Voice Conversion

Melissa Chen

ECE 477 Guest Lecture

11/14/2023

UNIVERSITY of ROCHESTER

Audio Information Research Laboratory

Table of Contents

About Me

I'm a 3nd-year Ph.D. student of ECE @ airlab

Working with Prof. Duan Zhiyao on speech synthesis topics, i.e. text-to-speech synthesis and voice conversion.

Current paper:
"ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Speed," Interspeech, 2023. [paper][demo][code]

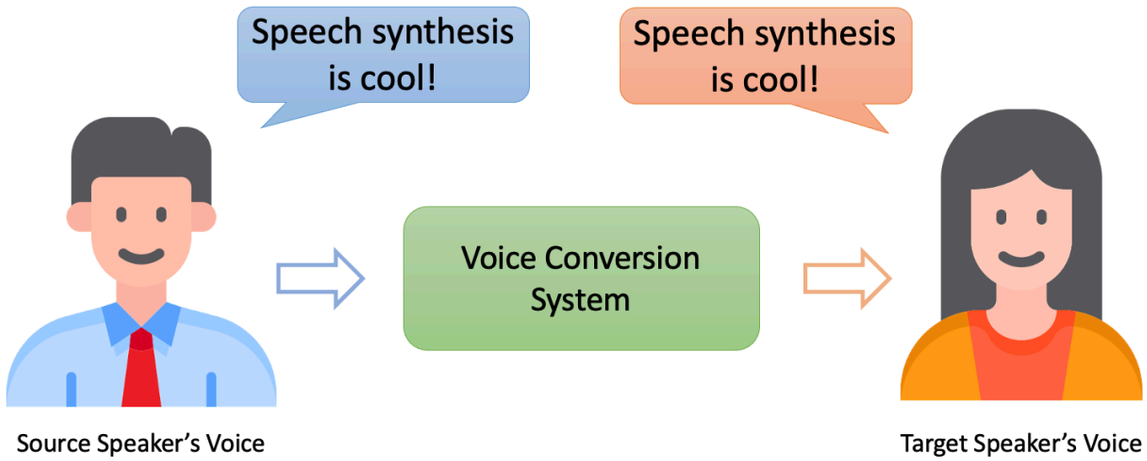
Internship Projects:
Meta Platforms., 06-09/2023: Self-supervised prosody learning for expressive TTS.
TikTok Inc., 05-present/2022: GAN-based speech enhancement and super-resolution.
Tencent America, 05-08/2021: Universal vocoder for TTS.
Kwai Inc., 05-12/2020: Singing voice synthesis; low-resources TTS.

Table of Contents

What is voice conversion?

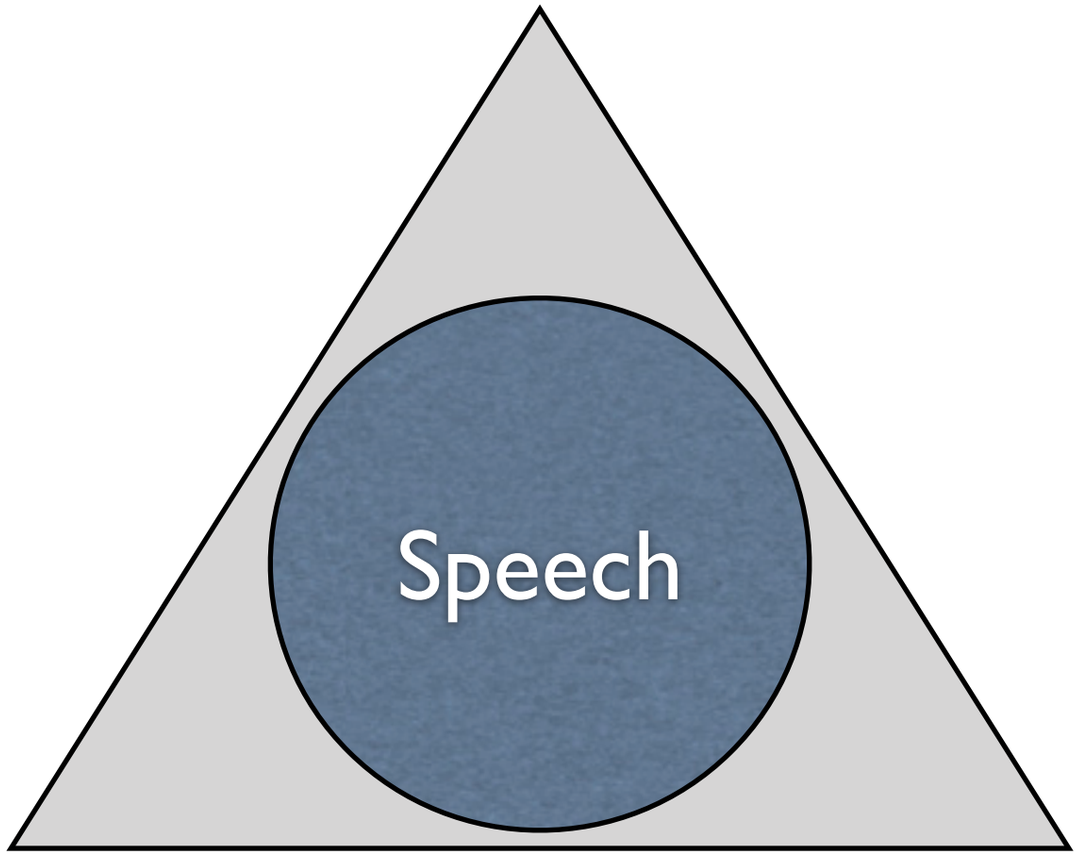
Voice Conversion is

the task of converting one's voice to sound like that of another's, while maintaining the content.



Components of Speech

Content (What)

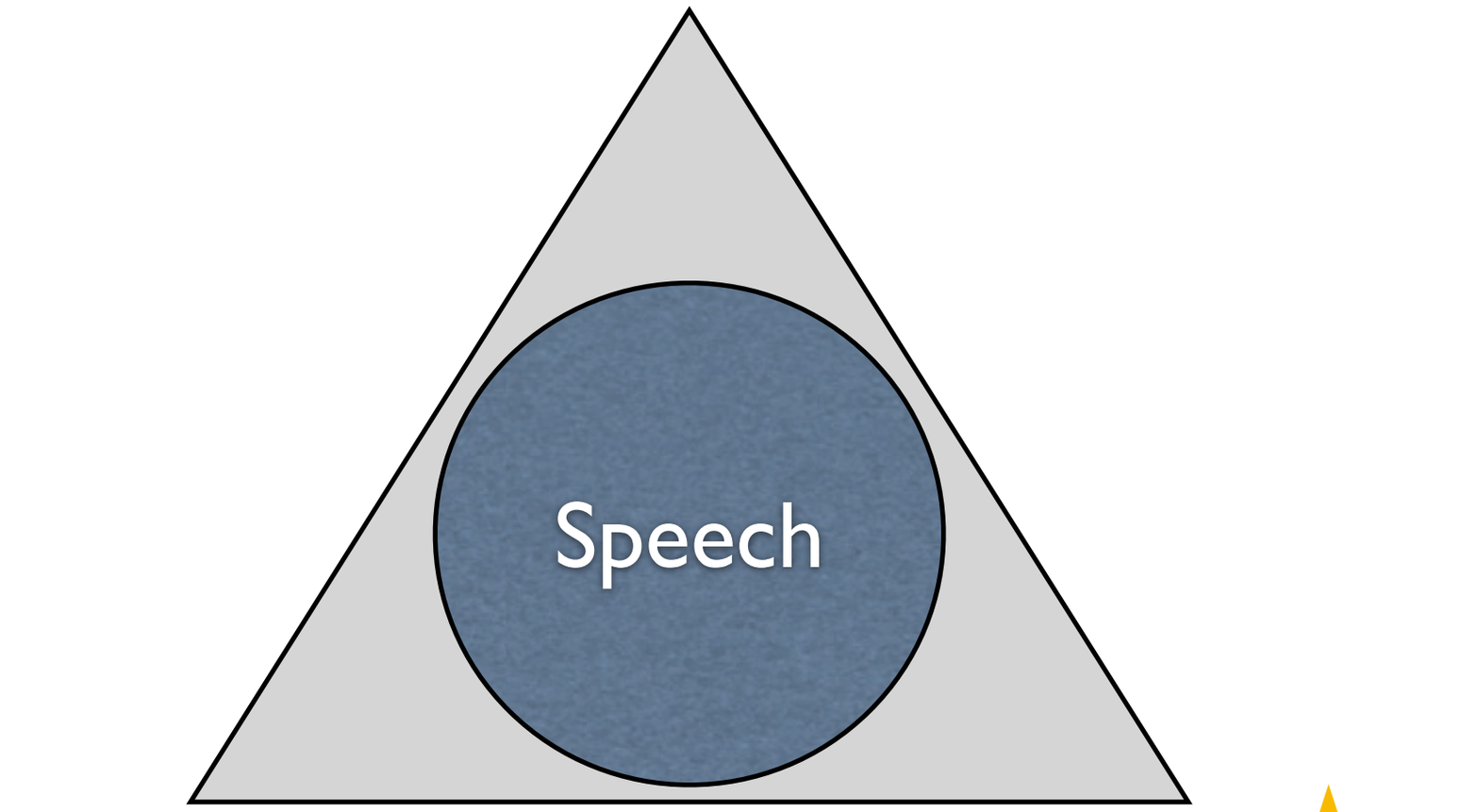


Prosody (How)

Timbre (Who)

Note that this is an approximate categorization. For example, prosody also tells a lot about who.

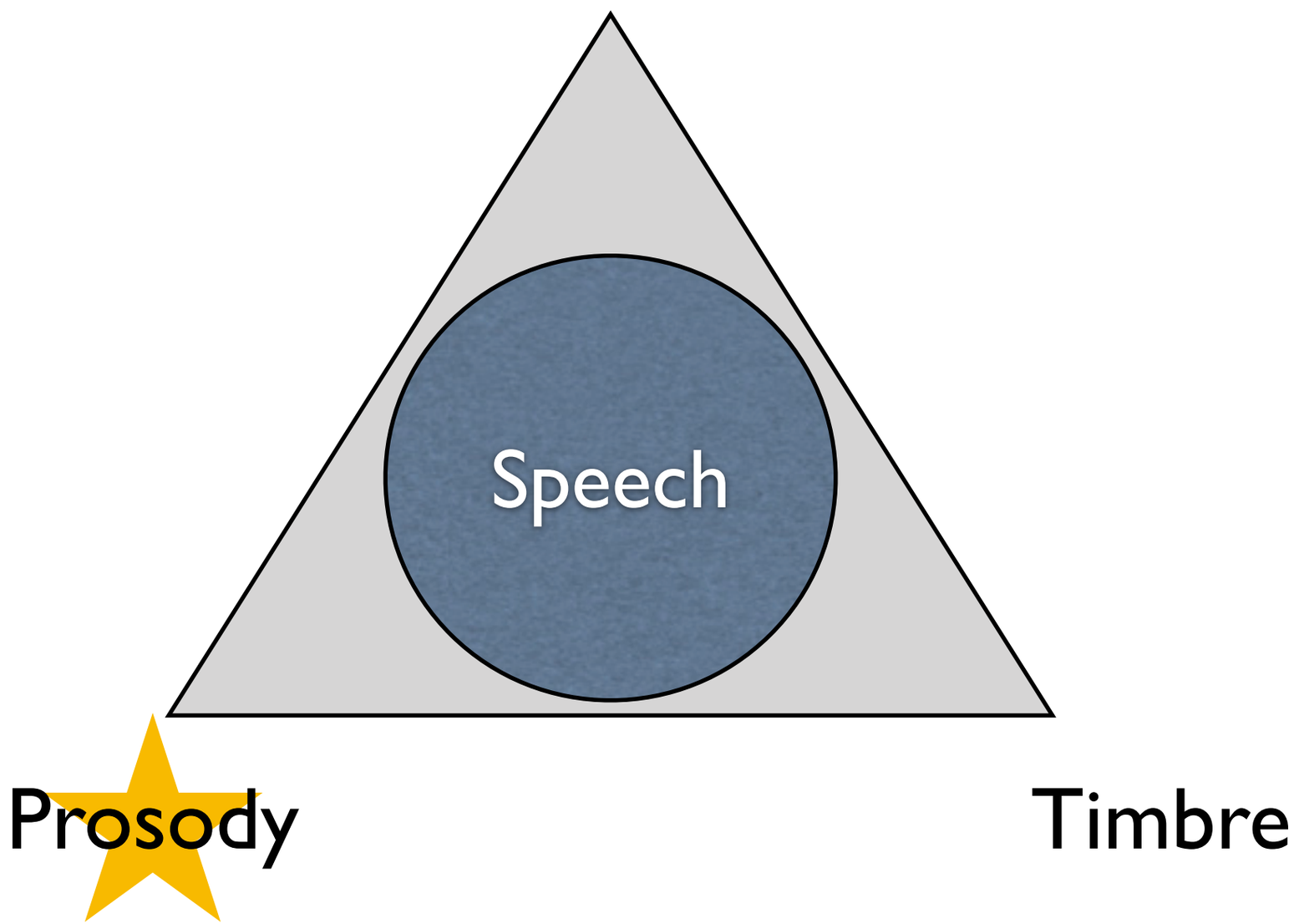
Components of Speech



What is voice conversion?

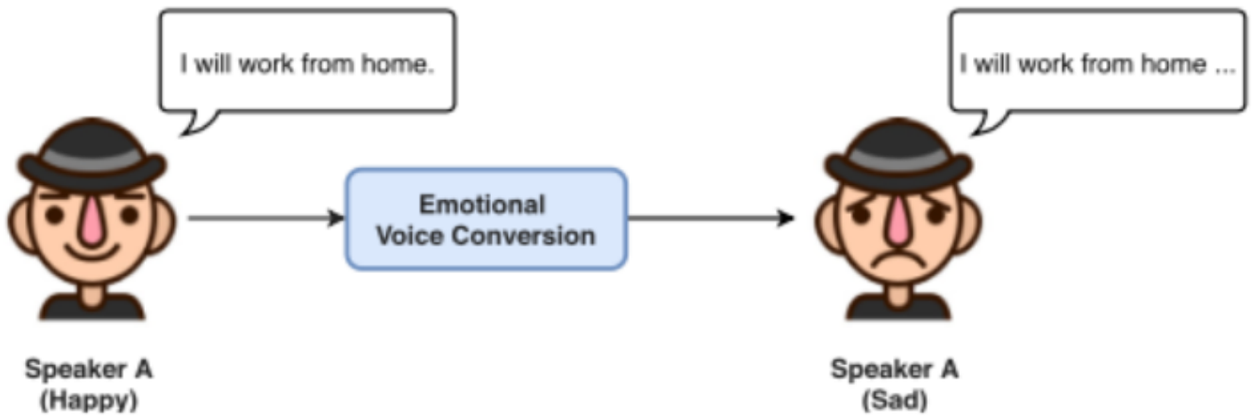
Leonardo DiCaprio

Conversion in Content



Conversion in Prosody (1)

Emotional voice conversion is
to convert an utterance from one emotion state to another,
while preserving the linguistic information and speaker identity.



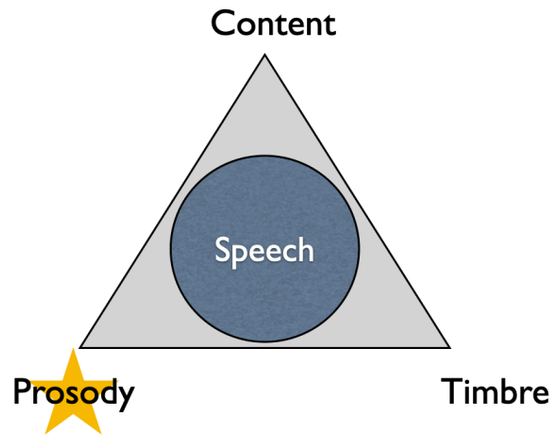


Figure from Sisman et al. "An overview of voice conversion and its challenges: From statistical modeling to deep learning.", TALSP 2020.

Conversion in Prosody (2)

Accent Conversion seeks to change the accent of speech from one to another while preserving the speech content and speaker identity.

Enhance understanding
Preserve speaker identity
Applications: Call center

Samples:
https://vcsamples.github.io/SPL2022AC/
https://tuannamnguyenkit.github.io/

Zhou, Yi, et al. "TTS-Guided Training for Accent Conversion Without Parallel Data." *IEEE Signal Processing Letters* (2023).
Nguyen, Tuan Nam, Ngoc-Quan Pham, and Alexander Waibel. "Accent Conversion using Pre-trained Model and Synthesized Data from Voice Conversion." *Proc. Interspeech*. Vol. 2022. 2022.

Conversion in Prosody (3)

Speech-to-singing voice conversion task aims to generate singing samples corresponding to speech recordings.

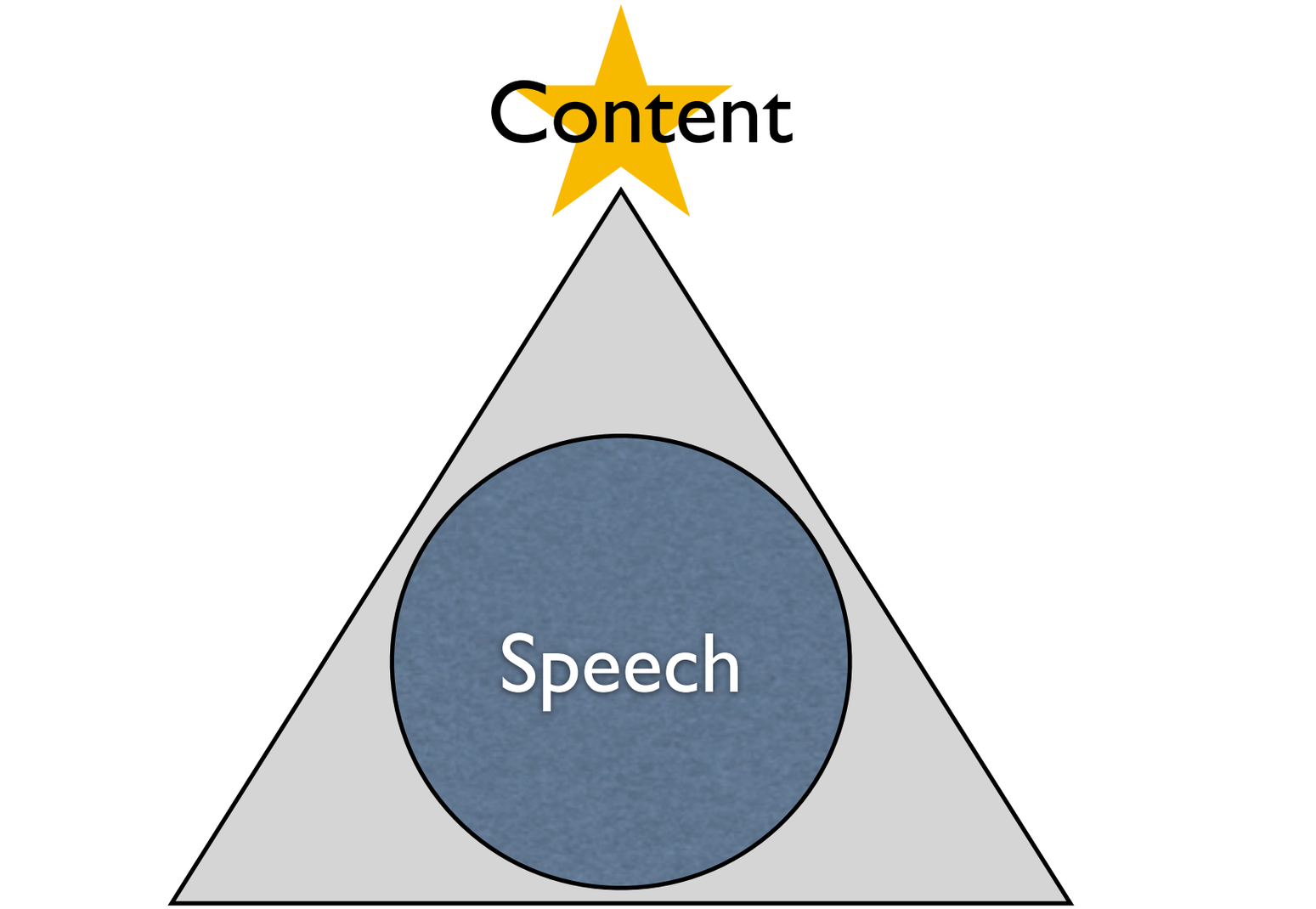
AlignSTS [1]: Real Speech Synthesized Singing Real Singing

Singing-to-singing

Singing Voice Conversion_2023 [2]
DiffSVC [3]: Sample https://liusongxiang.github.io/diffsvc/

[1] Li, Ruiqi, et al. "AlignSTS: Speech-to-Singing Conversion via Cross-Modal Alignment." *arXiv preprint arXiv:2305.04476* (2023).
[2] Huang, Wen-Chin, et al. "The Singing Voice Conversion Challenge 2023." *arXiv preprint arXiv:2306.14422* (2023).
[3] Liu, Songxiang, et al. "Diffsvc: A diffusion probabilistic model for singing voice conversion." *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* IEEE, 2021.

Conversion in Content



Conversion in Content (1)

Voice conversion can be used to improve speech quality:

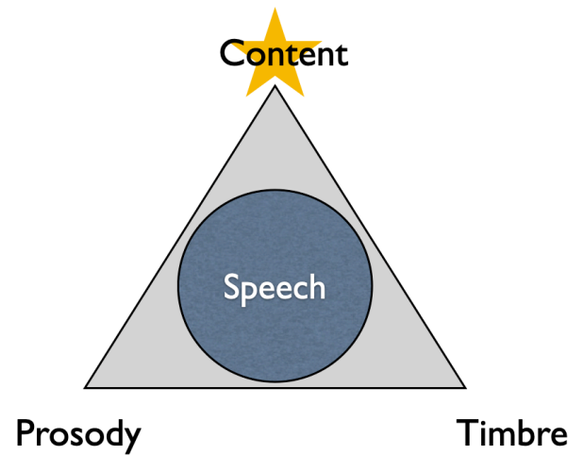
Help people who have difficulties vocalizing sound due to

- Injuries
- Surgeries
- Dysarthria
- …

Dysarthric speech reconstruction [Wang. et al, ICASSP 2020]

Original Converted

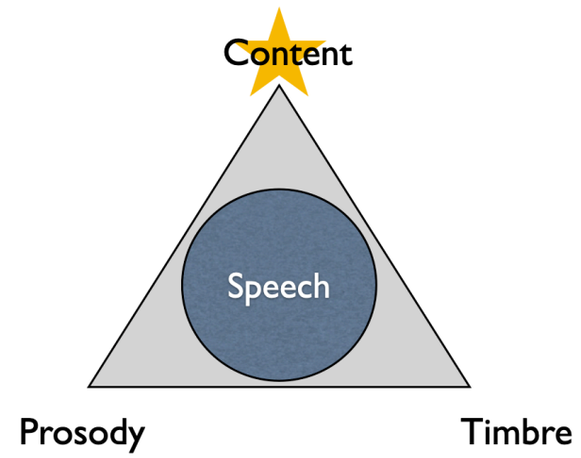
"Supreme"



Conversion in Content (2)

Cross-lingual voice conversion (XVC) transforms the speaker identity of a source speaker to that of a target speaker who speaks a different language.

β-VAE VC: Samples: https://beta-vaevc.github.io/



Zhou, Yi, et al. "Optimization of Cross-Lingual Voice Conversion With Linguistics Losses to Reduce Foreign Accents." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

Applications - Cross-lingual Dubbing

Compared to TTS

TTS (Text-to-speech Synthesis) and VC are closely related topics, sharing similar technical methodologies and applications.

VC has some advantages:

- Less data-hungry and less computationally demanding (known utterance structures)
- Flexible timbre change
- Expressive
- Natural prosody

Table of Contents

Acoustic Differences between Speakers

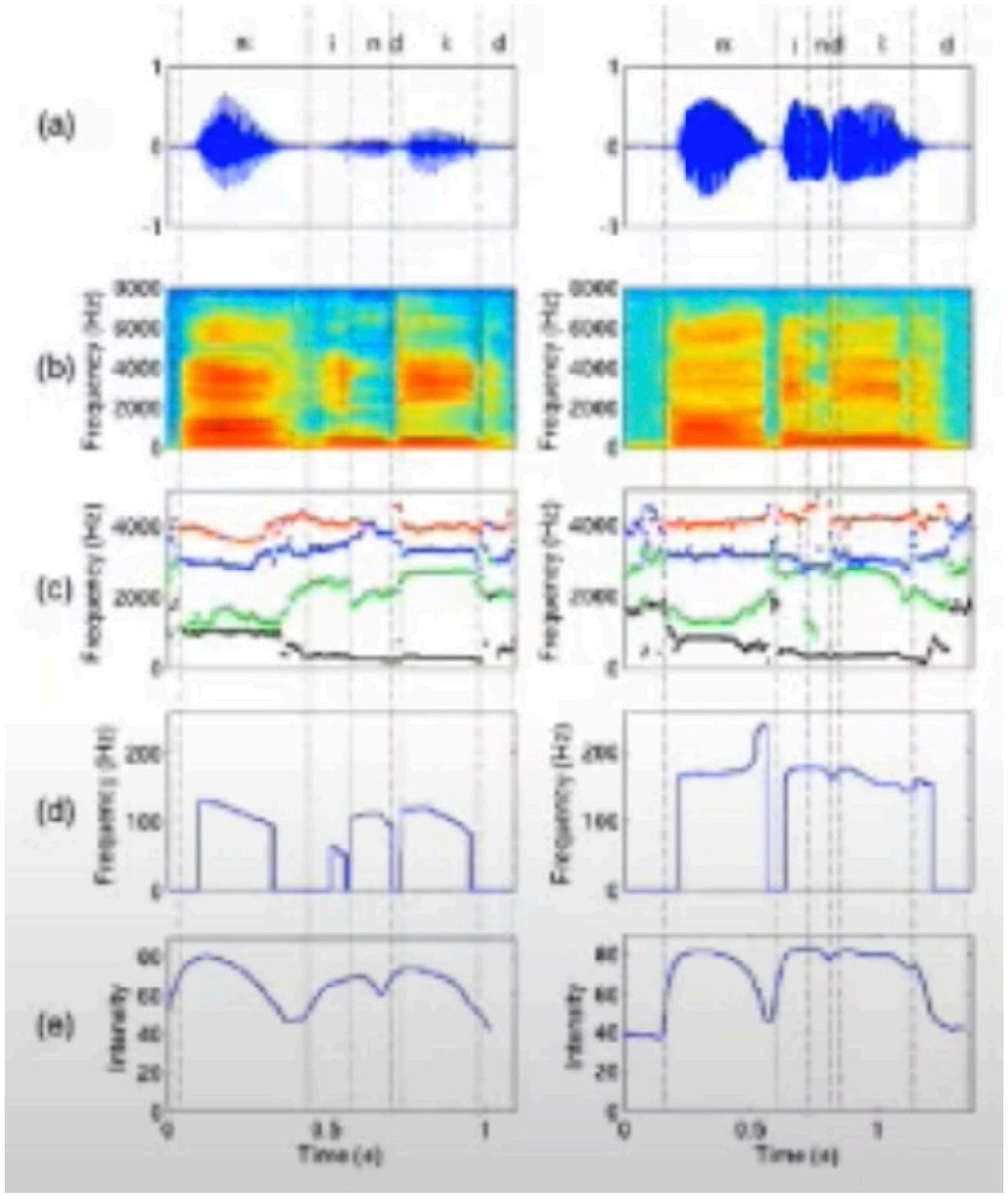
Wav

Spectrogram

Formants

Pitch

Energy

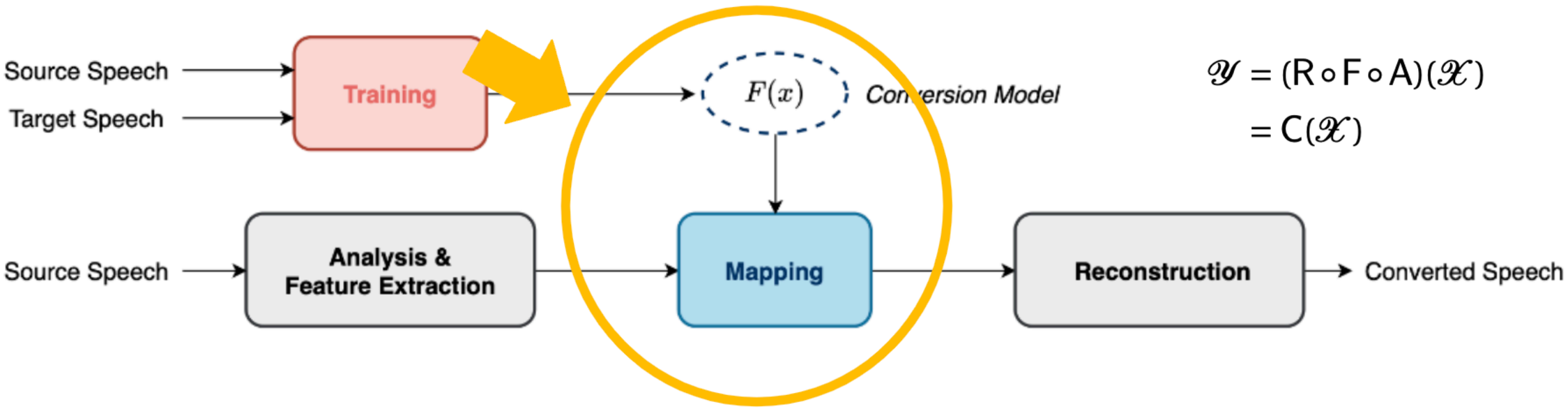


Timbre: Formants

Prosody: Pitch, Intensity, Duration

Figure from Li et al, APSIPA Tutorial 2020 (Theory and Practice of Voice Conversion)

Basic Structure of VC Systems



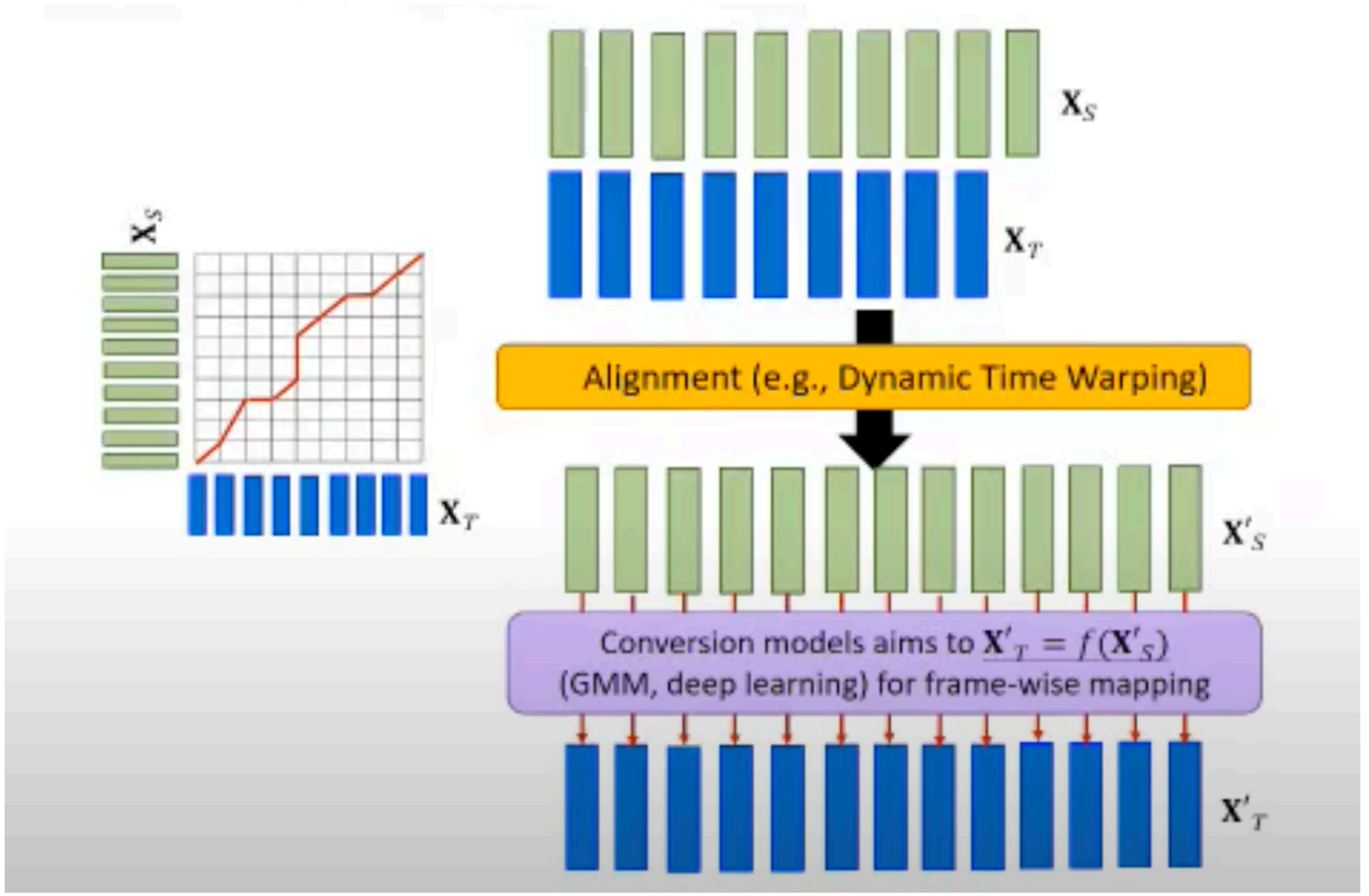
- Training time vs inference time
- Conversion: Mapping from source acoustic features to target acoustic features

Figure from Sisman et al. "An overview of voice conversion and its challenges: From statistical modeling to deep learning.", TALSP 2020.

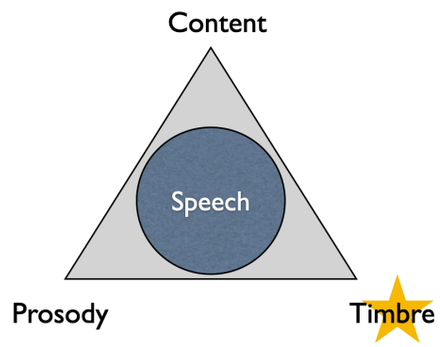
Table of Contents

Parallel Dataset

- Same linguistic content
- Convert speaker information
- Alignment (DTW)
- Frame-to-frame mapping



Courtesy Yu Tsao



Traditional Parallel Conversion

- Vector Quantization: maps codewords between source and target codebooks
- Gaussian Mixture Models [Toda et al, TALSP 2007]: represents the relationship between two sets of spectral envelopes
- Non-negative Matrix Factorization ENMF-VC [Wu et al, TASLP 2014]

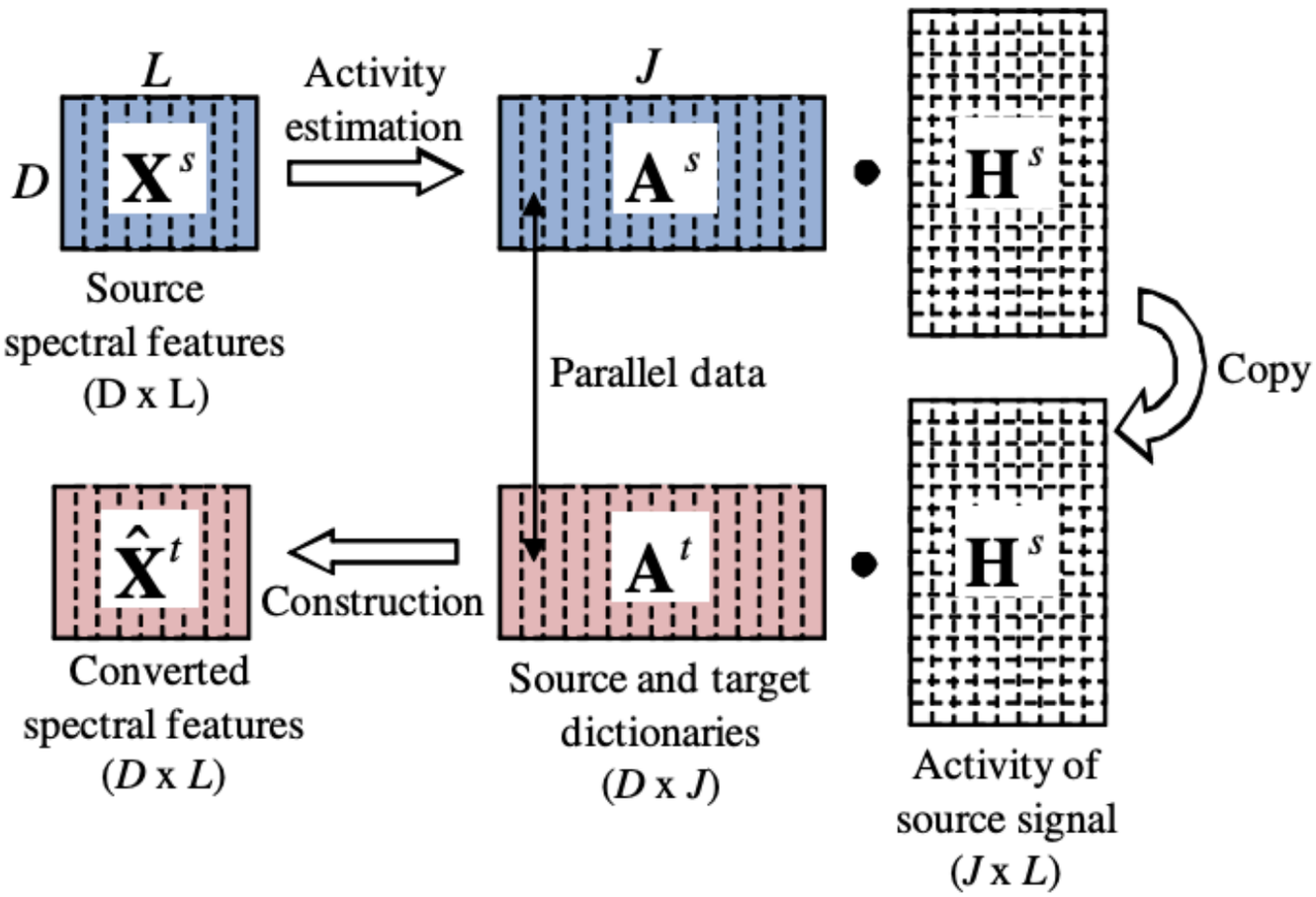
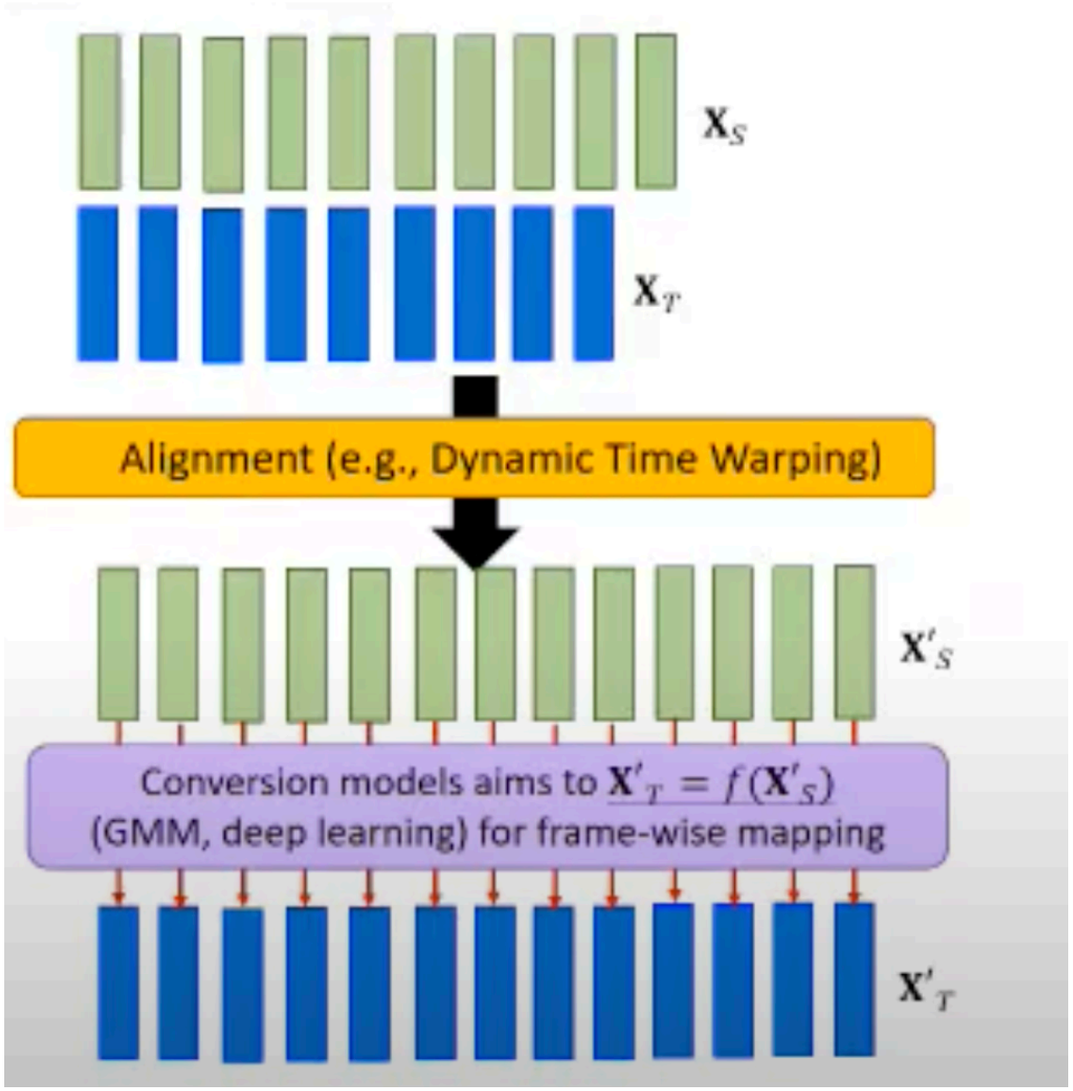


Figure from Aihara, Ryo, et al. "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary."

Deep Learning Parallel Conversion



Courtesy Yu Tsao

Conversion Models

CNN

[Chen. Et al, TALSP 2014]

[Desai. Et al, TALSP 2010]

RNN

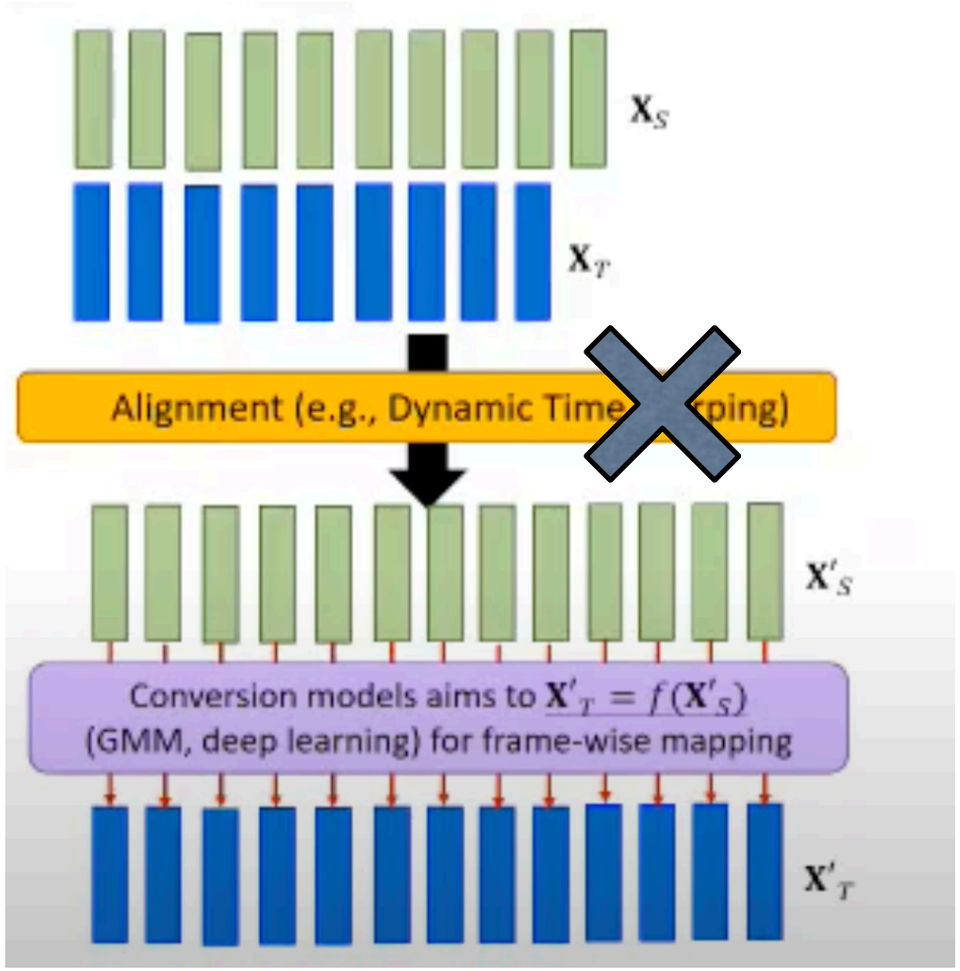
[Nakashika. Et al, interspeech 2014]

Transformer Attention

[ATTS2S-VC, Tanaka. et al]

Table of Contents

Non-parallel Dataset



Courtesy Yu Tsao

Non-parallel VC

Style Transfer Methods

Style transfer GANs

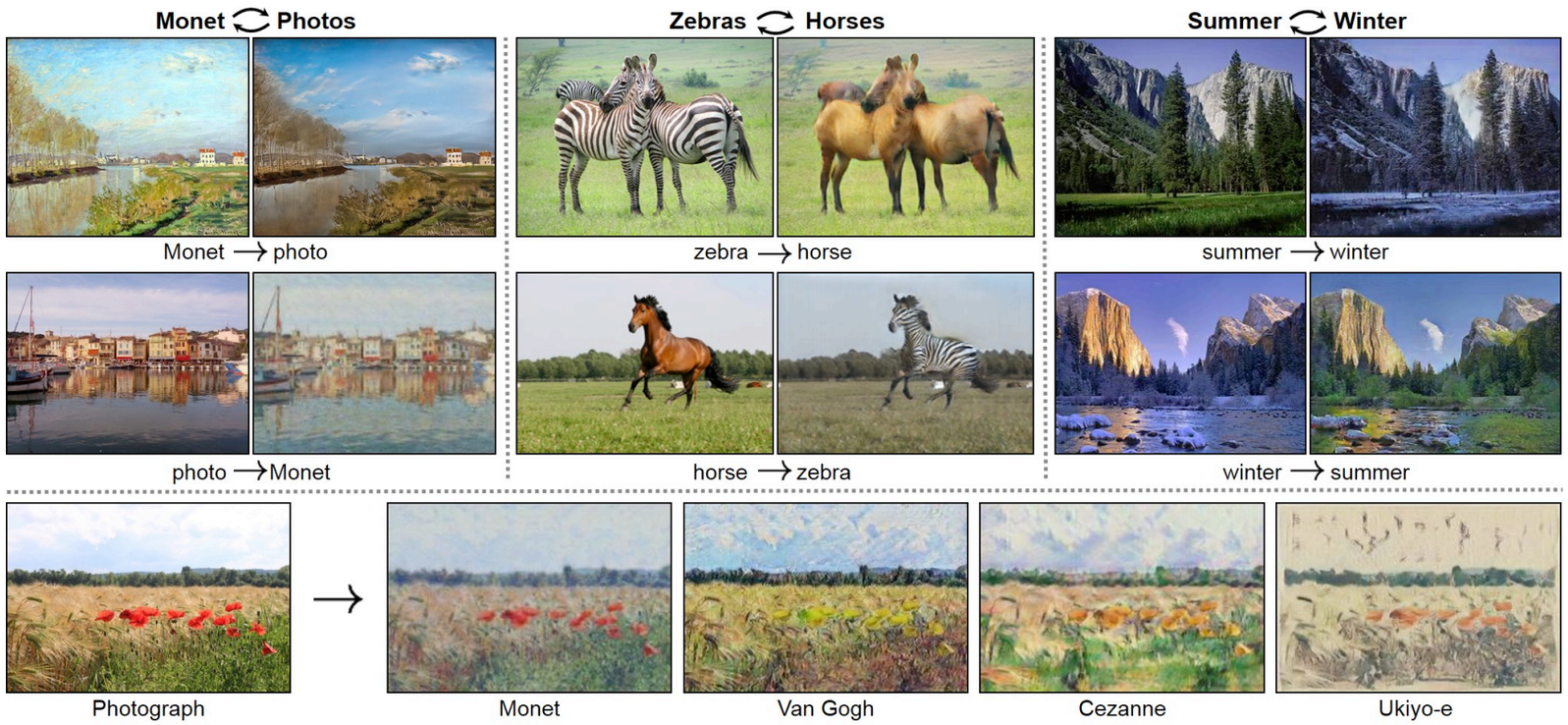
Disentanglement-based Methods

Information bottleneck
Pre-trained ASR
Self-supervised encoders
Latent diffusion and prompting

Table of Contents

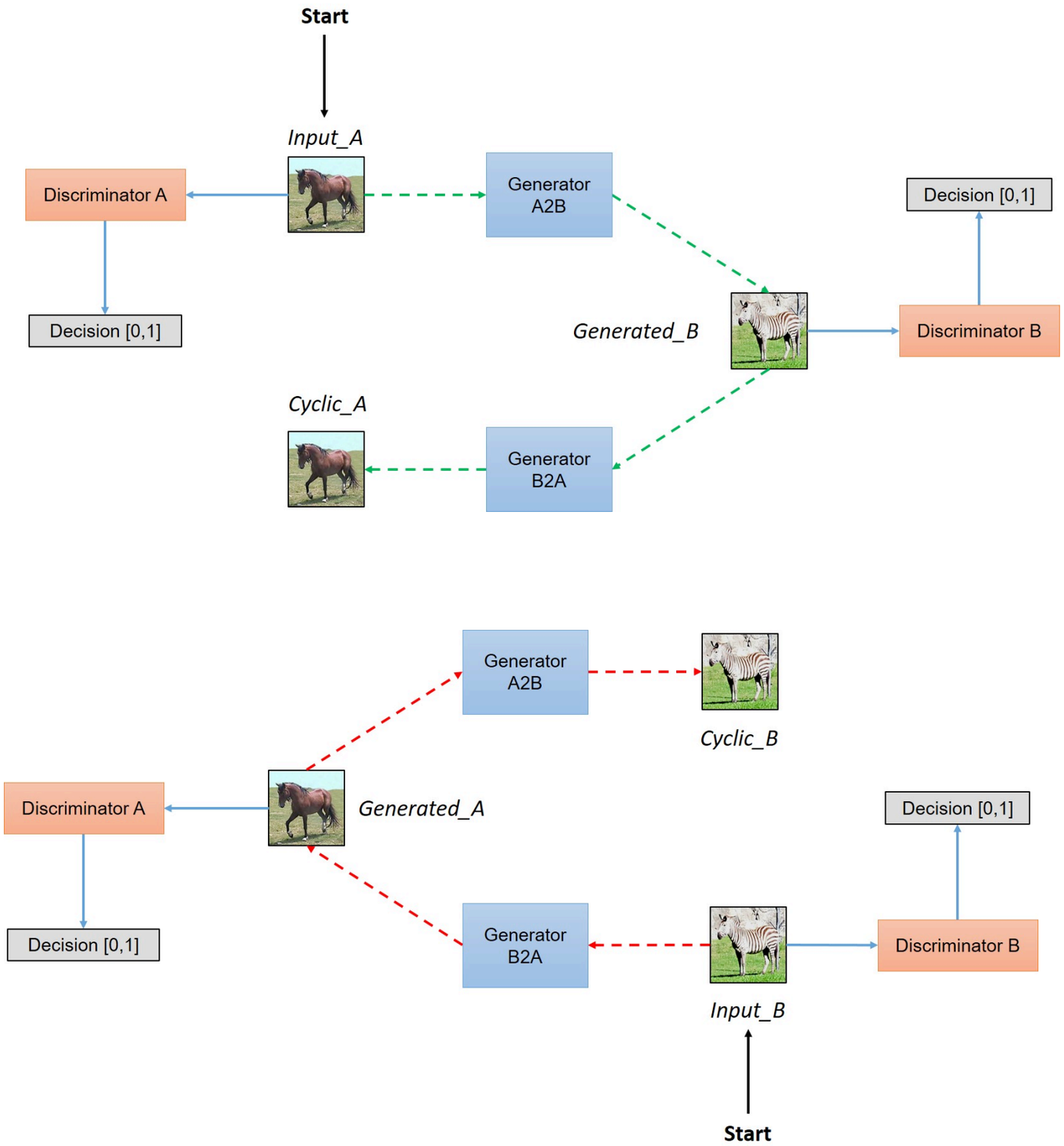
Style Transfer Methods

Image style transfer task



CycleGAN: Unpaired Image-to-Image Translation. Figure from: https://hardikbansal.github.io/CycleGANBlog/

Style Transfer Methods - CycleGAN

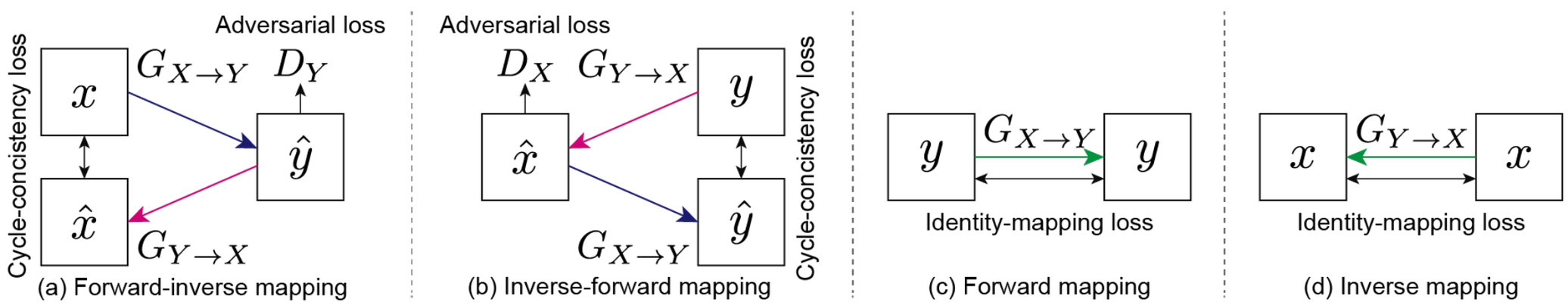


CycleGAN: Unpaired Image-to-Image Translation. Figure from: https://hardikbansal.github.io/CycleGANBlog/

Style Transfer Methods

CycleGAN-VCs: CycleGAN-VC (EUSIPCO 2018),
CycleGAN-VC2, CycleGAN-VC3, MaskCycleGAN-VC (samples)
StarGAN-VCs: StarGAN-VC (SLT 2018), StarGAN-VC2, StarGAN-VC++
WaveCycleGANs: WaveCycleGAN, WaveCycleGAN2

Inspired by image style transfer task
Match the distribution



Source Target Converted

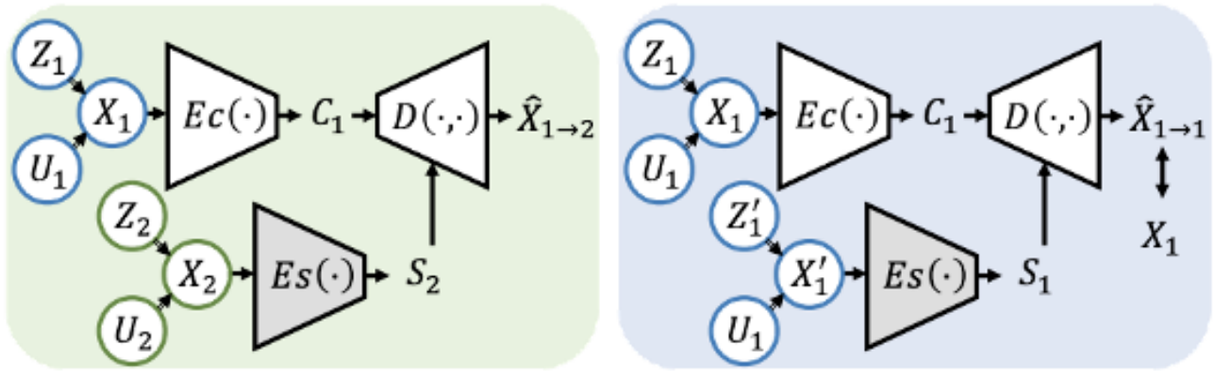
Table of Contents

Disentanglement-based Methods

Information bottleneck: VQVAE

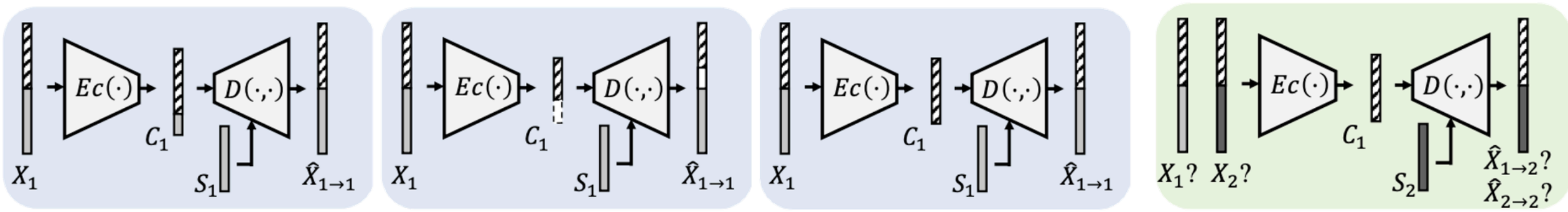
AutoVC [Qian et al, ICML 2019]

Source Target Converted



(a) Conversion (b) Training

Style transfer autoencoder framework



(a) Bottleneck too wide (b) Bottleneck too narrow (c) Bottleneck just right (d) Conversion

An intuitive explanation of how AutoVC works

SpeechSplit (2020), AutoPST (2021), SPEECHSPLIT2.0 (2022)

Disentanglement-based Methods

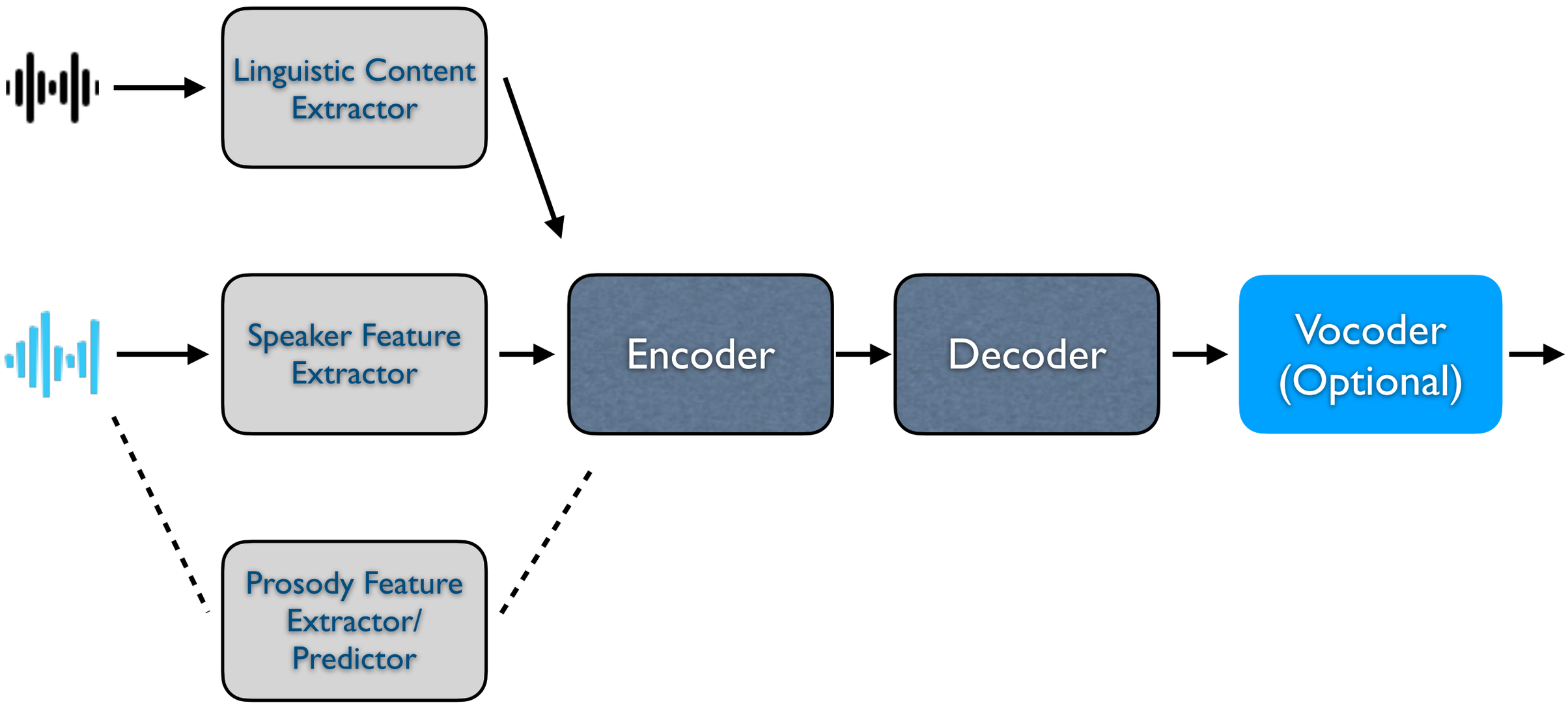
Pre-trained ASR: ASR + TTS

Intermediate representation: text, phoneme posterior gram(PPG), batch normalization (BN)
Best disentanglement performance

Samples

Disentanglement-based Methods

Encoder-decoder Architecture (Usually Text-free)



Self-supervised encoders

ControlVC

Recent developments in neural speech synthesis and vocoding have sparked a renewed interest in voice conversion (VC). Beyond timbre transfer, achieving controllability on para-linguistic parameters such as pitch and rhythm is critical in deploying VC systems in many application scenarios.

However, existing studies:

- Only provide utterance-level global control
- Lack of interpretability on the controls

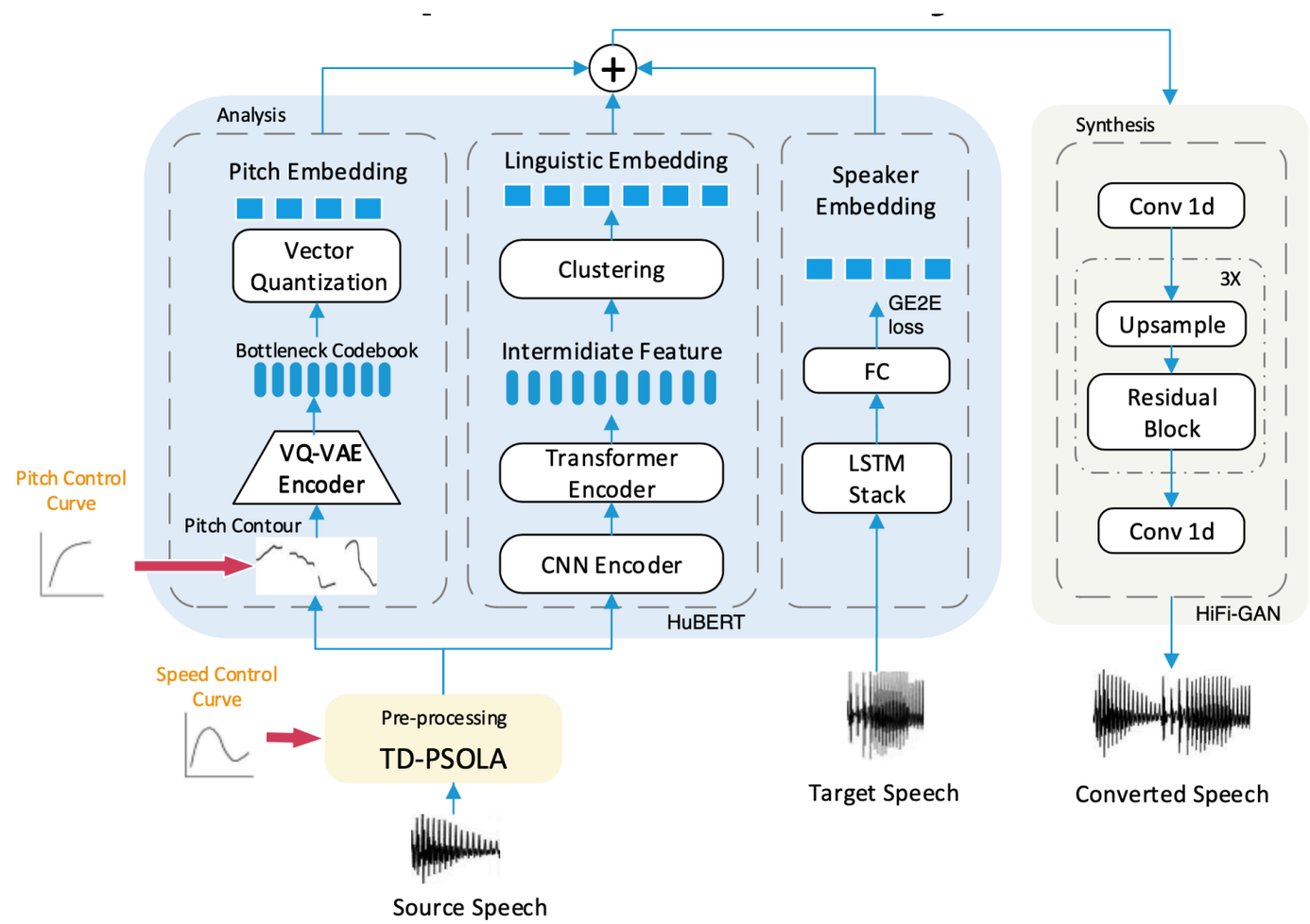
We propose ControlVC,

the first end-to-end and zero-shot

neural VC system that achieves:

- Time-varying controls on pitch and rhythm
- Intuitive control using user input curve

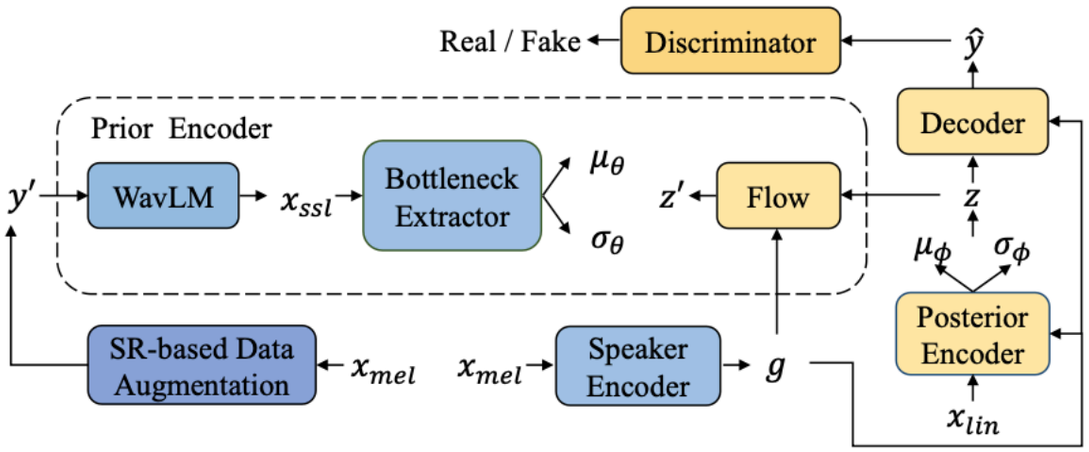
[paper][demo][code]



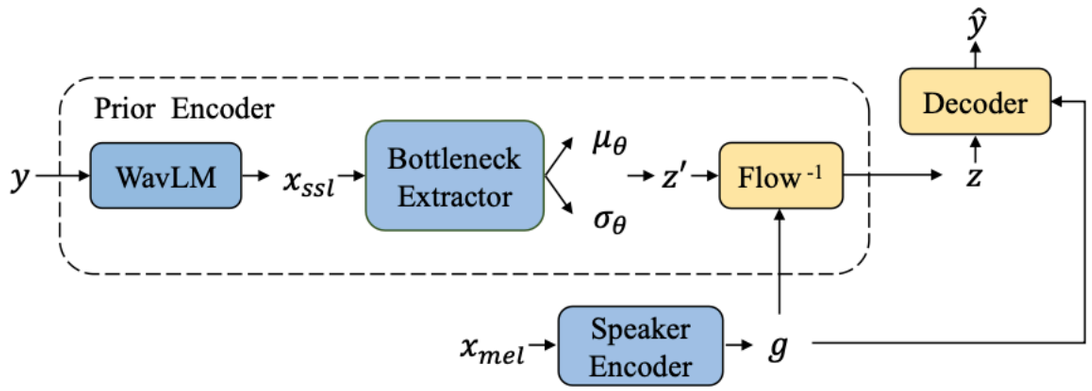
Disentanglement-based Methods

Self-supervised encoders

FreeVC



(a) Training procedure



(b) Inference procedure

Fig. 1: Training and inference procedure of FreeVC. Here y denotes source waveform, y' denotes augmented waveform, $\hat{y}$ denotes converted waveform, x_{mel} denotes mel-spectrogram, x_{lin} denotes linear spectrogram, x_{ssl} denotes SSL feature, and g denotes speaker embedding.

Li, Jingyi, Weiping Tu, and Li Xiao. "Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

Disentanglement-based Methods

Latent diffusion and prompting [Samples]

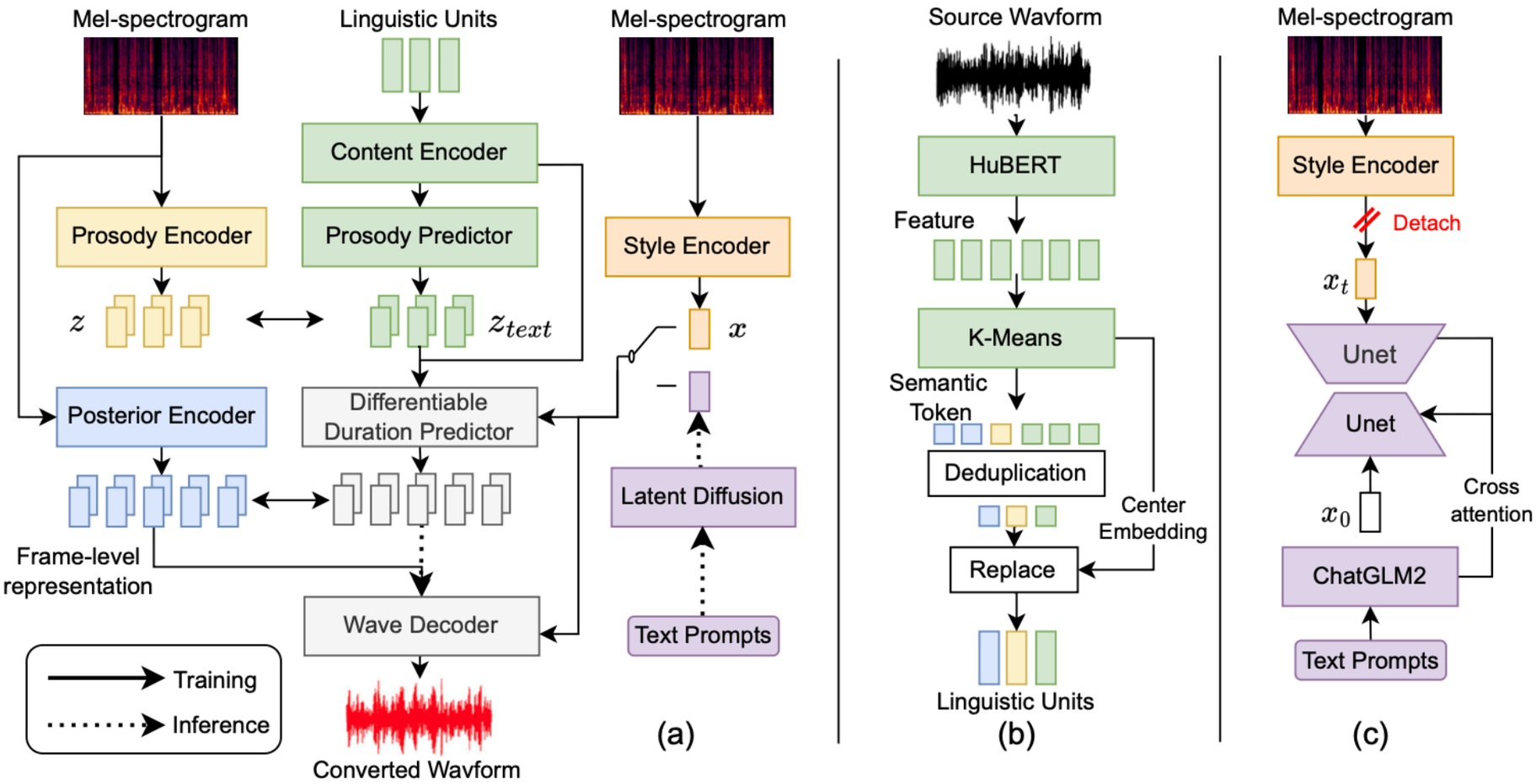


Fig. 1. The details of our proposed approach. Subfigure (a) is the architecture of PromptVC. The solid line indicates the training stage while the dashed line represents the inference stage. Subfigures (b) and (c) illustrate the process of linguistic unit extraction and the training procedure of the latent diffusion model, respectively.

Yao, Jixun, et al. "PromptVC: Flexible Stylistic Voice Conversion in Latent Space Driven by Natural Language Prompts." arXiv preprint arXiv:2309.09262 (2023).

Table of Contents

Evaluation of Conversion Quality

Naturalness

Similarity

Objective Evaluation

Spectral Conversion

- Mel-Cepstral distortion (MCD) [Kubichek et al, 1993]
  - M is Mel-cepstral coefficients
  - K is frame
  - I is dimension of m
  - The lower the better

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{24}(m_{k,i}^{t} - m_{k,i}^{c})^2}$$

Prosody Conversion: phonetic duration, pitch contour, energy contour

- Person coefficients (PCC)
  - The higher the better

$$\rho(F0^c, F0^t) = \frac{cov(F0^c, F0^t)}{\sigma_{F0^c}\sigma_{F0^t}}$$

- RMSE
  - The lower the better

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F0_k^c - F0_k^t)^2}$$

- Not always correlated with human perception
- Subjective evaluation is needed

Subjective Evaluation

MOS: Mean Opinion Score (MOS)
In MOS experiments, listeners rate the quality of the converted voice using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. Similarity and naturalness.

Similar metrics: MUSHRA, requires fewer participants

AB test:
Listeners are presented with two speech samples and asked to indicate which one has more of a certain property.

Similar metrics: ABX test

BWS: Best-Worst Scaling
Listeners are presented only with a few randomly selected options each time.

Evaluation

Neural network based:

MOSNET [Lo et al, Interspeech 2019]

STOI-NET [Zezario et al, APSIPA 2020]

STOI: short-time objectivity intelligibility
Highly correlated with the intelligibility of degraded speech signals, e.g., due to additive noise, single/multi-channel noise reduction, binary masking and vocoded speech.

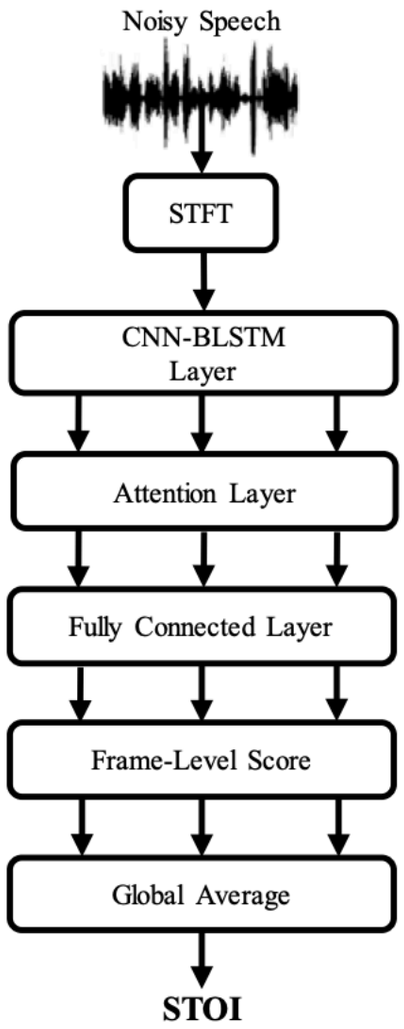


Fig. 1: Architecture of the STOI-Net model.

Figure from STOI-NET [Zezario. et al, APSIPA 2020]

STOI explanation from https://ceestaal.nl/code/

What we covered today…

1. Introduction
2. Speech Disentanglement
3. Applications
4. Parallel Dataset
5. Parallel VC (traditional & DL)
6. Non-parallel VC (DL)
7. Objective & Subjective Evaluation

References

1. Li et al, APSIPA Tutorial 2020 (Theory and Practice of Voice Conversion)
2. Sisman et al. "An overview of voice conversion and its challenges: From statistical modeling to deep learning.", TALSP 2020.
3. Wang, Disong, et al. "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction." ICASSP 2020.
4. Kaneko, Takuhiro, and Hirokazu Kameoka. "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks." 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.
5. Zezario, Ryandhimas E., et al. "STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model." 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020.
6. Lo, Chen-Chou, et al. "Mosnet: Deep learning based objective assessment for voice conversion." arXiv preprint arXiv:1904.08352 (2019).
7. 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.
8. Aihara, Ryo, et al. "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.
9. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen 'A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech', ICASSP 2010, Texas, Dallas.

Thank you!

Melissa Chen

meiying.chen@rochester.edu

https://github.com/MelissaChen15

Feel free to contact me!