An Unified Understanding of Voice Conversion and its Medical Application

Wen-Chin Huang
Nagoya University, Japan

名古屋大学
NAGOYA UNIVERSITY

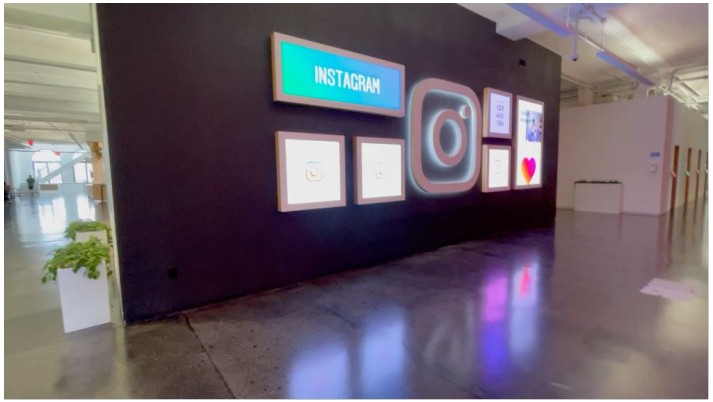
Who am I?

A PhD student in Japan who spent most of his graduate school life in Taiwan.

A 10 years street dancer.

A trilingual speaker. (Mandarin, English, Japanese)

A two-time Meta intern.



An unified understanding of voice conversion techniques

Just a personal opinion!

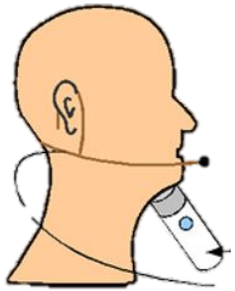
- Definition and goal
- Sidestory:
  An observation in top conf papers
- My understanding of VC techniques

Voice conversion (VC)

Definition: converts one kind of speech to another while keeping the linguistic content.

Doesn't have to be speaker conversion! **Applications**:

- Speaker conversion (Detective Conan, deepfake, etc.)
- Accent conversion (international customer service)
- Electrolarygeal conversion (speech organ disability)
- ...and more



Sanas aims to convert one accent to another in real time for smoother customer service calls

Devin Coldewey @techcrunch / 7:23 PM EDT • August 31, 2021 Comment

https://techcrunch.com/2021/08/31/sanas-aims-to-convert-one-accent-to-another-in-real-time-for-smoother-customer-service-calls/

Forbes

CYBERSECURITY • EDITORS' PICK

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

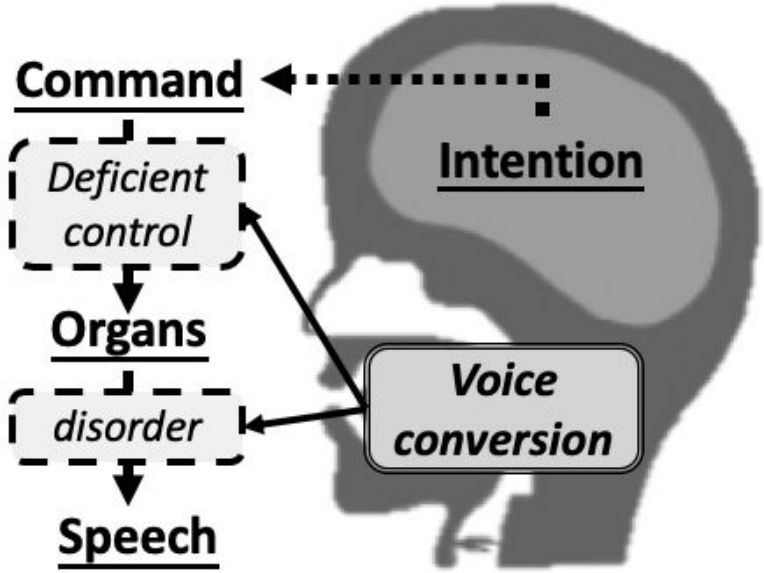
https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=3eb9bf375591

Ultimate goal of voice conversion: augmented communication

Physical condition of the human body often limits the production of speech.

Ex 1. Deficient control of the organs
→ accented voice
Solution: Convert into native speech

Ex 2. Damaged speech organs
→ severe vocal disorders
Solution: Speaking aid devices to restore natural voice.



It's easy to think of new VC application!

Side story... you might have heard these papers:

Autovc: Zero-shot voice style transfer with only autoencoder loss

K …

Multi-target voice conversion without parallel data by adversarially learning di…

… J …

☆ …

Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks

T Kaneko, H Kameoka - 2018 26th European Signal …, 2018 - ieeexplore.ieee.org

… VC method [10] even though CycleGAN-VC is trained under disadvantageous conditions (non… In Section III, we review the CycleGAN and explain our proposed method (CycleGAN-VC). …

☆ 儲存 99 引用 被引用 183 次 相關文章 全部共 5 個版本

Or… wi…

J C…

… This kind of factorization enable our model to perform one-shot voice conversion as follows: with one utterance from source speaker and another utterance from target speaker, we first …

☆ 儲存 99 引用 被引用 119 次 相關文章 全部共 9 個版本 »

StarGAN-VC does not require any … Although the concept is similar to our StarGAN-VC approach, …

☆ 儲存 99 引用 被引用 284 次 相關文章 全部共 5 個版本

Speech-related papers getting accepted to top confs! But...

IMPROVING ZERO-SHOT VOICE STYLE TRANSFER VIA DISENTANGLED REPRESENTATION LEARNING

Published as a conference paper at ICLR 2021 https://arxiv.org/pdf/2103.09420.pdf

5.3 STYLE TRANSFER PERFORMANCE

We test our model with four competitive baselines: Blow (Serrà et al., 2019)[3], AUTOVC (Qian et al., 2019), AdaIN-VC (Chou & Lee, 2019) and StarGAN-VC (Kameoka et al., 2018). The de-

Metric	Objective		Subjective	
	Distance	Verification[%]	Naturalness [1–5]	Similarity [%]
StarGAN	6.73	71.1	2.77	51.5
AdaIN-VC	6.98	85.5	2.19	50.8
AUTOVC	6.73	89.9	3.25	55.0
Blow	8.08	-	2.11	10.8
IDE-VC (Ours)	**6.70**	**92.2**	**3.26**	**68.5**

Representation learning is hot! But...

Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia. https://arxiv.org/pdf/2110.14513.pdf

total. We trained three baseline models with official implementations - VQVC+ [51], AdaIN [10], AUTOVC [36] - using the same dataset and mel spectrogram configuration as NANSY. For a fair

	M2M			A2M			A2A		
	CER[%]	MOS[1-5]	SSIM[%]	CER[%]	MOS[1-5]	SSIM[%]	CER[%]	MOS[1-5]	SSIM[%]
SRC as TGT	n/a	4.23 $\pm$ 0.05	0	n/a	4.28 $\pm$ 0.09	0.60	n/a	4.26 $\pm$ 0.07	0.25
TGT as TGT	n/a	4.32 $\pm$ 0.05	94.9	n/a	4.29 $\pm$ 0.05	92.4	n/a	4.27 $\pm$ 0.07	96.2
VQVC+	54.0	1.76 $\pm$ 0.05	54.5	74.7	1.73 $\pm$ 0.11	15.6	69.3	1.83 $\pm$ 0.09	13.8
AdaIN	62.9	2.22 $\pm$ 0.07	24.0	79.6	1.92 $\pm$ 0.12	18.1	59.3	2.12 $\pm$ 0.10	21.2
AUTOVC	31.7	3.41 $\pm$ 0.06	47.3	36.1	2.74 $\pm$ 0.11	33.2	28.2	2.59 $\pm$ 0.08	23.3
NANSY	**7.5**	**3.79** $\pm$ 0.07	**91.4**	**7.6**	**3.73** $\pm$ 0.05	**88.1**	**8.6**	**3.44** $\pm$ 0.07	**64.6**

Here's a recent one...

DIFFUSION-BASED VOICE CONVERSION WITH FAST MAXIMUM LIKELIHOOD SAMPLING SCHEME

ICLR 2022.
https://arxiv.org/pdf/2109.13821.pdf

4.2 ANY-TO-ANY VOICE CONVERSION

We chose four recently proposed VC models capable of one-shot many-to-many synthesis as the baselines:

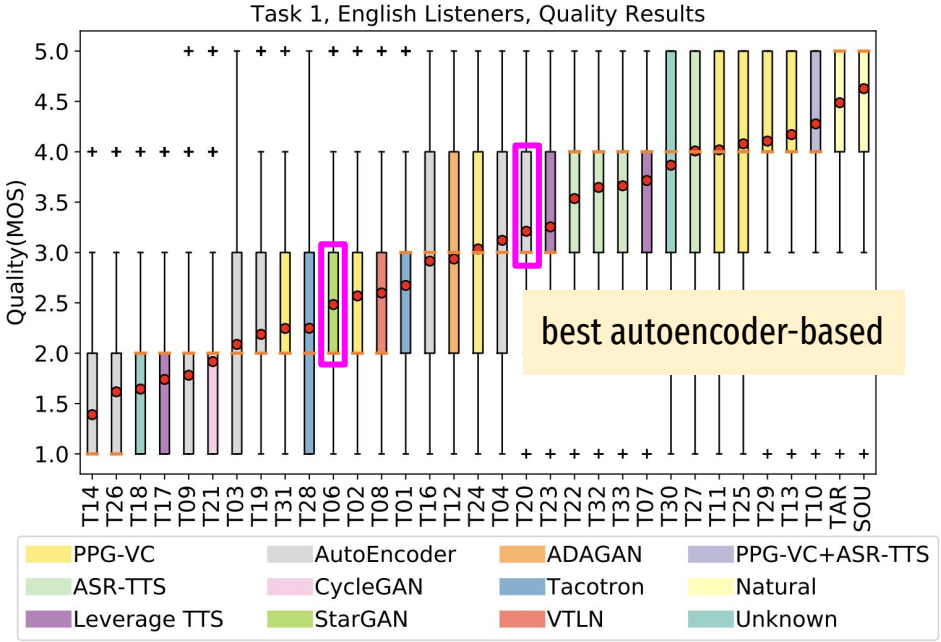
- *AGAIN-VC* (Chen et al., 2021b), an improved version of a conventional autoencoder AdaIN-VC solving the disentanglement problem by means of instance normalization;
- *FragmentVC* (Lin et al., 2021), an attention-based model relying on wav2vec 2.0 (Baevski et al., 2020) to obtain speech content from the source utterance;
- *VQMIVC* (Wang et al., 2021), state-of-the-art approach among those employing vector quantization techniques;

	VCTK test (9 speakers, 54 pairs)		Whole test (25 speakers, 350 pairs)	
	Naturalness	Similarity	Naturalness	Similarity
AGAIN-VC	1.98 ± 0.05	1.97 ± 0.08	1.87 ± 0.03	1.75 ± 0.04
FragmentVC	2.20 ± 0.06	2.45 ± 0.09	1.91 ± 0.03	1.93 ± 0.04
VQMIVC	2.89 ± 0.06	2.60 ± 0.10	2.48 ± 0.04	1.95 ± 0.04
Diff-VCTK-ML-6	$\mathbf{3.73 \pm 0.06}$	3.47 ± 0.09	3.39 ± 0.04	$\mathbf{2.69 \pm 0.05}$
Diff-VCTK-ML-30	$\mathbf{3.73 \pm 0.06}$	$\mathbf{3.57 \pm 0.09}$	$\mathbf{3.44 \pm 0.04}$	$\mathbf{2.71 \pm 0.05}$
Ground truth	4.55 ± 0.05	4.52 ± 0.07	4.55 ± 0.05	4.52 ± 0.07

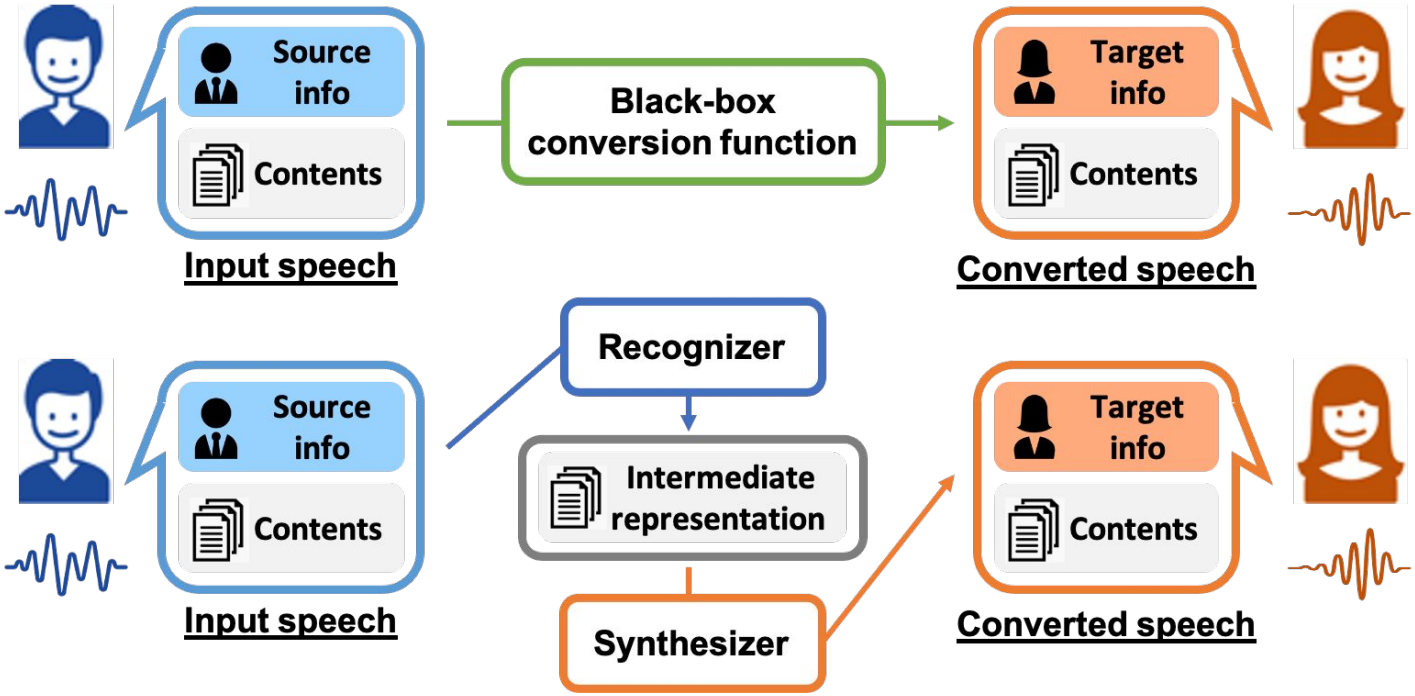
The truth is...

Voice Conversion Challenge 2020
- Intra-lingual semi-parallel and cross-lingual voice conversion -

Zhao Yi[1⋆], *Wen-Chin Huang*[2⋆], *Xiaohai Tian*[3⋆], *Junichi Yamagishi*[1⋆],
Rohan Kumar Das[3], *Tomi Kinnunen*[4], *Zhenhua Ling*[5], *Tomoki Toda*[2]



To me, there are only two approaches to voice conversion…



Black-box function: how people did voice conversion 30 years ago

Voice conversion through vector quantization
M Abe, S Nakamura, K Shikano… - Journal of the Acoustical …, 1990
… VOICE CONVERSION THROUGH VECTOR QUANTIZATION Our consists of two steps:a learning step and a conversion-synthesis step
☆ 儲存 99 引用 被引用 817 次 相關文章 全部共 12 個版本

Continuous probabilistic transform for voice conversion
Y Stylianou, O Cappé… - IEEE Transactions on …, 1998 - ieeexplore.ieee.org
… For the same reason, voice conversion techniques would … Finally, it is interesting to note that the voice conversion problem … and voice conversion is that in the case of voice conversion, …
☆ 儲存 99 引用 被引用 1292 次 相關文章 全部共 14 個版本

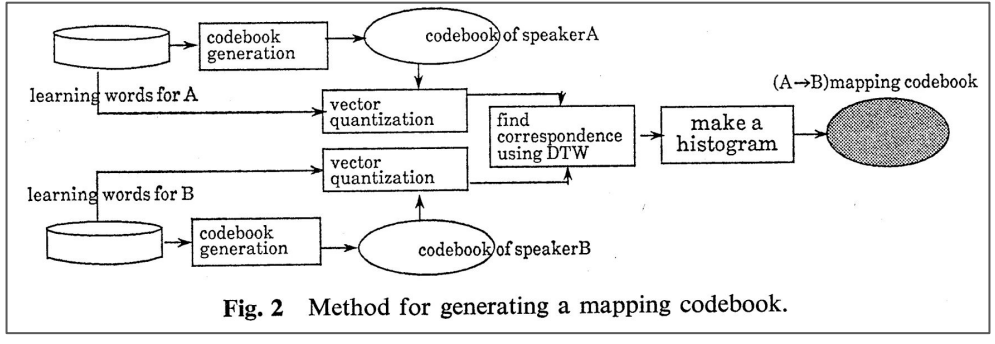


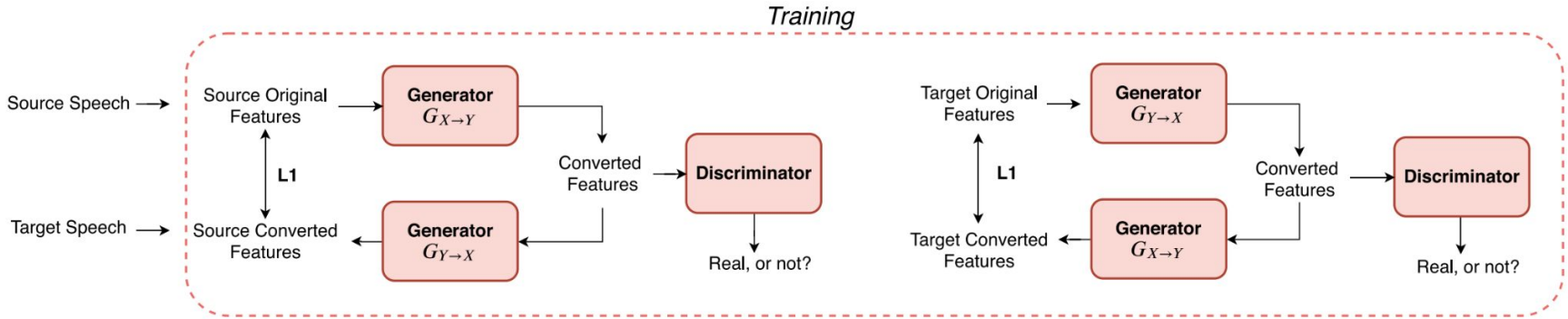
Fig. 2 Method for generating a mapping codebook.

Given an input speech segment (a frame, several frames, the whole sentence…), convert without knowing what the content is.

Requires a parallel dataset to learn the mapping (a.k.a. parallel VC)

Only black-box VC without parallel data: Cycle-GAN VC

Parallel VC: requires a training dataset, with same contents from source and target speakers.
Nonparallel VC: does not require the above.



https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9262021

What would you do if you were asked to perform VC ?

BUSINESS | TECH | ARTIFICIAL INTELLIGENCE

This AI startup claims to automate app making but actually just uses humans

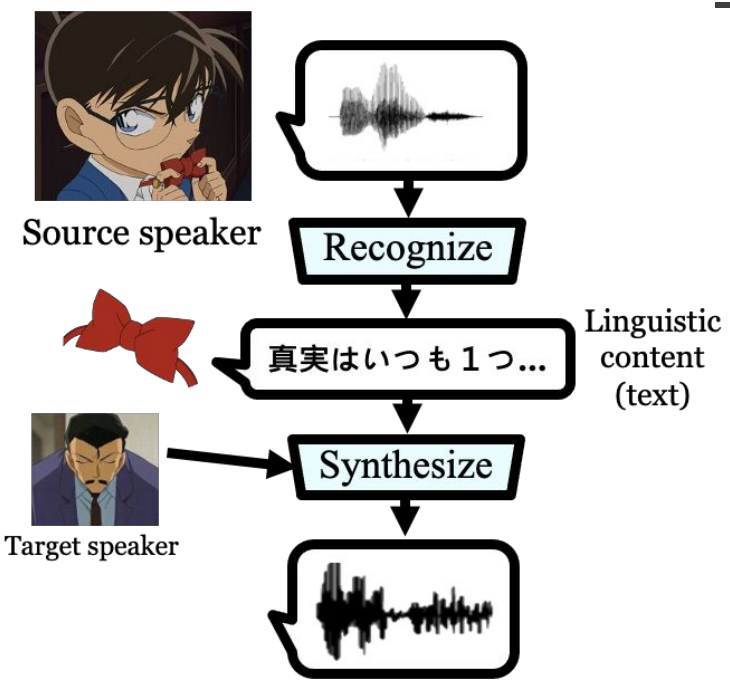
Who could have seen that coming?

By Nick Statt | @nickstatt | Aug 14, 2019, 1:58pm EDT | 11 comments

https://www.theverge.com/2019/8/14/20805676/engineer-ai-artificial-intelligence-startup-app-development-outsourcing-humans

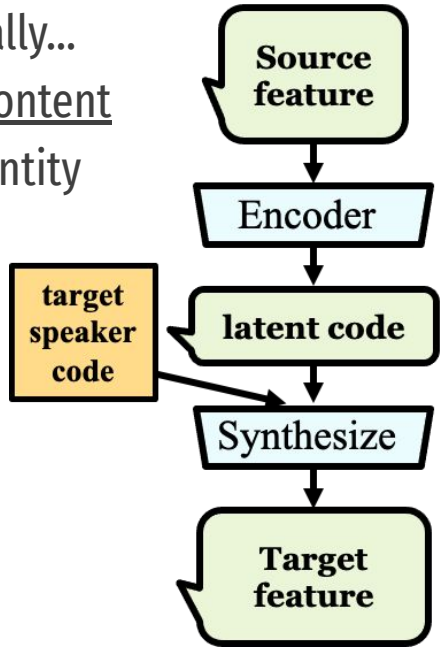
Arturo de Albornoz@Flickr

Conan's example



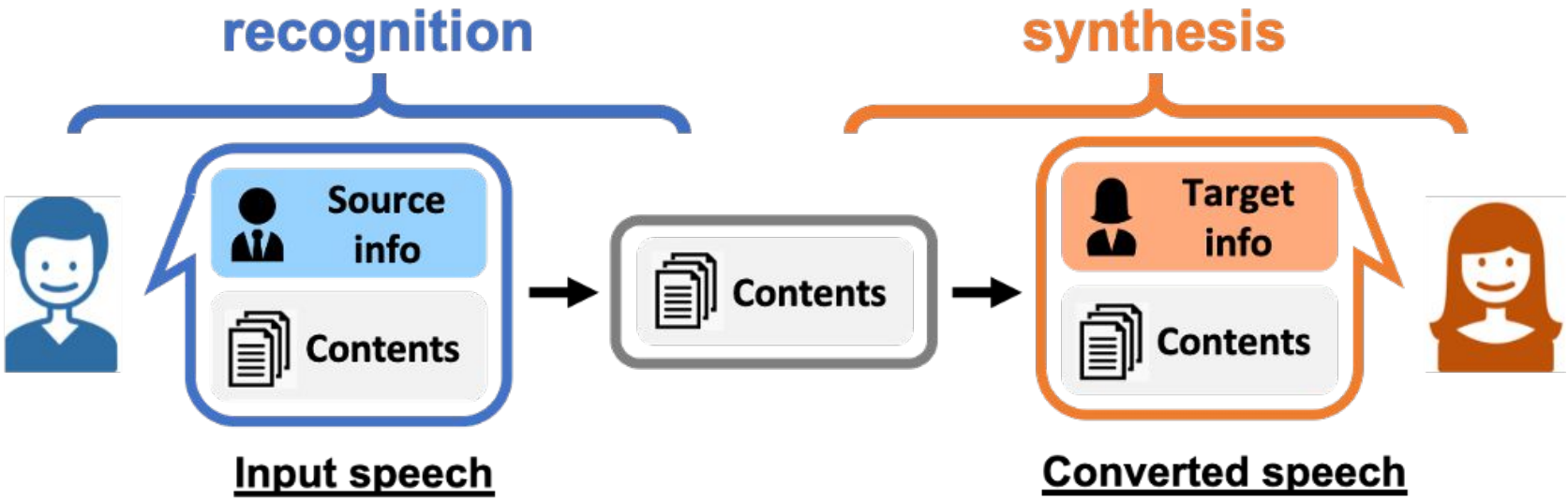
VAE-VC [Hsu; '16]

- Learns to automatically...
  - extract linguistic content
  - model speaker identity



C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational autoencoder," Proc. APSIPA 2016

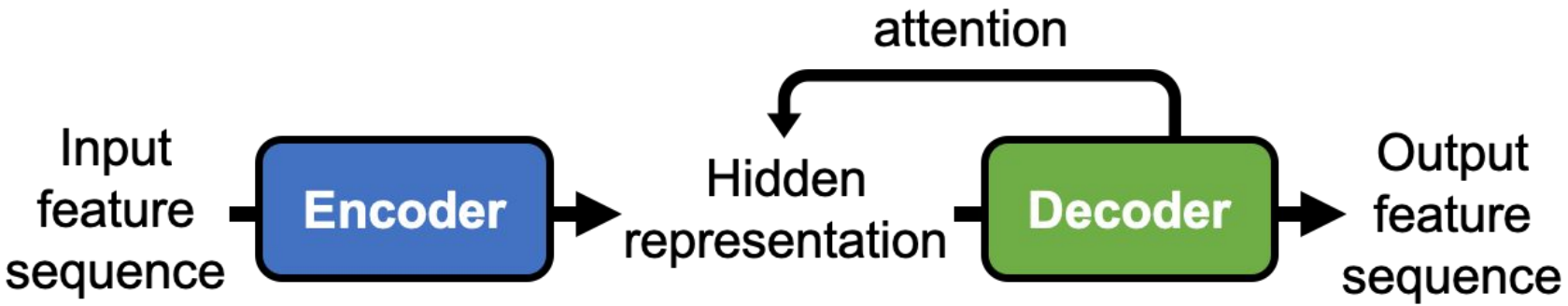
Recognition-synthesis based voice conversion (Rec-syn VC)



Rec-syn VC is … everywhere!

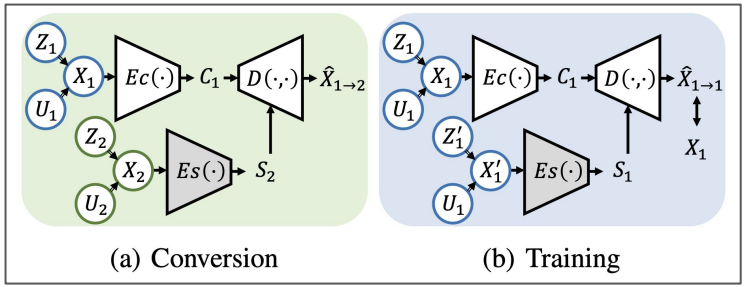
Best model architecture for ***parallel VC*: Voice Transformer Network**

Huang, W., Hayashi, T., Wu, Y., Kameoka, H., Toda, T. (2020) Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining. Proc. Interspeech 2020, 4676-4680



Rec-syn VC is ... everywhere!

And of course... the **autoencoder** family.



AUTOVC

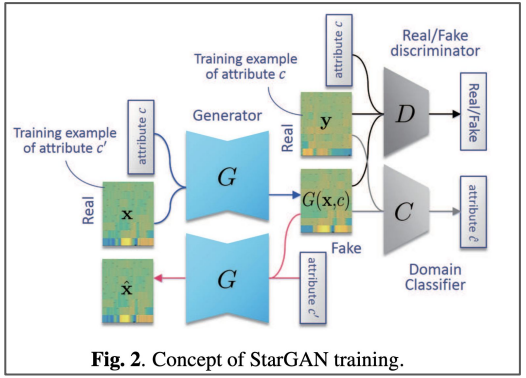
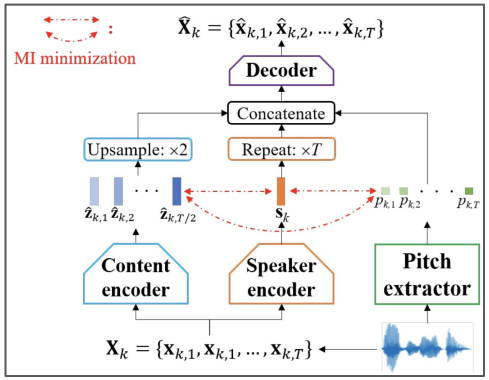
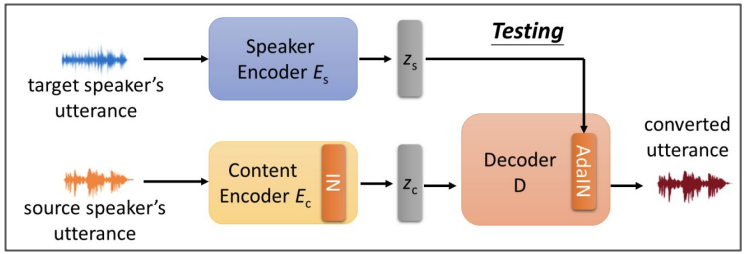


Fig. 2. Concept of StarGAN training.

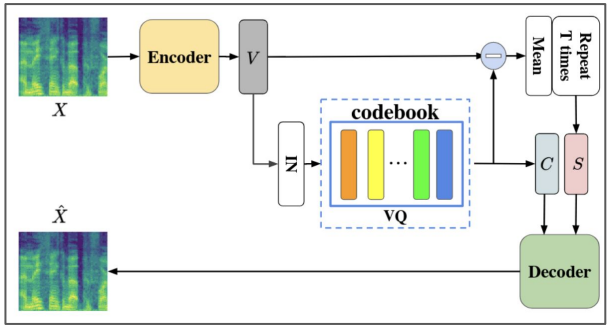
STARGAN-VC



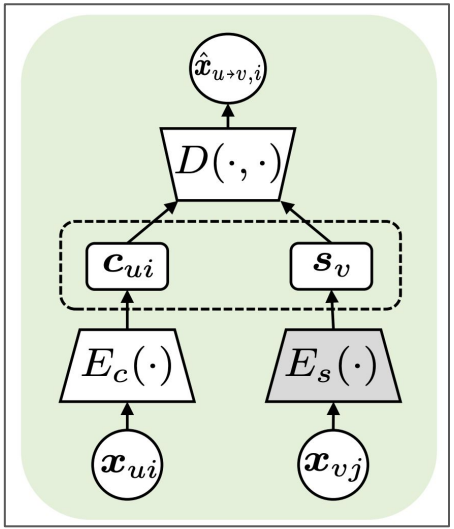
VQMIVC



AdaIN-VC



VQVC

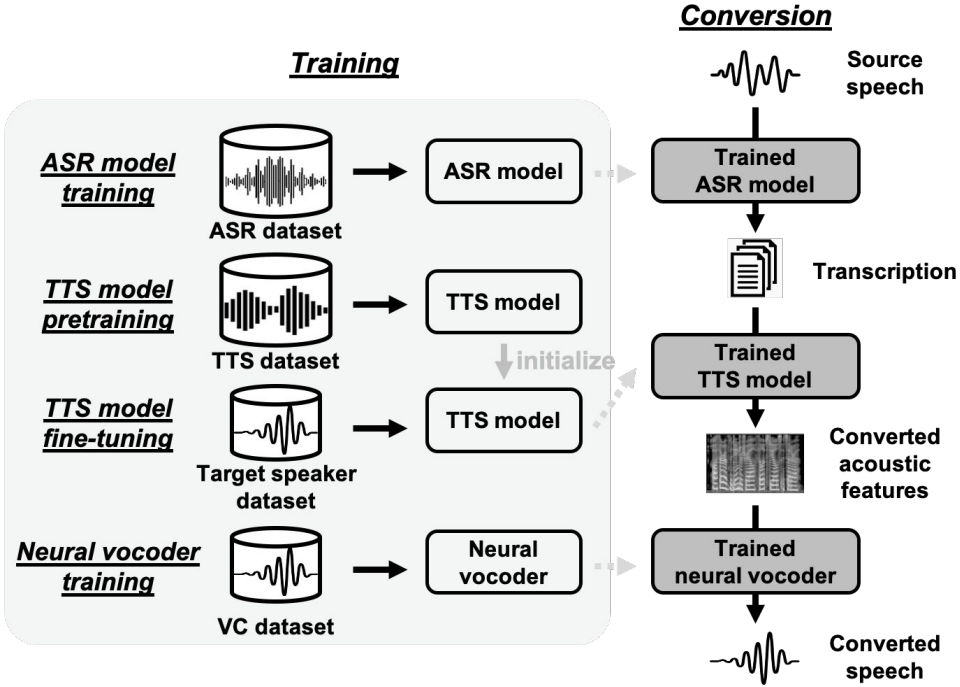


IDE-VC

Rec-syn VC is … everywhere!

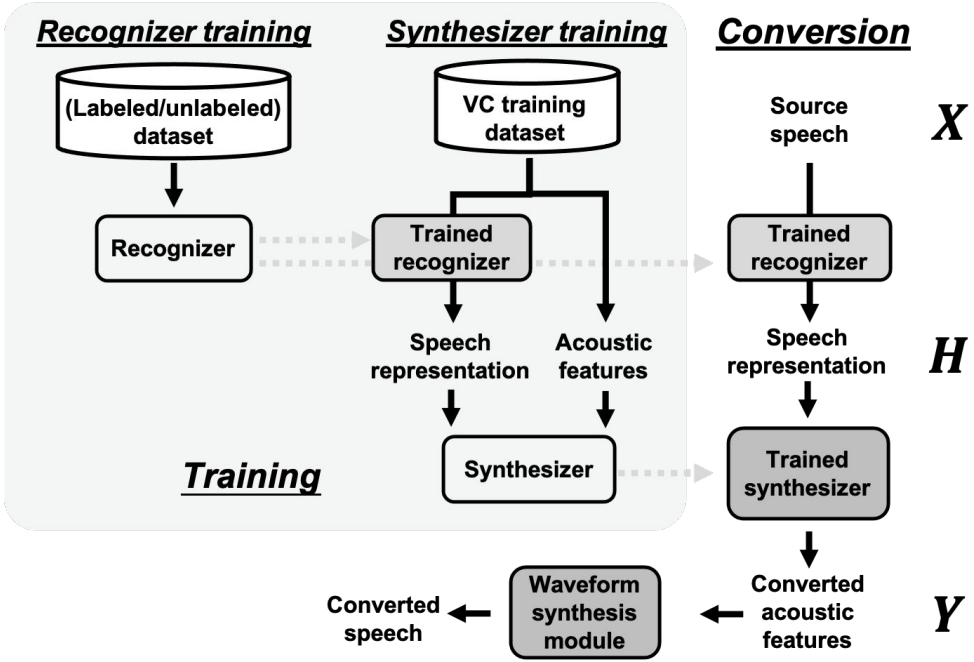
How about just concatenating **separately** trained ASR and TTS model?

Huang, W., Hayashi, T., Watanabe, S., Toda, T. (2020) The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS. Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 160-164



Rec-syn VC is … everywhere!

A more general idea of concatenating separately trained recognition and synthesis models

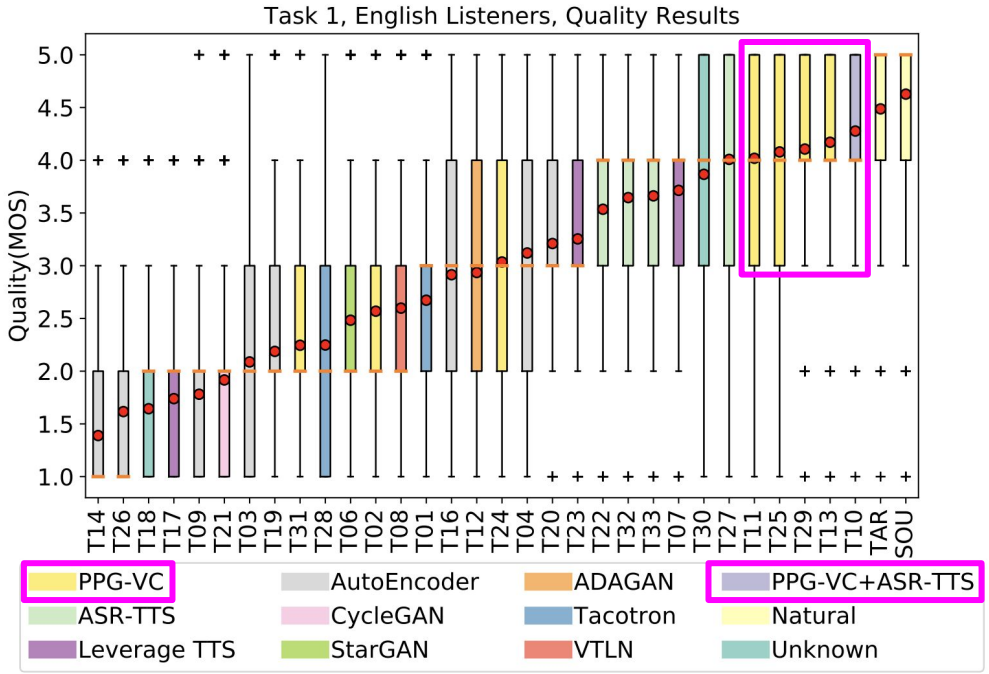


Rec-syn VC is … everywhere!

One can use different intermediate representations.

Separately trained rec-syn VC systems are still dominant even up to now.

Representation	Text	Phonetic Posteriorgram	Self-supervised speech representations
Extractor	ASR model		self-supervised model
Training data	labeled data		unlabeled data
Resolution	token level	frame level	



Voice conversion for dysarthric speech

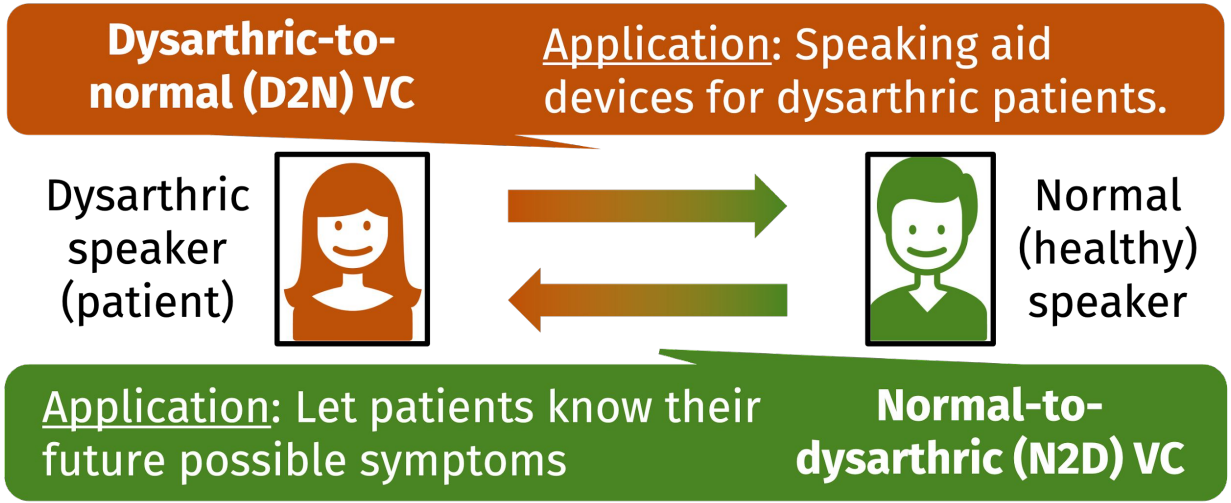
A medical application

What is dysarthria?

A type of speech disorder caused by disruptions in the **neuromotor** interface

Dysarthric speech: unnatural and unintelligible speech, ex. phoneme loss, unstable prosody, and imprecise articulation.

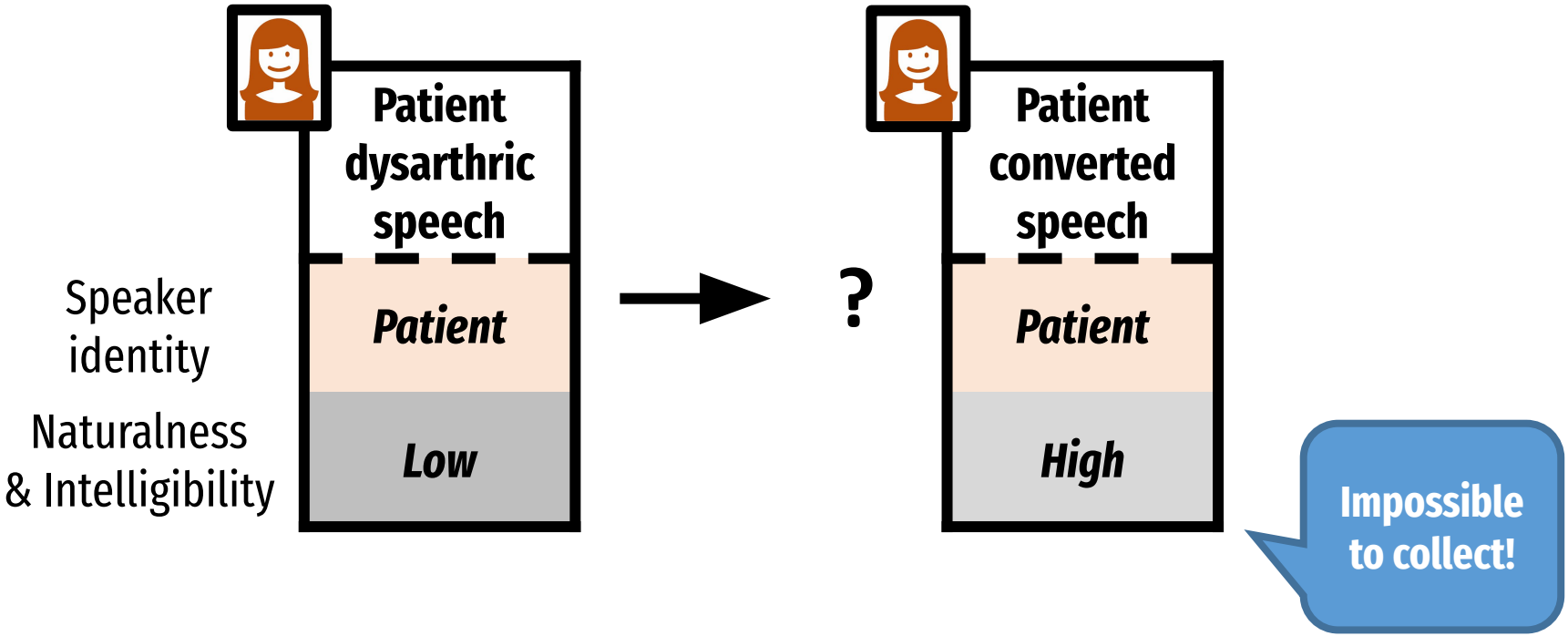
Dysarthric voice conversion: two directions



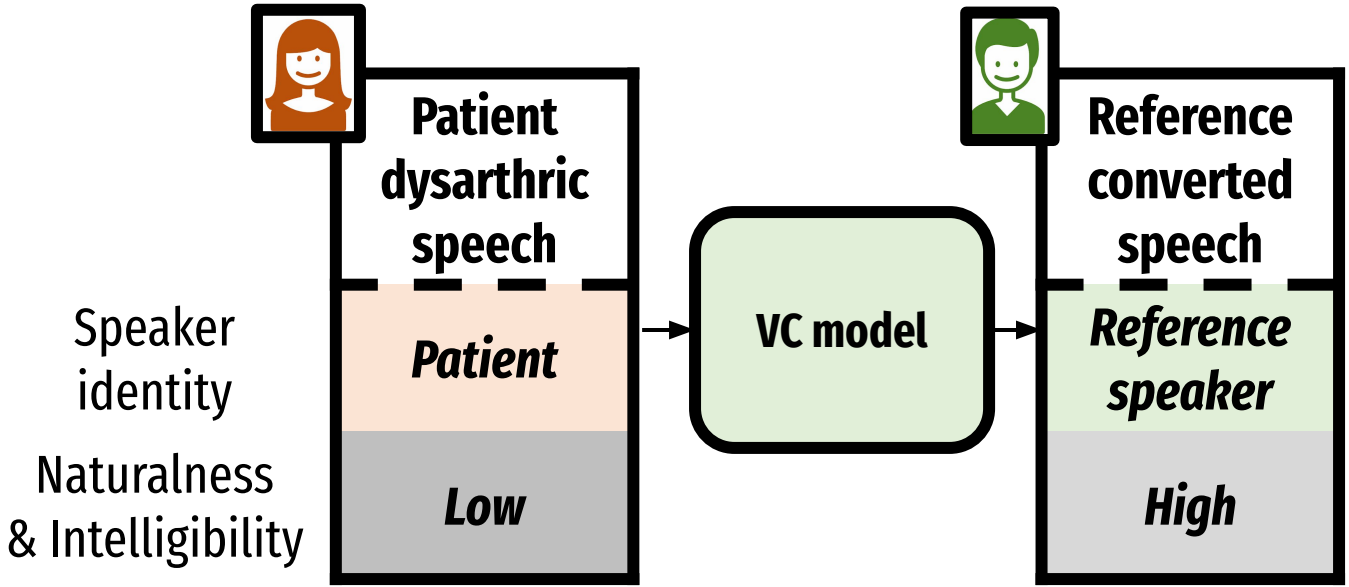
Goal:

1. High quality naturalness & correct intelligibility level
2. Maintaining the speaker identity of the source speaker

Problem: naturally ~~low~~no-resourced

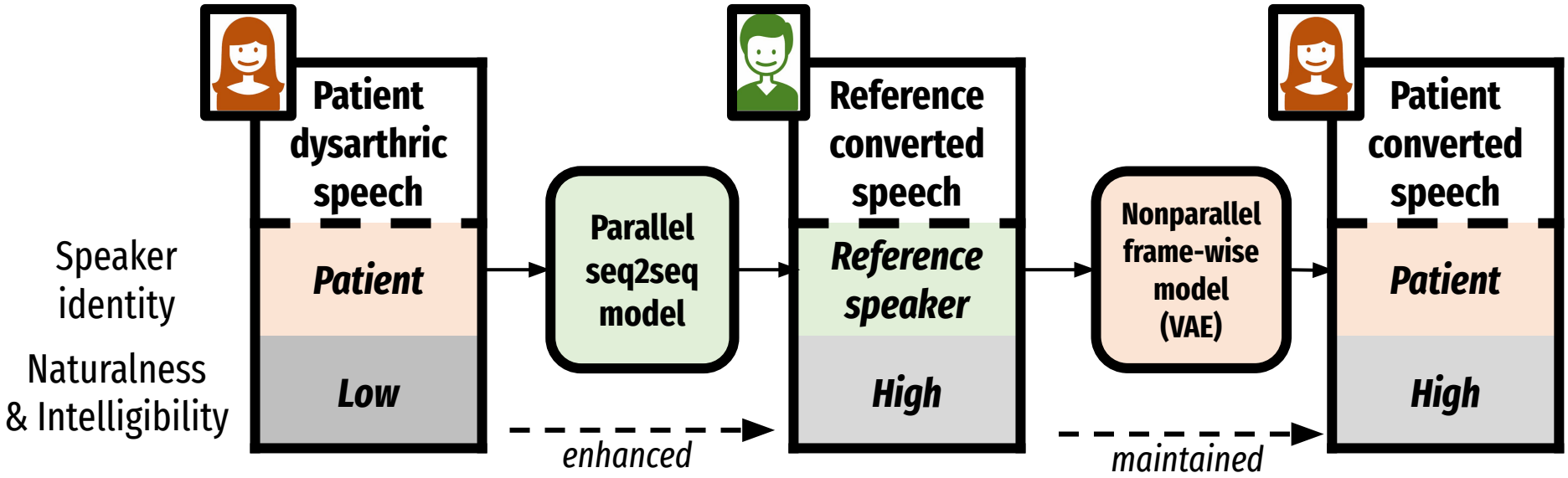


Past work: record a parallel corpus from a normal reference speaker



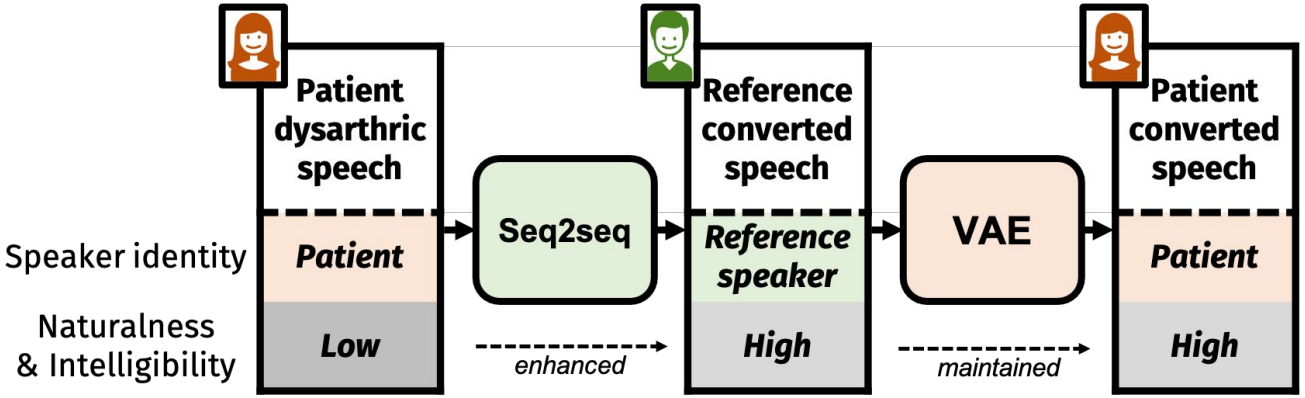
We want the converted speech to sound like the patient!

How about... a two-stage approach for maintaining identity?

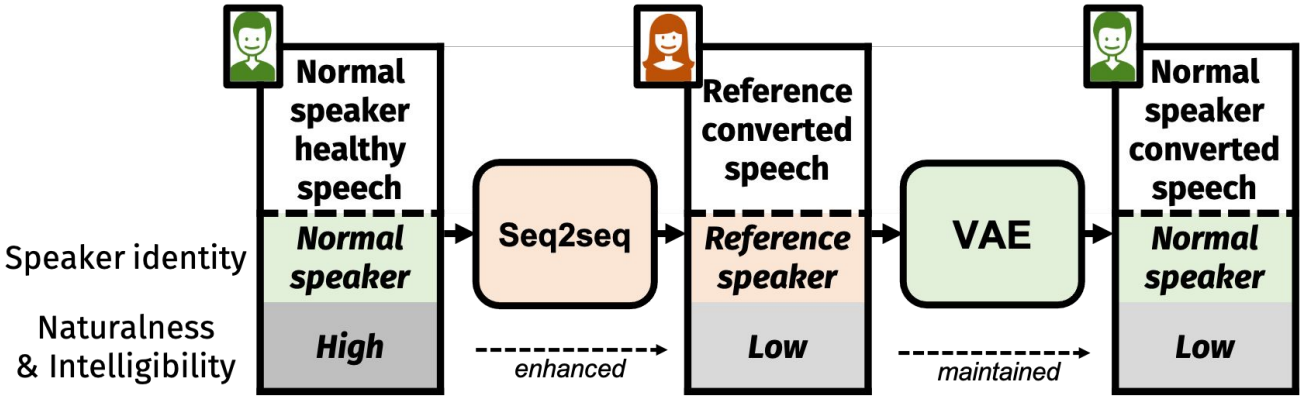


Generalizable to both directions!

D2N VC



N2D VC



Evaluation

D2N VC

Improve intelligibility (WER: 94% → 75.8%)

Improve naturalness (MOS: 2.37 → 2.65)

Barely maintain patient's identity (49%)

N2D VC

Achieve good naturalness results

Mimic the dysarthric characteristics

Convert away from the reference speaker's identity

Poorly maintain the source speaker's identity

Demo samples

D2N VC: https://unilight.github.io/Publication-Demos/publications/dvc-vtn-vae/index.html

N2D VC: https://unilight.github.io/Publication-Demos/publications/n2d-vc/index.html

Takeaways

Ultimate goal of voice conversion: augmented communication

There are two kinds of voice conversion techniques: black-box, recognition-synthesis

Recognition-synthesis is the mainstream.

Separately training is currently still better than joint training.