音楽情報科学　29－16
（1999. 2. 19）

5kHz 帯域低ビットレート音声符号化
― 帯域感と主観品質の関係 ―

原田 登　大室 仲

NTT サイバースペース研究所

〒180-8585 東京都武蔵野市緑町 3-9-11

harada@splab.hil.ntt.co.jp

あらまし　近年、低ビットレートの音声符号化方式が必要とされる一方で、より情報量の多い自然な音声に対する要求も次第に高まっている。本稿では、帯域を混在させた評価実験によって帯域感と主観品質の関係を明らかにすると同時に、そこで得られた知見から低ビットレートと自然性を両立する選択肢のひとつとして 5kHz 帯域の音声を対象とする符号化方式を提案する。主観品質評価の結果、提案する 5kHz, 7.8kbit/s の符号化方式は、G.729 (3.4kHz, 8kbit/s), G722 (7kHz, 48kbit/s)と同等以上の高い評価を得た。

キーワード　音声符号化、5kHz 帯域、低ビットレート

A 5-kHz-Bandwidth Low-Bit-Rate Speech Coder

Noboru Harada and Hitoshi Ohmuro

NTT Cyber Space Laboratories
3-9-11 Midori-cho,
Musashino-shi, Tokyo 180-8585, Japan
Harada@splab.hil.ntt.co.jp

Abstract　In this paper, we propose a 5-kHz-bandwidth speech coder. In order to achieve both low bit-rate and naturalness, 5-kHz-bandwidth speech signals are used (instead of 3.4 kHz or 7 kHz). The speech signals under consideration are band-limited to 5 kHz and are sampled at 11.025 kHz. Subjective tests (CMOS and MOS) indicated that a 5-kHz-bandwidth is effective for the speech coders. It makes the speech much more natural than 3.4kHz coders, and operates at a lower bit rate than that of 7-kHz-bandwidth coders. The MOS and CMOS showed that the quality of this coder (5 kHz, 7.8 kbit/s) is better than that of the G.729 (3.4 kHz, 8 kbit/s), G.722 (7 kHz, 48 kbit/s) and equivalent to that of G.729 annex E (3.4 kHz, 12 kbit/s). We also determine the relationship among characterizations of objective quality, bandwidth and S/N ratios.

key words　speech coder, 5kHz-bandwidth, low bit rate

1. はじめに

従来、音声符号化方式に関する研究は低ビットレート化の方向に進んできた。

低ビットレートな音声符号化方式では、符号化に必要な情報量との兼ね合いやアナログ回線時代の歴史的な経緯から、3.4kHz 程度に制限された電話帯域の音声信号を用いるのが一般的である。電話帯域の音声を対象とする符号化方式としては、ITU-T の標準方式である G.729 (3.4kHz, 8kbit/s), G.723.1 (3.4kHz, 6.3/5.3kbit/s)などがある。現在は 4kbit/s の標準方式の策定も行われている。

しかし、低ビットレートの音声符号化方式が必要とされる一方で、近年のインターネット、モバイル情報端末の普及等、マルチメディア通信基盤の発達を背景に、より情報量の多い自然な音声に対する要求も高まっていることから、7kHz 帯域の広帯域音声を対象とする符号化方式もさかんに研究されている。

このように、低ビットレートと自然性の両方を兼ね備えた符号化方式が求められているが、現状ではその両立は難しい。たとえば、電話帯域の符号化方式を用いた場合にはビットレートは低いが臨場感の面で不満が残り、周波数帯域を広くして情報量を増やせば臨場感や自然性は増すが、伝送に必要なビットレートも増えるという問題点がある。これに対して、対象とするビットレートや帯域幅を複数種類実装し、目的に応じて選択して使用するというような提案[2]もなされているが、低ビットレートモードでは電話帯域の制限された音声しか使用できないため、自然性の高い音声を低ビットレートで利用したいという根本的な要求条件を満たすことはできない。

本報では、上記のような問題点を解決するための選択肢のひとつとして、5kHz 帯域の低ビットレート CELP 音声符号化方式を提案する。

5kHz 帯域の音声は、3.4kHz 帯域の音声に比べて格段に自然性が高いにもかかわらず、その性質は 3.4kHz 帯域の音声に非常に似ているため、7kHz 帯域の音声に比べて、より少ないビットレートで符号化が可能である。今回は、11.025kHz でサンプリングされた音声を入力とし、7.8kbit/s で符号化する方式を実装した。また、主観品質評価実験によってその有効性を示すと同時に、帯域感と主観品質の関係についても考察する。

第 2 節では提案する 5kHz 帯域の音声符号化方式について説明する。第 3 節ではＭＯＳ主観品質評価試験の結果を示す。また、周波数帯域と SN 比、主観品質評価値の間の関係についても考察する。第 4 節で CMOS による対比較試験の結果を示す。最後に第 5 節でこれらをまとめる。

2. 5kHz 帯域音声符号化方式の提案

本研究では次のような方針で低ビットレートと自然性の問題を解決する。

(1)帯域を 3.4kHz から 7kHz に増やすのではなく 5kHz に広げることで必要な情報量とビットレートの間のバランスをとる。

(2)帯域を広げたことによる情報量の増加分は、量子化器の効率を上げることで吸収しビットレートは電話帯域の従来法と同程度に抑える。

上記方針で設計した 5kHz 帯域、7.8kbit/s の CELP 音声符号化方式のビット配分を表 1 に示す。

入力された信号は 11.025kHz でサンプリングされ、前処理として 5kHz のローパスフィルタに通される。LSP の次数は 14 次、フレーム長 10ms (110samples)、サブフレーム長 5ms、先読み 5ms である。

表 1 ビット配分

LSP	18 (1+7+5+5)	
Pitch	8	5
Pulses	17	17
Gain	6	6
Parity	1	0
Total	7.8 kbit/s	

LSP 量子化

LSP の量子化では 4 次の移動平均予測を用いた 2 段ベクトル量子化[3][4]を行った。移動平均予測モードの切り替えに 1 bit、1 段目の量子化に 7bit、2 段目は 14 次の LSP 係数を下位 7 次と上位 7 次に分け、それぞれ 5bit で量子化する。

適応符号帳

適応符号帳では第 1 サブフレームに 8bit を割り当て、[27·1/3, 85·2/3]の範囲では 1/3 精度で、[86, 165]の範囲は整数精度で探索する。第 2 サブフレームでは第 1 サブフレームで得られたピッチ遅延 T_1 に対して[$int(T_1)$-5·2/3, $int(T_1)$+4·2/3]の範囲で 1/3 精度の探索を行う。ここで $int(T_1)$は第 1 サブフレームの分数ピッチ遅延 T_1 の整数部分である。

固定符号帳

固定符号帳は ACELP 構造[4][5]で、5ms (55 samples)のサブフレームに対して 4 本のパルスを立てる。表 2 にパルス配置を示す。

表 2 パルス配置

Pulse	Sign	Position
i0	±1	0,7,14,21,28,35,42,49
i1	±1	1,8,15,22,29,36,43,50
i2	±1	3,10,17,24,31,38,45,52
i3	±1	4,11,18,25,32,39,46,53 5,12,19,26,33,40,47,54

ゲイン量子化

固定符号帳ゲイン g_fは、4 次の移動平均予測でサブフレーム間の相関を取り除き、適応符号帳ゲイン g_a の値と組にして 6bit でベクトル量子化する。

ゲイン探索は歪み最小基準で行い、距離計算には次のような式を用いる。

$$d = \left\| \boldsymbol{X} - g_a \boldsymbol{Hp} - g_f \boldsymbol{Hc} \right\|^2 \quad (1)$$

ここで d はターゲットベクトルと候補ベクトルの間の距離、$\boldsymbol{X}$, $\boldsymbol{H}$, $\boldsymbol{p}$, $\boldsymbol{c}$ はそれぞれターゲットベクトル、インパルス応答、適応符号ベクトル、ピッチ周期化された固定符号ベクトルである。

なお、ゲイン符号帳学習時に(1)式を用いて、学習用データ系列に対する総歪みが最小となるように学習した場合には、符号帳のセントロイドベクトルがパワーの大きな要素に偏って学習されてしまうため、ビットレートを削減した場合に、比較的音量が小さな場所で聴覚的な品質劣化が生じる等の問題があった。この問題を解決するため、ゲイン符号帳の学習時には(2)式に示すような距離尺度を用いて学習用データに対する歪み d'の総和が最小になるように符号帳を学習した。

$$d' = \frac{(|\boldsymbol{X}|^2)^{0.3}}{(|\boldsymbol{X}|^2)} \left\| \boldsymbol{X} - g_a \boldsymbol{Hp} - g_f \boldsymbol{Hc} \right\|^2 \quad (2)$$

上式はターゲットベクトルのパワーが小さいほど大きな重みをつけて学習することに相当する。これによってパワーの小さな部分に対応する符号帳要素の比率が増加し、符号化品質が向上する。ただし符号化時の距離計算には従来と同じ(1)式の距離尺度 d を用いる。

ポストフィルタ

サブフレームあたりの固定符号帳パルス本数が少ないことに起因する歪みを解消するため、ポストフィルタには位相拡散フィルタ[6][7]を用いた。

5kHz 帯域を用いる利点

5kHz 帯域の音声を入力対象とすることによって、3.4 kHz 帯域に対して次のような利点がある。

(1) 3.4kHz～5.5kHz 帯域の部分は、3.4kHz 以下の成分と相関が高く、LSP 予測次数を上げても VQ 符号帳のビット配分は同程度で良い。

(2) 3.4kHz～5.5kHz 帯域の部分はピッチ予測可能な範囲内なので適応符号帳のビット配分も 3.4kHz と同程度でよい。それに加えて、サンプリングレートが上がったことによって相対的に適応符号帳の時間分解能があがるという効果がある。

(3) (1),(2)の理由で LSP と適応符号帳の効率が上がっていることに加えて、パルス位置の時間分解能が上がっていることから、固定符号帳のビット数も電話帯域の符号化方式と同程度で良く、非常にスパースに立てることができる。

3. MOS評価試験結果と考察

符号化方式の性能を評価するために、フラット特性を持った周波数帯域 7kHz, 5kHz, 3.4kHz, 1.8kHz の音声信号と、3.4kHz 音声に Modified IRS フィルタをかけた音声信号(MIP 音声)のそれぞれについて原音と MNRU(振幅相関雑音付加音声) 40dB, 30dB, 20dB, 10dB を用意し、3.4kHz 帯域の MIP 音声を入力とした符号化方式には G.723.1 (5.3/6.3kbit/s)および G.729 (8kbit/s)を用いた。3.4kHz 帯域の音声符号化方式として G.726 (32kbit/s), G.723.1 (5.3/6.3kbit/s), G.729 (8kbit/s), G.729 Annex E (12kbit/s)を、7kHz 帯域の符号化方式としては G.722 (64kbit/s, 56kbit/s, 48kbit/s)を用いた。また、比較のために 3.4kHz 帯域の符号化方式に 11.025kHz サンプリングの音声を入力として符号化した G.723.1 Base 11kHz (7.29/8.66kbit/s), G.729 Base 11kHz (11kbit/s)を使用し、第2節で述べた 5kHz 帯域 7.8kbit/s の提案方式とともに、男性、女性あわせて 14 音声の評価用音声を用いて、被験者14名による5段階絶対評価の MOS 評価試験を行った。ヘッドホン受聴によって得られた帯域感と S/N、および各符号化方式における主観品質の関係を図1に、スピーカ受聴によって得られた帯域感と雑音間、および各符号化方式における主観品質の関係を図2に示す。尚、評価には背景雑音等を含まないクリーンスピーチ音声を使用した。

帯域感と雑音感、主観品質評価値の関係

グラフから、帯域感が広いほど主観品質評価値が高く、3.4kHz と 5kHz では原音レベルで明らかな品質の差があることがわかる。同じ帯域同士を比較した場合には、従来の試験結果と同様の傾向がある。今回は実験の枠組みとして、様々な帯域、様々な性質のノイズを含む音声データをバランス良く用いた。ITU-T による標準化時の試験では 3.4kHz 帯域の原音より 48kbit/s の G.722 の方が評価値が高いとされているが[8]、データセットなど実験の枠組みによる違いとも考えられる。

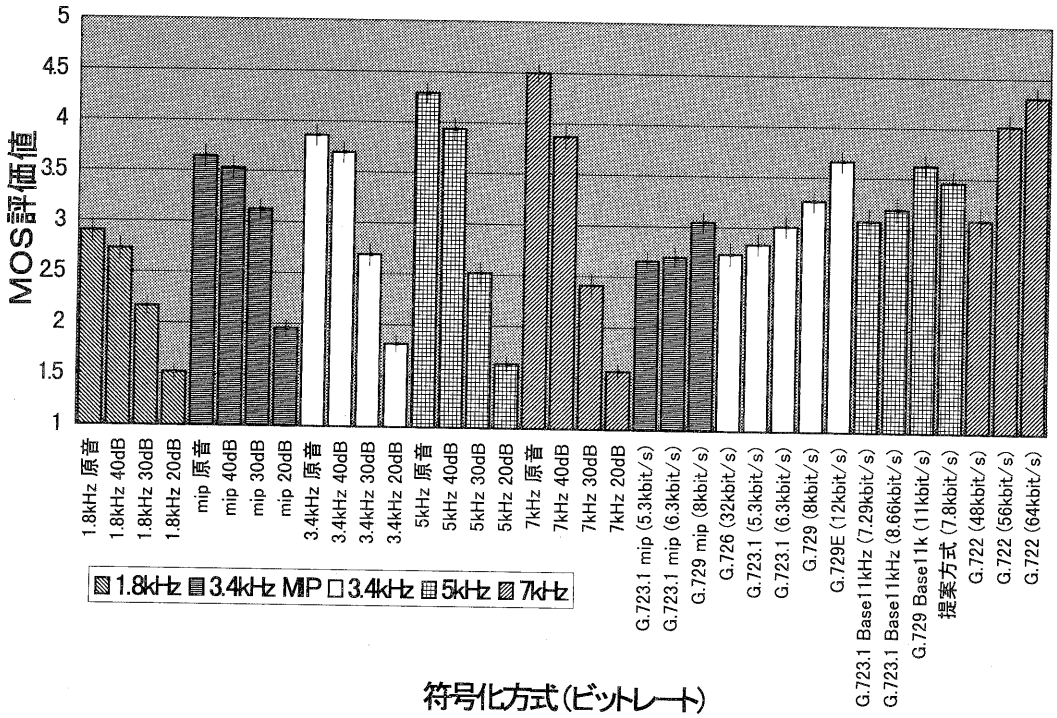

図1　MOS評価試験結果（ヘッドホン受聴）

MIP音声とフラット音声の比較

MIP音声は、3.4kHz帯域の原音にModified IRSフィルタをかけたもので、低域をカットして中域を持ち上げているため、フラットな音声に比べて明瞭度があがっているのが特徴である。MIP音声と、フラット特性を持った3.4kHz帯域の原音では、フラットな音声の方が高い評価を得、今回のような枠組みでは明瞭度よりも自然性が重視される結果となった。これには、ヘッドホンおよびスピーカの伝達特性、被験者の聴覚特性等も影響していると考えられるが、スピーカを用いるような拡声系の通信ではフラット音声を用いる方が良い品質が得られるといえる。

ヘッドホン受聴とスピーカ受聴の比較

原音およびMNRU音声では、ヘッドホン受聴でもスピーカ受聴でもほぼ同様の傾向があらわれた。これに対して、符号化音声の評価では多少傾向の違うものもあった。符号化歪みは単にS/Nだけで計れるものではなく、帯域感が狭まったり、自然性を損なうような特殊な歪みとして知覚されることが原因であると考える。ヘッドホン受聴では細部までS/Nを重視した評価となり、スピーカ受聴ではごく微細な雑音よりも、帯域感および自然性を重視した評価となる傾向がある。

提案方式の評価

5kHz, 7.8kbit/sの提案方式は、G.729 (3.4kHz, 8kbit/s), G.722 (7kHz, 48kbit/s) よりも評価値が高かった。スピーカ受聴では特に本方式の評価が高く、G.729 Annex E (3.4kHz, 12kbit/s)とほぼ同等の主観品質評価を得た。また、従来の3.4kHz帯域の符号化方式に11.025kHzサンプリングの音声を入力したものと比較した場合にも、同程度のビットレートでより高い評価値を得たことから、提案方式の有効性が確認された。

これらの結果から、本提案方式は電話帯域の音声符号化方式に対して同程度の低いビットレートで、より自然性の高い音声を符号化することが可能であるといえる。

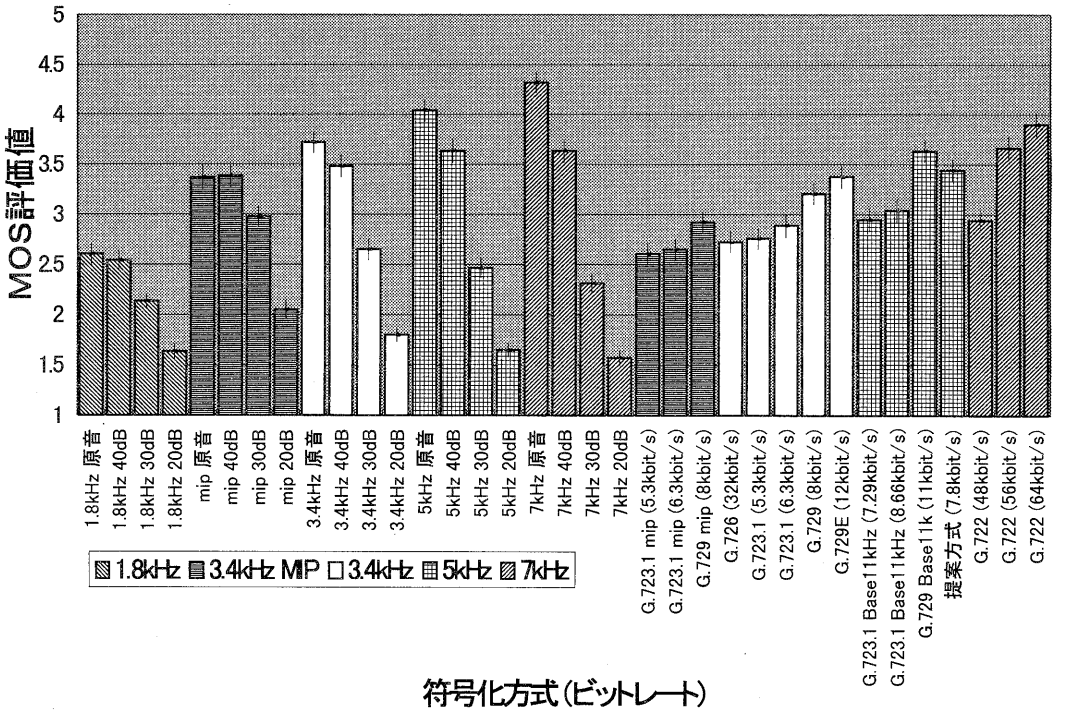

図2　MOS評価試験結果（スピーカ受聴）

4. CMOS評価試験

提案方式の品質を別の観点から確かめるために、対比較による 7 段階評価試験(CMOS)を行った。提案方式(7kHz, 7.8kbit/s) と G.729 (3.4kHz, 8kbit/s), G.729 Annex E (3.4kHz, 12kbit/s), G.722 (7kHz, 48kbit/s)のそれぞれを比較した試験の結果を表 3 に示す。

この結果、ヘッドホン受聴とスピーカー受聴の双方で、提案方式は比較した従来の符号化方式よりも品質が高いか、または少なくとも同等であるという結果を得た。

CMOS では、MOS に比べて提案方式と G.722 (48kbit/s)との差が少なくなっている。G.729 Annex E では評価が逆転している。このように、スピーカ受聴、ヘッドホン受聴に関わらず、対比較評価の場合には、絶対評価に比べて帯域感の差がより鮮明に知覚される傾向にある。

表 3 CMOS 評価試験結果

		Headphones		Speakers	
X	Y	CMOS (Y-X)	Std.	CMOS (Y-X)	Std.
提案方式(5kHz,7.8kbit/s)	G.729 (3.4 kHz,8 kbit/s)	-0.843	0.88	-0.979	0.935
	G.729 Annex E (3.4 kHz,12 kbit/s)	-0.244	0.817	-0.302	0.924
	G.722 (7 kHz, 48kbit/s)	-0.099	1.088	-0.171	1.074

5. まとめ

本報では、帯域感と主観品質の関係について調べ、そこで得られた知見から、自然性と低ビットレートを両立させる選択肢として 5kHz 帯域の音声符号化方式を提案した。主観品質評価実験によって 5kHz, 7.8kbit/s の提案方式は、G.729 (3.4kHz, 8kbit/s), G.722 (7kHz,48 kbit/s) よりも評価値が高い結果を得た。特に、スピーカ受聴では G.729 Annex E (3.4kHz,12kbit/s)とほぼ同等の高い評価を得た。

これらのことより、5kHz 帯域の音声を入力とする本提案方式の有効性が示された。

提案方式は、5kHz 帯域の音声を用いることで、従来の 3.4kHz 帯域の符号化方式と比べては、同程度のビットレートでより自然性の高い音声を符号化でき、7kHz 帯域の符号化方式と比べては非常に低ビットレートを実現できることから、マルチメディア用途のアプリケーションに適応可能である。

参考文献

[1] W. B. Kleijn, K. K. Paliwal, "Speech Coding and Synthesis," 1995

[2] 野村俊之, 岩垂正弘, 田中直哉, "MPEG-4/CELP 音声符号化方式", 信学技報, SP98-89, pp.19-26, Nov. 1998

[3] 大室仲, 守谷健弘, 間野一則, 三樹聡, "移動平均型フレーム間予測を用いる LSP パラメータのベクトル量子化", 信学論 A Vol.J77-A No.3 pp.303-313, 1994

[4] ITU-T COM 15-152-E, "G.729 - Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited-Linear-Prediction (CS-ACELP)," Jul.1995.

[5] R. Salami, C. Laflamme, and J-P. Adoul, "8 kbit/s ACELP Coding of Speech with 10 ms Speech-Frame: a Candidate for CCITT Standardization," IEEE Proc. ICASSP '94, pp. II-97, 1997.

[6] 安永他, "パルス拡散構造音源を併用する ACELP 符号化", 信学総大,D-14-11, pp.253, 1997.

[7] R. Hagen et al., "Removal of sparse-excitation artifacts in CELP," Proc. IEEE ICASSP '98, pp. 145-148, 1998.

[8] P. Mermelstein: "A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals," IEEE Communication Magazine, January, pp.8-15, 1988.