論　文

全波整流に基づくステガノグラフィを用いたG.711音声の一帯域拡張法

青木　直史[†a)]

A Band Extension Technique for G.711 Speech Using Steganography Based on Full Wave Rectification

Naofumi AOKI[†a)]

あらまし　狭帯域コーデックにより符号化された電話音声は，帯域制限のため，こもった品質として知覚されやすい．本研究では，最も基本的な狭帯域コーデックであるG.711により符号化された音声データに対して擬似的に帯域拡張を施す信号処理手法について検討している．提案法は，帯域拡張に必要な補助情報をステガノグラフィにより音声データそのものに埋込伝送する．そのため，提案法は，補助情報を利用することによる効果的な帯域拡張を実現できるばかりでなく，通話における見かけ上の伝送量を増加させる必要がないという利点を有する．

キーワード　G.711，帯域拡張，ステガノグラフィ，VoIP

1. まえがき

狭帯域コーデックによって伝送される電話音声の周波数帯域は，音声データの標本化周波数を8 kHzとし，音声データの最高周波数を4 kHzに制限したものになっている．

本来，狭帯域コーデックは，従来のアナログ固定電話の通話品質を目標として策定されているため，4 kHz以上の高域成分については符号化の対象になっていない．そのため，こうした帯域制限により，ハイファイ感の乏しい，こもった品質となってしまうことが，狭帯域コーデックの問題とされている[1], [2].

本研究では，狭帯域コーデックにより符号化された音声データに対して擬似的な帯域拡張を施し，4 kHz以上の高域成分を付加する信号処理手法について検討している[3], [4]．この信号処理手法は，音声データの標本化周波数を8 kHzから16 kHzに，すなわち音声データの最高周波数を4 kHzから8 kHzに拡張するものである．

† 北海道大学大学院情報科学研究科，札幌市
Graduate School of Information Science and Technology, Hokkaido University, N14 W9, Kita-ku, Sapporo-shi, 060–0814 Japan
a) E-mail: aoki@nis-ei.eng.hokudai.ac.jp

具体的な狭帯域コーデックとして，本研究ではG.711 [5]を対象とした．G.711は標本化周波数8 kHzの音声データを1サンプルごとに対数量子化し，8 bitで符号化を行うものである．G.711は圧縮率が低く冗長度の高いコーデックであるが，実装が簡単であることから幅広く利用されており，様々な端末間の相互接続性に優れている．そのため，近年普及が進むVoIP（Voice over IP）では，G.711を必須コーデックとして位置づけている[6].

音声データを擬似的に帯域拡張する信号処理手法としては，これまでに全波整流を利用した手法が提案されている[7]．この従来法は，低域における調波構造を高域にまで外挿する手法であり，調波構造を示す音声データの帯域拡張に有効であると報告されている．

しかし，従来法は帯域拡張におけるパラメータを時不変にしているため，音声データの種類によっては，ハイファイ感が十分に得られないことが報告されている[7]．そこで，本研究では，あらかじめ送信側で帯域拡張に必要なパラメータを抽出し，これを補助情報として伝送することにより，従来法の改善を試みている．

なお，補助情報を伝送する手法として，本研究の提案法では，ステガノグラフィと呼ばれる情報埋込技術により，補助情報を音声データそのものに埋込伝送している．こうした工夫により，提案法は，補助情報を

利用することによる効果的な帯域拡張を実現できるばかりでなく，通話における見かけ上の伝送量を増加させる必要がないという利点を有するものになっている．

2. 全波整流による帯域拡張

従来法は，全波整流による帯域拡張を基本としている [7]．この手法は，音声データをはじめ多くのサウンドデータでは，低域から高域まで基本周波数とその倍音から構成される調波構造が見られること，すなわち高域成分と低域成分には高い相関があるという事実を原理としている．こうした特徴を考慮し，低域成分から高域成分を擬似的に生成するには，低域の調波から高域の調波を発生させる非線形処理の適用が有望であると考えられる [8]．

図 1 に，全波整流による帯域拡張の処理手順を示す．従来法では，まず，BPF（Band-Pass Filter）によって 2 kHz から 4 kHz までの成分 $s_b(n)$ を分離し，これを全波整流することで，4 kHz 以上の高域成分を含んだ信号 $s_{br}(n)$ を生成する．次に，この信号から，HPF（High-Pass Filter）によって 4 kHz 以上の高域成分 $s_{brh}(n)$ を分離し，これに対して適当なゲインを乗じることによって重み付けを行う．その後，この信号を本来の 4 kHz 以下の低域成分 $s_l(n)$ に加算することで帯域拡張を実現する．

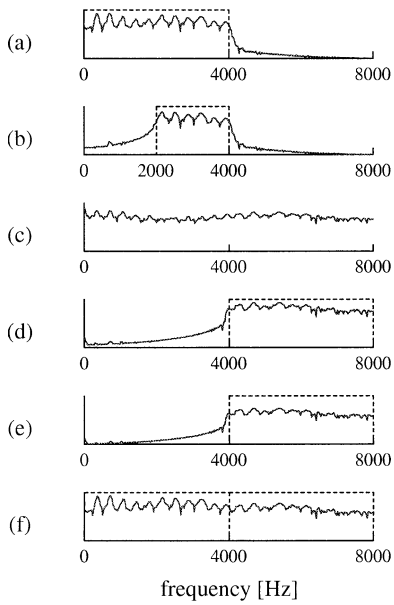

図 1　全波整流による帯域拡張の処理手順 (a) 原音声，(b) 帯域通過フィルタリング，(c) 全波整流，(d) 高域通過フィルタリング，(e) ゲイン制御，(f) 低域成分と高域成分の加算

Fig. 1　Procedure of the band extension technique based on full wave rectification: (a) original speech, (b) band-pass filtering, (c) full wave rectification, (d) high-pass filtering, (e) gain control, and (f) addition of the low and high bands.

ゲインを g とすると，従来法は次のように定義される．

$$s(n) = s_l(n) + g \cdot s_{brh}(n) \tag{1}$$

なお，従来法では経験的にゲインを 0.5 に設定し，これを時不変としている [7]．

3. 提 案 法

従来法では，高域成分のゲインを 0.5 に固定しているため，無声子音など，低域成分に比べて高域成分のパワーが大きくなる傾向を示す音声データでは十分な結果が得られないことが報告されている [7]．

また，有声音ではあっても，高域に明確な調波構造が見られない場合がある [9]．しかし，全波整流による帯域拡張を行うと，そうした音声であっても低域の調波構造に基づいた周波数特性が高域に生成されることになる．そのため，場合によってはトーナル的な成分が高域に生成され，聴感上の自然性を低下させる可能性がある．

提案法では，こうした問題に対処するために，あらかじめ送信側において原音声の高域成分のパワーとトーナリティを求めておき，これらの補助情報に基づいて，より効果的な帯域拡張を試みている．

図 2 に提案法の処理手順を示す．提案法では，送信側において原音声の高域成分から 20 ms のフレームごとに補助情報の抽出を行っている．

提案法では，高域成分のパワーを表すパラメータとして，高域成分の平均振幅を用いている．原音声から HPF によって分離した 4 kHz 以上の高域成分を $s_h(n)$ とすると，$k(\geq 1)$ 番目のフレームにおける $s_h(n)$ の平均振幅 $M_s(k)$ は次のように定義される．

$$M_s(k) = \frac{1}{L}\sum_n |s_h(n)|,\ (k-1)L \leq n < kL \tag{2}$$

ここで，L はフレーム長を表す．

また，受信側において，従来法と同様の処理により生成された 4 kHz 以上の高域成分を $s_{brmh}(n)$ とすると，$k(\geq 1)$ 番目のフレームにおける $s_{brmh}(n)$ の平均

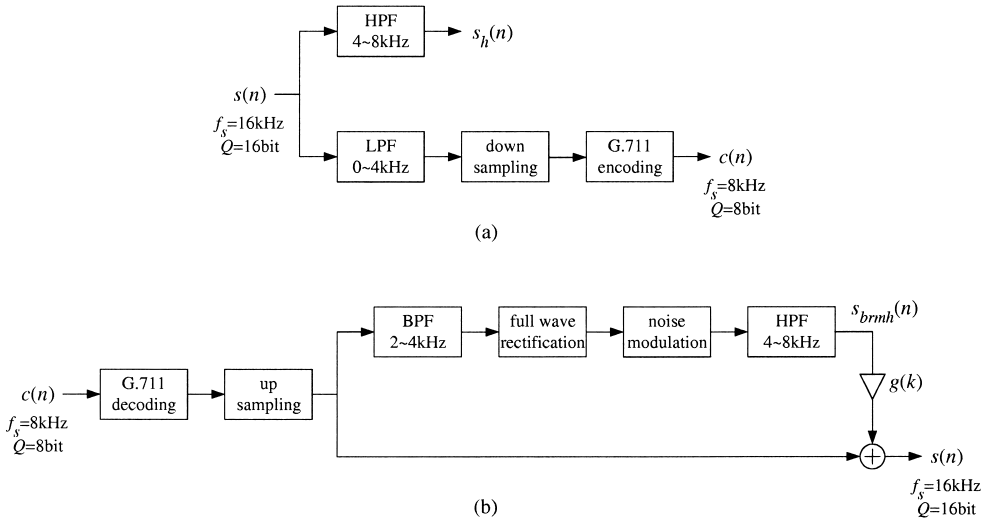

図 2　提案法の処理手順 (a) 送信側，(b) 受信側
Fig. 2　Procedure of the proposed technique at (a) sender and (b) receiver.

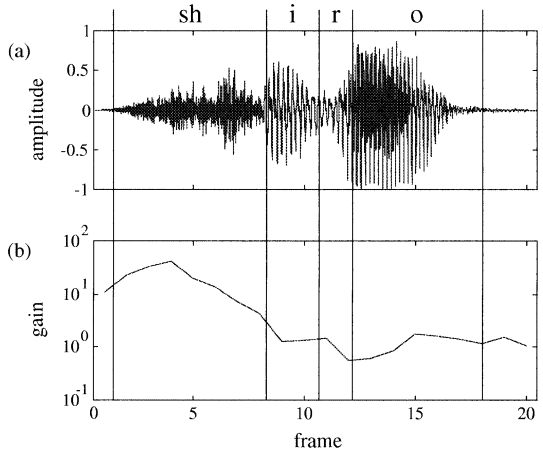

図 3　(a) 音声データ，(b) 高域成分のゲイン
Fig. 3　(a) Speech data and (b) gain of high band.

振幅 $M_r(k)$ は次のように定義される．

$$M_r(k) = \frac{1}{L}\sum_n |s_{brmh}(n)|,\ (k-1)L \le n < kL \tag{3}$$

したがって，高域成分のゲインは，フレームごとに次のように定義できる．

$$g(k) = M_s(k)/M_r(k) \tag{4}$$

図 3 に，[shiro] と発音された実際の音声データから計算された高域成分のゲインを示す．高域成分のゲインは，従来法では 0.5 に固定されているが，この図に示すように，本来は時変のパラメータであり，[sh] などの無声子音では，[i]，[r]，[o] などの有声音よりも大きくなる傾向を示す．

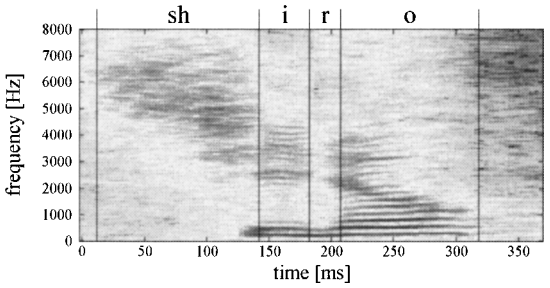

図 4　原音声 [shiro] のスペクトログラム
Fig. 4　Spectrogram of the original speech [shiro].

図 4 に，この音声データのスペクトログラムを示す．この音声データを G.711 で符号化した後，従来法のゲイン制御によって帯域拡張を行った場合のスペクトログラムを図 5 に示す．また，提案法のゲイン制御によって帯域拡張を行った場合のスペクトログラムを図 6 に示す．これらの結果から，従来法に比べて，提案法は [sh] などの無声子音における高域成分のパワーをより適切に制御できる可能があることが推察される．

提案法では，高域成分のトーナリティを表すパラメータとして，高域成分の正規化相互相関関数を用い

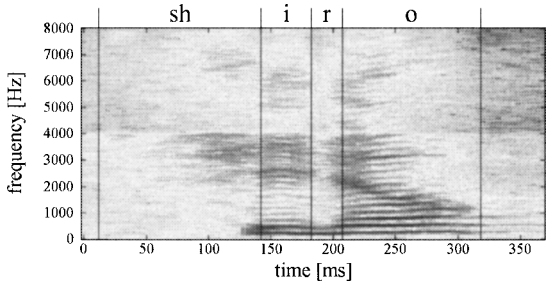

図 5　従来法のゲイン制御によって帯域拡張を行った音声 [shiro] のスペクトログラム
Fig. 5　Spectrogram of the band-extended speech [shiro] with the conventional gain control.

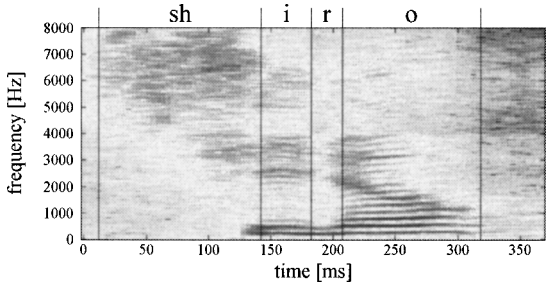

図 6　提案法のゲイン制御によって帯域拡張を行った音声 [shiro] のスペクトログラム
Fig. 6　Spectrogram of the band-extended speech [shiro] with the proposed gain control.

ている．原音声から HPF によって分離した 4 kHz 以上の高域成分を $s_h(n)$ とすると，$k(\geq 1)$ 番目のフレームにおける $s_h(n)$ の正規化相互相関関数 $r(m)$ は次のように定義される．

$$r(m) = \frac{\sum_n s_h(n)s_h(n+m)}{\sqrt{\sum_n s_h^2(n)}\sqrt{\sum_n s_h^2(n+m)}},$$
$$(k-1)L \leq n < kL \qquad (5)$$

提案法では，m がピッチ周期に等しくなるときの正規化相互相関関数の値を用いて，フレームごとに次のようにトーナリティを定義している．

$$t(k) = \max(r(p), 0) \qquad (6)$$

ここで，p はピッチ周期を表す．なお，ピッチ周期は低域成分の正規化相互相関関数から推定した値を用いている．

図 7 及び図 8 に，実際の音声における周波数特性と高域成分の例を示す．これらの図に示すように，高域

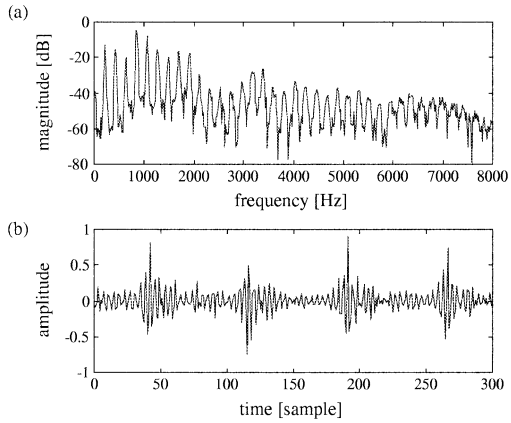

図 7　(a) 有声音の周波数特性，(b) 高域成分（トーナリティ 0.97）
Fig. 7　(a) Frequency characteristics of voiced speech and (b) its high band signal (tonality 0.97).

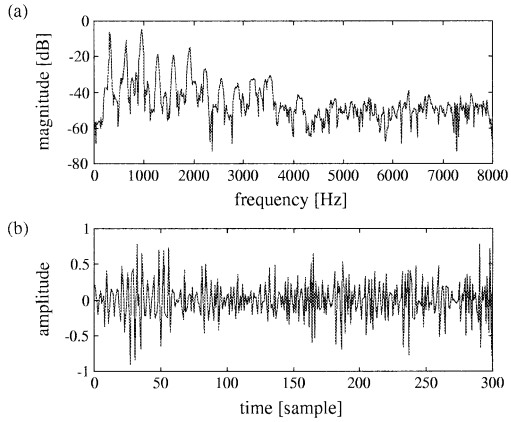

図 8　(a) 有声音の周波数特性，(b) 高域成分（トーナリティ 0.45）
Fig. 8　(a) Frequency characteristics of voiced speech and (b) its high band signal (tonality 0.45).

まで明確に調波構造が見られる場合は，高域成分における時間構造のエンベロープにピッチ成分の影響による周期性が見られる．また，トーナリティは大きくなる傾向を示す．一方，高域に明確な調波構造が見られない場合は，高域成分における時間構造のエンベロープに周期性が見られず，トーナリティは小さくなる傾向を示す．

提案法では，こうした特徴を考慮し，トーナリティを考慮した高域成分の生成を行っている．図 2 に示すように，提案法は，BPF によって 2 kHz から 4 kHz までの成分 $s_b(n)$ を分離し，これを全波整流することで $s_{br}(n)$ を生成する．提案法は，この信号をエンベ

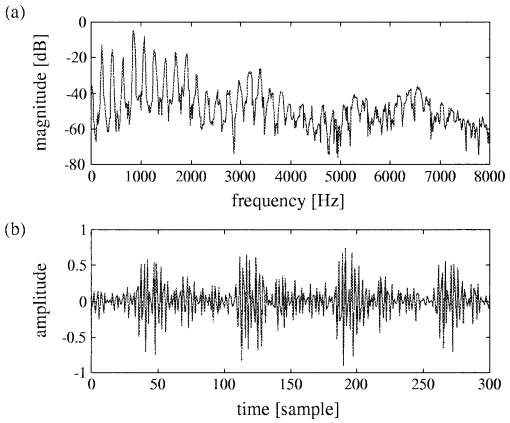

図 9　(a) 提案法によって帯域拡張を行った有声音の周波数特性，(b) 高域成分（トーナリティ 1）
Fig. 9　(a) Frequency characteristics of voiced speech extended by the proposed technique and (b) its high band signal (tonality 1).

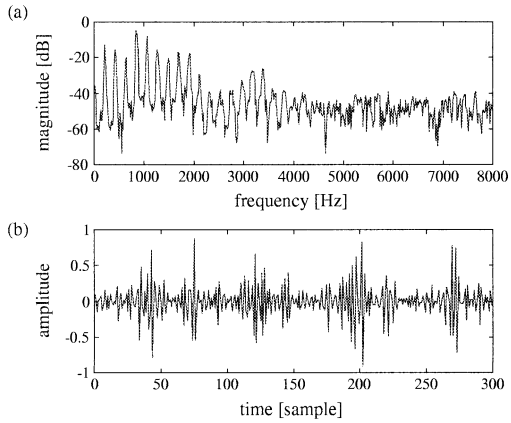

図 10　(a) 提案法によって帯域拡張を行った有声音の周波数特性，(b) 高域成分（トーナリティ 0）
Fig. 10　(a) Frequency characteristics of voiced speech extended by the proposed technique and (b) its high band signal (tonality 0).

ロープとし，トーナリティを考慮して混合した白色雑音 $w(n)$ を乗じることで，4 kHz 以上の高域成分を含んだ信号 $s_{brm}(n)$ を生成している．

$$s_{brm}(n) = s_{br}(n)(t(k) + (1 - t(k))w(n)),$$
$$(k-1)L \leq n < kL \tag{7}$$

この処理は，トーナリティが 1 のときは全波整流，トーナリティが 0 のときはノイズ変調に等しくなる [1]．

ノイズ変調は，音声データの帯域拡張の一手法として提案されたものである．この手法は，高域成分の知覚には，時間構造のエンベロープが重要な役割を担っており，適切なエンベロープによって変調したノイズによって，高域成分をモデル化できるという仮定に基づいている．

ただし，この手法はあくまでも高域成分をノイズによってモデル化しているため，高域まで明確に調波構造が見られ，トーナリティの大きい音声データの場合には，必ずしも適切な手法とはいえない [10]．提案法は，こうした特徴を考慮し，トーナリティによってノイズ変調の度合を制御できるようにしている．提案法は，トーナリティが大きいときは調波成分を生成する全波整流の効果を強調し，トーナリティが小さいときはノイズ変調の効果を強調する手法となっている．

図 9 及び図 10 に，トーナリティを変化させて生成した高域成分の例を示す．これらの図に示すように，提案法は，ノイズ変調の度合を制御することによって，トーナリティに応じた高域成分を生成することができる．

図 2 に示すように，提案法は，ノイズ変調によって得られた $s_{brm}(n)$ から HPF によって 4 kHz 以上の高域成分 $s_{brmh}(n)$ を分離し，これに対して式 (4) で定義されるゲインを乗じることによって重み付けを行う．その後，この信号を本来の 4 kHz 以下の低域成分 $s_l(n)$ に加算することで帯域拡張を実現する．すなわち，提案法は次のように定義される．

$$s(n) = s_l(n) + g(k) \cdot s_{brmh}(n),$$
$$(k-1)L \leq n < kL \tag{8}$$

提案法により実際に帯域拡張を行った音声データを試聴したところ，適切なゲイン制御によって高域成分のハイファイ感が向上すること，また，トーナリティを考慮し，ノイズ変調の度合を制御することによって，全波整流のみを適用した従来法よりも音色がソフトになる可能性があることが分かった．

4. ステガノグラフィによる補助情報の伝送

提案法で必要となるパラメータのうち，$M_r(k)$ は受信側で計算することができるが，$M_s(k)$ と $t(k)$ は送信側でのみ計算することができるため，これらのパラメータについては送信側で求めたものを受信側に伝送する必要がある．

これらの補助情報を伝送する手段として，提案法ではステガノグラフィと呼ばれる情報埋込技術を利用している．本研究では G.711 により音声データを符号化

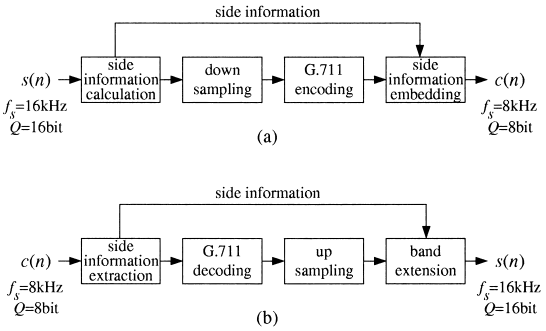

図 11 提案法の処理手順 (a) 送信側，(b) 受信側
Fig. 11 Procedure of the proposed technique at (a) sender and (b) receiver.

しているが，G.711 は冗長度の高いコーデックであることから，ステガノグラフィによる補助情報の埋込が容易に行えるという利点がある [11].

図 11 に示すように，提案法は，送信側で抽出した補助情報を G.711 によって符号化された音声データそのものに埋込伝送している．その際，提案法は，音声データの品質劣化を低減するために，音声サンプルの LSB（Least Significant Bit）に対して次のように補助情報を埋め込んでいる．

$$\mathrm{LSB}(c(n)) = \begin{cases} 0 & (b=0) \\ 1 & (b=1) \end{cases} \tag{9}$$

ここで，$\mathrm{LSB}(c(n))$ は音声サンプル $c(n)$ の LSB，b は 0 または 1 の値をとる 1 bit の補助情報である．このように，提案法における補助情報の埋込は，音声サンプルの LSB のビット値を補助情報のビット値に置き換える処理となっている．

なお，提案法では，[0, 32768] の値をとる $M_s(k)$ と，[0, 1] の値をとる $t(k)$ をそれぞれ線形量子化により 15 bit と 3 bit で符号化し，補助情報としている．すなわち，提案法における補助情報は 1 フレーム当り 18 bit となる．提案法では，この補助情報を 1 フレーム当り 18 個の音声サンプルに埋め込んでいる．

ステガノグラフィによる音声データの品質劣化をできる限り低減するために，提案法では，G.711 における対数量子化の特性を考慮した選択的埋込を行っている [11].

G.711 により符号化された音声サンプルは，LSB の重みが音声サンプルの振幅に依存し，音声サンプルの振幅の絶対値が大きいものほど LSB の重みが大きくなる特徴がある．そこで，提案法では，図 12 に示すように，ソートによってフレーム内の音声サンプルを並べ換え，振幅の絶対値が小さい音声サンプルから順番に補助情報の埋込を行っている．

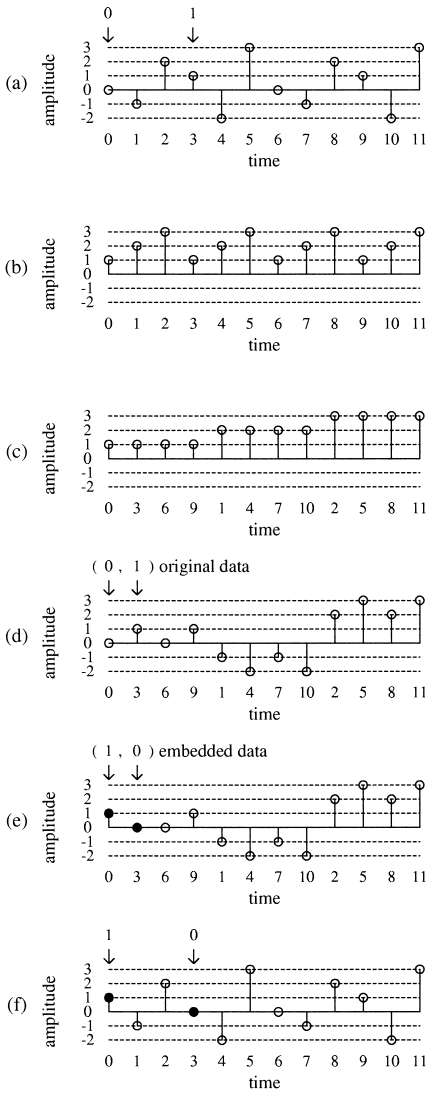

図 12 選択的埋込による 2 bit データ (1, 0) の埋込 (a) 原音声，(b) 振幅の絶対値，(c) ソート，(d) ソート後の音声データ，(e) 埋込，(f) 埋込後の音声データ
Fig. 12 Procedure of selective embedding method for 2 bit data (1, 0): (a) original speech, (b) absolute amplitude, (c) sorting, (d) sorted speech, (e) embedding, and (f) embedded speech.

こうした選択的埋込において，埋込を行った音声サンプルを正しく同定するために，提案法では，音声サンプルの振幅の絶対値については，次のように LSB

を無視したものを定義している．

$$|c(n)| = \begin{cases} 2\lfloor c(n)/2 \rfloor + 1 & (c(n) \geq 0) \\ -2\lfloor c(n)/2 \rfloor & (c(n) < 0) \end{cases} \quad (10)$$

ここで，$c(n)$ は G.711 により符号化された音声データである．また，$\lfloor x \rfloor$ は x を超えない最大の整数を表す．

図 13 は，選択的埋込とランダム埋込によってそれぞれ埋込を行った音声データの DSNR（Decoded-Signal to Noise Ratio）である．DSNR は，G.711 により復号化された音声データをリファレンスとして計算した SNR である．これは，符号化により発生する劣化の影響を除外し，埋込により発生する劣化のみを評価するための指標である．なお，ランダム埋込は，振幅の絶対値の大小にかかわらずランダムに選択した音声サンプルの LSB に対して情報を埋め込む手法である．

ここで，フレーム長については 20 ms にしているため，標本化周波数 8 kHz ではフレーム内の音声サンプルは合計 160 個となる．そのため，この図に示すように，1 フレームに埋め込める情報の最大値は 160 bit となる．

この図から示唆されるように，選択的埋込の方が，ランダム埋込よりも品質劣化が低減される可能性がある．特に，埋め込む情報が少ないときは選択的埋込が有利であり，提案法における合計 18 bit 程度の補助情報の埋込では，音声データの品質を極端に劣化させる可能性は少ないことが予想できる．

実際に，音声データベース [12] から無作為に選んだ 10 個の音声データに対して，提案法により補助情報の埋込を行った場合，SNR の平均は 37.78 dB となった．一方，補助情報の埋込を行わない場合は 37.85 dB となった．なお，ここでは，送信側における符号化前の音声データをリファレンスとして，符号化により発生する劣化も含め SNR を計算している．

この結果は，音声データにおける品質の劣化は，符号化により発生する劣化が圧倒的であり，補助情報の埋込により発生する劣化はほとんど無視できる程度であることを示唆している．

提案法は，ステガノグラフィを利用しているため，通話における見かけ上の伝送量を増加させる必要がないという特徴を有している．また，補助情報の埋込による劣化はほとんど無視できる程度であるため，提案法を実装していない一般の端末と通話する場合であっても，音声データを支障なく再生することができる．こうした互換性があることも提案法の特徴の一つと考えられる．

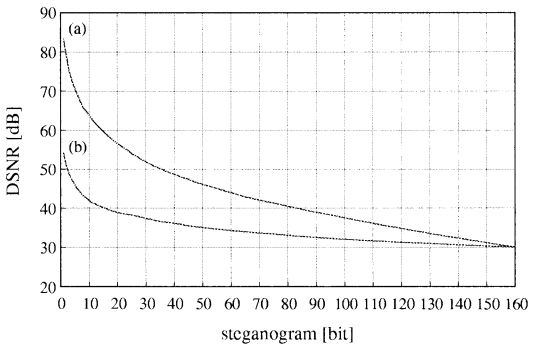

図 13　DSNR (a) 選択的埋込，(b) ランダム埋込
Fig. 13　DSNR: (a) selective embedding method and (b) random embedding method.

5. 評価実験

提案法の有効性を検証するため，CMOS（Comparison Mean Opinion Score）[13] による主観評価実験を行った．

実験には，音声データベース [12] から無作為に選んだ男性話者 5 個（m1〜m5），女性話者 5 個（f1〜f5）の音声データを用いた．また，実験には，音響心理実験の非専門家 10 名に参加してもらった．

それぞれの実験では，リファレンス音，刺激音 A，刺激音 B を，この順番で被験者に呈示した．ここで，リファレンス音は原音声，刺激音 A と刺激音 B は従来法または提案法で帯域拡張した音声データである．被験者には，表 1 の基準に従って，刺激音 A に対する刺激音 B の品質について評価を行ってもらった．

なお，刺激音 A と刺激音 B の組合せについては，呈示する順番を逆にした場合も含めて，同じ組合せを 2 回呈示した．したがって，それぞれの組合せは，10 名の被験者によって合計 20 回の評価を受けたことになる．

表 1　CMOS における評価点の定義
Table 1　Seven-point scale in CMOS.

point	quality
+3	much better
+2	better
+1	slightly better
0	about the same
−1	slightly worse
−2	worse
−3	much worse

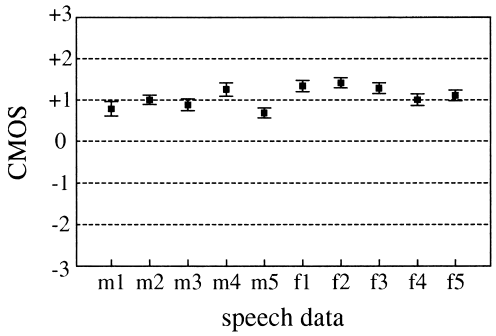

図 14　評価実験の結果（提案法 vs. 従来法）
Fig. 14　Experimental result of the subjective evaluation (proposed technique vs. conventional technique).

図 14 に，従来法に対する提案法の評価を示す．この図では，平均値とともに，平均値の 95%信頼区間も併せて示している．結果として，従来法に比べて提案法の方が，より原音声らしい品質を再現できる可能性があることが分かった．

6. むすび

評価実験の結果から，音声の高域成分における特徴を表すゲインやトーナリティといったパラメータを適切に制御することにより，従来法を改善できる可能性があることが示唆された．また，ステガノグラフィにより帯域拡張に必要な補助情報を伝送することで，通話における見かけ上の伝送量を増加させずに，付加価値の高い通信を実現できる可能性があることが示唆された．今後は，実際のシステムに提案法を適用し，その有効性について更に検討していくことを予定している．

謝辞　本研究の一部は平成 18 年度科学研究費（課題番号 18760263）により行われた．ここに謝意を表する．

文　献

[1] A. McCree, “A 14 kb/s wideband speech coder with a parametric highband model,” Proc. Int. Conf. Acoust., Speech, Signal Processing, pp.1153–1156, Istanbul, Turkey, 2000.
[2] R. Taori, R.J. Sluijter, and A.J. Gerrits, “Hi-BIN: An alternative approach to wideband speech coding,” Proc. Int. Conf. Acoust., Speech, Signal Processing, pp.1157–1160, Istanbul, Turkey, 2000.
[3] 青木直史，“ステガノグラフィを用いた VoIP における音声の広帯域化に関する一検討，” 信学技報，SP2003-72, 2003.
[4] N. Aoki, “A band extension technique for G.711 speech using steganography,” IEICE Trans. Commun., vol.E89-B, no.6, pp.1896–1898, June 2006.
[5] ITU-T G.711, “Pulse code modulation (PCM) of voice frequencies,” 1988.
[6] 千村保人，村田利文，SIP 教科書，IDG ジャパン，2003.
[7] R.M. Aarts, E. Larsen, and D. Schobben, “Improving perceived bass and reconstruction of high frequencies for band limited signals,” Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA2002), pp.59–71, Leuven, Belgium, 2002.
[8] U. Zölzer (ed), DAFX - Digital Audio Effects, John Wiley & Sons, 2002.
[9] A.V. McCree and T.P. Barnwell, III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE Trans. Speech Audio Process., vol.3, no.4, pp.242–250, 1995.
[10] Y. Qian and P. Kabal, “Classified highband excitation for bandwidth extension of telephony signals,” Proc. European Signal Processing Conf., 2005.
[11] 青木直史，“ピッチ波形複製法に基づくステガノグラフィを用いた VoIP におけるパケット損失の一隠蔽法，” 信学論（B），vol.J86-B, no.12, pp.2551–2560, Dec. 2003.
[12] ATR 音声翻訳通信研究所，研究用自然発話音声データベース，1997.
[13] ITU-T P.800, “Methods for subjective determination of transmission quality,” 1996.

（平成 18 年 12 月 1 日受付，19 年 2 月 26 日再受付）

青木　直史　（正員）

平 7 北大・工・電子卒．平 9 同大大学院修士課程了．平 12 同大大学院博士課程了．同年同大大学院工学研究科助手．平 19 同大大学院情報科学研究科助教．工博．平 10 カナダ・サスカチュワン大学・テレコミュニケーション研究所客員研究員．平 11〜12 日本学術振興会特別研究員．平 14 フィンランド・ヘルシンキ工科大学・音響研究所客員研究員．音声情報処理，画像情報処理，ヒューマンインタフェースに関する研究に従事．日本音響学会，信号処理学会各会員．