Vol.2016-MUS-111 No.38
2016/5/21

スペクトル位相復元を用いたケプストラム領域ピッチ操作

林耕平[†1]　松原聖人[†1]　光本大記[†1]　濱田康弘[†1]　小野順貴[†2]　嵯峨山茂樹[†1]

概要：音声のケプストラム領域において，高次部分に現れる音源成分の位置に対する操作をすることで，ピッチ変更の効果を与えることができる．ただし，パワースペクトル領域を経由した時点で位相が失われてしまうため，従来のケプストラムからの波形生成には MLSA フィルタなどが用いられていたが，source-filter モデル特有の品質劣化が生じやすかった．本研究では，スペクトル位相復元技術を用いることによって，フィルタを用いないノンパラメトリックな音声合成を検討した．また，Lag 窓を利用した処理を施すことにより，音質の向上を目指した．

キーワード：音声合成，ケプストラム分析，スペクトル位相復元，Lag 窓

1. はじめに

本稿ではケプストラム領域による基本周波数（F_0）成分操作と，スペクトル位相復元技術[1]を組み合わせた，ノンパラメトリックでフィルタによらない音声合成法の可能性を検討する．

ケプストラム分析[2]では，音声スペクトルの包絡成分が低ケフレンシ領域に，微細構造が高ケフレンシ領域に表出され，音声の声道特性と声帯振動を（ほぼ）正確に分離できることが利点である．また，音声の F_0 成分がピークとして現れ，これを操作して音高を変えることができるため，音声加工に有用である可能性があり，パラメトリックなモデルに基づかぬため，原音の情報を保持できる可能性がある．

ケプストラムから波形を生成するには，パワースペクトル領域を経由する必要があるが，パワースペクトル領域では位相が失われてしまい，Oppenheim らの複素ケプストラムに基づく準同型処理[3]を別にすると，ケプストラムから波形領域に変換することができなかった．そのため，従来のケプストラムからの波形生成では，今井らの対数振幅特性近似（LMA）フィルタ[4]やメル対数スペクトル近似（MLSA）フィルタ[5]が用いられてきた．

これらのフィルタは，線形予測分析合成（LPC）などと同様に，音源モデルと少数のスペクトル包絡パラメータを用いて巡回型フィルタにより信号を生成する source-filter モデル方式で波形を生成する．ただし，この方式では，時間特性による時間遅れや，共振特性により振幅異常が生じて人工的な印象を与えやすい．また，少数パラメータに依る音声合成では，有声音なら F_0 とパワー，無声音なら白色雑音のパワーのみがパラメータとして扱われるため，人間の持つ微妙な声質情報が失われがちである．

近年，パワースペクトルから無矛盾な位相を有する波形生成が可能なスペクトル位相復元技術が開発された．そこで，ケプストラム領域で F_0 成分を操作し，スペクトル位相復元技術を組み合わせることにより，人間の声質情報を活かした，フィルタに依らない音声合成法を検討する．

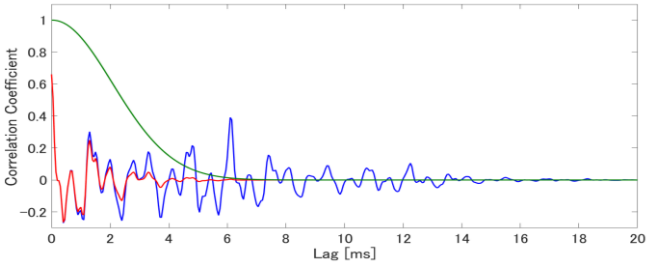

Figure 1　Lag 窓を掛けた自己相関関数

2. 非フィルタケプストラム分析合成系

2.1 ケプストラムドメインでの音声加工

音声のケプストラムは，ケフレンシ軸の低次部分は主に声道特性に対応するスペクトル包絡を反映し，高次部分は声帯振動に対応するスペクトル微細構造を反映し，特に F_0 に対応してピークを生じるものと考えられて来た．この F_0 成分の位置に関して操作を行えば，音声に対してピッチ変更が可能と考えられる．

つまり，入力音声をケプストラム分析し，それによって取得した情報を F_0 成分に関して操作したあと，パワースペクトルから音声波形に戻すことができれば，ピッチ変更などの変化を入力音声に与えることができる．

しかし，ケプストラムは対数パワースペクトル領域からの逆フーリエ変換なので，位相情報を持たないため，ケプストラムを操作できたとしても，音声信号に復元することができない．本稿ではそこにスペクトル位相復元技術を用いて，スペクトルに対して無矛盾な位相を付加することに

†1 明治大学
〒164-8525　東京都中野区中野 4-21-1
Email: {ev40556, ev40550, ev30626, sagayama}@meiji.ac.jp
†2 国立情報学研究所 / 総合研究大学院大学
〒101-8430 東京都千代田区一ツ橋 2-1-2
Email: onono@nii.ac.jp

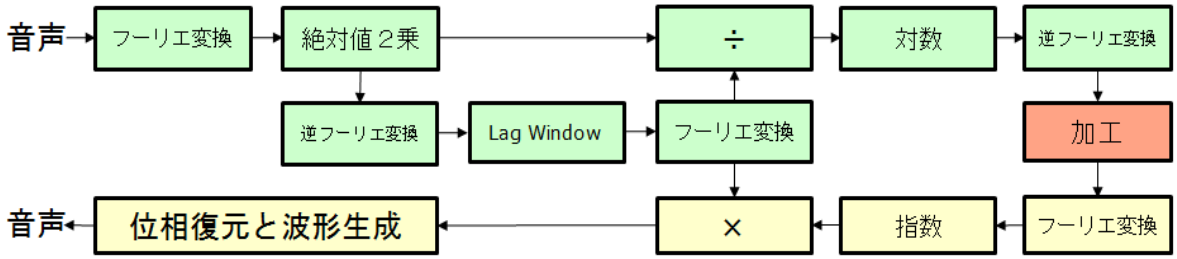

Figure 2　ケプストラム非フィルタ分析合成系

より，ケプストラム領域からの波形生成を可能とする．

2.2　Lag 窓によるスペクトル平滑化の利用

ケプストラム分析では，リフタによって，低ケフレンシ部分の声道特性と高ケフレンシ部分の音源特性に分離することがよく行われるが，低ケフレンシ部分はパワースペクトルの包絡を表現しているわけではなく，謂わば微細構造の山谷の中心を辿る成分を表している．そこで，包絡成分の分離の精度を向上させるために，パワースペクトルの自己相関関数（Fig.1．青線）に Lag 窓[6,7]（Fig.1．緑線）を掛けたもの（Fig.1. 赤線）をフーリエ変換することにより，微細構造の谷を埋めて，パワースペクトルの包絡成分を直接求めることができる（Wiener-Khinchin の定理と畳み込み定理の応用）. これによりパワースペクトルを割って微細構造を得ることで，パワースペクトルから直接ケプストラムを求める方法に比べより正確な分離が可能になる．

2.3　ケプストラム非フィルタ分析合成処理の流れ

2.1 節，2.2 節より，分析合成系における F_0 成分の移動の流れは以下の様になる．(Fig.2)

(1)　時間信号のフーリエ変換の絶対値の２乗によりパワースペクトルを得る．
(2)　逆フーリエ変換をし，短時間自己相関関数を得る．
(3)　上記に Lag 窓をかける．（今回はガウス窓）
(4)　これをフーリエ変換することで，パワースペクトルの包絡成分を得る．パワースペクトルをこの成分で除して微細構造を得る（ここでは「F_0 成分スペクトル」と呼ぶ，Fig.3 では黄線）.
(5)　上記の対数をとり逆フーリエ変換して，F_0 成分ケプストラムを得る（Fig.4．上）.
(6)　このケプストラムをケフレンシ軸で線形補間により伸縮操作する（Fig.4．下）.
(7)　上記をフーリエ変換し指数関数変換し，(4)の包絡成分と掛け合わせることでパワースペクトル領域へ戻す．

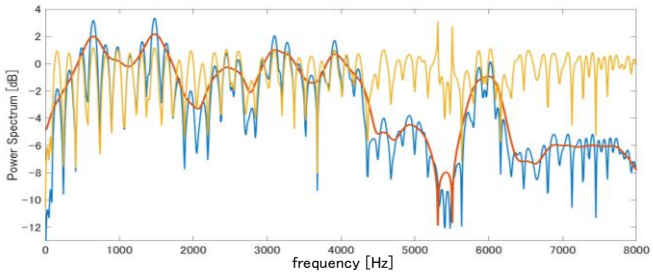

Figure 3　音声のスペクトルと音源スペクトル（青線: パワースペクトル，赤線: 包絡成分，黄線: F_0 成分スペクトル）

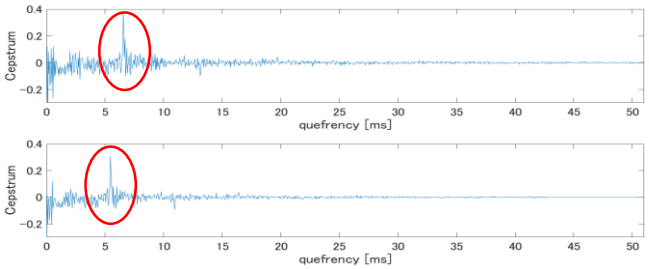

Figure 4　　上: F_0 成分ケプストラム
下: 伸縮した F_0 成分ケプストラム

(8)　位相復元により，時間信号を得る．

3.　F_0 加工実験と定性的評価

3.1　F_0 加工の方法

予備実験として，ケフレンシ軸において低周波領域と F_0 成分のピークを示す部分以外を 0 にしたところ，著しく品質が損なわれ，ケフレンシ軸のピーク以外の細かな上下も音声を加工する際に必要な情報であるということがわかった．そこで，ケプストラム分析によって得られた情報をなるべく保持するため，F_0 成分を伸縮する際には，ケフレンシ軸において，微細構造に対応する部分を線形圧縮・伸張した．その際，以下の条件で比較した．

① Lag 窓を用いた処理を施していないもの
② Lag 窓を用いた処理を施したもの

3.2 検証結果と考察

条件①, ②の両方で, ケフレンシ軸で F_0 成分に対応する部分の伸縮操作を行うことでピッチが変更された. この方式による合成音声は, 原音と比べると全体的に濁りが感じられたが, source-filter 方式による合成音声より人工的な印象は少なかった. また, 条件①よりも条件②の方がノイズは少なく, Lag 窓により高品質な音声加工が可能になったといえる.

MLSA フィルタなど source-filter モデルを用いた従来法は, 失われた位相に対して少数パラメータで制御される巡回型フィルタを用いることで音声加工を施していた. 本稿での source-filter モデルを用いない形式でのボコーダは, そのような従来法に依って生じていた合成音声独特の機械的な特徴をより肉声に近づけることができる原理的可能性を持っていると考えられる. 合成したケプストラムからの音声合成[8]では濁りが感じられないことから, フレーム間の揺らぎを押さえることで濁りを軽減できる可能性がある.

4. 結論

本稿では, ケプストラム分析とスペクトル位相復元技術を組み合わせることで, 非フィルタでノンパラメトリックな音声加工法を提案した. また, Lag 窓を用いた処理を施すことで, ケプストラムの包絡と微細構造のより正確な分離を行った.

ケプストラム領域での F_0 成分の伸縮操作により, 出力音声ではピッチ変更の効果が与えられることが確認された. ただし, フレーム間の揺らぎと思われる濁りが聴かれたため, 今後の展望として, ピッチ同期のケプストラム分析を行うことにより濁りを軽減する音声加工法を検討する予定である.

謝辞 この研究は, 日本学術振興会科学研究費補助金基盤研究(A)課題番号 26240025「音楽信号・作曲理論・演奏の数理モデルを融合する音楽音響情報処理の研究」の部分的支援を得て行われた.

参考文献

[1] J. Le Roux, N. Ono, and S. Sagayama, “Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction,” in Proc. SAPA, Sep. 2008.
[2] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, “The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking,” Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, pp. 209-243. New York: Wiley, 1963.
[3] B. ゴールド, C. M. レイダー（石田訳）, 電子計算機による信号処理, 共立出版, pp. 280-285, 1972.
[4] 今井, 北村, “対数振幅特性近似フィルタを用いた音声の分析合成系,” 電子通信学会論文誌, Vol. J61-A(6): 527-534, 1978.
[5] 今井, 住田, 古市, ”音声合成のためのメル対数スペクトル近似（MLSA）フィルタ,” 信号処理学会論文誌, Vol. J66-A(2), pp. 122-129, 1983.
[6] Y. Tohkura, F. Itakura, S. Hashimoto, “Spectral Smoothing Technique in PARCOR Speech Analysis-Synthesis,” IEEE Trans. ASSP, Vol. 26(6), pp.587-596, 1978.
[7] 嵯峨山, 古井, “ラグ窓を用いたピッチ抽出の一方法,” 電子情報通信学会全国大会予稿集, 1235, Vol. 5, p. 263, 1978.
[8] 濱田, 小野, 嵯峨山, “無矛盾位相復元を用いたケプストラム特徴量からの音声合成,” 情報処理学会第 78 回全国大会講演論文集, Vol. 2, pp. 15-16, 2016.