WORLD による合成音声の位相特性の観察 *

☆高津航輝, 矢田部浩平 (農工大)

1 まえがき

WORLD に代表される信号処理ベースの音声分析合成システム（本稿ではボコーダと呼称する）は，深層学習ベースの手法に比べて計算量が少なく動作が軽い．しかし，ボコーダによって合成された音声には，音声がブザーのようになる，いわゆるバジー感がある．このバジー音はボコーダの位相特性の悪さに起因すると言われている [1]．ボコーダの位相特性の悪さを改善できれば，ボコーダは計算量が少ない音声合成手法として深層学習ベースの手法との棲み分けが期待できる．本研究では，ボコーダのバジー感を軽減するための初期検討として，WORLD の再合成音声と元音声の位相を比較した．また，その位相の違いを時変オールパスフィルタによって補償し．音質の変化を調べた．

2 時間周波数解析

音声の性質を観察するために，信号を離散ガボール変換（DGT）して得られる複素スペクトログラムの位相を利用する．信号 x の DGT の定義式を示す [2]．

$$X[m,n] = \sum_{l=0}^{L-1} x[l]\overline{w[l-an]}e^{-2\pi iml/L} \qquad (1)$$

ただし w は長さ L の窓，a は時間方向の間引き量である．位相情報を画像として表示するには，式 (1) で得られた位相を HSV 色空間に対応させる．まず，位相と色相はともに 360° で循環する．これを利用して，HSV の色相には位相を反映する．次に，振幅の大きい周波数帯域が目立つように，振幅の大きさを HSV の明度に反映する．最後に，彩度は 100 % に固定した．このように定めた HSV を RGB に変換して，位相スペクトログラムを描画した．

3 時変オールパスフィルタ

WORLD では，再合成音声に最小位相 [3] を与える．しかし元音声には最小位相以外の成分も含まれるため，再合成音声と元音声の位相の性質は異なる．この位相の性質の違いは，WORLD 再合成音声中にバジー感として現れると言われている [1]．本研究ではこれを仮定して，WORLD 再合成音声の位相を元音声のものに近づけたときのバジー感の変化を観察する．具体的には，WORLD 再合成音声の振幅を変えずに位相のみを変えることで，元音声の位相情報に近づけた．

信号の振幅を変えずに特定の周波数の位相を変化させるフィルタとして，オールパスフィルタが挙げられる．時不変なオールパスフィルタの周波数特性は次式で与えられる [4]．

$$H_A(\omega) = \frac{r + ae^{-j\omega} + e^{-j2\omega}}{1 + ae^{-j\omega} + re^{-j2\omega}} \qquad (2)$$

ただし，式中の a は

$$a = -(1+r)\cos(\omega) \qquad (3)$$

で表され，ω がフィルタの中心角周波数を，$0 < r < 1$ が位相特性の変化の急峻さを表す．

音声信号では基本周波数 f_o が時々刻々と変化する．そのため，特定の調波の位相を変化させるにはオールパスフィルタの特性も時変である必要がある．本研究では，WORLD で推定した f_o 配列を使用して時変オールパスフィルタを設計し，これを再合成音声に適用する．時変オールパスフィルタの時刻 k における周波数特性は以下の式で表される [5]．

$$H_k(\omega) = \frac{b_k + c_k e^{-j\omega} + e^{-j2\omega}}{1 + c_k e^{-j\omega} + b_k e^{-j2\omega}} \qquad (4)$$

ただし，式中の b_k，c_k は，時変定数 p_{1k}，p_{2k}，時不変定数 p を用いて

$$b_k = \frac{-p + p_{1k} + p_{2k}}{p + p_{1k} + p_{2k}} \qquad (5)$$

$$c_k = \frac{2(p_{1k} - p_{2k})}{p + p_{1k} + p_{2k}} \qquad (6)$$

で表される．ここで簡単のために時不変定数 $p = 1$ と定めると，式 (2) と式 (3) より，時変定数 p_{1k} と p_{2k} は r と ω_k を用いて

$$p_{1k} = \frac{-(r - \cos\omega_k - r\cos\omega_k + 1)}{2(r-1)} \qquad (7)$$

$$p_{1k} = \frac{-(r + \cos\omega_k + r\cos\omega_k + 1)}{2(r-1)} \qquad (8)$$

と書き換えることができる．このため，本研究で設計する時変オールパスフィルタは，時不変なオールパスフィルタの式 (2) に基づいて各時刻 k で個別に設計することが可能である．

---

* Observation of phase characteristics of speech synthesized by WORLD. By Koki TAKATSU and Kohei YATABE (Tokyo University of Agriculture and Technology).

4　調波間の位相関係の観察

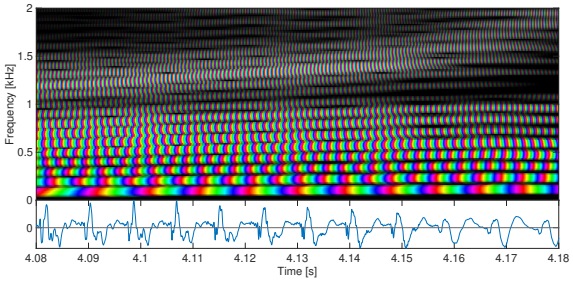
まず WORLD 再合成音声と合成元音声について，目視で位相スペクトログラムを，主観的な印象で音声のバジー感を比較した．次に，WORLD 再合成音声の調波間の位相関係と合成元音声の調波間の位相関係を揃えるために，時変オールパスフィルタを WORLD 再合成音声に適用した．以降，フィルタ適用後の WORLD 再合成音声をフィルタ適用音声と呼ぶ．最後に，フィルタ適用音声と WORLD 再合成音声，そして合成元音声の 3 つの音声について，目視で位相スペクトログラムを，主観的な印象で音声のバジー感を比較した．合成元音声には SiSEC2018 が提供する男声（`dev1_male4_src_2.wav`）を使用した [6].

WORLD 再合成音声と合成元音声の位相スペクトログラムを**図**–1 に示す．同じ色の箇所は位相が同じであることを表している．この位相スペクトログラムを目視で比較する．各音声，4.028 秒付近の位相に注目する．この時刻において，合成元音声では f_o 以外の高調波の位相が揃っている．一方，WORLD 再合成音声は，およそ第 5 調波から f_o に向かって同位相のタイミングが徐々にズレている．音声の主観的な印象では，WORLD 再合成音声にはバジー音が多く含まれているように感じられた．

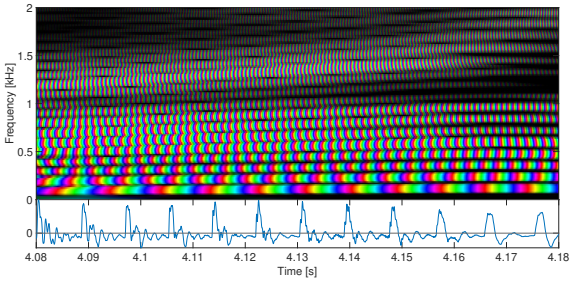
時変オールパスフィルタの設計では，同位相タイミングの階段状のズレ軽減と，同時刻での f_o の位相の合致を目標に定めた．設計したフィルタは，中心周波数 $f = 0.25 f_\mathrm{o}$ で $r = 0.8$ のもの，$f = f_\mathrm{o}$ で $r = 0.98$ のもの，$f = 6 f_\mathrm{o}$ で $r = 0.92$ のもの 3 種類から構成される．

フィルタ適用音声の位相スペクトログラムを**図**–2 に示す．図–1 の WORLD 再合成音声や合成元音声の位相スペクトログラムと目視で比較した．フィルタ適用音声の 4.157 秒付近の位相に注目すると，合成元音声ほど揃ってはいないが，階段状のズレが軽減されていることがわかった．また，同時刻での f_o の位相は目視ではほぼ一致していた．

次に，WORLD 再合成音声や合成元音声を主観的な印象で音声のバジー感を比較した．フィルタ適用音声には未だにバジー感が残るものの，WORLD 再合成音声に比べてバジー感が軽減した．特に，1 kHz 以下の帯域のみに注意して聞くとその傾向を強く感じられた．しかし，フィルタ適用音声には，WORLD 再合成音声になかった低音の異音が加わっていることが確認できた．これは，適用した時変オールパスフィルタのフィルタ係数が時変であることに由来すると考えられる．

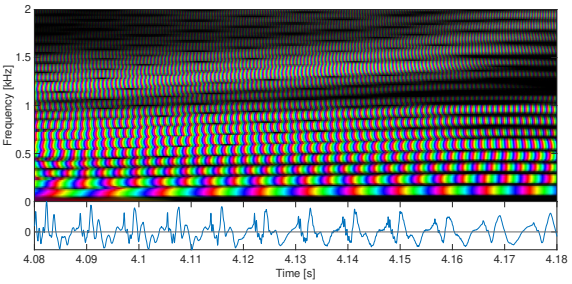


(a) 合成元音声



(b) WORLD 再合成音声

図–1　合成元音声と WORLD 再合成音声の位相スペクトログラム（周波数サンプリング点数 $M = 1024$，時間方向の間引き量 $a = 1$）



図–2　フィルタ適用音声の位相スペクトログラム（$M = 1024$，$a = 1$）

5　むすび

WORLD 再合成音声に時変オールパスフィルタをかけたときの，調波間の位相関係とバジー感の変化を観察した．今後はバジー感をさらに軽減するフィルタを検討する．

参考文献

[1] H. Kawahara, K. Sakakibara, M. Morise, H. Banno, T. Toda and T. Irino, “Frequency Domain Variants of Velvet Noise and Their Application to Speech Processing and Synthesis,” Proc. Interspeech 2018, pp.2027–2031, Sep 2018.

[2] 矢田部浩平，“第三回：短時間フーリエ変換，”日本音響学会誌，vol.77，no.6，pp.396–403，June. 2021.

[3] 森勢将雅，“話声の合成における基盤技術，”日本音響学会誌，vol.75，no.7，pp.387–392，July. 2019.

[4] 田中聡久，川村新，音声音響信号処理の基礎と実践，コロナ社，東京，2021.

[5] S. Bilbao, “Time-varying Generalizations of Allpass Filters,” in IEEE Signal Process. Lett., vol. 12, no. 5, pp. 376-379, May 2005.

[6] SiSEC2018, Underdetermined-speech and music mixtures, Available: `http://www.irisa.fr/metiss/SiSEC10/underdetermined/dev1.zip`, 2018.