音声チャット　マルチチャネル符号化　バイノーラル信号処理

Technology Reports

モバイル多地点音声チャットのための サラウンド音声伝送技術

円滑で快適な遠隔音声コミュニケーションの実現を目指し，多地点音声チャットにおいて，参加者の各音声を任意方向から立体的に分離再生が可能なサラウンド音声伝送技術を，モバイル向けに開発した．本技術により，1人ひとりの好みにあった受聴環境で，ユーザにとって直感的に聞き分けやすく聞き疲れしにくい音声チャットが可能になる．

先進技術研究所　飯塚 真也　菊入 圭　仲 信彦

1. まえがき

近年，コンテンツシェア[*1]やオンラインゲームなど，複数人が同時に参加可能なコミュニケーションサービスが注目されている．このようなサービスにおいて，リアルタイムに感情や感動を共有できる多地点音声チャットは，充実したコミュニケーションの実現に重要な役割を果たす機能の1つである．

一方，通信網の高速大容量化を背景に，より広い帯域の音声信号を伝送し，自然で臨場感のある音声コミュニケーションを目指した研究開発や実用化が行われている．ドコモにおいても，LTE（Long Term Evolution）[*2]や第4世代移動通信方式など高速無線パケットアクセス回線上でのVoIPサービスの利用を想定し，ビットレート48k～64kbit/sで超広帯域音声（上限周波数10kHz以上）が伝送できる高音質音声符号化を開発した[1]．

このようなモバイル環境での音声コミュニケーションにおけるユーザ体験品質（QoE：Quality of Experience）向上の取組みの一環として，多地点音声チャットのような多人数通話でも快適な受聴環境を提供することを目的に，高音質音声符号化の機能拡張としてサラウンド音声伝送技術を開発した．

1対1の通話と異なり，複数の話者が混在する環境では，「話者の識別」や「会話の聞分け」が困難になるという課題が新たに生じる．これらの課題に対して，バイノーラル信号処理[*3]技術によって各話者の音声信号に方向や奥行きなどの空間情報を付加し，立体的に再生することが有効であることが知られている[2]．

従来立体再生は，臨場感や場の共有といった，現実もしくは仮想的な音響空間の再現に使われることが多かった[3]．今回開発したサラウンド音声伝送技術では，さらにそれぞれのユーザが聞きやすさに応じて参加者の音声を自由に配置できることを目指した．各参加者の音声を受聴者の好みに応じて立体再生できる音声チャットシステムには，大きく3つの方式がある（**表1**）．クライアント処理方式は，各参加者の音声データを直接送受信し，クライアント上でそれぞれ立体再生するため，クライアントのみで実現可能であるが，参加者数に応じて情報伝送量やクライアントの演算負荷が増大する．サーバ処理方式は，各参加者の音声データをネットワーク上のサーバで多重化し立体再生信号を生成するため，情報伝送量やクライアントの演算負荷の低減が可能であるが，クライアントからサーバへ空間情報を管理するための制御情報を送信するバックチャネルが必要になる．ハイブリッド処理方式は，サーバでの圧縮多重およびクライアントでの立体

＊1　**コンテンツシェア**：ネットワークを介し，動画や画像などの情報を共有するサービス．
＊2　**LTE**：3GPPで検討されている第3世代移動通信方式の拡張規格．ドコモで提唱しているSuper3Gと同義．
＊3　**バイノーラル信号処理**：両耳間での音の聞こえの違いを人工的に付加することで，モノラル音を立体的に再生する信号処理．

再生を併用しており，サーバでの圧縮処理によって他の方式に比べ音質が劣化する可能性はあるが，情報伝送量の圧縮，演算負荷の分散およびクライアントに閉じた処理が可能である．

開発したサラウンド音声伝送技術は，無線回線容量やクライアントでの処理性能に制約のあるモバイル環境を考慮し，ハイブリッド方式に基づいている．また本技術の特長は，聴覚特性の利用によって，ハイブリッド方式の課題とされる音質劣化を抑えつつ，複数の高音質音声符号化データをサーバ上で48k～96kbit/sに圧縮多重が可能なマルチチャネル符号化[*4]処理と，復号および立体再生の効率的統合によってクライアント上での低演算量での動作が可能な立体復号処理である．

本技術が実用化されれば，モバイル環境での多地点音声チャットにおいて，ユーザにとって直感的に聞き分けやすい円滑な音声コミュニケーションを提供することが可能になる．

本稿では，開発したサラウンド音声伝送技術の概要およびその品質評価結果について解説するとともに，本技術を採用したモバイルVoIP多地点音声チャットプロトタイプについて述べる．

表1　立体再生音声チャットシステムの方式

	クライアント処理方式	サーバ処理方式	ハイブリッド処理方式
サーバ	不要	必要	必要
バックチャネル	不要	必要	不要
下り回線の伝送情報（情報伝送量）	全参加者の音声データ（参加者数により増大）	立体再生音声データ（一定）	多重化データ（一定）
端末での処理（演算負荷）	各音声データに対する復号および立体再生（参加者数により増大）	立体再生音声データの復号（一定）	多重化データに対する復号および立体再生（一定）

2. サラウンド音声伝送技術

2.1 サラウンド音声伝送技術の構成

サラウンド音声伝送技術は，音声符号化処理，マルチチャネル符号化処理および立体復号処理から構成されている．このうち，音声符号化処理および立体復号処理はクライアント上で行い，マルチチャネル符号化はサーバ上で行う（**図1**）．

音声符号化処理は，ドコモで開発した高音質音声符号化を用いる．本符号化は，入力された時間領域の音声信号を，修正離散コサイン変換（MDCT : Modified Discrete Cosine

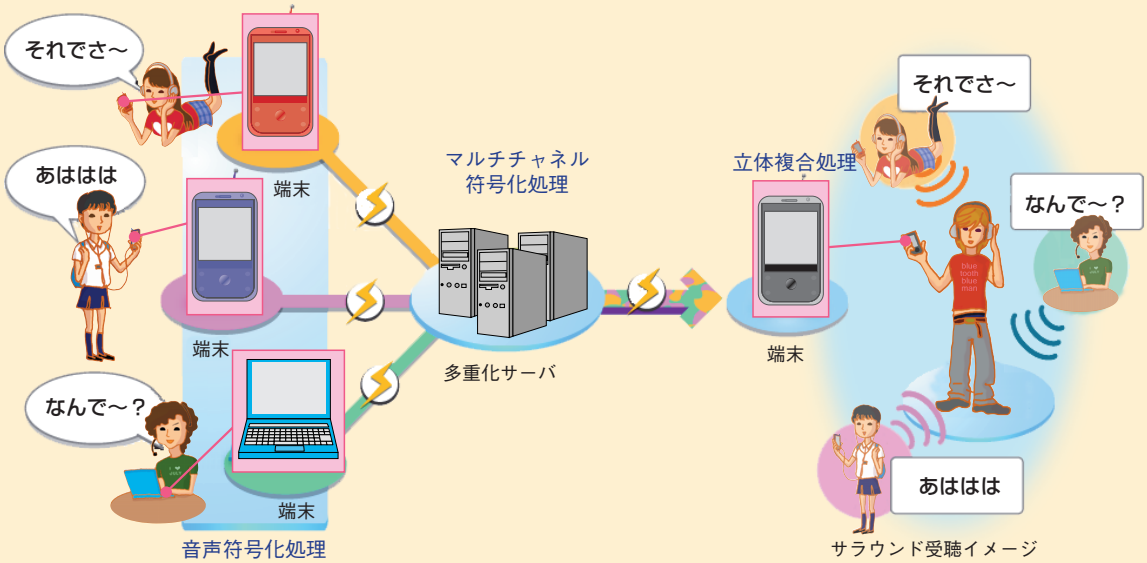

図1　サラウンド音声伝送技術の構成

＊4　**マルチチャネル符号化**：複数系統の入力信号を1系統に多重化し情報圧縮する信号処理．

Transform）[*5]によって周波数領域の係数へ変換し，聴覚的な重要度に応じた精度で各係数を量子化する．超広帯域音声を数十msの低遅延かつ従来の音声符号化と同程度の演算量で符号化することが特徴である．

マルチチャネル符号化処理は，クライアントから受信した複数の高音質音声符号化データを復号し，周波数領域の係数の比較によって聴覚的に重要な成分のみを抽出し圧縮多重することで，1つの符号化データに圧縮する（**図2**）．

立体復号処理は，マルチチャネル符号化処理によって圧縮多重された符号化データを，各参加者の音声に分離し周波数領域の係数に復号するとともに，立体化を行う．

ここで，人間が音源の位置を認知する仕組みを**図3**に示す．ある音源から発生した音波は空気を伝わって両耳に達する．このとき，音源から両耳への距離差によって生じる両耳間強度差（IID：Inter-aural Intensity Difference）および両耳間時間差（ITD：Inter-aural Time Difference）を手がかりとして，音波の到来方向を認知する．つまり，モノラル音声信号に対して，これらIIDおよびITDを信号処理により擬似的に付加した信号をヘッドホンで左右の耳にそれぞれ提示することで，受聴者は音声信号を立体的に感じる．

従来の処理では，符号化データを時間領域の音声信号に復号した後，立体化を施す必要があったが，本技術における立体復号処理は，符号化データの復号処理の過程で周波数領域の係数を直接操作することで立体化する方式を開発した（**図4**）．この音声復号処理と立体化処理との一体化によって，立体再生にかかる演算量を従来比約30～50％に抑えることが可能となった．

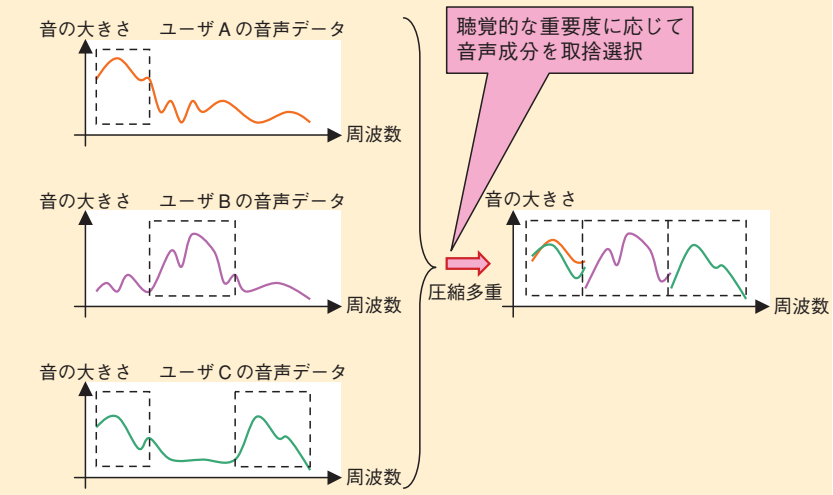

図2　マルチチャネル符号化における多重化処理

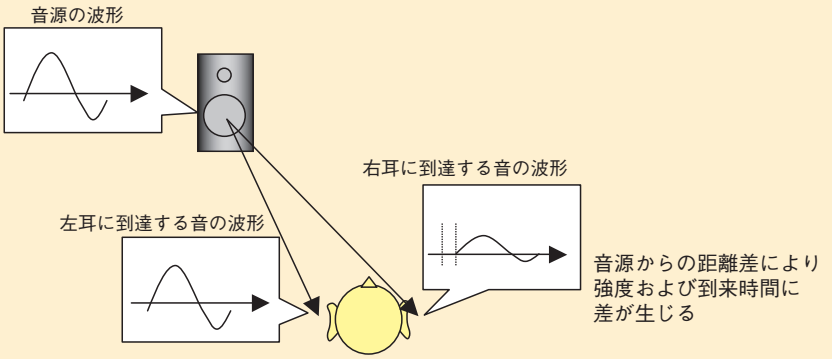

図3　方向認知の仕組み

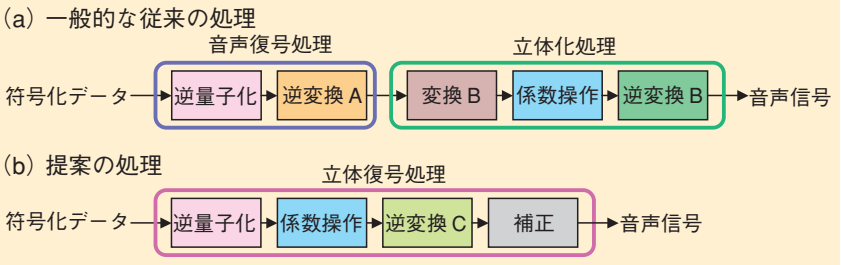

図4　立体復号処理の構成

＊5　**修正離散コサイン変換（MDCT）**：時系列信号を周波数成分に変換する手法の1つ．前後の隣接する変換ブロックと重ね合わせる変換で，情報の無駄なくブロック境界の歪みを防ぐことができることから，音響符号化において広く利用されている．

2.2 音声品質の検証

サラウンド音声伝送技術による立体再生音声の品質を検証するため，主観評価試験を行った．試験条件を**表2**に示す．試験方法には，評価対象音（原音を含む）を0～100点で採点するMUSHRA（Multi Stimulus test with Hidden Reference and Anchor）法[4]を採用した．

主観評価試験結果を**図5**に示す．図中の誤差範囲はそれぞれの評価値の95％信頼区間[*6]を示している．会話音声Aは瞬間的に同時発言を含む音声，会話音声Bは2名以上が常時同時発言状態の音声である．会話音声Aでは64kbit/s，会話音声Bでは96kbit/sで，マルチチャネル符号化による圧縮多重を行わない高音質音声符号化（チャネル当り64kbit/s）と同等の主観品質といえる．すなわち，サラウンド音声伝送技術では，それぞれの会話音声について，マルチチャネル符号化によって下り情報伝送量を非圧縮多重時の20～25％に削減することが可能である．

3. プロトタイプ

本技術を，SIP（Session Initiation Protocol）[*7]を用いたVoIPベースの多地点音声チャットシステムとして実装した．本実装では，サーバ機能はWindows®[*8]対応，クライアント機能はWindows Mobile®[*9]対応のアプリケーションソフトウェアとしてそれぞれ動作する．クライアント用ソフトウェアはFOMA PROシリーズHT-01Aで動作可能であることを確認した（**写真1**）．各クライアントは，サーバ内に設定された会議室への発呼により，音声チャットに参加する．クライアント画面には，参加者のリスト表示および発話状態の表示が可能であり，参加者名を選択後，左右ボタンで

表2　主観評価試験条件

試験方法	MUSHRA法
被験者数	10名
音源	会話音声A（参加者5名，同時発言頻度：低） 会話音声B（参加者6名，同時発言頻度：高）
参照音声 （サンプリング周波数）	原音を個別に立体化した立体再生音 （22.05kHz）
符号化音声 （ビットレート／ サンプリング周波数）	サラウンド音声伝送による立体再生音 （48k, 64k, 96kbit/s / 22.05kHz）
非圧縮多重符号化音声	高音質音声符号化音声（64kbit/s / 22.05kHz）を個別に 立体化した立体再生音
帯域制限音声	7kHz帯域，3.5kHz帯域
受聴方法	ヘッドホン両耳受聴

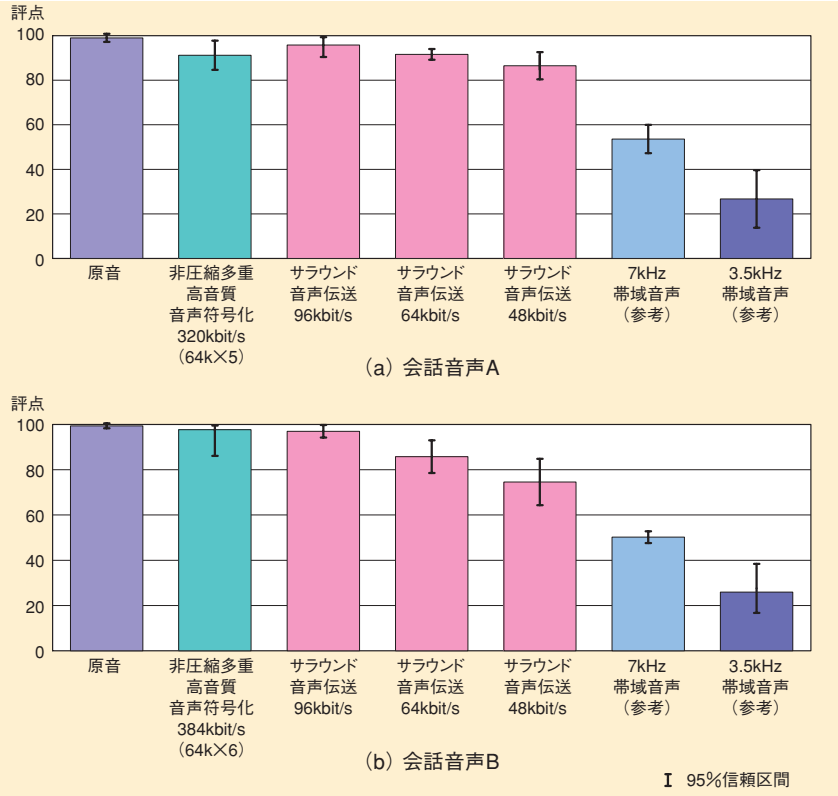

図5　主観評価試験結果

＊6 **95％信頼区間**：サンプルが特定の分布に従って存在すると仮定した場合，サンプルの95％が含まれる区間．

＊7 **SIP**：VoIPを用いたIP電話などで利用される，IETF（Internet Engineering Task Force）で策定された通信制御プロトコルの1つ．

＊8 **Windows®**：米国Microsoft Corporationの米国およびその他の国における登録商標または商標．

＊9 **Windows Mobile®**：米国Microsoft Corporationの米国およびその他の国における登録商標または商標．

写真1　プロトタイプソフトウェア表示例

対象参加者の音声の方向を，上下ボタンで対象参加者の発話音声の音量を操作できる．

4. あとがき

本稿では，円滑で快適な多人数音声コミュニケーションの提供を目指し，モバイル環境での多地点音声チャットにおいて，参加者の各音声を受聴者の好みに応じた任意方向から立体的に再生が可能なサラウンド音声伝送技術について解説した．主観評価の結果から，マルチチャネル符号化によって，主観品質を保ちながら情報伝送量を20～25％に削減可能であることが示された．また，本技術をVoIPベースの多地点音声チャットシステムとして実装したプロトタイプについて解説した．

参加者の音声の方向・音量を操作可能なサラウンド音声伝送技術は，音声チャットにおける受聴の快適性を向上させるほか，空間の共有感や臨場感の向上を活かしたアプリケーションへの応用も期待される．今後は，サラウンド効果の個人差への対応など本技術の立体化性能の向上を目指し検討を進めていく．

文　献

[1] 菊入，ほか：“高音質音声符号化，”本誌，Vol.15, No.2, pp.36-39, Jul. 2007.
[2] R.Drullman and A. W. Bronkhorst：“Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation,” J. Acoust. Soc. Am., 107, pp.2224-2235, 2000.
[3] 安田，ほか：“リアリティ音声音響通信技術，”本誌，Vol.11, No.1, pp.55-62, Apr. 2003.
[4] ITU-R Recommendation BS.1534-1：“Method for the subjective assessment of intermediate quality level of coding systems,” 2003.