End-to-End Neural Formant Synthesis Using Low-Dimensional Acoustic Parameters

Sumiharu Kobayashi
Graduate School of Science and Engineering
Yamagata University
Yonezawa, Japan
t232525m@st.yamagata-u.ac.jp

Tetsuo Kosaka
Graduate School of Sci. & Eng.
Yamagata University
Yonezawa, Japan

Tasashi Nose
Graduate School of Engineering
Tohoku University
Sendai, Japan

***Abstract*—Neural vocoders can synthesize high-quality speech waveforms from acoustic features, but they cannot control by acoustic parameters, such as F_0 and formant frequencies. Although analysis-synthesis based on signal processing can be controlled using acoustic parameters, its speech quality is inferior to that of neural vocoders. This paper proposes *End-to-End Neural Formant Synthesis* for generating high-quality speech waveforms with controllable acoustic parameters from low-dimensional representations. We compared three models with different structures, and investigated their synthesis quality and controllability. Experimental results showed that the proposed method performed as well as or better than conventional methods in terms of speech quality and controllability.**

***Index Terms*—speech synthesis, formant synthesis, neural vocoder, generative adversarial networks.**

I. Introduction

A neural vocoder is a deep-learning-based waveform generation method widely used in text-to-speech and voice conversion. Recent neural vocoders, such as HiFi-GAN [1], prioritize preserving speaker similarity and enhancing the naturalness of synthesized speech.

HiFi-GAN generates high-quality speech waveforms from high-dimensional acoustic features, such as mel-spectrogram. However, HiFi-GAN cannot control the phonetic properties of speech using low-dimensional acoustic parameters, such as pitch and formant frequencies. Controlling these parameters is crucial for achieving diverse voice expressions in practical applications. On the other hand, analysis-synthesis methods based on signal processing, such as formant synthesis, can control the phonetic properties of speech using low-dimensional acoustic parameters. These methods are less prevalent today due to limitations in synthesized speech quality and speaker similarity compared to neural vocoders.

To address this issue, *Neural Formant Synthesis* [2] which is a deep-learning-based formant synthesis method, has been proposed. This method employs a feature-mapping network to convert acoustic parameters into acoustic features, subsequently synthesizing speech waveforms using a neural vocoder.

The separation of the feature-mapping network and vocoder raises concerns about potential synthetic speech quality degradation. Therefore, this study investigated end-to-end speech synthesis using low-dimensional acoustic parameters as input to a neural vocoder to replace conventional acoustic features. Furthermore, we aimed to enhance the controllability of acoustic parameters using a source-filter model-based structure. Our method achieved high speech quality while maintaining controllability comparable to or superior to that of the conventional methods.

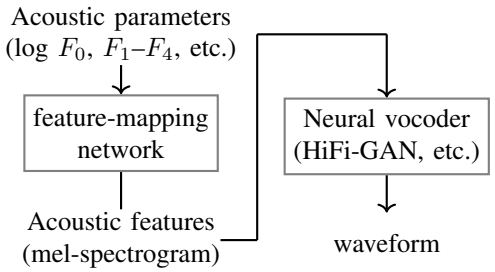

Fig. 1: Model architecture of *Neural Formant Syntheis*.

II. Related work

A. Neural Formant Synthesis

Neural Formant Synthesis (NF) is a speech synthesis system that controls the phonetic properties of speech using acoustic parameters (Fig. 1). The architecture of NF comprises a feature-mapping network and a neural vocoder. NF is speaker-independent and enables faithful reproduction of natural speech from given acoustic parameters, unlike general neural vocoders. Moreover, unlike analysis-synthesis methods, NF allows for the independent control of each acoustic parameter.

NF employs a two-stage architecture, comprising a feature-mapping network and a neural vocoder. The feature-mapping network converts low-dimensional acoustic parameters into high-dimensional acoustic features to learn their correspondence. The neural vocoder then generates speech waveforms from the converted acoustic features.

However, the two-stage structure raises concerns regarding the potential speech quality degradation. This discrepancy arises from the mismatch between the acoustic features transformed by the feature-mapping network and those used to train the neural vocoder.

III. Proposed Methods

A. E2E-NF/E2E-NF+

The feature-mapping network demonstrated that low-dimensional acoustic parameters can be converted into high-dimensional acoustic features. This is because low-dimensional acoustic parameters provide sufficient information for generating high-dimensional acoustic features.

Based on this hypothesis, we propose *End-to-End Neural Formant Synthesis* (E2E-NF) (Fig. 2 (a)). The structure of E2E-NF is the same as HiFi-GAN v1, a neural vocoder that uses acoustic features as input. Following the previous work, we employed nine acoustic parameters: V/UV, log

F_0, F_1–F_4, spectral tilt, spectral centroid, and energy. To accommodate these parameters, the input dimension of the first convolutional layer in the HiFi-GAN Generator was modified from 80 to 9. The intermediate feature dimension was set to 512, and the upsampling scales were set to (5, 5, 4, 3) with corresponding kernel sizes of (10, 10, 8, 6).

Furthermore, we introduce E2E-NF+, an extension of E2E-NF that incorporates layers mimicking the feature-mapping network of NF (Fig. 2 (b)). E2E-NF+ has a gated structure similar to the feature-mapping network employing a non-causal WaveNet-style gated convolutional architecture. This structure incorporates six ResBlocks, each consisting of a 1D-CNN, a Gated Linear Unit (GLU), and another 1D-CNN, placed before the upsampling blocks of E2E-NF, inspired by the approaches in [3] and [4]. The ResBlock structure offers reduced computational complexity compared with the feature-mapping network.

B. *E2E-SiFi-NF*

The E2E-NF structure is not considered optimal because it was not originally designed to use acoustic parameters.

Therefore, we propose *End-to-End Source-Filter NF* (E2E-SiFi-NF) (Fig. 2 (c)). This model employs a structure similar to SiFi-GAN [5], which simulates the vocalization process based on the source-filter model. SiFi-GAN has demonstrated superior performance compared to HiFi-GAN in terms of synthetic speech quality and F_0 control. This structure consists of a Source Network for sound source generation, and a Filter Network representing the vocal tract filter. The Source Network generates a latent source representation from acoustic features and F_0. The Filter Network synthesizes a waveform from this latent representation and the acoustic features. These networks employ 43-dimensional acoustic features: 40-dimensional mel-generalized cepstral coefficients and three-dimensional band aperiodicity.

E2E-SiFi-NF utilized seven acoustic parameters: F_1–F_4, spectral tilt, spectral centroid, and energy, derived from the original nine parameters. Therefore, the input dimension for the first convolutional layer in both Source and Filter Networks was reduced from 43 to 7. Additionally, the Source Network employed a sine wave generated from continuous F_0 as input. The intermediate feature dimension was set to 512, and the upsampling scales were set to (5, 5, 4, 3), with corresponding kernel sizes of (10, 10, 8, 6).

IV. Experimental Evaluation

We evaluated the synthesized speech quality and the controllability of acoustic parameters using our proposed method. A comparative analysis was conducted against the baseline provided by NF. Both analysis-synthesis and control of the acoustic parameters were evaluated. The training conditions and evaluation results for each model are as follows.

A. *Experimental setup*

The JVS corpus [6], which comprised recordings of 100 Japanese speakers (male and female) at 24 kHz, served as the dataset. The data were divided into training, validation, and evaluation sets at a ratio of 90:5:5 for each speaker's 100 utterances. The evaluation was conducted using four speakers: two male (JVS001 and JVS003) and two female (JVS002 and JVS004).

The acoustic parameters were obtained from acoustic analysis using a hop size of 300 samples and a window width

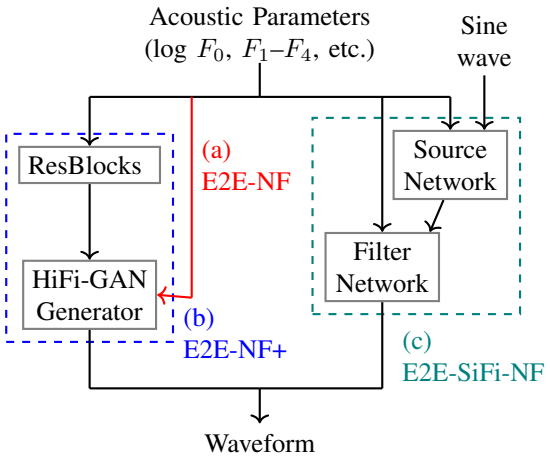

Fig. 2: Architecture of proposed models.
(a) E2E-NF, (b) E2E-NF+, and (c) E2E-SiFi-NF.

of 1200 samples. F_0 was extracted using WORLD [7] with lower and upper frequency limits of 75 Hz and 600 Hz, respectively. The unvoiced segments in the F_0 contour were interpolated linearly to ensure continuity. The formants were extracted using Praat [8] with the maximum number of formants set at five, a window width of 600, and a pre-emphasis of 50 Hz. The upper formant frequency limits during the extraction were set to 5000 Hz for male speakers and 5500 Hz for female speakers. The missing formant values were linearly interpolated. All parameters, except for V/UV, were standardized to have a zero mean and unit variance.

E2E-NF, E2E-NF+, and E2E-SiFi-NF were trained for 400 K steps each, using a batch size of 16 and a mini-batch length of 8,400. The HiFi-GAN v1 employed for NF as a vocoder was trained using 80-dimensional mel-spectrograms and the same dataset and training parameters as the proposed methods. The Univ Net multi-period and multi-resolution discriminator [9] was employed for all vocoders. NF training consisted of 99 K steps with a batch size of 128, which is consistent with previous work.

B. *Analaysis synthesis*

We compared our method with both conventional methods and HiFi-GAN, which synthesizes speech waveforms using acoustic features.

The evaluation metrics were the mean opinion score (MOS) predicted by an opinion score prediction system [10] (UTMOS), log F_0 root mean square error (RMSE), and V/UV error rate (V/UV). Table I shows the evaluation results.

Among the proposed methods, E2E-NF+ and E2E-SiFi-NF outperformed conventional methods in terms of UTMOS and V/UV. Additionally, in most cases, E2E-NF+ and E2E-SiFi-NF demonstrated superior performance compared to HiFi-GAN. These are considered to be the results of consistent learning using an end-to-end structure.

However, E2E-NF exhibited the lowest UTMOS. This is likely because the HiFi-GAN Generator, which assumes acoustic features, receives acoustic parameters as input. The evaluation results of E2E-NF+, which incorporates a structure analogous to a feature-mapping network, further support this explanation.

These results demonstrate that low-dimensional acoustic parameters retain sufficient information, thereby enabling end-to-end speech synthesis.

TABLE I: Results of objective and subjective evaluations for analysis synthesis.

Model	UTMOS	V/UV [%]	RMSE [cent]
GT	3.45 ± 0.41	-	-
HiFi-GAN	2.81 ± 0.25	**3.66 ± 0.69**	43.10 ± 19.00
NF (baseline)	2.90 ± 0.25	4.11 ± 1.12	39.57 ± 16.46
E2E-NF	2.64 ± 0.22	4.10 ± 0.66	38.24 ± 6.79
E2E-NF+	2.97 ± 0.37	3.93 ± 0.95	42.76 ± 9.44
E2E-SiFi-NF	**3.10 ± 0.35**	3.90 ± 0.49	**28.25 ± 9.32**

C. Parameter manipulated synthesis

Next, we evaluated the synthesized speech when the acoustic parameters were manipulated.

The acoustic parameters were manipulated by multiplying them by a constant factor (0.7, 0.8, 0.9, 1.1, 1.2, 1.3) across the entire utterance, as in previous research. Scaling was applied to only one acoustic parameter for each utterance, while the non-controlled acoustic parameters remained unchanged. As reported in [11], F_3, along with F_1–F_2, significantly influences the phonetic properties of speech. Therefore, the acoustic parameters selected for evaluation were F_0, F_1, F_2, and F_3.

UTMOS and mean squared error (MSE) were calculated as evaluation criteria for the manipulated acoustic parameters. The acoustic parameters, extracted from the synthesized speech using the same method as in the training, were regularized to have a zero mean and unit variance. MSE were calculated using only the voiced segments common to both the input and synthesized speech, was converted to a log scale. Voiced segments were determined by the V/UV flags after the input speech frame length was aligned to the synthesized speech frame length using dynamic time warping (DTW). Table II and Fig. 3 show the UTMOS results and the MSE plots, respectively.

In most cases, E2E-SiFi-NF showed the highest UTMOS, followed by E2E-NF+. This superior performance is attributed to SiFi-GAN's architecture, which enables higher-quality speech synthesis compared to HiFi-GAN.

Due to its focus on F_0 control, the SiFi-GAN generator exhibited remarkably low MSE for E2E-SiFi-NF during F_0 manipulation. For F_1–F_3 manipulations, the proposed method demonstrated comparable performance to the conventional method. These results suggest the feasibility of end-to-end acoustic parameter-controlled speech synthesis.

However, similar to the conventional method, the performance of the proposed methods deteriorated with large formant shift widths and higher-order formants. This limitation likely arises from the data-driven nature of the formant-speech correspondence, which lacks robustness and controllability beyond the training data distribution.

V. CONCLUSION

In this study, we proposed E2E-NF, E2E-NF+, and E2E-SiFi-NF, end-to-end speech synthesis models that utilize low-dimensional acoustic parameters as input. The results of the analysis-synthesis demonstrate the feasibility of end-to-end speech synthesis using low-dimensional acoustic parameters instead of acoustic features. Experiments on acoustic parameter manipulation demonstrate that the proposed methods synthesize speech with controllable acoustic properties, yielding quality and controllability comparable to or superior to those of conventional approaches.

However, the correspondence between the acoustic parameters and waveforms in our models is data-driven. Consequently, the controllability of acoustic parameters remains limited, and degradation in speech quality is observed when they are manipulated. Future work will investigate dataset augmentation and novel loss functions to improve the control of acoustic parameters and synthesized speech quality during manipulation.

TABLE II: UTMOS results of acoustic parameter manipulation.

F_0

Model	× 0.7	× 0.8	× 1.2	× 1.3
NF (baseline)	2.73 ± 0.23	2.84 ± 0.20	2.46 ± 0.33	2.23 ± 0.35
E2E-NF	2.68 ± 0.16	2.67 ± 0.21	2.34 ± 0.31	2.08 ± 0.36
E2E-NF+	2.81 ± 0.18	2.84 ± 0.27	2.48 ± 0.28	2.36 ± 0.37
E2E-SiFi-NF	**3.15 ± 0.28**	**3.17 ± 0.29**	**2.82 ± 0.35**	**2.48 ± 0.43**

F_1

Model	× 0.7	× 0.8	× 1.2	× 1.3
NF (baseline)	**2.38 ± 0.17**	2.84 ± 0.20	2.84 ± 0.25	2.73 ± 0.27
E2E-NF	2.17 ± 0.16	2.67 ± 0.21	2.64 ± 0.26	2.57 ± 0.27
E2E-NF+	2.36 ± 0.16	2.84 ± 0.27	2.77 ± 0.25	2.75 ± 0.28
E2E-SiFi-NF	2.31 ± 0.22	**3.17 ± 0.29**	**3.13 ± 0.39**	**3.03 ± 0.40**

F_2

Model	× 0.7	× 0.8	× 1.2	× 1.3
NF (baseline)	2.68 ± 0.19	2.79 ± 0.23	2.89 ± 0.24	2.88 ± 0.26
E2E-NF	2.54 ± 0.21	2.57 ± 0.20	2.67 ± 0.21	2.63 ± 0.20
E2E-NF+	2.83 ± 0.29	2.89 ± 0.32	2.97 ± 0.41	2.91 ± 0.39
E2E-SiFi-NF	**2.93 ± 0.29**	**3.01 ± 0.33**	**3.09 ± 0.36**	**3.05 ± 0.35**

F_3

Model	× 0.7	× 0.8	× 1.2	× 1.3
NF (baseline)	2.47 ± 0.20	2.60 ± 0.22	2.64 ± 0.24	2.48 ± 0.26
E2E-NF	2.11 ± 0.19	2.34 ± 0.20	2.43 ± 0.25	2.23 ± 0.25
E2E-NF+	2.56 ± 0.20	2.99 ± 0.36	2.61 ± 0.30	2.40 ± 0.28
E2E-SiFi-NF	**2.73 ± 0.29**	**3.01 ± 0.33**	**2.79 ± 0.33**	**2.55 ± 0.29**

REFERENCES

[1] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[2] P. P. Zarazaga, Z. Malisz, G. E. Henter, and L. Juvela, "Speaker-independent neural formant synthesis," in *Proc. Interspeech*, 2023, pp. 5556–5560.

[3] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, 2018, pp. 2100–2104.

[4] M. Tanaka, T. Nose, A. Kanagaki, R. Shimizu, and A. Ito, "Scyclone: High-quality and parallel-data-free voice conversion using spectrogram and cycle-consistent adversarial networks," *arXiv preprint, 2005.03334*, 2020.

[5] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. ICASSP*, 2023, pp. 1–5.

[6] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free japanese multi-speaker voice corpus," *arXiv preprint, 1908.06248*, 2019.

[7] M. Morise, F. Yokomori, and K. Ozawa, "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, 2016.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.4.13," retrieved 10 June 2024 from http://www.praat.org/.

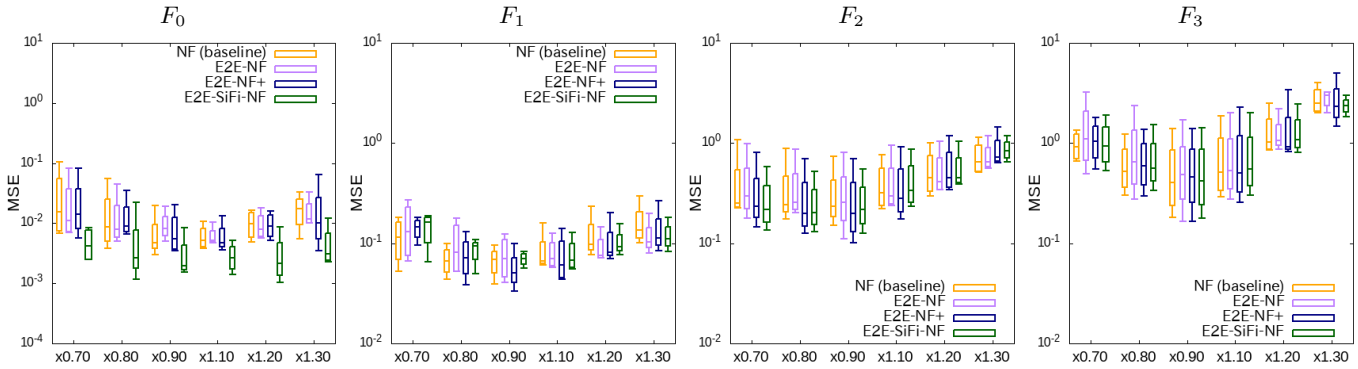

Fig. 3: Log-scale plot of MSE for manipulated acoustic parameters.

[9] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint, 2106.07889*, 2021.

[10] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint, 2204.02152*, 2022.

[11] K. Satou, "Importance of higher formants in discriminating /i/ and /e/ in Nagai, Yamagata: A perceptual experiment using synthesized speech," *Japanese Linguistics*, no. 132, p. p123〜110, 1983, (in Japanese).