振幅スペクトログラム量子化と位相復元復号化に基づく音響信号符号化の検討*

☆藤田 卓 (東大・工), 中野 允裕, 小野 順貴, 嵯峨山 茂樹 (東大・情報理工)

1 はじめに

現在, 音声符号化には時間領域での線形予測を利用したソースフィルタモデルに基づいたアルゴリズムが活躍しており, 音響符号化においては周波数領域における人間の音響心理のモデルを陽に用いたアルゴリズムが主流である [1].

一方で, 音声音響信号の解析技術として時間周波数表現であるスペクトログラム領域において信号を扱う研究が近年目覚ましい発展を遂げている. 特に, 振幅やパワースペクトログラムに対する非負値行列分解 (Nonnegative matrix factorization: NMF) [2] や振幅スペクトログラムからの位相復元 [3] はスペクトログラムを用いた新しい音響信号の符号化方法の可能性を与えてくれると考えられる.

本稿では, このような信号処理手法を利用することで, 元の複素スペクトログラムから位相情報を捨てた振幅スペクトログラムに対して圧縮を行い, 復号には位相復元を用いるという符号化の構想を提案し, 基礎的な実験を通してその有用性を検討する.

2 スペクトログラム音響信号符号化

2.1 提案手法の概略

近年, 振幅もしくはパワースペクトログラムの大域的な性質に着目して, 全体を少ないパラメータで表現しようとする研究が盛んに行われており, それらの手法を用いることでスペクトログラムを用いた信号の圧縮の可能性が生まれてきたと考えることが出来る. 代表的な手法としては, NMF によるスペクトログラムの基底分解などが挙げられる. NMF はスペクトログラムを非負値行列 $\mathbf{Y}$ と見なし, $\mathbf{Y} \approx \mathbf{HU}$ のように 2 つの非負値行列 $\mathbf{H}$ と $\mathbf{U}$ の積に近似しパラメータを削減するものである. さらに, 位相情報を失った振幅スペクトログラムに対し, 適切な位相を付加し時間領域の信号に戻す高速な位相復元アルゴリズム [4] も提案されている.

そこで本稿では, 図 1 に示すように, 元信号の複素スペクトログラムから位相情報を捨てた振幅スペクトログラム領域での量子化と, 量子化されたスペクトログラムに対し適切な位相を付加し信号に戻す復号化による新しい符号化手法を提案する.

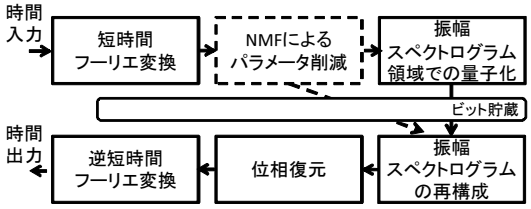

Fig. 1 スペクトログラムを用いた符号化の概略

2.2 振幅スペクトログラム量子化

符号化の観点からは, 聴感上の音質の劣化を抑えながら出来るだけ振幅スペクトログラムを圧縮することが求められている. これにはさまざまな方法が考えられ, 検討事項としては

1. スペクトログラムの振幅, パワー, それらの対数, もしくは Bark scale など, どの領域でどのように量子化するのが良いのか
2. NMF によるパラメータ削減において, どのような距離尺度を用いて近似を行うか

などを挙げることが出来る.

2. においては近年, 人間の音響心理モデルを考慮しスペクトログラムの近似を行う新しい距離尺度 [5] も提案されている. さらに分解されたパーツやモデル化誤差の圧縮しやすさの観点から新しい距離尺度が生まれる可能性もある.

本稿では NMF によるパラメータ削減を用いず, 単に対数振幅スペクトログラム $\mathbf{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$ を各周波数の平均 μ_ω と分散 σ_ω^2 によって n ビットに量子化する方法を試みた. 簡単のために, 量子化された各時間周波数の要素 $\hat{Y}_{\omega,t}$ はパラメータ α を用いて $\hat{Y}_{\omega,t} = (2^{n-1}/\alpha\sigma_\omega)\,(Y_{\omega,t} - \mu_\omega + \alpha\sigma_\omega)$ と与えた. ただし, $[\cdot]$ はガウス記号を表し, $\hat{Y}_{\omega,t} < 0$ のときは $\hat{Y}_{\omega,t} = 0$, $\hat{Y}_{\omega,t} > 2^n - 1$ のときは $\hat{Y}_{\omega,t} = 2^n - 1$ と与えた. α は各周波数において標準偏差の α 倍までを量子化の対象にしていることを意味している. 復号化の際には $Y_{\omega,t}$ はパラメータ α を用いて $Y_{\omega,t} = \mu_\omega - \alpha\sigma_\omega + (\alpha\sigma_\omega/2^{n-1})(\hat{Y}_{\omega,t} + 0.5)$ とした.

2.3 振幅スペクトログラムの位相復元による復号化

フレームのオーバーラップに起因する複素スペクトログラムの冗長さを利用し, 振幅スペクトログラムに対して適切な位相を反復的に推定し付加する手法

---

*Audio Coding based on Encoding magnitude spectrogram and Decoding with phase reconstruction by Suguru FUJITA, Masahiro NAKANO, Nobutaka ONO and Shigeki Sagayama (The University of Tokyo)

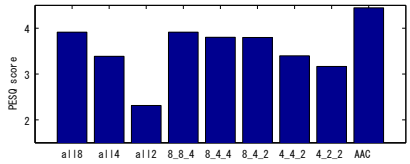

Fig. 2　音声に対する提案手法と AAC の PESQ 値

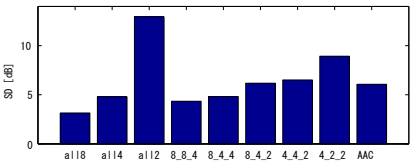

Fig. 3　音声に対する提案手法と AAC の SD

が提案されている [3]. これによって, 振幅スペクトログラム領域において量子化 (加工) された振幅スペクトログラムに適切な位相を付加し, 復号化, すなわち信号に戻すことが可能となる.

3　実験

実験に用いた信号はモノラル, 16000 kHz に統一し, スペクトログラム作成のためにフレーム長 32ms, フレームシフト 8ms, ハミング窓の短時間フーリエ変換を用いた.

提案法の有用性を確認するための基礎実験として, 単に振幅スペクトログラムを量子化するだけでどの程度の音質の劣化となるのかについて客観評価を行った. 2.2 の量子化において $\alpha = 3$ とし, 全ての帯域を 8 ビット, 4 ビット, 2 ビットで量子化した $all8, all4, all2$ と, 低域, 中域, 高域をそれぞれ左から順に数字のビットで量子化した $8\_8\_4, 8\_4\_4, 8\_4\_2, 4\_4\_2, 4\_2\_2$ と AAC-LC の 64 kbit/s の 9 つの符号化方法を評価した. 位相復元は [3] を用い, 全てランダムな初期値から 500 回の反復を行った. 信号は ATR database の A セットから無作為に選んだ 6 つを用いた. 音質に関する客観評価には ITU-T 勧告 P.862 PESQ (Perceptual Evaluation of Speech Quality) と SD (Spectral Distortion measure) を用い, 6 つの信号への平均値を図 2, 図 3 に示した. SD は参照信号と評価したい信号の振幅スペクトログラム $Y_{\omega,t}, \bar{Y}_{\omega,t}$ に対して, $SD^2 = (1/\Omega T)\sum_{\omega,t} 100(\log_{10} Y^2_{\omega,t} - \log_{10} \bar{Y}^2_{\omega,t})^2$ として算出した. 提案法の圧縮率についてはまだ議論出来る段階ではないが, 全帯域を 4 ビット量子化したスペクトログラムを zip 圧縮することで 80 kbit/s 相当になることを確認している. PESQ 値を見る限り, 単純な対数振幅スペクトログラムの符号化だけでは AAC に及ばないことが分かったが, SD のように別の指標においては勝る点もあり, 量子化方法について今後さらなる検討の余地が残されていると考えられる.

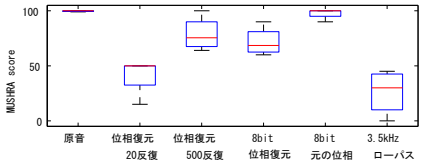

Fig. 4　位相復元による音質への影響 (クラシック)

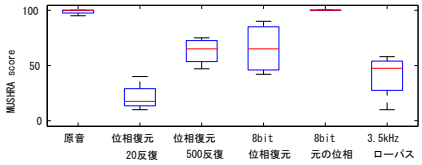

Fig. 5　位相復元による音質への影響 (ジャズ)

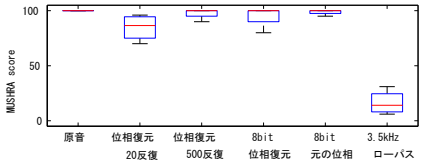

Fig. 6　位相復元による音質への影響 (ポップス)

4　おわりに

本稿では, 振幅スペクトログラム領域での量子化と位相復元による復号化を用いた音響信号の符号化の新しい枠組みを提案した. 今後は量子化手法の改善や NMF の導入を行い, 先行技術と性能比較を行っていきたい.

参考文献

[1] P. Motlicek *et al.*, EURASIP, pp. 1–14, 2010.
[2] P. Smaragdis and J. C. Brown, WASPAA, 2003.
[3] D. W. Griffin and J. S. Lim, IEEE Trans. ALSP, vol. 32, no. 2, pp. 236–243, 1984.
[4] J. Le Roux *et al.*, SAPA, 2008.
[5] J. Nikunen and T. Virtanen, ICASSP, 2010.
[6] M. Goto *et al.*, ISMIR, 2002.

付録

位相復元による音質の劣化の影響を調べるため, 小規模ではあるが MUSHRA 法を用いた評価結果を掲載する. アンカーには 3.5 kHz 帯域制限信号を用いた. 元の振幅に対し反復 20 回の位相復元だけ行ったもの, 反復 500 回の位相復元だけ行ったもの, 対数振幅を 2.2 の方法による $\alpha = 3$ として 8 ビット量子化し反復 500 回の位相復元を行ったもの, 8 ビット量子化し元の位相を付加したものの 4 つに対し評価を行った. RWC database [6] のクラシック, ジャズ, ポップスから 8 秒に切り出した信号を 1 つずつ用意し, 4 名の被験者にヘッドフォン装着にて評価を行ってもらい, その結果を図 4, 図 5, 図 6 に示した. ポップスは特に位相復元の影響が少ないと予想される.