Vol.2015-MUS-107 No.8
2015/5/24

音声音響符号化技術の最近の話題

守谷　健弘

音声音響符号化技術に関する進展の経緯を簡単に振り返り，あわせて最近の標準化，実用化の話題を紹介する．このなかに昨年 IEEE マイルストンに認定された LSP，昨年日本の超高精細度テレビの高音質サービス用に選定された MPEG-4 ALS，昨年携帯電話用に 3GPP で標準化された EVS を含む．

Recent Topics of Speech and Audio Coding Technologies

TAKEHIRO MORIYA

History and recent topics of speech and audio coding technologies are briefly explained. Topics include LSP, which has been certified as IEEE milestone, MPEG-4 ALS, which has been defined for the Japanese Ultra High Definition broadcasting, and EVS, which has been standardized in 3GPP for future mobile phone.

1. はじめに

音声音響符号化技術は図 1 に示すように，さまざまな基盤技術に支えられつつ，すでに多くの実用用途に使われてきている．そのひとつは携帯電話や IP 電話むけの双方向通信用の音声符号化である．これは通常低ビット，低遅延，音声専用の符号化で出力音声の帯域は狭いものである．二番目の分類としてあげられるのは放送や音楽の蓄積や再生のための片方向通信用の音響符号化である．これは一般にビットレートが高く，遅延も大きいが，出力音声帯域も広く，音楽なども含めた高品質の符号化が中心である．これらの二つの符号化は圧縮率優先の高圧縮符号化で，ビットレートの制約のもとに品質はできるだけ劣化させないことを狙うが，これとは別に歪を許さないカテゴリがある．これはロスレス符号化または可逆符号化とも呼ばれ，歪を許さないという制約の中で情報量をできるだけ削減することを狙うものである．

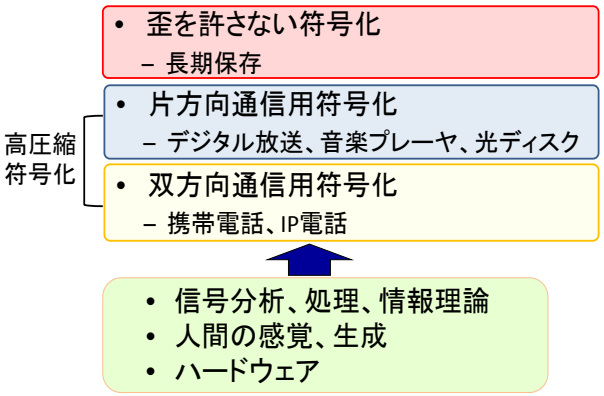

図 1　音声音響符号化の基盤技術と応用用途

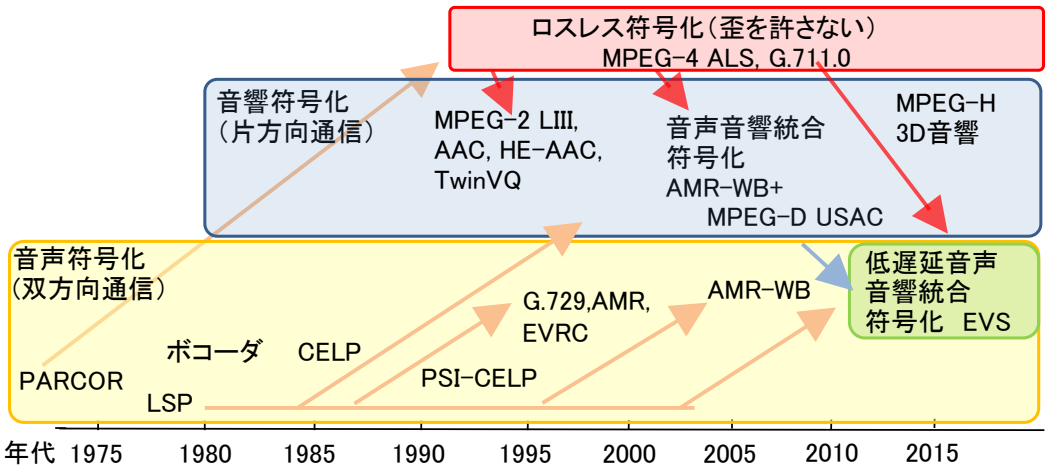

図 2　音声音響符号化技術の進展

日本電信電話(株) NTT コミュニケーション科学基礎研究所
Nippon Telegraph and Telephone Corporation, .
NTT Communication Science Labs.

図 2 はこれらの音声音響符号化技術の概要を時間軸上に表示したものである．音声信号を対象にしたデジタル信号処理の研究は AT&T ベル研や NTT の研究所などが中心になって 1960 年代から開始された. なかでも線形予測の基本技術は予測だけでなく周波数領域のスペクトルの推定理論によって音声の分析，合成，符号化，認識までの基本技術となった．音声分析，符号化の分野では予測パラメータの実用的で効率的な表現法として PARCOR（Partial Auto Correlation）や LSP (Line Spectrum Pair) といった技術が考案されてきた.

1990 年代には携帯電話のデジタル化，インターネットの普及に伴って，デジタル携帯電話，IP 電話の実用化が急速に進展した．これをささえる高能率の音声符号化アルゴリズムを規定するために，日米欧の携帯電話用の標準化，さらに ITU-T などの国際標準化がさかんに行われた，音響符号化の分野でも 1990 年代に ISO/IEC MPEG (Moving Picture Expert Group) で映像の圧縮符号化と並行して音響符号化の国際標準化が行われた．なかでも MP3 と呼ばれる MPEG-2 Layer III[1]や MPEG-2 AAC (Advanced Audio Coder)[2] はデジタル放送，音楽配信，プレーヤなどで広く使われるようになった.

21 世紀になると MPEG-4 ALS (Audio Lossless coding) [3] などの歪のない符号化の国際標準化も行われ，長期保存や配信，放送などに使われようとしている． 3GPP の AMR-WB+ [4] や MPEG-D USAC (Unified Speech and Audio Coding) [5-9] といった音声音響の統合符号化標準化されるようになった．ただいずれも長いフレーム長を要する遅延の大きい音響符号化に音声符号化を統合したものである. これに対し，2014 年に完成した 3GPP (3rd Generation Partnership Project) の EVS (Enhanced Voice Services) は双方交通信に使える遅延の短い音声符号化に，音響符号化を統合し，双方向通信に使える低遅延音声音響統合符号化になった.

次節以下で最近話題になっている 3 つの技術について紹介する．一つは昨年 IEEE Milestone に認定された LSP, 二番目は日本の 4K/8K 超高精細度テレビ放送の高音質サービス用に認定された歪のない符号化 MPEG-4 ALS，最後は将来の携帯電話用に昨年制定された 3GPP の EVS 規格であり. EVS については他より詳しく紹介する.

2. LSP

LSP は 1975 年 NTT 研究所に在籍中の板倉文忠先生によって考案され，管村昇先生や嵯峨山茂樹先生によって研究開発された全極型フィルターの係数と等価なパラメータセットである．図 3 に示すように，音声の生成過程を声帯振動を駆動信号として，声道の周波数特性をフィルター特に全極型フィルターとするモデルが広く使われている．LSP は全極型フィルターを安定に，能率よく符号化できる．LSP のパラメータ自体が周波数軸の値を持っており，スペクトル包絡との対応が素直であるため，LSP の値そのものの補間，高能率符号化，予測が可能であることが他のパラメータにない利点である．このため，音声合成や低ビットの分析合成(駆動音源をパルスと雑音でモデル化するボコーダ)符号化に有効であった．同時に駆動音源波形を高能率に符号化する CELP (Code Excited Liner Prediction) に代表される符号化でも有効であったため，携帯電話用として現在でも世界で広く使われている．また全極型モデルや LSP は音声に限らず，優れたスペクトル表現法として音響符号化にも使われ始めている．これらの世界の人たちへの幅広い貢献が認められ 2014 年には電気電子通信分野の技術の世界遺産に相当する IEEE Milestone に選定された[10]. 図 4 と 5 は記念に授与された銘盤と日本語訳である.

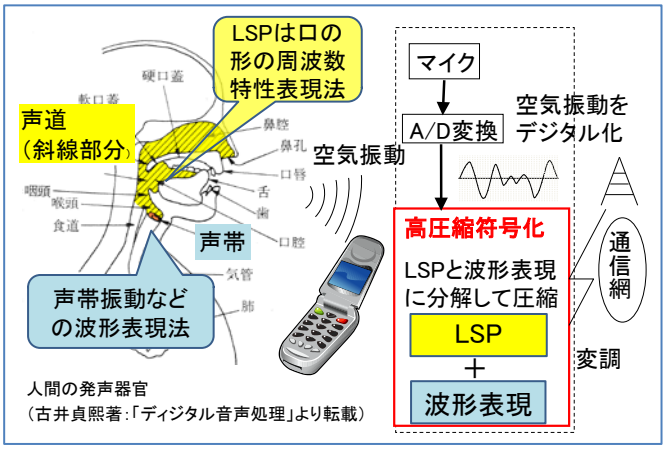

図 3　音声生成モデルと高圧縮符号化

図 4　IEEE マイルストン銘盤の写真

高圧縮音声符号化のための線スペクトル対(LSP), 1975

1975年にNTTで考案された線スペクトル対は音声合成や符号化のための重要な技術である。1980年には線スペクトル対に基づく音声合成チップが作成された。1990年代には、この技術はほぼすべての国際音声符号化標準に必須の要素技術として採用され、世界中の移動体やインターネットでの デジタル音声通信の高品質化に貢献した。

2014年5月

図 5　IEEE マイルストン銘盤の和訳

3. ALS

ALS は原音の線形 PCM 信号波形を 1 ビットの歪もなく圧縮できる符号化で，MPEG-4 ALS[3]として国際標準規格として制定されているものである．その背景と特徴を図 6 に示す．これまでの日本のデジタル放送の音声部分は AAC で符号化されていた．AAC の聴感品質は多くの主観評価の結果，ステレオ音声に対して 144kbps 以上であれば ほとんどの入力に対し，“原音との差はわかるが，劣化としては感じられない”とされている．図 7 は音楽信号の 1 フレームのパワースペクトルを原音と AAC で比較したものである．AAC は与えられたビットの範囲で，ステレオ信号での周波数領域の聴感モデルに基づいて，耳に差が分かりにくい成分を適応的に省略する．このため聴感的劣化はほとんどなく，もとの線形 PCM の 1.5Mbps のビットレートの 1/10 以下のビットレートに圧縮できている．

ただ音楽の製作者は AAC による劣化の程度よりはるかに細かい調整を行っていることや，音源位置が偏ったり動く信号，拍手など音源数がきわめて多い信号などでの聴感的差がわかる場合があり，放送局や制作関係者からは原音のままの放送が望まれていた[11]．放送のビデオ部分の規格が 4K/8K に対応する機会に，ビデオのビットレート増加に伴う全体のビットレートの増加や大画面の音像定位の改善のため，歪のない音声符号化が選択可能となった．

ALS はデジタル信号を完全に再構成するという制約のために，ビットレートは線形 PCM の約半分程度にしか圧縮できない．このため AAC のビットレートより数倍多くなってしまうが，これは全体のビットレートの増加のバランスで吸収されうるものである．またビットレートは入力音声の性質で時間的に変動するが，この変動はデータ放送の伝送ビットに有効利用でき，放送の運用には支障がないことが ARIB での実証実験で示されている．

これらの問題が解消され，放送スタジオで丹念に作りこまれた音声や音楽がありのまま，衛星放送や IPTV で家庭に届けられるようになる見込みである．また ALS はサンプリングレート，遅延，チャンネル数が自由に選択できるので，放送だけでなくさまざまな高品質コンテンツの伝送蓄積に柔軟に利用できる．また医療用信号，環境センサー信号，変調後の電波のデジタル信号の圧縮にも利用できる．

- 2002年 NTT、ベルリン工大、I2R、RealAudio標準化を提案
- 2005年 MPEG-4 (ISO/IEC 14496-3) ALSとして標準完成
- 2014年 ARIB標準、総務省令で超高精細度テレビ(4K/8K) 放送の高音質サービス用に制定
  - MPEGビデオ符号化、多重化標準と整合
  - 圧縮後のビットレートはもとの線形PCMの平均半分程度
  - ハイレゾ、低遅延、多チャネル等に柔軟に対応
  - 臨場感伝送、生体、環境、電波信号への展開可能

図 6　ALS の背景と特徴

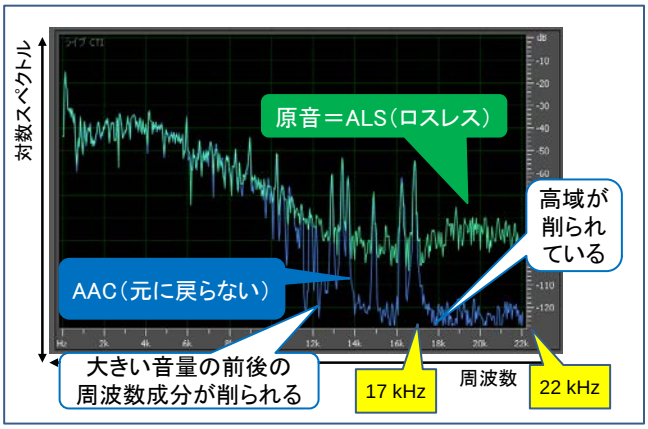

図 7　AAC と原音(=ALS) とのスペクトル比較

4. EVS

4.1 EVS の経緯

現在世界で使われている大多数の携帯電話は第 3 世代 (3G) の規格で，その音声符号化方式は 1995 年ころの技術による標準化方式がそのまま使われている．音声帯域は 0.3 から 3.5 kHz までで，音楽に対して不愉快な出力となるといった問題があった．この間，歪のない符号化，帯域を拡張する符号化，音声と音響を統合する符号化，ビットエラーでなくパケット消失の劣化を改善する符号化などの開発や標準化が進展した．しかし携帯電話のような大規模に使われている音声符号化方式を更改することはシステム全体の大幅な変更や円滑な移行手順が伴うことになり，容易ではない．携帯電話が 3G の回線交換から LTE になる機会に音声符号化も新たな統一規格の制定が望まれてきた．3GPP では 2010 年より，LTE を想定した EVS 規格の制定を目標に活発な活動が開始され 2014 年に完成した[12-22].

4.2 EVS の特徴

まず，世界の有力機関 12 社による統一規格ができたことが大きな特徴である．また表 1 のように幅広い入力の帯域とビットレートに対応している，（NB：狭帯域 8kHz サンプル，WB：広帯域 16kHz サンプル，SWB：超広帯域 32kHz サンプル，FB：全帯域 48kHz サンプルで，5.9 kbps 以外はフレームごとの固定ビット割り当てで，チャネルアウェアとは伝送状況に合わせた適応符号化モード選択可能）．また図 7 に示されるように，ごく最近使われ始めた AMR-WB と完全互換性を保ったまま性能が改善された AMR-WB 互換モードや（DTX：Discontinuous Transmission）に対応する無音圧縮符号化なども備えている．

符号化の性能の観点では，EVS 規格はこれまでの標準規格で達成されなかった 4 つの符号化性能のすべてを満たす最初の標準となった．すなわち，ビットレートは低いものまで対応する“高圧縮”，双方向通信を実現する“低遅延”，高いサンプリング周波数まで対応できる“広帯域”，音楽も

含めた再生音声の品質が高い“高音質”のすべてが満たされていることになる．これらの特徴を次節以下で紹介する．

表 1　EVS が対応するビットレートと音声帯域

ビットレート(kbps)	入力音声帯域
5,9 (可変レート)	NB，WB
7,2	NB，WB
8.0	NB，WB
9,6	NB，WB，SWB
13,2	NB，WB，SWB
13,2 (チャネルアウェア)	WB，SWB
16,4	NB，WB，SWB，FB
24,4	NB，WB，SWB，FB
32	WB，SWB，FB
48	WB，SWB，FB
64	WB，SWB，FB
96	WB，SWB，FB
128	WB，SWB，FB

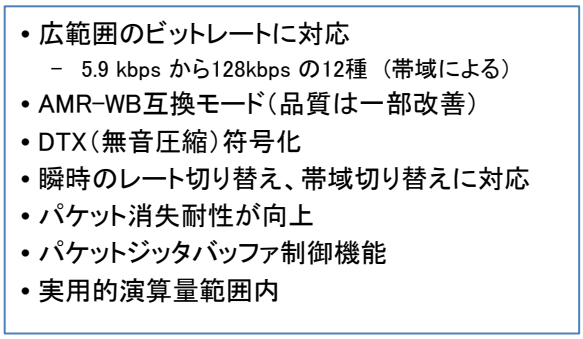

図 7　EVS の主な機能

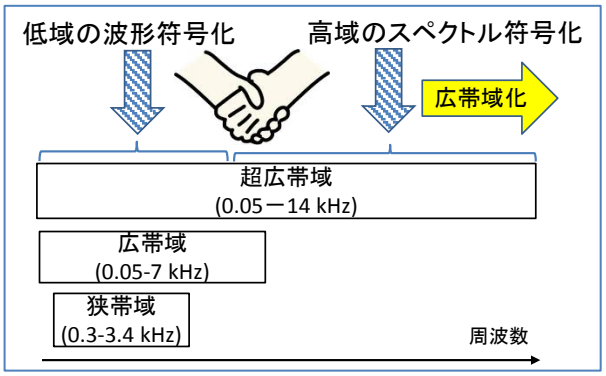

図 8　高域のスペクトル符号化との統合

4.3 国際協調

EVS の標準化作業は要求条件などの制定，競争的予備試験，最終選択試験，特性評価試験を経て，結果的に有力 12 社による共同案が標準化仕様となった．この 12 社の中には移動通信サービスを行っている Orange，NTT DOCOMO，通信機器を扱う Ericsson，Huawei，Nokia，ZTE，電話機やチップを製造する Qualcomm，Samsung，Panasonic，研究開発機関である Fraunhofer，VoiceAge，NTT が含まれている．通常はこれらの組織のエンジニアはライバルであり，各社の利害は激しく対立し，符号化技術の策定は難航したが，世界最高の品質を世界の人たちにタイミングよく提供しようという熱意という点では結束することができた．この結果タイミングよく市場で使われる可能性が高まった．

4.4　技術の統合（広帯域）

入出力の信号帯域がひろがり，サンプリング周波数が高くなると，単純に符号化情報量が増える．情報量をあまり増やさずに広帯域化するには，低域の波形符号化と高域のスペクトル符号化の統合が有効な手段で，1980 年代から提案はあったが，2000 年以降，盛んに研究が進んだ．別の言葉では，聴覚心理の知見を活用して高域の信号の位相成分を無視することに相当する．図 8 に示されるように高域は入力の信号を分析し，そのスペクトル包絡を非常に少ないパラメータで表現する．この手法の実現手段としては，時間領域でスペクトルを折り返すとか，サブバンド分割フィルタによる時空間表現とか，MDCT 係数での処理などがある．

4.5　技術の統合（高圧縮化）

高圧縮符号化のさらなる効率化のために，図 9 のように歪のない符号化でさかんに使われている可変長符号化が多面的に組み込まれている．可変長符号は別名エントロピー符号とも呼ばれ，パラメータなどの情報の冗長性を排除して本来の情報量に近い情報量までの圧縮を実現する．この一方，符号語の単位の長さ(ビット数)が固定でなくなり，符号誤りが混入すると符号語の単位の境界も誤り，ビット列全体の復号結果は無意味になってしまう可能性がある．従来の音声符号化は回線交換を前提として，デジタル圧縮ビット列に符号誤りが生じることは避けられなかったため，可変長符号は利用されることはなかった．一方 LTE ではパケットベースでの伝送であるため，上位のレイヤーで誤りが制御される．したがって，パケットの中の符号列に誤りが含まれることはなくなり，可変長符号を使って圧縮率を高めることが可能となった．

EVS ではフレームごとのビット数は通常固定で，可変長符号を使いつつフレーム内での総ビットを一定に保つためのさまざまな工夫が行われる．

EVS では符号化に先立つ前処理で多面的な分析が行われ，そのパラメータに依存した多数の符号化モードの切り替えが行われる．入力に適合した圧縮符号化が選択できるので高圧縮化が可能となる．このような高度化ももしモード切替情報に符号誤りが生じた場合の被害が大きいので，これまでの回線交換用音声符号化では使われていなかった．

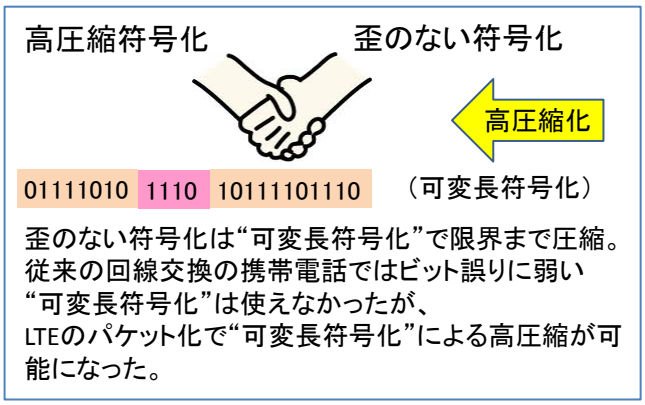

図 9　歪のない符号化技術の統合

4.6　技術の統合（低遅延化）

音声入力に対しては時間領域の CELP 方式のほうが品質が高く，音響信号が入力の場合は MDCT を使った周波数領域の符号化の品質が高くなることが知られている．これらの符号化を完全に統合することは一つの夢であるが，現状の品質維持のためには，時間領域と周波数領域の組み合わせが現実的選択となり，適応的分類，モード切替，なめらかな遷移などの工夫が必要になる．

さらに周波数領域の音響符号化はこれまでたとえば 80ms のフレームを使わないと能率的に符号化できないと考えられてきた．ところが EVS ではフレーム長 20ms，サンプルの先読み，オーバーラップも含めて符号化のアルゴリズム遅延が 32ms 以内であることが必須の要求条件と決められている．このため音響符号化の低遅延化が必須である．調波成分がはっきりする音楽では低遅延化による劣化が著しく，これを解決するために，周波数領域での量子化歪削減，調波成分の効率的表現などの新たな考案により劣化を抑えることができた．また周波数領域で線形予測モデルを用いるスペクトル表現を行う MDCT 符号化と不等間隔の帯域ごとのエネルギーでスペクトル表現を行う MDCT 符号化はこれまでの音響符号化の 2 つの主流であったが，EVS では両方を備えて，より性能の高いと推定される方法を適応的に選択する．

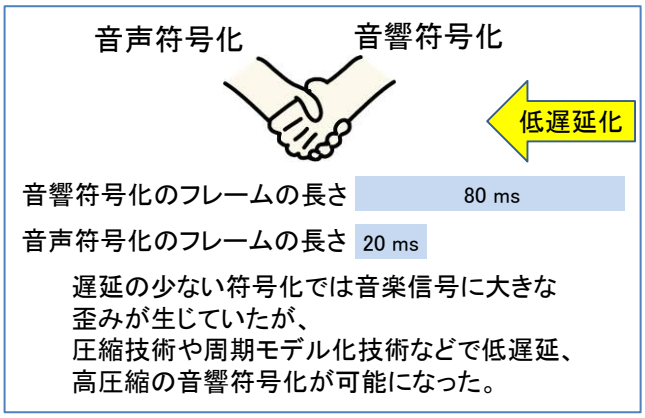

図 10　周波数領域の低遅延符号化

表 2　EVS の性能の達成

	電話用符号化	放送用符号化	音声音響符号化	低遅延音響符号化	VoLTE用符号化	
レート [kbps]	10	48	16	24	5.9–128	高圧縮
遅延 [ms]	30	80	80	30	32	低遅延
帯域 [kHz]	4	24	24	24	24	広帯域
音楽品質	劣化	良好	良好	良好	良好	高音質
標準例	ITU-T G.729 3GPP AMR	MPEG MP3 MPEG AAC	AMR-WB+ MPEG USAC	AAC-ELD	3GPP EVS	

4.7　EVS での達成

前節までの技術的特徴を統合して，これまでの主な標準化方式とおおざっぱに比較したものが表 2 である．オレンジ色の枠はより好ましい性能を示すものである．これまでのさまざまな用途に開発されてきた標準化方式では，高圧縮，低遅延，広帯域，(音楽の)高音質の 4 つの要件を同時に達成できるものは存在しなかった．今回の EVS で初めて達成できたことになる．

4.8　EVS の評価

EVS では大規模な性能評価試験を行い，その結果と分析をテクニカルレポートで公開している[13]．図 11 はこれらの試験とは別に独自に行った主観評価結果を示したものである．雑音下の日本語音声（女声 3 種，男声 3 種）を入力とし，帯域の異なる 3 つの符号化を混在させて参照音と比較して同時に評価しており，横軸はビットレート，縦軸は一般の評価者 16 名の平均主観評点である．符号化方式の一つ目は 3G 携帯電話で広く使われている狭帯域（8kHz サンプル）AMR[23]，二つ目は現在 VoLTE で広く使われている広帯域（16kHz サンプル）AMR-WB [24]，三つ目が超広帯域（32kHz サンプル）の EVS である．12～13 kbps 付近の類似のビットレートで比較すると，EVS の品質の高さが明らかである．この傾向は音楽や音声音楽の混合入力などでも確認されている．

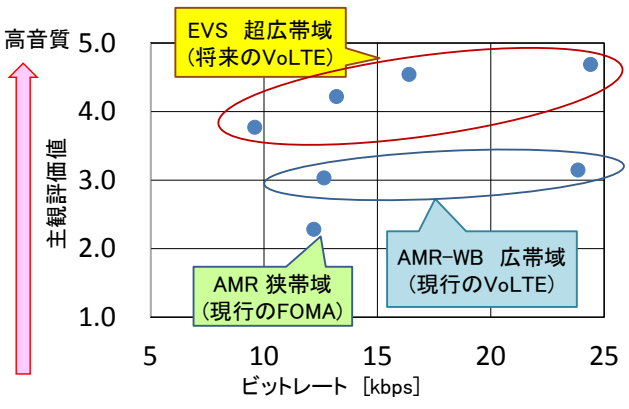

図 11　これまでの携帯電話と比較した EVS の主観品質

2.9　EVS の応用用途

電話の長い歴史で音声の周波数帯域を振り返ってみたい．1876 年に電話が発明されてからおよそ 100 年はアナログ伝送で，音声帯域はほぼ 3.5 kHz であったと思われる．その後伝送や交換がデジタル化されるなど，さまざまな技術の進歩があったが，おもに経済的理由で音声帯域は変わらなかった. 技術的には 1980 年代には広帯域音声符号化が標準化されたが，対応する電話器が一部の TV 電話だけにとまっていた．自分の電話機が広帯域になっても相手の電話機が対応していないと有効に使えないことが普及の障害になったと考えられる．EVS 規格という国際統一規格として，情報量をほとんど増やさないで超広帯域化ができ，比較的短時間で買い替えられるスマホに搭載されるという条件がそろってはじめて，超広帯域の電話をたくさんの人に使ってもらえる可能性が出てきた．電話の発明以来，百数十年をへてようやく電話の音声帯域が大きく広がることになるのではないだろうか.

EVS は LTE を目標に設計, 最適化された符号化であるが，音声や音楽の圧縮伝送蓄積の部品として幅広く有効利用できる．これらの用途の中には移動通信に限らず，特に IP を伝送プロトコルとする幅広い一般の通信, テレビ電話, web 会議などがある．さらに低遅延の音楽の符号化がスマホでできることで，ゲーム，ユーザー作成の音楽関係のアプリなどへの用途が考えられる．

5. おわりに

音声音響符号化の概略を振り返り, 特に 2014 年に話題となった LSP，ALS，EVS について紹介した．筆者の偏った想像を許していただけるなら，LSP は引き続きさまざまな符号化で使われ続けると思われる．ALS は日本の放送で衛星放送だけでなくさらに地上波に，さらに国外の放送にも使われている可能性がある．EVS は世界の携帯電話，さらに一般の電話などにも広く使われている可能性がある．

音声音響に限らないが, サービスの性能や品質の改善努力が停止することはありえないが，これまでのような形式の標準化活動の意義は低下するものと思われる．今後はソフトウェアの共有で互換性を保証する形で高品質化が進展するものと思われる. EVS の次に思いをはせて、”Simplicity is the seal of truth”という格言を思い起こせば，EVS はまだまだ発展途上と思わざるを得ない.

謝辞

LSP にかかわる技術開発，IEEE Milestone 認定にご尽力いただいた関係者，ALS の共同開発と標準化，日本での規格化に尽力いただいた関係者，EVS の共同開発と標準化にご尽力いただいた関係者一同に感謝する．

参考文献

1) ISO/IEC 13818-3，- Information technology - Generic coding of moving pictures and associated audio information，Part 3: Audio，1994.

2) ISO/IEC 13818-7:，- Information technology - Generic coding of moving pictures and associated audio information Part 7: Advanced Audio Coding, 1997.

3) ISO/IEC 14496-3:, - Information technology - Coding of Audiovisual Objects，Part 3: Audio，Subpart 11, Audio Lossless Coding，2005.

4) 3GPP TS 26.290 Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions.

5) ISO/IEC 23003-3 - Information technology - MPEG audio technologies Part 3: Unified speech and audio coding，2011.

6) M. Neuendorf, et. al, “Unified speech and audio coding scheme for high quality at low bitrates,” Proc. ICASSP, 2009.

7) 守谷，"音声・音響符号化技術と標準化動向"，電子情報通信学会技術研究報告. SP, 音声 109(57), 55-58, 2009.

8) 菊入，仲，“音声と音楽を効率的な圧縮を実現する MPEG 標準・音声音響統合符号化方式” NTT DoCoMo テクニカルジャーナル vol. 19, No. 3, pp.18-23, 2011.

9) 菊入, “音声音響符号化技術”, 電子情報通信学会誌 Vol.96, No.11, pp. 882-887, 2013

10) 守谷 "高圧縮音声符号化の必須技術 : 線スペクトル対（LSP）", NTT 技術ジャーナル, Vol.26, No.9, pp. 58 - 60, 2014.

11) 総務省, “放送システム委員会報告（案）に対する意見の募集（超高精細度テレビジョン放送システムに関する技術的条件について）”，2014.

12) 3GPP TS 26.445，“Codec for Enhanced Voice Services (EVS); Detailed Algorithm Description (Release12),” 2014.

13) 3GPP TR 26.952、“Codec for Enhanced Voice Services (EVS); Performance characterization”

14) 守谷，鎌本，原田，菊入，堤，仲，大崎，江原，三田，河嶋，中尾, “3GPP 標準 EVS コーデックの概要－VoLTE 用高性能音声音響符号化－”，信学技法 2015

15) 三田，菊入，原田，“3GPP 標準 EVS コーデック向け通信制御技術”，電子情報通信学会総合大会, B-6-35, 2015.

16) 江原，三田，河嶋，スリカンス，リウ，中尾，“3GPP 標準 EVS コーデック向け低レート超広帯域 MDCT 符号化 “，電子情報通信学会総合大会 D-14-9, 2015

17) 守谷，鎌本，原田, ”3GPP 標準 EVS コーデック用 MDCT ベースの TCX “，電子情報通信学会総合大会，D-14-10. 2015.

18) 菊入，3GPP 標準 EVS コーデックにおける時間領域帯域拡張向け TEC 技術 “，電子情報通信学会総合大会, D-14-1, 2015.

19) 堤，菊入，“VoLTE のさらなる高音質化と音楽の活用を実現する 3GPP 標準音声符号化方式 EVS”、NTT DoCoMo テクニカルジャーナル vol. 22, No. 4, pp.18-23, 2015.

20) SS-L6:Special Session, Enhanced Voice Services I, ICASSP 2015

21) SS-P1:Special Session, Enhanced Voice Services II, ICASSP 2015

22) SS-P2:Special Session, Enhanced Voice Services III, ICASSP 2015.

23) 3GPP TS 26.071 Mandatory speech CODEC speech processing functions; AMR speech CODEC.

24) 3GPP TS 26.190 Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions