Non-autoregressive Any to Many Voice Conversion

Yansen Zhou, Lejun Peng, Qi Chen*
School of Cyber Science and Engineering, University of International Relations, Beijing, China
**Email: zhouys@uir.edu.cn*

Abstract

Aiming at **the problems of Transformer model**, a voice conversion model based on gated attention unit (GAU) is proposed.

The gated attention unit integrates **the gated linear unit (GLU)** and **the self attention mechanism**.

Compared with the Transformer model, **the model has fewer parameters, lower memory occupation and faster training speed**. On the CSTR VCTK and CMU ARCTIC speech data sets, the subjective and objective evaluation of the two models proposed is carried out.

The experimental results show that the proposed two models have faster training speed, faster prediction speed and more stable prediction quality while ensuring the quality of Mel spectrum.

Introduction

The existing voice conversion technologies still face the following issues:

1) Traditional parallel voice conversion methods rely on parallel voice datasets, which limits their practicality;

2) one-to-one conversion models require separate training for each target speaker, making them inefficient for real-world applications.

Methods

- A. Non-autoregressive voice conversion model
- The model uses log F0, voiced/voiceless flag, and speaker ID as input, transformed by linear layers and positional encoding. Encoder and decoder (N layers, Transformer- or GAU-based) process features in parallel. Decoder output passes through a linear layer and a 5-layer 1D convolution Postnet to generate the Mel spectrum.

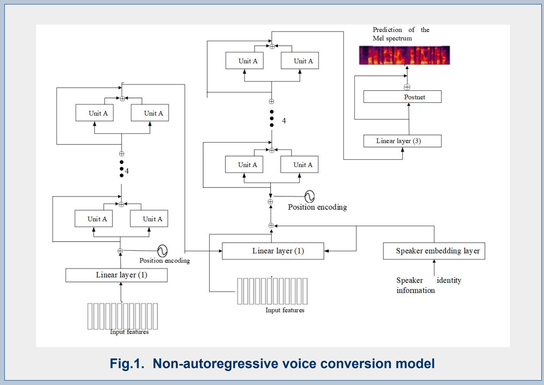

Fig.1. Non-autoregressive voice conversion model

- B. Transformation model based on gated attention unit GAU
- The GAU-based non-autoregressive voice conversion model improves the Transformer model using gated attention units and 1D convolutions, offering fewer parameters, faster speed, and better conversion quality while capturing local frame-level information.

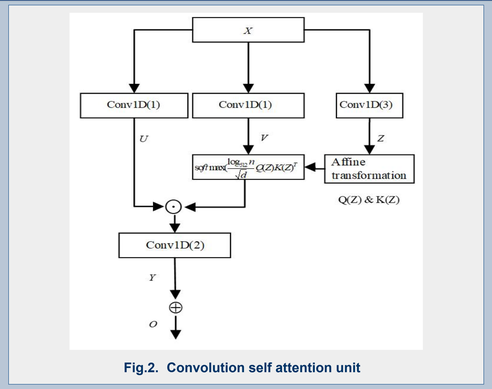

Fig.2. Convolution self attention unit

Results

- A. Speech conversion quality assessment
- The proposed model was evaluated on arbitrary-to-multilingual and single-sample VC tasks using MCD (objective) and MOS (subjective) metrics. Two baselines, DNN-VC (BiLSTM) and PPG-VC (Tacotron2-based), were compared across four conversion types (F-F, F-M, M-F, M-M). Results in Table IV show superior MCD and MOS performance of the proposed model.

Table 1. Subjective evaluations and objective evaluations

Method	MCD(dB)				MOS			
	DNN-VC	PPG-VC	Transformer-VC	GAU-VC	DNN-VC	PPG-VC	Transformer-VC	GAU-VC
F-F	6.931	6.821	6.843	**6.615**	2.93	2.91	**2.97**	2.91
F-M	6.752	6.706	6.639	**6.487**	2.88	3.07	2.94	**3.12**
M-F	6.723	6.803	6.856	**6.688**	2.67	2.94	2.90	**2.98**
M-M	6.681	6.743	6.75	**6.356**	2.96	3.22	3.24	**3.25**
Average	6.771	6.768	6.772	**6.536**	2.86	3.03	3.01	**3.06**

- B. Ablation experiments
- Three experiments were conducted: (1) without the arbitrary-to-one model, (2) without LogF0 & UV, and (3) replacing GAU with multi-head attention. Results (Table 6) show that the proposed feature processing and GAU design achieve the best performance, confirming their effectiveness

Table 2. Ablation experiments

Method	MCD(dB)				MOS			
	GAU-VC	1)	2)	3)	GAU-VC	1)	2)	3)
F-F	**6.615**	6.623	6.642	6.843	2.91	2.41	2.96	**2.97**
F-M	**6.487**	6.563	6.543	6.639	**3.12**	2.97	3.02	2.94
M-F	6.688	6.734	**6.601**	6.856	**2.98**	2.30	2.88	2.90
M-M	**6.356**	6.427	6.431	6.75	**3.52**	3.17	3.28	3.24
Average	**6.536**	6.586	6.554	6.772	**3.06**	2.71	3.03	3.01

Conclusions

This paper proposes a non-autoregressive voice conversion model based on non-parallel speech, which improves training and prediction speed while maintaining or surpassing the quality of autoregressive models. However, performance may decline with large speaking rate differences between speakers. Future work will address this limitation.

Acknowledgment

Supported by Research Funds for NSD Construction, University of International Relations