Vol.2022-MUS-134 No.38
Vol.2022-SLP-142 No.38
2022/6/18

変調周波数伝達特性と周波数応答で音声処理を調べよう

河原 英紀[1,a)]　矢田部 浩平[2,b)]　榊原 健一[3,c)]　北村 達也[4,d)]　坂野 秀樹[5,e)]　森勢 将雅[6,f)]

概要：新しい計測信号として著者らが提案した CAPRICEP を用いることにより，ピッチ抽出器などの，音声信号処理系を周波数領域において詳細にかつ客観的に調べることができるようになった．ここでは，その効果と使い方を紹介し，議論したい．

Investigate speech signal processing regarding modulation frequency transfer function and responses

1. はじめに

2016 年の WaveNet の発表 [1] を嚆矢として急速に広がった深層学習を用いた研究により，合成音声の品質は飛躍的に向上した [2]．現在では，人間による音声と区別できないレベルの品質が実現されている [3]．その過程を通じて，これまでの大規模システムの実現方法の常套手段であったモジュール化と分割統治の弱点が顕在化し，目的を直接実現しようとする end-to-end システムが主流になりつつある．しかし，その品質を実現している機構が科学的な意味で理解できている訳ではない．機構の理解にはモジュール化と分割統治が（少なくとも現在では）必要である．システムの実現において，モジュール化と分割統治の（深層学習に基づく様々な方法に対して）弱点が顕在化した背景には，歴史的に貧弱な計算能力という制約の下で，モジュールの性能を（暫定的に）与えられた指標に関して部分最適化せざるを得なかったという歴史的経緯があるのではないかと考えている．ここでは，モジュールを構成する音声処理機構を，音声の物理的属性を計測し変換する装置として捉えて評価する方法を提案する．こうして処理装置としての性能を保証されたモジュールを用いることで，人間による音声生成と知覚と，さまざまな音声処理技術による物理属性の操作との関係を深く理解することにつなげたいと考えている．本資料では，まず，基本周波数を抽出し処理するモジュールの評価について議論する．

[1] 和歌山大学
Wakayama University
[2] 東京農工大学
Tokyo University of Agriculture and Technology
[3] 北海道医療大学
Health Science University of Hokkaido
[4] 甲南大学
Konan University
[5] 名城大学
Meijo University
[6] 明治大学
Meiji University
[a)] kawahara@wakayama-u.ac.jp
[b)] yatabe@go.tuat.ac.jp
[c)] kis@hoku-iryo-u.ac.jp
[d)] t-kitamu@konan-u.ac.jp
[e)] banno@meijo-u.ac.jp
[f)] mmorise@meiji.ac.jp

2. 背景

基本周波数は有声音の重要な属性の一つである．信号の物理的属性である基本周波数は，音の高さの知覚的属性であるピッチと密接に関係している．音声信号から基本周波数を求める方法は，『ピッチ抽出』と呼ばれており[*1]，音声処理の初期 [5] から現在に至るまで研究され続けている．これまでの研究では，音声合成や音声認識への応用を目的とするものが多く，評価もそれらの応用において問題となる項目が中心であった．音声生成の研究のための測定装置としての研究は，1993 年の Titze によるもの [6] 以降，本格的には行われていないようである．

持続母音を一定のピッチで発声しているときに，同時に基本周波数を変調した音を聴かせると，発声した音声の基本周波数が，その変動を補償するような応答を示す [7]．この応答を定量的に調べるためには，基本周波数の変化を正確に測定できるピッチ抽出器が必要である．最初に，ピッ

[*1] 厳密には誤用であるが，ここでは慣用に従う．ただし基本周波数については慣用である『F0』という表記ではなく『f_o』（エフ・オーと読む）を用いる [4]．

チ抽出器を調べようとした動機は，この要請に基づいている．この現象を調べるために用いた実験の枠組みは [8], [9]，ピッチ抽出器を測定器として評価するための枠組みとしても用いることができる．その鍵となるのが，以下で説明する CAPRICEP（cascaded all-pass filters with randomized center frequencies and phase polarities）である [10].

3. CAPRICEP

オールパスフィルタのインパルス応答は時間的に広がっている．この時間的に広がったインパルス応答と，時間反転したインパルス応答とを畳み込むことにより，インパルスを復元することができる．この性質を利用することにより，実使用には不適切なインパルスそのものを用いずに，対象とする音響システムなどのインパルス応答を測定することができる．オールパスフィルタを従属接続したものもオールパスフィルタとなる．この性質を利用し，IIR(Infinite Impulse Response) フィルタをある規則に基づいて多数従属接続したものが CAPRICEP である．この規則を適切に設計することにより，CAPRICEP のインパルス応答の包絡の形状を設計することができる．

CAPRICEP のインパルス応答の設計には，乱数を用いて（通常）数千個の値を設定する．そのため，異なった乱数から設計された CAPRICEP のインパルス応答を，直交系列を用いて符号を設定して時間軸上に周期的に配置することにより，直交する信号の組を作ることができる．この信号の組を加算して作成した測定用信号を用いることにより，対象とするシステムのインパルス応答に加え，システムの非線形性に起因する信号依存応答および，時変ランダム応答を同時に測定することができる．

3.1 ピッチ抽出器の測定

基本周波数の周波数変調に対する発声された音声の基本周波数の応答を調べる際には，静的な応答に加えて，周波数変調に対する変調周波数伝達特性を求めることが必要になる．基本周波数を前で説明した測定用信号で周波数変調し，ピッチ抽出器で求められた基本周波数の系列を応答として変調周波数特性を求めることができそうである．しかし，音声信号には多数の調波成分が含まれているため，周波数変調に用いる信号に基本周波数よりも高い周波数の成分が含まれていると不都合が生じる．また，周波数変調は非線形処理であるので，変調を深くすると側波帯が広がるという不都合が生じる．測定対象であるピッチ抽出器には非線形性が含まれていると考えられる．必要なのは音声の基本周波数変動の測定であるので，実際の音声の基本周波数の変動と同じ性質を有する試験信号を用いて，ピッチ抽出器の振る舞いを調べる必要がある．しかし，それらの性質は測定により調べられるべきものであり，現状では測定法が確立していないため，いくつかの暫定的な値を用いて

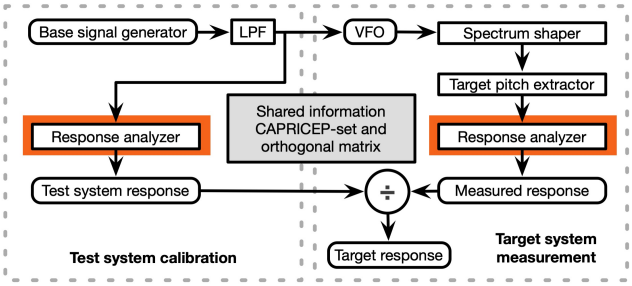

図 1 Schematic diagram of response measurement procedure[9]

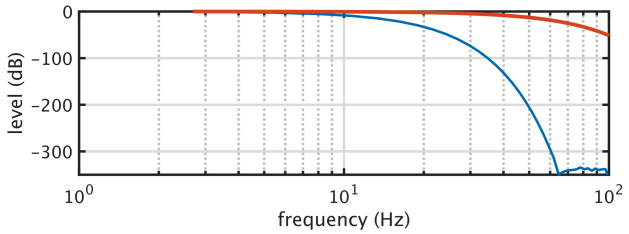

図 2 Frequency response of the LPF used in [9] (blue line) and this report (red line).

測定を行うこととした [9]．この引用した資料の執筆時に用いた値は，その後の測定により，不適切であることが明らかとなった．この資料では，前述の資料の問題点を指摘し，より適切な値を用いた測定について説明する．

3.2 周波数変調に対する応答の測定

図 1 に測定法の仕組みを説明する資料 [9] の図を再掲する．図の右側が測定対象となるピッチ抽出器を含む系，左側が，前の節で説明した制約を考慮するために導入した低域通過フィルタ（LPF）の特性を測定する系である．ピッチ抽出器を含む系の測定結果を，低域通過フィルタの（逆）特性を用いて補償することにより，ピッチ抽出器自体の変調周波数伝達特性を求める．しかし，資料 [9] では，聴覚的に提示した基本周波数変調音に対する発声された音声の基本周波数の応答を求める実験で用いていた LPF をそのまま用いていたため，ピッチ抽出器の特性の正確な測定に失敗していた．

図 2 に，前回の資料で用いた LPF と今回用いる LPF の特性を示す．前回の資料の LPF では高い変調周波数成分のレベルが低過ぎたため，主要な線形時不変応答が正確に求められないという問題が生じていた．

図 3 に，Neural ではない VOCODER として広く用いられている WORLD[11] のピッチ抽出器である Harvest[12] の測定を，前回の資料の LPF と今回のものを用いて行った結果を示す．それぞれ太い実践が線形時不変応答，破線が非線形に起因する信号依存の応答，点線がランダムおよび時変応答を表している．前回の資料の LPF を用いた測定結果では，上段に示したように，非線形性に起因する応答やランダム応答が変調周波数とともに急速に増加するた

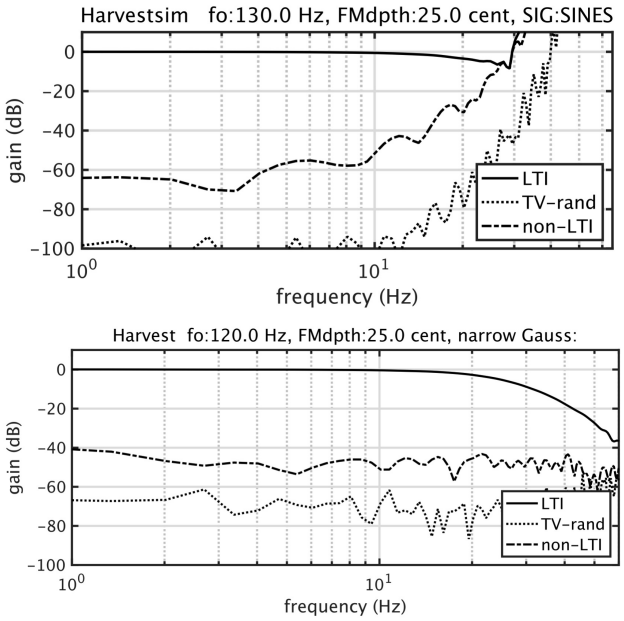

図 3 Modulation transfer function of Harvest pitch extractor. The upper plot shows the result reported in [9]. The lower plot shows the revised result using the revised LPF.

めに，25 Hz 以上の変調周波数における線形時不変応答の様子を見ることができない．今回の LPC を用いた結果では，表示されたすべての帯域において，線形時不変応答の様子を調べることができている．

3.3 ピッチ抽出器の測定例

前回の資料の問題点を解消したピッチ抽出器を測定するための枠組みを用いて、実装を入手できるいくつかのピッチ抽出器の特性を測定した。測定信号の標本化周波数は 44100 Hz、LPF は、前の節で示したものを用い、周波数変調の深さを RMS 値で 25 cent として、基本周波数を 80 Hz から 400 Hz まで 1/48 オクターブ毎に設定した。スペクトルは日本語母音/a /を用いた。

用いたピッチ抽出器は以下である。

MATLAB に用意されているもの　MATLAB では幾つかのピッチ抽出器が実装されている。まず `pitch` という関数のオプションとして以下が実装されている。CEP のオプションでは、ケプストラムに基づく方法 [13]、LCF のオプションでは、線形予測分析に基づく方法 [14]、LHS のオプションでは、調波性を利用したハーモニックサムによる方法 [15]、PEF のオプションでは、比較的最近の耐雑音性を考慮した方法 [16] や、SRH のオプションでは、同様に最近の調波性を利用した方法 [17] が実装されている。その他に、CREPE と呼ばれる深層学習に基づく方法も `pitchnn` として提供されている [18]。

YINestimate　シフト差分に基づく方法である YIN[19] は、著者らによる実装が最新の MATLAB では動かないため、別のパッケージの実装を用いた [20]。

SWIPEP　鋸歯上波のパワースペクトルの性質を利用した方法 [21], [22] として、引用されることの多い方法である。

RAPT (VOICEBOX)　RAPT は自己相関と後処理に基づく方法であり [23]、ここでは VOICEBOX で提供されている実装 [24] を用いた。

REAPER　REAPER は、声門閉止の検出と有声音／無声音の判定を統合した方法であり、ここではオープンソースとして提供されているものを用いた。

Praat　ここでは広く用いられている言語学的研究のためのツール Praat の既定値として用意されているピッチ抽出器 [25] を用いた。

openSMILE　paralinguistic の研究に用いられる openSMILE[26] に用意されている調波構造を利用するピッチ抽出のオプション prosodyShs.conf を SHS とし、自己相関を用いるオプション prosodyAcf.conf を ACF とした。

STRAIGHT (NDF, XSX)　従来型の VOCODER である STRAIGHT に用意されている 2 種類のピッチ抽出器を分析した。NDF[27] は、legacy-STRAIGHT[28] の最新のオプションであり、XSX は、TANDEM-STRAIGHT[29] のピッチ抽出器である。

WORLD (Harvest)　VOCODER である WORLD のピッチ抽出器である。

NINJAL　国立国語研究所の CSJ コーパスの分析のために開発した方法である [30]。ここでは、測定結果に基づいて内部の時定数を調整した版を NINJALX2 としている。

図 4 は、これらの方法を分析した結果から作成したムービーのスナップショットである。図は、上下に、二つのスナップショットを示している。上の図は、男性の基本周波数に相当する約 120 Hz についての結果を示し、下の図は、女性の基本周波数に相当する約 240 Hz についての結果を示している。

CEP や LCF などの初期の方法で変調周波数の帯域幅が狭く、非線形性やランダム応答が多いことは想定されたことではあるが、CREPE など最近の方法も初期の方法と比較して大きく異なっていないのは予想外であった。測定器として、現象を定量的に精密に記述することが目的ではなく、音声認識など、目的とする応用の性能を向上させるように全体として最適化を図ったことによるのではないかとも思われる。

それらの方法に対して、VOCODER への応用を目的として作成された、NDF、Harvest が広い帯域幅と少ない非線形性、ランダム応答であることが目立っている。広く用いられている Praat が比較的良好なことと、openSMILE の SHS が良好なことは、それらの方法の利用者が多いことを

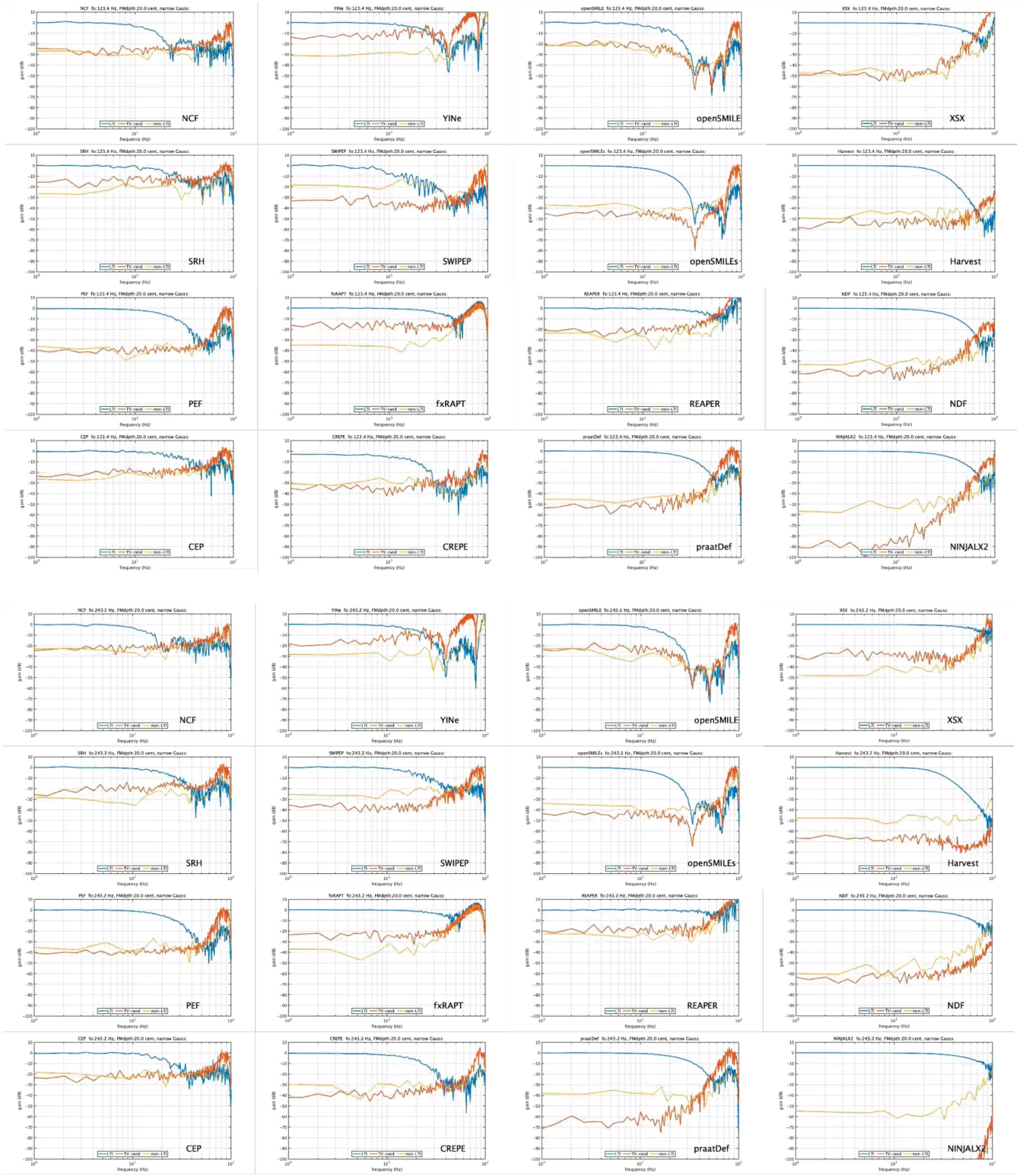

図 4 Snapshots of visualization movie. The movie displays sixteen pitch extractors frequency responses for different fundamental frequencies.

考慮すると朗報である。

なお、国立国語研究所のコーパス分析用に開発した方法を、今回紹介した測定法を用いて調整した NINJALX2 が、突出して広い帯域幅と少ない非線形性、ランダム応答を有することも目立っている。この方法は、対数周波数軸上で同一の形状のフィルタ出力の瞬時周波数を求めるという素直な方法である。測定器としてのピッチ抽出器としては適切であるが、雑音環境下でのピッチ抽出など、応用のための構成要素としては弱点が顕在化すると思われる。

図 5 に、測定信号に pink noise を加え SNR を 20dB と

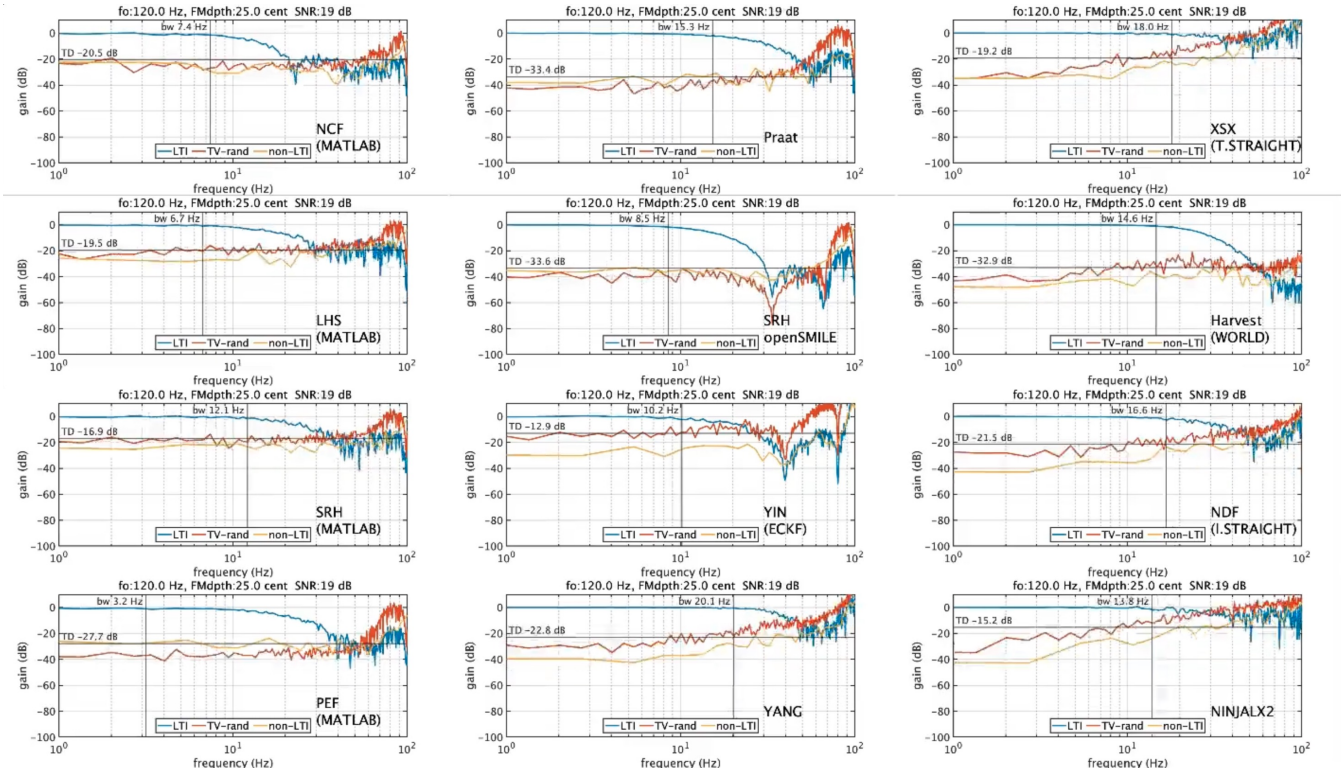

図 5 Snapshot of visualization movie. This movie shows SNR effects.

した場合の例を示す。ここでは、Praat や openSMILE が、耐雑音性を示している。

4. おわりに

本資料では，音声処理の重要なモジュールであるピッチ抽出器の物理特性を，著者らが開発した信号を用いて測定する方法を説明した．ここで紹介した方法を用いることにより，ピッチ抽出器を周波数変調の復調器として捉えた場合，変調周波数伝達特性に加え，信号依存応答，ランダム応答を求めることができる．これまでに提案されたピッチ抽出器をここで紹介した方法を用いて分析することにより，様々なピッチ抽出器の特徴を一覧することができるようになった．分析のために用いた方法は MATLAB を用いて実装されており，オープンソースとして公開している．また，筆頭著者の YouTube Channel には，様々なピッチ抽出器の動作を比較したムービーを置いた．

謝辞 本研究は，科学研究費基盤研究 (C)18K00147, 18K10708，挑戦的研究（萌芽）19K21618，基盤研究 (B)21H03468, 21H00497，基盤研究 (A)21H04900 による．

参考文献

[1] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. and Kavukcuoglu, K.: WaveNet: A generative model for raw audio., *SSW*, Vol. 125, p. 2 (2016).

[2] 全　炳河：深層学習によるテキスト音声合成の飛躍的発展，電子情報通信学会誌， Vol. 105, No. 5, pp. 413–417 (2022).

[3] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S. and Liu, T.-Y.: NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality, *arXiv preprint arXiv:2205.04421* (2022).

[4] Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S. and Wolfe, J.: Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization, *J. Acoust. Soc. Am.*, Vol. 137, No. 5, pp. 3005–3007 (online), DOI: 10.1121/1.4919349 (2015).

[5] Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177 (1939).

[6] Titze, I. R.: Comparison of Fo Extraction Methods for High-Precision Voice Perturbation Measurements An Optimizer-Simulator for Phonosurgery View project, *Article in Journal of Speech and Hearing Research*, Vol. 36, pp. 1120–1133 (online), DOI: 10.1044/jshr.3606.1120 (1993).

[7] Kawahara, H., Matsui, T., Yatabe, K., Sakakibara, K.-I., Tsuzaki, M., Morise, M. and Irino, T.: Mixture of orthogonal sequences made from extended time-stretched pulses enables measurement of involuntary voice fundamental frequency response to pitch perturbation, *Proc. Interspeech*, pp. 3206–3210 (2021).

[8] Kawahara, H., Matsui, T., Yatabe, K., Sakakibara, K.-I., Tsuzaki, M., Morise, M. and Irino, T.: Implementation of Interactive Tools for Investigating Fundamental Frequency Response of Voiced Sounds to Auditory Stimulation, *Proc. APSIPA ASC*, pp. 897–903 (2021).

[9] 河原英紀，矢田部浩平，榊原健一，北村達也，坂野秀樹，

森勢将雅：音声の基本周波数の周波数変調に対するピッチ抽出法の線形・非線形・ランダム応答の同時測定について-拡張された時間伸長パルス系列の直交化の応用-，信学技報 SP2021-44， Vol. 121, No. 282, pp. 27–32 (2021).
[10] Kawahara, H. and Yatabe, K.: Cascaded all-pass filters with randomized center frequencies and phase polarity for acoustic and speech measurement and data augmentation, *Proc. ICASSP*, pp. 306–310 (2021).
[11] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Trans. Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).
[12] Morise, M.: Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals, *Proc. Interspeech*, pp. 2321–2325 (online), DOI: 10.21437/Interspeech.2017-68 (2017).
[13] Noll, A. M.: Cepstrum Pitch Determination, *J. Acoust. Soc. Am.*, Vol. 41, No. 2, pp. 293–309 (online), DOI: 10.1121/1.1910339 (1967).
[14] Atal, B. S.: Automatic speaker recognition based on pitch contours, *J. Acoust. Soc. Am.*, Vol. 52, No. 6B, pp. 1687–1697 (1972).
[15] Hermes, D. J.: Measurement of pitch by subharmonic summation, *J. Acoust. Soc. Am.*, Vol. 83, No. 1, pp. 257–264 (1988).
[16] Gonzalez, S. and Brookes, M.: PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise, *IEEE/ACM Trans. ASLP*, Vol. 22, No. 2, pp. 518–530 (online), DOI: 10.1109/TASLP.2013.2295918 (2014).
[17] Drugman, T. and Alwan, A.: Joint robust voicing detection and pitch estimation based on residual harmonics, *Proc. Interspeech*, pp. 1973–1976 (online), DOI: 10.21437/Interspeech.2011-519 (2011).
[18] Kim, J. W., Salamon, J., Li, P. and Bello, J. P.: CREPE: A convolutional representation for pitch estimation, *Proc. ICASSP*, pp. 161–165 (2018).
[19] de Cheveigné, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *J. Acoust. Soc. Am.*, Vol. 111, No. 4, pp. 1917–1930 (2002).
[20] Das, O., Smith III, J. O. and Chafe, C.: Improved Real-Time Monophonic Pitch Tracking with the Extended Complex Kalman Filter, *J. Audio Engineering Society*, Vol. 68, No. 1/2, pp. 78–86 (2020).
[21] Camacho, A.: SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech And Music, *Ph. D. Thesis, University of Florida* (2007).
[22] Camacho, A. and Harris, J. G.: A sawtooth waveform inspired pitch estimator for speech and music, *J. Acoust. Soc. Am.*, Vol. 124, No. 3, pp. 1638–1652 (online), DOI: 10.1121/1.2951592 (2008).
[23] Talkin, D. and Kleijn, W. B.: A robust algorithm for pitch tracking (RAPT), *Speech coding and synthesis*, Vol. 495, p. 518 (1995).
[24] Brucal, S. G. E., Africa, A. D. M. and Dadios, E. P.: Female voice recognition using artificial neural networks and MATLAB voicebox toolbox, *J. Telecommunication, Electronic and Computer Engineering*, Vol. 10, No. 1-4, pp. 133–138 (2018).
[25] Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *Proc. ICPhS*, Vol. 17, No. 1193, pp. 97–110 (1993).
[26] Eyben, F., Wöllmer, M. and Schuller, B.: openSMILE – The munich versatile and fast open-source audio feature extractor, *Proc. 18th ACM Multimedia*, pp. 1459–1462 (2010).
[27] Kawahara, H., de Cheveigne, A., Banno, H., Takahashi, T. and Irino, T.: Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT, *Proc. Interspeech*, pp. 537–540 (online), DOI: 10.21437/Interspeech.2005-335 (2005).
[28] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207 (1999).
[29] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, *Proc. ICASSP*, pp. 3933–3936 (online), DOI: 10.1109/ICASSP.2008.4518514 (2008).
[30] Kawahara, H.: Application of Time-frequency Representations of Aperiodicity and Instantaneous Frequency for Detailed Analysis of Filled Pauses, *Journal of the Phonetic Society of Japan*, Vol. 21, No. 3, pp. 63–73 (2017).