音声変換の紹介
(Introduction to voice conversion)

HUANG Wen-Chin（ホワン　ウェンチン）
名古屋大学　情報学研究科　戸田研究室　助教

自己紹介

- **HUANG Wen-Chin (ホワン　ウェンチン)**
- **出身：台湾　台北市**
- **略歴**
  - 2018.06　　　台湾大学　情報学部　卒業(学士)
  - 2021.03　　　名古屋大学大学院 情報学研究科 博士前期課程 修了(修士)
  - 2024.03　　　名古屋大学大学院 情報学研究科 博士後期課程 修了(博士)
  - 2024.04-現在 名古屋大学大学院 情報学研究科 戸田研究室 助教
- **研究分野：音声合成・変換・評価**

Outline

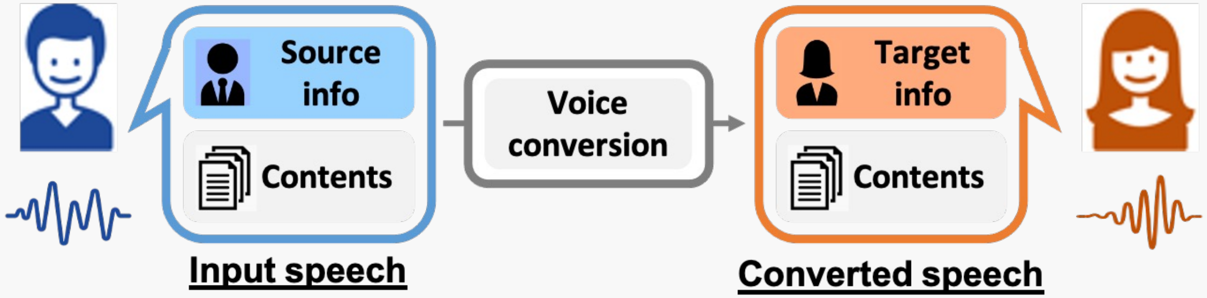
- What is voice conversion (VC)? Why do we need it?
- How do we build a VC system?
- What is difficult in VC?

Outline

- **What is voice conversion (VC)? Why do we need it?**
- How do we build a VC system?
- What is difficult in VC?

What is voice conversion (VC)?

- **Definition:**
  Converts one kind of speech to another while **keeping the linguistic content**.



- **The most common type: speaker conversion**
  - Convert between two speakers
  - Famous example (in Japan): Detective Conan

“VC” basically refers to “speaker conversion” if not specified

VC can be dangerous: deep fake (singer copying)

Original song
From Taiwanese singer
Jay Chou

Classical song from
Singapore singer
Stefanie Sun

Copied song

The singer did not know her voice was being copied!

VC can be dangerous: fraud

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

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

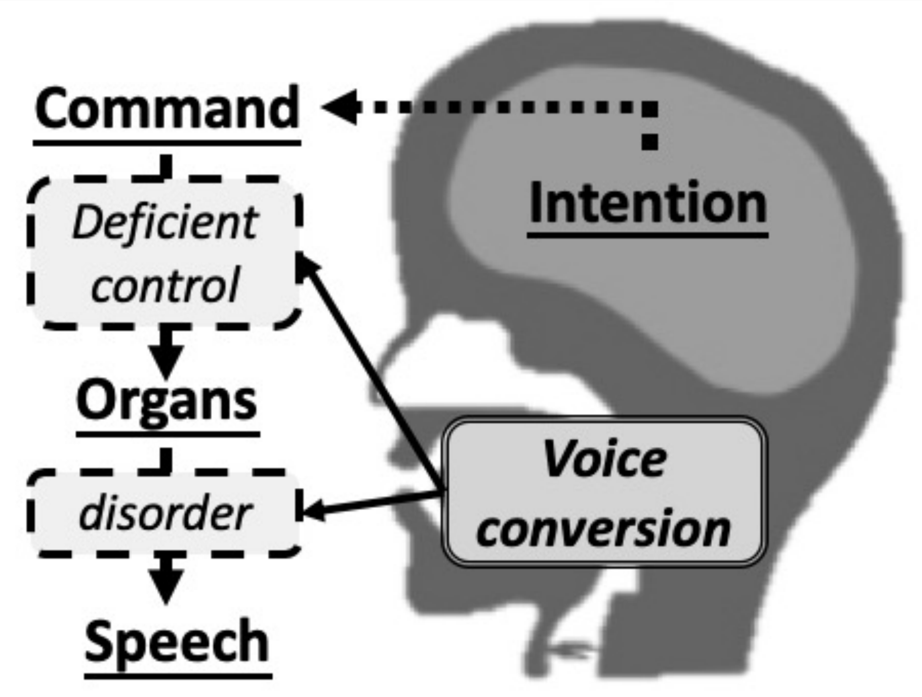
It should be used in a good way!

So why do we need VC? – An ultimate goal [Toda '14]

- **Augmented Communication**
  - Physical condition of the human body limits the production of speech.
  - VC can be used to break the barrier.

Ex 1. Deficient control of the organs
Result: accented voice

Ex 2. Damaged speech organs
Result: severe vocal disorders

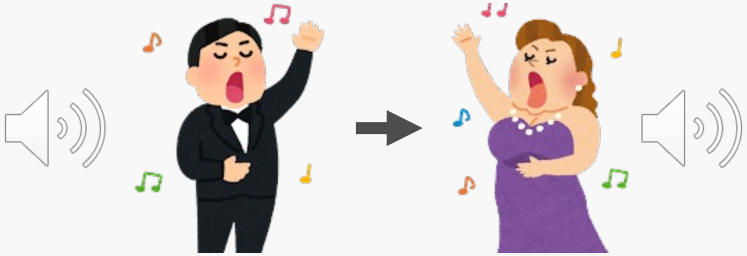


We can think of a lot of applications…

Let's start by asking ourselves: what can't I do?

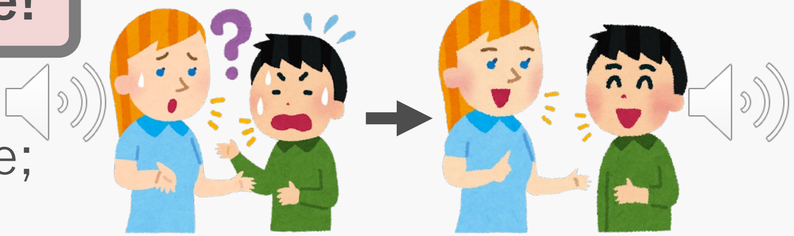
I cannot sing!

→ Singing voice conversion
- Application: autotune,Vtuber, etc.



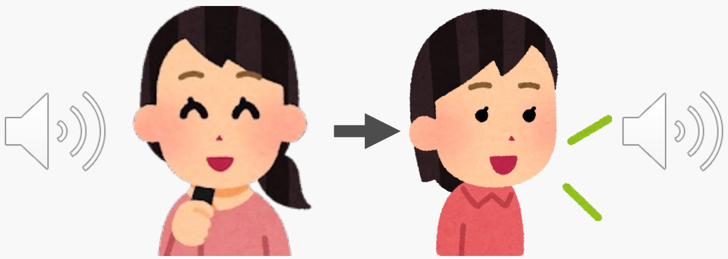
I cannot speak good Japanese!

→ Accent conversion
- Application: customer service; language learning



I cannot speak! (maybe in the future)

→ Electrolarygeal VC
- Application: speaking-aid



Interesting application that I really like: dubbing

[Gan+ '22]

Original

Converted

https://wblgers.github.io/IQDUBBING-VC.github.io/

Outline

- What is voice conversion (VC)? Why do we need it?
- How do we build a VC system?
- What is difficult in VC?

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

Parallel VC: how people did VC research 30 years ago

[Abe+ ‘90] [Stylianou+ ’98]

- **Idea: collect a <u>parallel corpus</u>, and find a mapping function**

たったひとつの真実見抜く

たったひとつの真実見抜く

見た目は子ども、頭脳は大人

見た目は子ども、頭脳は大人

その名は名探偵コナン！

その名は名探偵コナン！

Collect utterances with same contents by the source & target speaker

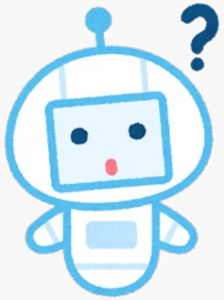
Parallel VC: how people did VC research 30 years ago

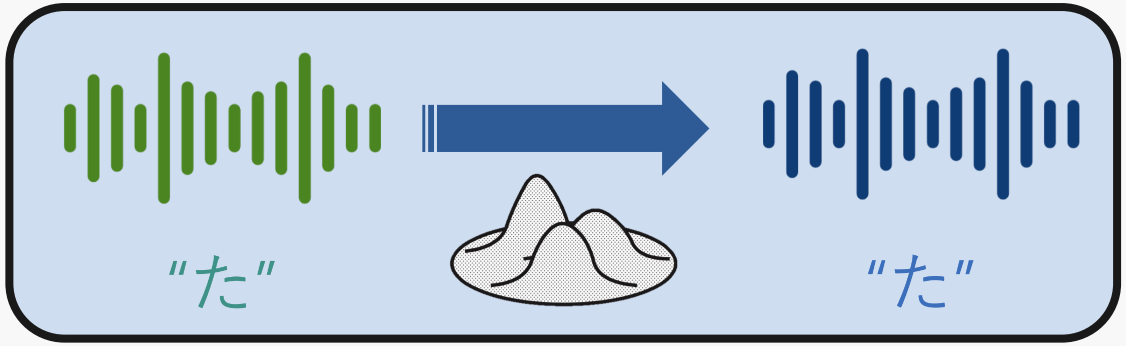
[Abe+ '90] [Stylianou+ '98]

- **Idea: collect a parallel corpus, and find a mapping function**

たったひとつの真実見抜く

たったひとつの真実見抜く





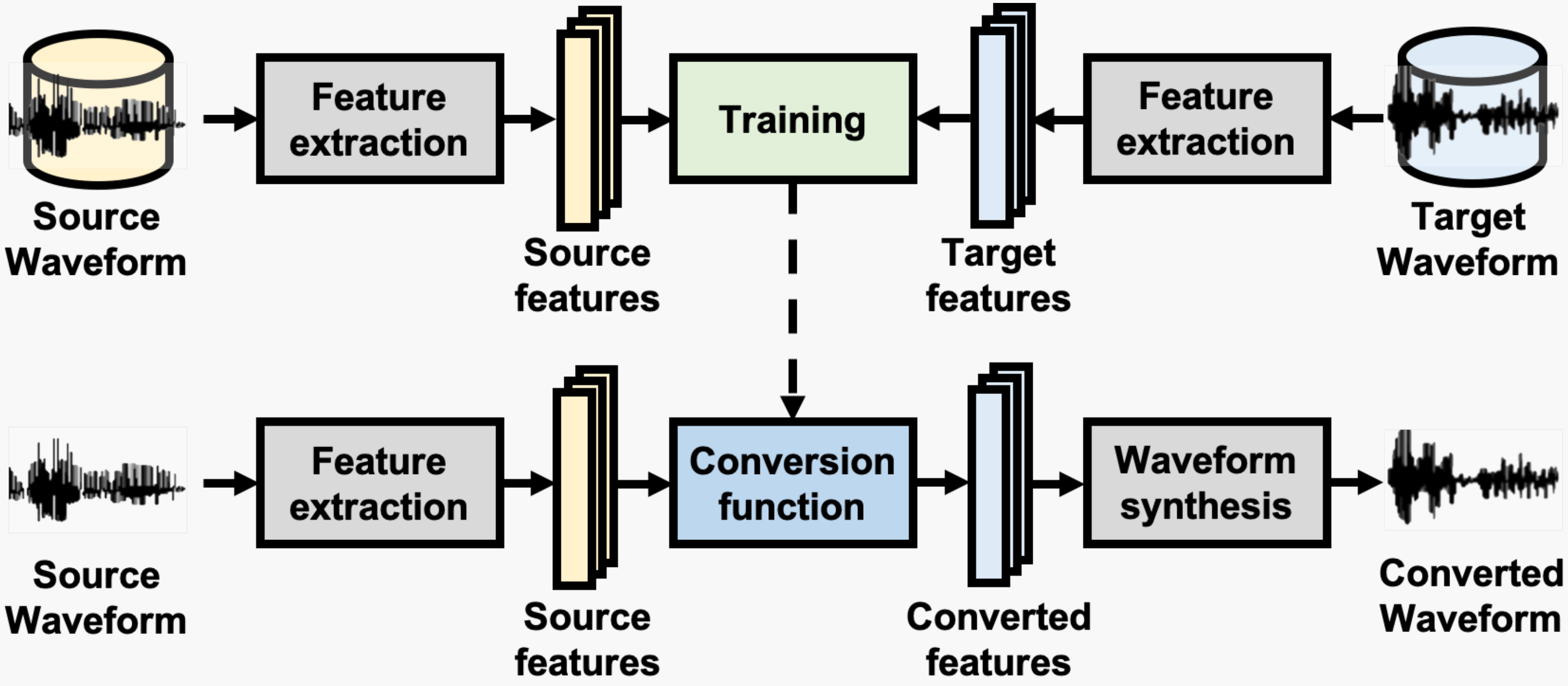
Note: the model does not necessary know the "words" in the utterance

Parallel VC: how people did VC research 30 years ago

[Abe+ ‘90] [Stylianou+ ’98]

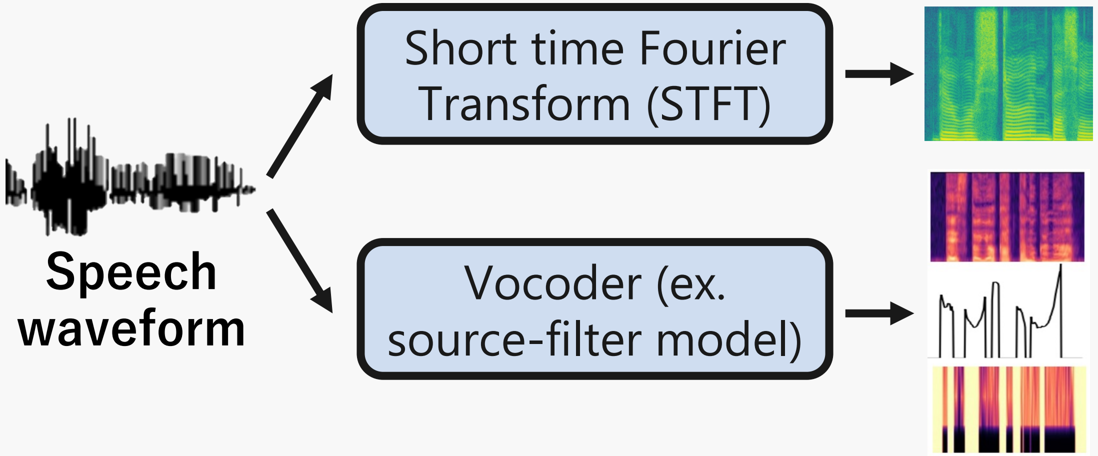
- **In practice, a typical VC system has three components**
  - Feature extraction, conversion, waveform synthesis (all are important!)

usually this is a machine learning model



Review: what is “feature”?

- **Speech is a kind of waveform = super dense 1-D data**
  - Telephone speech: 8000 Hz = 8000 samples per second
  - Singing voice (or music): 44100 Hz = 44100 samples per second!
  - Super complicated, too difficult for machine learning models!
- **It is easier to model in the frequency domain**



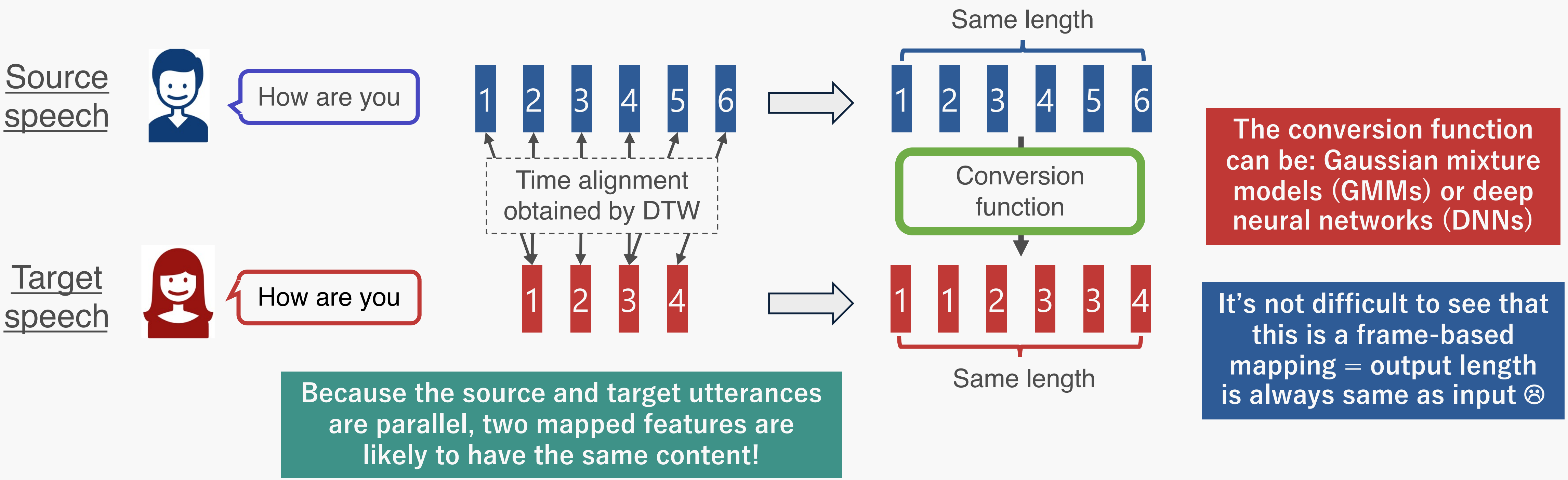
Time-frequency representation
- **Spectrogram**
- **Mel spectrogram (= spectrogram → mel filter bank)**

Vocoder parameters
- **Spectral envelope**
- **Fundamental frequency (f_0)**
- **Aperiodicity signal**

Parallel VC: how people did VC research 30 years ago

[Abe+ '90] [Stylianou+ '98]

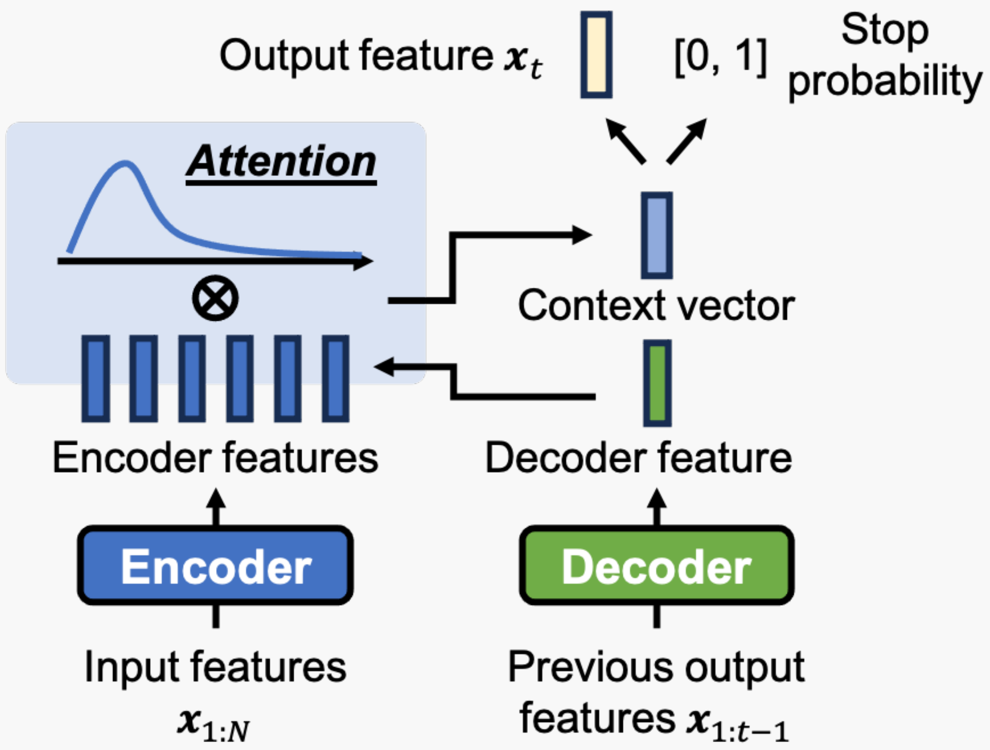
- **An important question: time alignment**
  - Source and target utterances can be of different lengths.
  - To calculate loss, we need to align the features.



State-of-the-art parallel VC : sequence-to-sequence modeling

[Tanaka+ '19] [Huang+ '20]

- ☺ **can model duration = can model prosody**
  - Accent, speaking rate, etc. are important factors to conversion similarity



source

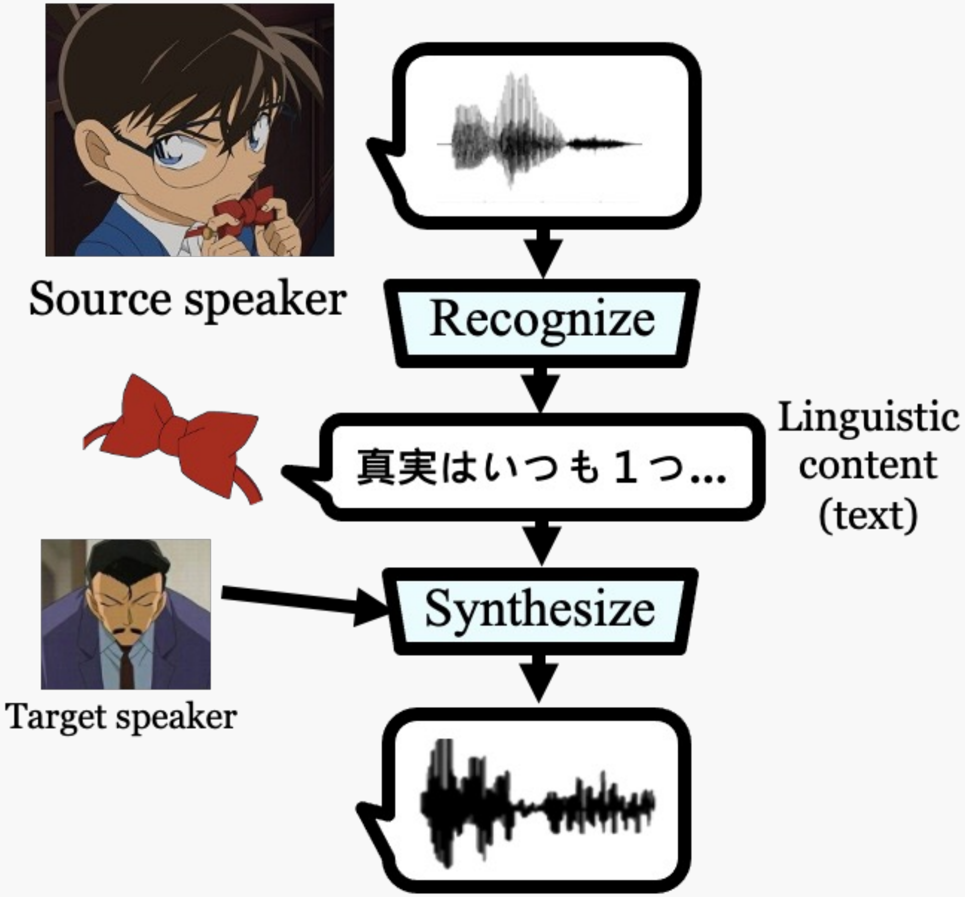
target

converted

Biggest advantage: parallel corpus is difficult to collect

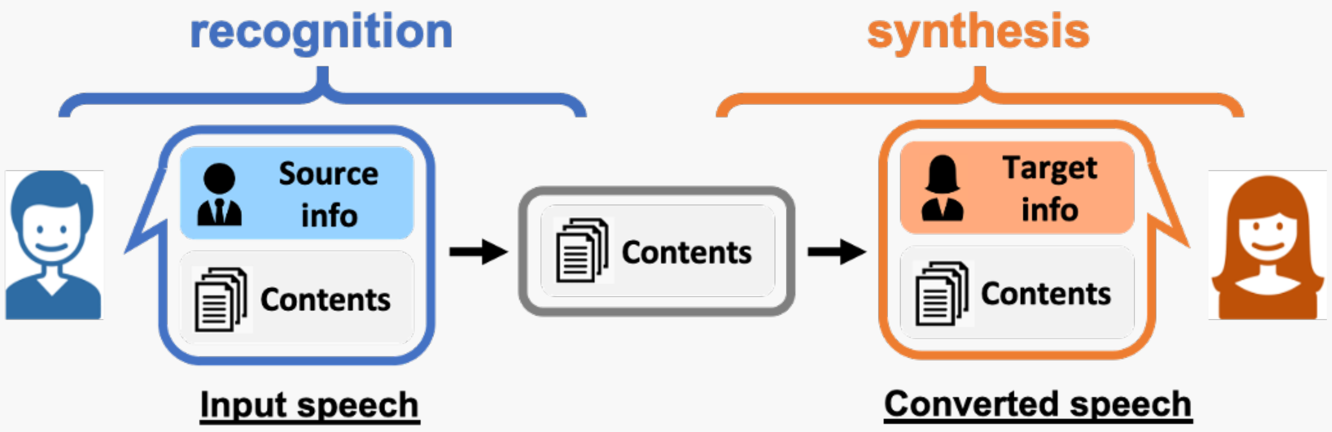
- **Think about it: How much data are <u>you</u> willing to record?**
  - Experimentally, (I think) we need at least one hour of parallel data
  - In practice, 5 minutes is considered "okay"
  - But many people will not be willing to record even 5 minutes of data!

Let me ask again: what would you do if YOU were asked to perform VC?



This is called "recognition-synthesis" VC

- Recognition = (1) extract "desired" information
  (2) eliminate "unwanted" source information
  - Ex., in speaker conversion, extract contents, eliminate speaker info
- Synthesis = Inject condition (target) information

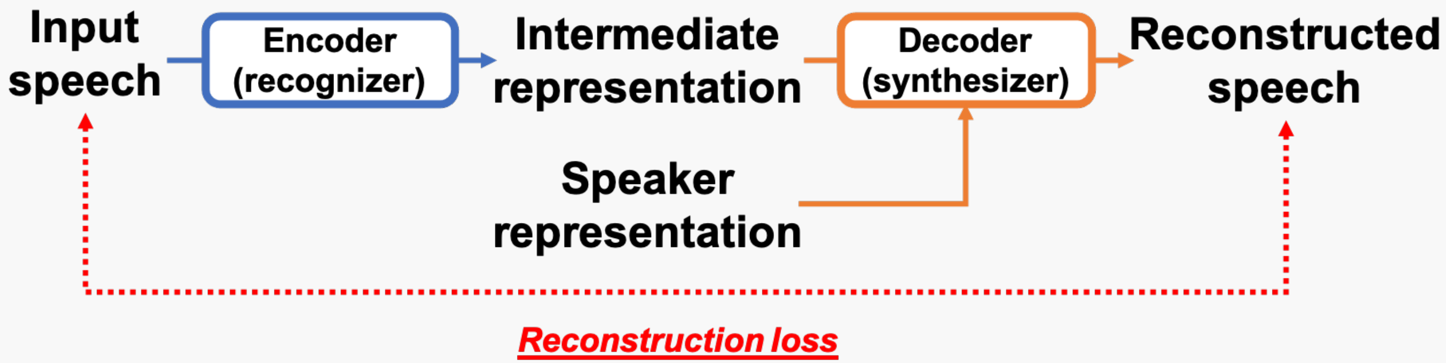


Two ways to categorize recognition-synthesis VC

- First category: joint training of the recognizer & synthesizer
  - Also called "auto-encoder" based VC
  - During training, the model tries to reconstruct the input speech.
    - Don't need parallel datasets anymore!

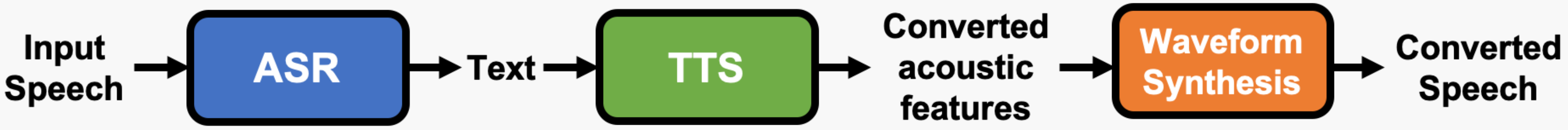
Parallel VC ↔ Nonparallel VC

  - The point is to design a good information bottleneck.
    - Ex., variational autoencoder (VAE), vector quantization (VQ), …



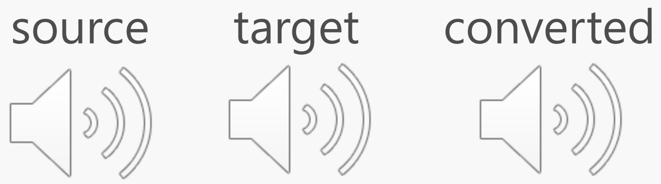
Two ways to categorize recognition-synthesis VC

- **Second category: separate training**
  - Train the recognizer to extract the desired information
  - Train the synthesizer to generate desired speech with condition
  - Similar to auto-encoder based VC -- don't need parallel datasets!
- **Straightforward example: cascade ASR+TTS** [Huang+ '20]
  - Directly use pre-trained automatic speech recognition and text-to-speech models



  - Problem: error propagation

Ground truth: In reality the European Parliament is practising delay tactics.
Recognized: In reality the European parliament is practising dialectics

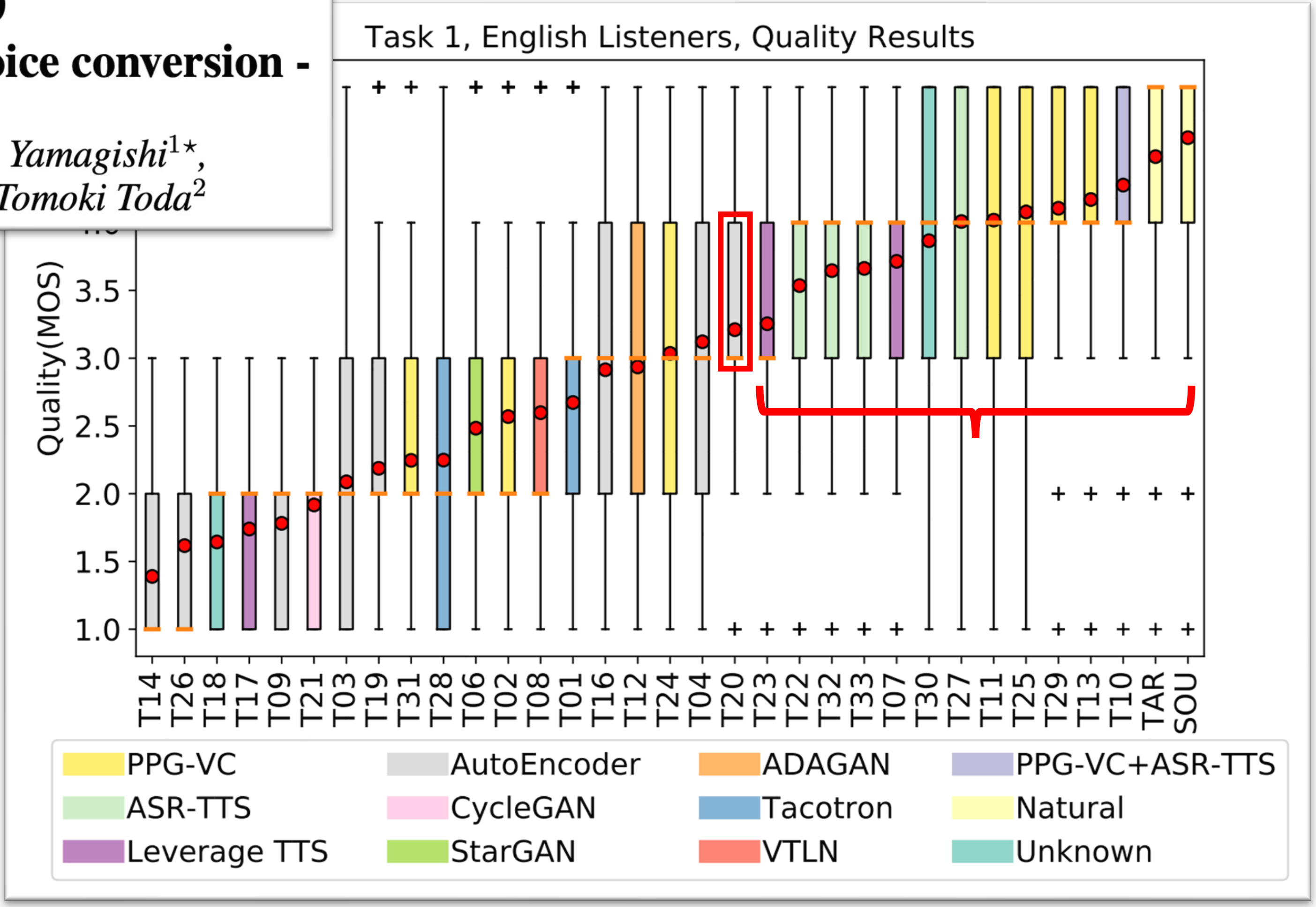


Auto-encoder based VC was outperformed [Zhao+ '20]

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]

- The best auto-encoder ranked 13th in 2020 ☹
- Separate training has thus become the mainstream



Content representation is important

- **Why does error propagation happen?**
  Ans: the recognizer throws away too much information

	Speech waveform	Spectrogram	SSL features	PPG (ASR encoder outputs)	Text
Resolution	16000 Hz	160~320 Hz	160~320 Hz	160~320 Hz	1~2 Hz (1~2 words per second)
Speaker information	Complete	Complete	Much ~ few	Almost none	Almost none

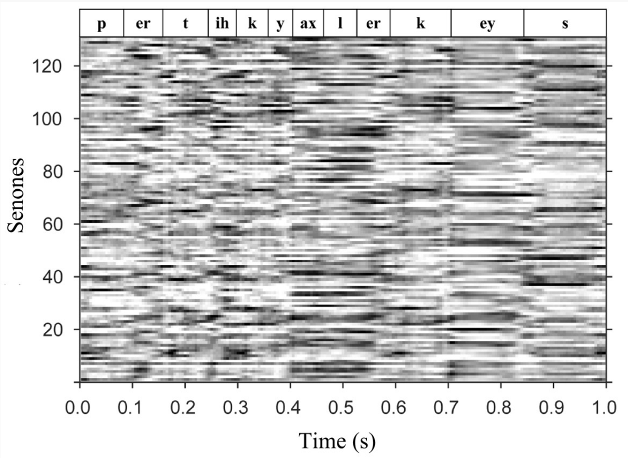
Can we find some features in the middle?

Phonetic posteriorgram (PPG)/ ASR encoder outputs

[Sun+ '16] [Liu+ '21]

- **What is PPG?**
  - A time-class matrix of the posterior probabilities of each phonetic class for each specific time frame.

  → A frame-based pure content representation
  - PPG is a natural by-product of traditional ASR (= difficult to collect nowadays)



- **Alternative: ASR encoder outputs**
  - Also contains pure content information
  - Ex. Whisper (strong ASR from OpenAI)

[Radford+ '23]

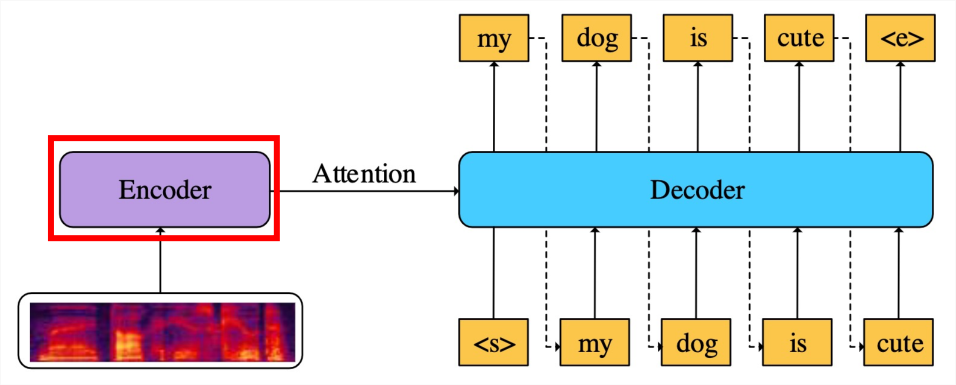


Figure from [Bai+ '21]

Self-supervised learning (SSL) features

- **What is SSL?**

  In terms of speech: no text, no speaker …

  - Learning useful features without label using some well-designed loss
  - With "SSL", usually we think of a two-stage framework:

    Self-supervised pre-training → **Supervised fine-tuning**

  - "Useful" = better than raw features (waveform, spectrogram, …)

- **Famous SSL models for speech**
  - Contrastive learning based: wav2vec 2.0 [Baevski+ '21]
  - Masked language modeling based: HuBERT, WavLM [Hsu+ '21][Chen+ '22]
  - Due to the natural of the pre-training loss, these features contain **rich content information** → suitable for recognition-synthesis VC!

S3PRL-VC

[Huang+ '22]

- **Compare how different SSL features perform in VC**

System	MCD	WER	ASV	Nat.	Sim.
Intra-lingual A2O					
mel	8.47	38.3	77.25	2.61 ± .11	35% ± 3%
PPG (TIMIT)	7.18	33.6	99.75	3.32 ± .10	58% ± 4%
PASE+	8.66	30.6	63.20	2.58 ± .12	31% ± 3%
APC	8.05	27.2	87.25	2.92 ± .11	43% ± 4%
VQ-APC	7.84	22.4	94.25	3.08 ± .10	40% ± 4%
NPC	7.86	30.4	94.75	2.98 ± .11	46% ± 3%
Mockingjay	8.29	35.1	79.75	2.81 ± .12	42% ± 4%
TERA	8.21	25.1	83.75	2.91 ± .12	37% ± 4%
Modified CPC	8.41	26.2	71.00	2.74 ± .11	33% ± 3%
DeCoAR 2.0	7.83	17.1	90.75	3.04 ± .11	43% ± 4%
wav2vec	7.45	10.1	98.25	3.40 ± .05	52% ± 2%
vq-wav2vec	7.08	13.4	100.00	3.59 ± .10	59% ± 4%
wav2vec 2.0 B.	7.50	10.5	98.00	3.36 ± .06	51% ± 2%
wav2vec 2.0 L.	7.63	15.8	97.25	3.26 ± .10	50% ± 4%
HuBERT B.	7.47	8.0	98.50	3.48 ± .10	55% ± 4%
HuBERT L.	7.22	9.0	99.25	3.47 ± .10	54% ± 4%

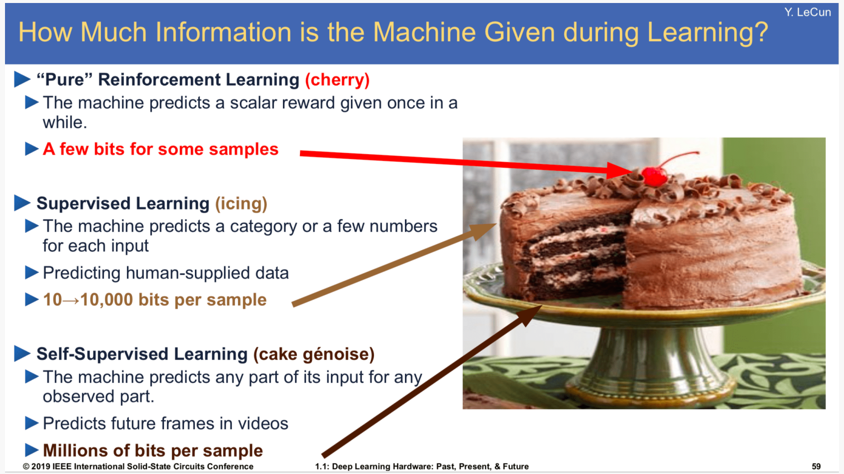
A comparison of content representations

	Speech waveform	Spectrogram	SSL features	PPG (ASR encoder outputs)	Text
Resolution	16000 Hz	160~320 Hz	160~320 Hz	160~320 Hz	1~2 Hz (1~2 words per second)
Speaker information	Complete	Complete	Much ~ few	Almost none	Almost none

System	MCD	WER	ASV	Nat.	Sim.	
vq-wav2vec	7.08	13.4	100.00	$3.59 \pm .10$	$59\% \pm 4\%$	SSL feature
USTC-2018† [31]	–	6.5	99.00	$4.20 \pm .08$	$55\% \pm 4\%$	PPG
USTC-2020 [23]	6.98	5.4	100.00	$4.41 \pm .07$	$\boxed{82\% \pm 3\%}$	text
SRCB [25]	8.90	11.5	92.00	$4.16 \pm .08$	$68\% \pm 3\%$	PPG
CASIA [26]	7.13	11.0	98.25	$4.25 \pm .08$	$61\% \pm 4\%$	PPG
ASR+TTS [22]	6.48	8.2	100.00	$3.84 \pm .09$	$\boxed{75\% \pm 3\%}$	text

Opinion: how people approach AI problems today

- **Use as much data as possible**
  - Data is usually unlabeled
  - Unsupervised/self-supervised learning
  - In VC: from parallel VC → nonparallel VC



Slides by Yann LeCun in NIPS 2016

- **Use "human knowledge" to make models learning easier**
  - In VC: spectrogram → text, PPG, SSL features…

Outline

- What is voice conversion (VC)? Why do we need it?
- How do we build a VC system?
- What is difficult in VC?

What are some unsolved problems in VC?

- Improve the quality of the converted voice
- Flexible learning
- New applications
- Evaluation

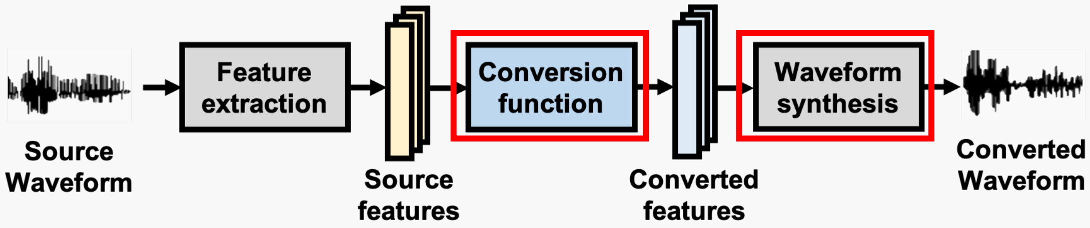
We will only be “touching the surface” of these topics.

What are some unsolved problems in VC?

- **Improve the quality of the converted voice**
- Flexible learning
- New applications
- Evaluation

Improve the quality of the converted voice

- **VC is a subfield of speech generative modeling**
- **Generative modeling (or generative AI, 生成式AI) = learn (approximate) the distribution of the data**
  - The better the model captures the distribution, the better the quality
- **(As mentioned before) speech is super difficult to model**
  - We always need better modeling techniques!
  - These techniques are used in these two parts:



Popular technique 1: autoregressive modeling

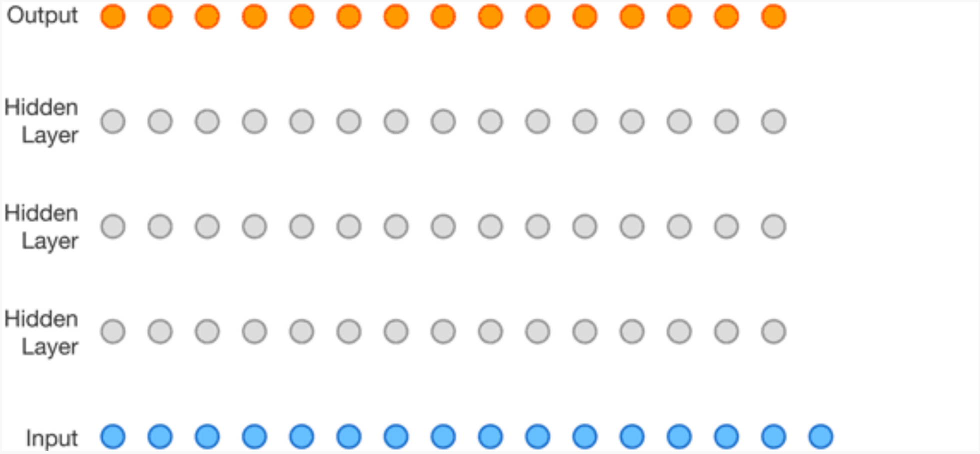
- Autoregressive modeling is accurate because it is the exact likelihood, which comes from chain rule:

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_n|x_1, \ldots, x_{n-1})$$

- Example: WaveNet, GPT-4o

[van den Oord+ '16]

Disadvantage: slow



Popular technique 2: generative adversarial net (GAN)

[Goodfellow+ '14]

- **Discriminator: try to distinguish real and fake**
  Generator: try to fool the discriminator
  - In fact there is a very rigorous mathematical formulation behind GAN…
  - I recommend reading this tutorial: https://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2018/Lecture/GANtheory%20(v2).pdf
- **GAN used to be famous for "hard to train"**
  - Difficult to set the hyper-parameters
  - But now it has become very simple thanks to researchers' efforts!
- **Famous "GAN in VC" papers:** [Hsu+ '17] [Chou+ '18] [Kaneko+ '18]

Disadvantage: the formulation of GAN is an "approximation"

Popular technique 3: normalizing flow

[Dinh+ '14]

- **Transforms a simple distribution into a complex one by applying a sequence of invertible transformation functions.**
  - There is also a rigorous math formulation…
  - Recommended tutorial: https://blog.evjang.com/2018/01/nf1.html

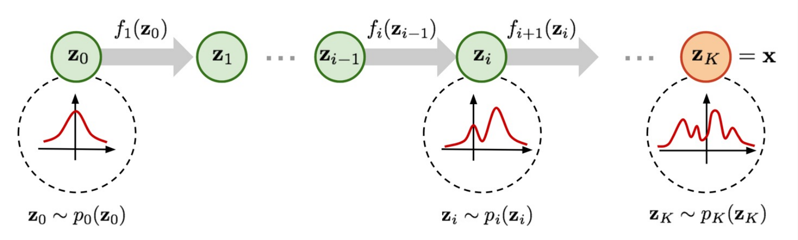


Figure from https://lilianweng.github.io/posts/2018-10-13-flow-models/

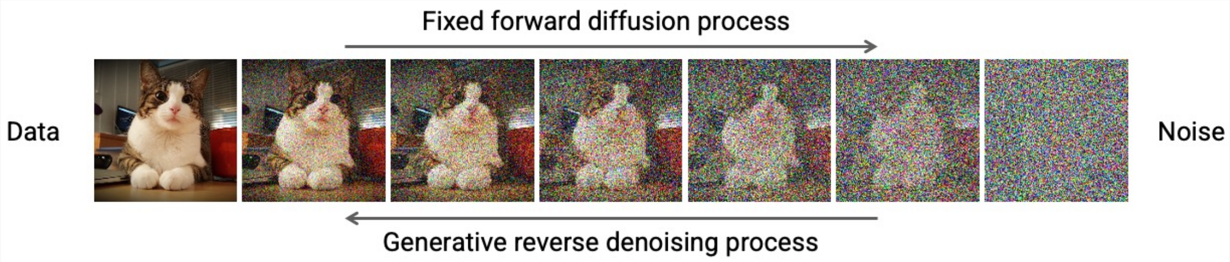
- **Advantage: exact likelihood, and it's faster than auto-regressive modeling**
- **Famous "flow in speech synthesis" papers:** [Pregnar+ '19][Kong+ '21]

Disadvantage: requires a large model, difficult to train

Popular technique 4: diffusion modeling

[Ho+ ‘20][Yang+ ‘21]

- Transform from noise to data (similar to flow!)
  - More math… even thermodynamics…



https://deeplearning.cs.cmu.edu/F23/document/slides/lec23.diffusion.updated.pdf

- Advantages: better quality than GAN, easier to train than flow
- Famous “diffusion in speech synthesis” papers: [Liu+ ‘22] [Ju+ ‘24]

Disadvantage: the number of diffusion steps makes generation slow

What are some unsolved problems in VC?

- Improve the quality of the converted voice
- Flexible learning
- New applications
- Evaluation

Categorize VC based on "what speaker can be handled"

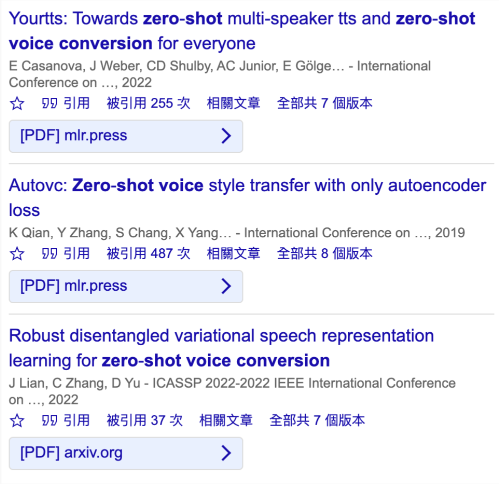
- **One-to-one VC**
  - Can convert <u>one training</u> source speaker to one training target speaker
  - Traditional parallel VC falls in this category
- **Many-to-one VC**
  - Can convert <u>multiple training</u> source speakers to one training target speaker

Can you guess what one-to-many VC means?

- **Any-to-one VC**
  - Can convert <u>any unseen</u> source speaker to one training target speaker

This categorization is based on the data used in model learning

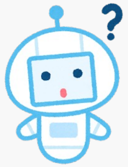
Ultimate goal: any-to-any VC


Yourtts: Towards **zero-shot** multi-speaker tts and **zero-shot voice conversion** for everyone
E Casanova, J Weber, CD Shulby, AC Junior, E Gölge… - International Conference on …, 2022
☆ 引用 被引用 255 次 相關文章 全部共 7 個版本
[PDF] mlr.press

Autovc: **Zero-shot voice** style transfer with only autoencoder loss
K Qian, Y Zhang, S Chang, X Yang… - International Conference on …, 2019
☆ 引用 被引用 487 次 相關文章 全部共 8 個版本
[PDF] mlr.press

Robust disentangled variational speech representation learning for **zero-shot voice conversion**
J Lian, C Zhang, D Yu - ICASSP 2022-2022 IEEE International Conference on …, 2022
☆ 引用 被引用 37 次 相關文章 全部共 7 個版本
[PDF] arxiv.org

- **Also known as one-shot VC**
  - Can convert to any unseen speaker with only one utterance
  - Ex. Conan wants to copy Hattori's voice now
  - But it is difficult to collect a lot of his data…



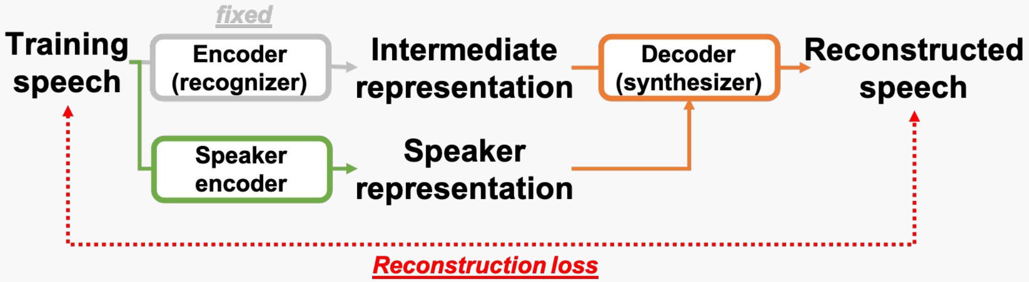
たったひとつの真実見抜く

There are many papers use the term "zero-shot VC". Is the term "zero-shot" proper? Why?

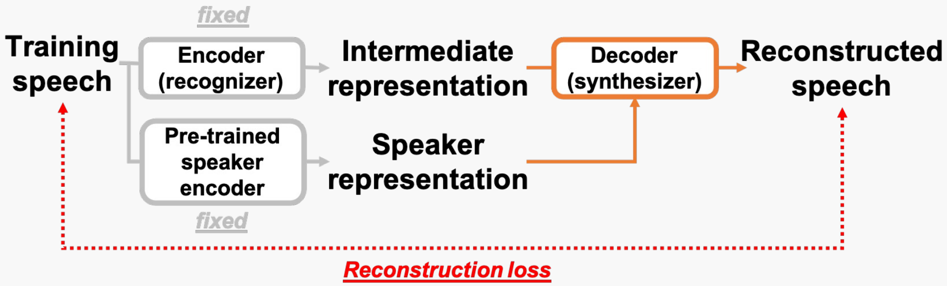
- **Any-to-any VC is easy for human, but not for machine!**
  - By listening to one speech clip, human can easily imagine how this person talks!
  - Human is smart, but machine is not…

Two common practices for speaker representation in recognition-synthesis VC

1. **Train a speaker encoder jointly with the synthesizer**
   - Advantage: better conversion quality [Chien+ '21]



2. **Pre-trained speaker encoder (ex., d-vector, x-vector)** [Variani+ '14] [Snyder+ '18]
   - Advantage: better generalization ability



This is still one of the most researched direction in VC

What are some unsolved problems in VC?

- Improve the quality of the converted voice
- Flexible learning
- New applications
- Evaluation

New application 1: VC in noisy environment

- **Sometimes we want to do VC with noise (or music, etc.)**
  - Back to the dubbing example…

https://yaoxunji.github.io/background_sound_vc/

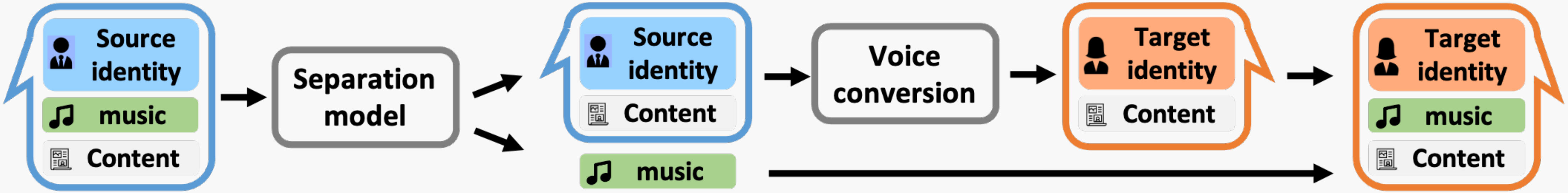
[Yao+ '23]

Original

Converted

- **Straightforward idea**

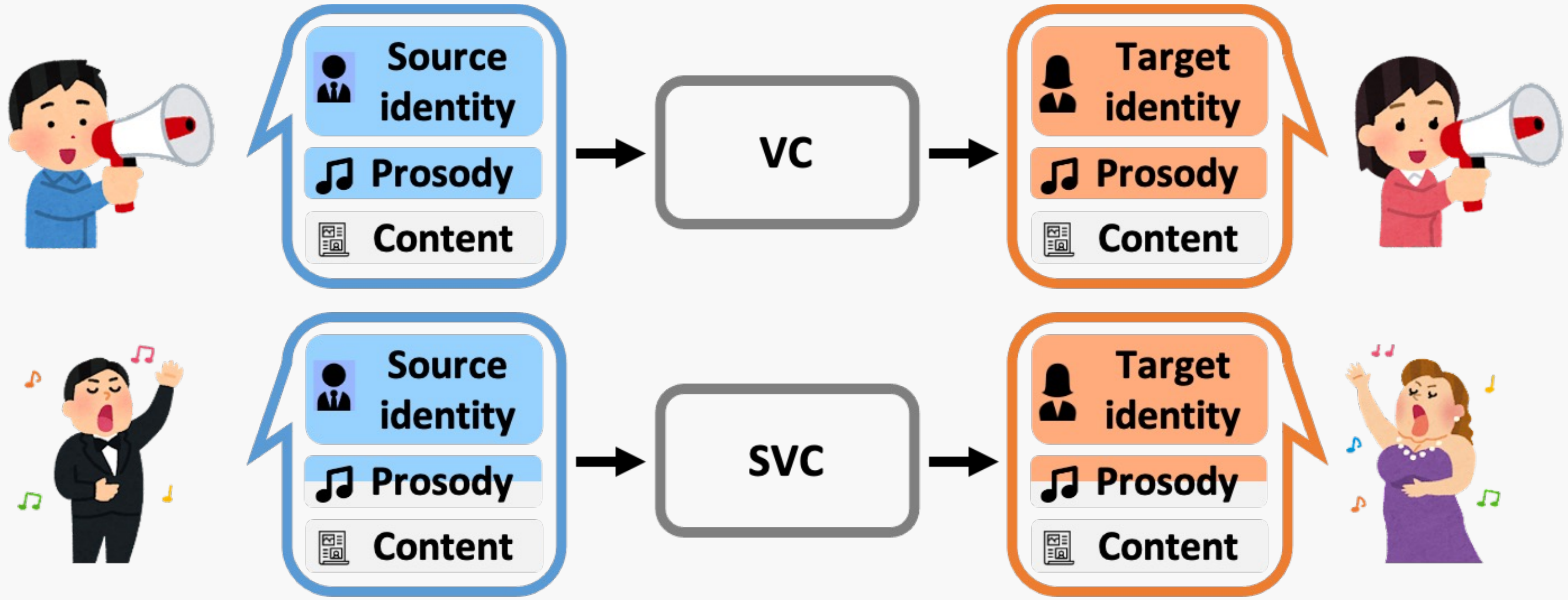
Difficulty: separation is hard!



New application 2: Singing VC

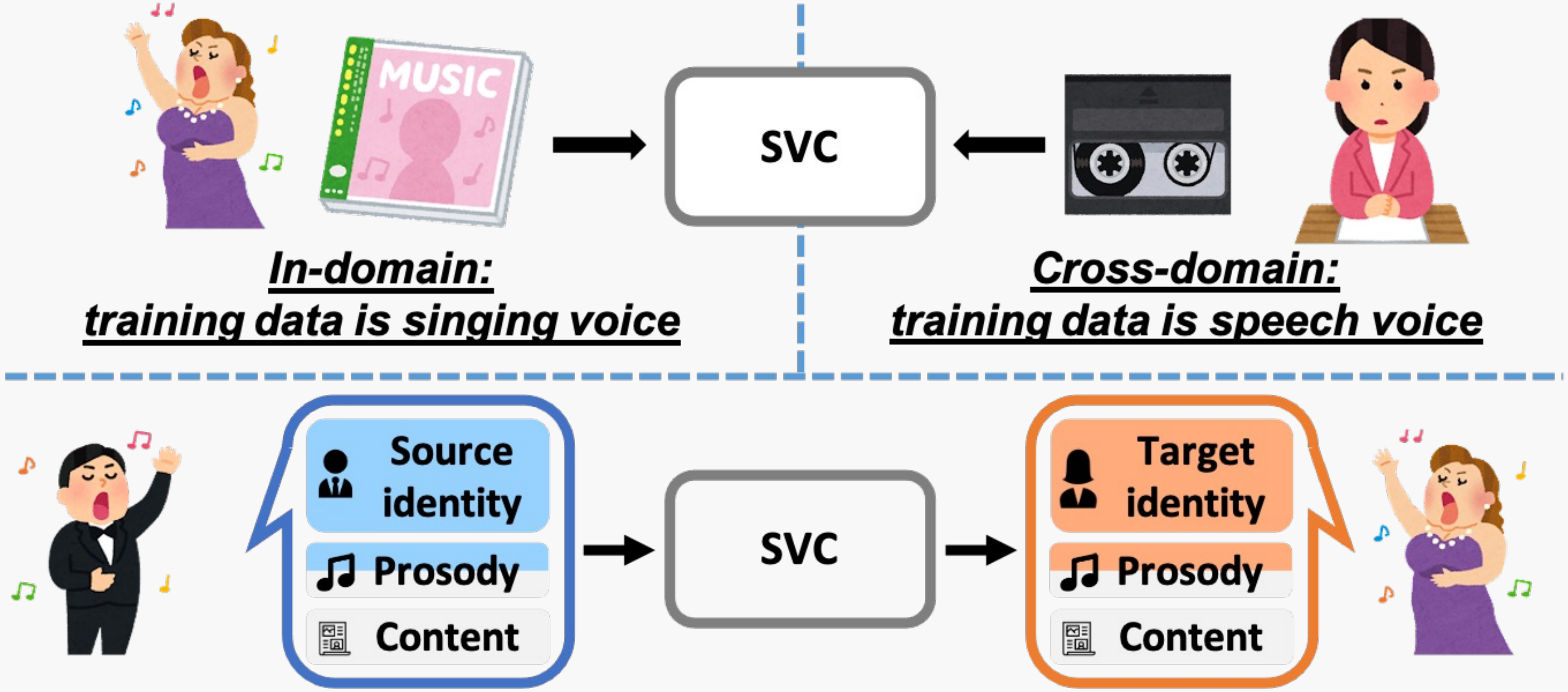
[Villavicencio+ '10]

- **What is the difference between normal VC and singing VC?**
  - In singing VC, we want to change "singing style" but with same notes
    = add vibrato, falsetto, ⋯, but sound like the same song



New application 2: Singing VC

- **What's more difficult: cross-domain singing VC** [Huang+ '23]

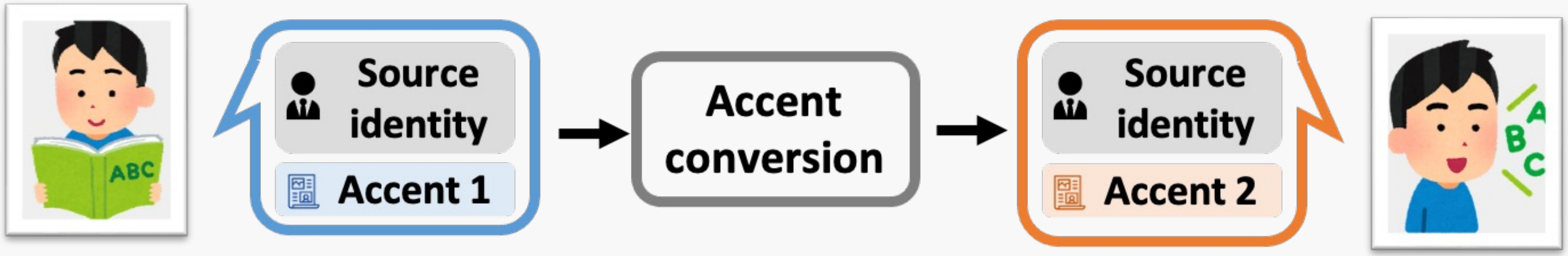


そもそも is this possible?

New application 3: accent conversion

1. **Convert from non-native speech → native speech**
   - Also known as foreign accent conversion (FAC) [Zhao+ '18] [Zhao+ '21]
2. **Convert from accent 1 → accent 2** [Ezzerg+ '23]
   - Ex., American accent → British accent; 九州弁 → 関西弁

- **Difficulty: usually we want to main the speaker identity!**



What are some unsolved problems in VC?

- Improve the quality of the converted voice
- Flexible learning
- New applications
- Evaluation

Evaluation in VC (and speech synthesis in general)

- **Subjective evaluation (human judgement) is considered the golden standard to evaluate machine-generated speech**
  - Because it is human who will be listening to these speech samples
- **Objective evaluation metrics are just for reference** [Wagner+ '19]
  - Ex., mean square error of generated speech and ground truth speech
  - They are considered to be not aligned well with human judgements.
    ⇨ A > B in objective metric does not always mean A > B by human.
- **Usually done through a mean opinion score (MOS) test**
  - Listen to sample ⇨ choose a rating (usually 1-5)

Aspects to evaluate in VC

- Naturalness := how natural the generated sample sounds
  - There is a tendency to use "naturalness" over "quality", which is vague.
  - Naturalness is a "basic aspect" shared by all speech synthesis tasks
- In speaker conversion, conversion similarity is also important
  - How to evaluate? a common approach:

1. Let the listener listen to (1) converted sample (2) reference sample
   - (1) and (2) can be of the same contents (some people choose different contents)
2. Ask the listener "do you think these two samples are spoken by the same person?"
   - The listener does not know which one is the reference

If (1) and (2) are of the same contents, does it make it easier to think they come from the same speaker?

Different VC application require different evaluation aspects

- Can you tell these two samples are from the same person?

The same singer can sound very differently in different parts of the same song

- Can you rate which sample has a heavier accent?

Even native speakers can have a hard time giving a judgement

- Can you rate which sample is closer to the reference?

Even in speaker conversion, It can be just… difficult

Outline

- What is voice conversion (VC)? Why do we need it?
- How do we build a VC system?
- What is difficult in VC?

Concluding remarks

- **VC has many applications → it is considered an important and fundamental technique**
- **Many people think VC is solved… is it?**
  - Most papers use a very ideal experimental setting!
- **VC is useful only when used to improve people's lives**
  - Same for any AI technique!

Useful materials

- Advanced Voice Conversion
  - https://www.slideshare.net/slideshow/advanced-voice-conversion/107923321#2
- An Overview of Voice Conversion and its Challenges: From Statistical Modeling to Deep Learning
  - https://arxiv.org/abs/2008.03648

References (in chronological order)

[Abe+ ’90] Abe, Masanobu, et al. "Voice conversion through vector quantization." Journal of the Acoustical Society of Japan (E) 11.2 (1990): 71-76.
[Stylianou+ ‘98] Stylianou, Yannis, Olivier Cappé, and Eric Moulines. "Continuous probabilistic transform for voice conversion." IEEE Transactions on speech and audio processing 6.2 (1998): 131-142.
[Villavicencio+ ‘10] Villavicencio, Fernando, and Jordi Bonada. "Applying voice conversion to concatenative singing-voice synthesis." Interspeech. 2010.
[Toda ‘14] T. Toda, "Augmented speech production based on real-time statistical voice conversion," 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, USA, 2014, pp. 592-596
[Goodfellow+ ‘14] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
[Dinh+ ‘14] Dinh, Laurent, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation." arXiv preprint arXiv:1410.8516 (2014).
[Variani+ ‘14] Variani, Ehsan, et al. "Deep neural networks for small footprint text-dependent speaker verification." Proc. ICASSP 2014.
[Sun+ ‘16] L. Sun, K. Li, H. Wang, S. Kang and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," Proc. ICME, 2016, pp. 1-6.
[van den Oord+ ‘16] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).
[Hsu+ ‘17] Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M. (2017) Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks. Proc. Interspeech 2017, 3364-3368
[Chou+ ‘18] Chou, J.-c., Yeh, C.-c., Lee, H.-y., Lee, L.-s. (2018) Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations. Proc. Interspeech 2018, 501-505
[Kaneko+ ‘18] Kaneko, Takuhiro, and Hirokazu Kameoka. "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks." Proc. EUSIPCO, 2018.
[Snyder+ ‘18] Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." Proc. ICASSP 2018
[Zhao+ ‘18] Zhao, Guanlong, et al. "Accent conversion using phonetic posteriorgrams." Proc. ICASSP, 2018

References (in chronological order)

[Wagner+ '19] Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., Voße, J. (2019) Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10), 105-110

[Prenger+ '19] R. Prenger, R. Valle and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," Proc. ICASSP 2019, pp. 3617-3621

[Tanaka+ '19] K. Tanaka, H. Kameoka, T. Kaneko and N. Hojo, "ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," Proc. ICASSP, 2019, pp. 6805-6809

[Huang+ '20] W.-C., Hayashi, T., Wu, Y.-C., 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

[Zhao+ '20] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R.K. Das, T. Kinnunen, Z. Ling, T. Toda, "Voice Conversion Challenge 2020 -- intra-lingual semi-parallel and cross-lingual voice conversion --" Proc. Joint workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, pp. 80-98, Oct. 2020.

[Ho+ '20] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

[Baevski+ '21] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[Hsu+ '21] W. -N. Hsu, Y. -H. H. Tsai, B. Bolte, R. Salakhutdinov and A. Mohamed, "Hubert: How Much Can a Bad Teacher Benefit ASR Pre-Training?," Proc. ICASSP, 2021, pp. 6533-6537.

[Liu+ '21] Liu, Songxiang, et al. "Any-to-many voice conversion with location-relative sequence-to-sequence modeling." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1717-1728.

[Bai+ '21] Bai, Ye, et al. "Integrating knowledge into end-to-end speech recognition from external text-only data." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1340-1351.

References (in chronological order)

[Kong+ '21] Jaehyeon Kim, Jungil Kong, Juhee Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," Proc. ICML, 2021, 5530-5540.
[Yang+ '21] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." Proc. ICLR; 2021
[Chien+ '21] C. -M. Chien, J. -H. Lin, C. -y. Huang, P. -c. Hsu and H. -y. Lee, "Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech," Proc. ICASSP 2021, pp. 8588-8592
[Zhao+ '21] G. Zhao, S. Ding and R. Gutierrez-Osuna, "Converting Foreign Accent Speech Without a Reference," in IEEE/ACM TASLP, vol. 29, pp. 2367-2381, 2021
[Huang+ '22] Huang, Wen-Chin, et al. "A comparative study of self-supervised speech representation based voice conversion." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1308-1318.
[Gan+ '22] Gan, Wendong, et al. "Iqdubbing: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion." arXiv preprint arXiv:2201.00269 (2022).
[Chen+ '22] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505-1518, Oct. 2022.
[Liu+ '22] Liu, Jinglin, et al. "Diffsinger: Singing voice synthesis via shallow diffusion mechanism." Proc. AAAI, Vol. 36. No. 10. 2022.
[Radford+ '23] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International conference on machine learning. PMLR, 2023.
[Yao+ '23] Yao, Jixun, et al. "Preserving background sound in noise-robust voice conversion via multi-task learning." Proc. ICASSP, 2023.
[Huang+ '23] Huang, Wen-Chin, et al. "The singing voice conversion challenge 2023." Proc. Automatic Speech Recognition and Understanding Workshop (ASRU). 2023.
[Ezzerg+ '23] Ezzerg, Abdelhamid, et al. "Remap, warp and attend: Non-parallel many-to-many accent conversion with normalizing flows." Proc. IEEE Spoken Language Technology Workshop (SLT). 2023.
[Ju+ '24] Ju, Zeqian, et al. "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models." arXiv preprint arXiv:2403.03100 (2024).