振幅スペクトログラムに基づく波形合成音の音質評価と音響信号符号化への検討*

☆佐藤 匠 (東大院・情報理工), 鎌本 優 (NTT・CS 研), 小野 順貴, 嵯峨山 茂樹 (東大院・情報理工)

1 はじめに

近年，携帯型音楽プレーヤなどが身近となり，ディジタル化された音響データを利用する機会が増加している．これらのディジタル音響データは，情報量を削減するために信号の相関や人間の聴覚特性を用いるなどの圧縮を行った非可逆圧縮データとして取り扱われている．MP3 や AAC などで使われているこれらの方法は，既に多くの研究がされており，世界中で使われている．

我々は，さらなる高音質かつ高圧縮を目指すには，大きく異なる方針での圧縮が必要であると考え，短時間フーリエ変換 (STFT) の振幅スペクトログラムの符号化に基づく新たな符号化の枠組みを検討している [1]．この枠組みでは，人間の聴覚は位相情報に対して敏感ではないことに着目し，符号器では STFT の対数振幅スペクトログラムだけを符号化し，復号器では振幅スペクトログラムから位相復元 [2] を用いて推定された位相スペクトログラムを用いて波形を合成することによって情報を削減するものであり，振幅スペクトログラムの冗長性を利用することによって，さらに高い圧縮率の実現が期待できる．

本研究では基礎的検討として，符号化において対数振幅スペクトログラムに主成分分析 (PCA) を用いて次元削減を行った後に，位相復元を用いて復号化する手法を検討した．また，位相復元による影響と，符号化に用いる主成分に対する量子化条件の違いにより復元された音質の違いを主観評価実験により検証したので報告する．

2 振幅スペクトログラムに基づく音響信号符号化

2.1 対数振幅スペクトログラムへの PCA の適用

離散時間信号 x_n(n はサンプル点のインデックス) を STFT して得られる対数振幅スペクトログラムを $\mathbf{Y} = (\mathbf{y}_1, \cdots \mathbf{y}_K)^{\mathrm{T}} = (Y_{m,k})_{M\times K} \in \mathbb{R}^{\geq 0, M\times K}$ とする．なお，STFT のフレームのインデックスを m，周波数軸方向の座標を k とした．ここで，対数にしているのは，人間の耳が周波数を対数的に知覚するためである [3]．$\mathbf{Y}$ に対して周波数ビン間の相関から PCA による圧縮を考える．階数が R である $\mathbf{Y}$ を中心化したものを $\mathbf{Y}_{mean} = ((Y_{mean})_{m,k})_{M\times K} \in \mathbb{R}^{M\times K}$ とするときに，$\mathbf{Y}_{ave} = ((Y_{ave})_{m,k})_{M\times K} \in \mathbb{R}^{M\times K}$, $(Y_{ave})_{m,k} = E[\mathbf{y}_m]$ を考えると，以下を得る．

$$\mathbf{Y}_{mean} = \mathbf{Y} - \mathbf{Y}_{ave} \quad (1)$$

$$\mathbf{Y}_{mean}^{\mathrm{T}} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \quad (2)$$

ここで，$\mathbf{I}$ を単位行列とすると，$\mathbf{U} = (U_{k,r})_{K\times R} \in \mathbb{R}^{K\times R}, \mathbf{V}^{\mathrm{T}} = (V_{r,m})_{R\times M} \in \mathbb{R}^{R\times M}$ は正規直交ベクトルを列ベクトルにもつ行列 ($\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$), $\Sigma \in \mathbb{R}^{R\times R}$ は特異値 $\lambda_1, \cdots, \lambda_R (\lambda_1 \geq \cdots \geq \lambda_R > 0)$ を対角要素にもつ対角行列である．$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_R), \mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_R)$ とおくと，$\mathbf{u}_r, \mathbf{v}_r$ は特異ベクトルである．ここで，$\mathbf{A} = (A_{k,r})_{K\times R} = \mathbf{U}\Sigma = (\mathbf{a}_1, \cdots, \mathbf{a}_R) \in \mathbb{R}^{K\times R}$ とおくと，以下の関係を得る．

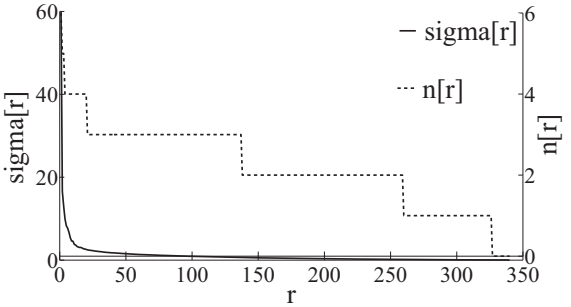

Fig. 1 σ_r と n_r の対応関係

$$\mathbf{Y}_{mean}^{\mathrm{T}} = \mathbf{A}\mathbf{V}^{\mathrm{T}} \quad (3)$$

$\mathbf{V}^{\mathrm{T}}$ は，コードブックとして符号器，復号器の双方で保持しておくことが考えられるので，今回は量子化の対象とせず，本研究では $\mathbf{A}$ を量子化することを考える．まず，$\mathbf{a}_r$ の平均 μ_r と標準偏差 σ_r を求める．次に，$\mathbf{a}_r$ の要素を $\pm\alpha\sigma_r$ の範囲で n_r ビットで量子化することを考える．$\mathbf{a}_r$ の要素は一様分布であることを仮定すると，量子化誤差のパワー $P_e = (\alpha\sigma_r^2/2^{n_r})/3 = \{\alpha^2/(3\cdot 4^{n_r})\}\sigma_r^2$ が求められる．ここで量子化誤差を一定以下に抑える条件：$P_e < P_{\max}$ を考えると，$n_r = [\log_2(C\sigma_r)]$ が必要なビット数となる．ただし，

$$C = \alpha(3P_{\max})^2 \quad (4)$$

である．$\mathbf{a}_r$ を $0\sim 2^{n_r}-1$ で等間隔量子化すると，

$$\hat{A}_{k,r} = [(2^{n_r-1}/\alpha\sigma_r)(A_{k,r} - \mu_r + \alpha\sigma_r)] \quad (5)$$

を得る．ただし、n_r と式 (5) の $[\cdot]$ はガウス記号を表わす．$A_{k,r} < 0$ のときは,$A_{k,r} = 0$, $A_{k,r} > 2^{n_r} - 1$ のときは $A_{k,r} = 2^{n_r} - 1$ と与えた．ここでの 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意味している．図 1 は，ポップス (mono, sampling freq.:16kHz, duration: 10sec., RWC-MDB-G-2001-No.1) を用いて生成された対数振幅スペクトログラムに PCA を行って得られた σ_r と n_r の関係を表わすグラフである．r が大きくなるにつれて，量子化ビット数が小さくなっていくのが分かる．

2.2 位相復元を用いた復号化

式 (5) を復号するには，式 (6) を用いる．

$$A_{k,r} = \mu_r - \alpha\sigma_r + (\alpha\sigma_r/2^{n_r-1})(\hat{A}_{k,r} + 0.5) \quad (6)$$

これより，式 (1) 〜 (3) から対数振幅スペクトログラムが復号化できる．この対数振幅スペクトログラムに，位相復元の技術 [2] を用いることで位相スペクトログラムを求め，元の音響信号を復元することができる．近年では，高速での位相復元が研究がされているために本研究の実用化の際には有用である [4].

---

*Sound Quality Assessment of Synthesized Wave Based on Magnitude Spectrogram and Its Application to Audio Coding. by Sho SATO (The University of Tokyo), Yutaka KAMAMOTO (NTT Communication Science Laboratories), Nobutaka ONO and Shigeki SAGAYAMA (The University of Tokyo)

Table 1　位相復元の影響についての MUSHRA 法による評価

(a) ポップス

	平均	最高	最低
A	100	100	100
B	27	47	0
C	23	30	14
D	50	70	34
E	48	60	35
F	41	55	28

(b) ジャズ

	平均	最高	最低
A	100	100	100
B	51	69	36
C	8	13	0
D	56	79	32
E	61	90	44
F	47	64	17

(d) クラシック

	平均	最高	最低
A	100	100	100
B	38	71	20
C	36	69	5
D	61	82	50
E	55	81	40
F	47	76	30

Table 2　対数振幅スペクトログラムに PCA を用いた符号化についての MUSHRA 法による評価

(a) ポップス

	BitRate	平均	最高	最低
a	256	53	76	21
b	※ 32	23	26	20
c	※ 64	44	51	38
d	※ 96	38	40	35
e	64	72	85	53
f	63	55	61	37
g	126	70	74	57

(b) ジャズ

	BitRate	平均	最高	最低
a	256	50	76	38
b	※ 32	46	54	41
c	※ 64	70	80	59
d	※ 96	55	57	51
e	64	80	88	77
f	63	63	74	49
g	126	61	79	55

(c) クラシック

	BitRate	平均	最高	最低
a	256	42	62	14
b	※ 32	52	59	50
c	※ 64	83	90	77
d	※ 96	90	91	88
e	64	93	95	90
f	63	63	66	52
g	126	75	81	64

Table 3　実験条件

	量子化ビット	シフト長
A	-	-
B	-	-
C	2	64
D	4	128
E	8	256
F	16	512

3　振幅スペクトログラムの量子化と位相復元による波形合成音の音質主観評価実験

3.1　実験条件

対数振幅スペクトログラムの量子化と位相復元の音質への影響を調べるために，小規模（被験者 4 人）な主観評価実験を行った．STFT によって得られた対数振幅スペクトログラムのみを符号化し，位相復元 [2] により位相スペクトログラムを推定した後に，時間領域の信号へ復元させた音質を調べる．本実験では，信号として 3 曲 (RWC-MDB-G からポップス，ジャズ，クラシックを 1 曲ずつ選択．それぞれ mono, sampling freq.: 16 kHz，duration: 10 sec.) を用い，ビットレートを 256 kbps，フレーム長を 1024 点 に固定して量子化ビット，シフト長 (オーバーラップ率) を変えた場合 (表 3) の検証をした．なお，STFT にはハミング窓を用い $\alpha = 3$，位相復元の反復演算は 200 回とした．評価の手法として，MUSHRA 法 [5] を用いた．この方法は，比較したい音 (C〜F) と元信号 (A, 16 kHz, 16 bit)，アンカー (B, sampling freq.:8 kHz の帯域制限信号，16 bit) を用意し，被験者は元信号とアンカーを含めてランダムに提示される音をヘッドフォンで聴き，元信号と比較した音質を，0〜100 の連続値で評価するものである．

3.2　実験結果

結果を表 1 に示す．どのジャンルでも，C のように 2 bit で量子化するのは粗すぎることと，F のようにオーバーラップが小さいと位相復元が精度よくできていないことが分かる．しかし，D と E は似たような結果が得られている．したがって，曲ごとに適切な量子化ビット，シフト長を選択する必要があることがわかる．また，本実験においては平均値の最高が 61 点であり，音質の劣化が無視できないことがわかる．

4　PCA を用いた量子化に基づく主観評価実験

4.1　実験条件

本実験では，PCA を用いた量子化の有用性を確認する．信号は 3.1 と同じ 3 つの楽曲を用いた．対数振幅スペクトログラム作成のためには，フレーム長 1024 点, フレームシフト 512 点, ハミング窓の STFT を用いた．対数振幅スペクトログラムを PCA して得られた $\mathbf{A}$ に対して，式 (4) の定数 C を調節してビットレートが 32 〜 96 kbps (b〜d) となるように符号化した場合の 3 通り，PCA をせずに全てを 4 bit (f)，8 bit (g) で符号化した場合の 2 通り，さらに 64 kbps の AAC (e) を加えた計 8 通りの符号化方法を MUSHRA 法を用いて評価した．位相復元と MUSHRA 法の条件は 3.1 と同様であり，a はアンカーの 8 kHz 帯域制限信号である．被験者についても同様に 4 人である．表 2 の※は $\mathbf{V}$ は符号化せず，$\mathbf{A}$ のみを符号化したときのビットレートを表わしている．

4.2　実験結果

表 2 のジャズとクラシックにおいて，ビットレートが近い c と f を比べたときに PCA を用いた c の平均点が高くなっており，PCA による情報圧縮の効果が確認できる．本実験では，$\mathbf{V}$ は曲毎に求めたが，実際には符号器側，復号器側で共通のコードブックとして持つことを想定している．コードブックの個数，学習方法および量子化精度についても，今後，検討していく予定である．

5　おわりに

本研究では，対数振幅スペクトログラムのみを量子化・符号化したものを復号化した場合における位相復元の音質への影響と，PCA を用いた量子化・符号化について検討する主観評価実験を行った．今後は，元の位相情報を補助情報として符号化することによる音質劣化の軽減，PCA 以外の次元削減法の適用，聴覚心理モデルを用いたビット割り当ての導入，などを検討していきたい．

参考文献

[1] 藤田 他，音講論 (秋)，pp. 1391-1392，2010.
[2] D. W. Griffin and J. S. Lim, IEEE Trans. ALSP, vol. 32, no. 2, pp. 236-243, 1984.
[3] M.Bosi and R.E.Goldberg, "Introduction to digital audio coding and standards," Kluwer Academic Publisher, 2003
[4] J. Le Roux et al., SAPA, 2008.
[5] ITU-R Rec. B. 1534-1