系列変換型声質変換におけるアライメント方式の比較*

◎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，声質変換 (Voice Conversion: VC) 技術 [1] が発展し，実際のアプリケーションでも用いられている．特に，非自己回帰型 (Non-AR) の系列長変換 (Sequence-to-Sequence:S2S) モデル [2] は，CycleGAN [3] などをベースとしたフレームバイフレーム方式と違い，話速や韻律も制御でき，高品質な変換を高速で可能にした．Non-AR S2S モデルを学習するためには，ソース話者音声とターゲット話者音声間のアライメントを生成する必要がある．Fastspeech2 [4] を基に VC モデルへ変更した FS2-VC では別途学習した AR モデルの出力からアライメントを生成する．しかし，AR モデルのアライメントの精度や別に AR モデルを学習する必要があるなどの問題点から，Monotonic alignment search (MAS) [5] を用いる Alignment module を使用することで音響モデルの学習と同時にアライメントを学習できるモデルとして，JETS-VC [6] や AAS-VC [7] などが提案された．MAS では CTC-Loss [8] を用いて対角なアライメントを生成していくことで安定した学習を可能にしているが，アライメント生成に Viterbi algorithm を使用する Hard alignment であるため，Alignment path が粗くなりやすく学習が間違っているときの修正が難しい，1 度間違った Path を選択したときに後のフレームまでその影響が残ってしまうなどの問題がある．そこで，Scaled-dot product attention [9] ベースの Soft alignment を用いることで Alignment path を滑らかにし，前フレームの影響を受け無くしたモデルとして Eden-VC [11] が提案された．Soft alignment の実装によって Eden-VC では，アライメントの柔軟性が向上し声質変換時の音声の崩れを抑えることができたが，学習が不安定で，安定して正しいアライメントを生成することが難しかった．

そこで，Soft alignment の重みを Hard alignment から生成する新たな Alignment module とそれを用いた WSAS-VC (Weighted Soft Alignment Search Voice Conversion) を提案する．WSAS-VC は Soft alignment を Hard alignment で条件付けするため，Soft alignment の柔軟性と Hard alignment の学習の安定性を両立していることが期待される．本稿では WSAS-VC によって，VC の変換精度を向上させることを目的とする．

2 モデル詳解

2.1 FS2-VC-MAS

FS2-VC-MAS は高品質な音声を高速で合成可能な Text-to-Speech モデルである Fastspeech2 の入力をテキストからメルスペクトログラムに変えることで VC モデルとした FS2-VC に，MAS によるアライメント学習機構を追加したモデルであり，図 1(a) に示すように AAS-VC と似た構造を持つ．これによって，音響モデルと Alignment module を同時に学習可能になっている．JETS-VC と同様に Variance predictor は Gaussian upsampling 後に追加し，その際の教師にはターゲット音声の f_o/Energy を用いた．

Monotonic alignment search (MAS)

MAS は Glow-TTS [18] で提案された Alignment module であり，図 1(b) に示すように，2 つの Alignment encoder から構成されている．Alignment encoder の出力をそれぞれ x^{src} と y^{trg} とすると，数式 (1)(2) のように A_{soft} を求める．

$$D_{n,t} = dist_{L2}(x_n^{src}, y_t^{trg}) \tag{1}$$

$$A_{soft} = softmax(-D, dim = 0) \tag{2}$$

A_{soft} は CTC-Loss によって対角が強調されるように学習される．その後，Viterbi algorithm を用いて実際の Hard alignment を生成する．CTC-Loss を使用するため，ソース側音声がターゲット側音声と比べて必ず短い必要がある．このため，MAS を使用する際はソース話者側の音声を Reduction factor [10] を用いて短くする．AAS-VC では Reduction は Encoder 後に行う Post-reduction を採用しているが，FS2-VC-MAS では参考文献 [12] の結果に，隠れ特徴量ではなくメルスペクトログラム同士を直接 MAS に入力することで精度が上がることが報告されているため，本実験においてもそのように Encoder 前に Reduction を行う Pre-reduction とした．

2.2 Eden-VC

Eden-VC は EdenTTS [13] を基に VC モデルへ変更した Scaled-dot product attention ベースの Alignment module を持つモデルであり，図 2 に示す構造を持つ．学習時は，まず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

*Comparison of Alignment Methods in Sequence-to-Sequence Voice Conversion. by YAMASHITA, Haruki[1,2], OKAMOTO, Takuma[1], TAKASHIMA, Ryoichi[1], OHTANI, Yamato[2], TAKIGUCHI, Tetsuya[1], TODA, Tomoki[3,2], KAWAI, Hisashi[1] ([1]Kobe Univ, [2]NICT, [3]Nagoya Univ)

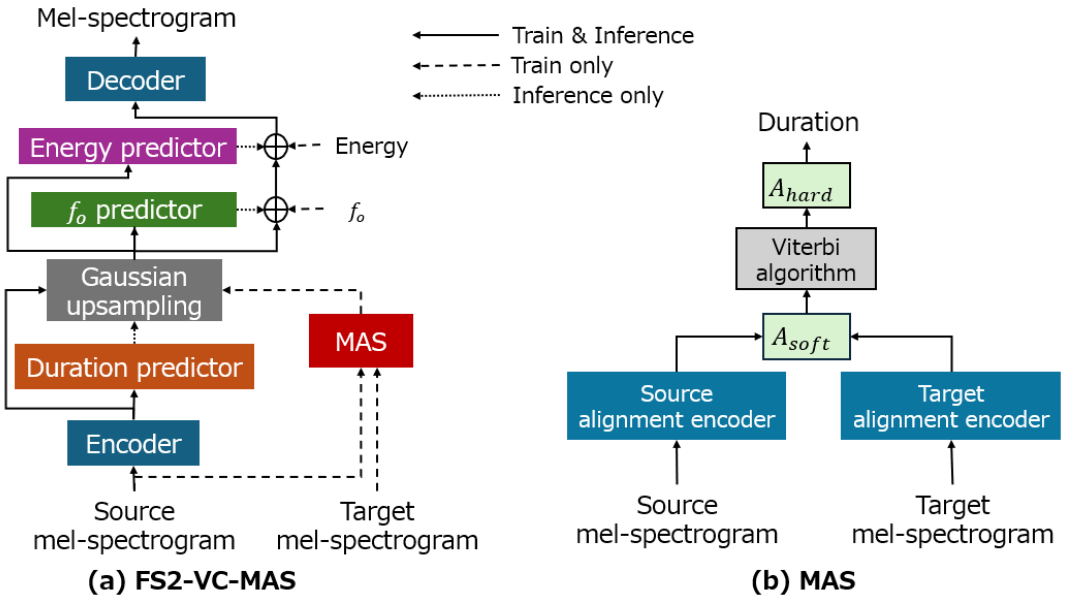

Fig. 1　Network architectures of (a) FS2-VC-MAS and (b) MAS.

encoder に入力し潜在変数 **h**，**m** を取り出す．**h**，**m** を用いて，Guided aligner にて数式 (3) の重み付き Scaled-dot product attention を用いたアライメントを生成する．

$$\alpha_{n,t} = \frac{e^{-(w_{n,t}\mathbf{h}_n \cdot \mathbf{m}_t)/\sqrt{D}}}{\sum_{i=0}^{N-1} e^{-(w_{i,t}\mathbf{h}_i \cdot \mathbf{m}_t)/\sqrt{D}}} \tag{3}$$

ここで，D は次元数であり，$w_{n,t}$ は数式 (4) で定義される α が単調になるような制約となる重みである．

$$w_{n,t} = e^{-(n/(N-1)-t/(T-1)^2/(2g^2)} \tag{4}$$

ここで，$g = 0.2$ とした．その後，Duration extractor で数式 (5) によって Duration を生成する．

$$d_n = \sum_{i=0}^{T-1} \alpha_{n,i}, n = 0, 1, \ldots, N-1 \tag{5}$$

この d_n を用いて，Duration predictor を学習する．

2.3　WSAS-VC (proposed)

WSAS-VC は図 2 に示すように基本構造は Eden-VC と同じであるが，α の生成時の重みとして，MAS で得られた Alignment path 周辺を強調する重みを利用するモデルである．MAS から得られたパスから重みを生成し使用することで，MAS の安定した学習を維持したまま，ソフトアライメントの柔軟性が利用できると期待される．重みは数式 (6) を用いて生成する．

$$w_{n,t} = e^{(-p_t-n)^2/2g_{mas}^2} \tag{6}$$

ここで，p_t はターゲット側の t フレーム目における MAS の出力パスを表している．また，$g_{mas} = 5$ とした．

3　実験

3.1　実験条件

データセット：英語音声データセットである CMU ARCTIC [14] から男性話者 (bdl) と女性話者 (slt) をそれぞれ 1 名ずつ選び，男性話者から女性話者への変換と，女性話者から男性話者への変換を行った．学習データは 1,090 文とし，サンプリング周波数は 24 kHz とした．

モデル設定：Eden-VC と WSAS-VC の実装について，EdenTTS の実装に従い，Conv1d，Linear，LayerNorm を全て weights を正規分布に従って初期化し，bias を 0 で初期化した．これらの変更によって，モデルは適切にアライメントを学習できるようになる．Reduction factor は Eden-VC のみ 1 とし，ほかのモデルは 3 で実験を行った．各 VC モデルは Pytorch ベースのオープンソースである ESPnet2-TTS [16] を利用し，それぞれ 300 epoch の学習を行った．

ボコーダには高品質な音声を高速で合成できるモデルとして MS-FC HiFi-GAN [15] を用いた．MS-FC HiFi-GAN の ResBlock のカーネルサイズは (4, 4) とし，CMU ARCTIC から選ばれた男性話者 (bdl) と女性話者 (slt) を用いて学習を行った．音響特徴量は Frame shift を 10 ms とし 8 kHz まで帯域制限したメルスペクトログラムを用いた．

評価手法：合成品質の客観評価には，メルケプストラム歪み (MCD)，対数 f_o の二乗平均誤差 ($\log f_o$ RMSE)，また変換による音声の崩れを評価するため音声認識モデルによる文字誤り率 (CER) 用いた．客観評価には学習に用いていないデータのうち 20 文を使用して，MCD と $\log f_o$ RMSE の計算にはオープンソフトの ESPNet2-TTS を利用した．音声認識モデルには Transformer ベースで librispeech [17] で学

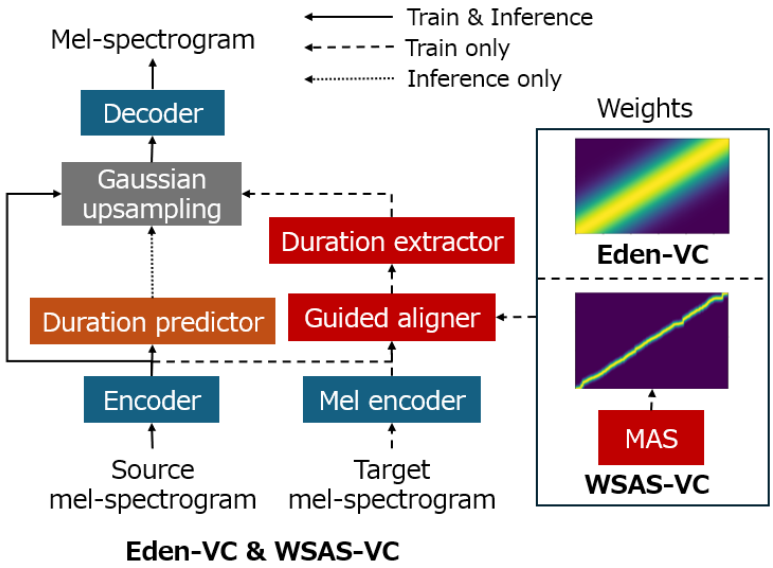

Fig. 2　Network architectures of Eden-VC and WSAS-VC.

習されたモデルを用いた．

主観評価では音質評価実験として平均オピニオン評点 (MOS) を用い，また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データのうち 10 文を用いて，6 名の日本人話者にヘッドフォン聴取を行った．MOS は 5 段階評価で測定し，5 が最も音質がよいとした．Similarity 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 の値とした．

3.2　実験結果

実験結果を表 1 に，実際に学習時された Weights と Alignment path を図 3 に示す．まず，FS2-VC-MAS の Alignment path は Viterbi の出力になっておりハードで粗い．Eden-VC では Soft alignment であるため滑らかな Alignment path ではあるが強調されていないことがわかる．その一方で WSAS-VC の Alignment path は FS2-VC-MAS と同程度強調されている上，Eden-VC のような滑らかさを持っており，提案手法が予想通りの安定した Alignment を出力していることがわかる．次に CER についてみてみると，提案手法の WSAS-VC が最もよい結果となり，原音と同等の値となっている．これは，Alignment path が既存手法と比較して本来のものに近づいた結果，変換時の音声の崩れが少なくなったからであると考えられる．また，MCD の結果においても WSAS-VC が他のモデルと比較して最もよく，原音に近い音声になっていることがわかる．$\log f_0$ RMSE の結果は，3 モデルで全て同程度であり，話者性も大きく差がないことがわかる．

次に，MOS については t 検定を行ったところ 3 モデル間に有意差はなく，同程度の音質であると言える．Simirality についても同様で，全てのモデルで同程度の話者性を有していることがわかる．

4　おわりに

本稿では，声質変換時の音声の崩れを防ぐモデルとして WSAS-VC を提案し，実験結果から提案手法は Alignment をより精度よくとることができ，音声の変換時の崩れを抑えることができていることが分かった．提案手法は Aard alignment と Soft alignment の両方の性質を併せ持つため，ノンネイティブ話者からネイティブ話者への変換といったより複雑な問題への適応も可能であり，今後の課題である．

参考文献

[1] 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.

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

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

[4] Y. Ren *et al.*, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, May 2021, pp. 1–15.

[5] R. Badlani *et al.*, “One TTS Alignment to rule them all,” in *Proc. ICASSP*, May 2022, pp. 6092-6096.

Table 1　Result of Experiment. ± represents standard error

Model	Source	Target	CER	MCD[dB]	$\log f_0$ RMSE	MOS	Similarity
FS2-VC-MAS	Male	Female	1.2	8.84	0.44	3.23±0.28	**3.35±0.20**
	Female	Male	2.1	5.74	0.20	**3.47±0.26**	**3.47±0.23**
Eden-VC	Male	Female	1.4	8.98	**0.42**	**3.73±0.26**	3.29±0.19
	Female	Male	1.4	5.54	**0.19**	3.23±0.28	3.38±0.22
WSAS-VC (proposed)	Male	Female	**0.7**	**8.79**	0.43	3.52±0.27	3.23±0.21
	Female	Male	**0.7**	**5.50**	**0.19**	3.25±0.24	3.25±0.20
Original male	-	-	0.5	-	-	4.70±0.15	-
Original female	-	-	0.8	-	-	4.35±0.22	-

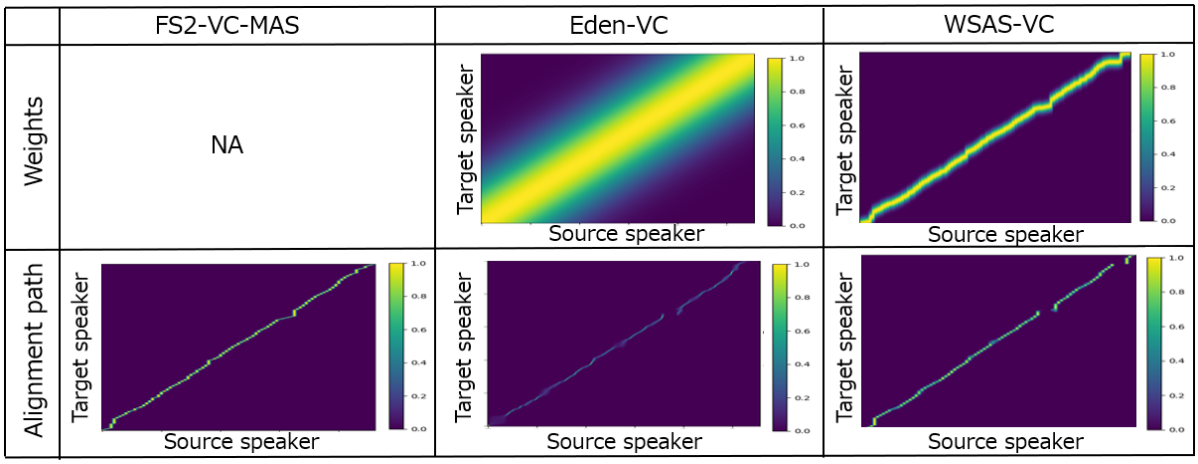

Fig. 3　Comparison of learned weights and alignment paths.

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

[7] W. Huang *et al.*,"AAS-VC: On the generalization ability of automatic alignment search based non-autoregressive sequence-to-sequence voice conversion" *arXiv:2309.07598*, 2023.

[8] A. Graves *et al.*, "Connec-tionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376

[9] A. Vaswani *et al.*, "Attention is all you need," in *NeuralIPS*, vol. 30, 2017

[10] W. Huang *et al.*, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech,* Oct. 2020, pp. 4676–4680.

[11] 山下ら, "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", 音講論, Mar. 2024.

[12] 山下ら, "系列変換型声質変換モデルにおける単調アライメント探索の改良", 音講論, Sep. 2024.

[13] 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.

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

[15] H. Yamashita *et al.*, "Fast eural speech waveform generative models with fully-connected layer-based upsampling," *IEEE Access*, vol. 12, pp. 31409–31421, 2024.

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

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

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