一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2017-145,SIP2017-154,SP2017-128 (2018-03)

高品質音声符号化のための スペクトル包絡・非周期性指標量子化の知覚的影響

宮下　玄太†　　森勢　将雅††

† 山梨大学大学院医工農学総合教育部　〒 400–8511 山梨県甲府市武田 4–3–11
†† 山梨大学大学院総合研究部　〒 400–8511 山梨県甲府市武田 4–3–11
E-mail: †{g17tk022,mmorise}@yamanashi.ac.jp

あらまし　本稿では，フルバンド音声をボコーダを用いて分析合成する際の，音声パラメータの量子化に伴う音質劣化について検討した．高品質な音声と，そのパラメータを用いる主観評価実験がある中で，音声パラメータのデータ量が，波形と比べて多いことが問題となっている．我々はフルバンド音声用の符号化のアプローチとして，音声分析により取得した，音声パラメータの符号化を行っている．本稿では，音声分析によって得られるパラメータを量子化した際に，どの程度の細かさなら高品質さを保ち，聴覚に与えないかを，スペクトル包絡と非周期性指標について，主観評価実験により検証する．実験の結果，元のパラメータ (64 bit) と音質の差異が見られなかったのは，それぞれスペクトルは 14 bit, 非周期性指標は 3 bit であった．今回の符号化は，実音声を対象とした場合，非現実的な範囲で量子化を行っている．これは，最悪の条件として，この bit 数表現に置き換えることに問題がないことを示唆する．この結果から，スペクトル包絡は 22%，非周期性指標は 5% 程度の表現に置き換えることが可能である．
キーワード　音声分析合成, 量子化, 音質, スペクトル包絡, 非周期性指標

Perceptual influence of spectral envelope and aperiodicity quantization for encoding high-quality speech

Genta MIYASHITA† and Masanori MORISE††

† Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi 4–3–11　Takeda, Kofu, Yamanashi, 400–8511 Japan
†† Integrated Graduate School, University of Yamanashi　4–3–11　Takeda, Kofu, Yamanashi, 400–8511 Japan
E-mail: †{g17tk022,mmorise}@yamanashi.ac.jp

Abstract　In this paper, we investigate the relationship between the degradation of sound quality and the parameter quantization in analysis/synthesis of full–band speech using vocoder. We have verified the coding by speech parameters as an approach for high–quality speech coding. A subjective evaluation by MUSHRA was carried out to verify the threshold that the listener cannot perceive the degradation. Experimental results showed that the thresholds of spectral envelope and aperiodicity were 14 and 3 bit, respectively. Since the proposed quantization was performed with a considerably wide range, the obtained thresholds are useful in all kinds of speech to synthesize speech without degradation.
Key words　speech analysis/synthesis, quantization, sound quality, spectral envelope, aperiodicity

1. はじめに

音声符号化は，古くから研究されている内容の 1 つである．これは主に，波形符号化と分析合成による符号化の 2 つに分類される．本研究では，音声分析合成で得られた音声パラメータの符号化に焦点を当て，本稿ではパラメータの量子化について実験した結果について述べる．

音声分析合成を用いた符号化は，主に電話音質程度の，16 kHz 以下のサンプリング周波数を対象とした技術として研究されている [1]．近年ではフルバンド音声用の符号化についても検

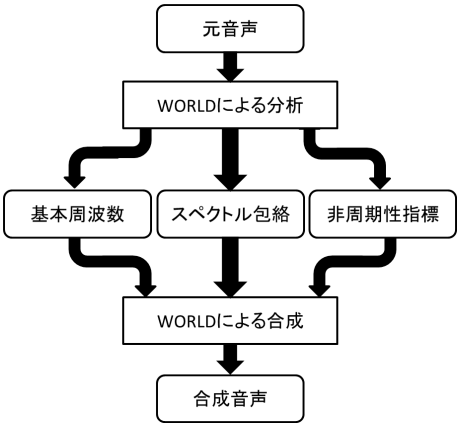

図 1　音声分析合成の概要.

Fig. 1　Overview of speech analysis/synthesis system.

討が進みつつある [2]．また，高品質音声の分析パラメータを用いた研究も存在する [3]．しかし，高品質な音声の音声パラメータを符号化した際の，音質劣化と符号化度合いとの関係性を明らかにしたものは少ない．

本稿では，サンプリング周波数が 40 kHz を超えるフルバンド音声を対象とした，音声符号化技術の開発を目標とする．また，各音声パラメータを符号化し，制御する際に音質の劣化しない境界値の指標を作成する．前報 [4], [5] ではそれぞれ，フルバンド音声の周波数軸方向，時間軸方向の符号化について検討を行った．本稿は，分析された音声パラメータについての量子化と，量子化ビット数による音質劣化について検討を行う．

2. 実 験 手 法

本稿では，主観評価実験により音声の評価を行い，聴覚的な音質の劣化度合いを確認する．主観評価実験に用いる音声として，フルバンド音声の音源を用意し，それらをボコーダを用いて分析する．次に，分析されたパラメータを量子化し，音声の再合成を行う．合成された音声の聴取比較を行う．音声分析合成に用いるボコーダには，高品質音声合成系 WORLD [6] (D4C edition [7]) を使用した．

2.1 音 声 分 析

WORLD による音声分析合成の概要を図 1 に示す．WORLD では，音声をフレームシフト幅毎の時間で分析し，フレーム毎に 3 つのパラメータを取得する．パラメータは，基本周波数 (Fundamental frequency: F0)，スペクトル包絡 (Spectral envelope: SP)，非周期性指標 (Aperiodicity: AP) の 3 種類である．これらのパラメータは，それぞれ音声の高さ，音声の音色，音声のかすれの程度に対応している．各パラメータは，それぞれ 1 次元，1025 次元，1025 次元であり，64 bit の倍精度浮動小数点型で構成されている．これらのパラメータの量子化ビット数を変更し，音質に与える影響を調査する．本稿では，スペクトル包絡と非周期性指標について量子化を行った．

2.2 音声パラメータ量子化

分析により得られたパラメータの量子化を行う．まず，スペクトル包絡についての概要を図 2 に示す．各パラメータの量子化を行う際に必要なパラメータは，量子化範囲と量子化ビット数である．

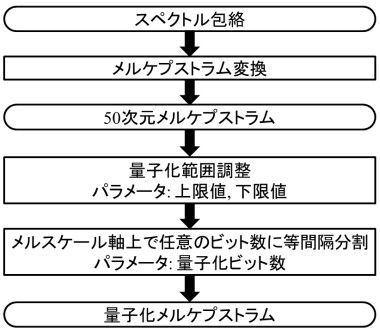

図 2　スペクトル包絡量子化の概要.

Fig. 2　Overview of speech parameter quantization for spectral envelope.

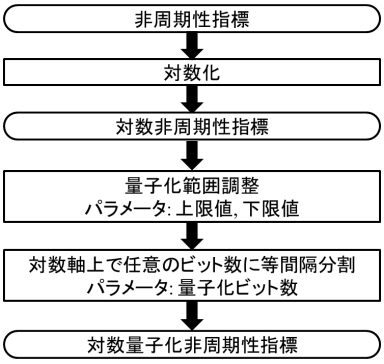

図 3　非周期性指標量子化の概要.

Fig. 3　Overview of speech parameter quantization for aperiodicity.

まず，スペクトル包絡を前報 [4] に基づき，50 次元のメルケプストラムに圧縮する．次に，各周波数のもつスペクトルのパワーの上限値から下限値を設定する．実音声の場合，振幅が取りうる範囲をある程度予測することは可能であるが，本稿ではあらゆる音声に対応するため，float 型の全範囲をカバーできる値の近似値とした．設定したスペクトル包絡の上限値は対数軸上で 88 , 下限値は-88 である．次に，メルスケール軸上の下限値から上限値の間を任意のビット数で等間隔に分割する．これによりパラメータを範囲内でのビット数表現に置き代えることができる．予備検討を実施し，その結果から量子化ビット数は 11，12，13，14 とした．最後に，指数化を行い元のパラメータの形式に戻す．

非周期性指標についての量子化の概要を図 3 に示す．非周期性指標は，文献 [7] より 5 次元に圧縮可能なことが示されている．まず，この 5 次元の非周期性指標を対数に変換する．次に，非周期性指標の上限値から下限値を設定する．非周期性指標は 0〜1 の範囲で表されるため，対数の上限は 0 dB となる．下限は −60 dB とした．そして，対数軸上の下限値から上限値の間を任意のビット数で等間隔に分割する．量子化ビット数は 2, 3，4，5 である．

表 1　実 験 条 件
Table 1　The experimental conditions

評価手法	
比較方法	MUSHRA 法
被験者	20 代男性 10 名
実験環境	防音室 (A–weighted SPL: 18 dB)
オーディオインターフェース	Roland QUAD–CAPTURE
ヘッドフォン	SENNHEISER HD650
使用音声	親密度別単語了解度試験用音声 データセット 2007(FW07) [8]
評価用音声	
発話者	4 人 (男女各 2 名)
A/D 変換	48 kHz/16 bit
音源数	全 20 音声 (各発話者 5 音声)
音声種	4 モーラ単語
比較条件	
比較用音声	分析合成音 5 種
リファレンス音声	非量子化分析合成音 (64 bit)
SP 量子化ビット数	14, 13, 12, 11 ms
SP 量子化範囲	−88〜88
AP 量子化ビット数	5，4，3，2 ms
AP 量子化範囲	−60〜0 dB

3. 実 験 条 件

主観評価の実験条件を表 1 に示す．実験の評価方法として MUSHRA 法 (Method for the subjective assessment of intermediate quality levels of coding systems) を用いた．MUSHRA 法とは，提示された音声の品質を評価する方法の 1 つで，被験者は GUI を用いて 0 から 100 の尺度で音声刺激を採点する．この手法は，一般的に音質の違いの評価に使用されていて，MOS 評価よりも差の検出力が高い手法と言われている．

実験は，暗騒音の A–weighted SPL が 18 dB の防音室を使用し，正常な聴力を有する 10 人が評価に参加した．音声刺激は，ヘッドホン (SENNHEISER HD650) を用いて与えた．主観評価に用いた音声刺激は，2 人の男性と 2 人の女性による各 5 発話の，計 20 発話である．サンプリング周波数は 48 kHz，量子化ビット 16 bit である．発話内容は，日本語による子音を含む 4 モーラ単語であり，親密度別単語了解度試験用音声データセット 2007 (FW07) に収録された音声を使用した [8]．被験者には，1 音源につき 1 種類のリファレンス音声と 4 種類の比較対象となる音声を同時に提示する．被験者は，まずリファレンス音声を聴取し，次にリファレンス音声を含む 5 種類の比較音声との比較を行う．被験者は，同時に提示された音声については任意に切り替えて聴取することができる．比較の結果，リファレンス音声よりも音質が下がる場合には，どの程度下がっているかを評点とする．音質が元音声と判別が出来ない場合，これを 100 点として評点を付ける．また，比較音声には必ずリファレンス音声が含まれるため，被験者は，必ず 1 つ以上の音声に 100 点を付ける必要がある．これを全 20 音源が終わるまで繰り返す．使用音声の順番は，音源，比較用音声共にランダム化されている．この評価により，被験者にリファレンス音声と比較音声との音質の差異を検出させる．比較音声 5 種類は，WORLD を用いて分析合成された量子化ビット数の異なる 5 個の音声刺激である．この実験を，スペクトル包絡と非周期性指標それぞれについて行う．

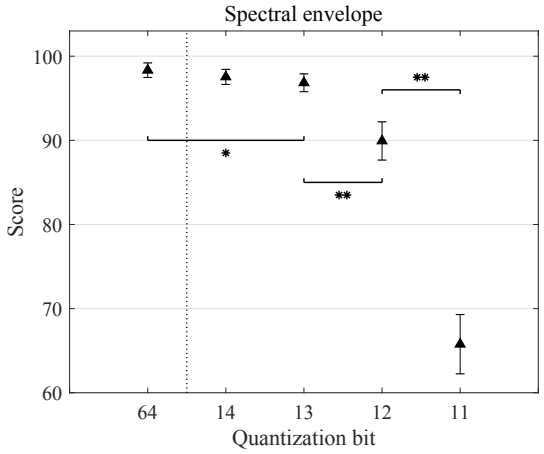

図 4　MUSHRA 法による，スペクトル包絡の量子化による音質劣化度合い評価結果．*は補正 p 値 < 0.05 を示す．**は補正 p 値 < 0.01 を示す．
Fig. 4　Sound quality evaluation results for quantization of spectral envelope by the MUSHRA method. Symbol * represent adjusted–p values lower 0.05. Symbol ** represent adjusted–p values lower 0.01.

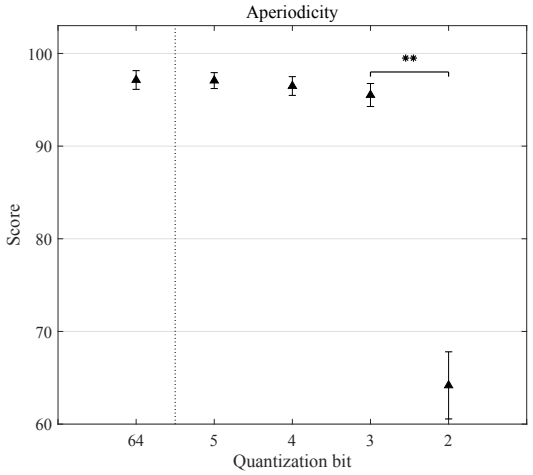

図 5　MUSHRA 法による，非周期性指標の量子化による音質劣化度合い評価結果．*は補正 p 値 < 0.05 を示す．**は補正 p 値 < 0.01 を示す．
Fig. 5　Sound quality evaluation results for quantization of aperiodicity by the MUSHRA method. Symbol * represent adjusted–p values lower 0.05. Symbol * represent adjusted–p values lower 0.01.

4. 実 験 結 果

実験結果を図 4,5 に示す．縦軸は MUSHRA 法による評点を，横軸は各パラメータに対応したフレームシフト幅を表す．

表 2　スペクトル包絡の量子化ビット数による音質劣化について，検証する組み合わせとその補正 p 値．補正 p 値が 0.05 を超える場合は n.s. (non significant) と表す．$p < 0.001$ は 0.001 以下と表す．

Table 2　List of adjusted–p values for spectral envelope.

組み合わせ	補正 p 値
64 bit，14 bit	n.s.
64 bit，13 bit	0.03
64 bit，12 bit	0.001 以下
64 bit，11 bit	0.001 以下
14 bit，13 bit	n.s.
14 bit，12 bit	0.001 以下
14 bit，11 bit	0.001 以下
13 bit，12 bit	0.001 以下
13 bit，11 bit	0.001 以下
12 bit，11 bit	0.001 以下

表 3　非周期性指標の量子化ビット数による音質劣化について，検証する組み合わせとその補正 p 値．補正 p 値が 0.05 を超える場合は n.s. (non significant) と表す．$p < 0.001$ は 0.001 以下と表す．

Table 3　List of adjusted–p values for aperiodicity.

組み合わせ	補正 p 値
64 bit，5 bit	n.s.
64 bit，4 bit	n.s.
64 bit，3 bit	n.s.
64 bit，2 bit	n.s.
5 bit，4 bit	n.s.
5 bit，3 bit	n.s.
5 bit，2 bit	n.s.
4 bit，3 bit	n.s.
4 bit，2 bit	n.s.
3 bit，2 bit	0.001 以下

誤差棒は 95%信頼区間を示す．まず，実験結果について統計分析を行う．複数の比較が必要であるため，Benjamini–Hochberg 法 [9] に基づく 2 段階線形上昇手順 [10] を実施した．

比較リストとその補正 p 値を表 2,3 に示す．Non significant (n.s.) は，音声が同じ音質を持っていることを保証するものではなく，MUSHRA 法の検出力においてその差が検出されなかったことを意味する．表 2 から，スペクトル包絡では 64, 14bit 間には有意差が無かった．64–13 bit 間に補正 p 値 0.03 の有意差があり，64–12, 13–12, 12–11 bit 間には補正 p 値 0.001 以下の有意差が検出された．表 3 から，非周期性指標では 64, 5, 4, 3 bit 間にそれぞれ有意差が無く，3–2bit 間には補正 p 値 0.001 以下の有意差が検出された．

5. 考　　察

本稿の条件で音声パラメータの量子化を行った場合，音質が劣化しない境界値は，スペクトル包絡で 14 bit, 非周期性指標で 3 bit となった．スペクトル包絡は，分析パラメータの中で最も次元数が多く，データ量に関係するパラメータである．このことから，最も符号化した際に影響が出やすいパラメータであると考えられる．しかし，データ量が大きいため，スペクトル包絡をより効率的に符号化することで，他のパラメータよりも効果的に音声全体の符号化を行うことができる．本稿ではスペクトル包絡の量子化範囲について，float 型の全範囲を用いた．しかし，通常発話におけるスペクトル包絡のパワーが取りうる範囲を分析し，設定することによって，より効率的に符号化ができると考えられる．

6. 終 わ り に

本稿では，フルバンド音声の分析合成における，パラメータの量子化度合と音質の関係について検証を行った．量子化はスペクトル包絡と非周期性指標に対してそれぞれ行い，量子化されたパラメータを用いた音声を作成した．主観評価実験を行い，各パラメータにおいて音質が劣化しない量子化度合いを策定した．その結果として，スペクトル包絡は 14 bit，非周期性指標は 3 bit に境界値が存在することを確認した．

今後の発展としては，前報と合わせたフルバンド音声の符号化形式を作成する．また，相互作用を確認するための実験を行い，音質を保つために必要な音声パラメータのデータ量を策定する．

a）謝　　辞

本研究は，JSPS 科研費 JP15H02726, JP16H05899, JP16K12511, JP16H01734 の支援を受けて実施された．

文　　献

[1] 北村正, 今井聖, 古市千枝子, 小林隆夫, “メルケプストラムを利用する音声の分析合成系と合成音声の品質,” 電子情報通信学会論文誌，vol.J68–A, vo.11, pp.957–964, 1985.

[2] 鎌本優, 杉浦亮介, 守谷健弘, 古角康一, 野口賢一, 兼清知之, “情報帯域制限下条件付無歪音声符号化（ＣＬＥＡＲ）の概要と音質評価,” 日本音響学会秋季研究発表会, pp.491–492, 2017.

[3] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King and P. Alku, “GlottDNN – A full-band glottal vocoder for statistical parametric speech synthesis,” in Proc. IEEE, pp.2473–2477, 2016.

[4] 宮下玄太, 森勢将雅, 小澤賢司, “フルバンド音声を対象とした音声分析合成システムに用いるスペクトル包絡の音質劣化のない低次元表現,” 情報処理学会音楽情報科学研究会, vol.2017–MUS–115, no.23, pp.1–6, 2017.

[5] 宮下玄太, 森勢将雅, “高品質音声分析合成による各パラメータのフレームシフト幅が音質に与える影響,” 電子情報通信学会技術報告書, vol.117, no.393, SP2017–72, pp.35-38, 2018.

[6] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol.E99-D, no.7, pp.1877–1884, 2016.

[7] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol.84, pp.57–65, 2016.

[8] T. Kondo, S. Amano, S. Sakamoto, and. Y. Suzuki, “Development of familiarity-controled word-lists (fw07),” IEICE Society Conference research report, vol.107, no.432, pp.43–48, 2008.

[9] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” J.R.Statist. Soc. Series B, vol.57, no.1, pp.289–300, 1995.

[10] Y. Benjamini, A. M. Krieger and D. Yekutieli, “Adaptive linear step-up procedures that control the false discovery rate,” Biometrika, vol.93, no.3, pp.491–507, 2006.