 Comparaison des deux types de configuration angulaire sur les résultats de la session de qualité audio du test wideband

A nouveau, un t-test a montré que les deux types d'angles ont obtenu des résultats similaires ($t(628)=1.68$, $p= 0,09$).

Nous présentons sur la Figure 4.41 les résultats obtenus pour les différents codeurs.

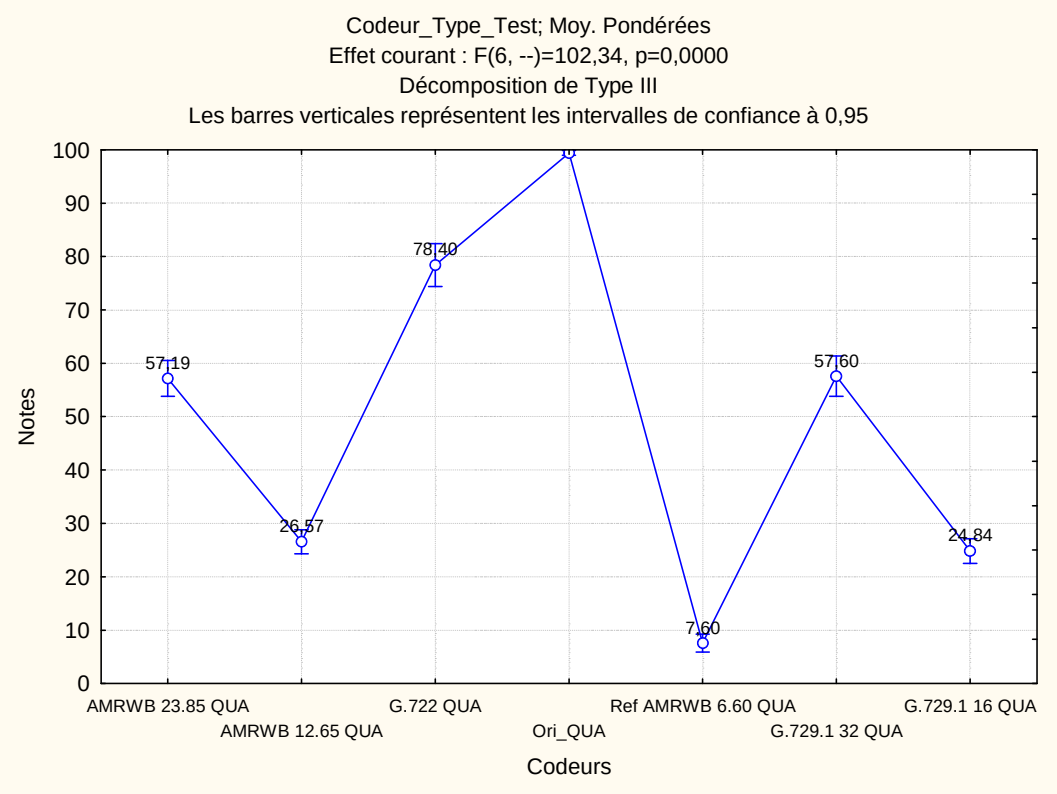

Figure 4.41 Résultats des codeurs wideband tous angles confondus pour la session de qualité audio

On peut remarquer que la référence haute (la condition n°1) est la condition la mieux notée. Pour une même famille de codeurs, les codeurs haut-débit sont logiquement mieux notés que les bas-débit.

On peut remarquer que seul un codeur a une qualité jugée bonne : le codeur G.722 (proche de la qualité excellente). Tous les autres codeurs ont une qualité jugée au mieux passable (proche de bonne pour les codeurs AMR-WB à 23.85 kbits/s et G.729.1 à 32 kbits/s).

Une analyse de variance ANOVA a confirmé l'effet significatif du codage et a montré qu'en wideband la combinaison d'angles avait eu un impact sur le jugement des sujets.

Facteurs	Degrés de liberté	F-ratio	p =Significativité
Codeurs	6	102.3442	0.00
Combinaison d'angles	9	8.3176	0.00
Sujets	8	1.8530	0.09
Codeurs* Combinaison d'angles	54	2.9785	0.00

Tableau 4.39 Effets des différents facteurs pour le test de conférence narrowband pour la session de qualité audio du test wideband

Il est à noter qu'il n'y a pas d'effet sujet dans ce test, ce qui signifie qu'ils ont noté de manière similaire la qualité audio.

Le test HSD de Tukey effectué sur ces données a montré que certains codeurs étaient équivalents. Le classement d'équivalence des codeurs est donc :

N°	Codeurs
1	Ori
2	G.722
3	G.729.1 32 ⇔ AMR-WB 23.85
4	G.729.1 16 ⇔ AMR-WB 12.65
5	Ref AMR-WB 6.60

Tableau 4.40 Classement des codeurs narrowband tous angles confondus pour la session de qualité audio

On obtient deux classes d'équivalence entre d'une part les codeurs CELP haut-débit et d'autre part les codeurs CELP bas-débit.

4.5.6.2 Résultats wideband de la session de spatialisation

Nous comparons les deux types de configuration angulaire sur les résultats de la session spatialisation du test wideband, Figure 4.42 :

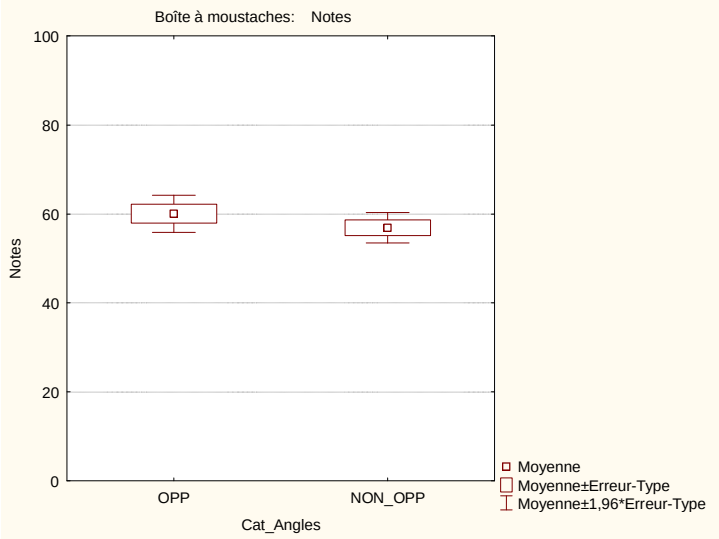

Figure 4.42 Comparaison des deux types de configuration angulaire sur les résultats de la session de qualité de spatialisation du test wideband

Un t-test a montré que les deux groupes d'angles ont obtenu des résultats similaires ($t(628)=1.15$, $p= 0,25$).

Nous présentons sur la Figure 4.43, les résultats obtenus pour les différents codeurs.

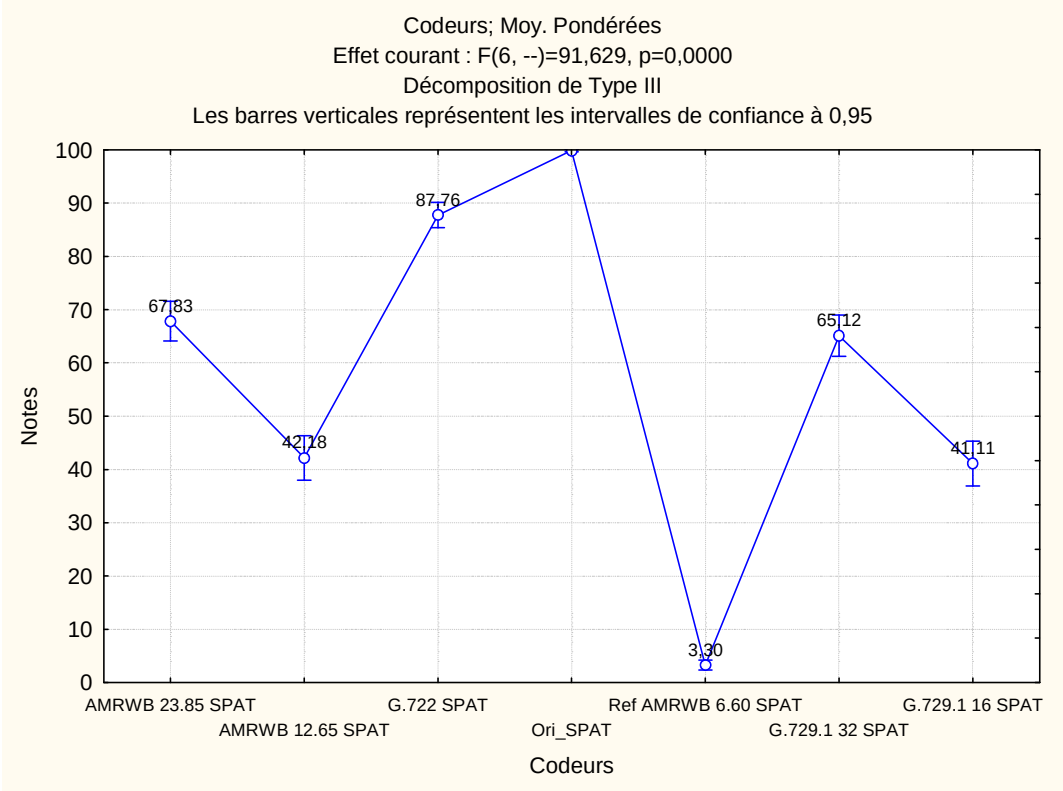

Figure 4.43 Résultats des codeurs wideband tous angles confondus pour la session de qualité de spatialisation

On peut remarquer que les références haute (la condition n°1) et basse (la condition n°7) sont bien retrouvées par les sujets. Pour une même famille de codeurs, les codeurs haut-débit sont mieux notés que les bas-débit.

On peut remarquer qu'un seul codeur a une spatialisation jugée excellente : le codeur G.722. Les codeurs haut-débit ont quant à eux une spatialisation jugée bonne. Tous les autres codeurs ont une qualité jugée au mieux passable.

Une analyse de variance a confirmé l'effet significatif ($p<0.05$) du codage mais aussi de la combinaison d'angles et cette fois des sujets.

Facteurs	Degrés de liberté	F-ratio	p =Significativité
Codeurs	6	91.6292	0.00
Combinaison d'angles	9	3.1395	0.00
Sujets	8	5.0373	0.00
Codeurs* Combinaison d'angles	54	1.5645	0.01

Tableau 4.41 Effets des différents facteurs pour le test de conférence narrowband pour la session de qualité de spatialisation du test narrowband.

Le test HSD de Tukey effectué sur ces données a montré que certains codeurs étaient équivalents. Le classement d'équivalence des codeurs est donc :

N°	Codeurs
1	Ori
2	G.722
3	G.729.1 32 ⇔ AMR-WB 23.85
4	G.729.1 16 ⇔ AMR-WB 12.65
5	Ref AMR-WB 6.60

Tableau 4.42 Classement des codeurs narrowband tous angles confondus pour la session qualité de spatialisation

On obtient deux classes d'équivalence entre d'une part les codeurs CELP haut-débit et d'autre part les codeurs CELP bas-débit.

4.5.7 Discussion et perspectives

Tout d'abord, il est à noter que ce test a été effectué en utilisant uniquement deux phrases (une pour chaque locuteur) afin d'éviter une durée de test trop importante. Cela peut donc avoir un impact sur les résultats obtenus ici. Néanmoins comme nous allons le voir dans cette section, les différences entre les codeurs sont relativement nettes et recoupent certains résultats vus auparavant.

Le premier résultat, qui ressort de ces tests, est la forte similitude entre les classements de qualité audio et de qualité de spatialisation, quel que soit le test. Deux tests de Pearson ont montré une forte corrélation entre les moyennes obtenues par chaque codeur dans chacune des deux sessions, de l'ordre de 95 % ($p<0.0011$) pour le test narrowband et de l'ordre de 97 % ($p<0.001$) pour le test wideband. Soulignons que la plupart des sujets ont relevé la difficulté de noter uniquement la spatialisation en présence de dégradations apportées par le codage dual-mono sur un contenu binaural multi-locuteur. Le codage dual-mono a donc dégradé à la fois la qualité de la spatialisation et la qualité audio dans des proportions similaires.

Un résultat intéressant est que la spatialisation en période multi-locuteur pourrait peut-être même accentuer l'impression de dégradation globale suite à l'encodage dual-mono (voir les résultats de qualité audio ressentie pour le codeur AMR à 4.75 kbits/s, suivant que la spatialisation des deux locuteurs est faite ou non aux mêmes emplacements). Afin de pouvoir mieux évaluer l'impact de la spatialisation sur les résultats en termes de qualité audio, un test de transport de contenu monophonique multi-locuteur a été mené dans la section 4.5.8. Nous obtiendrons ainsi en même temps un classement de nos codeurs en multilocuteur sur un contenu monophonique.

Il est par contre possible, que dans le cadre de la conférence audio distribuée, la qualité audio ressentie suite à la spatialisation de signaux dégradés serait jugée meilleure que leur simple somme monophonique. Cela est à tester.

Concernant les classements de qualité audio, les résultats obtenus en narrowband sont identiques à ceux obtenus dans la section 4.3.5 (transport de contenu binaural mono-locuteur sans perte de trame). Les résultats ne nous permettent pas de montrer clairement que l'écart s'est creusé entre notamment le codeur G.711 et les autres entre ce test et le test 4.3, puisque les deux types de test sont différents. Néanmoins, les notes DMOS obtenues pour les codeurs G.729.1 à 12 kbits/s et AMR à 12.2 kbits/s dans la section 4.3.5 sont supérieures à 4 ce qui sous-entend des dégradations audibles mais non gênantes. Les résultats que nous obtenons pour ce test multi-locuteur laissent à penser que les différences perçues sont plus importantes.

En wideband, toujours pour la qualité audio perçue, nous n'avons plus l'équivalence entre les codeurs G.722, AMR-WB à 23.85 kbits/s et le G.729.1 à 32 kbits/s que nous avions dans la section 4.3.6 en présence d'un seul locuteur. Les résultats du G.722 sont en effet nettement meilleurs que ceux des deux autres pour le transport d'un contenu binaural multi-locuteur. On peut ici en déduire que le G.722 s'est bien mieux adapté pour le transport d'un contenu binaural multi-locuteur que les codeurs CELP.

Il résulte que les codeurs G.711 ou G.722, par leur nature, sont les plus adaptés à transporter des contenus binauraux multi-locuteur en termes de qualité audio (proche d'excellent) et de qualité de spatialisation (excellent). La qualité audio et la qualité de spatialisation des codeurs CELP sont jugées nettement en dessous de celles de ces codeurs (au mieux passable pour la qualité audio et bonne pour la qualité de spatialisation). A la vue des résultats obtenus ici et de ceux obtenus dans la section 4.3, la qualité audio ressentie suite à l'encodage dual-mono d'un contenu binaural semble très impactée par la présence d'un second locuteur.

Il faut de plus tenir compte que l'encodage-décodage du premier étage de la configuration centralisée n'a pas été appliqué. Or, à la vue des tests précédents (sections 4.4.5 et 4.4.6), les codeurs G.711 et G.722 sont jugés équivalents en configuration centralisée et en configuration distribuée avec la spatialisation ce qui veut dire qu'ils sont performants lors du transcodage ce qui n'est pas le cas des autres codeurs. Il est donc fort probable que la qualité audio et la qualité de spatialisation de ces deux codeurs subissent moins de dégradations que les autres codeurs avec l'ajout du premier étage.

4.5.8 Test complémentaire

Nous avons réalisé un nouveau test MUSHRA (même interface, même matériel et même échelle) pour évaluer la qualité audio ressentie sur le même enchainement de phrases. Cependant, dans ce test, on somme uniquement les deux contenus monophoniques sans les spatialiser, avant de coder le mixage résultant par un codeur monophonique.

Cette configuration est équivalente à un traitement d'un contenu sur deux canaux par un codage dual-mono puisque les sujets écoutent au casque le même contenu sur chaque canal.

Nous n'avons donc plus que deux extraits : un avec une bande de qualité narrowband et l'autre avec une bande de qualité wideband.

4.5.8.1 Conditions du test et procédure expérimentale

Les schémas de traitement sont semblables à ceux des Figure 4.33 et Figure 4.34 mise à part qu'il n'a plus les blocs de spatialisation et que les traitements après le mixage sont sur un seul canal. Il est à noter que les écoutes n'ont pas été effectuées dans une salle acoustiquement isolée.

Voici les différentes conditions des deux extraits narrowband et wideband :

Numéros des conditions	Codeur
1	Ori
2	G.711
3	G.729.1 8
4	G.729.1 12
5	AMR 4.75
6	AMR 12.2

Tableau 4.43 Liste des conditions de transport de contenu monaural narrowband multi-locuteur

Numéros des conditions	Codeur
1	Ori
2	G.722
3	G.729.1 16
4	G.729.1 32
5	AMR-WB 12.65
6	AMR-WB 23.85

Tableau 4.44 Liste des conditions de transport de contenu monaural wideband multi-locuteur

19 sujets ont passé notre test dans lequel il leur était demandé de classer les codeurs narrowband et les codeurs wideband suivant la qualité audio. Tout sujet mettant moins de 90 à la référence haute est exclu des résultats. Nous en avons retenu 16 pour le test wideband et 15 pour le test narrowband.

Les sujets avaient 2 classements à effectuer, présentés de manière aléatoire (un pour chaque bande de qualité).

4.5.8.2 Résultats obtenus

Les notes obtenues pour chaque codeur narrowband sont données sur la Figure 4.44.

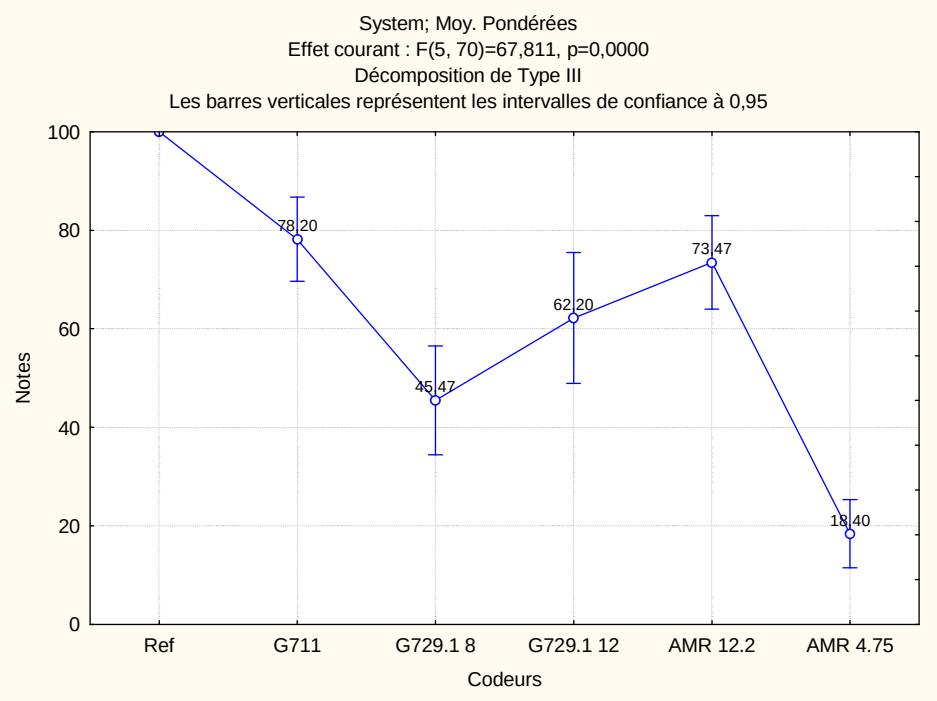

Figure 4.44 Résultats des codeurs narrowband pour le transport multi-locuteur monophonique

Les résultats montrent que la référence est la condition la mieux notée et que l'impact du débit est évident. On retrouve le même ordre que dans la section 4.5.5 (session de qualité audio

ou de qualité de spatialisation) néanmoins les différences ne sont pas aussi significatives comme le montre le classement Tableau 4.45 obtenu suite à un test HSD de Tukey.

N°	Codeurs
1	Ori
2	G.711 ⇔ AMR 12.2
3	AMR 12.2 ⇔ G.729.1 12
4	G.729.1 8
5	AMR 4.75

Tableau 4.45 Classement des codeurs narrowband tous angles confondus pour la session de qualité de spatialisation

Les notes obtenues pour chaque codeur wideband sont données sur la Figure 4.45.

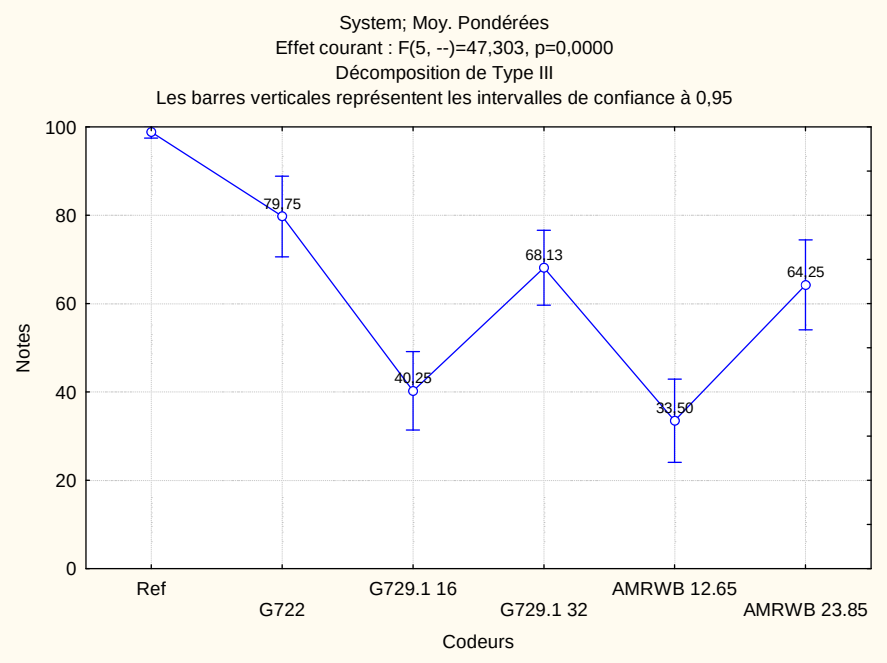

Figure 4.45 Résultats des codeurs wideband pour le transport multi-locuteur monophonique

Les résultats montrent que la référence est la condition la mieux notée et que l'impact du débit est évident. On retrouve le même ordre que dans la section 4.5.6 (session de qualité audio ou de qualité de spatialisation), néanmoins les différences ne sont pas aussi significatives comme le montre le classement Tableau 4.45 obtenu suite à un test HSD de Tukey. Nous obtenons ici une équivalence entre le G.722 et le G.729.1 à 32 kbits/s que nous n'avions pas dans la section 4.5.6. Par contre, à la vue de la Figure 4.45, il semblerait, en augmentant le nombre d'échantillons de sujets ou d'échantillons de test, qu'il soit possible que ces deux codeurs deviennent significativement différents.

N°	Codeurs
1	Ori
2	G.722 ⇔ G.729.1 32
3	729.1 32 ⇔ AMR-WB 23.85
4	G.729.1 16 ⇔ AMR-WB 12.65

Tableau 4.46 Classement des codeurs narrowband tous angles confondus pour la session de qualité de spatialisation

4.5.8.3 Discussion

Globalement, on peut remarquer que les codeurs les mieux notés sont le G.711 et le G.722 même si cela n'est pas toujours de manière significative par rapport aux autres codeurs.

Ces résultats s'expliquent par le fait que les codeurs (G.729.1, AMR-WB et AMR) sont des codeurs CELP. Par leur fonctionnement, ces derniers cherchent à extraire le pitch et les formants de la parole d'un locuteur. Dans le cas de la parole simultanée, il y a une superposition de deux types de pitch et de deux types de formants et ces codeurs ne sont pas adaptés à cette configuration. Cela peut donc aboutir à une dégradation de la qualité perçue par rapport à un codeur G.711 (PCM) ou G.722 (ADPCM) qui ne fonctionnent pas selon ce modèle.

Nous rappelons que la configuration testée est équivalente à la transmission d'un contenu multi-locuteur non spatialisé par un codage dual-mono. Les classements obtenus dans ce test sont relativement semblables à ceux déjà obtenus en qualité audio dans les sections 4.5.5 et 4.5.6 pour le transport par un codage dual-mono d'un contenu spatialisé multi-locuteur. Ils ne nous permettent pas de montrer clairement que la spatialisation a pu avoir un impact sur la qualité perçue et donc sur le classement des codeurs suite à cet encodage dual-mono multi-locuteur (hypothèse du test précédent issue des meilleurs résultats de la version "locuteurs spatialisés à 0°" de l'AMR à 4.75 par rapport à celles où les locuteurs sont distincts). Des comparaisons par paire testant la qualité audio ressentie par les sujets entre les versions spatialisées et non spatialisées seraient peut-être plus adaptées pour mieux évaluer cet impact.

4.6 Conclusion et perspectives

Offrir un service de conférence audio spatialisée binaurale en utilisant des codeurs monophoniques standard est possible avec une bonne voire excellente qualité audio et une excellente qualité de spatialisation. L'avantage principal de ces choix réside dans l'interopérabilité avec l'existant et le fait que les encodeurs-décodeurs soient normalisés. Nous avons notamment montré que notre solution de codage dual-mono avec ces codeurs monophoniques était possible avec une bonne qualité audio et de spatialisation.

Les codeurs G.711 et G.722 semblent les plus adaptés à la conférence centralisée narrowband et wideband avec une qualité jugée nettement supérieure aux codeurs CELP. Ces deux codeurs sont en plus de faible complexité, robustes au transcodage et à la perte de trames, au transport de contenu binaural, et au transport de contenus multi-locuteur. Pour sa part, le G.711 permet en plus une bonne interopérabilité en termes de qualité avec le RTC. Leur principal désavantage est une bande passante importante entre le pont et les terminaux (128 kbits/s) comparée à celles des codeurs CELP. Ces derniers sont donc à utiliser uniquement en cas de fortes contraintes de débit.

En présence de perte de trames, la reconstruction se fait de manière indépendante sur chaque canal. Une recherche intéressante pourrait être de travailler à une reconstruction cohérente sur les deux canaux en même temps suite à des pertes de trames sur un contenu codé en dual-mono. Nous avons de même vu que le principal problème pour la conférence centralisée est lorsque le contenu est multi-locuteur. Nous supposons aussi que la spatialisation peut dans certains cas accentuer la perception de dégradation de la qualité audio suite à un encodage dual-mono.

Les codeurs AMR-WB à 23.85 kbits/s, G.729.1 à 32 kbits/s et G.722 en wideband et les codeurs G.711, AMR à 12.2 kbits/s, et G.729.1 à 12 kbits en narrowband sont les plus adaptés à la conférence distribuée avec une bonne qualité audio, une bonne résistance à la perte de trames et évidemment une excellente qualité de spatialisation. Ils présentent des atouts différents en termes de complexité et de débit qui permettent de s'adapter aux contraintes du réseau ou du matériel.

A plus long terme, un codeur stéréo normalisé (encodeur et décodeur) adapté à transporter des contenus multi-locuteur et utilisant des techniques telles que le parametric stéréo est une piste intéressante à creuser en tenant compte de locuteurs simultanés et de la PdT.

Nous avons aussi montré l'impact non négligeable du type d'écoute (monaurale ou diotique), et que notre perception de la qualité est symétrique (gauche/droite).

Une publication est à venir concernant l'ensemble des résultats de ces tests.

Conclusion

Contributions de la thèse

Fondés sur les orientations définies en introduction à ce document, ces travaux de thèse ont cherché d'une part à définir les architectures adaptées à la conférence audio spatialisée et à en vérifier la validation par une importante série de tests, et d'autre part à établir les spécifications nécessaires au contrôle et la gestion du son spatialisé.

Notre première contribution a consisté en la définition d'architectures adéquates à la conférence audio spatialisée en se basant sur les architectures connues que sont les architectures centralisées et distribuées. Nous avons cherché à tenir compte des différents traitements jouant sur la qualité que sont la détection vocale, le contrôle automatique de gain, la commutation de flux, la sélection de flux, la réduction de bruit ou de parole, etc., cela afin de proposer des solutions complètes.

Nous avons notamment montré qu'il semblait difficile par exemple d'utiliser conjointement sur un pont de conférence la spatialisation et la commutation de flux. Par contre la solution utilisant un pont mixeur ne présentait pas d'inconvénient pour inclure la spatialisation tout en conservant l'ensemble des améliorations de qualité. Par cette configuration, nous garantissons en plus l'interopérabilité avec les réseaux voix existants et la possibilité d'envoyer à tout terminal une spatialisation adaptée.

Les solutions distribuées sont tout autant réalisables dans la théorie mais pâtissent actuellement des limites des terminaux. Pour une interopérabilité, il est en plus nécessaire d'avoir une entité de mixage pour créer un contenu monophonique (pour l'exemple du RTC, cette fonction serait à intégrer dans une passerelle IP/RTC).

Par la suite, nous avons souligné les avantages et les inconvénients de l'utilisation de pont de conférence de type mixeur et de type répliquant. Pour obtenir le meilleur parti de ces architectures, nous avons proposé une solution de pont mixte. Cette solution fonctionne tantôt en mode répliquant, tantôt en mode mixeur suivant les capacités des terminaux. Cette architecture innovante a fait l'objet d'un brevet.

Afin de réduire la bande passante d'un pont répliquant vers un terminal, nous avons proposé une méthode de sélection de flux basée sur le masquage auditif. Cette méthode fonctionnant par sous-bande évalue quelles trames sont masquées afin d'éviter de les transmettre vers les terminaux. Cette solution a obtenu de bons résultats aussi bien théoriques que pratiques. Cette méthode a fait l'objet d'une publication pour l'AES 30.

Notre seconde contribution a été la définition des extensions nécessaires à la gestion et au transport du son spatialisé. Nous avons tout d'abord défini les spécifications permettant de

commander un positionnement de locuteurs dans une conférence audio spatialisée. Nous avons cherché à présenter toutes les solutions possibles pour les gestions automatique ou manuelle.

Dans le cas de la spatialisation sur un pont de conférence, nous avons ainsi souligné le fait que cette gestion ne pouvait se faire par l'intermédiaire du protocole SIP, car ce n'est pas le rôle de ce dernier de transporter dans ses messages des informations sur les contenus ou des commandes de spatialisation. Nous avons proposé une solution basée sur ce qui se fait dans les conférences audio standard : une solution de web-pilotage certes propriétaire à chaque fournisseur de services mais en cohérence avec la gestion des protocoles de Voix sur IP.

Pour la conférence avec un pont mixeur, nous avons établi les paramètres du protocole de signalisation SIP nécessaires au transport de flux asymétriques tout en garantissant une interopérabilité avec les terminaux. La nécessité de transporter ces flux asymétriques est due à notre hypothèse de départ concernant l'équipement des terminaux : prise de son monophonique et restitution sur casque stéréophonique ou sur deux haut-parleurs. Cette spécification fera l'objet d'une action en normalisation courant 2008.

Nous avons enfin souligné l'importance de bien dissocier le contenu (type de spatialisation) du contenant (codeur, nombre de canaux). Il est en effet important de pouvoir identifier le contenu pour pouvoir par la suite le restituer de manière adéquate ou pour pouvoir le modifier à la guise de l'utilisateur.

Notre troisième contribution concerne les tests des architectures centralisée et distribuée retenues dans la première contribution. Ceux-ci ont nécessité une séparation des différents blocs de traitements pour pouvoir évaluer chacun d'entre eux et déterminer lesquels étaient les plus impactants. Ces tests nous ont amené à définir des nouveaux protocoles adaptés à ces architectures. En outre, nous avons retenu et validé la solution du codage dual-mono qui nous semblait à ce jour la plus adaptée pour transporter le contenu binaural par rapport aux codeurs stéréophoniques et les méthodes de compression de contenu multicanal.

Nous avons montré dans un premier temps que les codeurs n'étaient pas perçus de la même façon suivant que l'on écoute en écoute monaurale ou en écoute diotique (Publication AES 123). Nous avons de même montré que l'on percevait la qualité de manière symétrique (gauche/droite).

Dans un second temps, nous avons montré que le codage dual-mono était adéquat au transport de contenu binaural. Il ressort que les codeurs G.711 (PCM) et G.722 (ADPCM) sont les plus adaptés à la conférence audio centralisée car ils sont robustes au transcodage (l'impact du second étage étant on l'a vu assez fort), au transport de contenu binaural en dual-mono, à la perte de trames et au transport de contenus multi-locuteur. Ils ont une qualité jugée nettement supérieure à celle des codeurs CELP (AMR, AMR-WB, G.729.1). Ces derniers sont donc à utiliser de préférence uniquement lorsque les contraintes de débits sont fortes. Ces résultats ont souligné le lien entre le débit et la qualité : plus le débit est fort, meilleure est la qualité.

A plus long terme, un codeur stéréo normalisé (encodeur et décodeur) adapté à transporter des contenus multi-locuteur et utilisant des techniques telles que le parametric stéréo est une piste intéressante à creuser en tenant compte de locuteurs simultanés et de la correction de perte de trames.

Concernant la conférence audio distribuée wideband, les codeurs G.729.1 à 32 kbits/s, AMR-WB à 23.85 kbits/s et G.722 à 64 kbits/s semblent les plus adaptés quelle que soit la perte de trames. Ils ont une qualité jugée équivalente. En conférence audio distribuée narrowband, les codeurs G.711, AMR à 12.2 kbits/s et G.729.1 à 12 kbits/s obtiennent les meilleures notes de qualité, quelle que soit la perte de trames. Au final, dans tous les cas distribués, le choix du codeur dépendra des contraintes de l'application suivant un compromis complexité/débit.

L'ensemble de ces tests fera l'objet d'une publication courant 2008.

Perspectives de recherche

Nous avons restreint notre étude à la transmission d'un contenu monaural vers un pont ou vers des terminaux mais rien ne nous empêche de transmettre un contenu déjà spatialisé capturé par exemple avec une prise de son binaurale. En généralisant, la prise de son peut être multi-microphone et l'entité effectuant le mixage doit connaître la localisation de ces microphones pour pouvoir éventuellement transformer le contenu. Cette information sur les microphones doit être échangée entre le terminal et l'entité de mixage, avec un formalisme à établir. Enfin, les terminaux participant à une conférence pouvant être hétérogènes en termes de système de prise de son, la question de mixage audio de contenus différents (monophoniques, binauraux, multi-canal, etc.) se pose.

Nous avons testé pour la conférence audio centralisée avec pont mixeur un transcodage avec le même codeur sur les premier et second étages. Il conviendrait d'étudier les effets avec des codeurs différents.

Notre algorithme de positionnement automatique nécessite une évaluation. Son intérêt parait évident par rapport à un positionnement aléatoire, mais il nous faut à présent le valider par une méthode subjective. Cela sera l'un de nos prochains objectifs.

La spatialisation pouvant être déroutante pour les utilisateurs, de nombreuses études sont encore à mener concernant l'ergonomie. Nous avons proposé différentes méthodes pour le positionnement des locuteurs. Il convient à présent d'avoir un retour de la part des utilisateurs afin de valider les choix. Par exemple, il convient d'établir quelles règles de placements ont la préférence des utilisateurs selon le contexte de communication.

ARTICLES & BREVETS

Articles acceptés

- Nagle, A., Tsingos, N., Lemaitre, G. and Sollaud, A., On the fly auditory masking for scalable VoIP bridges, article présenté à la conférence AES 30 à Saariselkä (Finlande) en mars 2007.
- Nagle, A., Quinquis, C., Sollaud, A., Battistello, A. and Slock, D., Quality impact of diotic versus monaural hearing on processed speech, article présenté à la convention AES 123 à New York (USA) en octobre 2007.

Articles à soumettre

- Validation de notre brevet sur le positionnement automatique de locuteurs en conférence audio 3D.
- Présentation de la série de tests effectuée dans le chapitre 4.

Brevets déposés

- Procédé de détermination d'un mode d'encodage spatial de données audio. Déposé le 27/12/2005, n° FR 05 54106.
- Pont de conférence mixte. Déposé le 22/12/2006, n° FR 06 55908
- Positionnement automatique de locuteurs en conférence audio 3D. Déposé le 29/06/2007, n° FR 07 04712

Brevets en cours de dépôt

- Optimisation dynamique d'un système de conférence audio spatialisée à prises de son multicanal en cours de dépôt.

BIBLIOGRAPHIE

[1] Commutation de circuits et Commutation de paquets, http://www.commentcamarche.net/initiation/commutation-circuits-paquets.php3.

[2] H.323 : Systèmes de communication multimédia en mode paquet, ITU (Juillet 2003).

[3] 3GPP, TS 26.073 AMR speech Codec; C-source code. V.5.1.0, (2003).

[4] 3GPP, TS 26.173 ANSI-C code for the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec. Version 5.8.0, (2003).

[5] I.I. 23003-1:2007, MPEG Surround, (2007).

[6] K.S. Abouchacra, J. Breitenbach, T. Mermagen, T. Letowski, Binaural Helmet: Improving Speech Recognition in Noise with Spatialized Sound, Human Factors (2001).

[7] F. Andreasen, Session Description Protocol (SDP) Simple Capability Declaration, IETF RFC 3407 (2002).

[8] F. Andreasen, B. Foster, Media Gateway Control Protocol (MGCP) Version 1.0, IETF RFC 3435 (Janvier 2003).

[9] R. Appel, J.G. Beerends, On the quality of hearing one's own voice, JAES 50 (2002) 237-246.

[10] F. Audet, C. Jennings, Network Address Translation (NAT) Behavioral Requirements for Unicast UDP, IETF RFC 4787 (Janvier 2007).

[11] J.J. Baldis, Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences, in: Computer Human Interaction 2001 (Seattle, WA, USA, 2001).

[12] F. Baumgarte, C. Faller, Binaural Cue Coding, in: IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol. 11 (2003).

[13] D.R. Begault, E.M. Wenzel, Headphone Localization of Speech, The Human Factors and Ergonomics Society (1993).

[14] J. Blauert, Spatial Hearing (The MIT Press, London, England, 1983).

[15] R.S. Bolia, W.T. Nelson, M.A. Ericson, A Speech Corpus for Multitalker Communication Research, JASA (2000).

[16] M. Botte, G. Canévet, L. Demany, C. Sorin, Psychoacoustique et perception auditive (1989).

[17] J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers, Parametric Coding of Stereo Audio, EURASIP Journal on Applied Signal Processing 9 (2005) 1305-1322.

[18] M. Briand, Etudes d'algorithmes d'extraction des informations de spatialisation sonore : Application aux formats multicanaux, Signal, Image, Parole, Télécoms, , Thèse de Doctorat, Institut National Polytechnique de Grenoble, 2007.

[19] G. Camarillo, G. Eriksson, J. Holler, H. Schulzrinne, Grouping of Media Lines in the Session Description Protocol (SDP), IETF RFC 3388 (2002).

[20] G. Camarillo, A. Johnston, Conference Establishment Using Request-Contained Lists in the Session Initiation Protocol (SIP), IETF Draft draft-ietf-sip-uri-list-conferencing-01 (Work in Progress) (Janvier 2007).

[21] T.G. Champion, Multi-speaker conferencing over narrowband channels, in: Proc. IEEE Military Communications Conf. (Washington, D.C., 1991).

[22] A. Chraa, Encodage et restitution d'une scène sonore au format ambisonique : optimisation du décodage spatial pour l'écoute binaurale, Stage de master II, Université de Rennes I, Septembre 2004.

[23] N. Coté, Qualité perçue de parole transmise par voie téléphonique large-bande, Stage de master II, Université Pierre et Marie Curie, 2005.

[24] DECT-NG07_033, ITU-T G.722 PLC selection phase: additional information, in: NG_DECT #7 (2006).

[25] R. Drullman, A.W. Bronkhorst, Multichannel Speech Intelligibility and Talker Recognition using Monaural, Binaural and Three-Dimensional Auditory Presentation, JASA (1999).

[26] P. Escolano, Approche concrète du téléphone fixe - RTC, in: http://stielec.ac-aix-marseille.fr/cours/escolano/download/rtc.pdf (Ed.), (2006).

[27] R. Eslava, Telecoms Strategies & Trends : Strategic analyses inside the world of telecoms, in: (InfoCom, 2006).

[28] H. Fletcher, Auditory Patterns, Review of Modern Physics 12 (1940) 47-65.
[29] H. Fletcher, W. Munson, Loudness, its definition, measurement and calculation, JASA 5 (1933) 82-108.
[30] E. Gallo, G. Lemaître, N. Tsingos, Prioritizing audio signals for selective processing, in: International Conference on Audio Displays (Limerick, Ireland, 2005).
[31] W. Gardner, 3D audio using loudspeakers (1998).
[32] P. Gasser, Les formats MPEG audio, in: MSH Paris Nord - Plate-forme Arts, Sciences, Technologies (2006).
[33] N. Gleiss, Usability - Concepts and Evaluation (1992).
[34] J.C. Gruber, L. Strawczynski, Subjective Effects of Variable Delay and Speech Clipping in Dynamically Managed Voice Systems, IEEE Transactions on Communications 33 (1985) p801-808.
[35] M. Handley, V. Jacobson, C. Perkins, SDP: Session Description Protocol, IETF RFC 4566 (2006).
[36] S. Haykin, Adaptive Filter Theory (Prentice Hall, 2002).
[37] J. Herre, S. Disch, New concepts in parametric coding of spatial audio : from SAC to SAOC, in: International Conference on Multimedia and Expo (Pekin, Chine, 2007).
[38] O. Hersent, D. Gurle, J. Petit, La Voix sur IP : Codecs, H.323, SIP, MGCP, déploiement et dimensionnement (Dunod, 2004).
[39] Z. Huang, Conferencing forecasts, in: T.a.S.C. (OVUM)- (Ed.), (2006).
[40] ISO/IEC 14496-3, Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio, 2nd edition), (2001).
[41] ITU-R, Rec. BS.1534 : Method for the subjective assessment of intermediate quality level of coding systems, (2001).
[42] ITU-T, ITU-T Software Tool Library 2005 User's Manual (ITU, Geneva, 2005).
[43] ITU-T, Rec. E.800 Termes et définitions relatifs à la qualité de service et à la qualité de fonctionnement du réseau, y compris la sûreté de fonctionnement, (1994).
[44] ITU-T, REC. G.114 Temps de transmission dans un sens, (2003).
[45] ITU-T, Rec. G.701 Vocabulaire relatif à la modulation par impulsions et codage (MIC), au multiplexage et à la transmission numérique, (1993).
[46] ITU-T, REC. G.729.1 G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729, (2006).
[47] ITU-T, Rec. P.800 Methods for Subjective Determination of Transmission Quality, in: (1996).
[48] ITU-T, TD AH-06-42 : Draft Handbook STP-Handbook of subjective testing practical procedures (2006).
[49] W. Jesteadt, C. Wier, Comparison of monaural and binaural discrimination of intensity and frequency, JASA 61 (1977).
[50] D.L. Jones, K.M. Stanney, H. Foaud, An Optimized Spatial Audio System for Virtual Training Simulations : Design and Evaluation, in: International Conference on Auditory Display 2005 (Limerick, Ireland, Juillet 2005).
[51] C. Kayser, C. Petkov, M. Lippert, N. Logothetis, K., Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map, Current Biology, Current Biology 15 (2005).
[52] M.C. Kelly, A.I. Tew, The continuity illusion in virtual auditory space, in: 112th Audio Engineering Society Convention (Munich, Germany, 2002).
[53] M.C. Kelly, Tew, A.I., The continuity illusion revisited: coding of multiple concurrent sound sources., in: 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002) (Leuven, Belgium, 2002).
[54] R. Kilgore, M. Chignell, P. Smith, Spatialized Audioconferencing : What are the Benefits ?, in: Centre for Advanced Studies conference on Collaborative Research (2003).
[55] R. Kilgore, Chignell, M., Listening to Unfamiliar Voices in Spatial Audio : Does Visualization of Spatial Position Enhance Voice Identification ?, in: Human Factors in Telecommunication (2006).
[56] R. Kilgore, Chignell, M., Simple visualizations enhance speaker identification when listening to spatialized voices, in: Human Factors and Ergonomics Society (2005).
[57] R. Le Bouquin-Jeannes, P. Scalart, G. Faucon, C. Beaugeant, Combined Noise and Echo Reduction in Hands-Free Systems: A Survey, IEEE Trans. on Speech and Audio Processing 9 (November 2001) p808-820.
[58] M. Lutzky, G. Schuller, M. Gayer, U. Krämer, S. Wabnik, A guideline to audio codec delay, in: AES 116 (Berlin, 2004).
[59] R. Mahy, B. Campbell, R. Sparks, J. Rosenberg, D. Petrie, A. Johnston, O. Levin, A Call Control and Multi-Party usage framework for the Session Initiation Protocol (SIP), IETF Draft draft-ietf-sipping-cc-framework-07 (Work in Progress) (Mars 2007).

[60] M. Mealling, Dynamic Delegation Discovery System (DDDS) Part Three: The Domain Name System (DNS) Database, IETF RFC 3403 (Octobre 2002).
[61] H. Möller, Fundamentals of binaural technology, Applied Acoustics 36 (1992) pp. 171-218.
[62] S. Möller, Quality of Telephone-Based Spoken Dialogue Systems (New-York, 2005).
[63] E.D. Montag, Forced Choice and miscellaneous consideration, in: http://www.cis.rit.edu/people/faculty/montag/vandplite/pages/chap_4/ch4p5.html (Ed.), Psychophysics (2003).
[64] A. Nagle, C. Quinquis, A. Sollaud, D. Slock, Quality impact of diotic versus monaural hearing on processed speech, in: Convention AES 123 (New-York, 2007).
[65] A. Nagle, N. Tsingos, G. Lemaitre, A. Sollaud, On the fly auditory masking for scalable VoIP bridges, (2006).
[66] D. Nahumi, Conferencing arrangement for compressed information signals, in: (AT&T Corp, USA, 1995).
[67] W.T. Nelson, R.S. Bolia, M.A. Ericson, R.L. McKinley, Spatial Audio Displays for Speech Communications : A Comparison of Free Field and Virtual Acoustic Environments, in: Human Factors and Ergonomics Society (1999).
[68] R. Nicol, Restitution sonore spatialisée sur une zone étendue : Application à la téléprésence, Acoustique, Thèse de Doctorat, Université du Maine, Décembre 1999.
[69] J.E.M. Painter, S. Spanias, Perceptual Coding of Digital Audio, Proceedings of the IEEE 88 (2000).
[70] Project Group 841, ROBUST VOICE ACTIVITY DETECTION and Noise ReductionMechanism USING HIGHER-ORDER STATISTICS, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, 2005.
[71] A. Raake, Speech quality of VoIP : Assessment and Prediction (2006).
[72] R. Rabipour, P. Coverdale, Tandem-free VoX conferencing., in: (Nortel Networks, 1999).
[73] G. Reynolds, S. Stevens, Binaural Summation of Loudness, JASA 32 (1960).
[74] A.B. Roach, Session Initiation Protocol (SIP)-Specific Event Notification, IETF RFC 3265 (Juin 2002).
[75] J. Rosenberg, A Framework For Conferencing with the Session Initiation Protocol (SIP), IETF RFC 4353 (Février 2006).
[76] J. Rosenberg, Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols, IETF Draft draft-ietf-mmusic-ice-19 (Work in Progress) (Octobre 2007).
[77] J. Rosenberg, M. Handley, H. Schulzrinne, E. Schooler, SIP : Session Initiation Protocol, IETF RFC 2543 (Juin 1999).
[78] J. Rosenberg, C. Huitema, R. Mahy, D. Wing, P. Matthews, Session Traversal Utilities for NAT (STUN), IETF draft draft-ietf-behave-rfc3489bis-11 (Work in Progress) (Octobre 2007).
[79] J. Rosenberg, R. Mahy, C. Huitema, Traversal Using Relays around NAT (TURN): Relay Extensions to Session Traversal Utilities for NAT (STUN), IETF Draft draft-ietf-behave-turn-04 (Work in Progress) (Juillet 2007).
[80] J. Rosenberg, H. Schulzrinne, Reliability of Provisional Responses in SDP, IETF RFC 3262 (Juin 2002).
[81] J. Rosenberg, H. Schulzrinne, Session Initiation Protocol (SIP): Locating SIP Servers, IETF RFC 3263 (Juin 2002).
[82] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP : Session Initiation Protocol, IETF RFC 3261 (Juin 2002).
[83] J. Rosenberg, H. Schulzrinne, O. Levin, A Session Initiation Protocol (SIP) Event Package for Conference State, IETF RFC 4575 (Août 2006).
[84] J. Rosenberg, Schulzrinne, H., An Offer/Answer Model with the Session Description Protocol (SDP), IETF RFC 3264 (Juin 2002).
[85] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, RTP : A Transport Protocol for Real-Time Applications, IETF RFC 3550 (Juillet 2003).
[86] F. Simard, P.K. Edholm, N.K. Burns, Apparatus and method for packetbased media communications., in: (Nortel Networks, Canada, 2001).
[87] K. Siyan, TCP/IP (Janvier 2003) 706.
[88] P.J. Smith, Voice Conferencing over IP Networks, Department of Electrical & Computer Engineering, Master of Engineering - McGill University - Canada, 2002.
[89] P. Srisuresh, M. Holdrege, IP Network Address Translator (NAT) Terminology and Considerations, IETF RFC 2663 (Août 1999).
[90] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, V. Paxson, Stream Control Transmission Protocol, IETF RFC 2960 (Octobre 2000).

[91] J.D. Tardelli, P.D. Gatewood, E.W. Kreamer, P.A. La Follette, The benefits of multi-speaker conferencing and the design of conference bridge control algorithms, in: Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Vol. 2 (Minneapolis, USA, 1993) 435-438.

[92] N. Tsingos, Scalable Perceptual Mixing and Filtering of Audio Signals using an Augmented Spectral Representation, in: 8th Int. Conference on Digital Audio Effects (DAFx'05) (Madrid, Spain, 2005).

[93] N. Tsingos, E. Gallo, G. Drettakis, Perceptual Audio Rendering of Complex Virtual Environment, ACM Transactions on Graphics 23 (2004).

[94] N. Tsingos, O. Warusfel, J.C. Lombardo, J. Soula, B. Katz, A. Raake, H. Goidell, J. Hognon, A. Sollaud, A. Nagle, M. Emerit, OPERA : Optimisation PErceptive du Rendu Audio (2006).

[95] P. Vary, R. Martin, Digital Speech Transmission (John Wiley and Sons, UK - Chichester, 2006).

[96] Z.H. Zhou, The Role of 3D Sound in Human Reaction and Performance in Augmented Reality Gaming Environment, (2002).

[97] S. Znaty, Le réseau Sémaphore Numéro 7 : Principes, Architecture et Protocoles, in: http://efort.com/r_tutoriels/SS7_EFORT.pdf (Ed.).

[98] E. Zwicker, H. Fastl, Psychoacoustics: Facts and Model (1999).