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 *

◎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] が発展してきており，その中でも非自己回帰型の系列変換 (Sequence-to-Sequence: S2S) モデル [2] が注目されている。S2S モデルは，それまで主に用いられてきた Cycle-GAN [3] のフレームバイフレーム方式の VC モデルよりも話速や韻律の変換が可能であり，ノンネイティブ (L2) 話者からネイティブ (L1) 話者への変換といった問題において優位であると考えられる。しかし，従来の S2S モデルである JETS-VC [4] では，L1 話者同士の変換においては高品質な音声の変換が可能であったが，L2 話者から L1 話者への変換では Alignment Module [5] が変換に十分な教師アライメントを生成できず，音声の変換精度が劣化し自然な文章を生成できない問題があった [6]。これは，Alignment Module が Text-to-Speech (TTS) タスクで提案されたものであり，声質変換タスクに最適化されていないためであると考えられる。また，JETS-VC に用いられている Alignment Module よりも VC タスクに優位であると考えられるソフトアライメントを利用する Alignment Module を用いた Eden-VC [7] では，実際に JETS-VC よりも変換精度が良くなる結果が得られたが，JETS-VC に比べ実験条件によって学習が安定しない問題があった。

そこで，本稿では JETS-VC で用いられている Alignment Module に着目し，Alignment Encoder や特徴量について様々な条件で比較を行うことで，声質変換タスクにおいて最適化を行うことを目的とする。

2 モデル詳解

2.1 系列長変換モデル: FastSpeech2-VC

声質変換モデルにおける系列長変換モデルは JETS-VC や AAS-VC [9] が挙げられるが，これらのモデルは TTS モデルである FastSpeech2 [8] を基にしたモデルである。FastSpeech2 を基にした VC モデル (FastSpeech2-VC) の構造を Fig. 1 に示す。まずソース話者のメルスペクトログラムを Encoder に入力し隠れ特徴量 h を得る。次に，ソース話者のメルスペクトログラムから Encoder を用いて得られた隠れ特徴量 h をターゲット話者の系列長に変換するため，Gaussian Upsampling を用いたアップサンプリングを行う。アップサンプリングに用いる Duration

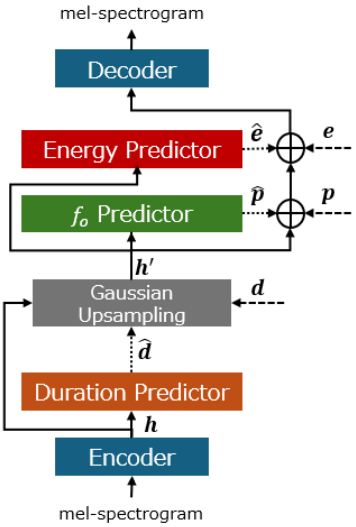

Fig. 1 FastSpeech2-VC architecture.

情報は，学習時は Alignment Module からの出力 $\mathbf{d}$ を，推論時には Duration Predictor の出力 $\hat{\mathbf{d}}$ を用いる。系列長を変換した h' には，ターゲット話者の Energy/基本周波数 (f_o) 情報を加算し，Decoder によってターゲット話者のメルスペクトログラムを生成する。Energy/f_o 情報には，学習時はターゲット話者から得られた Energy/f_o である $\mathbf{e}$，$\mathbf{p}$ を用い，推論時は Energy/f_o Predictor の出力である $\mathbf{p}$，$\hat{\mathbf{e}}$ を用いる。

2.2 Alignment Module

Alignment Module には Fig. 2 に示すような Glow-TTS [10] にて提案されたモジュールを用いる。

この Alignment Module は JETS-VC や AAS-VC でも用いられているが，ノンネイティブ話者からネイティブ話者への変換など，複雑なタスクにおいては十分な精度のアライメントが生成できず，変換精度が劣化するという問題がある。この原因として，Glow-TTS で提案されている Alignment Module は TTS タスク用に調整されており，VC タスク用に調整が必要であるからと考えられる。

まず，Glow-TTS で提案されている Alignment Module では $\mathbf{x}^{enc}$，$\mathbf{y}^{enc}$ を得るための Alignment Encoder にはそれぞれ 2 層，3 層の 1 次元畳み込みを用いている。これは，TTS タスクにおいては $\mathbf{x}$ にはテキスト特徴量を，$\mathbf{y}$ にはメルスペクトログラムを用いるため，複雑な特徴量であるメルスペクトログラムに対して多くの畳み込み層を通すことでより抽象化され

* Improving monotonic alignment search in sequence-to-sequence voice conversion models. 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)

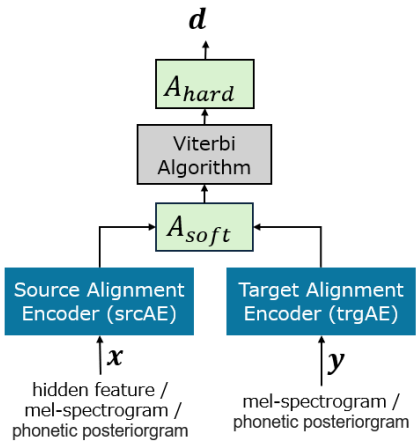

Fig. 2　Alignment Module architecture.

ることを目的としているからである。しかし，VC タスクにおいては $\mathbf{x}$，$\mathbf{y}$ の両方ともメルスペクトログラムなどの音響特徴量や，FastSpeech2-VC の Encoder に通して得られた隠れ特徴量であるため，同じ程度の抽象化をするべきである。したがって，本稿では両方の Alignment Encoder の層数を 3 にし，実験にて変換精度を確認する。

3　実験

本実験ではまず各声質変換モデルと Alignment Module，アライメント特徴量に関して客観評価実験を行う。この実験では，L1 に発音の近い 1 名の L2 話者と 2 名の L1 話者を用いる。その後，アライメント特徴量についてそのほかの L2 話者を用いて変換を行い，その変換精度を比較する。主観評価実験では，客観評価実験にて品質の良いモデルを選び，それらに対して行う。

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 [11] 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。また，L2 話者は，L2-ARCTIC [12] から，スペイン語を母語とする女性話者 (NJS) と男性話者 (ERMS)1 名ずつ，ベトナム語を母語とする女性話者 (PNV) と男性話者 (TLV) の 1 名ずつを用いた。この中で，スペイン語を母語とする女性話者は librispeech [13] で学習された Conformer ベースの音声認識モデルで認識したときの文字誤り率 (Charactor Error Rate: CER) が最も低いため，Alignment Module の変換精度の比較実験に用いた。学習データは 1,090 文とし，サンプリング周波数は 16 kHz とした。

モデル設定： メルスペクトログラムの変換には FasSpeech2-VC を用い，Vocoder には高品質な音声を高速で生成可能な MS-FC-HiFi-GAN [15] を用いた。Vocoder の各モデルの ResBlock のカーネルサイズは (4, 4) とし，学習には CMU ARCTIC から選ばれた男性話者と女性話者である bdl と slt を用いた。FastSpeech2-VC の学習には Pytorch ベースのオープンソースである ESPnet2-TTS [14] を利用し，各モデルは 100 epoch 学習した。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。

アライメント特徴量： アライメント特徴量には，FastSpeech2-VC の Encoder から得られる隠れ特徴量，メルスペクトログラム[1]，音素事後確率 (Phonetic PosteriorGram: PPG) の 3 種類を用いた。PPG の取得には PPG-VC[18] にて公開されている音声認識モデルを用いた。メルスペクトログラムは FFT の frame shift 量を 10 ms，20 ms の 2 種類とすることで，メルスペクトログラムの粗さがどの程度変換精度に影響するかを調べた。また，Viterbi Algorithm を用いるため，ソース話者のアライメント特徴量には reduction factor を 3 とすることで系列長が必ずターゲット話者よりも短くなるようにした。

3.2　客観評価実験結果

客観評価実験の結果を Table 1 に示す。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_0 の二乗平均誤差 (log f_0 RMSE)，また変換による音声の崩れの評価に CER 用いた。MCD と log f_0 RMSE の計算にはオープンソフトの ESPNet2-TTS を利用した。CER は librispeech [13] で学習した conformer ベースの音声認識モデルで測定した。客観評価には学習に用いていないデータのうち 20 文を使用した。

PPG をアライメント特徴量に用いたモデルは，ほとんどの条件において最も CER が良くなり，メルスペクトログラムや隠れ特徴量よりもアライメント特徴量に適していることが分かった。これは PPG が話者情報の除かれた言語情報であるためであると考えられる。次に，Alignment Encoder の層数はソース話者とターゲット話者で合わせることによって，L2 話者から L1 話者への変換において CER が向上することから，より複雑なアライメントがとりやすくなっていることが分かった。この結果は Eden-VC [7] より上回っており，変換精度と学習の安定性の両方において Eden-VC で提案されたソフトアライメントよりも優位である。また，隠れ特徴量をアライメント特徴量に用いるモデルはメルスペクトログラムを用いるモデルより，ほぼすべての条件で CER が悪い。これは，隠れ特徴量を生成する Encoder が Transformer ベースであり，遠くの時系列情報をまとめてしまうためであると考えられる。FFT の frame shift については 10 ms を用いた場合，20 ms よりも細かく音響特徴量が取れるため，CER が向上することが分かっ

[1] TTS タスクにおいても Encoder から得られる隠れ特徴量ではなく，音素埋め込みベクトルとする方式が検討されている [16, 17]。

Table 1　Result of objective evaluations for Alignment Module, where src is source, trg is target and AE is Alignment Module. Non-native speaker is Spanish speaker.

srcAE conv layers	trgAE conv layers	src alignment feature	trg alignment feature	FFT frame shift [ms]	src speaker	trg speaker	source original CER [%]	converted CER [%]	MCD [dB]	logf_oRMSE
2	3	hidden feature	mel-spectrogram	20	native male	native female	0.5	4.2	4.86 ± 0.34	0.19 ± 0.34
				20	native female	native male	0.8	4.1	5.55 ± 0.40	0.20 ± 0.07
				20	non-native female	native female	3.0	82.4	8.95 ± 0.82	0.21 ± 0.07
				20	non-native female	native male	3.0	21.2	6.14 ± 0.57	0.20 ± 0.06
2	3	hidden feature	mel-spectrogram	10	native male	native female	0.5	1.2	4.97 ± 0.29	0.18 ± 0.05
				10	native female	native male	0.8	4.1	5.47 ± 0.33	0.19 ± 0.08
				10	non-native female	native female	3.0	10.7	5.70 ± 0.38	0.21 ± 0.05
				10	non-native female	native male	3.0	12.0	6.14 ± 0.50	0.19 ± 0.06
3	3	hidden feature	mel-spectrogram	20	native male	native female	0.5	4.7	4.83 ± 0.28	0.18 ± 0.05
				20	native female	native male	0.8	4.8	5.46 ± 0.34	0.20 ± 0.07
				20	non-native female	native female	3.0	80.5	9.23 ± 0.79	0.22 ± 0.06
				20	non-native female	native male	3.0	17.5	6.14 ± 0.51	0.21 ± 0.06
3	3	hidden feature	mel-spectrogram	10	native male	native female	0.5	1.5	4.91 ± 0.32	0.18 ± 0.05
				10	native female	native male	0.8	3.3	5.53 ± 0.34	0.20 ± 0.07
				10	non-native female	native female	3.0	10.6	5.73 ± 0.36	0.20 ± 0.05
				10	non-native female	native male	3.0	10.6	6.21 ± 0.52	0.21 ± 0.06
3	3	mel-spectrogram	mel-spectrogram	20	native male	native female	0.5	2.7	4.69 ± 0.37	0.19 ± 0.06
				20	native female	native male	0.8	5.1	5.29 ± 0.36	0.19 ± 0.08
				20	non-native female	native female	3.0	14.1	5.54 ± 0.37	0.19 ± 0.06
				20	non-native female	native male	3.0	10.0	5.95 ± 0.52	0.19 ± 0.06
3	3	mel-spectrogram	mel-spectrogram	10	native male	native female	0.5	1.2	4.83 ± 0.34	0.20 ± 0.05
				10	native female	native male	0.8	2.1	5.32 ± 0.34	0.19 ± 0.07
				10	non-native female	native female	3.0	11.2	5.60 ± 0.41	0.19 ± 0.05
				10	non-native female	native male	3.0	8.5	6.10 ± 0.46	0.21 ± 0.07
3	3	PPG	PPG	10	native male	native female	0.5	0.7	4.79 ± 0.34	0.18 ± 0.05
				10	native female	native male	0.8	2.1	5.29 ± 0.34	0.17 ± 0.06
				10	non-native female	native female	3.0	8.4	5.58 ± 0.40	0.19 ± 0.05
				10	non-native female	native male	3.0	9.2	6.07 ± 0.47	0.21 ± 0.06

た。特に，L2 話者から L1 話者への変換においては 20 ms は十分な情報が取れず適切なアライメントが生成できないため変換後の音声が完全に崩れてしまう。

次に，アライメント特徴量による各 L2 話者からの変換精度の結果を Tabel 2 に示す。こちらの結果から，ソース話者の CER が低いときは PPG の方がメルスペクトログラムを用いるよりも変換後の CER が良くなるが，CER が高くなると L1 と発話情報が離れてくるため PPG がうまく取れなくなりメルスペクトログラムを用いる方が変換後の CER が良くなることがわかる。

3.3　主観評価実験結果

合成品質の主観評価には自然性評価に平均オピニオン評点 (MOS) と話者類似性評価実験 (Similarity) を用いた。主観評価には各モデルから 10 文ずつを抜き出し，日本人聴者 5 名がヘッドホンを用いて行った。また，話者類似性の評価には原音と変換後の音声が同じ話者かどうかを 4 段階で評価を行い，自然性評価は 5 段階で評価を行った。

主観評価実験の結果を Fig. 3 に示す。音質においてメルスペクトログラムを用いたアライメントを用いた方がどの条件においても安定し，高品質であることが分かった。また，話者性においては有意差は得られなかった。

4　おわりに

L2 話者から L1 話者への変換では，TTS タスクで提案された Alignment Module を用いるよりも，より VC タスクに適した形状にし，アライメント特徴量にメルスペクトログラムを用いることで変換精度が向上することが分かった。ただし，本稿の結果では CER は向上したものの，十分な精度ではないため Alignment Module の更なる改良が必要であり，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] T. Okamoto *et al.*, “E2E-S2S-VC: End-to-End sequence-to-sequence voice conversion,” in *Proc.Interspeech*, Aug. 2023.

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

Table 2　Result of Objective Evaluation for Alignment Feature.

source alignment feature	target alignment feature	source speaker	target speaker	source original CER [%]	converted CER [%]
mel-spectrogram	mel-spectrogram	non-native (Spanish) female	native female	3.0	11.2
		non-native (Spanish) female	native male	3.0	8.5
		non-native (Spanish) male	native female	6.4	31.3
		non-native (Spanish) male	native male	6.4	31.6
		non-native (Vietnamese) female	native female	7.1	24.4
		non-native (Vietnamese) female	native male	7.1	26.2
		non-native (Vietnamese) male	native female	13.7	31.9
		non-native (Vietnamese) male	native male	13.7	26.2
PPG	PPG	non-native (Spanish) female	native female	3.0	**8.4**
		non-native (Spanish) female	native male	3.0	**9.2**
		non-native (Spanish) male	native female	6.4	37.9
		non-native (Spanish) male	native male	6.4	41.1
		non-native (Vietnamese) female	native female	7.1	26.3
		non-native (Vietnamese) female	native male	7.1	23.3
		non-native (Vietnamese) male	native female	13.7	34.9
		non-native (Vietnamese) male	native male	13.7	36.8

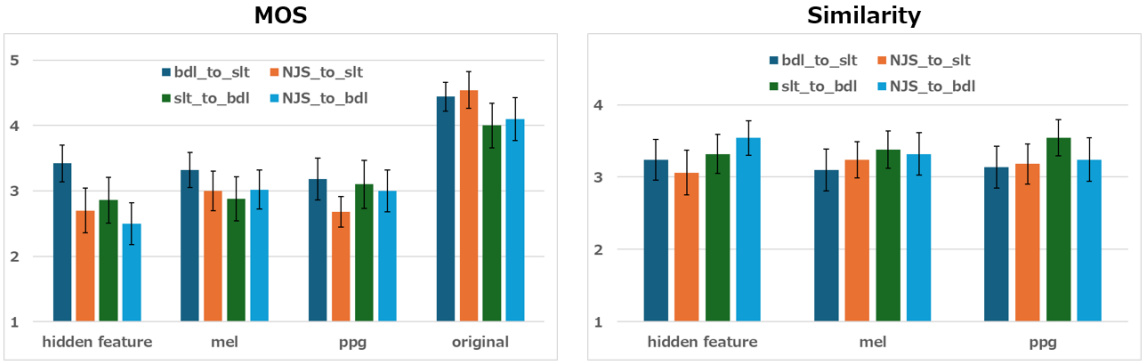

Fig. 3　Result of listening tests with 5 subjects. The error bars represent the 95 % confidence intervals. NJS is Spanish feamale, bdl is native male and slt is native female.

[6] 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.

[7] 山下ら, “EdenVC：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.

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

[9] 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.

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

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

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

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

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

[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. Okamoto *et al.*, “Challenge of singing voice synthesis using only text-to-speech corpus with FIRNet source-filter neural vocoder,” in *Proc. Interspeech*, Sept. 2024. (accepted, to appear)

[17] 小椋ら, “音素埋め込みスキップ接続を用いた継続長拡張に頑健な音声合成”, 音講論, Sept. 2024.

[18] S. Liu *et al.*, “Any-to-many voice conversion with location-relative sequence-to-sequence modeling,” in *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1717–1728, 2021.