Theory and Practice of Voice Conversion

Dr. Berrak Sisman (SUTD, Singapore)

Dr. Yu Tsao (Academia Sinica, Taiwan)

Dr. Haizhou Li (NUS, Singapore)

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

中央研究院 ACADEMIA SINICA

NUS National University of Singapore

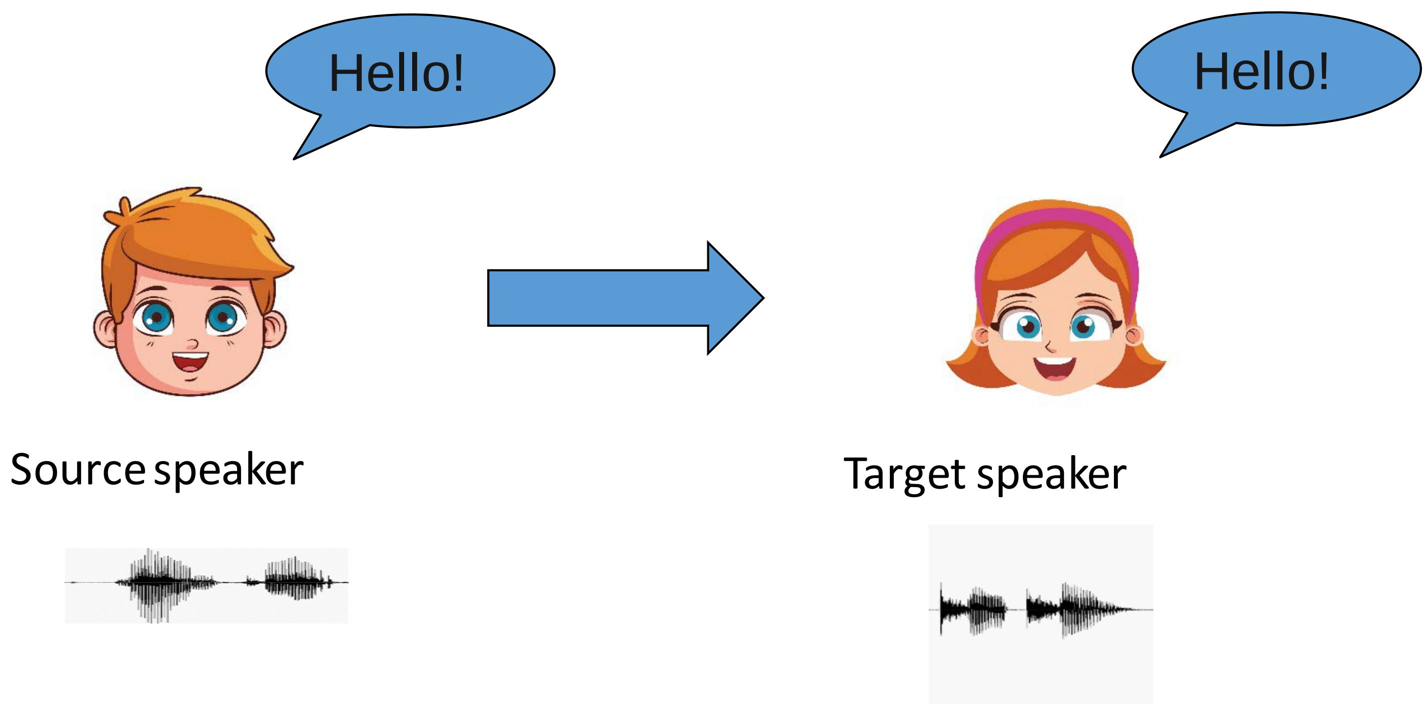
Content

- Introduction
- Voice Conversion History & Applications
- Parallel Data for Voice Conversion
  - Traditional Approaches
  - Deep Learning Era
- Beyond Parallel Data for Voice Conversion
  - Non-parallel data of paired speakers
  - Disentangling speaker from linguistic content
  - Leveraging TTS systems
  - Leveraging ASR systems
- Evaluation of Voice Conversion
  - Traditional methods
  - Neural approaches
- Voice Conversion Challenges
- Emotional VC and a new dataset (ESD)

Introduction

Voice conversion:

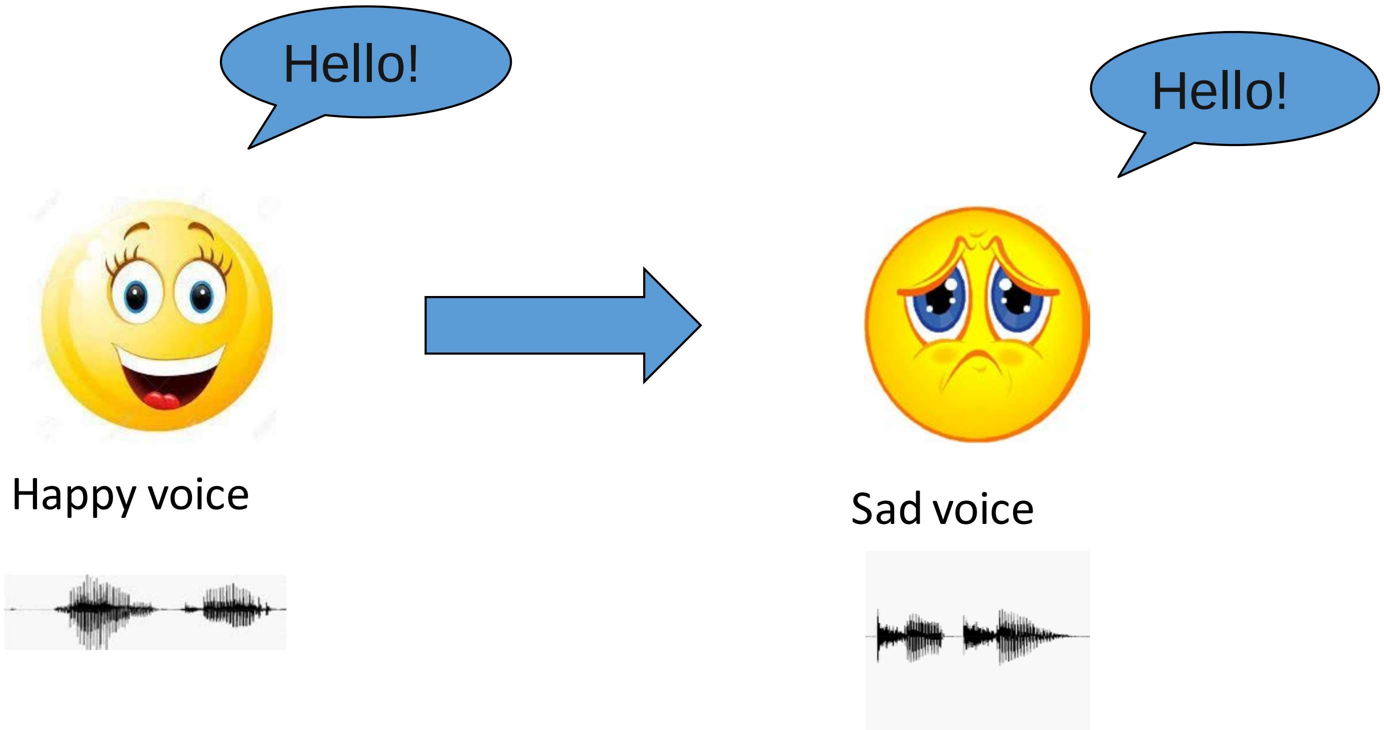
To convert one's voice to sound like that of another without changing the language content (with or without parallel data).



Introduction

Emotional Voice conversion:

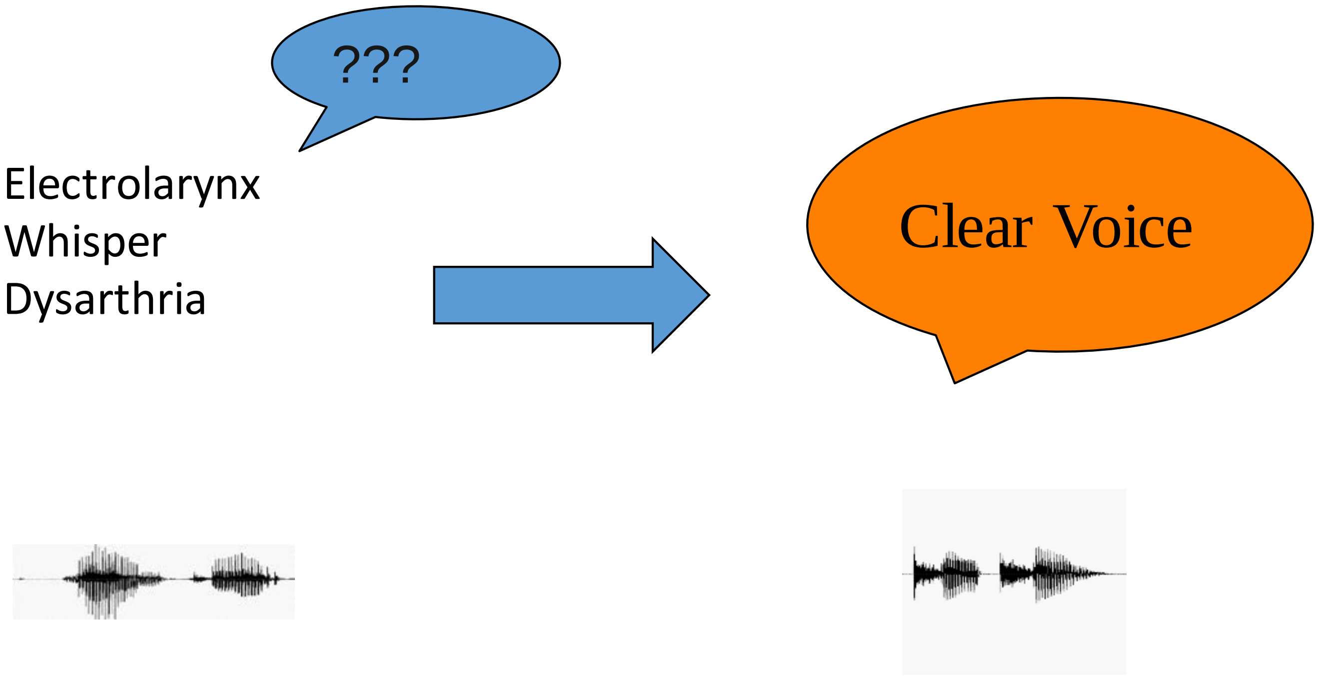
To convert one's voice from one emotion state to another.



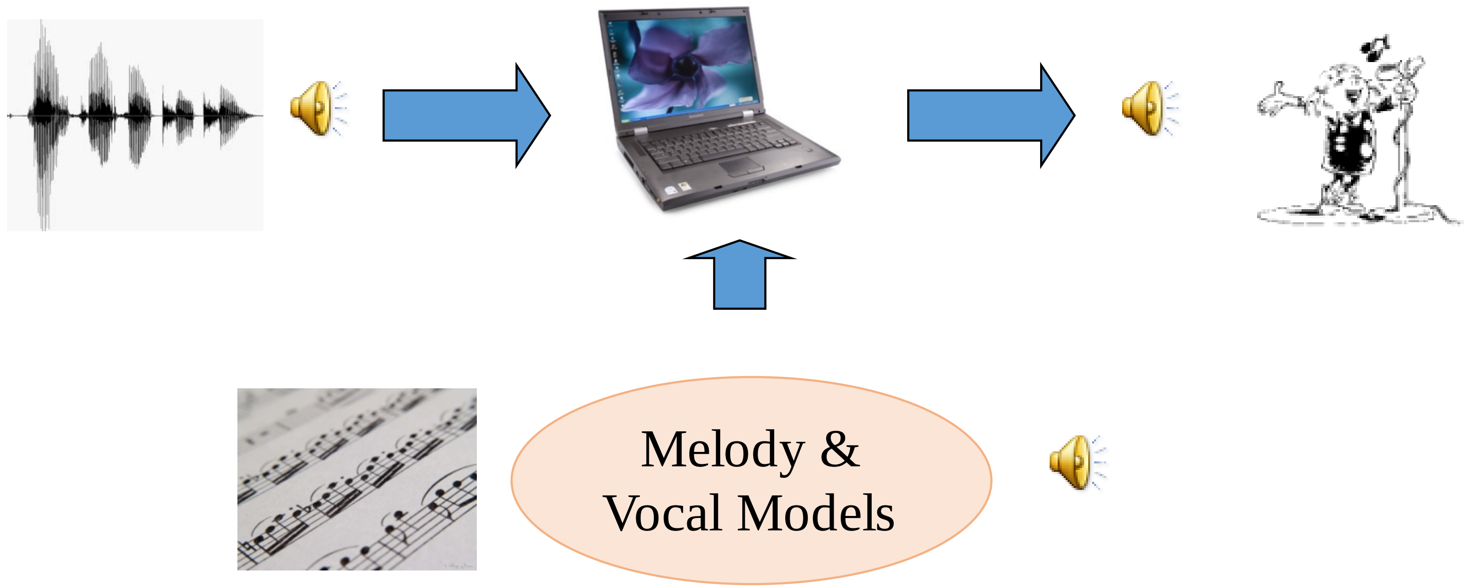
Introduction

Voice Conversion:

To improve perceptual quality of speech.



Voice Conversion Applications



- To convert speech to singing vocal
- Changing prosody and timbre from speech to singing, keeping the same person's voice

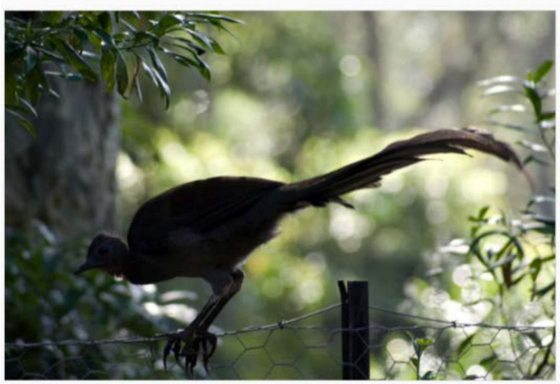
Siu Wa Lee, Ling Cen, Haizhou Li, Yaozhu Paul Chan, Minghui Dong, Method and system for template-based personalized singing synthesis, US Patent: 20150025892 A1

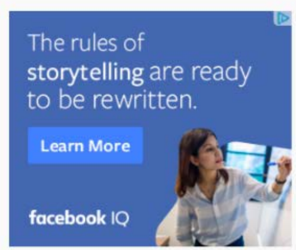
Voice Conversion Applications

- Personalized Text-to-Speech
- Dubbing of movies and games
- Speech emotion conversion
- Spoofing attack

Lyrebird is a voice mimic for the fake news era

Posted Apr 25, 2017 by Natasha Lomas (@riptari)





Crunchbase

Lyrebird –

After 20 Minutes of Listening, New Adobe Tool Can Make You Say Anything

Adobe promises never to abuse it as they use to abuse their host.

SHARE TWEET

Matthew Gault
Nov 6 2016, 3:00am

Hell no. Image: Adobe Creative Cloud/ YouTube

Differences between Speakers

Timbre (Spectrum)

Spectrogram

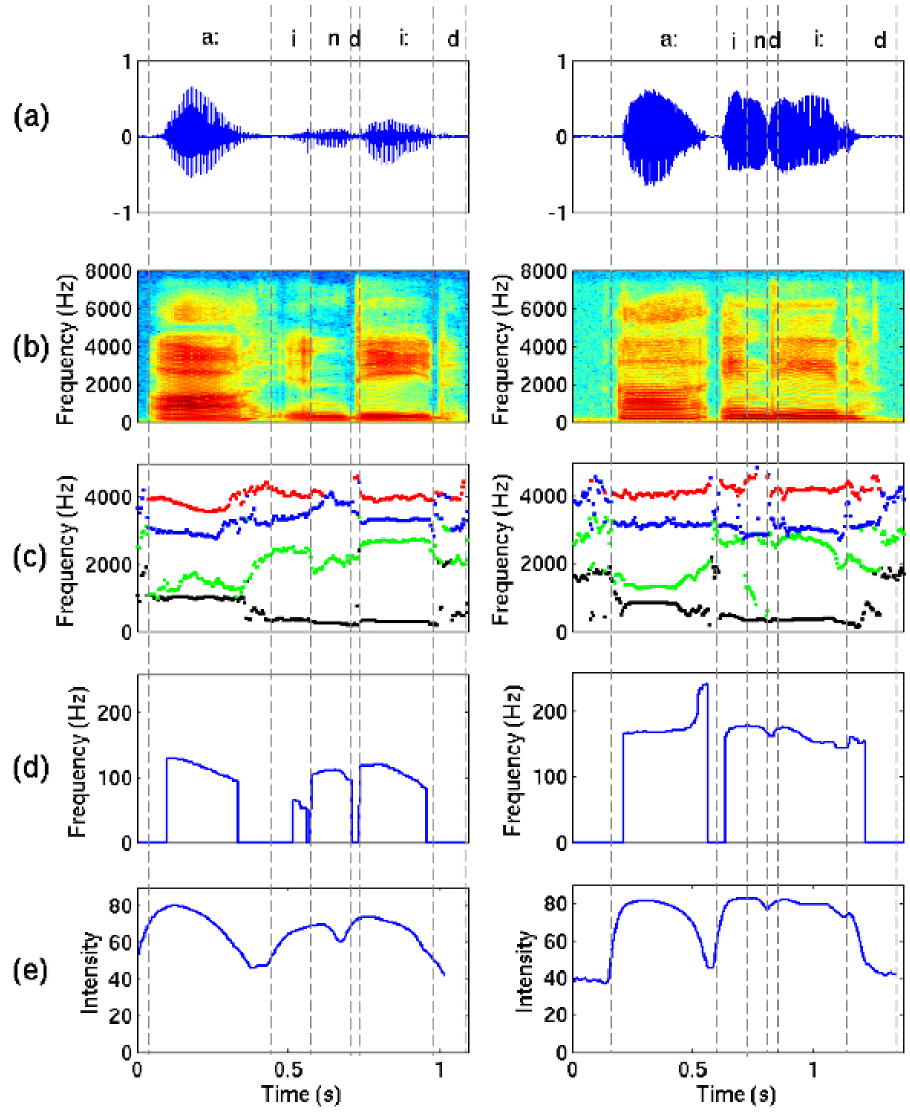
Formant

Prosody

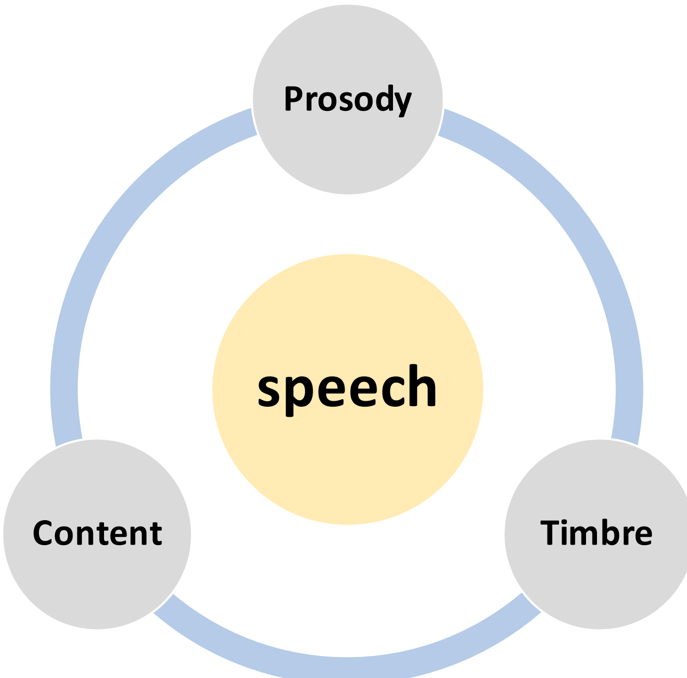
Fundamental Frequency (f0)

Intensity

Duration



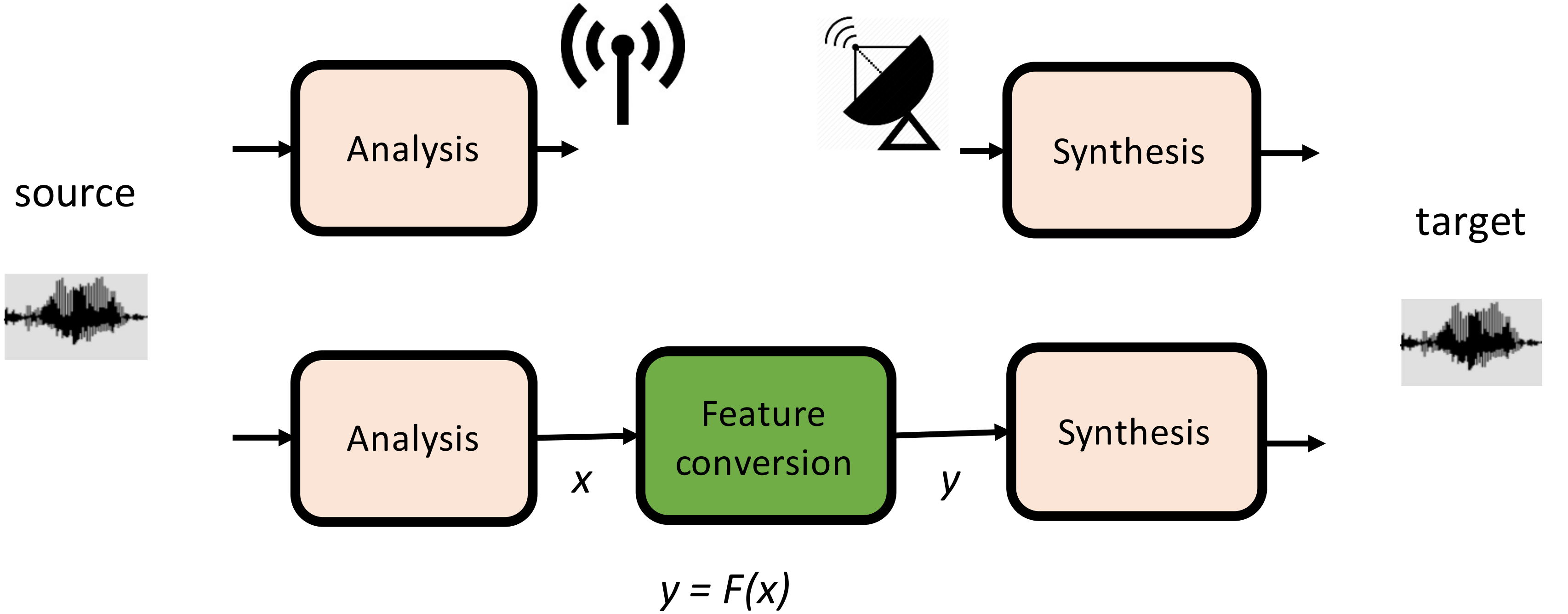
Elements of Speech



"Love stories - Why you are not alone," National Museum of Emerging Science and Innovation Tokyo , 23 April - 15 August 2006. The interface by Takashi Yamaguchi and auditory morphing sounds synthesized by Hideki Kawahara.

Vocoder

It analyses and synthesizes the human voice signal for audio data compression, multiplexing, voice encryption, voice transformation, etc.



Speech Synthesis Quality

Haskins, 1959

KTH – Stockholm, 1962

Bell Labs, 1973

MIT, 1976

MIT-talk, 1979

Speak 'N Spell, 1980

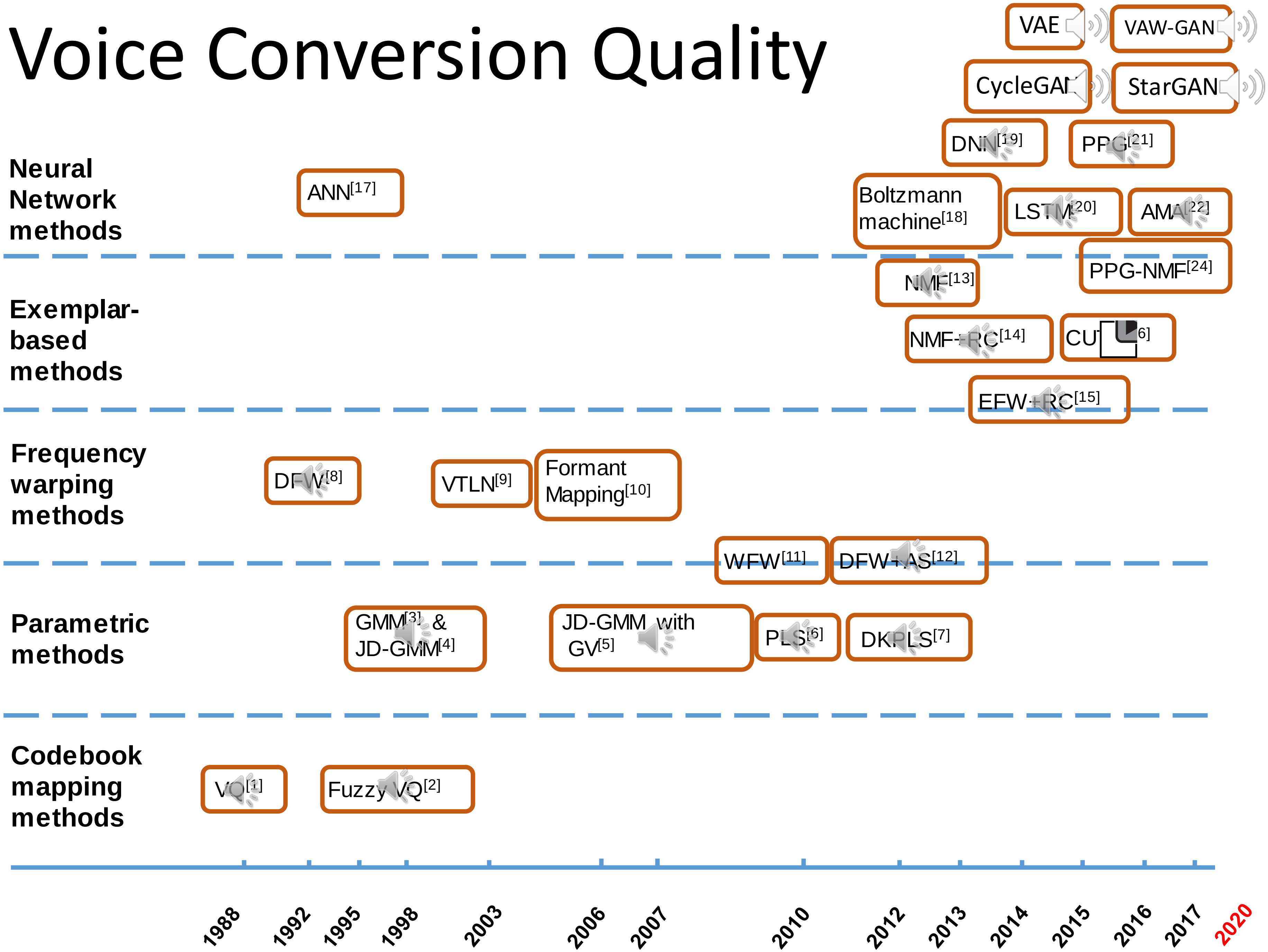
BELL Labs, 1985

DECtalk (voice morphing), 1987

Abacus 2013

Homer Dudley (1896-1987) , 1939 World Fair in New York City – Bell Labs VODER

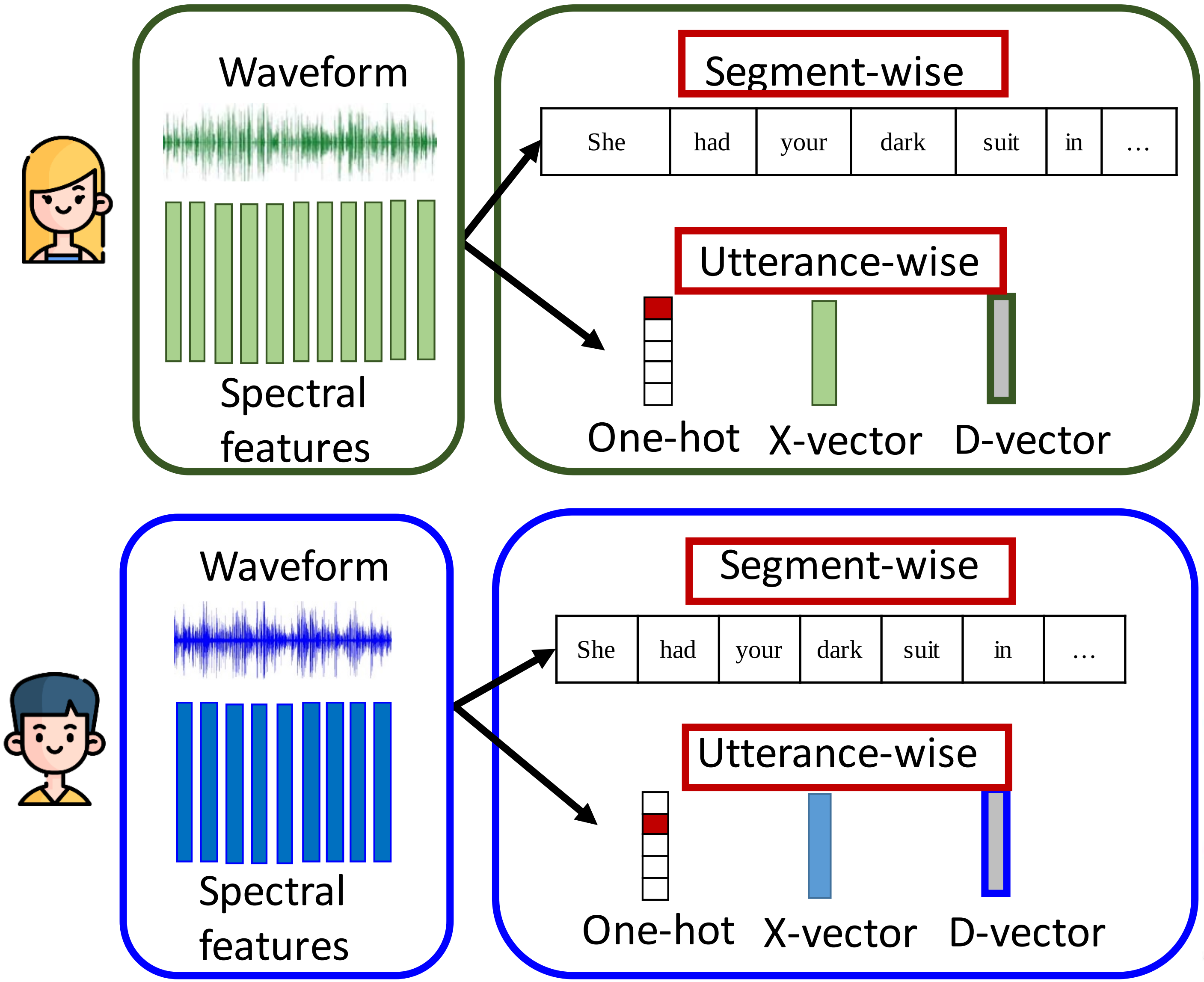
Voice Conversion Quality

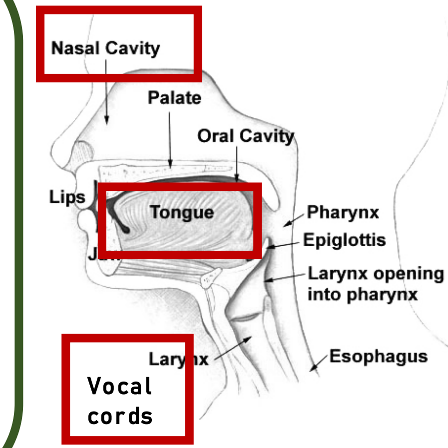


Parallel Data for Voice Conversion

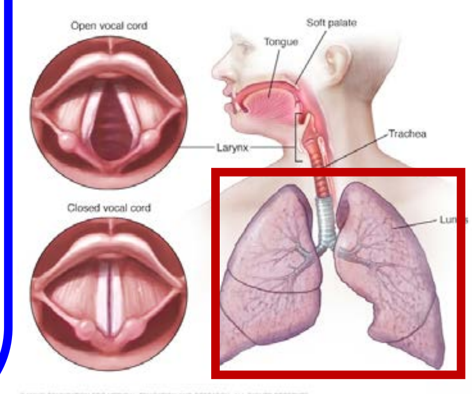
- Introduction
- Traditional Approaches
- Deep Learning Era

Segment and Utterance Information

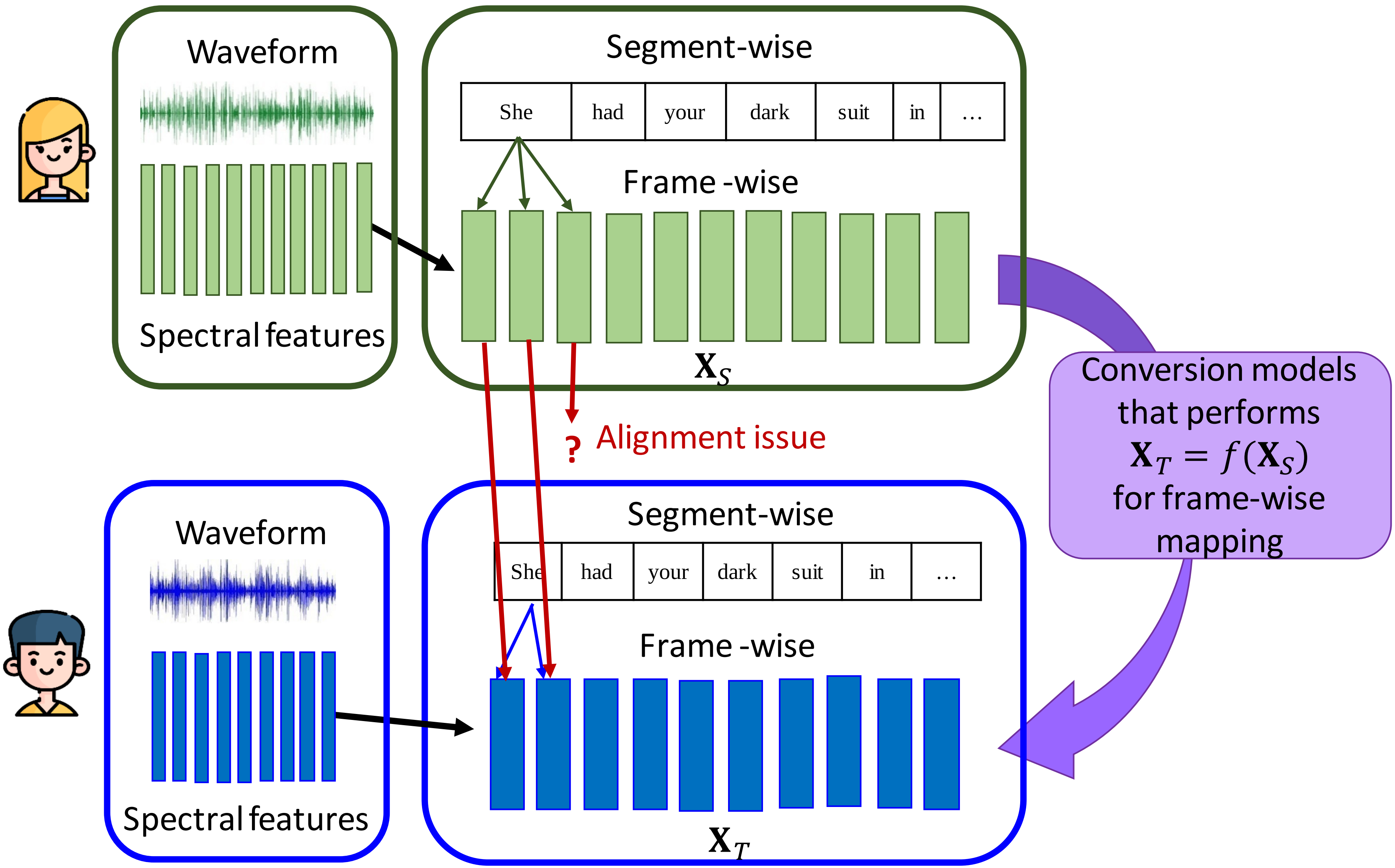




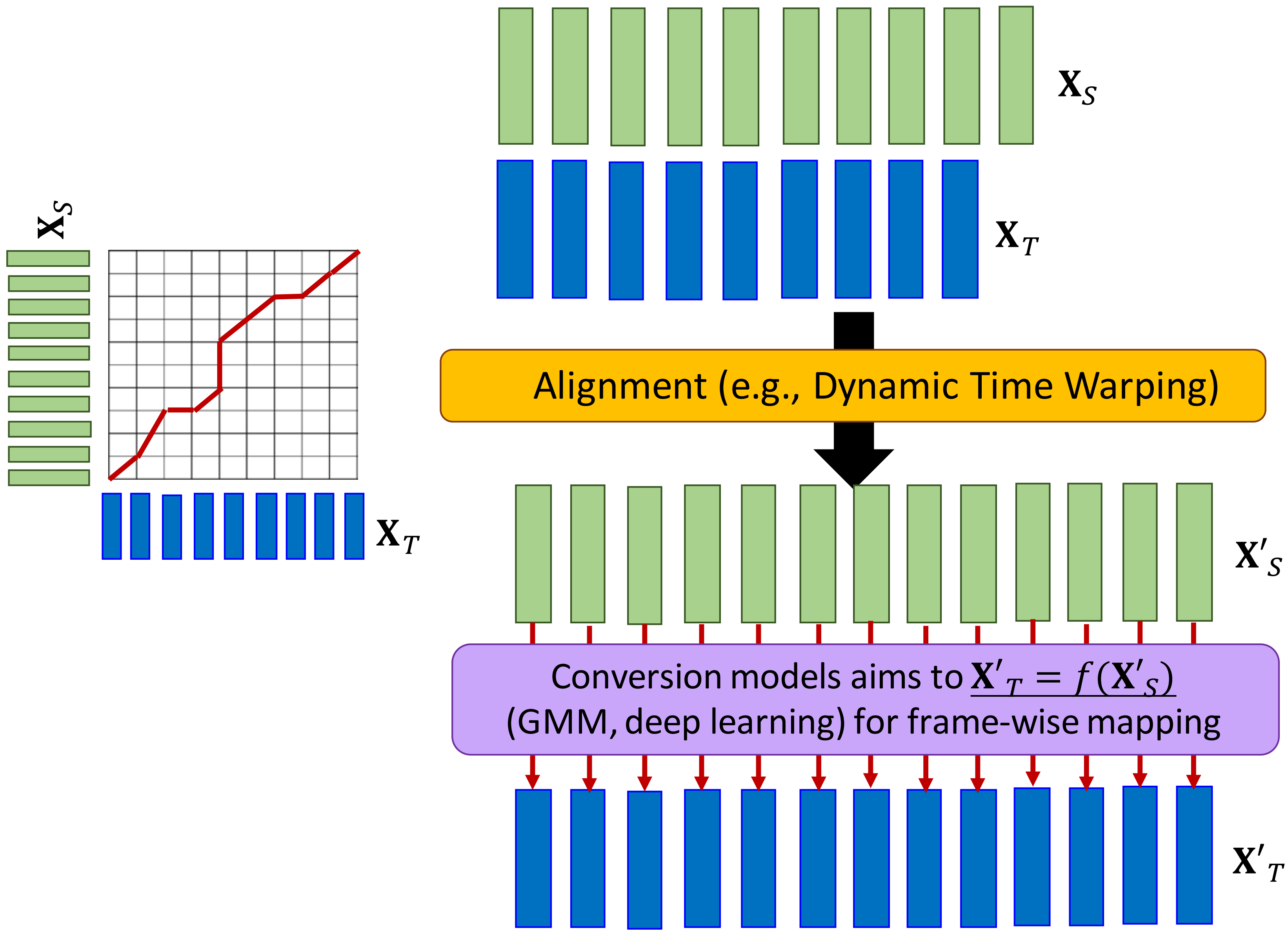
By Arcadian - http://training.seer.cancer.gov/head-neck/anatomy/overview.html



Parallel VC (Introduction)

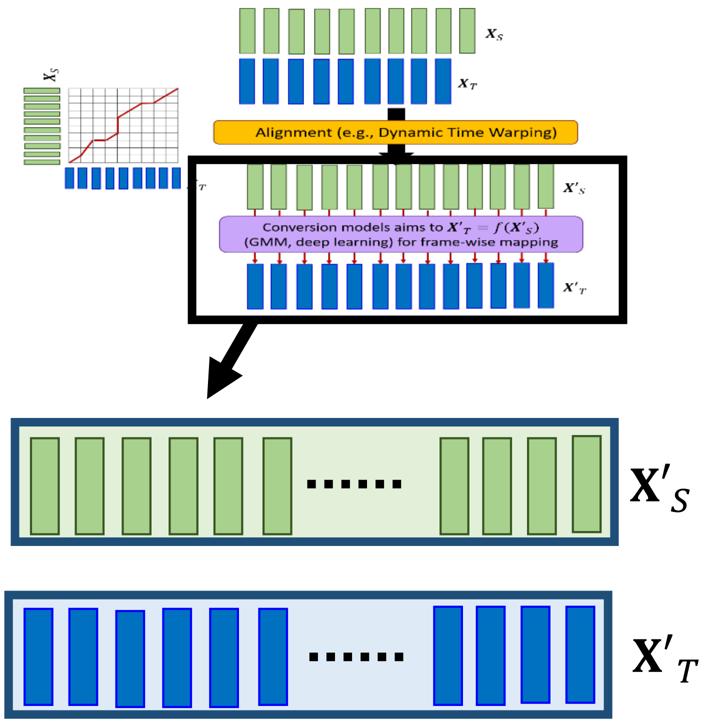


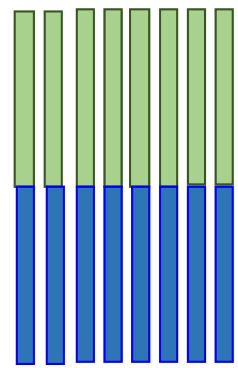
Parallel VC (Introduction)

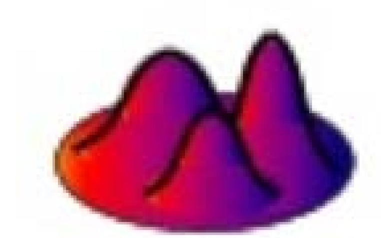


Parallel VC (Traditional Approaches)

- GMM-VC [Toda. et al, IEEE TASLP 2007]







(a) Model the joint vector by Gaussian mixture models

$$P(x_S^{(n)}, x_T^{(n)} | \Theta) = \sum_{m=1}^{M} N\left(\begin{bmatrix} x_S^{(n)} \\ x_T^{(n)} \end{bmatrix}; \begin{bmatrix} \mu_m^{(S)} \\ \mu_m^{(T)} \end{bmatrix}, \begin{bmatrix} \Sigma_m^{(SS)} & \Sigma_m^{(ST)} \\ \Sigma_m^{(TS)} & \Sigma_m^{(TT)} \end{bmatrix}\right)$$

(b) Estimating the converted speech by MMSE

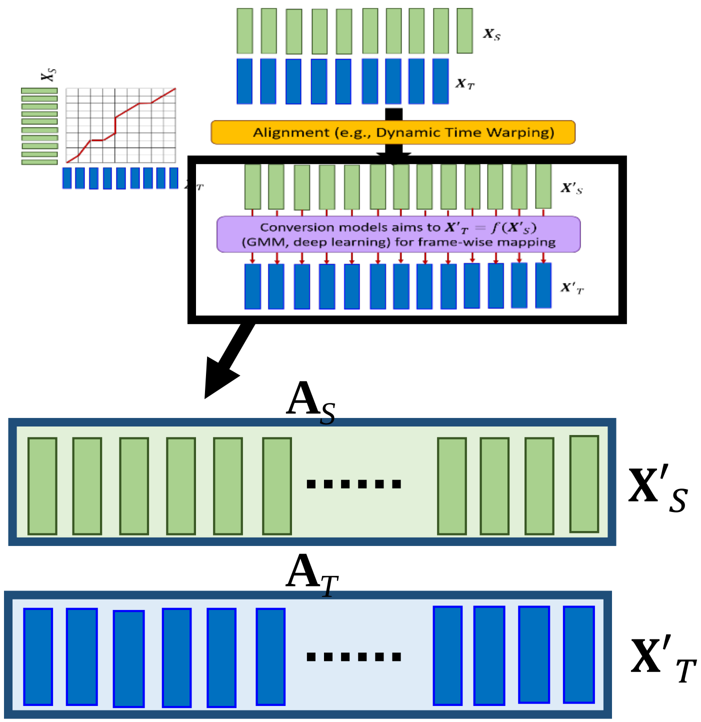
$$\hat{x}_T^{(n)} = E\left(x_T^{(n)} \middle| x_S^{(n)}\right) = \sum_{m=1}^{M} P\left(m \middle| x_S^{(n)}, \Theta\right) E_{m,t}^{T} \quad \text{or}$$

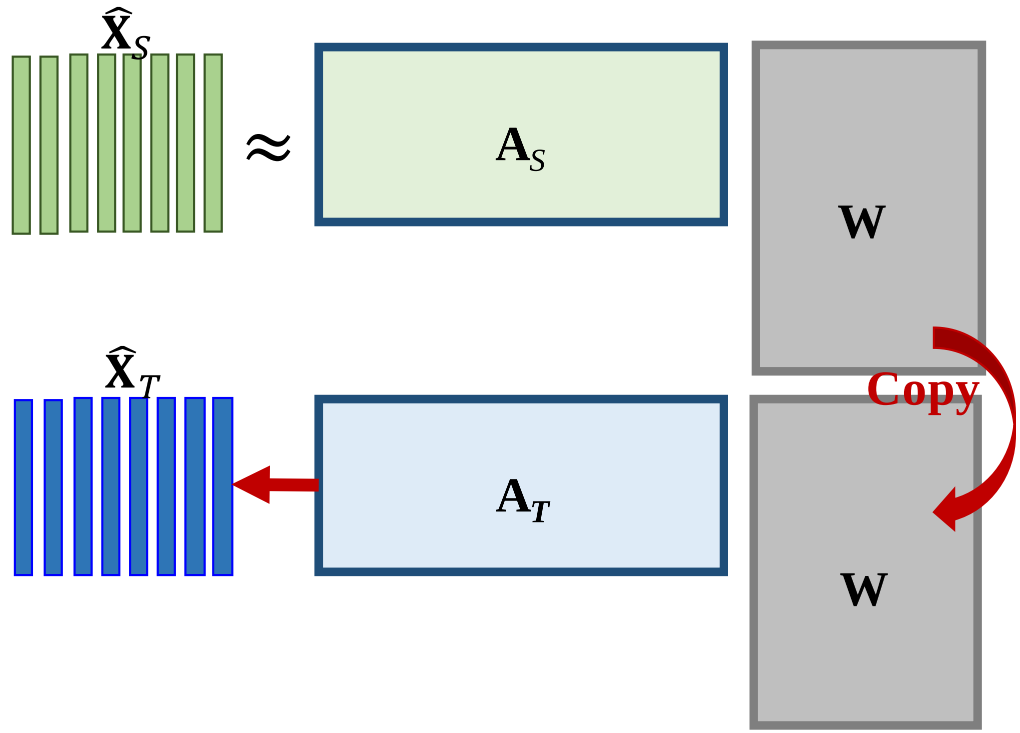
(b) Estimating the converted speech by MLPG

$$\hat{\mathbf{x}}_T = argmaxP(\mathbf{X}_T | \mathbf{X}_S, \Theta), \text{such that } \mathbf{X}_T = \mathbf{W}\mathbf{x}_T$$

Parallel VC (Traditional Approaches)

- ENMF-VC [Wu. et al, IEEE TASLP 2014]





(a) Estimating weights to reconstruct source speech

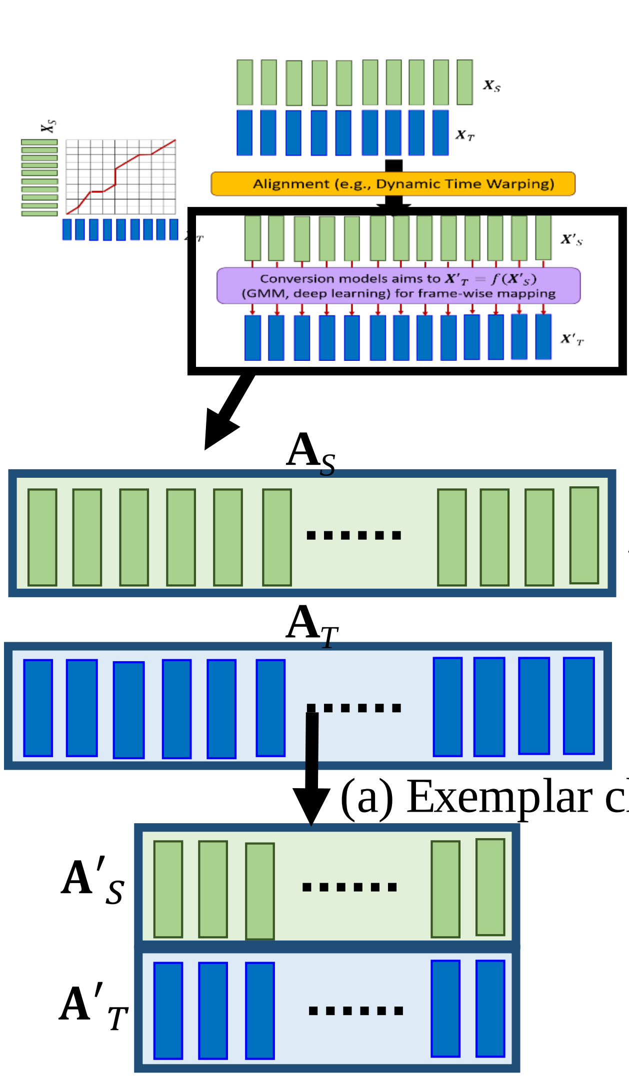
$$\mathbf{W} = argmin\ D\left(\mathbf{A}_S\mathbf{W}, \widehat{\mathbf{X}}_S\right) + \lambda\|\mathbf{W}\|_1$$

(b) Apply the weights to the target exemplars

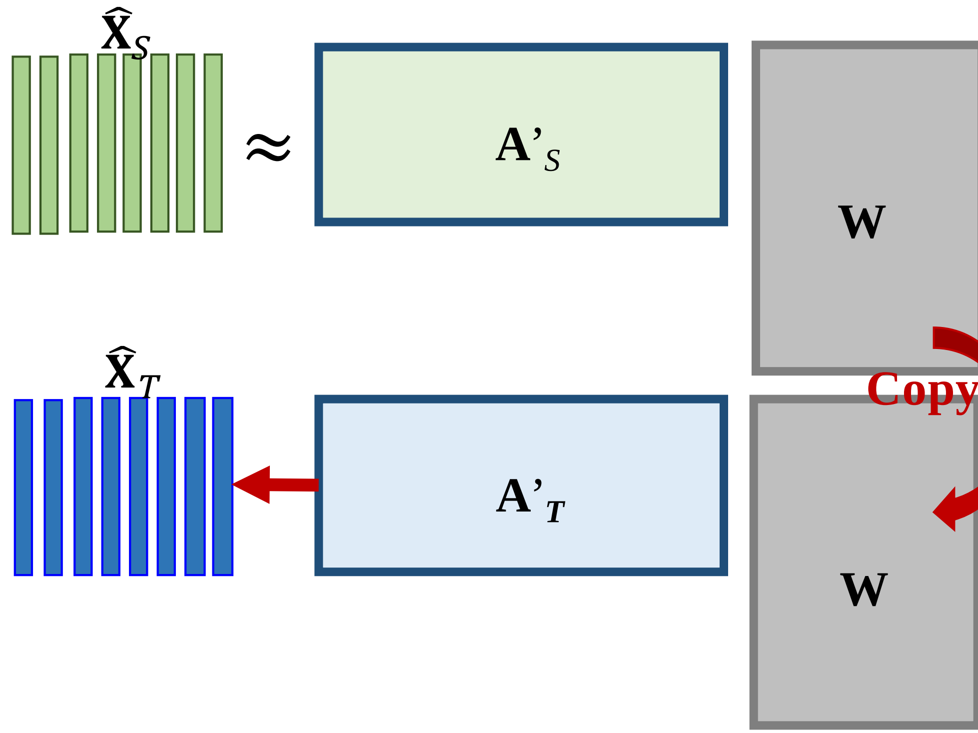
$$\mathbf{A}_S\mathbf{W} \longrightarrow \widehat{\mathbf{X}}_T$$

Parallel VC (Traditional Approaches)

- JDNMF-VC [Fu. et al, IEEE TBME 2016]



(a) Exemplar clustering



(b) Estimating weights to reconstruct source speech

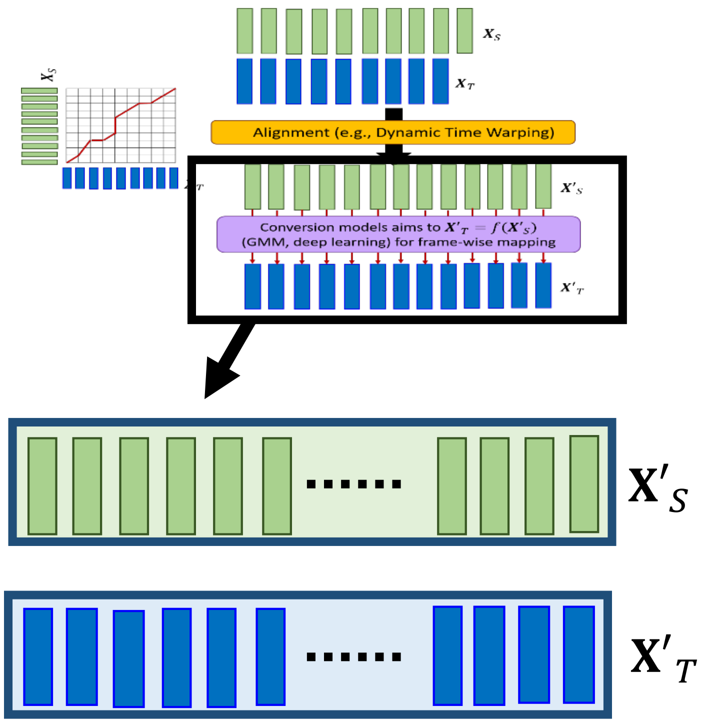
$$\mathbf{W} = argmin\ D\left(\mathbf{A}_S\mathbf{W}, \hat{\mathbf{X}}_S\right) + \lambda\|\mathbf{W}\|_1$$

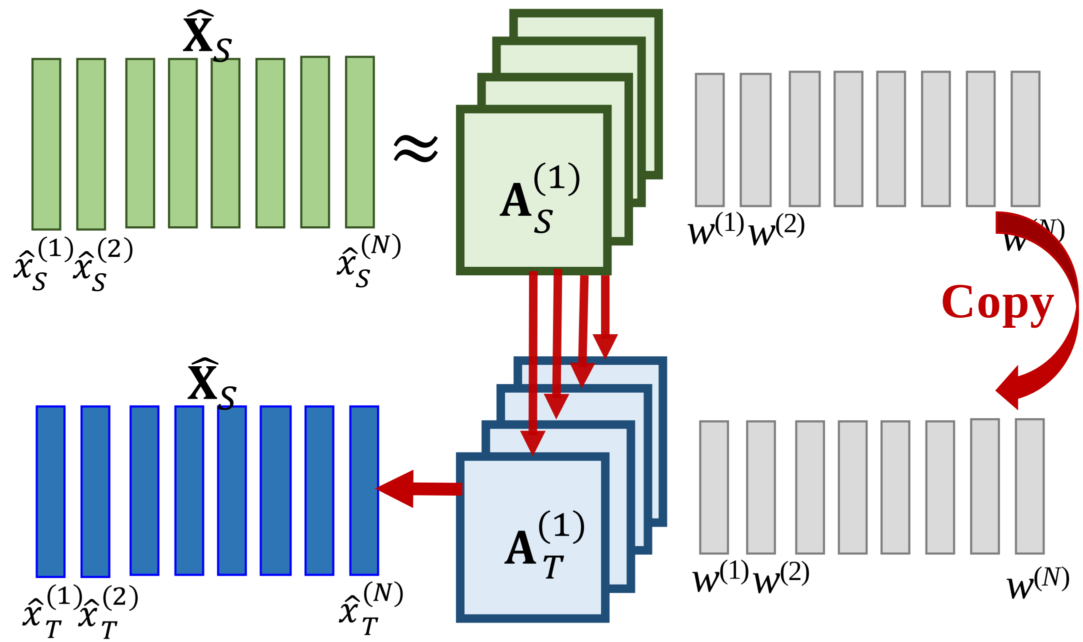
(c) Apply the weights to the target exemplars

$$\mathbf{A}_S\mathbf{W} \longrightarrow \hat{\mathbf{X}}_T$$

Parallel VC (Traditional Approaches)

- LLE-VC [Wu. et al, Interspeech 2016]





(a) Find the local patch (K nearest neighbors)

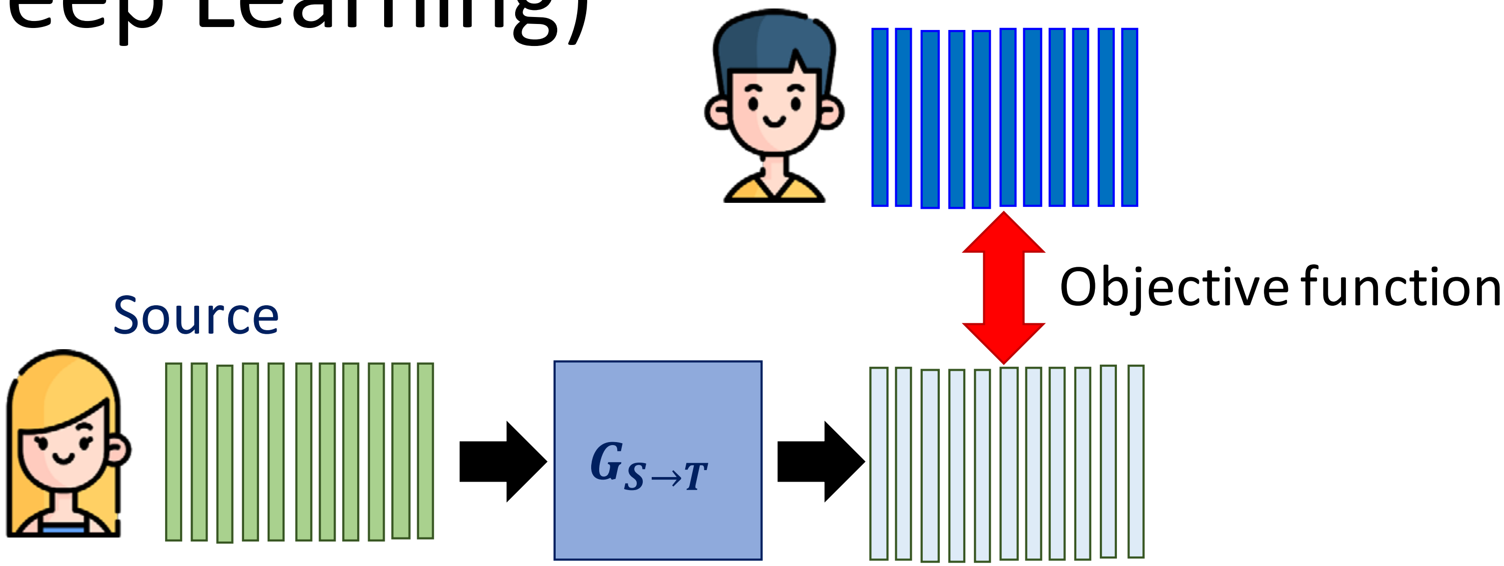
(b) Estimating weights to reconstruct source speech

$$\{\mathbf{A}_S^{(n)}, w^{(n)}\}_{n=1\ldots N} = argmin \sum_{n=1}^{N} \|\hat{x}_S^{(n)} - \mathbf{A}_S^{(n)} w^{(n)}\|^2$$

(c) Apply the weights to the target exemplars

$$\hat{x}_T^{(n)} = \mathbf{A}_T^{(n)} w^{(n)}, n = 1, 2, \ldots N$$

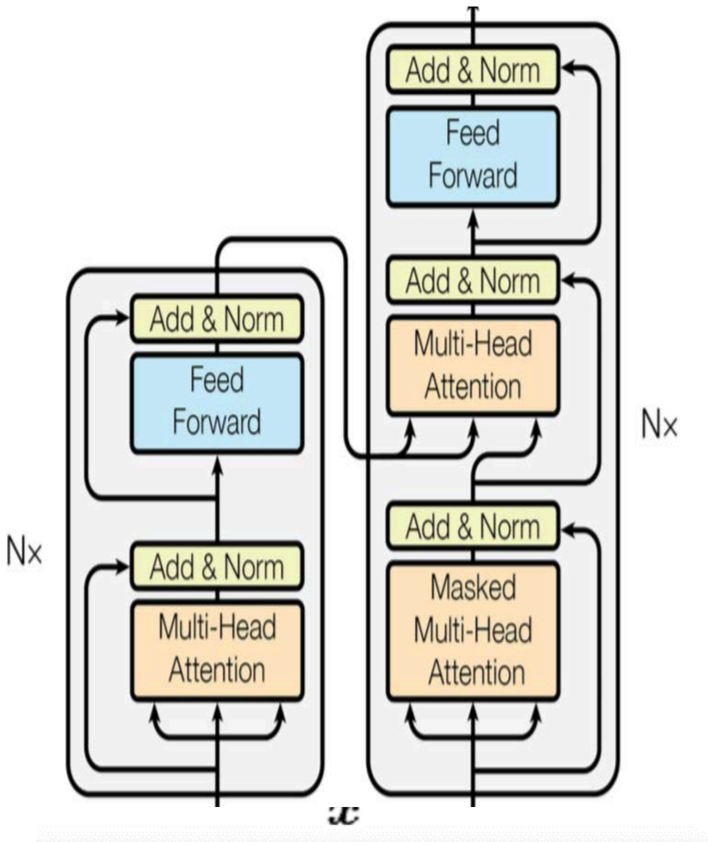
Parallel VC (Traditional Approaches vs Deep Learning)

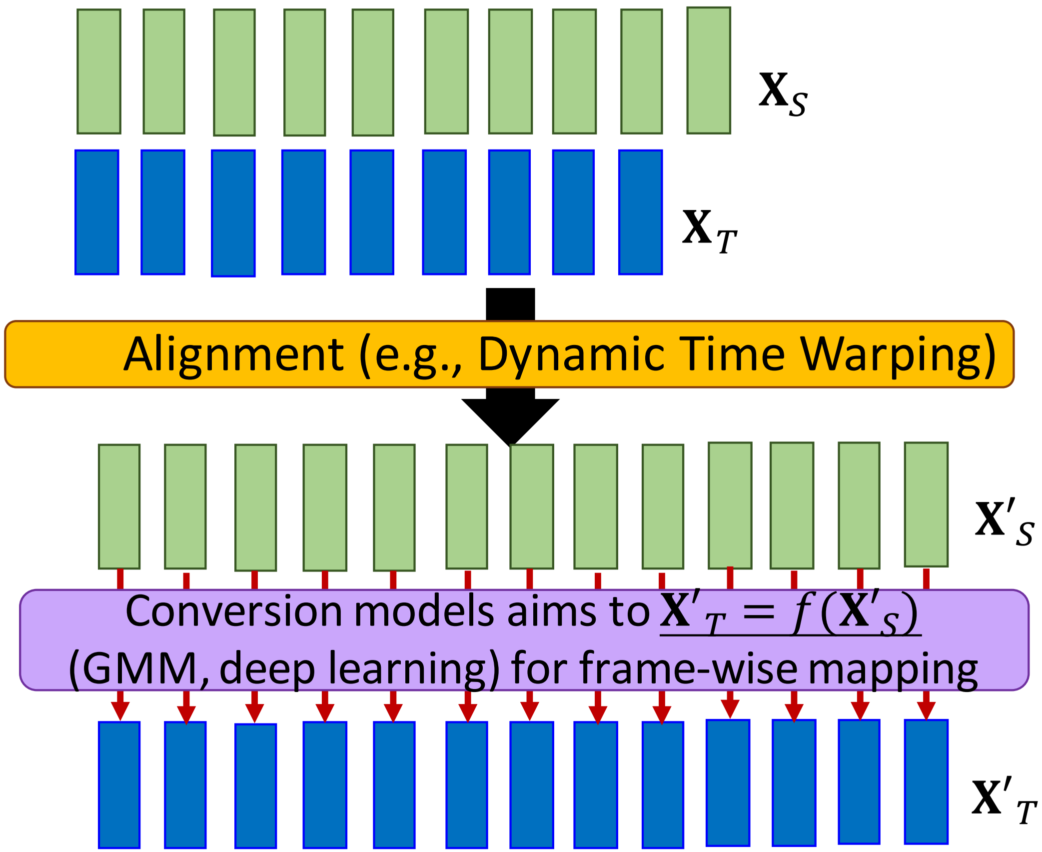


- ➢Gaussian mixture model (GMM) [Toda et al., TASLP 2007], non-negative matrix factorization (NMF) [Wu et al., TASLP 2014; Fu et al., TBME 2017], locally linear embedding (LLE) [Wu et al., Interspeech 2016].
- ➢Deep neural network models: restricted Boltzmann machine (RBM) [Chen et al., TASLP 2014], feed forward NN [Desai et al., TASLP 2010], recurrent NN (RNN) [Nakashika et al., Interspeech 2014], Transformer [Huang et al., Interspeech 2020],
- ➢Objective function: MMSE [Kain and Macon ICASSP 1998], maximum Likelihood parameter generation (MLPG) [Zen et al., TASLP 2011], minimum generation error (MGE) [Wu and Wang, ICASSP 2006], sequence error minimization (SEM) [Xie et al., Interspeech 2014].

Parallel VC (Deep Learning Era)

- RBM [Chen et al., TASLP 2014]
- FFNN [Desai et al., TASLP 2010]
- Transformer [Huang et al., Interspeech 2020]

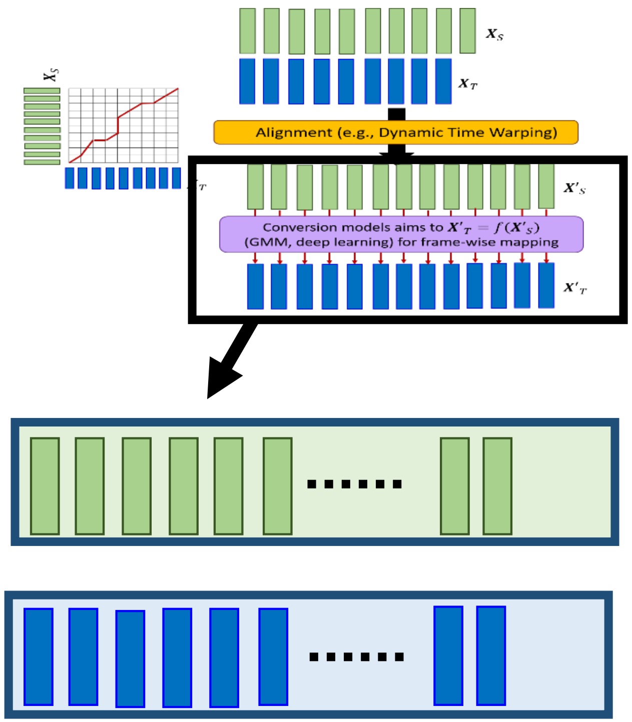




- The encoder and decoder are Transformer blocks [Vaswani et al., NIPS 2017]

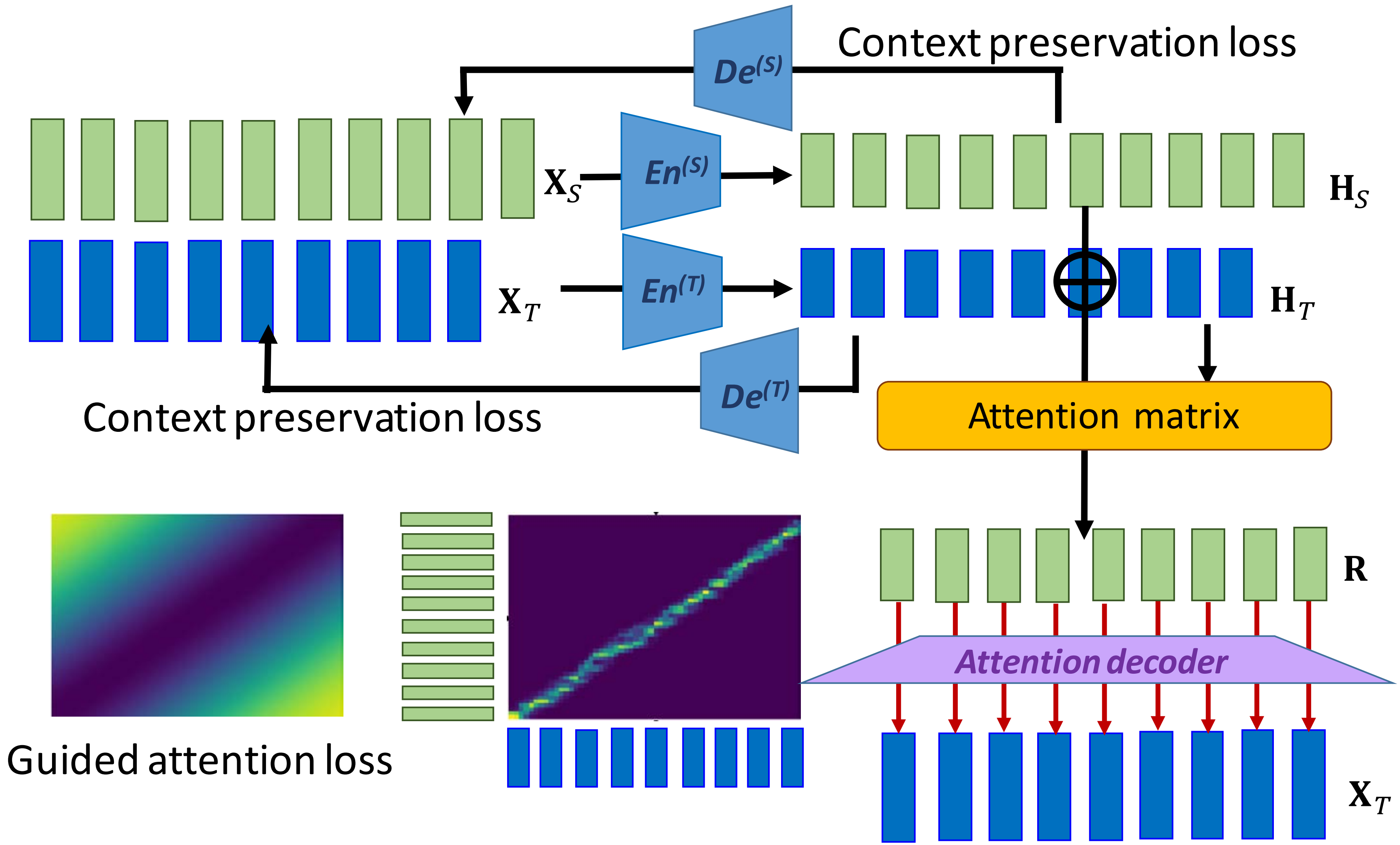
Parallel VC (Traditional Approaches)

- GMM-VC [Toda. et al, IEEE TASLP 2007]



Parallel VC (Deep Learning Era)

- ATTS2S-VC [Tanaka et al., ICASSP 2019]



Beyond Parallel Data for Voice Conversion

- Non-parallel data of paired speakers
- Disentanglement
- Leveraging TTS systems
- Leveraging ASR systems

Beyond Parallel Data

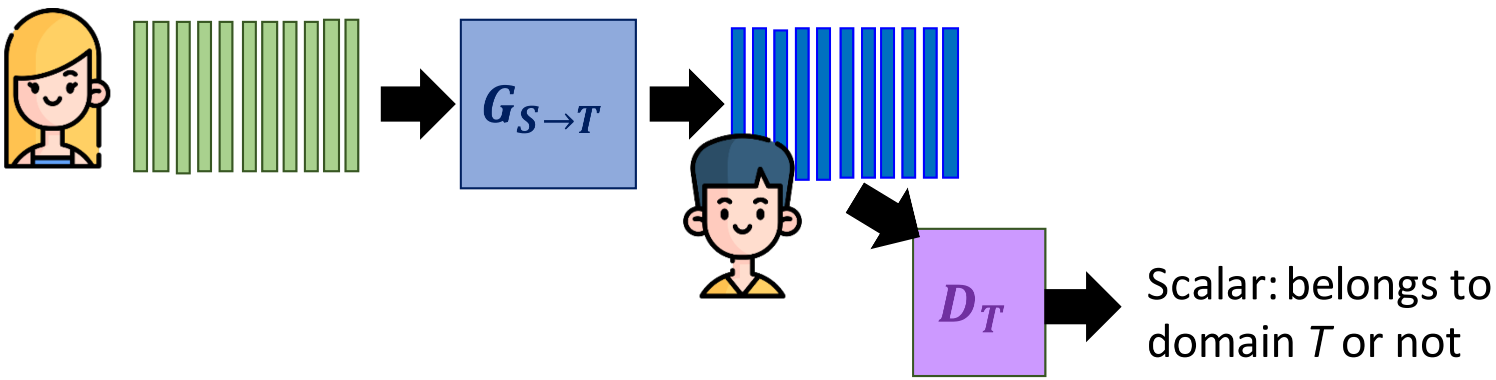
Non-parallel data of paired speakers

- CycleGAN-VC [Kaneko et al., Eusipco 2018]
- StarGAN-VC [Kameoka et al., SLT 2018]

Beyond Parallel Data

Non-parallel data of paired speakers

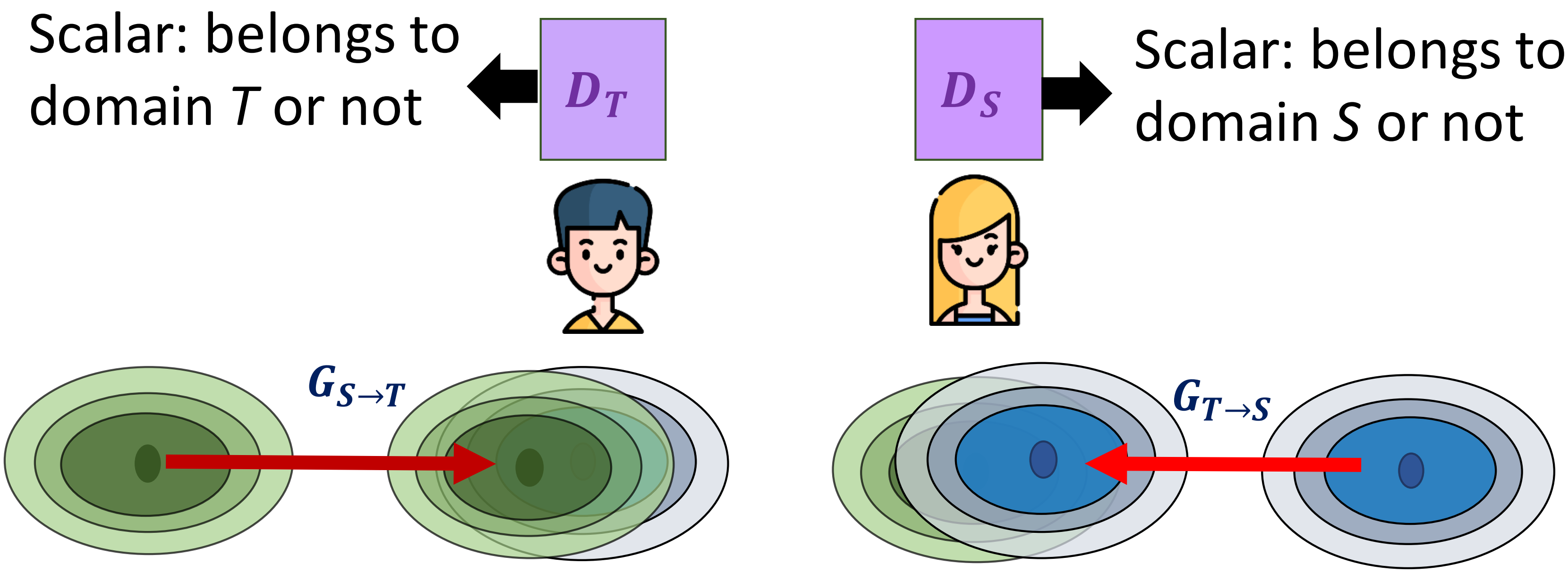
- CycleGAN-VC [Kaneko et al., Eusipco 2018]



Beyond Parallel Data

Non-parallel data of paired speakers

- CycleGAN-VC [Kaneko et al., Eusipco 2018]

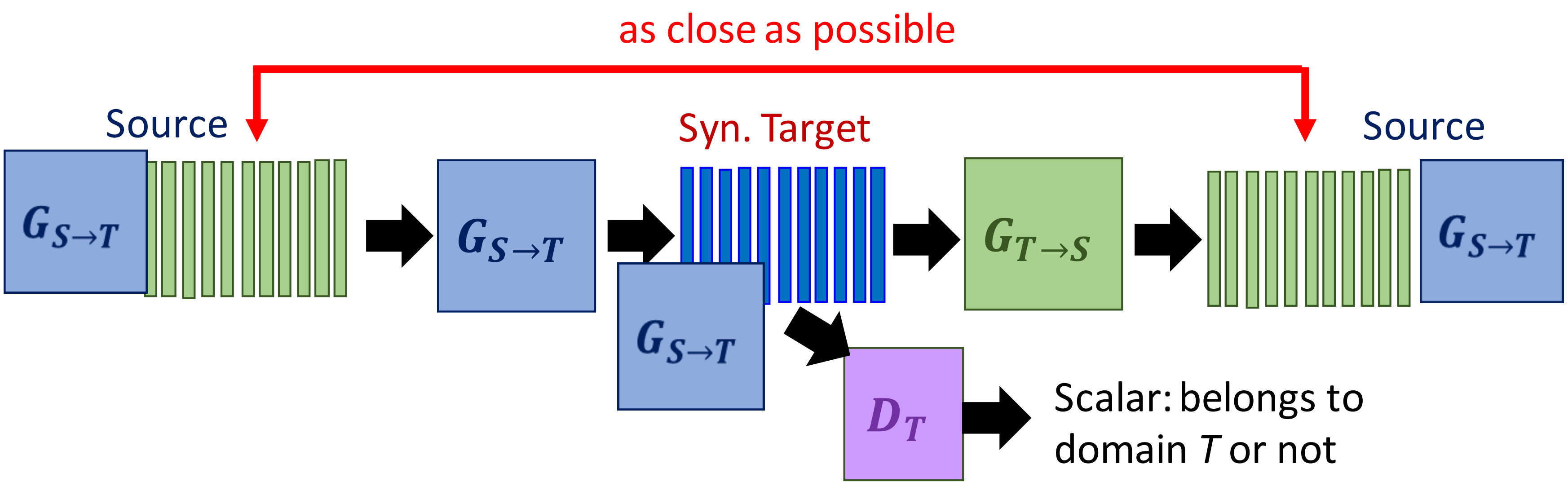


- Trained using many utterances from speaker *S*/*T*
- Speech contents are averaged out
- Cares more about speaker identity (ID)

Beyond Parallel Data

Non-parallel data of paired speakers

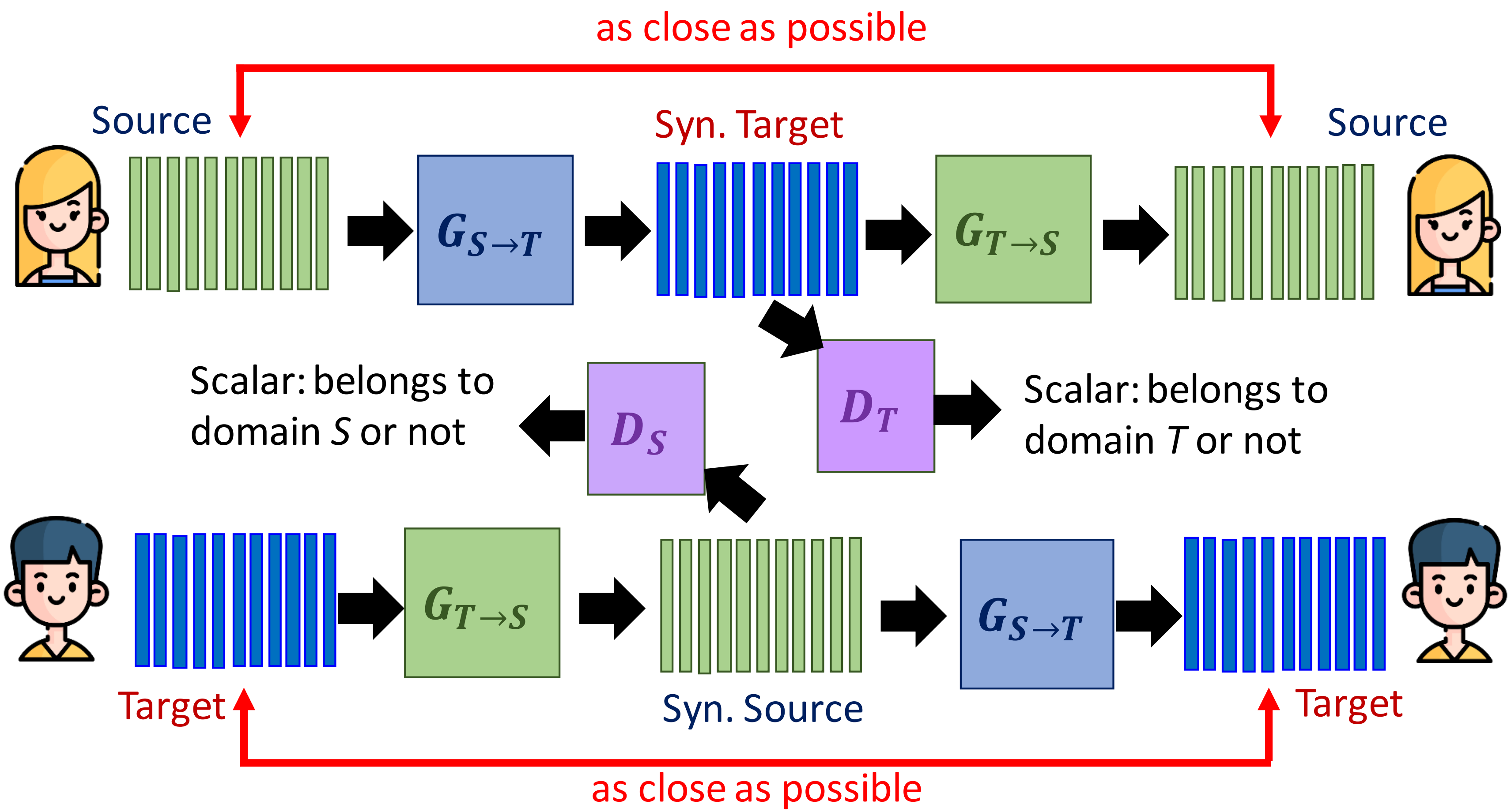
- CycleGAN-VC [Kaneko et al., Eusipco 2018]



Beyond Parallel Data

Non-parallel data of paired speakers

- CycleGAN-VC [Kaneko et al., Eusipco 2018]

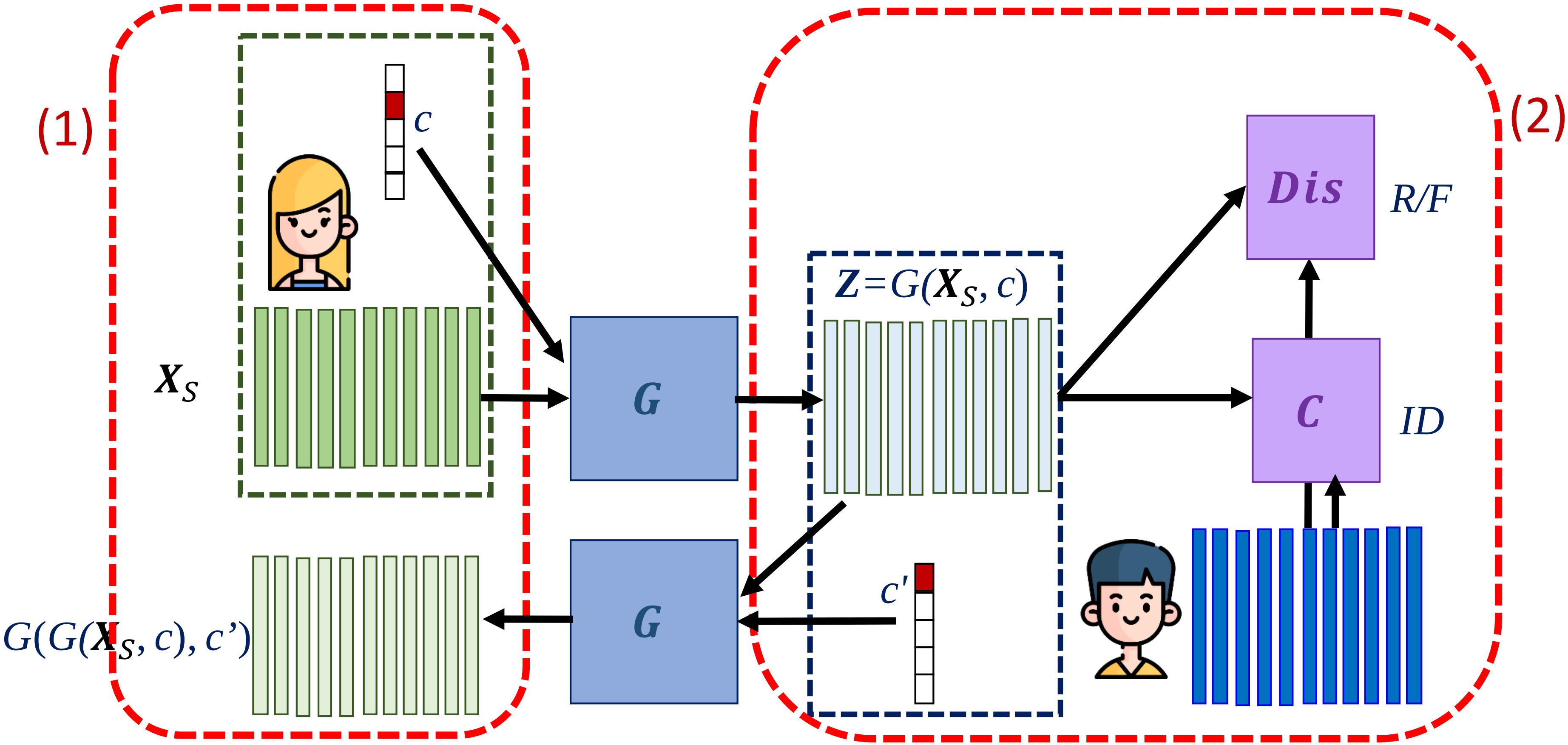


Cycle consistency loss: maintaining content information
Discriminative loss: changing speaker information

Beyond Parallel Data

Non-parallel data of paired speakers

- StarGAN-VC [Kamaoka et al., SLT 2018]



(1) Maintaining content information
(2) Changing speaker information

Beyond Parallel Data

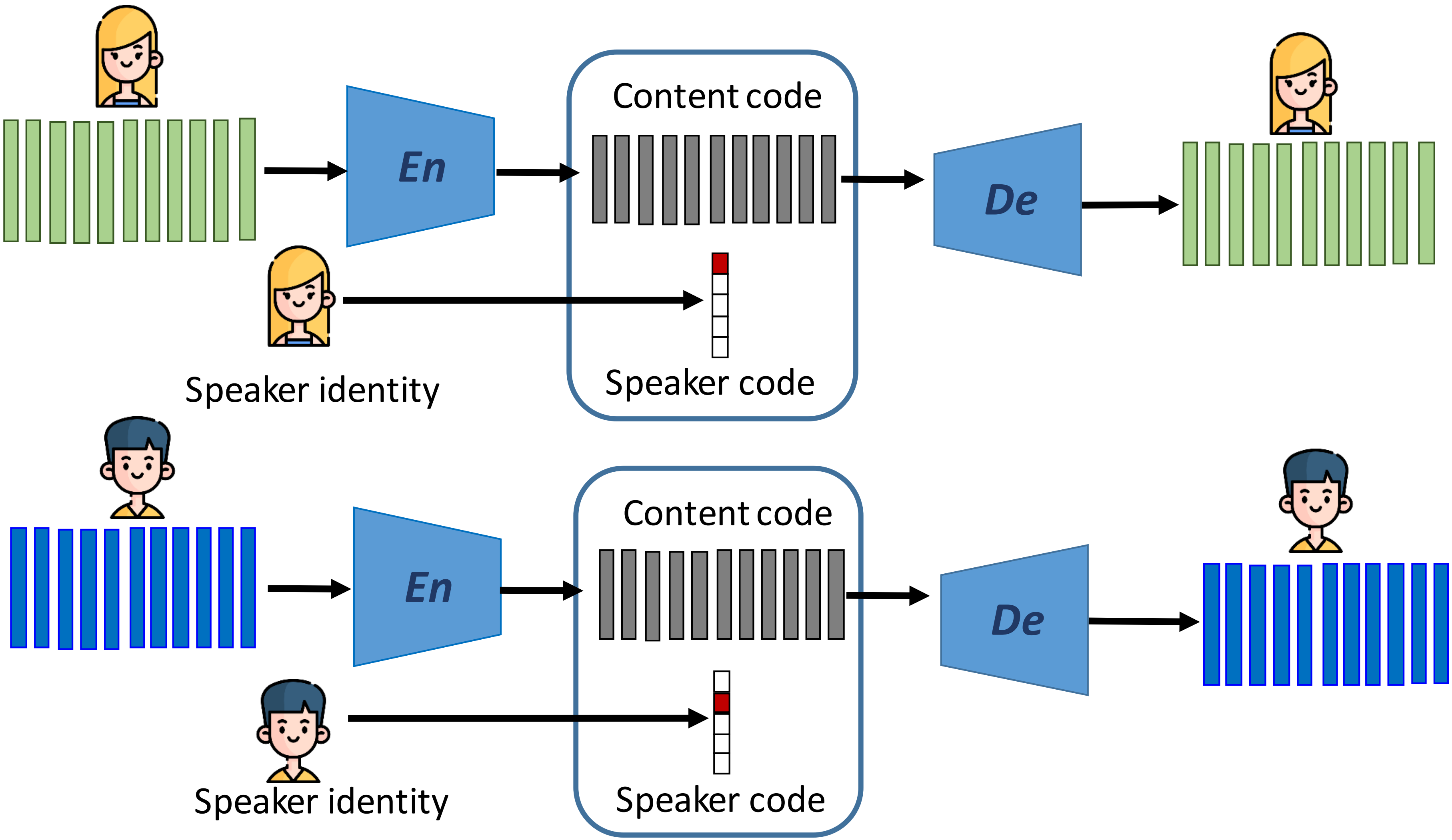
Disentanglement

- VAE-VC [Hsu et al., APSIPA 2016, Hsu et al., NIPS 2017, Oord et al., NIPS 2017]
- VAW-GAN-VC [Hsu et al., Interspeech 2017]
- MOEVC [Chang et al., ISCSLP 2021]
- CDVAE-VC [Huang et al., IEEE TETCI 2020]
- Multi-target VC [Chou et al., Interspeech 2018]
- IN-VC [Chou et al., Interspeech 2019]

Beyond Parallel Data

Disentanglement

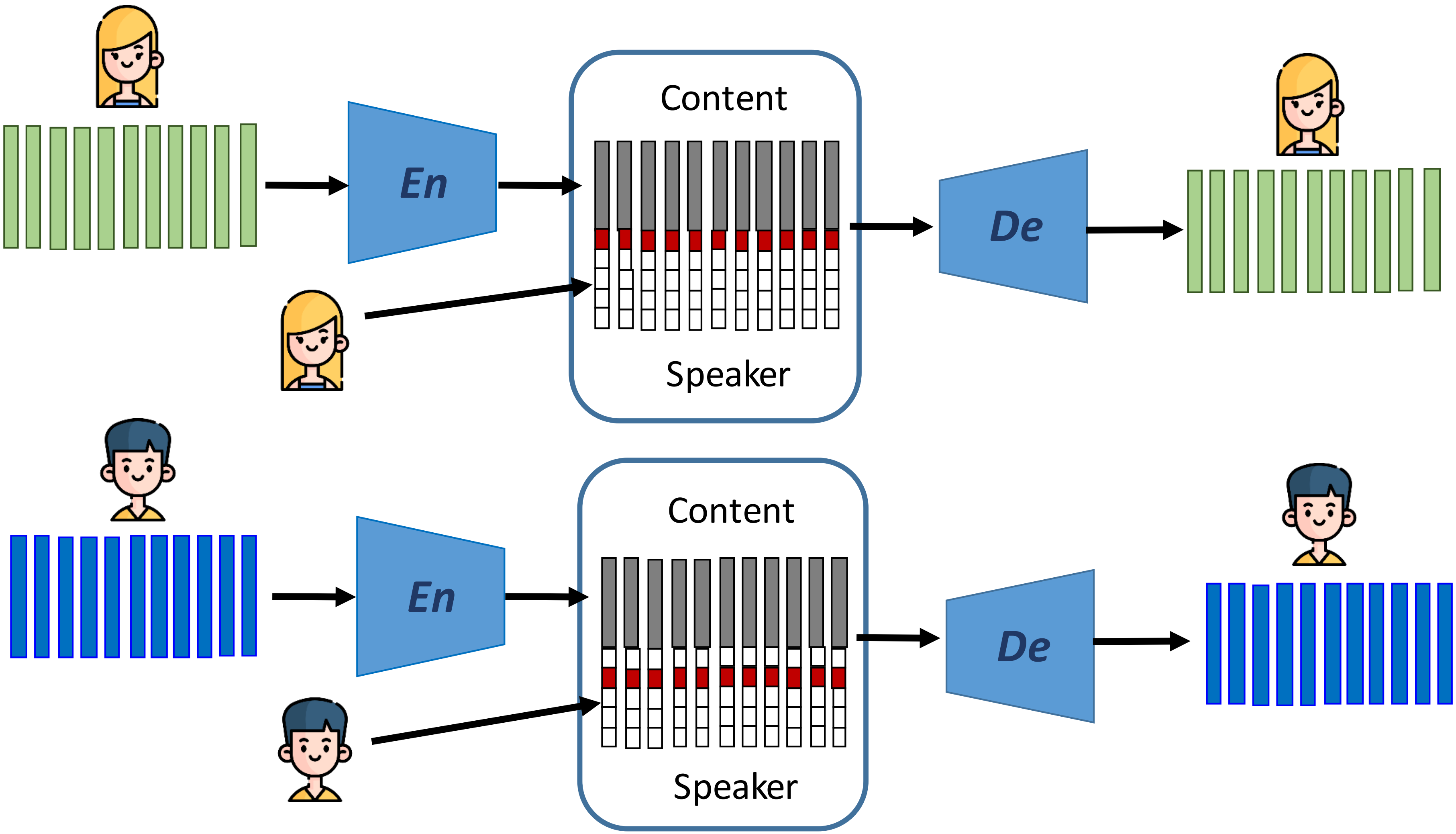
- VAE-VC [Hsu et al., APSIPA 2016, Hsu et al., NIPs 2017], AUTOVC [Qian et al., ICML 2019]



Beyond Parallel Data

Disentanglement

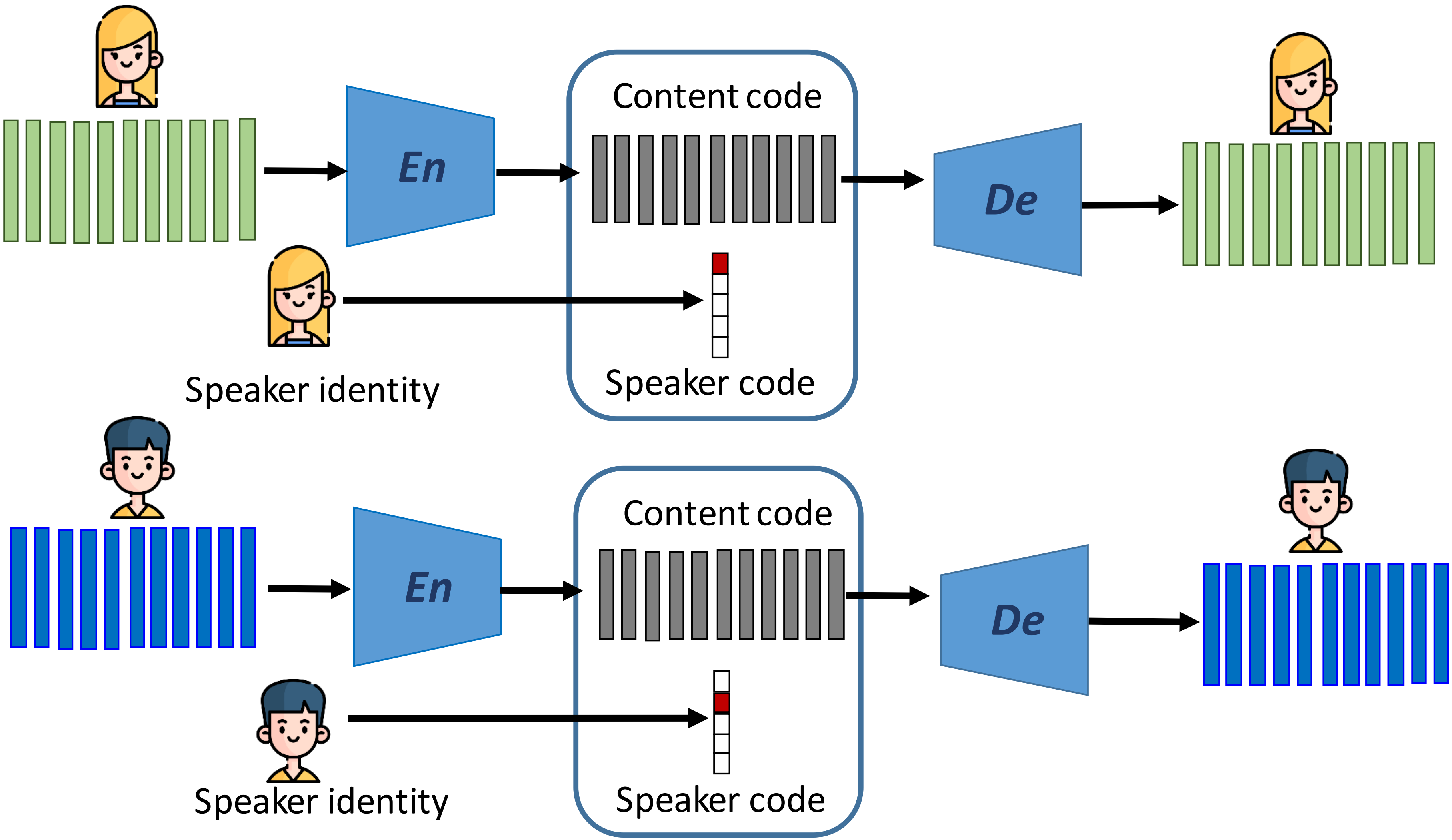
- VAE-VC [Hsu et al., APSIPA 2016, Hsu et al., NIPs 2017], AUTOVC [Qian et al., ICML 2019]



Beyond Parallel Data

Disentanglement

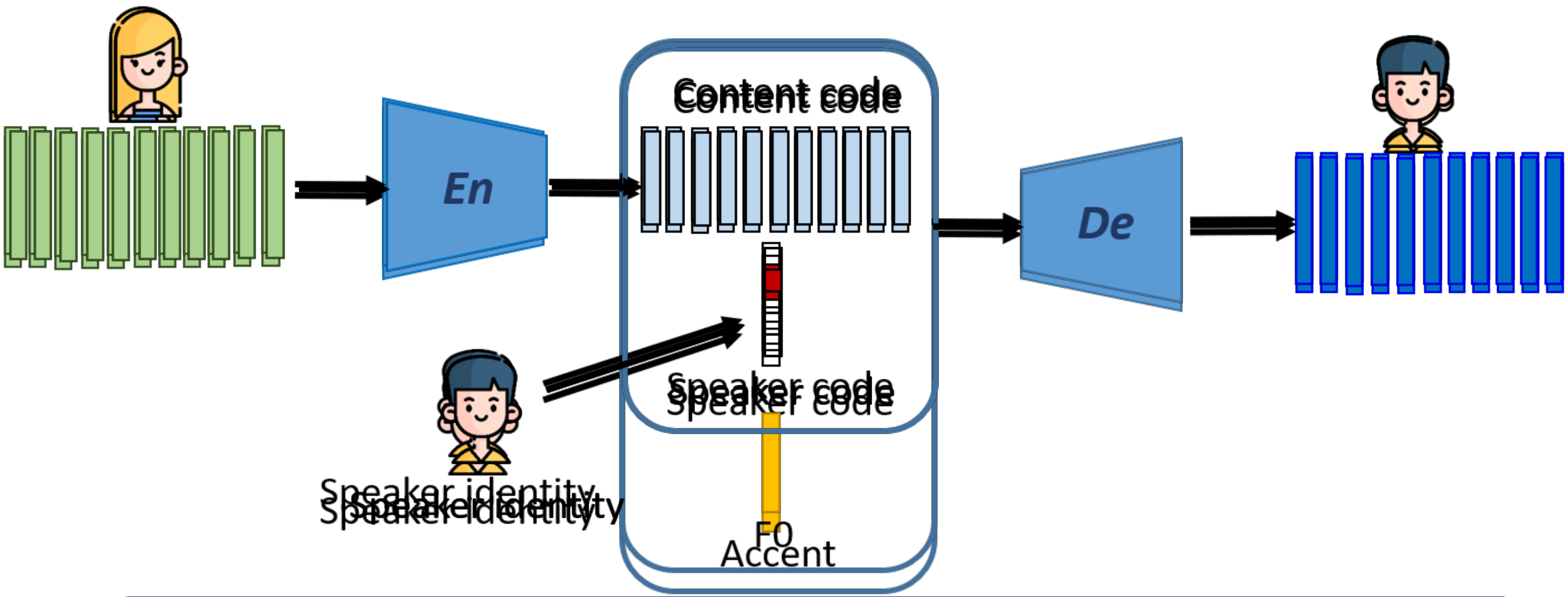
- VAE-VC [Hsu et al., APSIPA 2016, Hsu et al., NIPs 2017], AUTOVC [Qian et al., ICML 2019]



Beyond Parallel Data

Disentanglement

- [illegible] [Qian et al., ICML 2019] [Qian et al., ICASSP 2020], Emotional VC [Zhou et al., Interspeech 2020]

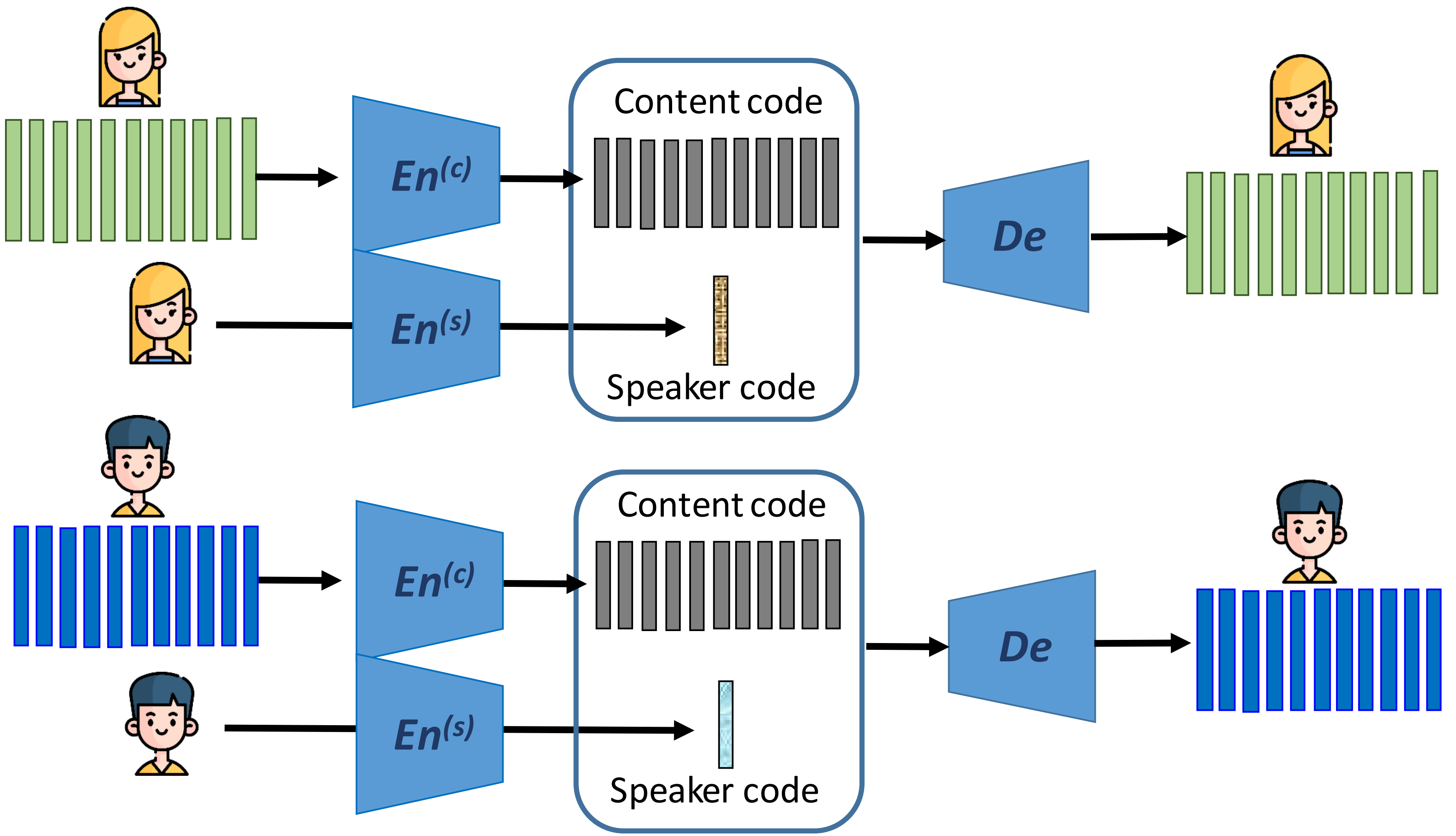


(1) Extracting content information (removing speaker info.)
(2) Adding speaker info. on the content

Beyond Parallel Data

Disentanglement

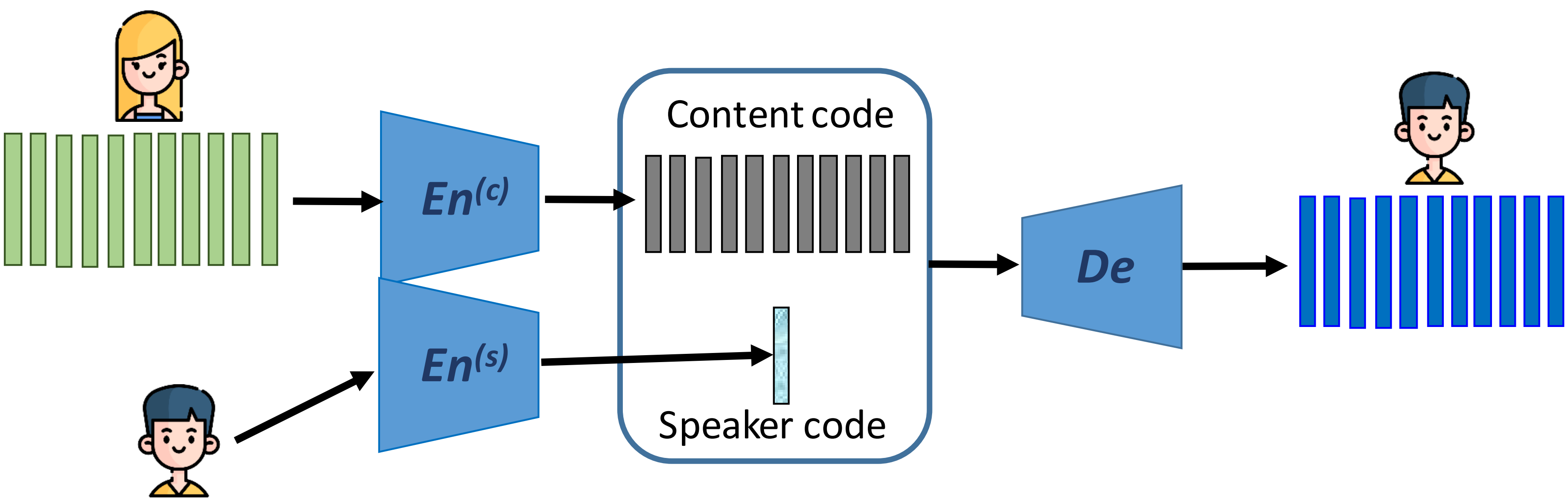
- VAW-GAN VC [Hsu et al., Interspeech 2017]



Beyond Parallel Data

Disentanglement

- VAW-GAN VC [Hsu et al., Interspeech 2017]

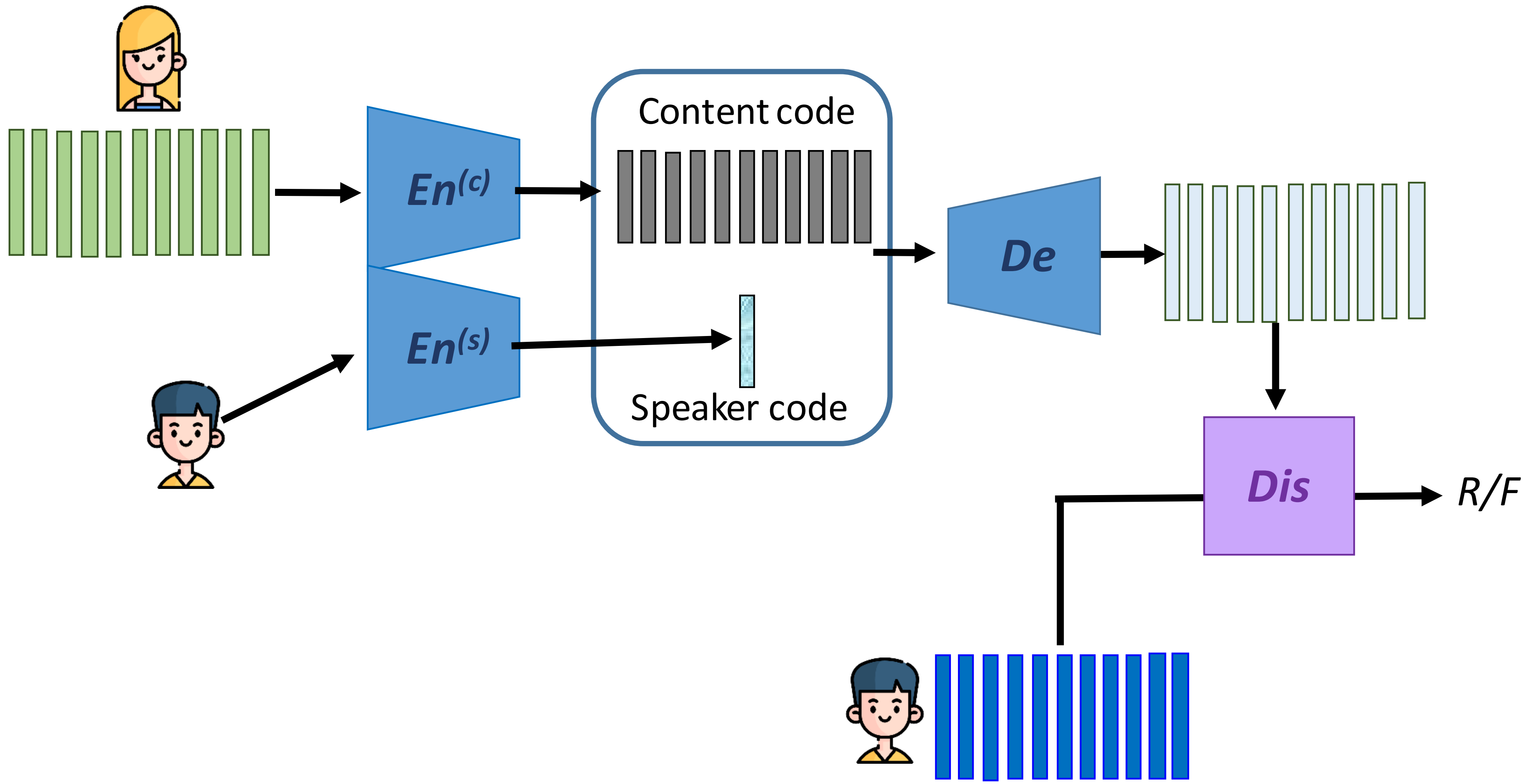


(1) Extracting content information (removing speaker info.)
(2) Adding speaker info. on the content

Beyond Parallel Data

Disentanglement

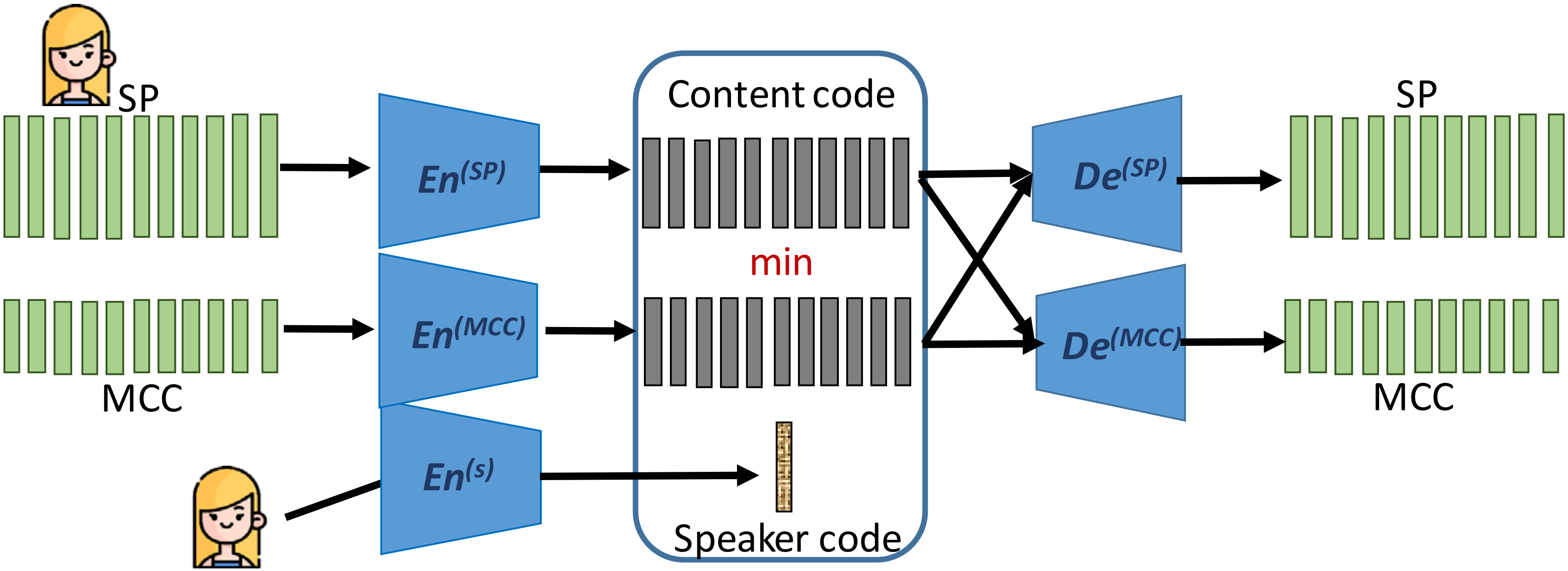
- VAW-GAN VC [Hsu et al., Interspeech 2017]



Beyond Parallel Data

Disentanglement

- CDVAE-VC [Huang et al., IEEE TETCI 2020]

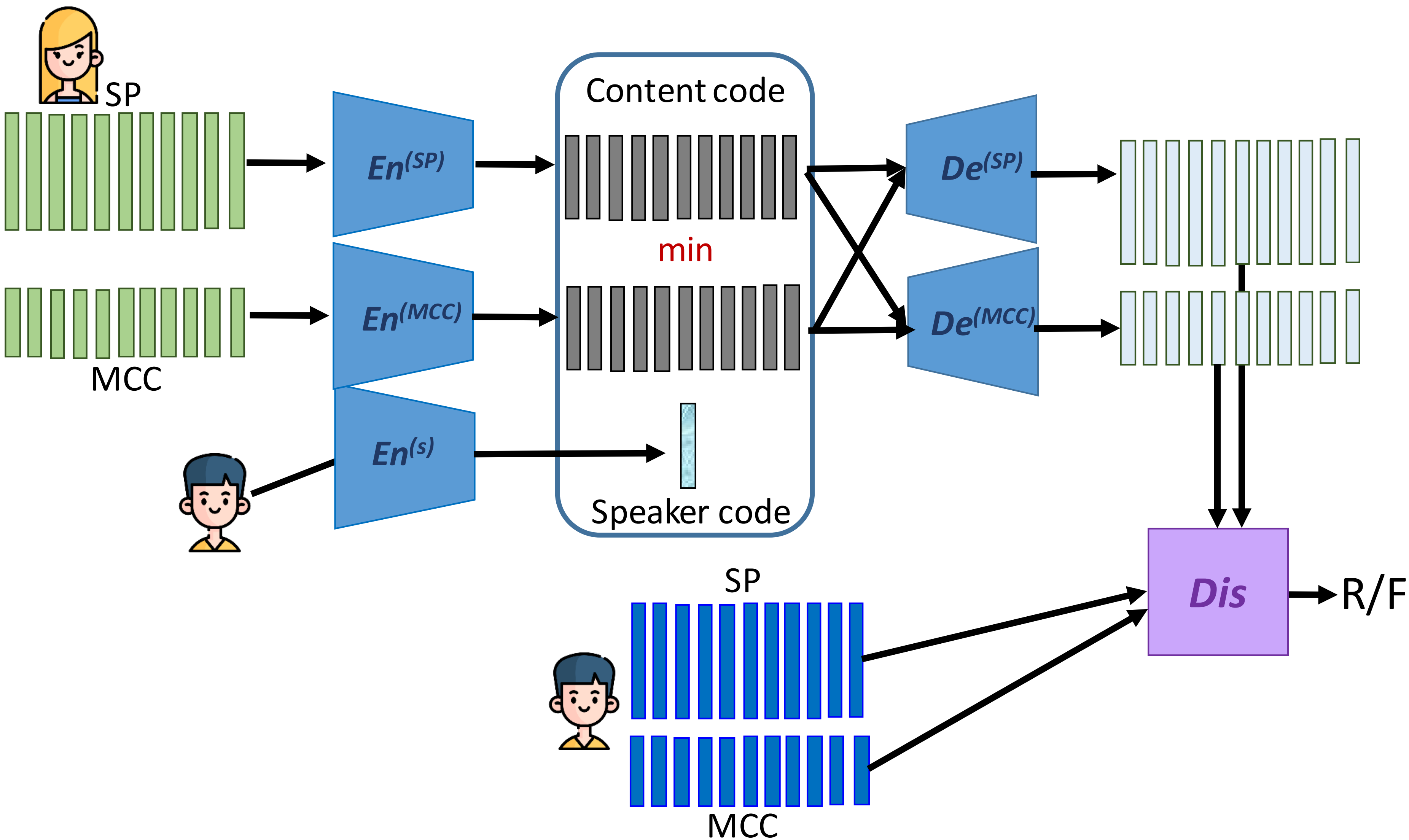


- Multi-task learning with low- and high-resolution features.
- SP (spectra): high-resolution spectral feature; MCC (Mel-cepstral-coefficients): low-resolution & designed based on human perception.
- Better disentanglement (content and speaker), and quality of converted speech by CDVAE.

Beyond Parallel Data

Disentanglement

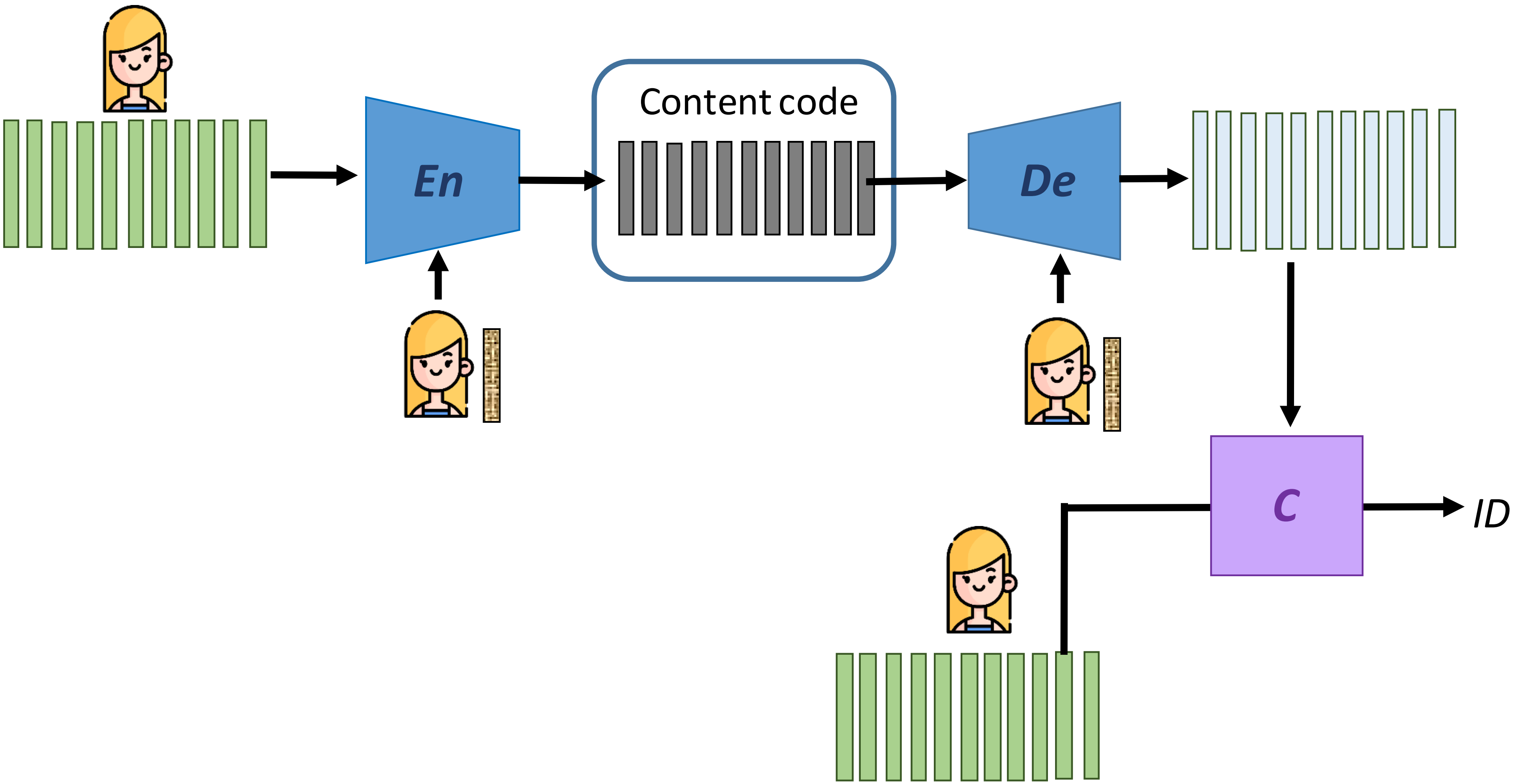
- CDVAE-VC [Huang et al., IEEE TETCI 2020]



Beyond Parallel Data

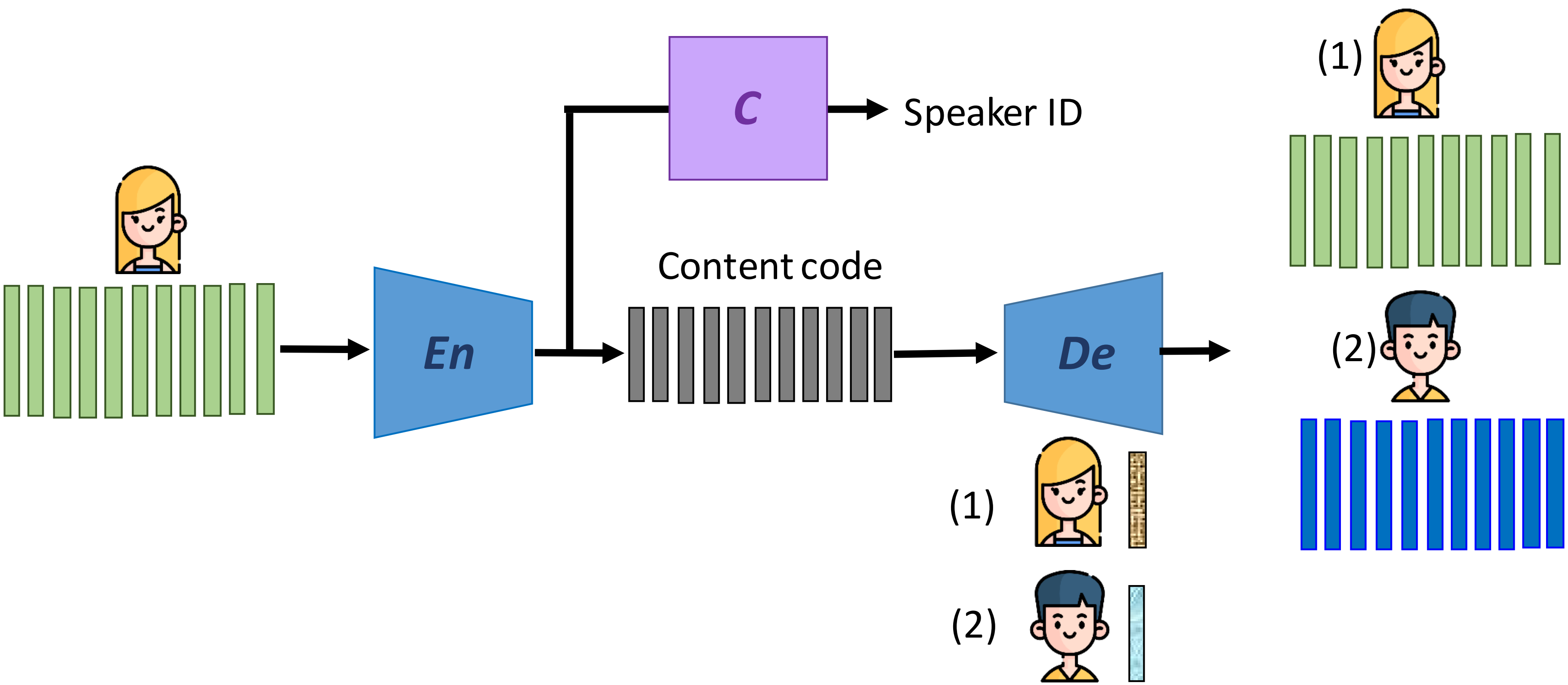
Disentanglement

- ACVAE-VC [Kameoka et al., IEEE TASLP 2019]
  - ➢ Using an auxiliary classifier



Beyond Parallel Data
Disentanglement

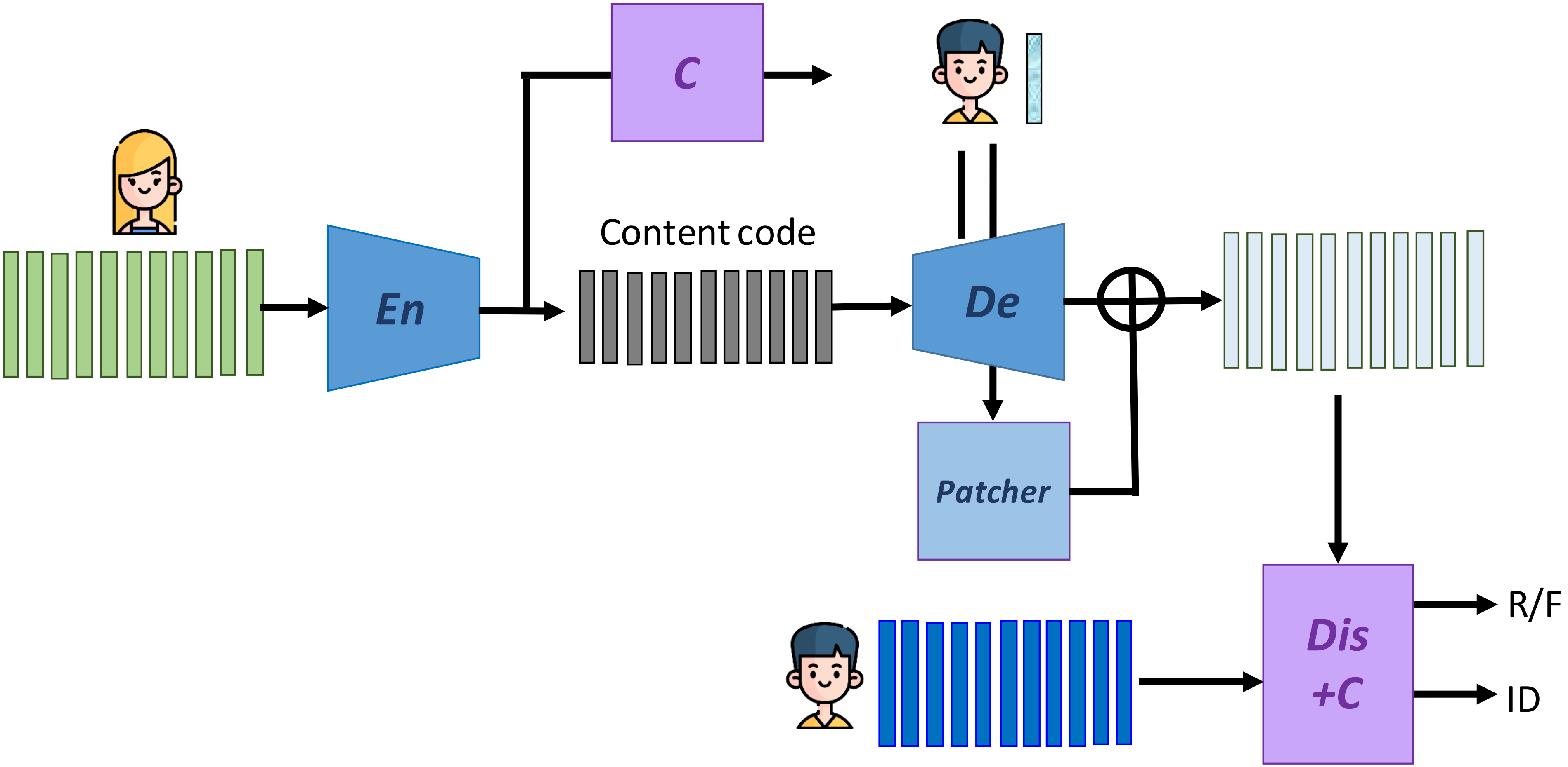
- Multi-target VC [Chou et al., Interspeech 2018]
  - ➢Training stage 1



Beyond Parallel Data

Disentanglement

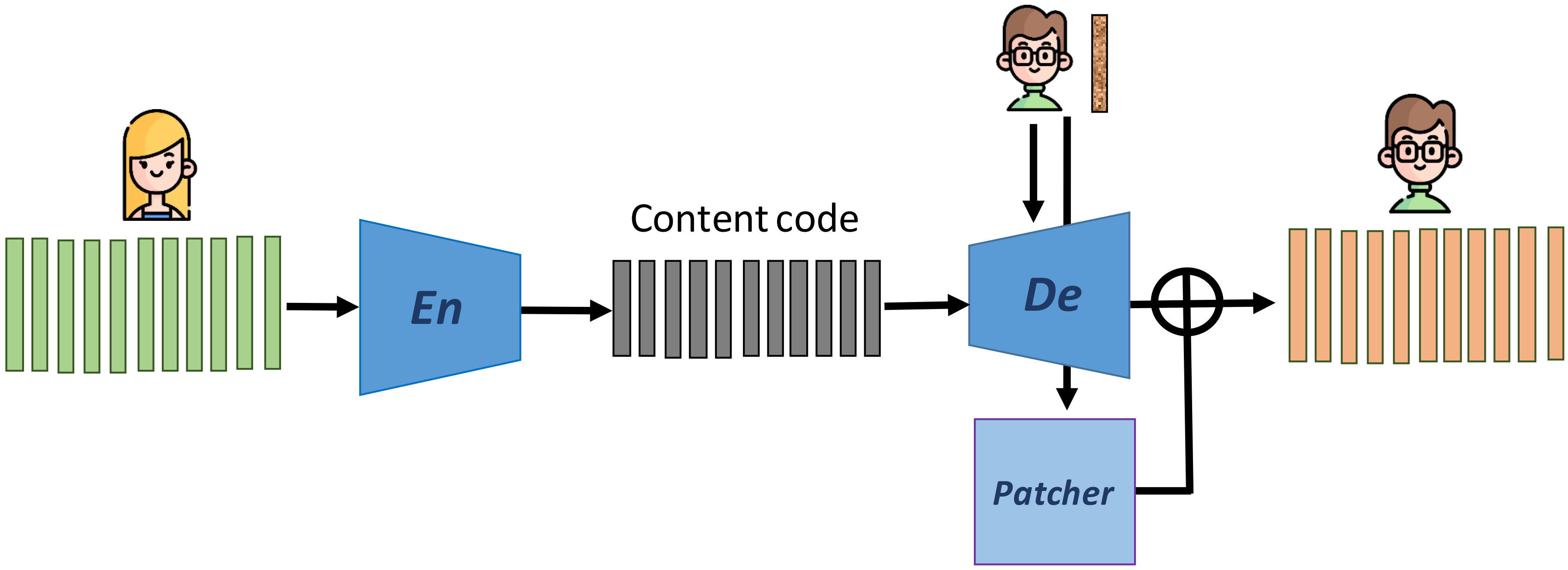
- Multi-target VC [Chou et al., Interspeech 2018]
  - ➢Training stage 2 (similar to ACVAE and VAW-GAN)



Beyond Parallel Data

Disentanglement

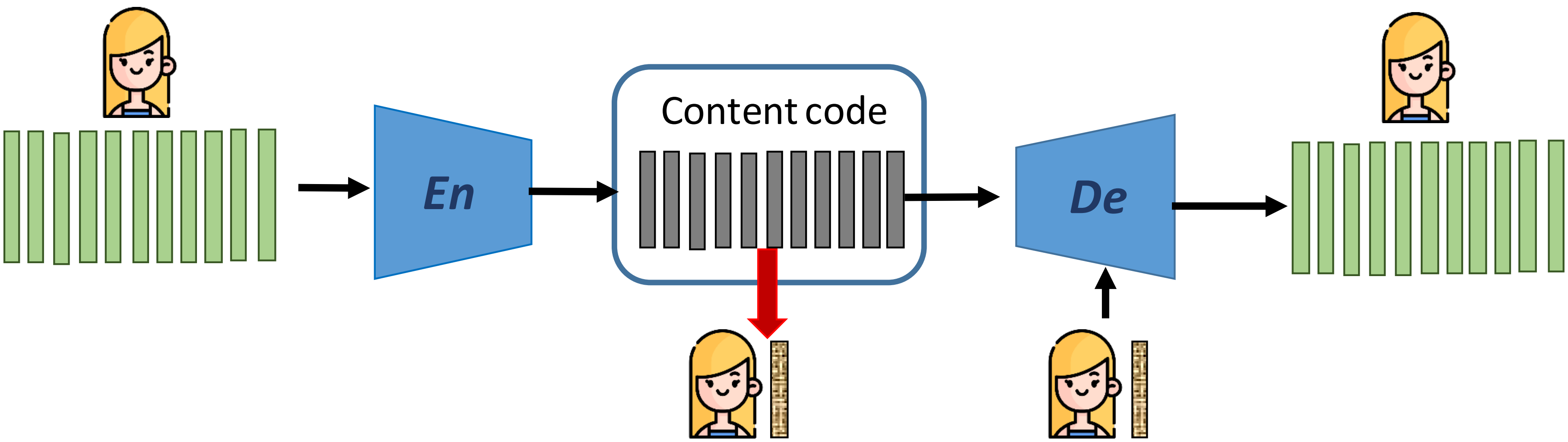
- Multi-target VC [Chou et al., Interspeech 2018]
  - ➢Conversion stage



Beyond Parallel Data

Disentanglement

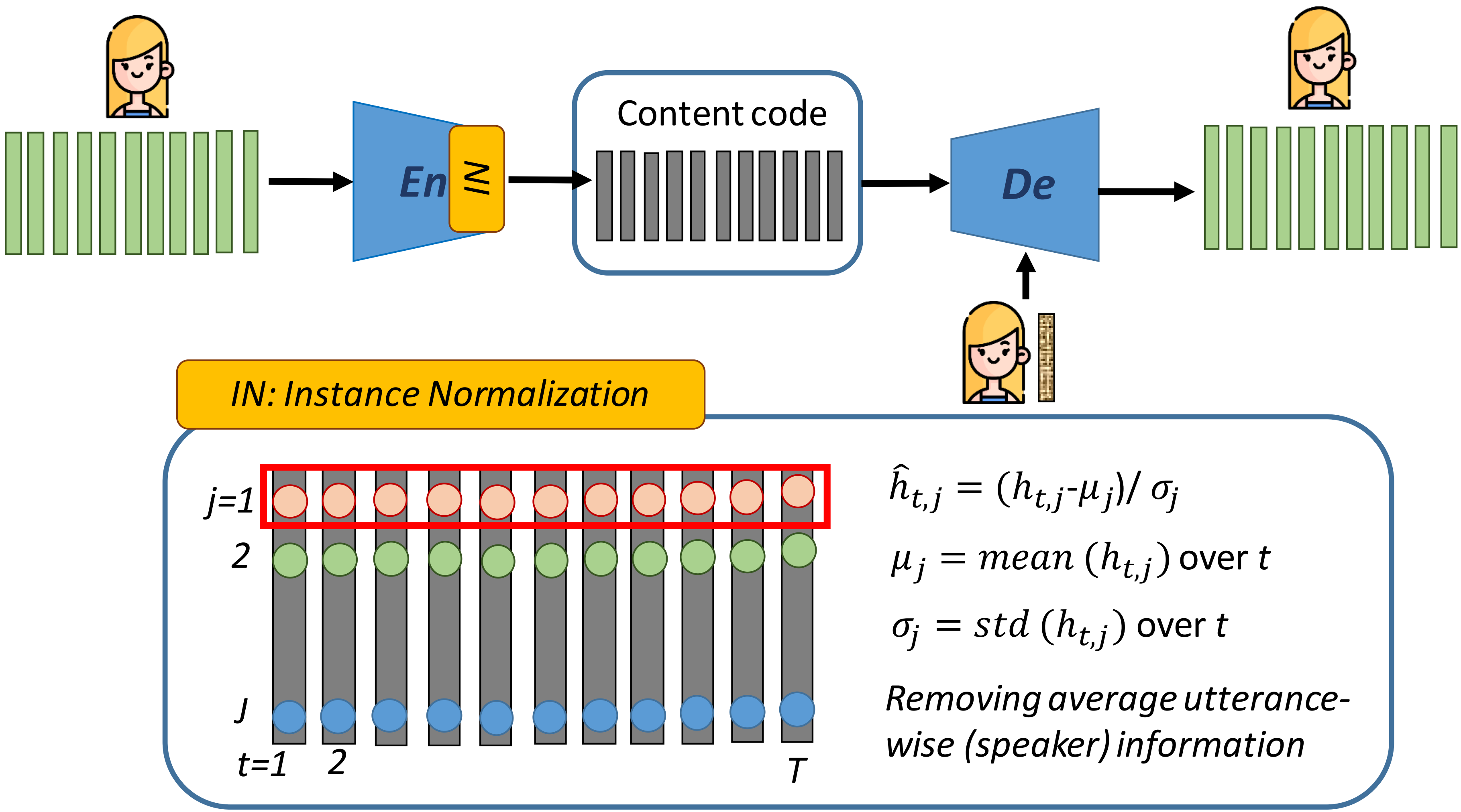
- IN [Chou et al., Interspeech 2019, Patel et al., APSIPA 2019]
  - Removing speaker information



Beyond Parallel Data

Disentanglement

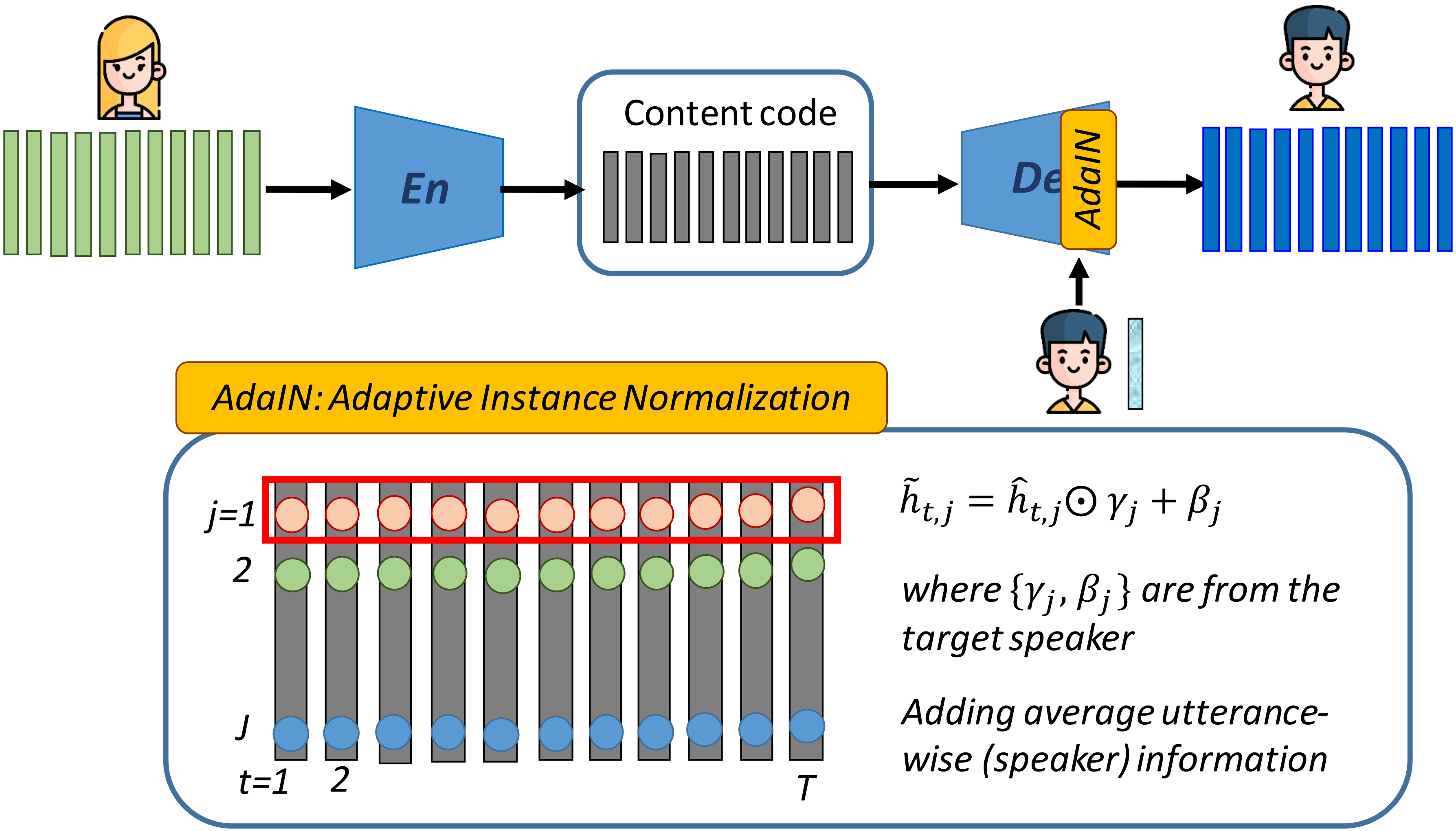
- IN [Chou et al., Interspeech 2019, Patel et al., APSIPA 2019]
  - ➢ Removing speaker information



Beyond Parallel Data

Disentanglement

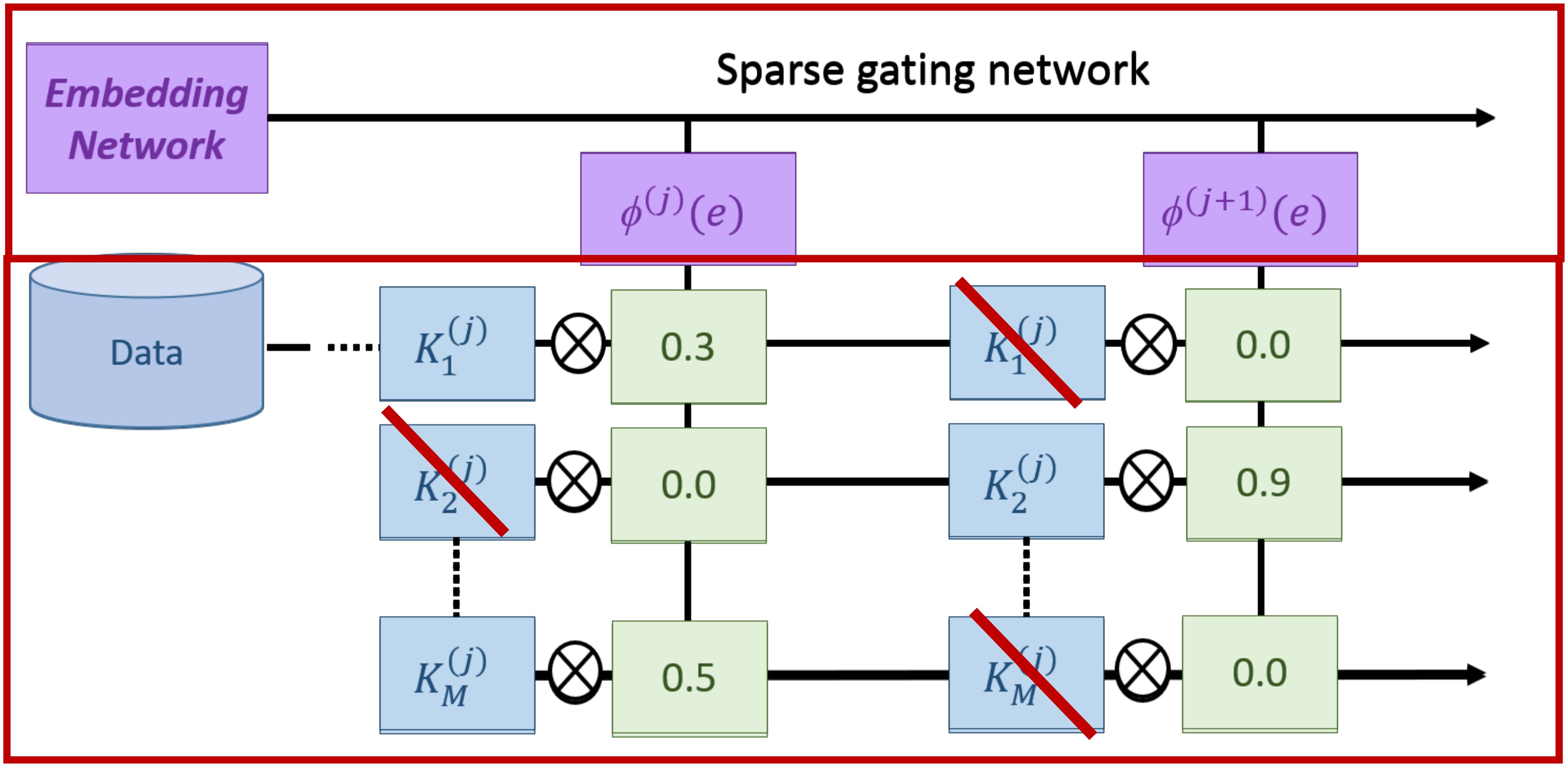
- IN [Chou et al., Interspeech 2019, Patel et al., APSIPA 2019]
    - Adding speaker information



Beyond Parallel Data

Disentanglement

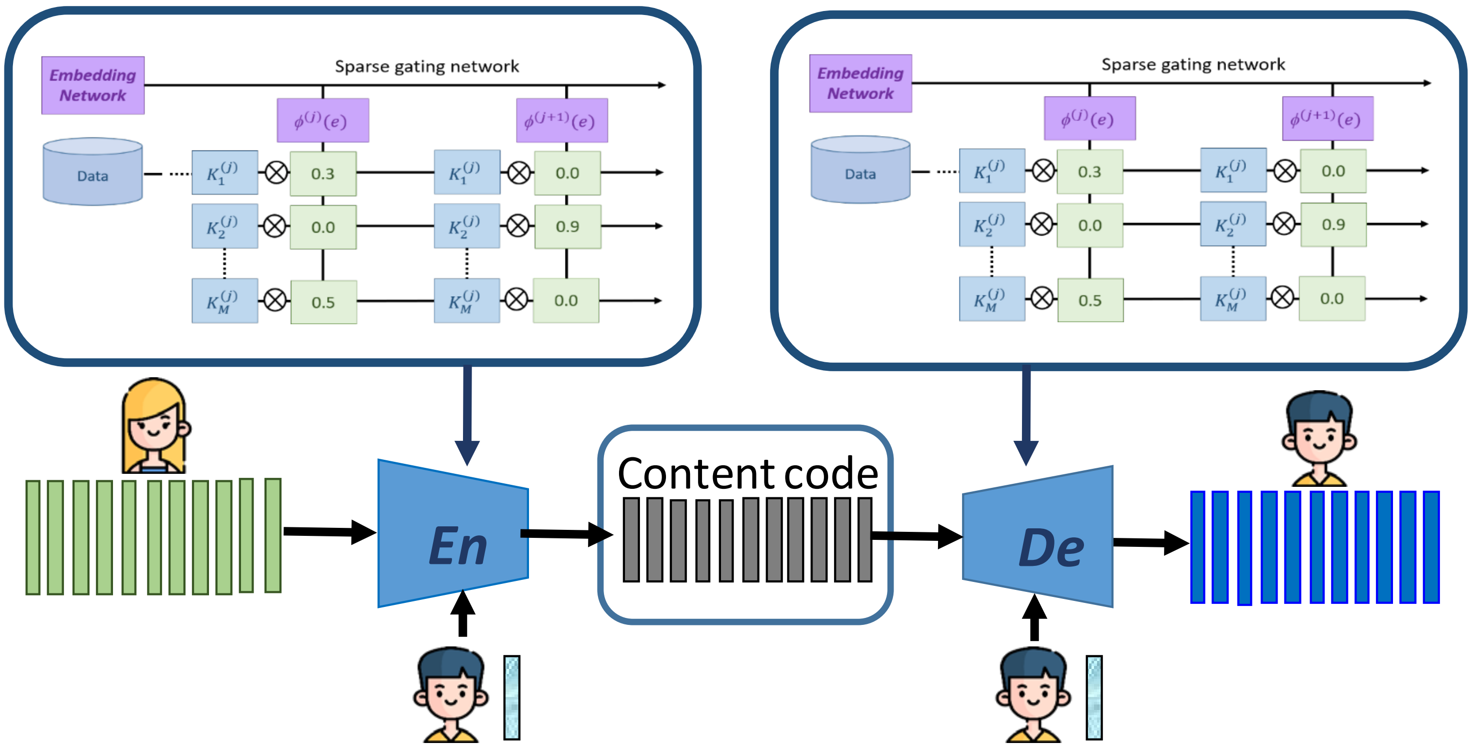
- MoE [Wang et al., UAI 2020]
  - ➢ Embedding network with sparse constraints



Beyond Parallel Data

Disentanglement

- MoE-VC [Chang et al., ISCSLP 2021]
  - ➢ VC online computation acceleration

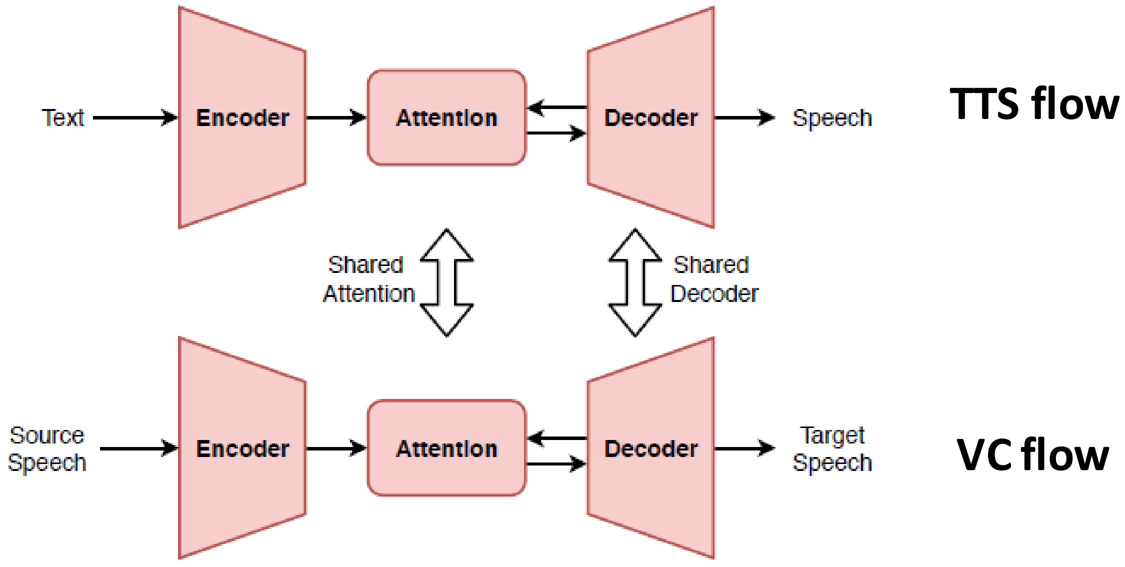


- MOE-VC can effectively reduce 70% FLOPs (namely floating point operations per second) with unnoticeable quality drop.
- A MOSNet is used to determine the optimal architecture.

Beyond Parallel Data
Leveraging TTS

The ideas to leverage TTS mechanism can be motivated in different ways.

- A TTS system is equipped with a quality attention mechanism that is needed by voice conversion, and
- A TTS system is trained on a large speech database that offers a high quality speech re-construction mechanism given the linguistic content.



Both follow similar encoder-decoder with attention architecture

Beyond Parallel Data

Leveraging TTS

Joint training of TTS & VC [Zhang et al., INTERSPEECH 2019]

- Both TTS and voice conversion can be benefited from each other.
- Both can be divided into two parts: an input encoder and an acoustic decoder.
- Even though various successful methods have been proposed for TTS and voice conversion, most of the systems can achieve only one task.

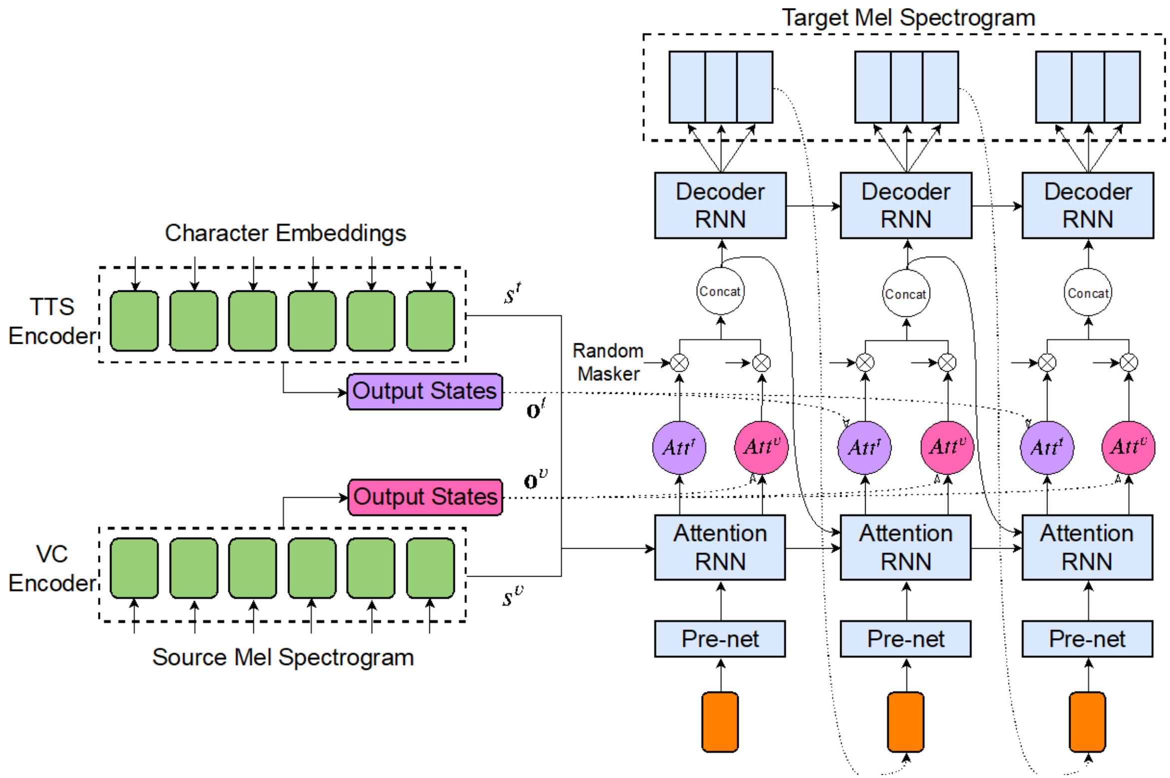
To construct one model shared for both TTS and voice conversion?

Beyond Parallel Data

Leveraging TTS

Joint training of TTS & VC [Zhang et al., INTERSPEECH 2019]

- An encoder-decoder model that supports multiple encoders.

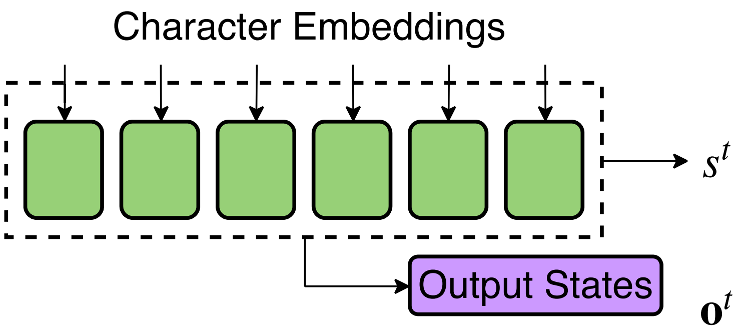


Improves the performance of VC compared with the stand-alone model!

Beyond Parallel Data

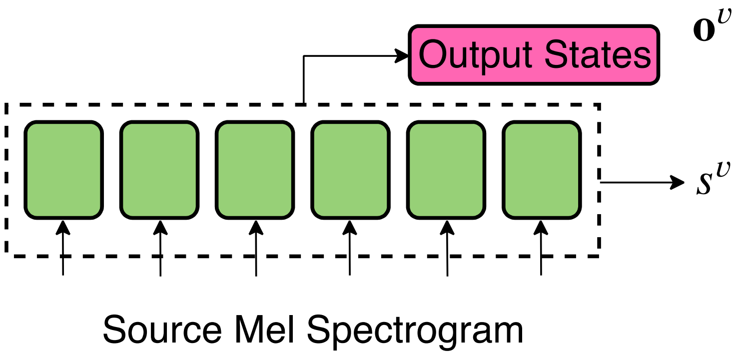
Leveraging TTS

Joint training of TTS & VC [Zhang et al., INTERSPEECH 2019]



TTS encoder

Given text characters as input, the model conducts end-to-end speech synthesis.



VC encoder

Given spectrogram, the model conducts seq-to-seq voice conversion.

Beyond Parallel Data

Leveraging TTS

Cotatron [Park et al., INTERSPEECH 2020]

- Transcription-guided speech encoder, based on multi-speaker TTS.
  - Encode an arbitrary speaker's speech into speaker-independent linguistic features, which are fed to a decoder for any-to-many VC.
- Cotatron VC is similar to PPG-based VC models.
  - Cotatron VC uses the TTS encoder to extract speaker-independent linguistic features, or disentangle the speaker identity.
  - The decoder then takes the attention aligned speaker-independent linguistic features as the input, and the target speaker identity as the condition, to generate a target speaker's voice.

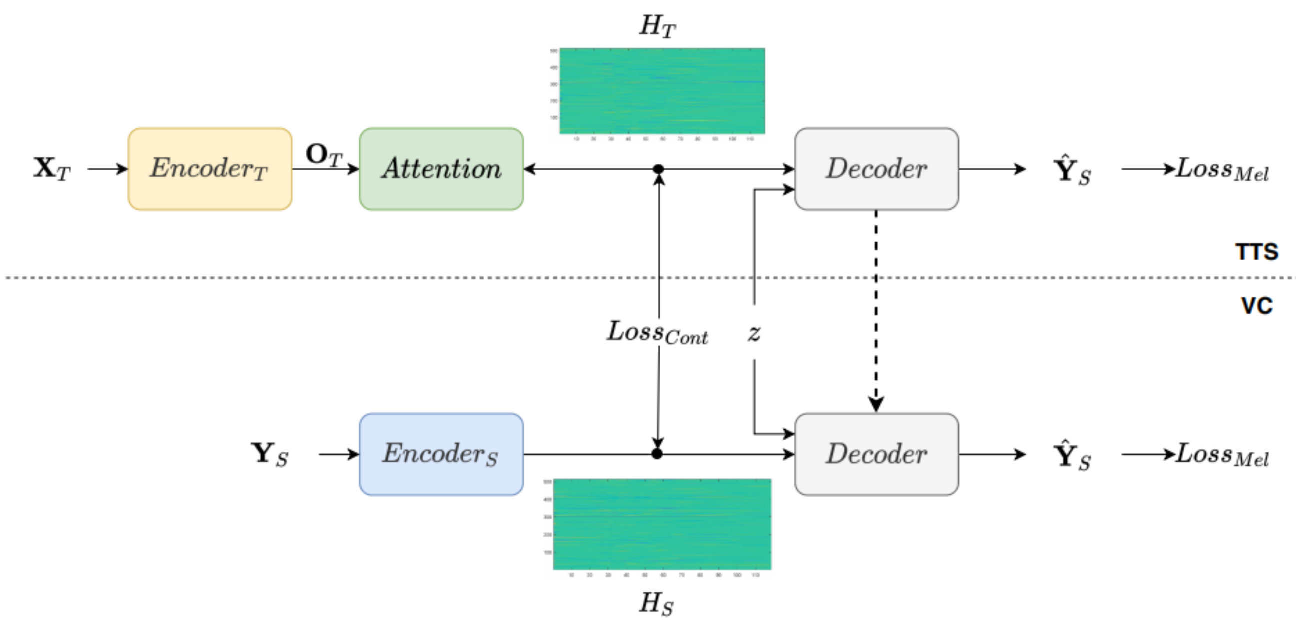
VC leverages the attention mechanism or shared attention from TTS.

Beyond Parallel Data

Leveraging TTS

TTS-VC transfer learning [Zhang et al., 2020]

- First train a multi-speaker Tactron-2 on a large database.
- Then transfer the TTS knowledge to an encoder-decoder architecture for voice conversion.

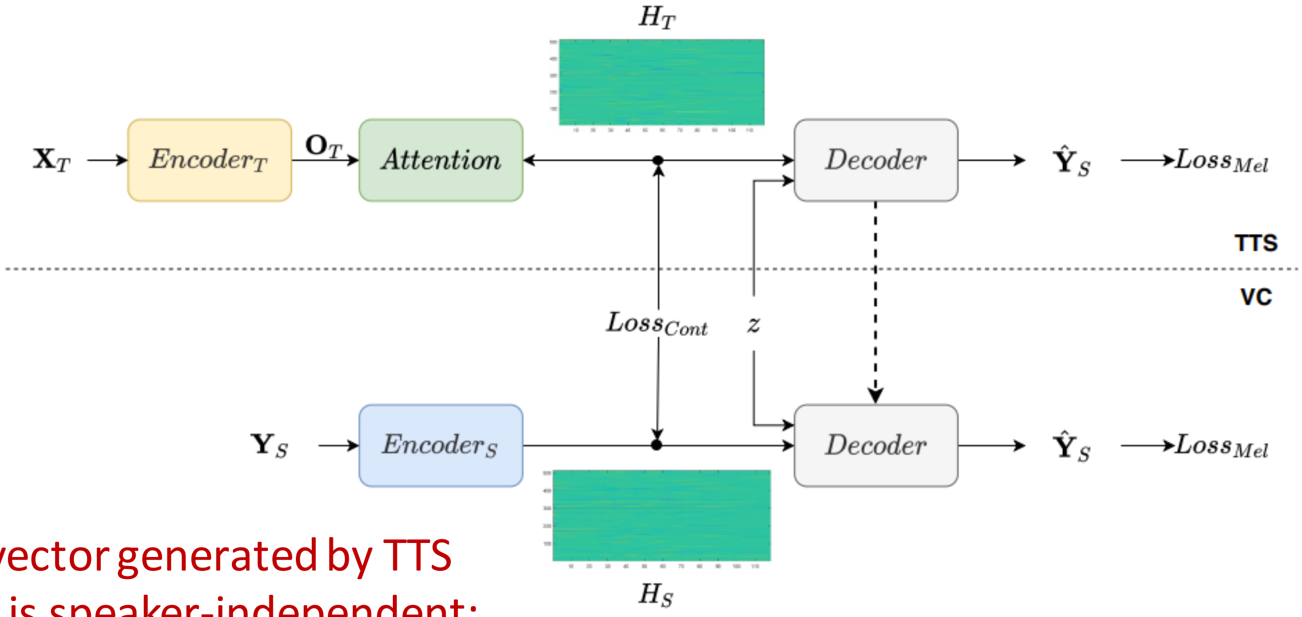


Beyond Parallel Data

Leveraging TTS

TTS-VC transfer learning [Zhang et al., 2020]

- First train a multi-speaker Tactron-2 on a large database.
- Then transfer the TTS knowledge to an encoder-decoder architecture for voice conversion.



Assumption:

1) the context vector generated by TTS text encoder is speaker-independent;
2) TTS decoder works for voice conversion

Beyond Parallel Data

Leveraging TTS

Pre-training TTS model for VC [Wen-Chin Huang et al., INTERSPEECH 2020]

- seq2seq models are data-hungry!
- seq2seq VC model based on the Transformer architecture with TTS pre-training.
- Simple yet effective pre-training technique to transfer knowledge from learned TTS models, which benefit from large-scale TTS corpora.
- Transferring knowledge from Transformer-based TTS models to a Transformer-based VC

Beyond Parallel Data

Leveraging TTS

Many other interesting approaches:

- Text supervision to improve seq2seq VC [Zhang et al., 2018]
- NAUTILUS: A voice cloning system [Hieu-Thi Luong and Junichi Yamagishi, ASRU 2019]
- Speaker adaptive TTS model for VC [Hieu-Thi Luong and Junichi Yamagishi, ASRU 2019]

Beyond Parallel Data

Leveraging TTS: Our perspective

- Deep learning has facilitated the interaction between TTS and voice conversion.
- By leveraging TTS systems, we hope to improve the training and run-time inference of VC!
- However, most of the techniques usually require a large training corpus.
- It deserves future studies as to how voice conversion can benefit from TTS systems without involving large training data.

Beyond Parallel Data

Leveraging ASR

- We know that most ASR systems are already trained with a large corpus.
- They already describe well the phonetic systems in different ways.

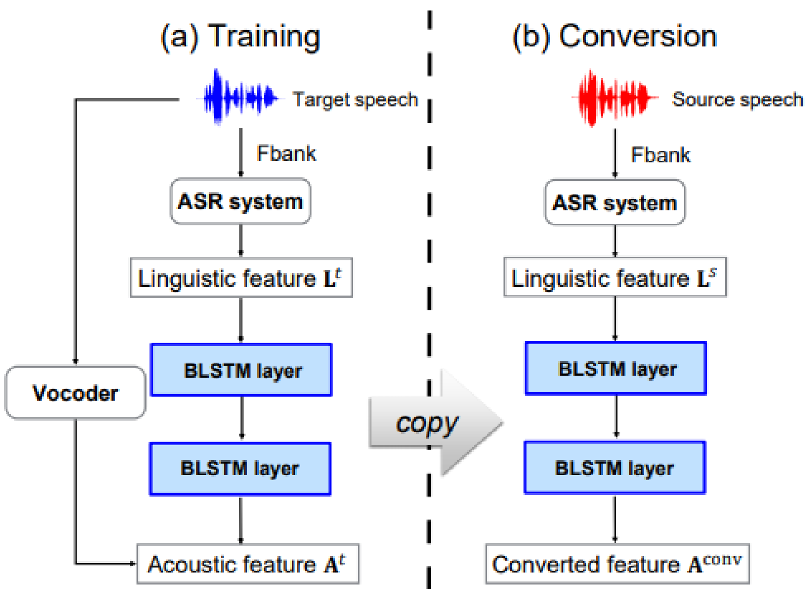
The question is how to leverage the latent representations in ASR systems for voice conversion…

Beyond Parallel Data

Leveraging ASR

Phonetic Posteriograms (PPGs) for VC [Lifa Sun et al., ICME 2016]

- To build a mapping function to convert phonetic posteriogram (PPG) [32] to acoustic features.
- The PPG features are derived from an ASR system, that can be considered as speaker independent.

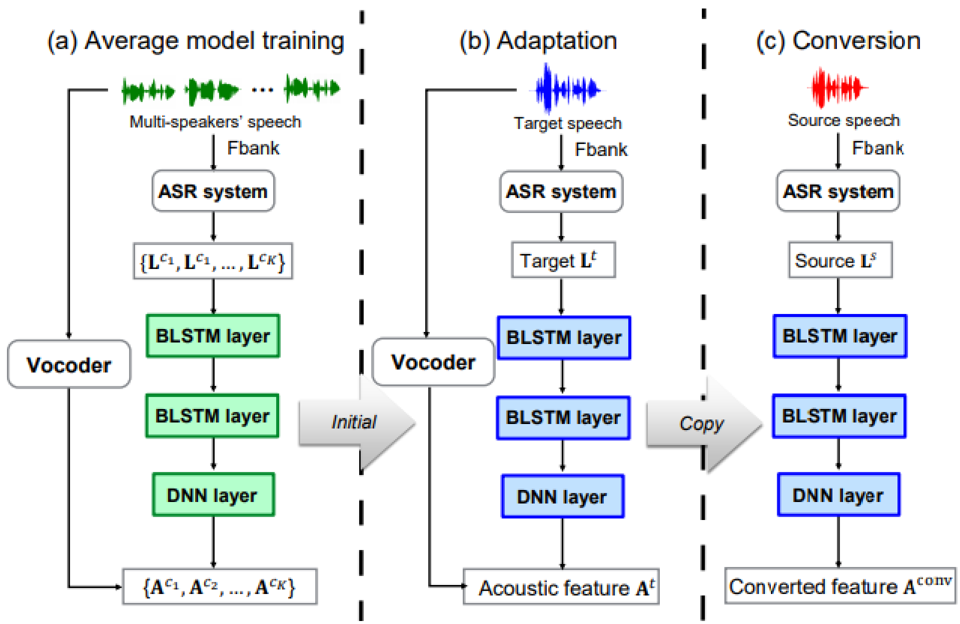


Beyond Parallel Data

Leveraging ASR

PPGs + average model adaptation for VC [Tian et al., Odyssey 2018]

- To build a mapping function to convert phonetic posteriogram (PPG) [32] to acoustic features.
- The average model can be adapted towards the target with a small amount of target speech.



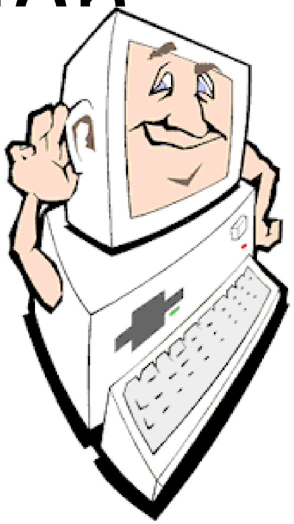
Beyond Parallel Data

Leveraging ASR

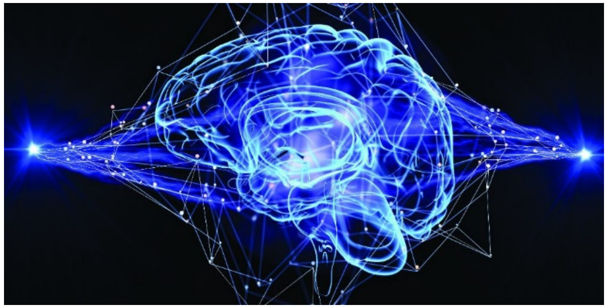
Average Modeling & PPGs for:

- PPG to waveform conversion with WaveNet [Tian et al., INTERSPEECH 2019]
- Emotional voice conversion [Liu et al., 2020]
- Cross-lingual voice conversion [Yi et al., ICASSP 2019]
- Monolingual voice conversion with limited data [Zhang et al., Speech Communication 2020]

Evaluation of Voice Conversion



- Objective Evaluation: Spectrum, Prosody
- Subjective Evaluation: Listening Experiments
- Neural Approaches



Objective Evaluation

- Mel-Cepstral distortion (MCD) [R. Kubichek, 1993]
  - Widely used for Spectrum conversion [Sisman et al., IEEE/ACM TASLP 2020] [Nakashika et al., IEEE/ACM TASLP 2014] [Zhang et al., IEEE/ACM TASLP 2019]

  - A lower MCD indicates better performance.
  - MCD value is **not** always correlated with human perception.
  - Subjective evaluations, such as MOS and similarity score, are also needed!

Objective Evaluation

- Log-Spectral Distance (LSD)
  - Widely-used for Spectrum conversion [Benisty et al., INTERSPEECH 2011] [Tian et al., IEEE/ACM TASLP 2017] [Xie et al., INTERSPEECH 2014] [Sisman et al., IEEE/ACM TASLP 2019].
  - A lower LSD indicates better performance.
  - Similar to MCD, LSD value is **not** always correlated with human perception.
  - Subjective evaluations, such as MOS and similarity score, are also needed!

Objective Evaluation

- Pearson Correlation Coefficient (PCC) [Benesty, Jacob, et al., 2009]
  - Used for F0 and energy contour conversion [Kun et al., INTERSPEECH 2020][Ming et al., ICASSP 2016] [Sisman et al., IEEE/ACM TASLP 2020].
  - A higher PCC value represents better conversion performance.

Objective Evaluation

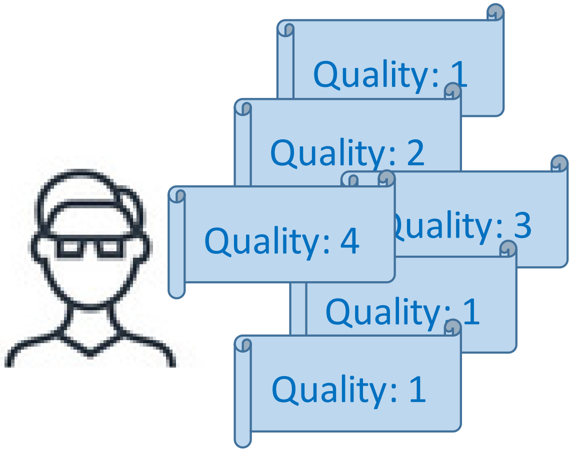
- Root Mean Square Error (RMSE)[Kenney, J. F. et al., 1962]
  - Used for F0 and energy contour conversion [Kun et al., INTERSPEECH 2020][Ming et al., ICASSP 2016] [Sisman et al., IEEE/ACM TASLP 2020].
  - A lower RMSE value represents better conversion performance

Objective Evaluation

- Other generally-accepted metrics for prosody transfer include:
  - F0 Frame Error (FFE) [Wei Chu and Abeer Alwan, ICASSP 2009], which reports the percentage of frames that either contain a 20% pitch error or a voicing decision error.
  - Gross Pitch Error (GPE) [Nakatani et al., Speech Communication 2008] which reports the percentage of voiced frames whose pitch values are more than 20% different from the reference.

Subjective Evaluation

- Mean Opinion Score (MOS)
  - 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.
  - Very widely-used! [Fan Zhang et al., 2014], [Sisman et al., IEEE/ACM 2019], [Toda et al., ICASSP 2005] [Zhao Yi et al., 2020]



Subjective Evaluation

- There are several evaluation methods that are similar to MOS, for example:
  - DMOS [Masatsune Tamura et al., 1998]

    a “degradation” or “differential” MOS test, requiring listeners to rate the sample with respect to this reference.

  - MUSHRA [Slawomir Zielinski et al., 2007]

    is MUltiple Stimuli with Hidden Reference and Anchor, and requires fewer participants than MOS to obtain statistically significant results.

Subjective Evaluation

- AB/ABX test
  - AB test [Sisman et al., IEEE/ACM 2019], [Toda et al., ICASSP 2005] [Zhao Yi et al., 2020]:

    listeners are presented two speech samples and asked to indicate which one has more of a certain property; for example in terms of naturalness, or similarity.

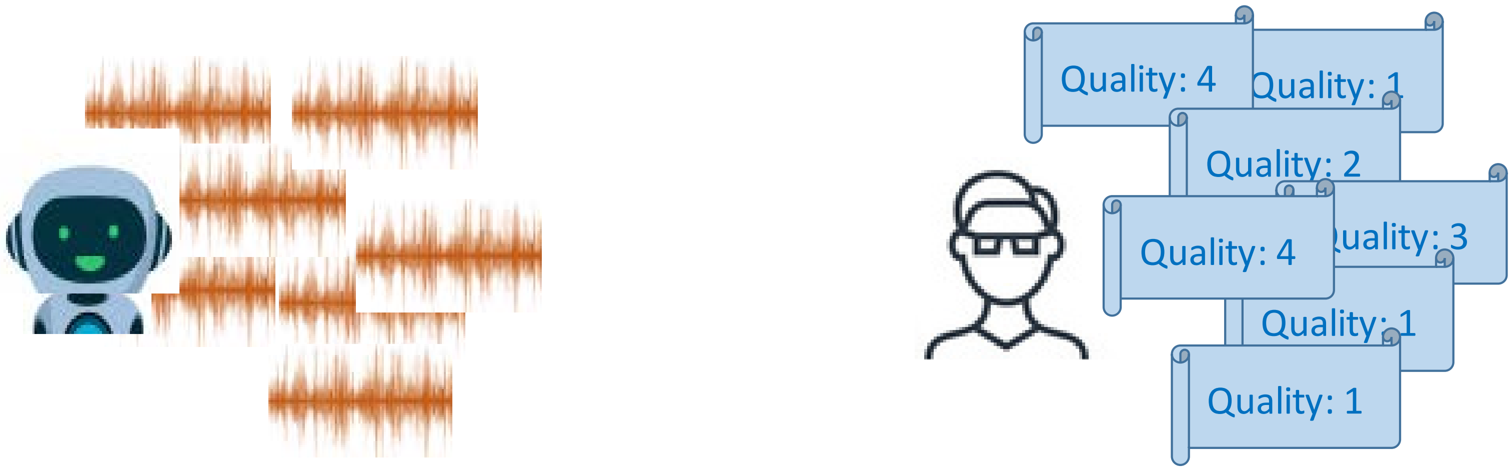
  - ABX test [Y Stylianou et al., IEEE TASLP 1998] [Sisman et al., IEEE/ACM TASLP 2020]:

    similar to that of AB, two samples are given but an extra reference sample is also given. Listeners need to judge if A or B more like X in terms of naturalness, similarity, or even emotional quality.

Very widely-used!

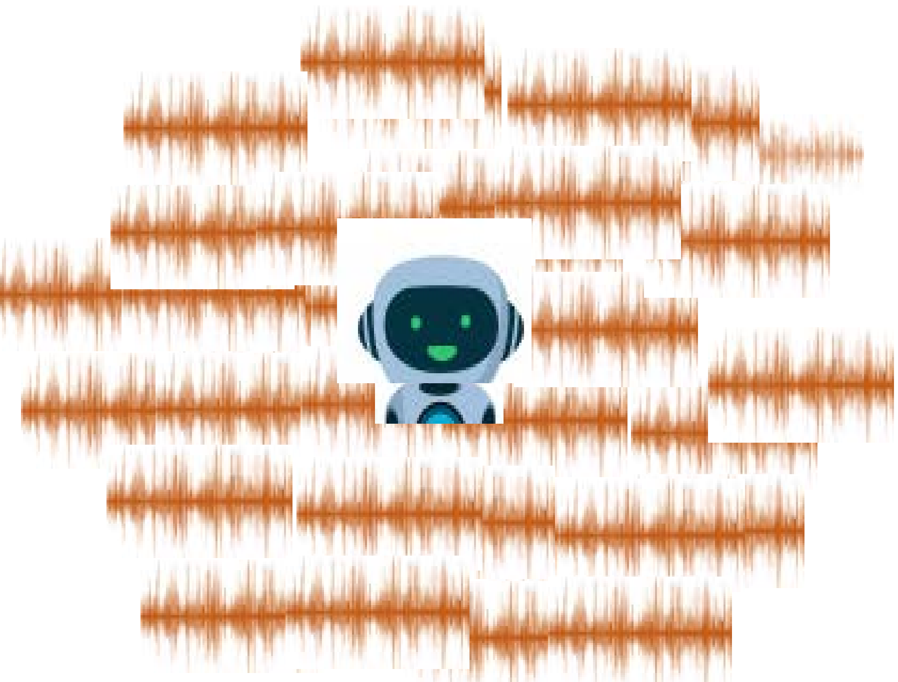
Neural Evaluation Metrics

- Listening tests



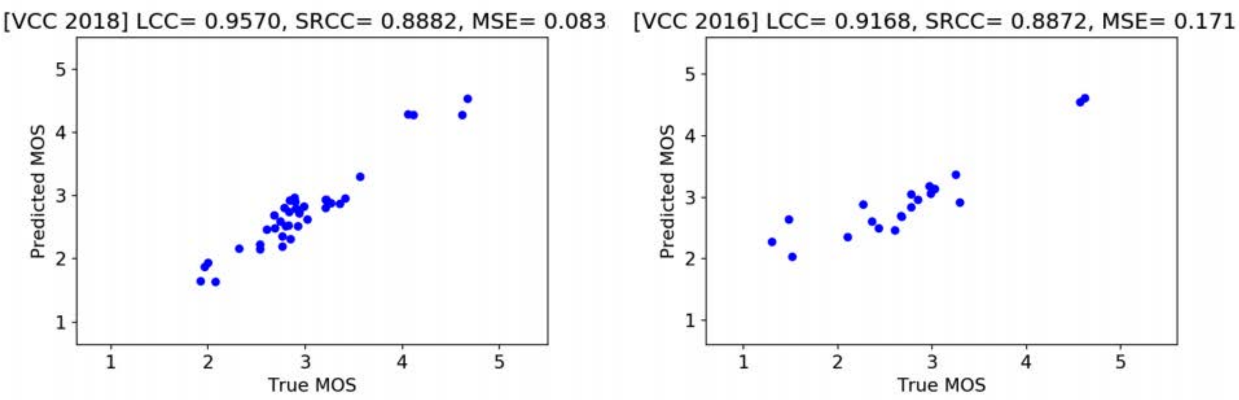
Neural Evaluation Metrics

- A lot of listening tests



Neural Evaluation Metrics

- MOSNet [Lo et al., Interspeech 2019]



VCC 2018 (matched) VCC 2016 (mismatched)

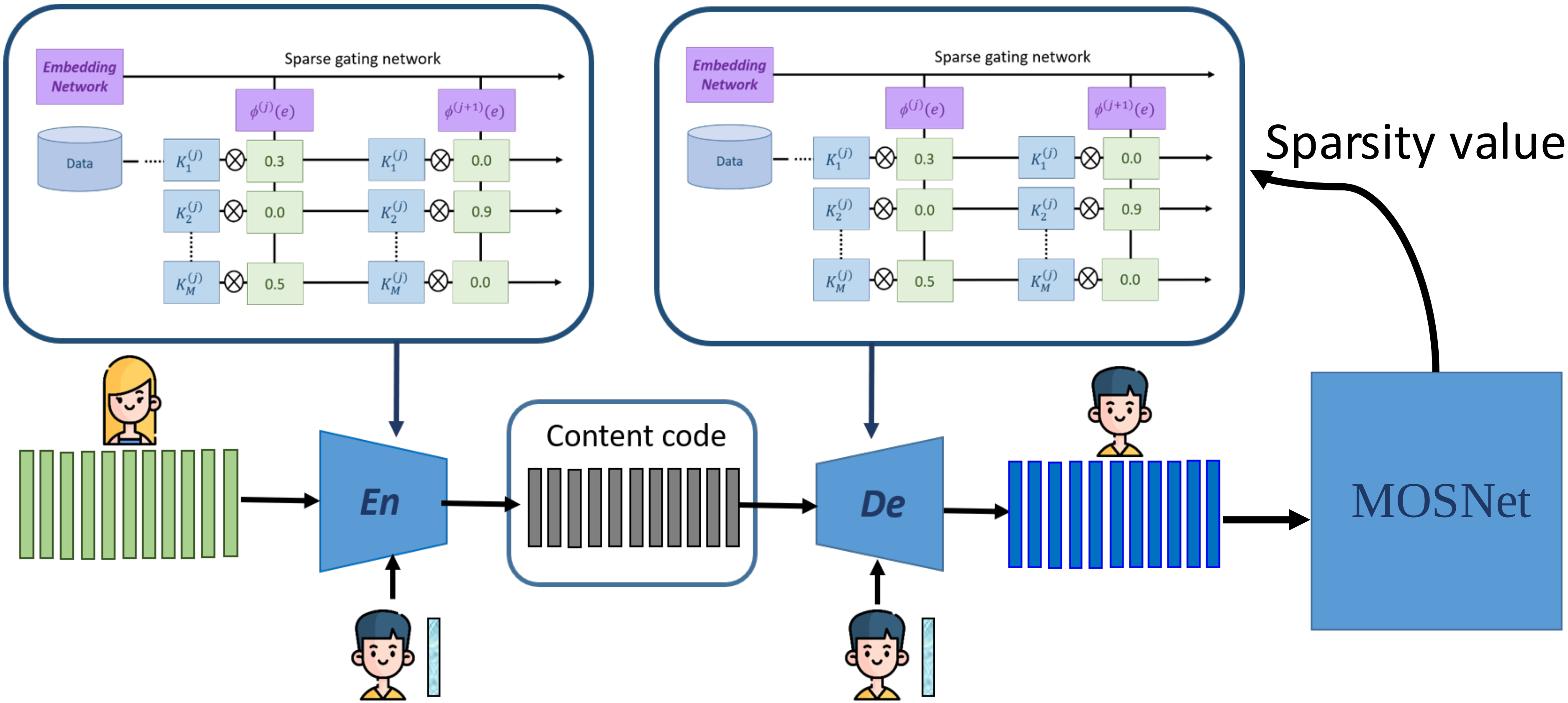
$$O = \frac{1}{S}\underbrace{\boxed{\sum_{s=1}^{S}(\hat{Q}_s - Q_s)^2}}_{\text{Utterance-level}} + \underbrace{\boxed{\frac{\alpha}{T_s}\sum_{t=1}^{T_s}(\hat{Q}_s - q_{s,t})^2}}_{\text{Frame-level}}$$

	LCC	SRCC	MSE
with frame MSE	**0.642**	**0.589**	**0.538**
without frame MSE	0.560	0.528	2.525

- The predicted MOS scores from MOSNet is highly correlated with ground-truth MOS scores.
- A combination of frame- and utterance-level losses achieves better performance.

Neural Evaluation Metrics

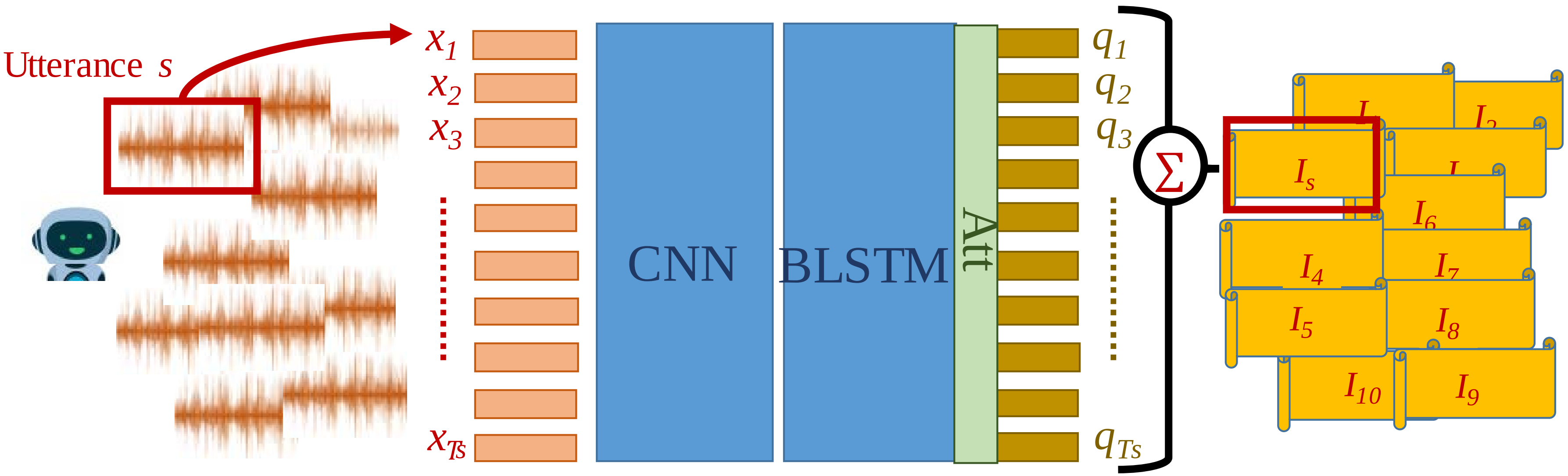
- MoE-VC [Chang et al., ISCSLP 2020]



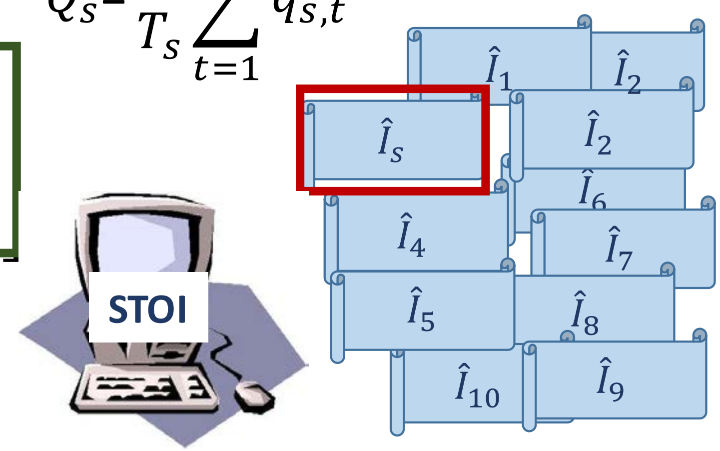
- MOSNet facilitates model architecture optimization online.
- MOSNet serves a new objective function to train VC models.

Neural Evaluation Metrics

- STOI-Net [Zezario et al., APSIPA 2020]



$$Q_s = \frac{1}{T_s}\sum_{t=1}^{T_s} q_{s,t}$$



$$O = \frac{1}{S}\left[\underbrace{\boxed{\sum_{s=1}^{S}\left(\hat{I}_s - I_s\right)^2}}_{\text{Utterance-level}} + \underbrace{\boxed{\frac{1}{T_s}\sum_{t=1}^{T_s}\alpha(\hat{I}_s)\left(\hat{I}_s - i_{s,t}\right)^2}}_{\text{Frame-level}}\right]$$

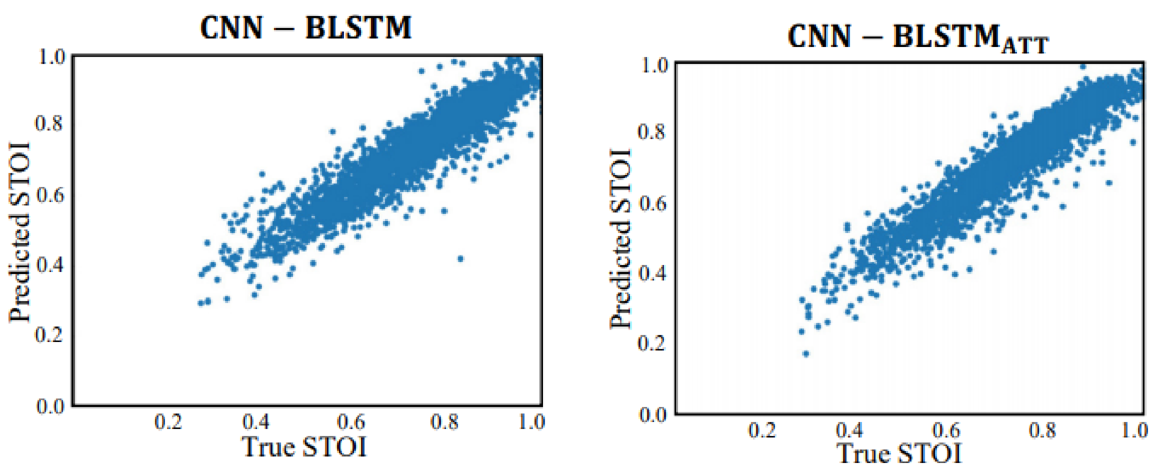
STOI: short-time objective intelligibility

Neural Evaluation Metrics

- STOI-Net [Zezario et al., APSIPA 2020]

LCC, SRCC, and MSE results of BLSTM, CNN-BLSTM, and CNN-BLSTM-ATT under the unseen test condition.

	LCC	SRCC	MSE
BLSTM	0.764	0.784	0.029
CNN-BLSTM	0.789	0.797	0.016
CNN-BLSTM+ATT	**0.827**	**0.815**	**0.015**



- CNN-BLSTM-ATT outperforms BLSTM and CNN-BLSTM.

Voice Conversion Challenges (VCC)

- VCC 2016
- VCC 2018
- VCC 2020

VCC 2016 [Toda et al., INTERSPEECH 2016]

- The first shared task in voice conversion.
- Parallel data voice conversion task.
- 17 participants submitted their conversion results.
- Two hundreds native listeners of English joined the listening tests.

VCC 2016 [Toda et al., INTERSPEECH 2016]

- The first shared task in voice conversion.
- Parallel data voice conversion task.
- 17 participants submitted their conversion results.
- Two hundreds native listeners of English joined the listening tests.

What did we learn?

- ➢ The best system uses GMM and waveform filtering.
- ➢ There is still a huge gap between target natural speech and the converted speech.
- ➢ It remains a unsolved challenge to achieve good quality and speaker similarity at that time.

For dataset, baseline and speech samples:
http://www.vc-challenge.org/vcc2016/summary.html

VCC 2018 [J. Lorenzo-Trueba et al., Odyssey 2018] [T. Kinnunen et al., Odyssey 2018]

Two tasks: parallel VC and non-parallel VC

- Parallel VC:
  - Similar to that of the VCC 2016.
  - VCC 2018 has a smaller number of common utterances uttered by source and target speakers.
- Non-parallel VC:
  - A non-parallel voice conversion task for the first time.
  - The same target speakers' data in the parallel task are used as the target.
  - The source speakers are four native speakers and different from those of the parallel conversion task.

VCC 2018 [J. Lorenzo-Trueba et al., Odyssey 2018] [T. Kinnunen et al., Odyssey 2018]

Two tasks: parallel VC and non-parallel VC

- Parallel VC:
  - Similar to that of the VCC 2016.
  - VCC 2018 has a smaller number of common utterances uttered by source and target speakers.
- Non-parallel VC:
  - A non-parallel voice conversion task for the first time.
  - The same target speakers' data in the parallel task are used as the target.
  - The source speakers are four native speakers and different from those of the parallel conversion task.

Important aspects

- ➢ To bridge the gap between the automatic speaker verification (ASV) and VC communities.
- ➢ To assess the spoofing performance of VC systems on the basis of anti-spoofing scores.

VCC 2018 [J. Lorenzo-Trueba et al., Odyssey 2018]

[T. Kinnunen et al., Odyssey 2018]

http://www.vc-challenge.org/vcc2018/index.html

Two tasks: parallel VC and non-parallel VC

- Parallel VC:
  - Similar to that of the VCC 2016.
  - VCC 2018 has a smaller number of common utterances uttered by source and target speakers.
- Non-parallel VC:
  - A non-parallel voice conversion task for the first time.
  - The same target speakers' data in the parallel task are used as the target.
  - The source speakers are four native speakers and different from those of the parallel conversion task.

Important aspects

- ➢ To bridge the gap between the automatic speaker verification (ASV) and VC communities.
- ➢ To assess the spoofing performance of VC systems on the basis of anti-spoofing scores.

VCC 2020

Two tasks:

- 1) parallel/non-parallel mono-lingual VC
- 2) non-parallel cross-lingual VC (English-Finnish, English-German, and English-Mandarin).

Baselines: CycleVAE, ASR + TTS based VC

Important aspects

In addition to the traditional evaluation metrics, the challenge also reports the **speech recognition**, **speaker recognition**, and **anti-spoofing evaluation results** on the converted speech.

For baselines and codes:
http://www.vc-challenge.org/

VCC 2016, 2018, 2020

Challenge	Language	Task	Training Data	# Speakers	Testing Data
VCC 2016	monolingual	parallel	162 paired utterances	4 source, 4 target	54 utterances
VCC 2018	monolingual	parallel	81 paired utterances	4 source, 4 target	35 utterances
	monolingual	nonparallel	81 unpaired utterances	4 source, 4 target	35 utterances
VCC 2020	monolingual	parallel + nonparallel	20 paired, 50 unpaired utterances	4 source, 4 target	25 utterances
	crosslingual	nonparallel	70 unpaired utterances	4 source, 6 target	25 utterances

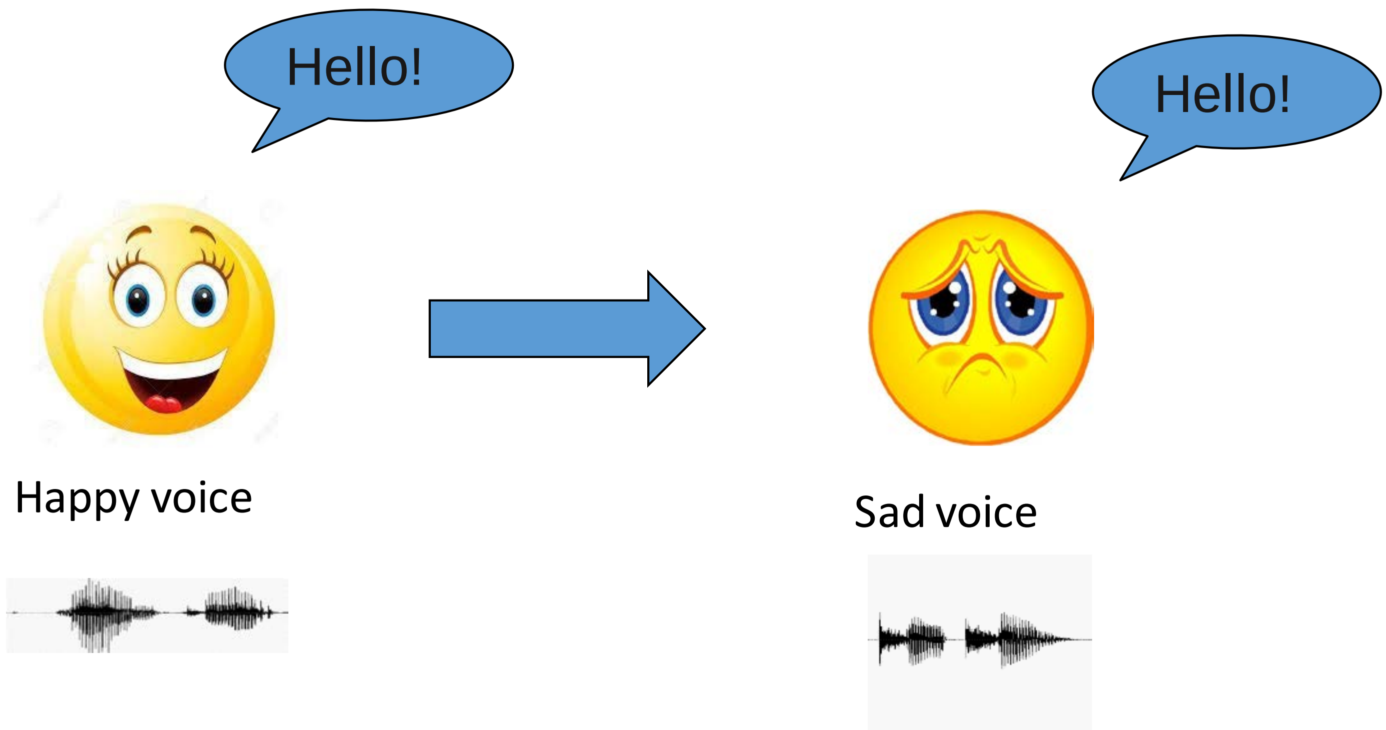
TABLE I: Summary of VCC 2016, VCC 2018 and VCC 2020.

An application of VC

Emotional Voice Conversion

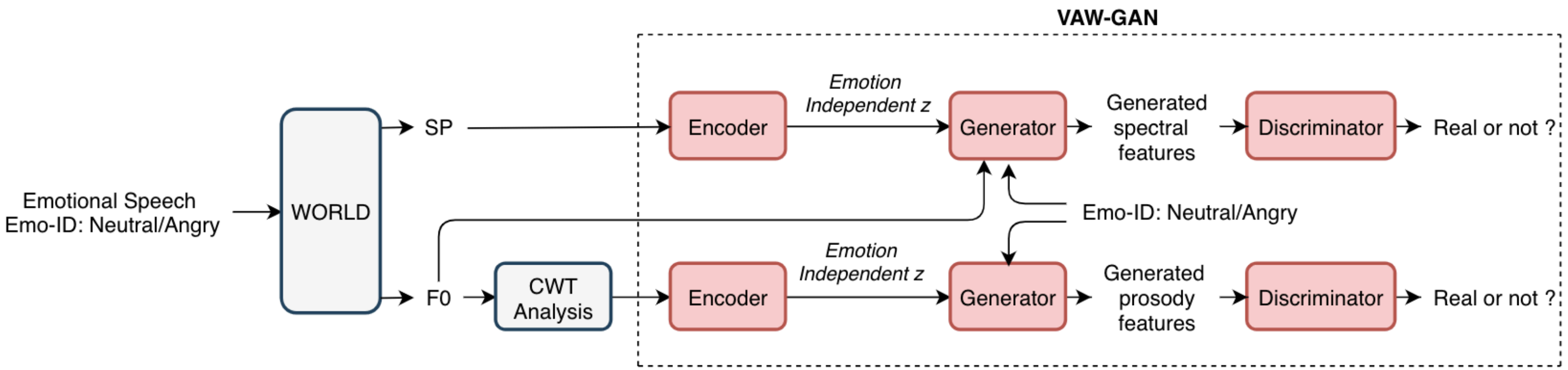
Emotional Voice Conversion

To convert one's voice from one emotion state to another, while protecting the speaker identity and linguistic content.



Emotional Voice Conversion

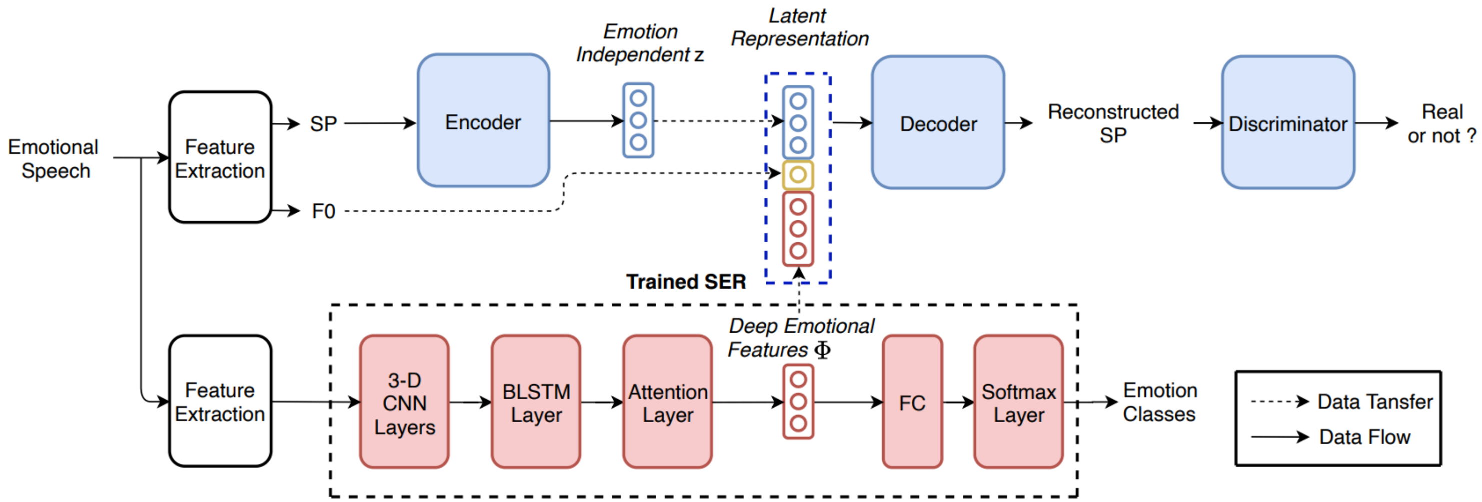
- Converting Anyone's Emotion [Zhou et al., INTERSPEECH 2020]



For publicly available codes and speech samples:
https://github.com/KunZhou9646/Speaker-independent-emotional-voice-conversion-based-on-conditional-VAW-GAN-and-CWT

Emotional Voice Conversion

- Seen and unseen emotional style transfer [Zhou et al., submitted to ICASSP 2020]



For publicly available codes and speech samples:
https://github.com/KunZhou9646/controllable_evc_code

Emotional Voice Conversion

Interesting approaches from recent years

[1] Zhaojie Luo et al., “Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019. https://ieeexplore.ieee.org/document/8740871

[2] Jian Gao et al., Nonparallel emotional speech conversion,” Proc. Interspeech 2019. https://arxiv.org/abs/1811.01174

[3] C. Robinson et al., “Sequence-to-sequence modelling of f0 for speech emotion conversion,” In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019. https://ieeexplore.ieee.org/document/8683865

[4] Rizos, Georgios et al., "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020. https://ieeexplore.ieee.org/document/9054579

[5] Shankar, Ravi et al., "Multi-speaker Emotion Conversion via Latent Variable Regularization and a Chained Encoder-Decoder-Predictor Network." Proc. Interspeech 2020. https://arxiv.org/abs/2007.12937

EVC Dataset?

Challenge: Lack of open-source emotional database

1. **RAVDESS database[1]:**
   2 different English sentences read by 24 actors in 8 emotions
2. **CREMA-D database[2]:**
   12 different English sentences recorded by 91 actors in 6 emotions
3. **Berlin Emotional Speech Dataset[3]:**
   10 different German sentences spoken by 10 actores in 7 emotions
4. **SAVEE database[4]:**
   15 different English sentences spoken by 4 male speakers in 7 emotions

Limited diversity of sentences!

5. **IMPROV database[5]:**
   9 hours of audio-visual data from 12 actors
6. **IEMOCAP database[6]:**
   12 hours of audio-visual data from 10 speakers

Contains over-lapping speech and external noise, thus not suitable for speech synthesis!

7. **EmoV-DB database[7]:**
   Around 300 different English utterances recorded by 4 speakers

Contains non-verbal expressions, limited diversity of speakers!

8. **CMU Arctic Speech Database[8]:** *All neutral utterances*
9. **AmuS database[9]:** *All amused utterances*

Only contains one single emotion!

Emotional Speech Dataset (ESD)

We release a new multi-lingual and multi-speaker parallel emotional speech dataset that can be used for various speech synthesis and voice conversion tasks:

- Mono-lingual VC
- Cross-lingual voice conversion,
- Emotional voice conversion (mono-lingual and/or cross-lingual)

350 parallel utterances by 10 native English, and 10 Mandarin speakers.

For each language, the dataset consists of 5 male and 5 female speakers with five emotions:

- neutral
- happy
- sad
- angry
- surprise

Emotional Speech Dataset (ESD)

Download from:

https://github.com/HLTSingapore/Emotional-Speech-Data

Zhou, Kun, Berrak Sisman, Rui Liu, and Haizhou Li. "Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset." arXiv preprint arXiv:2010.14794 (2020).

Conclusion

- Introduction to VC & History
- Parallel Data for VC
- Beyond Parallel Data for VC
- Evaluation of VC
- VC Challenges
- Emotional Voice Conversion (with codes)
- A New Dataset for VC and Emotional VC!

Acknowledgement

- Team members at NUS, SUTD, and Academia SINICA!
- Special thanks to our PhD student Zhou Kun (NUS)!

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

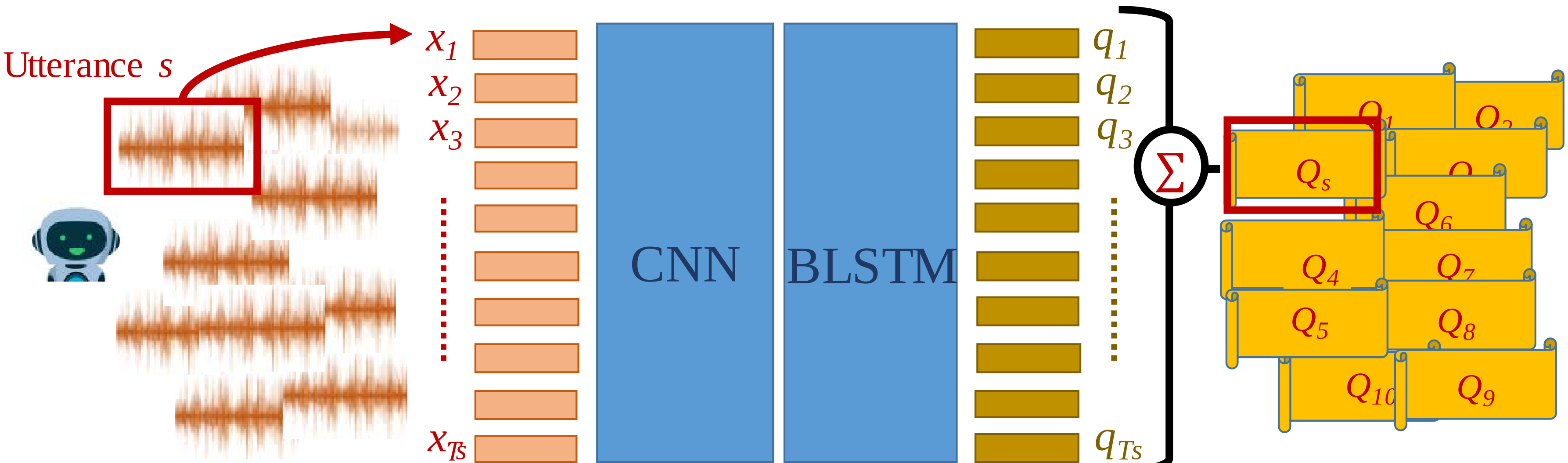
Established in collaboration with MIT

中央研究院 ACADEMIA SINICA

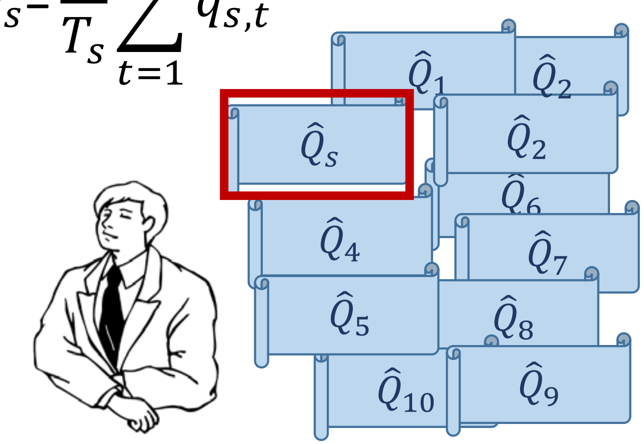
NUS National University of Singapore

Neural Evaluation Metrics

- MOSNet [Lo et al., Interspeech 2019]



$$Q_s = \frac{1}{T_s}\sum_{t=1}^{T_s} q_{s,t}$$

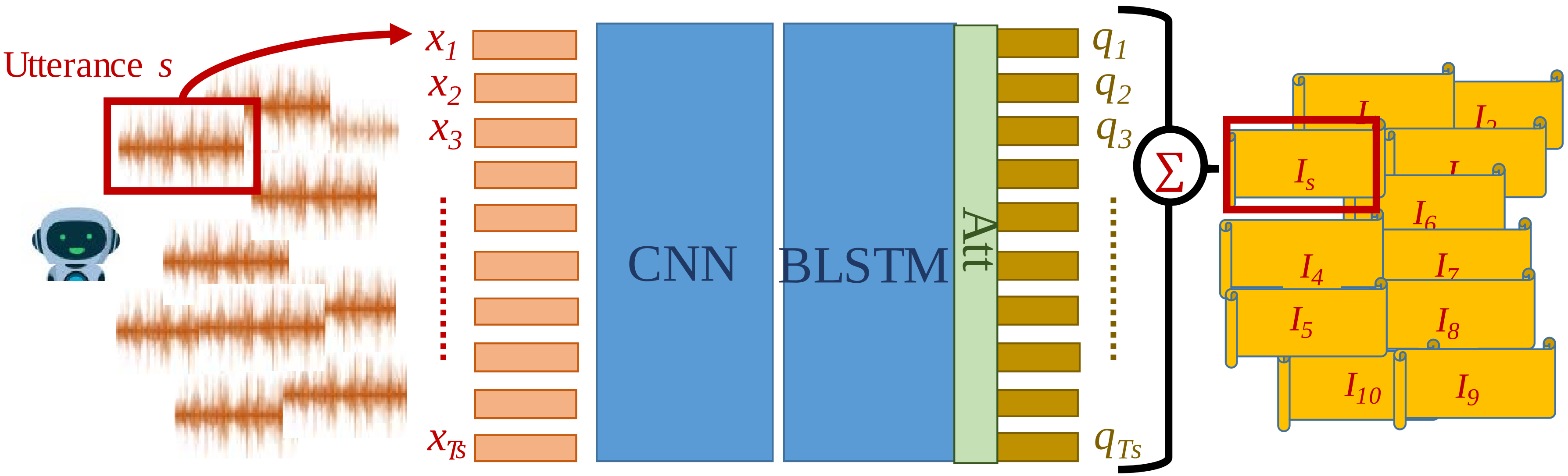


$$O = \frac{1}{S}\left[\boxed{\sum_{s=1}^{S}\left(\hat{Q}_s - Q_s\right)^2} + \boxed{\frac{\alpha}{T_s}\sum_{t=1}^{T_s}\left(\hat{Q}_s - q_{s,t}\right)^2}\right]$$

Utterance-level　　Frame-level

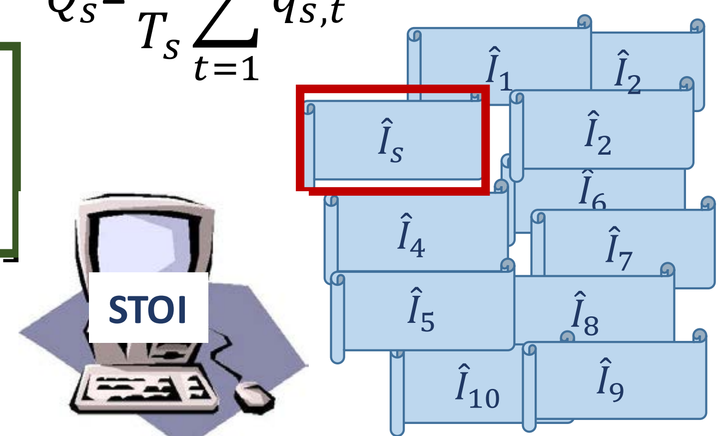
Neural Evaluation Metrics

- STOI-Net [Zezario et al., APSIPA 2020]



$$Q_s = \frac{1}{T_s}\sum_{t=1}^{T_s} q_{s,t}$$

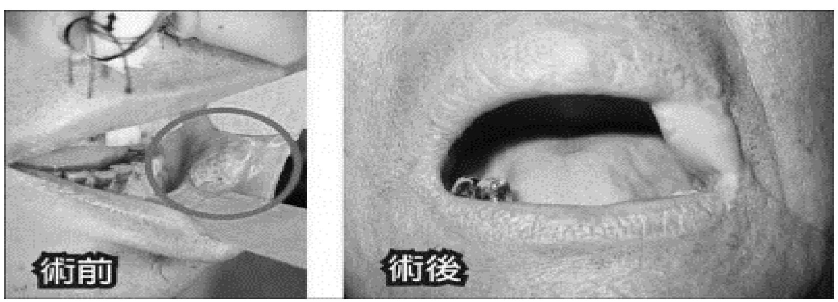
$$O = \frac{1}{S}\Bigg[\underbrace{\boxed{\sum_{s=1}^{S}\left(\hat{I}_s - I_s\right)^2}}_{\text{Utterance-level}} + \underbrace{\boxed{\frac{1}{T_s}\sum_{t=1}^{T_s}\alpha(\hat{I}_s)\left(\hat{I}_s - i_{s,t}\right)^2}}_{\text{Frame-level}}\Bigg]$$



STOI: short-time objective intelligibility

JDNMF for Impaired Speech Conversion

- **Task:** improving the speech intelligibility of surgical patients.
- **Target:** oral cancer (top five cancer for males in Taiwan).



Before After

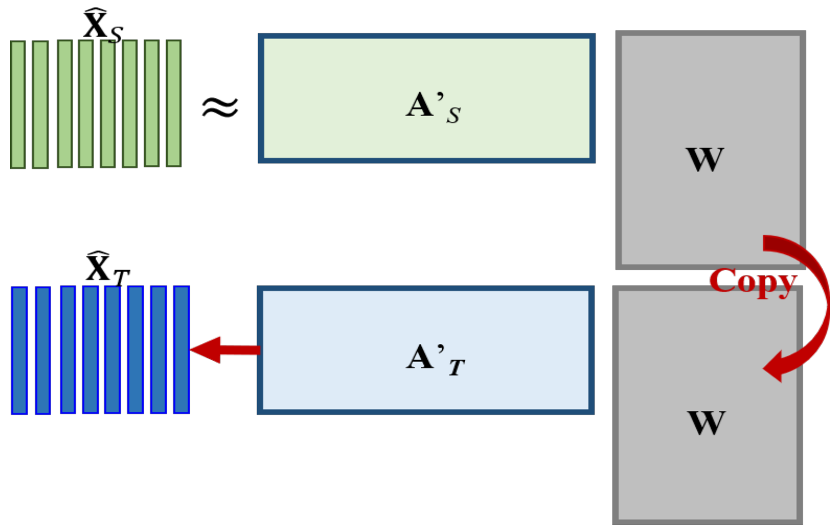
Liberty Times Ltd..

Before After

Taipei Veterans General Hospital

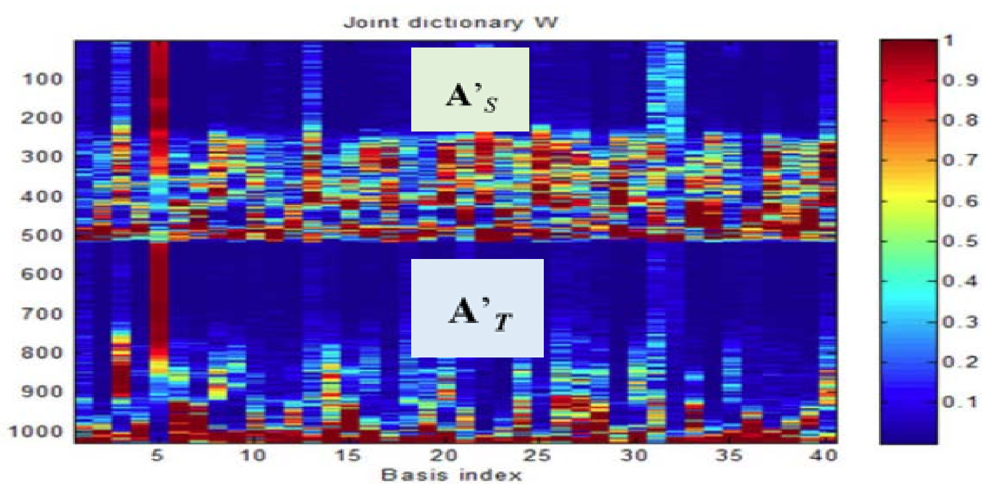
JDNMF for Impaired Speech Conversion

- Proposed: joint training of source and target dictionaries with non-negative matrix factorization (NMF):



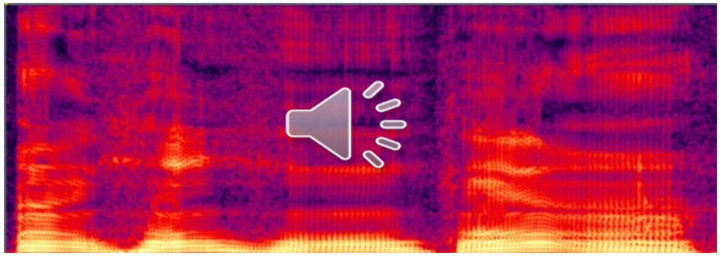
After Surgery

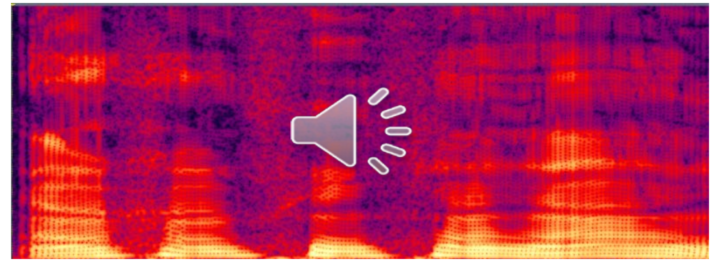
Before Surgery



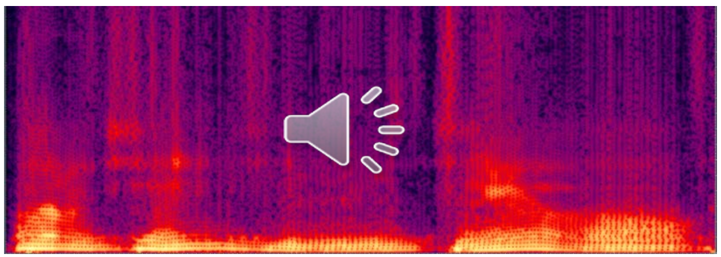
JDNMF for Impaired Speech Conversion

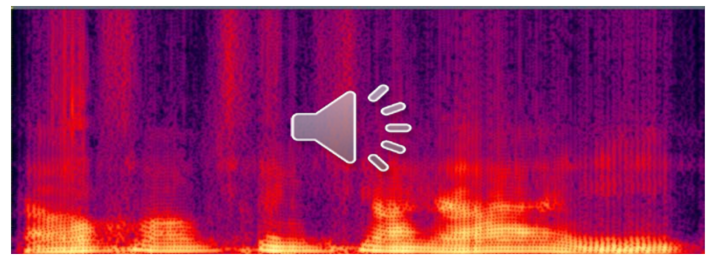
Original:





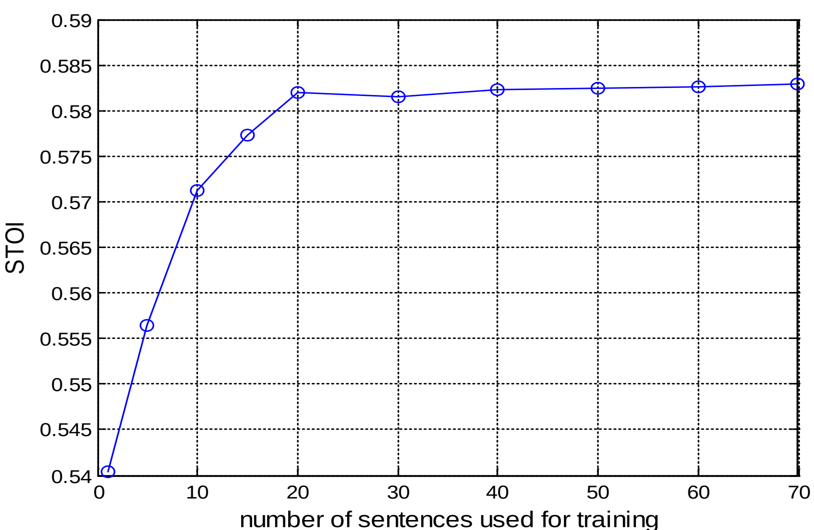
After Conversion:





衛生紙給我

遙控器在哪裡



Speech samples were from [Fu et. al., TBME 2017]