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
SP2017-72 (2018-01)

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

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　高品質音声をボコーダを用いて分析合成を行う際に音質の劣化が発生する．筆者らは，これまでの研究により，この音質の劣化と分析時のフレームシフト幅との関係について主観評価実験を行い，その結果から，音声分析の際に最適なフレームシフト幅が音声パラメータ毎に異なる可能性を明らかにした．本稿では基本周波数とスペクトル包絡に着目し，各音声パラメータの分析時のフレームシフト幅が音質に及ぼす影響について MUSHRA 法を用いた主観評価を行う．実験の結果，基本周波数パラメータは 20 ms まで有意な音質劣化が検出されなかった．一方で，スペクトル包絡パラメータは 5 ms で音質の劣化が検出された．この結果から，基本周波数パラメータの持つ時間軸方向のデータ量は，従来のものより少ない情報に符号化しても音質に影響がないことが確認された．
キーワード　音声分析合成, フレームシフト, 音質, 基本周波数, スペクトル包絡

Influence of frame shift in speech parameters on sound quality by high–quality speech analysis/synthesis system

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　Sound quality deterioration occurs when analyzing and synthesizing high–quality speech by using a vocoder. We conduct a subjective evaluation on the influence of the frame shift used for estimating each speech parameter on the sound quality. As a result, deterioration of the sound quality was not detected in cases where the frame shift of 20 ms is used in F0. On the other hand, deterioration of the sound quality was detected when using the frame shift of 5 ms in the spectral envelope.
Key words　speech analysis/synthesis, frame shift, sound quality, F0, spectral envelope

1. はじめに

音声符号化は音声の分析合成という分野において，古くから研究されている内容の一つである．ただし，これらの音声分析合成を用いた符号化は，主に電話音質程度の低い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．近年では高品質音声分析合成用の符号化についても検討が進みつつある [1]．

本研究では，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．前報 [2] 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，時間軸方向の符号化についての発展とし，より詳細な検討を行う．

音声分析合成において，音声のパラメータ化を行う際には，音声のパラメータ推定を実施する時間間隔が粗くなるほど，分析合成音声の音質が劣化する．本稿では，この間隔をフレームシフト幅とする．高品質音声合成システム STRAIGHT [3] では 1 ms，WORLD [4] では 5 ms がデフォルト値として設定されている．このフレームシフト幅と音質の関係について前報では検証を行った [2]．その結果として，5 ms 程度で 1 ms の分析合成音と有意差のない音質が保てることが確認された．これ

らの音声分析合成システムでは全ての音声パラメータについて同一のフレームシフト幅で分析を行っている．本稿では音声の各パラメータについて別のフレームシフト幅が適している可能性に着目して検証を行った．

本稿では，フルバンド音声の分析合成において，フレームシフト幅が音声の品質に与える影響を各音声パラメータについて検証する．また，各音声パラメータの符号化を行った際に，聴覚的に劣化のない音声合成をすることの出来る，最適なフレームシフト幅を策定する．

2. 関 連 研 究

GlottDNN [5] ボコーダはフルバンド音声に対応したボコーダである．GlottDNN において，DNN の学習にはフレーム毎に分析を行った音声パラメータが使用されており，1 フレームあたり 111 次元のパラメータを用いている．DNN に基づく音響特徴抽出・音響モデルを用いた統計的音声合成システムの構築 [6] では，パワースペクトルを用いた音響特徴量の抽出と，得られた特徴量からのモデル化を行っている．この実験ではスペクトル包絡を 59 次元に符号化していた．前報 [7] では，スペクトル包絡をメルケプストラムを用いて符号化した際の音質評価を行った．この実験では，メルケプストラム 40 次元にて，1025 次元の分析合成音との音質の有意差がないという結果が得られた．

これらの実験はどれも周波数方向の符号化については検討されているが，時間軸方向の符号化についてはされていない．本稿は，フルバンド音声の分析合成に必要なフレームシフト幅を求めることを目的としている．必要なフレームシフト幅について詳細な見解があれば，これらの手法もより効率的な学習が行える可能性がある．

3. 実 験 内 容

フルバンド音声である音源を用意し，それらをボコーダを用いてパラメータ化した後に再合成する．様々なパラメータを用いて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 [4](D4C edition [8]) を使用した．WORLD による音声分析合成の概要を図 1 に示す．音声分析合成においてパラメータ化する際に，フレームシフト幅の設定を行う．

3.1 音 声 分 析

WORLD を用いて音声の分析を行う．WORLD では音声をフレームシフト幅毎の時間に分割し，フレーム毎に 3 つのパラメータを取得する．パラメータはそれぞれ，基本周波数 (F0: Fundamental frequency)，スペクトル包絡 (SP: Spectral envelope)，非周期性指標 (AP: Aperiodicity) である．基本周波数は音声の高さ，スペクトル包絡は音声の音色，非周期性指標は音声のかすれに対応する．WORLD はデフォルトのフレームシフトパラメータを設定することにより，音声分析時のフレームシフト幅を変更することが出来る．前報では，このフレームシフトパラメータの変化による音質の差異を検証した．本稿では，フレームシフト幅を各パラメータにおいて別々の値を設定し，分析合成を行う．

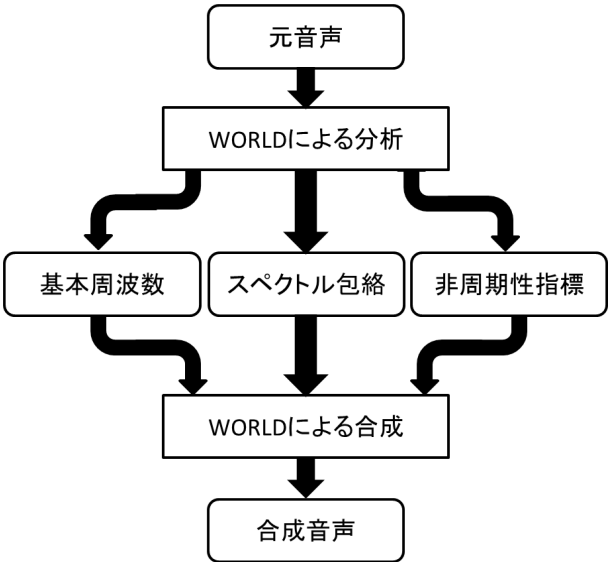

図 1　音声分析合成の概要．

Fig. 1　Overview of speech analysis/synthesis system.

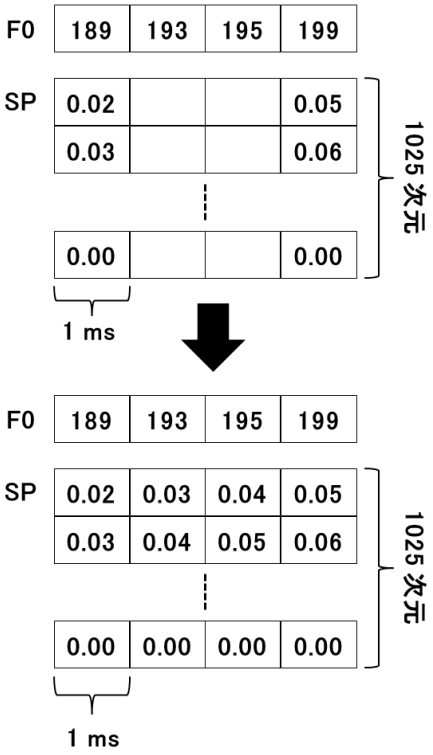

図 2　音声パラメータの補間の概要．

Fig. 2　Overview of interpolation of speech parameters.

3.2 パラメータ別のフレームシフト設定

音声の分析合成では，フレーム毎に音声パラメータを推定する必要がある．しかし，WORLD のプログラムの仕様上，異なるフレームシフト幅を持っている音声パラメータではそのまま合成することが出来ない．そのため，異なるフレームシフト幅をもつ音声パラメータの合成を行うにあたり，パラメータ調整が必要となる．本稿では，パラメータ間の線形補間を行うことで実現した．

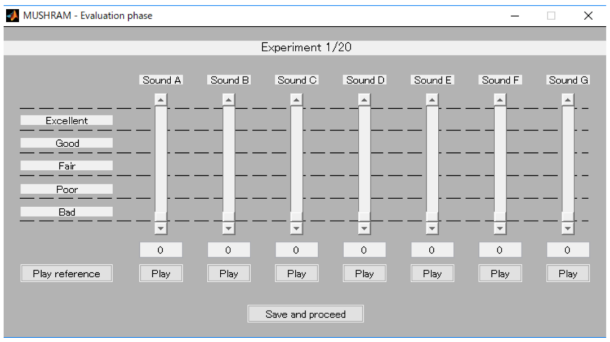

図 3　MUSHRA 法に用いる GUI.
Fig. 3　GUI used for the MUSHRA evaluation.

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 7 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) [9]
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 7 種
リファレンス音声	F0: 1 ms, SP: 1 ms
F0 フレームシフト幅	10，20，30 ms
SP フレームシフト幅	3，5，7 ms

図 2 に音声パラメータの補間の概要例を示す．WORLD で分析を行った際に，F0 パラメータを 1 ms で分析し，SP パラメータを任意のフレームシフト幅で分析する．次に，F0 パラメータの持つフレーム数に合わせて SP パラメータの線形補間を行い，疑似フレームを作成する．これらのパラメータをフレーム毎に合成することで最終的な分析合成音として出力する．同様に SP パラメータを 1 ms で分析し，F0 パラメータのフレームシフト幅を調整することができる．

本稿では，F0 と SP のパラメータのどちらがより音声分析合成の音質において支配的であるかを検討する．また，AP は F0 に依存してフレームシフト幅が変わるため，調整された F0 から分析を行う．そのため，今回は個別にフレームシフト幅の調整は行わない．

4. 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．

表 2　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 と表す．
Table 2　List of adjusted p values.

組み合わせ	補正 p 値
1 ms，F0–10 ms	n.s.
1 ms，F0–20 ms	n.s.
1 ms，F0–30 ms	<0.001
F0–10 ms，F0–20 ms	n.s.
F0–10 ms，F0–30 ms	<0.001
F0–20 ms，F0–30 ms	<0.001
1 ms，SP–3 ms	n.s.
1 ms，SP–5 ms	0.011
1 ms，SP–7 ms	<0.001
SP–3 ms，SP–5 ms	n.s.
SP–3 ms，SP–7 ms	<0.001
SP–5 ms，SP–7 ms	0.014

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–weighter 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) を使用した [9]．

実験に用いた GUI を図 3 に示す．被験者には，1 音声につき 1 種類のリファレンス音声と 6 種類の比較対象となる音声異を提示する．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 7 種類の比較音声との比較を行う．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，比較音声に必ずリファレンス音声が含まれるために，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．使用音声の順番はランダム化されており，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．これらの方法により，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 7 種類は，WORLD を用いて分析合成されたフレームシフト幅の異なる 7 個の音声刺激である．その内訳は，リファレンス音声としてフレームシフト幅 1 ms の分析合成音，比較音声として F0 のみをフレームシフト幅 10 , 20, 30 ms で分析した音声，SP のみをフレームシフト幅 3, 5, 7 ms で分析した音声である．これらのシフト幅は，予備実験にて決定した．

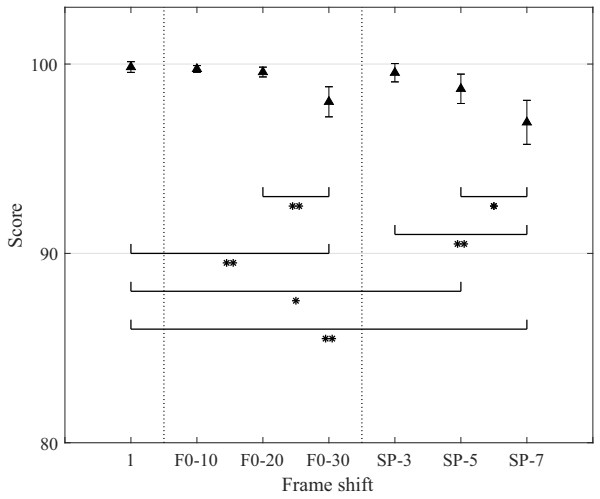

図 4　MUSHRA 法による各パラメータのフレームシフト幅毎の音質評価結果．*は Adjusted-$p < 0.05$ を示す．**は Adjusted-$p < 0.01$ を示す．

Fig. 4　Sound quality evaluation results for each frame shift of each parameter by the MUSHRA method.

5. 実験結果

実験結果を図 4 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．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 [10] に基づく 2 段階線形上昇手順 [11] を実施した．

比較リストとその補正 p 値を表 2 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，F0 では 20–30 ms 区間に，SP では 5–7 ms 区間に有意差が認められた．また，男女別の評点差の違いも検証を行ったが，そちらには差が見られなかった．これらの結果から，パラメータに必要なフレームシフト幅は SP の方がより支配的であり，細かい分析が必要なことが分かる．また，F0 の分析のみに関しては，20 ms まで品質の劣化がほとんどないという結果が明らかとなった．

6. 考　　察

F0 は 20 ms まで音質の劣化が検出されなかった．これは，基本周波数の変化が時間的に滑らかであるため，基本周期よりも長いシフト幅であっても品質劣化には繋がらないことを示唆する．SP は 7 ms まで音質の劣化が検出されなかった．これは SP が F0 よりも細かく分析する必要があることを示している．また，これらの結果より，前報 [2] で得られたフレームシフト幅による音質評価の結果が SP に依存するものであったことが確認できた．

F0 よりもスペクトル包絡の方が細かいフレームシフト幅での分析が必要であることが確認できた．F0 は 20 ms, SP は 3 ms で分析することによって音質劣化することなく分析合成を行うことができる．この結果より，既存のフレームシフト幅よりも，F0 のフレームシフト幅を長くすることによって音質を劣化させずにデータ量の削減が可能であることが確認された．

7. 終わりに

本研究では，フルバンド音声の分析合成における，パラメータ毎に必要なフレームシフト幅について検証を行った．フルバンド音声における分析合成音を用い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フレームシフト幅を策定した．その結果として，音声分析合成に必要なフレームシフト幅はスペクトル包絡に大きく依存し，より基本周波数パラメータよりも細かく分析する必要があることが確認された．

今後の発展としては，音声符号化を目的として，本研究で得られた実験結果を用いて音声パラメータの量子化を行う計画である．これらの結果から，品質劣化が生じないフルバンド音声を表現可能なデータ転送レート (bps: bit per second) を明らかにする．

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] 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.

[2] 宮下玄太, 森勢将雅, “フルバンド音声を対象とした品質劣化のない音声分析合成のためのフレームシフト幅の検証,” 日本音響学会秋季研究発表会, pp.259-260, 2017.

[3] H. Kawahara, I. Masuda-Kasuse and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds,” Speech Communication, vol.27, pp.187–207, 1999.

[4] 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.

[5] 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.

[6] S. Takaki and J. Yamagishi, “A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis,” in Proc. ICASSP, pp.5535–5539, 2016.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.