WaveRNNを利用した音声ロスレス符号化に関する検討と考察*

☆天田将太（筑波大）, 杉浦亮介, 鎌本優, 原田登, 守谷健弘（NTT）, 山田武志, 牧野昭二（筑波大)

1 はじめに

ネットワークやデジタル機器のブロードバンド化が進展するとともに、音声信号の高品質化（高サンプリングレート、高振幅分解能、多チャネル化）の要望が高まっており、蓄積や配信に必要な情報量が飛躍的に増大しつつある。信号の品質を全く損うことなく情報量を削減することができるロスレス符号化は、高品質が求められる近年において重要な役割を担っている。

普及しつつあるロスレス符号化方式として、動画や音声データを扱う国際標準化団体である MPEG (Moving Picture Experts Group) により MPEG-4 ALS(Audio Lossless Coding) [1–3] が規格化されている。MPEG-4 ALS はサンプリング周波数 192 kHz、量子化ビット数 32 bit にまで対応するなど、幅広い音声音響信号を扱うことができる。

ロスレス符号化では元の信号が完全に復元されるため、その性能指標としては圧縮率や演算量が注目される。MPEG-4 ALS では線形予測分析を用いて符号化を行なっており、残差信号を伝送することでロスレス符号化を行なっている。予測性能が良ければ残差信号の振幅は小さくなり、圧縮性能は高くなる。つまり、線形予測器の性能は圧縮性能に大きく影響する。ロスレス符号化における線形予測器を改善するための研究 [4–6] がこれまでなされてきたが、非線形予測器を用いることで線形に限らない柔軟な予測を行い、予測性能を向上させられる余地があると考えられる。

また近年、音声合成用に提案された Deep Neural Network(DNN) ベースの手法である WaveNet [7] を非線形予測器として利用し符号化を試みる検討がなされている [8]。上記で利用されている WaveNet は 8 bit 信号までの対応であることに加え、ネットワーク構造が大きく演算に時間がかかってしまう問題がある。この問題を解決する WaveRNN という手法が提案されている [9]。

WaveRNN は 16 bit/24 kHz の信号に対応した手法であり、前の時間のサンプル値を入力として次の時間のサンプル値の予測確率分布を出力するようにネットワークを学習する。エントロピーを最小化するように学習を行うため、符号化との相性が良いと考えられる。

本稿では、WaveRNN を非線形予測器として利用した音声のロスレス符号化を提案する。また提案法の性能を調査し、MPEG-4 ALS の後方適応予測モードと比較して考察する。

2 MPEG-4 ALS

MPEG-4 ALS において普及している Simple Profile では PARCOR 係数と線形予測残差信号を伝送し、元の信号を完全復元するという演算量の軽い方式が用いられている。一方、演算量が大きいため、これまで普及していないが、予測係数を伝送しない後方適応線形予測モードも定義されている [10]。提案手法は、MPEG-4 ALS の中でも後方適応線形予測モードと類似性が高いため、比較対象として用いることとする。

MPEG-4 ALS の後方適応予測モードのエンコーダのブロック図を Fig. 1 に示す。入力信号はそれぞれのフレームについて RLS-LMS 予測器により予測信号を取得し、入力信号と予測信号から予測残差信号を計算する。予測残差信号はエントロピー符号化され、圧縮信号として伝送される。

MPEG-4 ALS の後方適応予測モードのデコーダのブロック図を Fig. 2 に示す。伝送されてきた圧縮信号をエントロピー復号化し予測残差信号を得る。予測残差信号と RLS-LMS 予測器による予測信号から、元の信号を完全に復元する。

3 WaveRNN

WaveRNN のネットワーク構造を Fig.3 に示す。WaveRNN は 1 層の RNN(Recurrent Neural Network) 層と 2 層の NN(Neural Network) 層により構成されている。入力は予測するサンプル値の 1 つ前のサンプル値であり、粗い値 (coarse) と細かい値 (fine) に変換して利用される。coarse は 16 bit で量子化されたサンプル値の上位 (MSB)8 bit であり、fine は下位 (LSB)8 bit である。Fig.3 における c は coarse、f は fine を表している。出力は次のサンプル値における coarse、fine それぞれの出現確率である。

式 (1) は、前の時間 $t-1$ における coarse (c_{t-1}) と fine (f_{t-1}) から、次の時間 t における coarse (c_t) と fine (f_t) の出現確率 $P(c_t)$ と $P(f_t)$ を求める際の全体の流れを示している。

$$
\begin{aligned}
x_t &= [c_{t-1}, f_{t-1}, c_t] \\
u_t &= \sigma(R_u h_{t-1} + I_u^\star x_t) \\
r_t &= \sigma(R_r h_{t-1} + I_r^\star x_t) \\
e_t &= \tau(r_t \circ (R_e h_{t-1}) + I_e^\star x_t) \\
h_t &= u_t \circ h_{t-1} + (1 - u_t) \circ e_t \\
y_c, y_f &= split(h_t) \\
P(c_t) &= softmax(O_2\ relu(O_1 y_c)) \\
P(f_t) &= softmax(O_4\ relu(O_3 y_f))
\end{aligned}
\tag{1}
$$

ここで R_u, R_r, R_e は RNN 層の前の隠れ値に再帰重みをかけた値であり、それぞれ u_t, r_t, e_t の 3 つのために計算される。σ および τ は標準的な活性化関数であるシグモイド関数とハイパボリックタンジェント関数である。$split()$ は層を分割する関数を意味しており、coarse のための層と fine のための層に分割する働きがある。また、式 (1) における $\star$ はマスクされた

*Experimental evaluation of the audio lossless coding with WaveRNN. by Shota Amada (University of Tsukuba), Ryosuke Sugiura, Yutaka Kamamoto, Noboru Harada, Moriya Takehiro (NTT), Takeshi Yamada and Shoji Makino (University of Tsukuba).

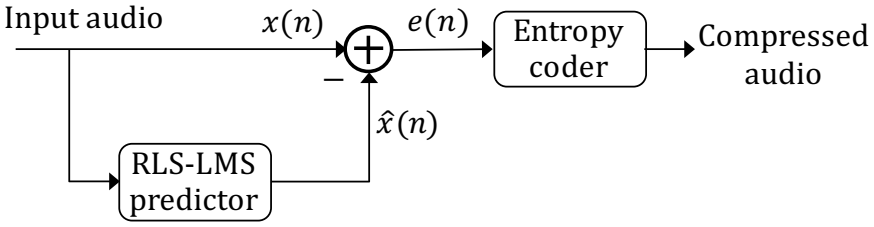

Fig. 1　MPEG-4 ALS Encoder.

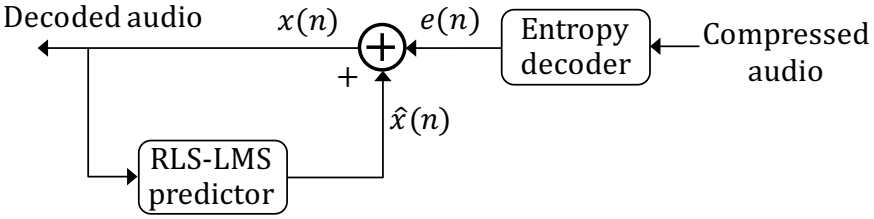

Fig. 2　MPEG-4 ALS Decoder.

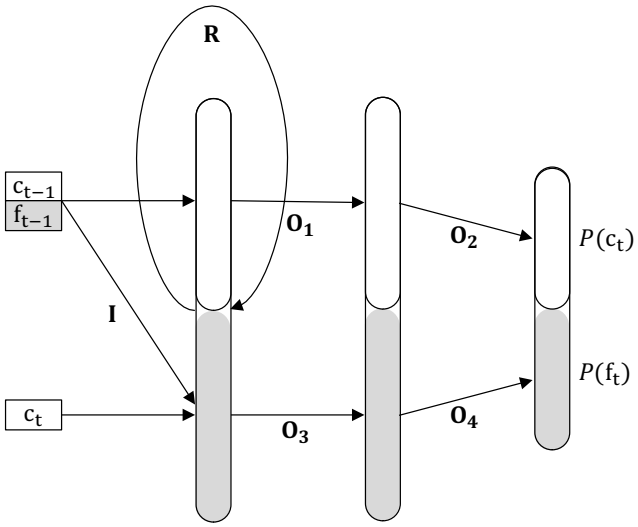

Fig. 3　WaveRNN のネットワーク構造

Table 1　Specifications of sound items.

言語数	8ヶ国
サンプリング周波数	16 kHz
量子化ビット数	16 bit
チャネル数	1 ch
1 音源の録音時間	3〜4 秒程度
全音源数	17,000

Table 2　Number of units in each network.

	入力ユニット数	出力ユニット数
RNN 層	896	2688
NN 層 O_1, O_3	488	488
NN 層 O_2, O_4	488	256

行列を表しており、これにより c_t は f_t の計算にしか利用されない。

学習時には真値の確率分布と予測確率分布の交差エントロピーを最小化するように重みの更新を行う。つまり、予測確率分布における真値の情報量を最小化する学習となる。

4　WaveRNN を利用したロスレス符号化

本稿で検討する WaveRNN を利用したロスレス符号化手法について説明する。本手法のエンコーダおよびデコーダのブロック図を Fig. 4, 5 に示す。Fig. 1, 2 における RLS-LMS 予測器を WaveRNN に置き換えるような形となる。信号の予測は WaveRNN 部のみで行うことができるため、予測のために伝送する信号は残差信号のみで良い。

エンコーダの WaveRNN において予測確率最大のサンプル値と真のサンプル値との差 (残差) を計算しておき、この残差をデコーダの WaveRNN に伝送すれば、デコーダにおける予測値を真値と同じ値に完全に復元できる。ここで伝送する残差は算術符号を用いれば予測確率における真値の確率によるエントロピーから理論上最適な符号を表現でき、学習時の交差エントロピーに近いビットレートを得られることが期待できる。

残差の計算と信号の復元を行う流れを Fig. 6, 7 に示す。デコード時において真値と同じ値に復元できれば、次の時間の予測にもエンコード時に用いた真値と同じ値を利用することができる。これを繰り返すことで、信号全体を完全に復元することができる。

5　実験結果

5.1　実験条件

本実験では、Multilingual Speech Database 2002 [11] の音声データを用いた。実験に用いた音源の情報を Table 1 に示す。本実験ではこの中から一部の音源のみを用いた。音源の前後の無発話区間を切り取って利用したため、各音源の録音時間にばらつきが生まれている。また、実験に用いた WaveRNN の構造情報を Table 2 に示す。

5.2　性能評価基準

本実験における圧縮性能の評価基準について説明する。本実験では、音源のエンコード時の平均符号語長により性能を評価する。平均符号語長とは信号 1 サン

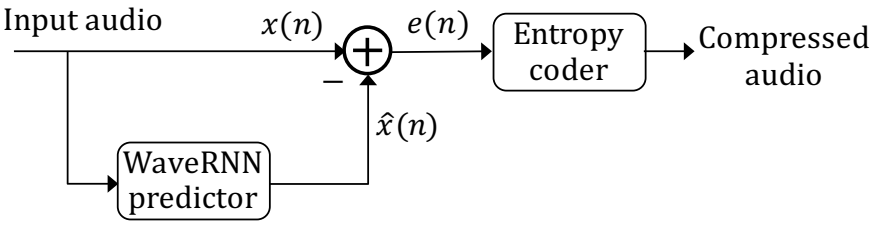

Fig. 4　Lossless Encoder with WaveRNN.

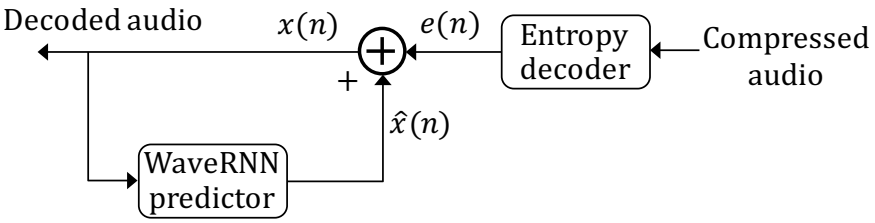

Fig. 5　Lossless Decoder with WaveRNN.

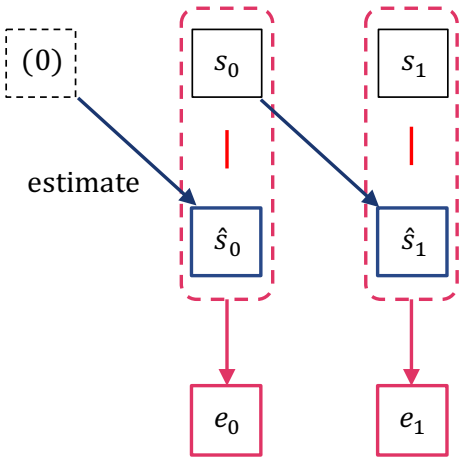

Fig. 6　Calculation of residuals with WaveRNN.

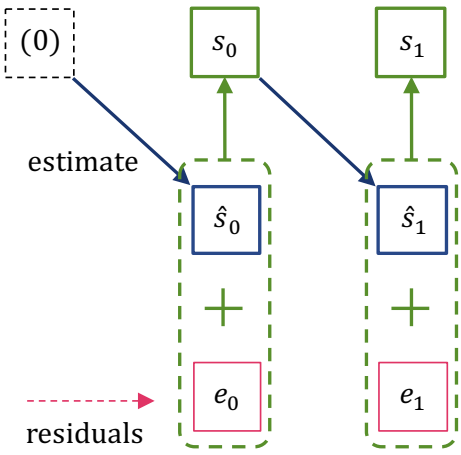

Fig. 7　Lossless reconstruction with WaveRNN.

プルあたりの平均交差エントロピーを示す値であり、圧縮をした場合の理論上のビットレート [bit/sample] を示す。平均符号語長 H を式 (2) に示す。ここで N はサンプル数、$p_i(x_i)$ は i 番目のサンプル値の予測確率分布 p_i における真値 x_i の出現確率である。

$$H = \frac{1}{N}\sum_{i=1}^{N} -\log_2 p_i(x_i) \tag{2}$$

また比較対象として、MPEG-4 ALS を用いて Multilingual Speech Database 2002 の米国男女 20 話者による 200 音源を圧縮をした場合の平均符号語長を Table 3 に示す。各フレーム長ごとに平均符号語長を検証したところ、フレーム長 512 サンプルの場合の 8.15 bit/sample が最高性能となった。

5.3　1 音源の学習

1 話者による 1 音源のみを学習データとして学習を行い、その学習用音源および学習に利用していない評価用音源をエンコードした場合の平均符号語長を Table 4 に示す。合計における () 内の数字は、MPEG-4 ALS における平均符号語長との差分である。

学習データをエンコードした場合 (「学習 : 1 話者 1 音源」) については、coarse, fine ともに良く予測できており、圧縮率は MPEG-4 ALS よりはるかに良い

Table 3　Mean code length compressed with MPEG-4 ALS

フレーム長	平均符号語長 [bit/sample]
256	8.18
512	**8.15**
1024	8.19
2048	8.29
4096	8.46
8192	8.69

性能であることがわかる。一方、学習に利用していない音源をエンコードした場合 (「評価 : 1 話者 1 音源」) ではうまくいっていないことが確認できる。学習データをエンコードした結果は他の入力信号に対する性能を犠牲にした結果であり、幅広い入力信号に対応できるようにするためには学習データ数を増加させる等の工夫をする必要がある。

5.4　同一話者音源の学習

米国女性 1 名の 150 音源を学習データ (「学習 : 1 話者 150 音源」) として学習を行なった。同女性の学習に使っていない 150 音源を「評価 : 1 話者 150 音源」、同女性を含めた米国男女各 10 名の計 20 名による 20 音源を「評価 : 20 話者 20 音源」として評価用音源を

Table 4　Mean code length of learning with only one sound source. The number in parentheses is the difference from MPEG-4 ALS.

	合計 [bit/sample]	coarse	fine
学習：1 話者 1 音源	1.295 (-6.85)	0.055	1.24
評価：1 話者 1 音源	43.272 (+35.12)	5.109	38.163

Table 5　Mean code length of learning with one person. The number in parentheses is the difference of mean code length from MPEG-4 ALS.

	合計 [bit/sample]	coarse	fine
学習：1 話者 150 音源	7.18 (-0.97)	0.61	6.57
評価：1 話者 150 音源	7.39 (-0.76)	0.69	6.69
評価：20 話者 20 音源	8.11 (-0.04)	1.17	6.94

用意し評価を行なった。結果を Table 5 に示す。

「評価：1 話者 150 音源」に対しては MPEG-4 ALS より 0.76 bit/sample 優れた性能となっていることが確認できる。一方、学習に用いた女性以外の人物の音源を含めた「評価：20 話者 20 音源」に対しては前者と比較して性能が悪い。さらに性能を向上させるためには、学習データ数や学習データの範囲 (話者の人数等) および学習回数を増やす必要があると考えられる。

5.5　複数話者音源の学習

米国男女各 10 名の計 20 名について、それぞれ 5 音源ずつ計 100 音源を学習データ (「学習：20 話者 100 音源」) として学習を行い、第 5.4 小節と同様の評価用音源を用意して評価を行なった。学習に用いた 20 名は、「評価：20 話者 20 音源」で利用した 20 名と共通である。結果を Table 6 に示す。

「評価：20 話者 20 音源」に対する性能は第 5.4 小節の結果から 0.49 bit/sample 程度改善し、MPEG-4 ALS より 0.53 bit/sample 優れていることが確認できる。対して、「評価：1 話者 150 音源」に対する性能は第 5.4 小節の結果より 0.34 bit/sample 悪化している。

学習データ数や学習データの範囲および学習回数を増やすことで、幅広い音源に対してさらなる性能向上を期待できる。一方で、第 5.4 小節のように学習データの範囲を絞れば、その範囲においては高い性能を実現できる。

6　おわりに

WaveRNN により符号化を行なった場合の平均符合語長の評価を 3 つの条件で行なった。1 音源のみの学習による評価では、MPEG-4 ALS を大きく超える圧縮率を実現できる可能性を確認した。同一話者音源、複数話者音源の学習による評価では、MPEG-4 ALS より 0.53〜0.76 bit/sample 程度優れた圧縮性能となっていることを確認した。MPEG-4 ALS による圧縮では平均符号語長 8.15 bit/sample であったところ、提案法により 7.62 bit/sample に改善した。

学習データ数や学習データの範囲および学習回数をさらに増加させることで、より幅広い音源に対しての性能向上が期待できる。一方で、学習データの範囲を絞ることによりその範囲においては高い性能を実現できる。適応範囲と圧縮性能のトレードオフの管理が課題となる。

Table 6　Mean code length of learning with 20 person. The number in parentheses is the difference of mean code length from MPEG-4 ALS.

	合計 [bit/sample]	coarse	fine
学習：20 話者 100 音源	7.18 (-0.97)	0.71	6.48
評価：1 話者 150 音源	7.73 (-0.41)	0.87	6.86
評価：20 話者 20 音源	7.62 (-0.53)	0.81	6.78

今後の展望として、学習データ数や学習データの範囲および学習回数の変更を検討する。

参考文献

[1] ISO/IEC 14496-3:2009/Amd 3:2015, 2009 Information technology – Coding of audio-visual objects – Part 3: Audio

[2] T. Liebechen,*et al.* "The MPEG-4 Audio Lossless Coding (ALS) Standard - Technology and Applications," in Proc. 119th AES Convention, Oct, 2005.

[3] T. Liebchen and Y. Reznik, "MPEG-4 ALS : an emerging standard for lossless audio coding," IEEE, Data Compression Conference, pp. 439-448, 2004.

[4] 天田将太, 他" 音響ロスレス符号化 MPEG-4 ALS のハイレゾ音源適応の検討と考察," 音講論, Mar, 2017.

[5] 天田将太, 他" 音響ロスレス符号化 MPEG-4 ALS におけるハイレゾ音源向け線形予測次数最適化に関する検討と考察," 音講論, Sep, 2017.

[6] Shota Amada, *et al.* "Experimental evaluation of encoding parameters of MPEG-4 ALS for high-resolution audio",In Proc. GCCE, Oct, 2017

[7] Aaron van den Oord,*et al.* "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO", arXiv:1609.03499 2016

[8] W. Bastiaan Kleijn,*et al.* "WAVENET BASED LOW RATE SPEECH CODING", arXiv:1712.01120, 2017

[9] Nal Kalchbrenner,*et al.* "Efficient Neural Audio Synthesis", arXiv:1802.08435, 2018

[10] Haibin Huang,*et al.* "Cascaded RLS-LMS Prediction in MPEG-4 Lossless Audio Coding", IEEE T AUDIO SPEECH,VOL. 16, NO. 3, pp. 554 - 562, Mar, 2008

[11] Multilingual Speech Database 2002, `http://www.ntt-at.com/product/speech2002/`