Harmonic-Net++: 基本周波数制御可能なメルスペクトログラム入力型高速ニューラルボコーダ *

☆清水聡太 [1,2]，岡本拓磨[2]，高島遼一[1]，滝口哲也[1]，戸田智基[3,2]，河井恒[2]
[1] 神戸大学，[2] 情報通信研究機構，[3] 名古屋大学

1 はじめに

近年，深層学習 (ニューラルネットワーク) を用いた音声合成の手法が盛んに研究されており，自然音声に近い高品質な音声を合成できるようになっている [1, 2]。音響特徴量から音声波形を再構成するニューラルボコーダは，従来のソースフィルタボコーダ [3] に比べ，合成音声の品質を大幅に向上させており，高速かつ高品質なニューラルボコーダが数多く提案されている [4, 5]。

ニューラルボコーダは従来のソースフィルタボコーダと同様に，基本周波数 (f_{o}) などの属性を柔軟に制御することが必要とされる。しかし多くのニューラルボコーダはデータ駆動型であるため，f_{o} の制御性能はソースフィルタボコーダに劣るのが一般的である。この問題を解決するために，いくつかのアプローチが提案されている [6, 7]。これらの手法では，高い f_{o} の制御性能を達成しているが，HiFi-GAN [8] のような純粋なデータ駆動型のボコーダと比較して合成品質が低い傾向がある。また，これらの多くは大規模な畳み込み層で構成されており，リアルタイム合成にはハイエンドな GPU が必要である。

f_{o} の制御性能を維持しつつ，高品質かつ CPU のみでもリアルタイム合成可能なモデルとして Harmonic-Net+ [9] が提案されている。高速かつ高品質なニューラルボコーダとして提案されている HiFi-GAN に対して，f_{o} に対応する励起信号およびその調波成分を入力するネットワークと，HiFi-GAN のようなアップサンプリング型モデルに対応したピッチ依存型拡張畳み込みネットワーク (Layerwise-Pitch-dependent dilated convolutional neural network: LW-PDCNN) を導入することで，高速かつ高品質な合成が可能となっている。しかし，LW-PDCNN の導入により，合成速度は元の HiFi-GAN の 2 倍近く遅くなっている。また，WORLD 特徴量の抽出には時間がかかるため，Harmonic-Net+は WORLD 特徴量の抽出を含むリアルタイム合成が可能でない。

本研究では，合成品質を維持したまま Harmonic-Net+の合成速度を向上させるため，WORLD 特徴量を入力として用いず，メルスペクトログラムを入力とする Harmonic-Net++を提案する。メルスペクトログラムはニューラルボコーダの入力特徴量として広く用いられており [10]，抽出時間は WORLD 特徴量と比べ極めて速い。Harmonic-Net++では，前処理ネットワークがメルスペクトログラムから WORLD 特徴量を推定し，事前学習した Harmonic-Net+が f_{o} の制御性能を維持したまま，高品質な合成を実現する。加えて，合成品質を維持しながらより高速な合成を実現するため，[11] で提案された，高速なアップサンプリング層を用いた Multi-stream (MS)-Harmonic-Net++を提案する。

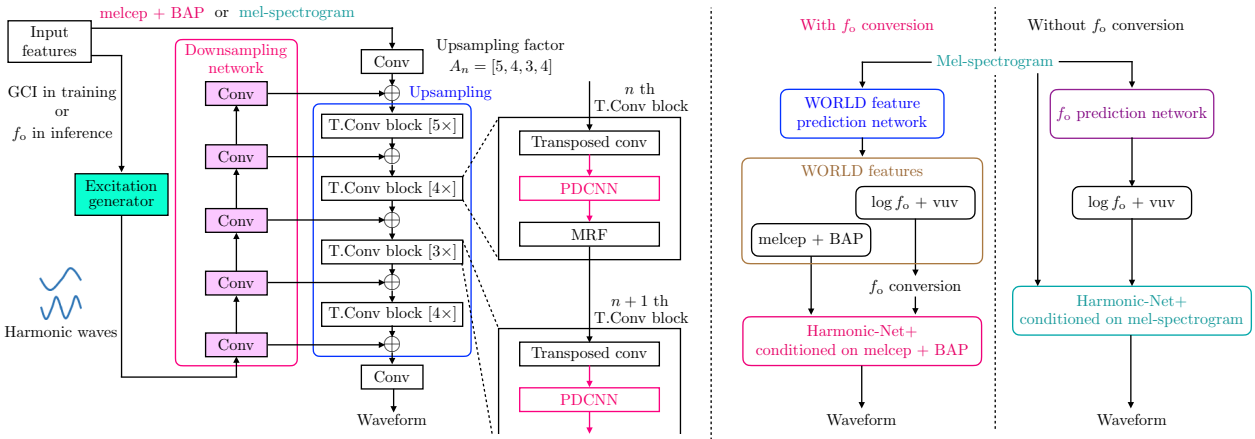
さらに，人間的には無理のある過剰な制御倍率 (0.5 や 1.5) ではなく，実際にありうる範囲で声の高さを変えて収録した PitchSpeech での評価も行った。PitchSpeech は ITA コーパス[1] 324 文を異なる話速で読み上げた SpeedSpeech [2] のパラレルとして収録し，音声合成研究の中でも特に f_{o} 制御に関する研究の促進を目的として，NICT から公開予定である。

2 Harmonic-Net+

Harmonic-Net+ (Fig. 1(a)) は HiFi-GAN に 2 つのネットワークを導入している。一つは f_{o} に対応する励起信号を階層的に受け取るダウンサンプリングネットワークである。励振信号を明示的に入力することで，PeriodNet[6] と同様にスケーリングした f_{o} を入力に用いた場合の合成品質を向上させる。もう一つは f_{o} に依存した受容野を持つ LW-PDCNN である。PDCNN は，音声波形のサンプリング周波数を時間分解能とする CNN ベースのニューラルボコーダ [12, 13] に向けて提案されたもので，層ごとに時間分解能の異なる HiFi-GAN にそのまま適用することができない。そこでアップサンプリング型モデルである HiFi-GAN に PDCNN を適用した LW-PDCNN が提案された [9]。LW-PDCNN により，f_{o} の揺らぎをモデル構造に取り込むことができ，スケーリングした f_{o} を入力に用いた場合の合成品質をさらに向上させる。

[1] https://github.com/mmorise/ita-corpus
[2] https://ast-astrec.nict.go.jp/release/speedspeech_ja_2022/download.html

* Harmonic-Net++: Fundamental frequency controllable fast neural vocoder with mel-spectrogram input. by SHIMIZU, Sota[1,2], OKAMOTO, Takuma[2], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2] and KAWAI, Hisashi[2] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)



(a) Harmonic-Net+ generator network　　(b) Proposed Harmonic-Net++

Fig. 1　Network architectures of (a) Harmonic-Net+ generator and (b) Harmonic-Net++.

3　提案手法

話者性を変化させずに f_o 制御を実現するためには，メルスペクトログラムを f_o とスペクトル包絡に分離し，WORLD 特徴量に変換することが必要である。メルスペクトログラムは短時間フーリエとメルフィルタバンクで計算可能なため，抽出時間は WORLD 特徴量に比べ高速である。そこで，入力のメルスペクトログラムから WORLD 特徴量を高速に推定する前処理ネットワークを導入した。Fig. 1(b) に，Harmonic-Net+ をベースにした提案手法である Harmonic-Net++の概要を示す。前処理ネットワークは畳み込み層と HiFi-GAN ベースの multiple receptive field (MRF) から構成される。Harmonic-Net++では，WORLD 特徴量 (f_o, 有声無声区間情報 (voiced/unvoiced vector: vuv), メルケプストラム (mel-cepstra: melcep), 非周期性指標 (binary-coded aperiodicity components: BAP)) またはメルスペクトログラム，f_o，vuv を条件とした 2 つの Harmonic-Net+を学習させる。予備実験で f_o 制御を行わない場合は前者と比較して後者の合成品質が同等以上であったため，前者を f_o 制御を行う場合，後者を f_o 制御を行わない場合に用いる。前処理ネットワークに関しても WORLD 特徴量を推定するものと，f_o と vuv のみを推定するものの 2 種類を学習させる。推定された特徴量を用いて高品質な合成を実現するため，Harmonic-Net+は推定された特徴量を用いてファインチューニングされる。

また，Harmonic-Net++の合成速度をさらに向上させるため，Multi-stream HiFi-GAN [11] で用いられる，高速な 4 倍アップサンプリング層を最終層に導入した MS-Harmonic-Net++を提案する。

これらの提案モデルにより，WORLD 特徴量の抽出を行わずメルスペクトログラム入力のみで，f_o 制御かつ特徴量抽出を含めたリアルタイム合成が可能である。

4　実験

4.1　実験条件

提案手法の性能を評価するため，サンプリング周波数 24 kHz の音声を用いた未知話者合成での評価実験を行った。データセットは JVS コーパス [14] より，100 名の日本人話者による各話者 130 文の音声を用いた。学習には 96 名 (jvs005〜jvs100) の 12477 文を学習に用い，残り 4 名 (jvs001〜jvs004) のノンパラレル 120 文を評価に用いた。f_o の制御倍率を 1.0 倍，0.5 倍および 1.5 倍の条件で評価した。さらに PitchSpeech を用いた，実際の低・高 f_o 制御倍率における評価も行った。PitchSpeech において，低・高 f_o 制御倍率での f_o の平均値を通常発話と比較すると，男性で 0.69 倍と 1.37 倍，女性で 0.87 倍と 1.19 倍となった。評価には，男女各 1 名の 20 発話 (RECITATION324_001〜020) を用いた。

入力特徴量には WORLD 特徴量と 80 次元ログメルスペクトログラムの 2 種類を用いた。WORLD 特徴量には 50 次元メルケプストラム，3 次元非周期性指標および対数連続 f_o を用いた。いずれも WORLD [3] を用いて窓長とフレームシフトを 42.7 ms と 10 ms に設定して抽出を行った。メルスペクトログラムの抽出は WORLD 特徴量と同様に窓長とフレームシフトを 42.7ms と 10ms に設定して行った。

比較対象には WORLD，HiFi-GAN，HN-uSFGAN，および Harmonic-Net+を [15] と同条件で用いた。Harmonic-Net+では，Harvest [16] を用いた f_o 抽出 (WORLD [Harvest]) に加え，DIO [3] を用いた f_o 抽出 (WORLD [DIO]) も行い，特徴量

抽出にかかる時間について評価を行った。

提案手法である Harmonic-Net++の実装は，Harmonic-Net+をベースに，畳み込み層 (カーネルサイズ 7，チャネル数 512)，MRF，畳み込み層 (カーネルサイズ 1，チャネル数 55 または 2) の順からなる前処理ネットワークを追加し，それぞれ WORLD 特徴量または f_0 と vuv を推定するものとした。Harmonic-Net++の入力特徴量としては，[8] で用いられている 80 次元ログメルスペクトログラムを用いた。MS-Harmonic-Net++ (+MS) は，[11] を元に 4 層目の高速な 4 倍アップサンプリング層を実装した。

4.2 実験結果

Table 1 に合成時の real-time factor (RTF) を示す。RTF は Intel Xeon 6152 CPU (1 コア) を用いて計測した。提案手法の Harmonic-Net++では特徴量抽出を含むリアルタイム合成を達成した。MS-Harmonic-Net++ではさらに高速なリアルタイム合成を実現した。また，DIO による特徴量抽出は Harvest よりも高速であったが，f_0 の制御性能は Harvest に比べ低いことが確認された。

聴取実験による平均オピニオン評定 (MOS) の結果を Table 2 に示す。被験者は 20 人でヘッドホン聴取により評価した。提案手法である Harmonic-Net++は Harmonic-Net+と同等の品質を達成し，多くの条件で HN-uSFGAN を大きく上回ることが示された。Harmonic-Net+と Harmonic-Net++は他のモデルと比較し高品質だが，PitchSpeech を用いた実験では，実際に低い声や高い声で話した原音には劣っていた。そのため合成品質をさらに向上させるには，実際に低い声や高い声で話した原音との比較が重要である。

最後に MS-Harmonic-Net++，Harmonic-Net++，HN-uSFGAN の合成品質を比較するため，一対比較実験を行った。被験者は 20 人でヘッドホン聴取により，通常の制御倍率 ($1.0 \times f_0$)，低・高 f_0 制御倍率の 3 条件で評価を行った。Table 3 に一対比較実験の結果を示す。MS-Harmonic-Net++は Harmonic-Net+との比較ではどちらでもない (Neutral) に多く投票されており，合成品質と f_0 の制御性能を維持したまま高速な合成を実現できることが示され，提案手法である MS-Harmonic-Net++の有効性を確認した。

5 おわりに

Harmonic-Net+の合成品質と f_0 の制御性能を維持したまま合成速度を向上させるため，メルスペクトログラムを入力とする Harmonic-Net++を提案した。実験結果により提案手法の有効性を確認した。

Table 1 Results of real time factors for feature extraction and synthesis with an Intel Xeon 6152 CPU.

Model	Feature extraction	Synthesis	Total
WORLD	WORLD [Harvest]	0.12	0.71
HiFi-GAN	WORLD [Harvest]	0.31	0.90
HN-uSFGAN	WORLD [Harvest]	3.67	4.26
Harmonic-Net+	WORLD [Harvest]	0.65	1.24
Harmonic-Net+	WORLD [DIO]	0.65	0.88
Harmonic-Net++	mel-spectrogram	0.75	0.75
+MS	mel-spectrogram	0.44	**0.44**

参考文献

[1] J. Kim *et al.*, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, July 2021, pp. 5530–5540.

[2] D. Lim *et al.*, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, Sept. 2022, pp. 21–25.

[3] M. Morise *et al.*, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, July 2016.

[4] T. Okamoto *et al.*, "Noise level limited sub-modeling for diffusion probabilistic vocoders," in *Proc. ICASSP*, June 2021, pp. 6014–6018.

[5] Y. Koizumi *et al.*, "WaveFit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," in *Proc. SLT*, Jan. 2023, pp. 884–891.

[6] Y. Hono *et al.*, "PeriodNet: A non-autoregressive raw waveform generative model with a structure separating periodic and aperiodic components," *IEEE Access*, vol. 9, pp. 137 599–137 612, 2021.

[7] R. Yoneyama *et al.*, "Unified source-filter GAN with harmonic-plus-noise source excitation generation," in *Proc. Interspeech*, Sept. 2022, pp. 848–852.

[8] J. Kong *et al.*, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17 022–17 033.

Table 2 Results of MOS tests in normal and f_o conversion conditions. Confidence level of the error bars is 95%.

JVS						
	$1.0 \times f_o$		$0.5 \times f_o$		$1.5 \times f_o$	
Model	male	female	male	female	male	female
Original	4.80 ± 0.12	4.59 ± 0.18	-	-	-	-
WORLD	-	-	3.36 ± 0.21	3.02 ± 0.23	3.72 ± 0.25	3.05 ± 0.26
HN-uSFGAN	3.58 ± 0.26	3.56 ± 0.21	3.17 ± 0.21	3.05 ± 0.15	3.52 ± 0.28	**3.44 ± 0.25**
HiFi-GAN	3.69 ± 0.33	3.57 ± 0.23	-	-	-	-
Harmonic-Net+	4.39 ± 0.21	**4.27 ± 0.11**	3.64 ± 0.32	**3.68 ± 0.20**	4.23 ± 0.28	**3.44 ± 0.25**
Harmonic-Net++	**4.49 ± 0.15**	4.25 ± 0.17	**3.77 ± 0.21**	3.53 ± 0.21	**4.30 ± 0.13**	3.28 ± 0.21
PitchSpeech						
	normal		low		high	
Model	male	female	male	female	male	female
Original	4.68 ± 0.12	4.55 ± 0.18	4.51 ± 0.14	4.74 ± 0.15	4.67 ± 0.12	4.66 ± 0.13
WORLD	-	-	3.87 ± 0.19	2.83 ± 0.21	3.67 ± 0.28	3.36 ± 0.20
HN-uSFGAN	3.73 ± 0.26	2.49 ± 0.23	3.66 ± 0.27	2.38 ± 0.16	3.76 ± 0.24	2.85 ± 0.24
HiFi-GAN	2.82 ± 0.23	2.94 ± 0.23	-	-	-	-
Harmonic-Net+	**4.46 ± 0.17**	3.76 ± 0.18	**4.27 ± 0.18**	**3.49 ± 0.20**	**4.31 ± 0.15**	**3.92 ± 0.23**
Harmonic-Net++	4.37 ± 0.21	**4.06 ± 0.19**	3.44 ± 0.24	3.30 ± 0.27	4.16 ± 0.19	3.81 ± 0.17

Table 3 Results of paired comparison tests in normal and f_o conversion conditions with (A): MS-Harmonic-Net++, (B): Harmonic-Net++, and (C): HN-uSFGAN.

Condition	A	B	Neutral	p-value
normal	155	67	**178**	8.964×10^{-10}
low	124	91	**185**	0.024
high	109	95	**196**	0.327
Condition	A	C	Neutral	p-value
normal	**321**	20	59	2.256×10^{-113}
low	**252**	63	85	3.320×10^{-32}
high	**231**	93	76	6.902×10^{-16}

[9] 松原ら，“Harmonic-Net+：高調波入力と Layerwise-Quasi-Periodic 畳み込みを用いた基本周波数制御可能な高速ニューラルボコーダ”，音講論，pp. 1133–1136，Sept. 2022.

[10] J. Shen *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.

[11] T. Okamoto *et al.*, “Multi-stream HiFi-GAN with data-driven waveform decomposition,” in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[12] Y.-C. Wu *et al.*, “Quasi-periodic WaveNet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1134–1148, 2021.

[13] ——, “Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 792–806, 2021.

[14] S. Takamichi *et al.*, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[15] 清水ら，“基本周波数制御可能なメルスペクトログラム入力型 HiFi-GAN の初期検討”，音講論，pp. 1137–1140，Sept. 2022.

[16] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.