Vol.2018-SLP-125 No.18
2018/12/10

音声分析変換合成法STRAIGHTの再構築について

河原 英紀[1,a)]

概要：新しいスペクトル包絡計算法，新しい瞬時周波数および群遅延計算法，周波数領域 velvet noise による新しい混合音源に基づいて，20 年前に構想された音声分析変換合成法 STRAIGHT を再構築する。本報告では、背景とそれらの構成要素を紹介し、検討課題と今後の進め方について議論する。

Reformulation of speech analysis, modification, and resynthesis framework STRAIGHT

Kawahara Hideki[1,a)]

1. はじめに

1997 年に発表された legacy-STRAIGHT [1,2] は、当時としては高い品質と操作の柔軟性を併せ持っていたことから広く引用されて現在に至っている。しかし、コードには様々なアドホックな処理が十分な根拠に基づかないまま実装されており [3]、動作の理解と改良を困難なものとしていた。legacy-STRAIGHT の登場から 20 年を経た現在では、標本化定理の見直し [4,5] や深層学習・機械学習の急速な発展 [6] により、音声分析合成を取り巻く状況は大きく変化している [7–9]。ここでは、legacy-STRAIGHT の背景となっている音声の役割に関する議論 [10,11] を踏まえて、音声の分析変換合成系の再構成を試みる。

背景で紹介するように、様々な偶然も重なって、STRAIGHT を再構成する材料が揃ってきた。以下では、背景に続けて、それぞれの材料の内容を紹介し、再構成の構想と課題について説明する。

2. 背景

声門の開閉により生ずるパルス状の呼気流は、声道の形状を反映した過渡応答により修飾されて有声音となる。声門の周期的な開閉により、声道形状の時間変化は有声音の特性の時間変化に反映される。見方を変えると、有声音は、背景にある滑らかな時間周波数表現を、時間、周波数のそれぞれで離散的に標本化する仕組みであると解釈することができる。この離散化による影響を取り除くため、legacy-STRAIGHT では、基本周波数 (fo) に応じたサイズを有する相補的時間窓を用意し、TANDEM-STRAIGHT では、基本周期の半分の間隔を隔てて時間窓を配置している。これらにより、時間方向の影響を除いたのち、周波数方向の平滑化と逆フィルタ（実質的には consistent 標本化による処理 [4]）によって、滑らかな時間周波数表現を復元している [2,12]。また WORLD では、窓関数を巧妙に設計することで、これらを簡単な処理で置き換えている [13,14]。

しかし、これらの方法によって復元される表現は、聴覚的に影響の大きなスペクトルのピーク周辺において、物理的な生成機構に対応する関数からの逸脱が大きくなるという問題を有していた [15]。ここで、前述の離散化による影響を取り除く新しい方法 [16] を発見したことがきっかけとなって、この問題を軽減することができるようになった [17]。

音声の駆動音源は、有声音の周期的駆動が情報を標本化するだけではなく、それ自身がコミュニケーションのための重要な情報を担っている。有声音の最も重要なパラメタである基本周波数については、現在でも多くの方法が研究され続けている（例えば [3] 参照）。しかし、実際の音声では、有声音の開始／終了部分や、強い表現を伴う音声ではその他の部分でも、周期性から大きく逸脱した駆動が生じ

[1] 和歌山大学
Wakayama University, Wakayama 640–8510, Japan
[a)] kawahara@sys.wakayama-u.ac.jp

る [18,19]。基本周波数とこれらに加え、開閉の状態による声質の変化や声道の狭窄により生ずる乱流雑音などの全てが、音声によるコミュニケーションに利用される。

これまで、これらの音源情報については、周期性は主に基本周波数として、周期性からの逸脱については主に声門閉止などのイベントや jitter や shimmer などの付属的な属性として、研究が進められてきた [20–24]。ここでも、声帯音源モデルの閉じた数式でのアンチエリアシングを検討する過程で設計された余弦級数 [25,26] を用いることで、音源の周波数方向と時間方向のパラメタを一括して求めることができるようになった [27]。さらに、合成の際に用いる非周期的駆動信号についても、系統的にランダム性を利用した疎な ±1 のパルスによる velvet noise [28,29] に着想を得て、スペクトルに影響を与えずに時間-周波数領域での分布を制御できる FVN(Frequency domain Velvet Noise) を発明することができた [30]。これは STRAIGHT や WORLD では表現されていなかった、破裂音やクリック、声門の開閉に伴う雑音の適切な実装の手段となる [30,31]。

3. 再構成のための素材

ここでは、再構成のための素材となる a) 時間周波数表現の復元、b) 音源情報の分析、c) 合成における音源情報拡張について、それぞれ簡単に説明する。詳しくは、参考文献に挙げた資料と関連する実装による。

3.1 周期性による干渉を排除した表現

ここでは、STRAIGHT や WORLD と同様に、時間方向での干渉を排除したパワースペクトルの計算と、周波数方向での干渉を排除した後処理を組み合わせる。まず、適切な窓関数を用いることにより、短時間スペクトル $S(\omega)$ から時間方向での干渉を排除したパワースペクトル $P_{\mathrm{TIF}}(\omega)$ を次式で求める。

$$P_{\mathrm{TIF}}(\omega) = P(\omega) + c_f^2 P_a(\omega) \tag{1}$$

$$= |S(\omega)|^2 + c_f^2 \left| \frac{dS(\omega)}{d\omega} \right|^2, \tag{2}$$

なお、ω は、基本周波数で正規化した角周波数を表す。

短時間スペクトル $S(\omega)$ の計算には、次式で示した初期値を余弦関数の半周期分とし B-spline 関数と畳み込むことで得られる窓関数を用いる。

$$v^{(0)}(x) = \begin{cases} \cos(\pi x) & -\frac{1}{2} \le x \le \frac{1}{2} \\ 0, & |x| > \frac{1}{2} \end{cases} \tag{3}$$

$$v^{(n)}(x) = v^{(0)}(x) * \underbrace{\beta^0 * \beta^0 * \cdots * \beta^0(x)}_{n \text{ times}}, \tag{4}$$

なお、ここでは x は、基本周期で正規化した時間軸に対応する。この関数を用いる場合の係数 c_f^2 は、閉じた形で求めることができ、資料 [17] に具体的な数値が示されている。

こうして求められた時間方向での干渉を排除したパワースペクトル $P_{\mathrm{TIF}}(\omega)$ から、次式により、さらに周波数方向の干渉を排除したパワースペクトル $P^{(n)}_{\mathrm{TFIF}}(\omega)$ を得る。

$$P^{(n)}_{\mathrm{TFIF}}(\omega) = \frac{1}{\omega_{\mathrm{o}}} \int_{-\omega_{\mathrm{o}}/2}^{\omega_{\mathrm{o}}/2} P^{(n)}_{\mathrm{TIF}}(\omega + \nu) d\nu \tag{5}$$

$$= \frac{1}{\omega_{\mathrm{o}}} \left(U^{(n)}_{\mathrm{TIF}}(\omega + \omega_{\mathrm{o}}/2) - U^{(n)}_{\mathrm{TIF}}(\omega - \omega_{\mathrm{o}}/2) \right) \tag{6}$$

$$U^{(n)}_{\mathrm{TIF}}(\omega) \equiv \int_{-\omega_{\mathrm{o}}}^{\omega} P^{(n)}_{\mathrm{TIF}}(\nu) d\nu, \tag{7}$$

ここでは、式 6 により畳み込みを簡単に実装している。窓関数に用いる B-spline の次数 (n) を 2 とすることで、基本周波数の推定誤差に耐性のある計算法（相対誤差 ±5%で平均対数スペクトルの誤差が 0.1 dB 以下）となる [17]。

こうした処理により求められる表現の形状は、特に聴覚的に影響が大きなスペクトルピークの近傍において、音声生成過程の物理的から求められるものと大きく異なっている。資料 [17] では、LPC 分析により求められる格子型逆フィルタで等価し、干渉の排除された表現を得た後に補償量を復元することを提案した。このように処理のパイプラインを構成することで、求められるスペクトルの誤差は、STRAIGHT や WORLD の 1/2 から 2/3 になる [17]。

3.2 基本周波数とイベントの分析

ここでは、声帯音源モデルの閉じた数式でのアンチエリアシングを検討する過程で設計された余弦級数 [25,26] と、最近のメディア処理用の計算の性能向上が鍵となる。逆三角関数などを含むベクトル処理が、高速化したことにより、瞬時周波数と群遅延の計算に、位相 VOCODER で導入された Flanagan の式 [32] が不要となった [27]。

用いた余弦級数は、以下で定義される。

$$w_e(t; f_c, c_{\mathrm{mag}}) = \sum_{k=0}^{K} a_k \cos\left(\frac{2\pi k f_c t}{K c_{\mathrm{mag}}} \right), \tag{8}$$

ここで、$K = 5$ として以下の値を係数 $\{a_k\}_{k=0}^5$ に用いることにより、最大のサイドローブのレベルが −114 dB で、漸近的な傾斜が −54 dB/oct の特性が得られる。

$$\{a_k\}_{k=0}^5 = \{0.2624710164, 0.4265335164, 0.2250165621, 0.0726831633, 0.0125124215, 0.0007833203\} \tag{9}$$

この余弦級数を包絡として複素指数関数との積で定義される以下の解析信号をインパルス応答とするフィルタバンクを構成する。

$$w(t) = w_e(t; f_c, c_{\mathrm{mag}}) \exp\left(j 2\pi f_c t\right), \tag{10}$$

ここで f_c は、フィルタの中心周波数、$c_{\mathrm{mag}} \le 1$ は、調整用の窓長の伸長係数を表す。このフィルタを通過した出力の瞬時周波数には、Hamming 窓、Blackman 窓、定義域が有限の場合に最小の時間周波数積を有する偏長楕円体波動

関数、その近似である Kaiser 窓を用いた場合に生ずる異常値は出現しない [26]。

このフィルタ群の出力の瞬時周波数 $\omega_i[n]$ と群遅延 $\tau_g[k]$ は、ほぼ定義通りに求められる。これらの値は、離散化されたフィルタ出力 $x[n]$ の標本毎に求めることができるため、計算をフレーム毎に行う必要はない。まず、瞬時周波数は次式による。

$$\omega_i[n] = \angle \left[\frac{x[n+1]}{x[n]} f_s \right], \tag{11}$$

ここで、n は離散時刻（音声信号の標本毎）を表す。k を、フィルタ番号を表す離散周波数とした時に、ある時刻における隣接するフィルタ出力を $X[k]$ と $X[k+1]$ とすると、群遅延は次式による。

$$\tau_g[k] = -\frac{1}{\Delta\omega} \angle \left[\frac{X[k+1]}{X[k]} \right], \tag{12}$$

ここで $\Delta\omega$ は、隣接する中心（角）周波数の差を表す。

瞬時周波数 $\omega_i[n]$ と群遅延 $\tau_g[k]$ の変化 $\Delta, \Delta\Delta$ の大きさは、信号が正弦波の場合には 0 となるため、基本周波数およびイベントの選択の指標となる。基本周波数の候補は、中心周波数から瞬時周波数への写像の不動点 [33]、イベントの候補は、離散時刻から（群遅延と出力のパワーから求められる）平均時刻への写像の不動点 [34] として求められる。なお、基本周波数候補の選択の指標として、フィルタのインパルス応答の包絡の時間長を操作して得られる二つのフィルタの出力 $y_N[n;k], y_W[n;k]$ から計算される $\sigma_{\mathrm{res}}[n;k]$ を加えることができる [27, 35]。

$$\sigma_{\mathrm{res}}[n;k] = \left| \frac{y_W[n;k]}{|y_W[n;k]|} - \frac{y_N[n;k]}{|y_N[n;k]|} \right|, \tag{13}$$

なお、$y_N[n;k]$ は長い時間長のフィルタ出力、$y_W[n;k]$ は短い時間長のフィルタ出力を表す。文献 [35] では、時間長の比を 2 としている。

3.3 駆動音源としての FVN の利用

ある規則でランダムに配置された ± 1 の疎なパルス列である velvet noise は、ほぼ 3,000 パルス／秒以上の密度の場合、白色の正規雑音よりも滑らかな雑音として知覚される [28, 29]。帯域通過フィルタによって帯域を制限した velvet noise の離散周波数軸上での信号は、局在する正規雑音とみなすことができる。FVN (Frequency domain Velvet Noise) は、この関係に注目し、時間軸と周波数軸を入れ替え、複素指数関数の位相のみを、前述の余弦関数を用いて操作することで生成される [30]。

具体的には、離散周波数軸を平均周波数間隔 F_d で分割し、それぞれの区間内での m 番目の離散周波数の位置 $k_{\mathrm{fvn}}(m)$ を、一様乱数の系列 $r_1(m)$ によって決める。

$$k_{\mathrm{fvn}}(m) = ||mF_d + r_1(m)(F_d - 1)||, \tag{14}$$

さらに、その位置に配置する位相操作関数 $w_p(k-k_c, B)$ の係数 $s_{\mathrm{fvn}}(m)$ を、もう一つの一様乱数の系列 $r_2(m)$ によって決める。

$$\varphi_{\mathrm{fvn}}(k) = \sum_{k_c \in \mathbb{K}} s_{\mathrm{fvn}}(k_c) \left(w_p(k-k_c, B) - w_p(k+k_c, B) \right), \tag{15}$$

$$s_{\mathrm{fvn}}(m) = (2||r_2(m)|| - 1)\, \varphi_{\max} \tag{16}$$

なお、ここで $\varphi_{\max}$ は、一個の位相操作関数の大きさを表す。

こうして求められた位相特性 $\varphi_{\mathrm{fvn}}(k)$ の和を以下により逆 Fourier 変換して、離散時間軸上での FVN を得る。

$$h_{\mathrm{fvn}}(n) = \frac{1}{K} \sum_{k=0}^{K-1} \exp\left(\frac{2kn\pi j}{KN} + j\varphi_{\mathrm{fvn}}(k) \right). \tag{17}$$

同一の乱数系列から作られた FVN を離散時間軸上に等間隔に配置すると、周期信号が得られる。これを Frozen FVN による系列とする。毎回、乱数を更新して FVN を作成して配置することにすると、同じ位置に FVN を配置しても雑音として聞こえる。これを Random FVN による系列とする。ここで、各配置毎に Frozen FVN の作成に用いられる位相特性と、Random FVN の作成に用いられる位相特性を内挿した位相特性を用いて FVN を作成することにより、雑音から周期信号まで、連続的に変化させることのできる信号を作ることができる。さらに、FVN の周波数毎の時間方向での広がりを、周波数軸の非線形伸縮により設計することができる。これらの操作を利用することにより、柔軟な操作が可能な合成音声の駆動音源とすることができる [30]。

FVN は、位相特性のみが周波数依存性を有する all-pass filter のインパルス応答でもある。したがって、短い持続時間の FVN であっても、パワースペクトルは周波数に依存しない。正規雑音を駆動音源の雑音として用いる場合、持続時間を短くすると、統計的な揺らぎにより、パワースペクトルの周波数特性は大きく変動し、合成音声のパワースペクトルを変化させてしまう。FVN を駆動音源の雑音として用いることで、この問題を回避することができる。

FVN は、基本周波数により振幅変調を受ける雑音源の実装に用いることもできる。この性質は、低いピッチの音声での影響が大きい時間マスキング [31] の調整に有用である。さらに、一個の FVN を時変フィルタとして VOCODER の後処理に用いることで、従来の VOCODER で問題となる buzz 感を簡単に取り除くことができる [30]。異なる乱数により作成される FVN は、互いにほぼ直交するため、この後処理を、処理音声であることを証明（署名）するために応用できる可能性がある。FVN の詳しい性質などは、文献 [30] への付録を付加した資料 [36] に紹介した。

4. 課題

End-to-End の手法に基づく合成音声の幾つかは、既に

自然音声に匹敵する品質を実現している [6,37,38]。品質を目的とするのであれば、STRAIGHT を再構築する必要はない。しかし音声コミュニケーションの各過程を理解し、診断や訓練に応用する際には有用である可能性はある。また、リアルタイム性が重要となるライブなどでの応用でも、軽く遅延時間の少ない実装とすれば、有用である可能性はある。それらと、現在では予想できない応用を期待して、紹介した要素をさらに検討し、次世代の STRAIGHT として再構築を進めることとしたい。

ただし、Alpha-Go から Alpha-Go master, Alpha-Go Zero とそれ以降の経緯 [39] を見ると、これまでの音声科学の蓄積に基づく STRAIGHT の構成と背景とする理解は、制約された研究手段と非常に少ないデータに基づいて形成された不適切なものである可能性がある [9]。この可能性を常に意識し、必要に応じて戦略を更新することとしたい。

5. おわりに

legacy-STRAIGHT の背景となった音声の役割の理解の下で、現在の技術に基づいて STRAIGHT を再構成する構想と、幾つかの予備的な検討について説明した。ここで紹介した構成要素は、legacy-STRAIGHT [40] や、音声生成／知覚の教育・研究用ツールである SparkNG [41] とともに、オープンソースとして公開する予定である。また、最終的に再構築された STRAIGHT も、同様に公開することとしたい。

謝辞 本研究は、科研費基盤 (A)16H01734 と、科研費基盤 (B)15H03207 の支援を受けた。

参考文献

[1] Kawahara, H.: Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited, *Proc. ICASSP 1997*, Vol. 2, pp. 1303–1306 (1997).
[2] 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).
[3] 森勢将雅: 音声分析合成，コロナ社 (2018).
[4] Unser, M.: Sampling-50 years after Shannon, *Proceedings of the IEEE*, Vol. 88, No. 4, pp. 569–587 (2000).
[5] 田中利幸: 圧縮センシングの数理，電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review， Vol. 4, No. 1, pp. 39–47 (2010).
[6] 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, *CoRR*, Vol. abs/1609.03499 (2016).
[7] 戸田智基，小林和弘: 統計的声質変換ソフトウェア入門，システム／制御／情報，Vol. 62, No. 2, pp. 69–75 (2018).
[8] 全 炳河: テキスト音声合成技術の変遷と最先端，日本音響学会誌， Vol. 74, No. 7, pp. 387–393 (2018).
[9] 河原英紀: また性懲りもなく VOCODER の話を：大人になるのはもったいない，日本音響学会誌，Vol. 74, No. 10, pp. 545–546 (2018).
[10] Irino, T. and Patterson, R. D.: Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform, *Speech Communication*, Vol. 36, No. 3, pp. 181 – 203 (2002).
[11] 河原英紀: Vocoder のもう一つの可能性を探る：音声分析変換合成システム STRAIGHT の背景と展開，日本音響学会誌， Vol. 63, No. 8, pp. 442–449 (2007).
[12] 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, *ICASSP 2008*, Las Vegas, pp. 3933–3936 (2008).
[13] Morise, M.: CheapTrick, a spectral envelope estimator for high-quality speech synthesis, *Speech Communication*, Vol. 67, pp. 1 – 7 (2015).
[14] Morise, M., Yokomori, F. and Ozawa, K.: 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).
[15] Kawahara, H., Morise, M., Toda, T., Nisimura, R. and Irino, T.: Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds., *Proc. Interspeech 2013*, Lyon, pp. 1921–1925 (2013).
[16] 河原英紀，森勢将雅，フアカンル: 周期信号の静的表現の VOCODER への応用について，聴覚研究会資料，Vol. 48, No. 5, pp. 429–434 (2018).
[17] Kawahara, H., Morise, M. and Hua, K.: Revisiting spectral envelope recovery from speech sounds generated by periodic excitation, *Proc. APSIPA ASC 2018*, Hawaii (2018). [Accepted].
[18] 榊原健一: 発声と声帯振動の基礎，日本音響学会誌，Vol. 71, No. 2, pp. 73–79 (2015).
[19] KAWAHARA, H.: 非周期性と瞬時周波数の時間-周波数表現の Filled pause の詳細な分析への応用について，音声研究， Vol. 21, No. 3, pp. 63–73 (2017).
[20] Naylor, P. A., Kounoudes, A., Gudnason, J. and Brookes, M.: Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 15, No. 1 (2007).
[21] Alku, P.: Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications, *Sadhana - Academy Proceedings in Engineering Sciences*, Vol. 36, No. October, pp. 623–650 (2011).
[22] Thomas, M. R. P., Gudnason, J. and Naylor, P. A.: Detection of glottal opening and closing instants in voiced speech using the YAGA algorithm, *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 20, pp. 82–91 (2012).
[23] Khanagha, V., Daoudi, K. and Yahia, H. M.: Detection of glottal closure instants based on the microcanonical multiscale formalism, *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 22, No. 12, pp. 1941–1950 (2014).
[24] Gangamohan, P. and Yegnanarayana, B.: A robust and alternative approach to zero frequency filtering method for epoch extraction, *Proc. Interspeech 2017*, pp. 2297–2300 (2017).
[25] Kawahara, H., Sakakibara, K.-I., Morise, M., Banno, H.,

Toda, T. and Irino, T.: A new cosine series antialiasing function and its application to aliasing-free glottal aource models for speech and singing synthesis, *Proc. Interspeech 2017*, pp. 1358–1362 (2017).
[26] 河原英紀：ディジタル信号処理の落とし穴，日本音響学会誌， Vol. 73, No. 9, pp. 592–599 (2017).
[27] KAWAHARA, H., SAKAKIBARA, K.-I., BANNO, H. and MORISE, M.: An analytic signal with a cosine series envelope for excitation source analyses of voiced speech – Practical substitute of Flanagan's equation –, 信学技報， Vol. 118, No. 193 (2018). [in Print].
[28] Järveläinen, H. and Karjalainen, M.: Reverberation modeling using velvet noise, *AES 30th International Conference, Saariselkä, Finland*, Audio Engineering Society,, pp. 15–17 (2007).
[29] Välimäki, V., Lehtonen, H. M. and Takanen, M.: A perceptual study on velvet noise and its variants at different pulse densities, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 7, pp. 1481–1488 (2013).
[30] Kawahara, H., Sakakibara, K.-I., Morise, M., Banno, H., Tomoki, T. and Irino, T.: Frequency domain variants of velvet noise and their application to speech processing and synthesis, *Proc. Interspeech 2018*, Hyderabad India, pp. 2027–2031 (2018).
[31] Skoglund, J. and Kleijn, W. B.: On time-frequency masking in voiced speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 4, pp. 361–369 (2000).
[32] Flanagan, J. L. and Golden, R. M.: Phase Vocoder, *Bell System Technical Journal*, Vol. 45, No. 9, pp. 1493–1509 (online), DOI: 10.1002/j.1538-7305.1966.tb01706.x (1966).
[33] Kawahara, H., Katayose, H., de Cheveigné, A. and Patterson, R. D.: Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity., *Proc. Eurospeech 99*, Budapest, Hungary, pp. 2781–2784 (1999).
[34] Kawahara, H., Atake, Y. and Zolfaghari, P.: Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay., *Icslp-2000*, Vol. 4, Beijing, China, pp. 664–667 (2000).
[35] Kawahara, H., Agiomyrgiannakis, Y. and Zen, H.: Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis, *9th ISCA Speech Synthesis Workshop*, pp. 221–228 (2016).
[36] Kawahara, H., Sakakibara, K.-I., Morise, M., Banno, H., Toda, T. and Irino, T.: Frequency domain variants of velvet noise and their application to speech processing and synthesis: with appendices, *arXiv preprint arXiv:1806.06812* (2018).
[37] van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D. and Hassabis, D.: Parallel WaveNet: Fast High-Fidelity Speech Synthesis, *CoRR*, Vol. abs/1711.10433 (2017).
[38] Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., Cao, Y., Jia, Y., Chen, Z., Shen, J. et al.: Hierarchical Generative Modeling for Controllable Speech Synthesis, *arXiv preprint arXiv:1810.07217* (2018).
[39] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al.: Mastering the game of Go without human knowledge, *Nature*, Vol. 550, No. 7676, p. 354 (2017).
[40] Kawahara, H.: Legacy STRAIGHT, *GitHub*, (online), available from ⟨https://github.com/HidekiKawahara/legacy_STRAIGHT⟩ (accessed: 8/Nov./2018).
[41] Kawahara, H.: SparkNG, *GitHub*, (online), available from ⟨https://github.com/HidekiKawahara/SparkNG⟩ (accessed: 8/Nov./2018).