Eden-VC: 音素継続長とアライメントの協調学習を用いた系列長変換型声質変換モデル *

◎山下陽生 [1,2]，岡本拓磨 [2]，高島遼一 [1]，大谷大和 [2]，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，ある話者の声質を別の話者の声質に変換する声質変換 (Voice Conversion: VC) の技術が発展している [1, 2]．一般的に VC 技術はネイティブ話者からネイティブ話者への変換が一般的であり，ノンネイティブ話者 (L2 話者) からネイティブ話者 (L1 話者) への変換についてはあまり議論が行われてこなかっが，L2 話者から L1 話者の発音へ変換が可能になれば国際的な場においてのコミュニケーションがより容易になっていくと考えられるため，重要な課題と言える．

VC モデルは大きく分けて，Sequence-to-Sequence (S2S) 方式 [4, 5, 6] と CycleGAN [3] などを用いたフレームバイフレーム方式の二つが存在している．前者は後者に比べ，話速や韻律の制御が容易に可能であり，この点が L2 話者の発音から L1 話者への発音への変換に優位であると考えられる．しかし，従来の S2S 方式の VC モデルでは L1 話者同士の変換では高品質かつ自然な音声が生成可能であったが，L2 話者と L1 話者の発音の変換を行うと，音素を誤って変換し自然な文章が生成できないという問題があった [7]．この原因の一つとして変換時のアライメントが十分でないことが挙げられる．従来の S2S 方式の VC モデルではソース音声とターゲット音声のアライメントに Monotonic Alignment Search (MAS) [8] を用いることが多く，これは動的計画法ベースの手法で単調アライメントを生成する．このような単純な手法であるため，L1 話者同士の変換では自然な音声が生成出来ていたが，L2 話者との変換のような複雑な問題の場合，図 1 に示すように同じ文章でも，音素継続長が違う，音素間のポーズ長が違う，同じ音が全く別の音のように発音される，といった問題が発生しアライメントが取りにくくなっている．

そこで本稿では，この問題を解決するために scale-dot attention を用いた柔軟なアライメントが生成可能な EdenTTS [9] を VC モデル化した Eden-VC を提案し，L2 話者と L1 話者の変換精度向上を目指す．

2 VC モデル

2.1 Eden-VC

Eden-VC は図 2(a) に示すように，EdenTTS を基にした scale-dot Attention ベースのアライメントモジュールを用いる S2S 型 VC モデルである．学習時は，ソース音声とターゲット音声のメルスペクトログラムをそれぞれ Encoder と Mel Encoder に入力し潜在変数 **h**，**m** を取り出した後，Guided Aligner で数式 (1) の重み付き scale-dot attention を用いてアライメントを生成する．

$$\alpha_{n,t} = \frac{e^{-(w_{n,t}\mathbf{h}_n \cdot \mathbf{m}_t)/\sqrt{D}}}{\sum_{i=0}^{N-1} e^{-(w_{i,t}\mathbf{h}_i \cdot \mathbf{m}_t)/\sqrt{D}}} \quad (1)$$

ここで，D は次元数であり、$w_{n,t} = e^{-(n/(N-1)-t/(T-1)^2/(2g^2)}$ は α が単調になるように制約を掛けており，Guided Aligner の収束が速くなっている．g はハイパーパラメータであり、0.2 とした．

Duration Extroctor は以下の式を用いて α から duration 生成する．

$$d_n = \sum_{i=0}^{T-1} \alpha_{n,i}, n = 0, 1, \ldots, N-1 \quad (2)$$

d_n は Duration Predictor の出力の教師として用いられる．Monotonic Aligner は d_n を用いて単調アライメント β を計算する．

$$\beta_{t,n} = \frac{e^{-\sigma^{-2}(t-\sum_{i=0}^{n} d_i - 0.5d_n)^2}}{\sum_{i=0}^{N-1} e^{-\sigma^{-2}(t-\sum_{i=0}^{n} d_i - 0.5d_n)^2}} \quad (3)$$

---

* Eden-VC:sequence-to-sequence voice conversion model with collaborative duration-alignment learning. by YAMASHITA, Haruki[1,2], OKAMOTO, Takuma[2], TAKASHIMA, Ryoichi[1], OHTANI, Yamato[2] TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], KAWAI, Hisash[1] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

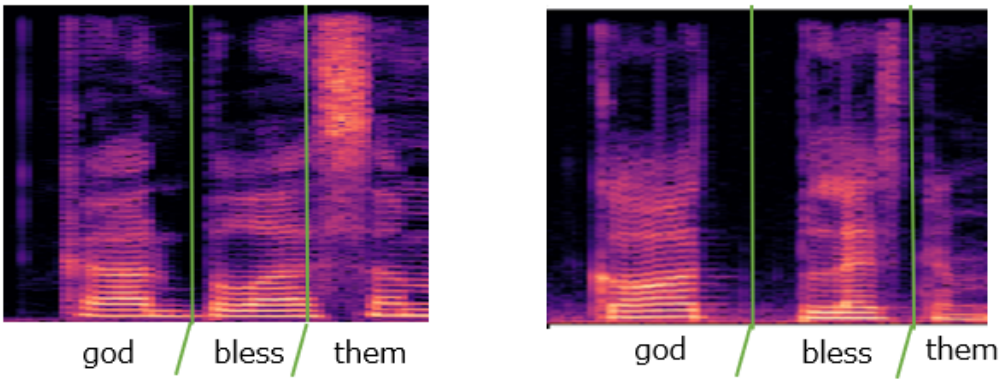

Fig. 1　Spectrograms of English L1 (left) and L2 (right) speakers.

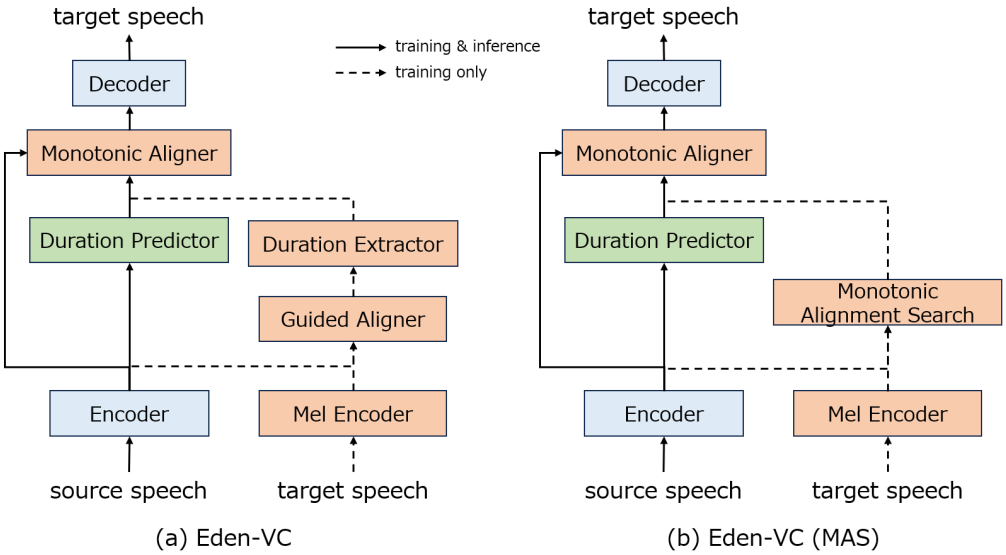

Fig. 2　Network architectures of (a) Eden-VC and (b) Eden-VC (MAS).

ここで，σ は 0.2 とした．この単調アライメントを用いて，以下の式を用いて Encoder の出力 $\mathbf{h}_n$ をアップサンプリングする．

$$\gamma_t = \sum_{n=0}^{N-1} \beta_{t,n}\mathbf{h}_n, t = 0, \ldots, T-1 \qquad (4)$$

scale-dot attention を用いたアライメントのため，MAS のように決定的なアライメントにならず，複雑な問題に対しても柔軟にアライメントが取れると考えられる．

2.2　Eden-VC (MAS)

Eden-VC (MAS) では図 2(b) のように Eden-VC のアライメントモジュールを MAS に変更したものである．MAS は動的計画法ベースのモジュールで，単調アライメントを生成する．Eden-VC (MAS) と Eden-VC を比較することで，MAS と scale-dot attention のどちらがより正確なアライメントが取れるかを比較する．

3　実験

実験は，音質評価のために日本語音声データセットを用いた日本語実験と L2→L1 の変換精度を確認する英語実験の二つを行う．

3.1　実験条件

日本語音声データセットには HiFi-CAPTAIN [10] の男性話者，女性話者の音声それぞれ 19056 文を用いて，Eden-VC と Eden-VC (MAS) をそれぞれ学習させ，客観評価実験と主観評価実験を行う．モデルの学習はどちらのモデルも 300 epoch 行った．

合成音声の客観評価実験には音声認識モデルを用いた CER の測定と，メルスペクトログラム歪み (MCD)，f_0 の対数二乗平均誤差 ($\log f_0$ RMSE) を用いた．MCD，$\log f_0$RMSE の計算には ESPNet2-TTS [15] を利用した．音声認識モデルは ESPNet2-TTS の評価で用いられている CSJ コーパスで学習された Transformer ベース

のモデルを用いた [15]．主観評価実験では平均オピニオン評点 (MOS) を用いた音質評価実験と話者類似性評価実験 (similarity) を行った．MOS, similarity の測定には各モデルから学習に用いていない 10 文を用いて，8 名の日本人話者にヘッドフォン聴取を行い求めた．MOS は 5 段階評価で測定し，5 が最も音質がよいとした．similaruty の評価には 4 段階評価を用いて，4 に近づくほど原音の話者と同じ話者と評価されるとし，その平均値を similarity の値として求めた．

L2→L1 の変換実験には，L1 話者として CMU-ARCTIC [12] から男性話者 (bdl)，女性話者 (slt) を 1 名ずつ (22.05 kHz，各 1131 文) 用い，L2 話者には，L2-ARCTIC [13] の 24 話者のうち，音声認識モデルを用いて最も CER が良かった女性話者 (NJS)1 名 (22.05 kHz, 1131 文) を選び，そのうちトレーニングには 1091 文を用いた．こちらの実験では客観評価実験のみを行った．客観評価実験には日本語事件の時と同様に CER, MCD，log f_o RMSE の測定を行い，CER の測定には LibriSpeech によって学習された Conformer ベースの音声認識モデルを用いた．

音響特徴量は 8 kHz まで帯域制限したメルスペクトログラムとした．また，Vocoder には ljspeech で事前学習した HiFi-GAN [14] を用いて，Eden-VC，Eden-VC (MAS) にて生成されたメルスペクトログラムから音声を生成した．

3.2 実験結果

表 1 の日本語実験の結果では，主観評価実験において Eden-VC は Eden-VC (MAS) と比較して有意に自然性が高いことが分かる．しかし，原音と比較するとどちらのモデルも有意に低く，これは vocoder が英語音声で学習されていることと，fine-tuning を行っていないためであると考えられる．また，話者類似性においても Eden-VC が Eden-VC (MAS) を上回り，話者性が正確に変換されていることが分かる．客観評価結果においても変換における CER の悪化は Eden-VC の方が少なく，正しい音素を維持したまま話者性のみを変換できていることが分かる．

次に表 2 の英語実験の結果では，L1 話者同士の変換では日本語実験と同様に CER の悪化はみられるものの，Eden-VC の方が Eden-VC (MAS) よりも変換精度が高いことが分かる．また，MCD と logf_o RMSE においてもほとんどの実験条件で Eden-VC が Eden-VC (MAS) よりも良く，正確に変換できていることが分かる．また，L2→L1 変換においては，NJS から bdl，slt どちらへの変換においても Eden-VC (MAS) の CER を上回り，特に NJS→bdl の変換では大幅に CER の劣化が改善されている．これは，EdenTTS にて提案された scale-dot attention ベースのアライメントモジュールを用いた方が複雑な変換においても MAS よりも高精度にアライメントを生成可能になったためであると考えられる．

4 おわりに

EdenTTS にて提案された scale-dot attention ベースのアライメント手法は，従来の S2S VC モデルで用いられてきた MAS よりも高精度にアライメントが生成可能であることが分かった．このアライメント手法を用いることで，声質変換においてもより正確な変換が可能である．また，このアライメント手法は L2 話者から L1 話者への変換といった複雑な問題においても柔軟にアライメントが生成でき，変換精度が向上することが分かった．しかし，変換後の自然性は原音から大きく劣り，vocoder との fine-tuning や一貫学習を行うことで合成品質を向上させる必要があり，今後の課題である．

参考文献

[1] S. H. Mohammadi and A.Kain, “An overview of voice conversion systems,” *Speech Commun.*, vol.88, pp.65–82, Apr. 2017.

[2] B. Sisman *et al.*, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Trans.Audio, Speech,Lang.Process.*, vol.29, pp.132–157, 2021.

[3] T. Kaneko *et al.*, “CycleGAN-VC: Non-parallel voice conversion Using cycle-consistent adversarial networks,” in *Proc. EUSIPCO*, 2018, pp.2114–2118.

[4] T. Hayashi *et al.*, “Non-autoregressive sequence-to-sequence voice conversion,” in *Proc. ICASSP*, June 2021, pp.7068–7072.

[5] P. Guo *et al.*, “Recent developments on ESPnet toolkit boosted by Conformer,” in *Proc. ICASSP*, June 2021, pp. 5874–5878.

Table 1 Result of experiment for Japanese L1 to L1 conversion.

Model	src	tgt	MOS	Similarity	CER	MCD[dB]	$\log f_o$ RMSE
Eden-VC	female	male	**2.81±0.19**	**3.05±0.21**	**0.9**	6.64±0.75	**0.23±0.05**
	male	female	**3.54±0.22**	**3.63±0.13**	**0.8**	8.18±0.71	0.27±0.07
Eden-VC (MAS)	female	male	1.70±0.18	2.61±0.20	1.2	**6.51±0.78**	0.24±0.06
	male	female	2.36±0.19	3.14±0.16	1.1	**8.09±0.78**	0.27±0.07
original female	-	-	4.97±0.03	-	0.0	-	-
original male	-	-	4.91±0.07	-	0.0	-	-

Table 2 Result of experiments for English L1 to L1 and L2 to L1 conversions. slt, bdl and NJS are L1 female, L1 male and L2 female speakers, respectively.

Model	src	tgt	CER	MCD[dB]	$\log f_o$ RMSE
Eden-VC	slt	bdl	**1.6**	**6.73±0.36**	0.25±0.06
	bdl	slt	**1.4**	9.60±0.78	**0.26±0.06**
	NJS	bdl	**13.9**	**7.32±0.54**	**0.26±0.07**
	NJS	slt	**14.8**	10.01±0.70	**0.26±0.06**
Eden-VC (MAS)	slt	bdl	3.9	6.79±0.34	0.25±0.07
	bdl	slt	2.3	**9.23±0.76**	0.28±0.07
	NJS	bdl	20.9	7.74±0.62	0.27±0.07
	NJS	slt	16.9	**9.77±0.58**	0.27±0.06
original slt	-	-	0.8	-	-
original bdl	-	-	0.5	-	-
original NJS	-	-	3.0	-	-

[6] T. Okamoto *et al.*, "E2E-S2S-VC: End-to-End sequence-to-sequence voice conversion," in *Proc.Interspeech*, Aug. 2023.

[7] 山下ら, "End-to-End 系列変換型声質変換の高速化およびノンネイティブ話者変換の検討", 音講論, Mar. 2023.

[8] J. Kim *et al.*, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *NIPS*, 2020, pp. 8067–8077.

[9] Y. Ma *et al.*, "EdenTTS: A simple and efficient parallel text-to-speech architecture with collaborative duration-alignment learning," in *Proc.INTERSPEECH*, 2023, pp.4449–4453.

[10] T. Okamoto, Y. Shiga and H. Kawai, "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," `https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/`, 2023.

[11] V. Panayotov *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *Proc.ICASSP*, 2015, pp.5206–5210.

[12] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *SSW5*, 2004, pp.223–224.

[13] G.Zhao *et al.*, "L2-ARCTIC: A non-native English speech corpus," in *Proc. Interspeech*, 2018, pp.2783–2787.

[14] J. Kong *et al.*, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17022–17033.

[15] T. Hayashi *et al.*, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv:2110.07840*, 2021.