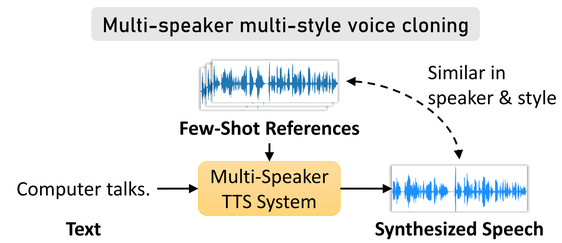
Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech

Chung-Ming Chien, Jheng-Hao Lin*, Chien-yu Huang*, Po-chun Hsu* and Hung-yi Lee
College of Electrical Engineering and Computer Science, National Taiwan University

Paper ID: 5685

* These authors contribute equally

I. Task Description



Challenges

- Extracting speaker and style information from limited references
- Generalized to different speakers/styles

ICASSP 2021 M2VoC Challenge

- Mandarin TTS
- Evaluated in
  - Quality
  - Speaker similarity
  - Style similarity
- 4 different tracks

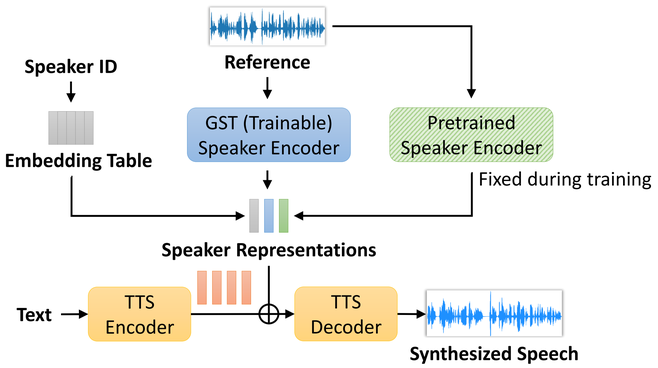
	Track 1A	Track 1B	Track 2A	Track 2B
# references	100	100	5	5
External data	No	Yes	No	Yes

II. Pretrained & Learnable Speaker Representations

Previous works

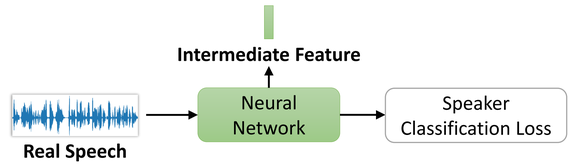
- Multi-speaker TTS by either a pretrained or a learnable speaker representation
- Pretrained:
  - D-vector [1]
  - X-vector [2]
- Learnable:
  - Embedding table
  - Global style token (GST) [3]

Proposed: use them all



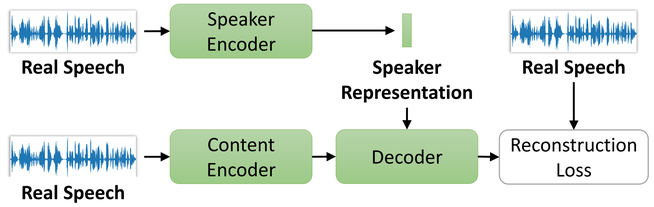
III. Generative Pretraining of Speaker Representations

Previous works: discriminative pretraining



D-vector / X-vector

Proposed: generative pretraining



AdaIN (one-shot) voice conversion (VC) [4]

IV. Experiments

- AIShell-3 + M2VoC official dataset (including the testing speakers)
  - 96 hours
  - 230 speakers

Performance of different speaker representations

Ratio of utterances passing a speaker verification system

TTS Model	Speaker representation					Results (↑)	
	Pretrained			Learnable		SV accuracy	
	D-vec	X-vec	VC	Embed	GST	Track 1	Track 2
Tacotron 2	√					.772	.367
		√				.785	.377
			√			**.942**	**.727**
				√		.630	.703
					√	.102	.050
FastSpeech 2	√					.977	.323
		√				.973	.623
			√			.980	**.837**
				√		**.988**	.490
					√	.778	.390

Combining multiple speaker representations

Ratio of utterances passing a speaker verification system

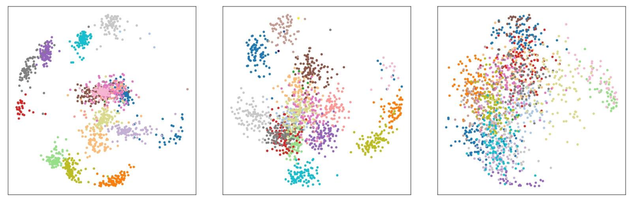
TTS Model	Speaker representation					Results (↑)	
	Pretrained			Learnable		SV accuracy	
	D-vec	X-vec	VC	Embed	GST	Track 1	Track 2
FastSpeech 2			√			.980	.837
	√		√			.978	.747
		√	√			.992	.860
			√	√		.983	**.937**
			√		√	.982	.783
			√	√	√	.988	.897
	√	√	√	√	√	.990	.887

Subjective evaluation

MOS on speaker similarity and naturalness

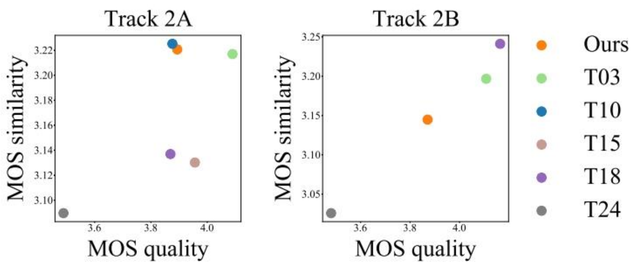
	X-vec	VC	Embed	VC+Embed
Quality (↑)	3.47±.13	3.61±.13	**3.65±.13**	3.55±.12
Similarity (↑)	3.25±.13	3.19±.14	3.27±.13	**3.38±.14**

Speaker representation scatter plots



(a) d-vector (b) x-vector (c) VC

Official subjective evaluation



V. Takeaway

- Combining Pretrained & learnable speaker representations are better than using a single representation in TTS
- Generative speaker pretraining (e.g. voice conversion) outperforms discriminative pretraining
- Ranked 2nd in track 2A and 3rd in track 2B of ICASSP 2021 M2VoC challenge

Code Audio samples

[1] Wan et al., Generalized End-to-End Loss for Speaker Verification
[2] Snyder et al., X-vectors: Robust DNN Embeddings for Speaker Recognition
[3] Wang et al., Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis
[4] Chou et al., One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization