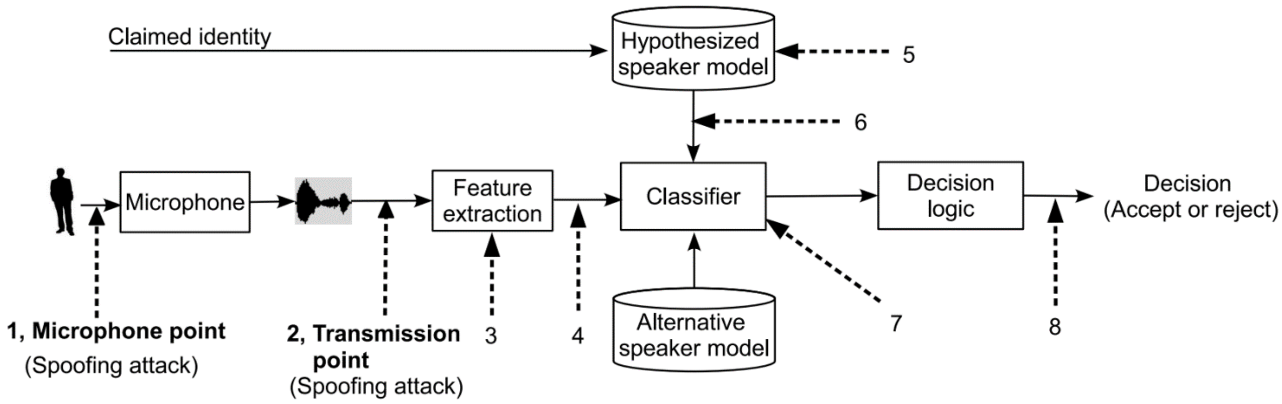
語音偽造辨識與偵測

2020/9/9

Outline

- Spoofing Attack and Automatic Speaker Verification
- Spoofing Attacks Methods
  - ➢ Replay attack
  - ➢ Impersonation (twins and siblings)
  - ➢ Cut and paste
  - ➢ Voice conversion (text-to-speech)
  - ➢ Acoustic scene conversion

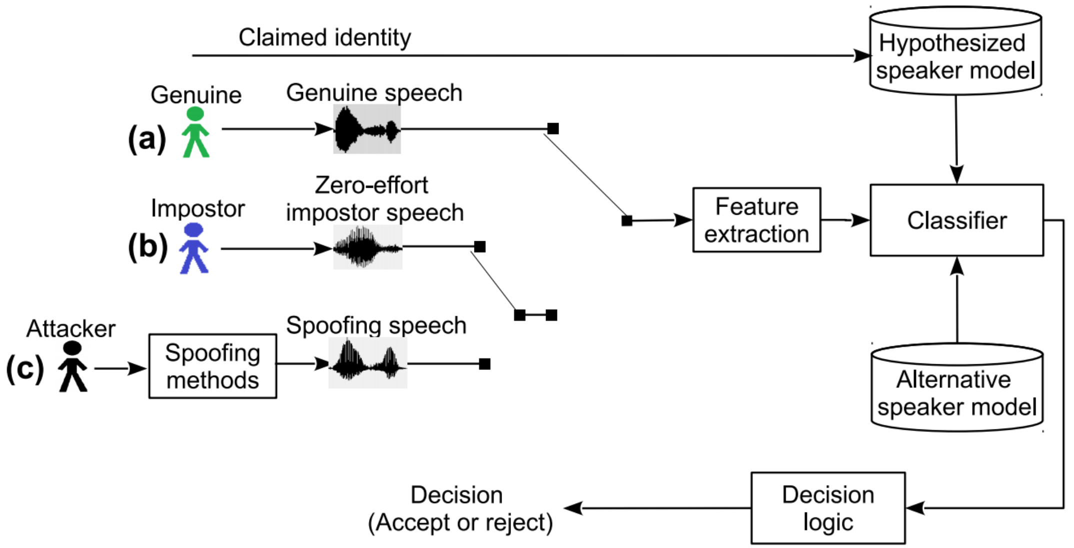
Automatic Speaker Verification (ASV) System With Eight Attack Points



- Direct attacks (spoofing attacks), can be applied at the microphone level as well as the transmission points 1 and 2.
- Indirect attacks (ASV system): points 3 to 8. They generally require system-level access, such as interfering with feature extraction (points 3 and 4), models (points 5 and 6) or score and decision logic computation (points 7 and 8).

Wu et. al., "Spoofing and countermeasures for speaker verification: a survey," Speech Communication 2015.

Automatic Speaker Verification (ASV) System With Eight Attack Points



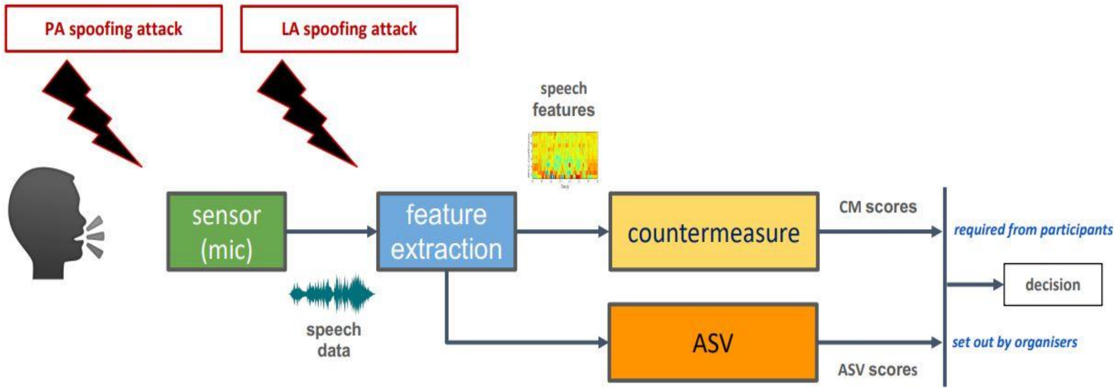
	Decision	
	Accept	Reject
Genuine	Correct acceptance	False rejection
Impostor	False acceptance	Correct rejection

- Standard ASV: an evaluation using (a) and (b).
- Spoofing and countermeasure: an evaluation using (a) and (c).
- (c) represents spoofed version of (b), and (b) has the same number of trials as (c).

Wu et. al., "Spoofing and countermeasures for speaker verification: a survey," Speech Communication 2015.

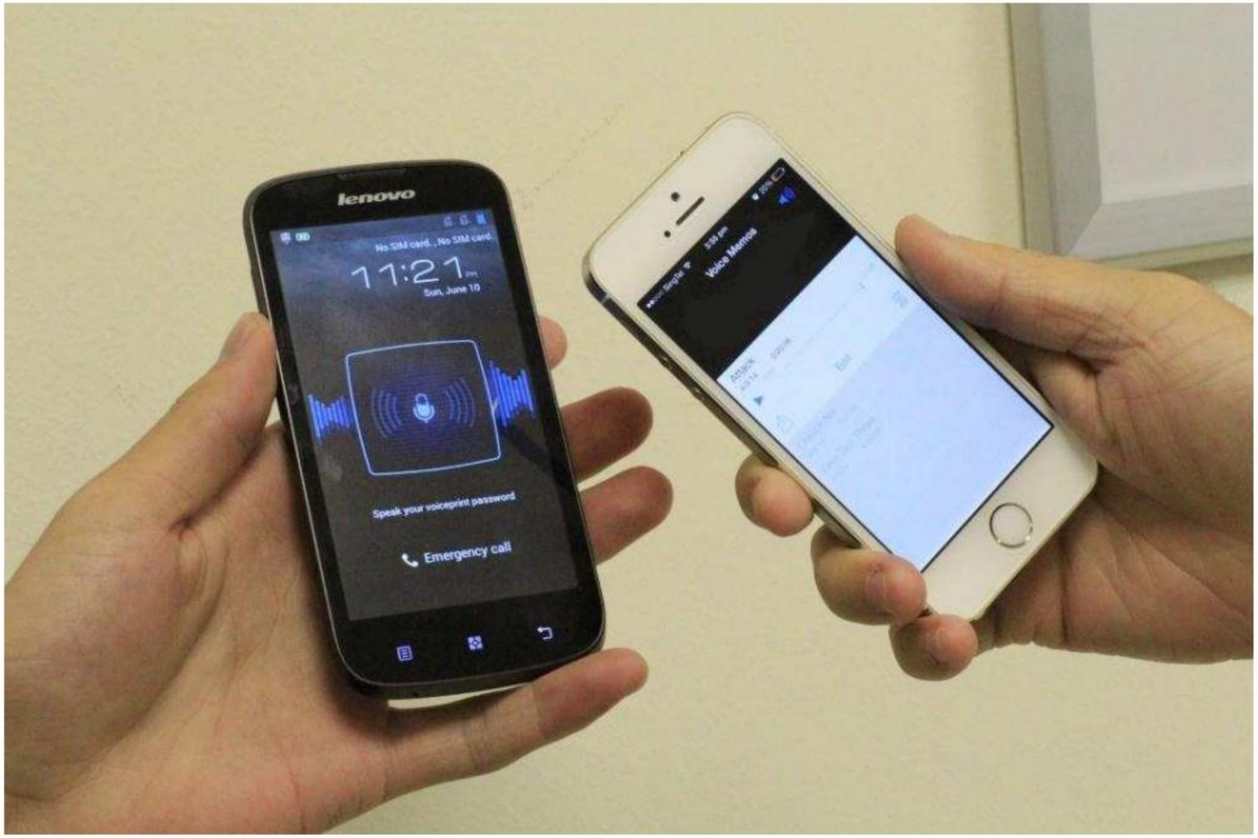
Spoofing Attack Methods

- Replay attack
- Impersonation (twins and siblings)
- Cut and paste
- Voice conversion (Text-to-speech)
- Acoustic scene conversion

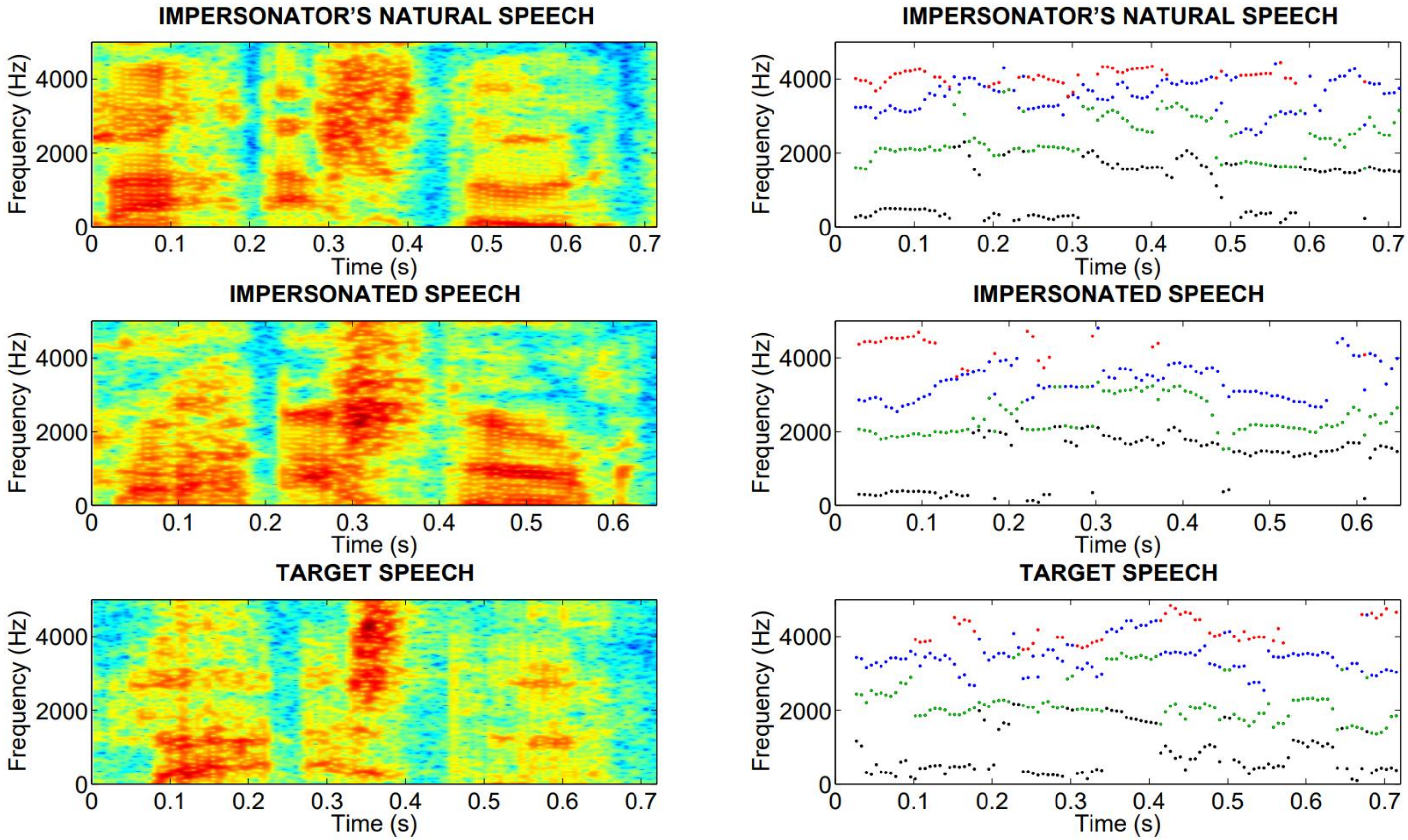


PA (physical access)
LA (logical access)

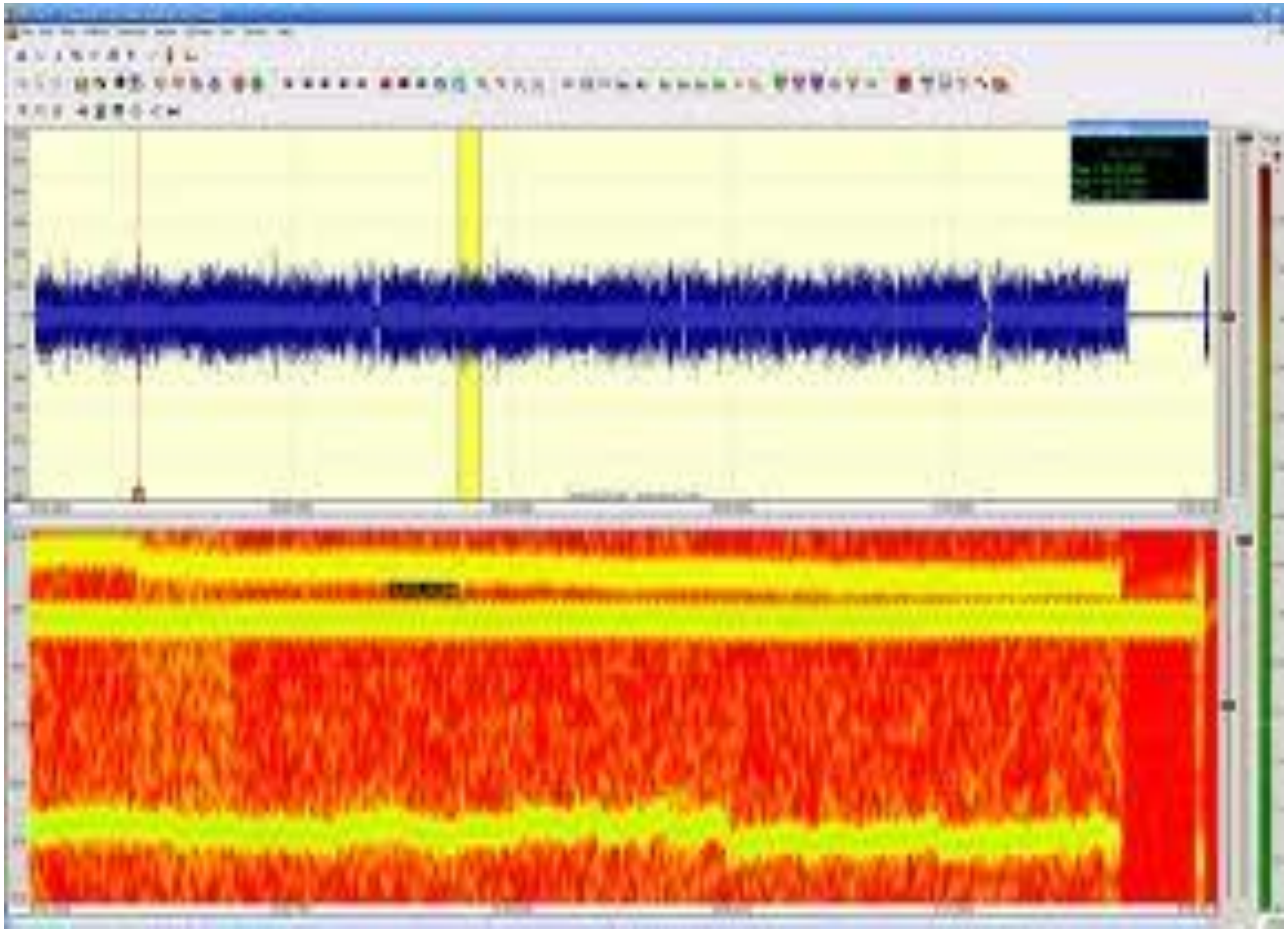
Replay Attack



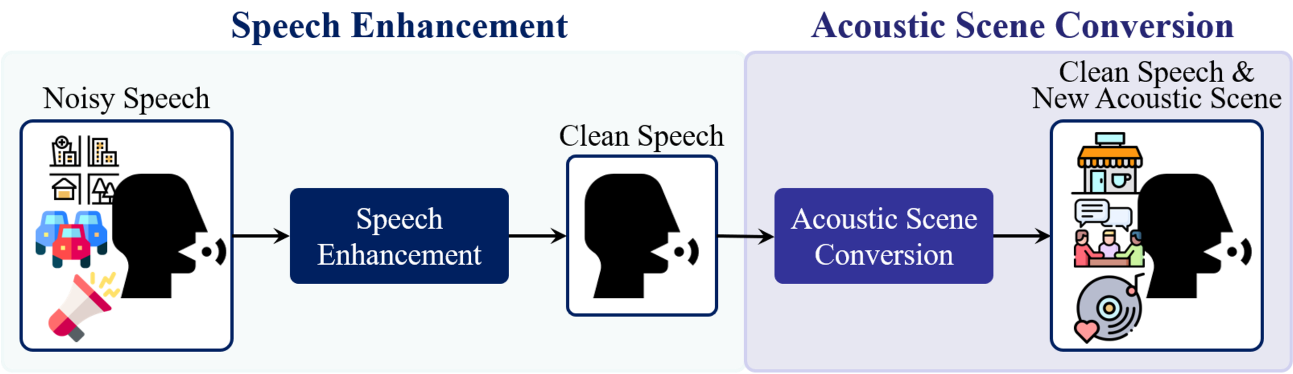
Impersonation (Twins and Siblings)

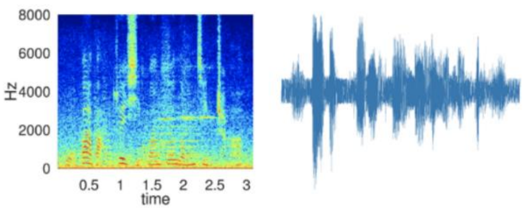


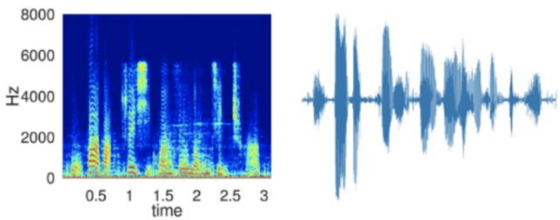
Cut and Paste

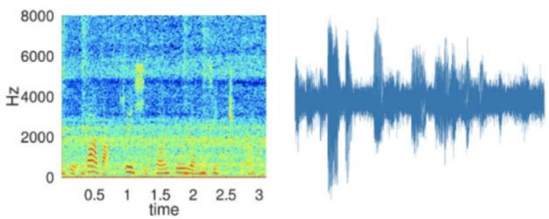


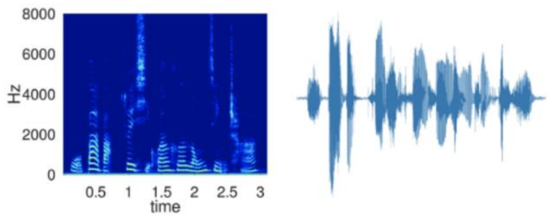
Acoustic Scene Conversion



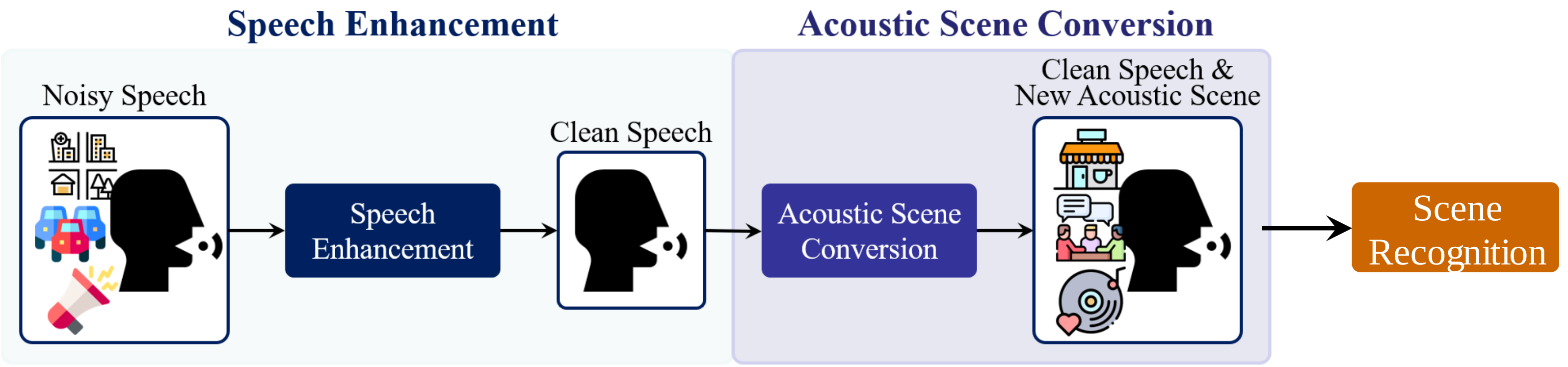








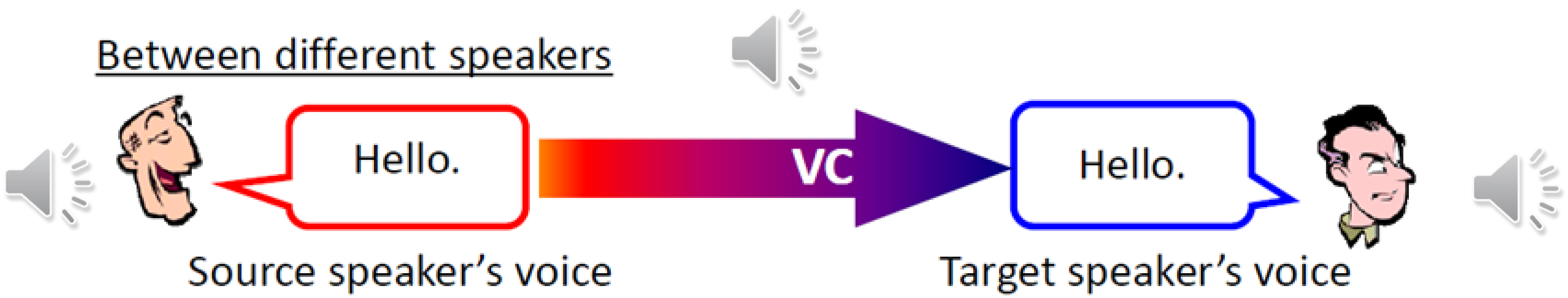
Acoustic Scene Conversion



	0dB	2dB	5dB
DDAE	100%	100%	100%
FCN	100%	100%	100%
MMSE	95.22%	94.94%	95.53%

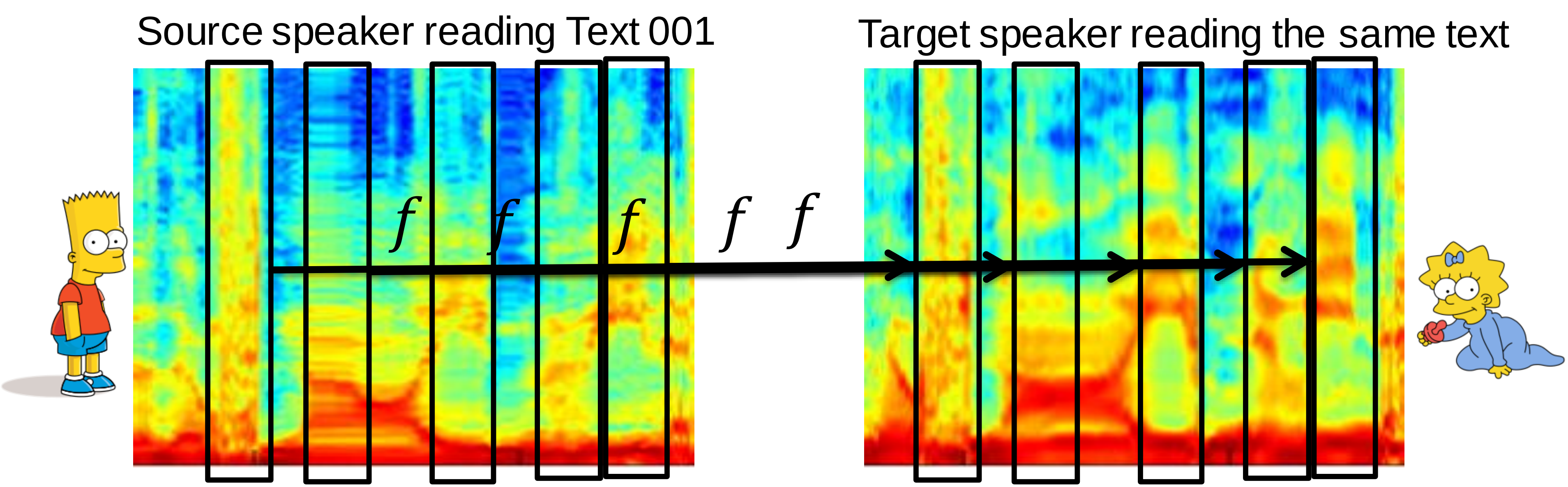
Voice Conversion

- Voice Conversion (VC) is a technique that converts one type of speech to another, without changing the linguistic content
- Applications:
  - Impaired speech to normal speech conversion
  - Narrowband speech to wideband speech conversion (bandwidth expansion)
  - Speech to singing conversion
  - Speaker voice conversion

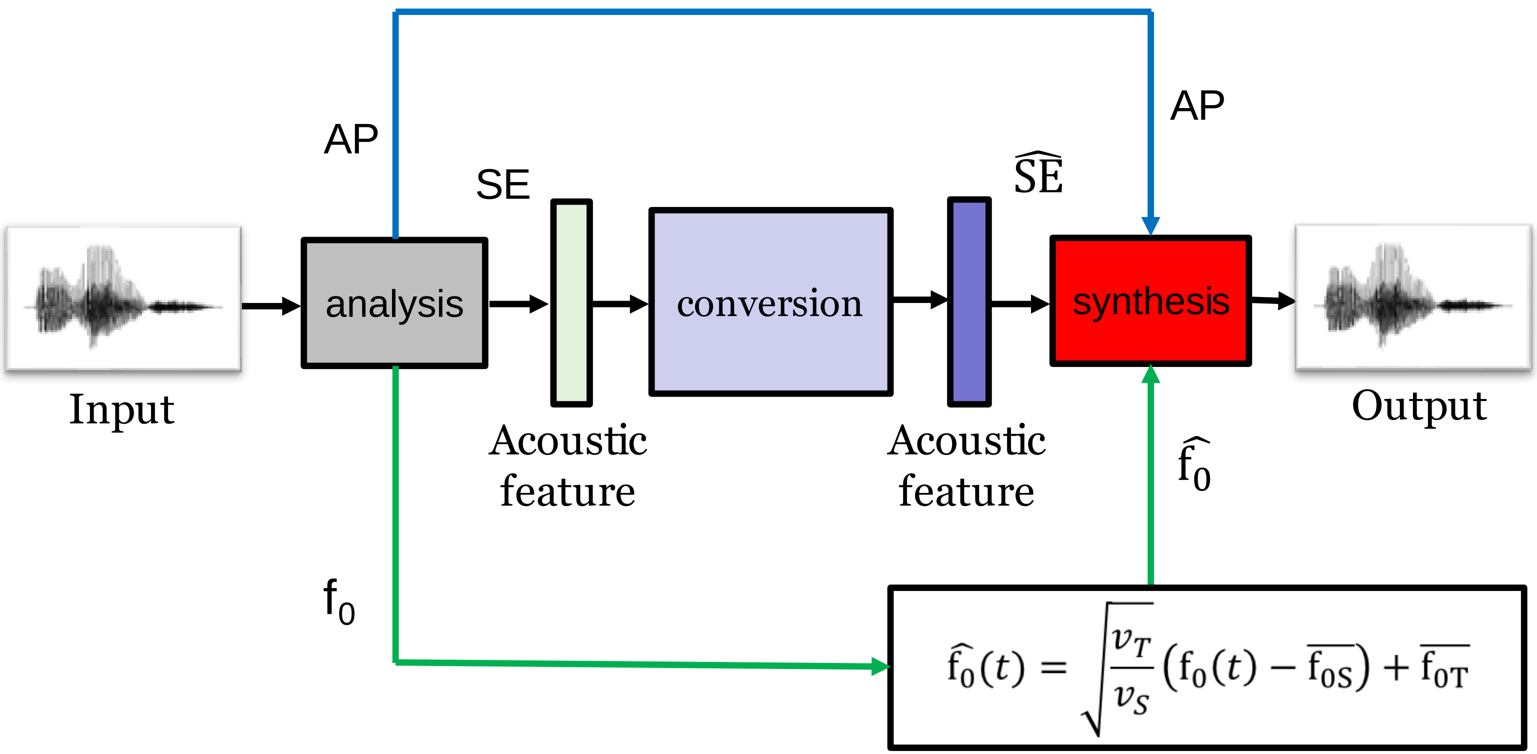


Spectral Conversion

- Convert the "spectrum" of a source to a target
- Standard procedures:
  - Parallel corpus available (same text for the source and target speakers)
  - Alignment (e.g., DTW)
  - Mapping function estimation: $\boldsymbol{x}_t = f(\boldsymbol{x}_s)$



Voice Conversion



Types of Voice Conversion

- One-to-one vs. Many-to-one
  - One-to-one VC: the source and target speech utterances are available in the offline stage
  - Many-to-one VC: the source speaker is not seen in the offline stage
    - The system can convert the speech of any arbitrary source speaker to that of a desired target speaker
  - One-to-many, Many-to-many

- Parallel vs. Non-parallel
  - Parallel: parallel speech corpora available in the offline stage
  - Non-parallel: parallel speech corpora not available in the offline stage

Parallel One-To-One VC Methods

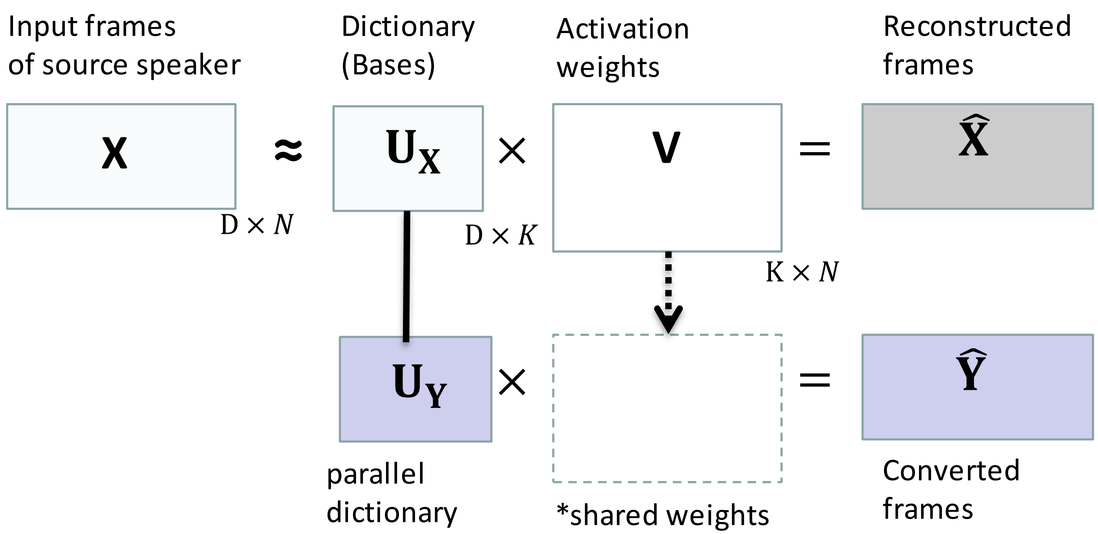
- Statistical methods:
  - Linear methods: Gaussian Mixture Model (GMM), partial least squares regression (PLS), etc
  - Nonlinear methods: dynamic kernel PLS (DKPLS), neural network (NN), etc

- Exemplar-based methods:
  - Nonnegative matrix factorization (NMF)
  - Locally linear embedding (LLE)

- Others:
  - Frequency Warping (FW), hybrid methods (e.g., FW+NMF), etc

ENMF-based VC (1/2)

- Exemplar-based NMF (ENMF) VC
  - Pre-select a **source** & **a target dictionary**
  - Obtain an **activation** by reconstructing the source input
  - Predict the output by activating the target dictionary



$$
\begin{aligned}
\boldsymbol{X} &= [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N], \\
\boldsymbol{U}_X &= [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K], \\
\boldsymbol{V} &= [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N], \\
\boldsymbol{X} &\approx \boldsymbol{U}_X \boldsymbol{V}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{V}^* &= \arg\min_{\mathbf{V}} \mathrm{D}(\hat{\mathbf{X}}, \mathbf{X}) + \mathrm{C}(\mathbf{V}) \\
&= \arg\min_{\mathbf{V}} \mathrm{D}(\mathbf{U}_X \mathbf{V}, \mathbf{X}) + \mathrm{C}(\mathbf{V})
\end{aligned}
$$

$$
\hat{\mathbf{Y}} = \mathbf{U}_Y \mathbf{V}
$$

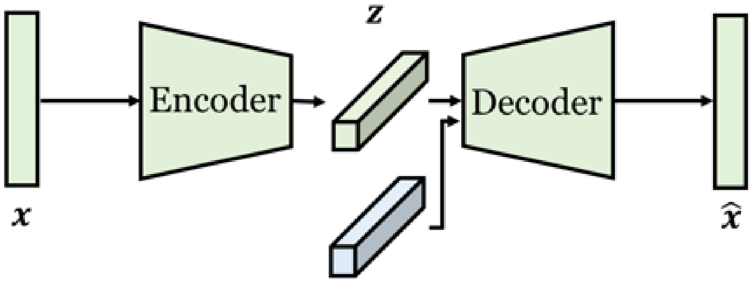
ENMF-based VC (2/2)

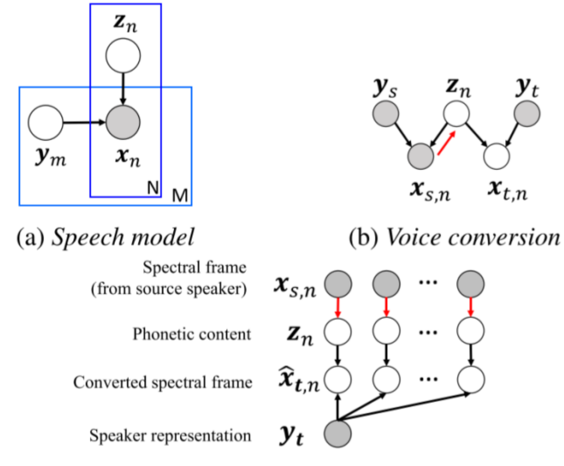
- Advantages
  - Acceptable quality
  - Reasonable similarity to the target speaker
  - Applicable on-the-fly (without training phases)
- Disadvantages
  - Slow conversion (solving V iteratively during conversion)
  - Less scalable (voice quality of output is somewhat proportional to the dictionary size)
  - Trade-off between performance and speed
- Dictionary learning ? $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \approx \begin{bmatrix} \mathbf{U}_{\mathrm{X}} \\ \mathbf{U}_{\mathrm{Y}} \end{bmatrix} \mathbf{V}$ high complexity
- Fast conversion ?

Non-parallel VAE-based VC

- VAE: variational autoencoder
- Nonparallel: no parallel speech corpora

$$z_n = f_\phi(x_n) \qquad \hat{x}_n = f_\theta(z_n, y_n)$$





(c) The conversion phase

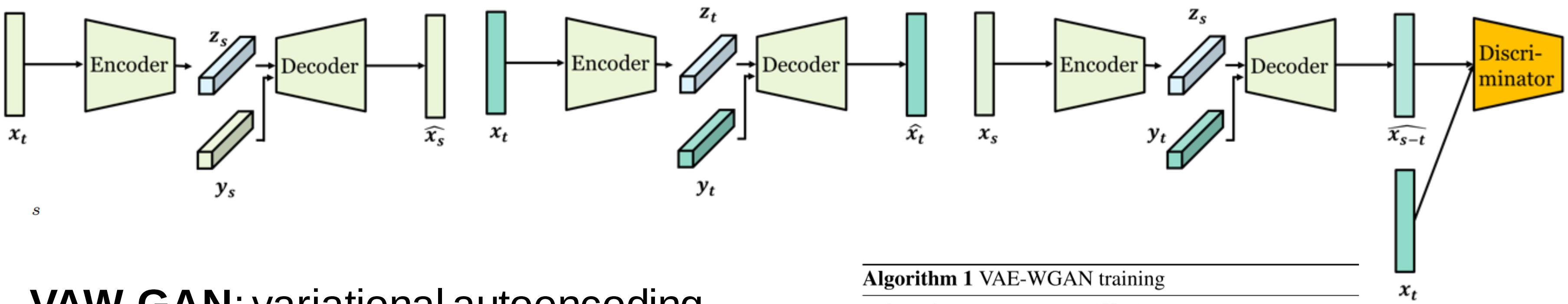
$$\hat{x}_n = \hat{f}(x_n, y_n) = f_\theta(z_n, y_n) = f_\theta(f_\phi(x_n), y_n)$$

- VAE modeling is to learn the encoding function and the decoding function through the process of encoding and decoding self-reconstruction

$$\mathcal{L}(\theta, \phi; x_n) = -D_{KL}(q_\phi(z_n|x_n)||p(z_n)) + \mathbf{E}_{q_\phi(z_n|x_n)}[\log p_\theta(x_n|z_n)]$$

- Speaker representation *y* can be a pre-defined one-hot representation or a learned representation

Non-parallel VAW-GAN-based VC



s

VAW-GAN: variational autoencoding Wasserstein generative adversarial network

$$\begin{aligned} J_{vawgan} = &- \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z}_n|\boldsymbol{x}_n)\|p(\boldsymbol{z}_n)\big) \\ &+ \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})\big] \\ &+ \alpha\,\mathbb{E}_{\boldsymbol{x}\sim p_t^*}\big[\mathcal{D}_\psi(\boldsymbol{x})\big] \\ &- \alpha\,\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\mathcal{D}_\psi(\mathcal{G}_\theta(\boldsymbol{z},\boldsymbol{y}_t))\big] \end{aligned}$$

$$J_{lat}(\phi;\boldsymbol{x}) = \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p_\theta(\boldsymbol{z})\big),$$

$$J_{obs}(\phi,\theta;\boldsymbol{x},\boldsymbol{y}) = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})\big],$$

$$J_{wgan} = \mathbb{E}_{\boldsymbol{x}\sim p_t^*}\big[\mathcal{D}_\psi(\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\mathcal{D}_\psi(\mathcal{G}_\theta(\boldsymbol{z}),\boldsymbol{y}_t)\big]$$

Algorithm 1 VAE-WGAN training

function AUTOENCODE(X, y)
- $Z_\mu \leftarrow \mathcal{E}_{\phi_1}(X)$
- $Z_\sigma \leftarrow \mathcal{E}_{\phi_2}(X)$
- $Z \leftarrow$ sample from $\mathcal{N}(Z_\mu,\ Z_\sigma)$
- $X' \leftarrow \mathcal{G}_\theta(Z,\ y)$
- **return** $X',\ Z$

$\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi} \leftarrow$ initialization

while not converged **do**
- $X_s \leftarrow$ mini-batch of random samples from source
- $X_t \leftarrow$ mini-batch of random samples from target
- $X_s^{'}, Z_s \leftarrow$ AUTOENCODE($X_s,\ y_s$)
- $X_t^{'}, Z_t \leftarrow$ AUTOENCODE($X_t,\ y_t$)
- $X_{t|s} \leftarrow \mathcal{G}_\theta(Z_s,\ y_t)$
- $J_{obs} \leftarrow J_{obs}(X_s) + J_{obs}(X_t)$
- $J_{lat} \leftarrow J_{lat}(Z_s) + J_{lat}(Z_t)$
- $J_{wgan} \leftarrow J_{wgan}(X_t,\ X_s)$
- // Update the encoder, generator, and discriminator
- **while** not converged **do**
  - $\boldsymbol{\psi} \xleftarrow{update} -\nabla_\psi(-J_{wgan})$
  - $\boldsymbol{\phi} \xleftarrow{update} -\nabla_\phi(J_{obs} + J_{lat})$
  - $\boldsymbol{\theta} \xleftarrow{update} -\nabla_\theta(J_{obs} + \alpha J_{wgan})$

Results

VCC2016 corpus

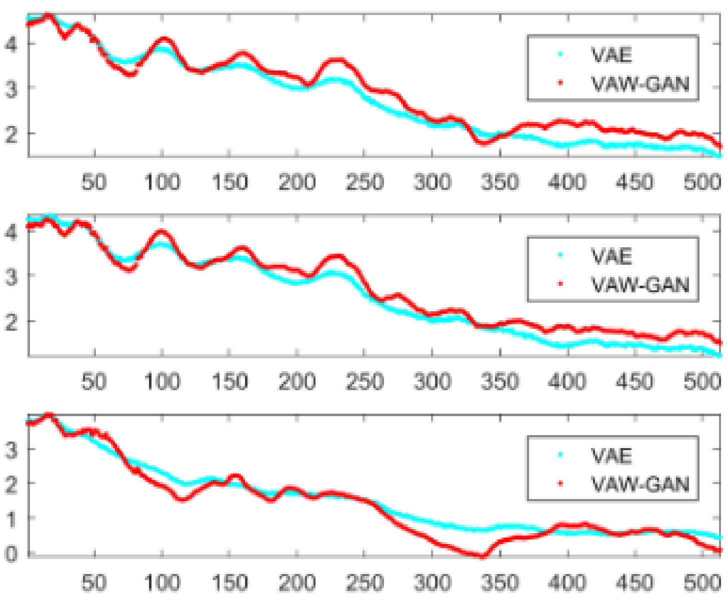


Figure 3: *Selected frames of the STRAGIHT spectra converted from SF1 to TM3. The spectral envelopes from the VAW-GAN outputs are less smooth across the frequency axis.*

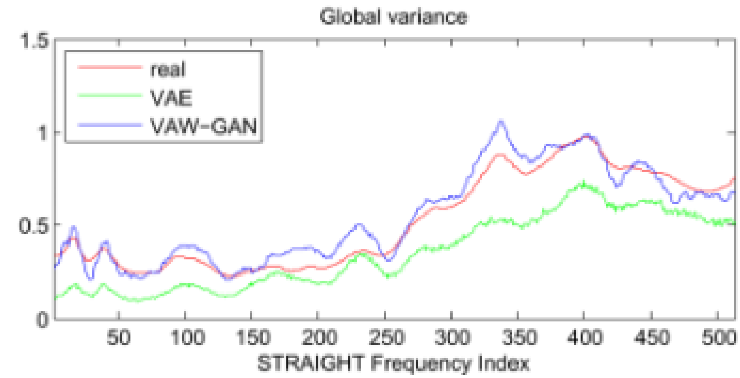


Figure 4: *Global variance computed from the $logSP_{en}$ over all non-silent frames from speaker TM3.*

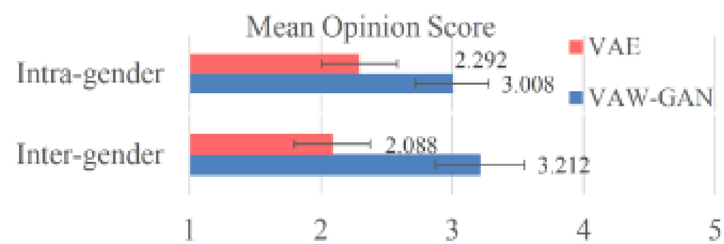
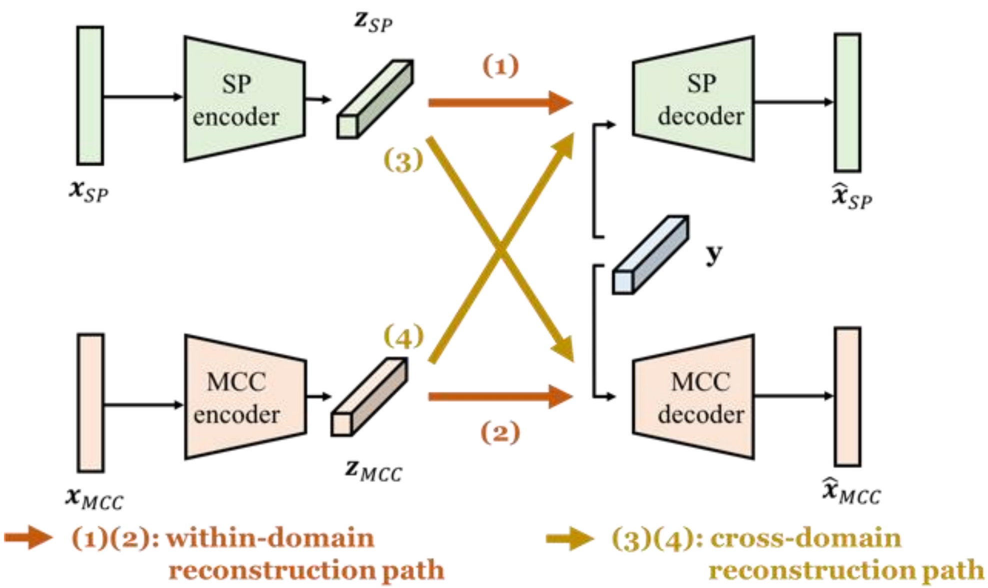
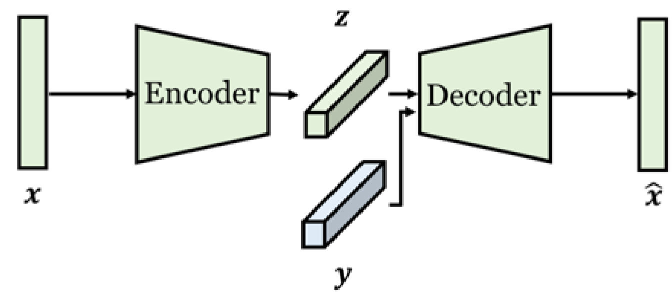


Figure 2: *MOS on naturalness. The source is SF1, and the targets are TF2 and TM3.*

Non-parallel CDVAE-based VC

CDVAE: cross-domain VAE





$$\mathcal{L}(\theta, \phi; \boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}_{recon}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}_{lat}(\boldsymbol{x}),$$

$$\mathcal{L}_{recon}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})\right],$$

$$\mathcal{L}_{lat}(\boldsymbol{x}) = -D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})),$$

$$\mathcal{L} = \mathcal{L}_{wi} + \mathcal{L}_{KLD} + \mathcal{L}_{cross} + \mathcal{L}_{sim}.$$

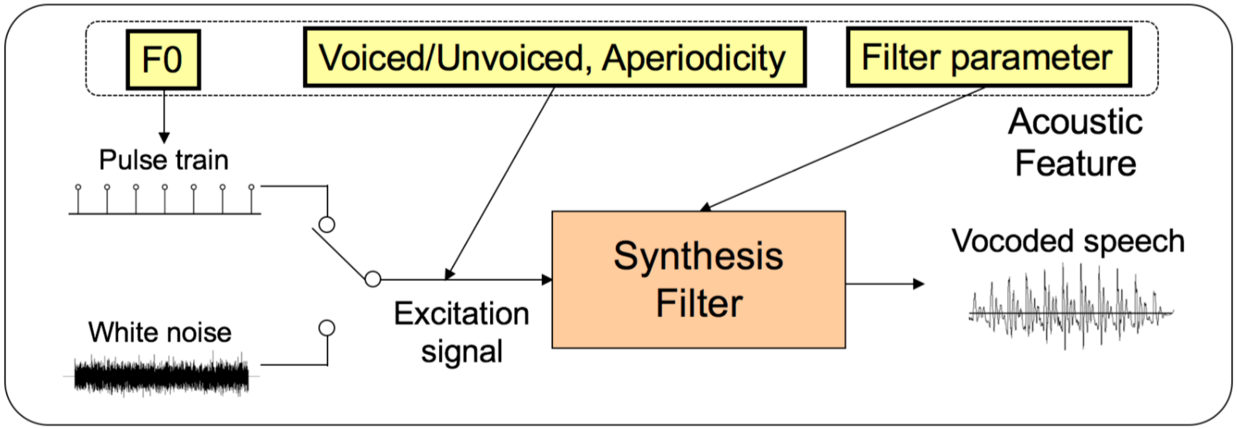
$$\mathcal{L}_{wi} = \overset{(1)}{\mathcal{L}_{recon}}(\boldsymbol{x}_{SP}, \boldsymbol{y}) + \overset{(2)}{\mathcal{L}_{recon}}(\boldsymbol{x}_{MCC}, \boldsymbol{y}),$$

$$\mathcal{L}_{KLD} = \mathcal{L}_{lat}(\boldsymbol{x}_{SP}) + \mathcal{L}_{lat}(\boldsymbol{x}_{MCC}),$$

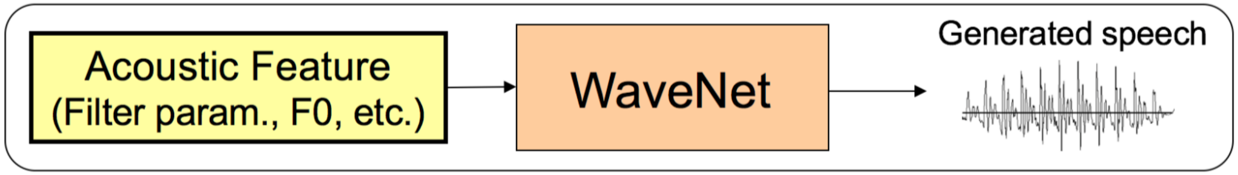
$$\mathcal{L}_{cross} = \overset{(3)}{\mathcal{L}_{recon}}(\boldsymbol{x}_{SP}, \boldsymbol{y}) + \overset{(4)}{\mathcal{L}_{recon}}(\boldsymbol{x}_{MCC}, \boldsymbol{y}).$$

$$\mathcal{L}_{sim} = \|\boldsymbol{z}_{SP} - \boldsymbol{z}_{MCC}\|_1.$$

Conventional Vocoder vs. WaveNet Vocoder



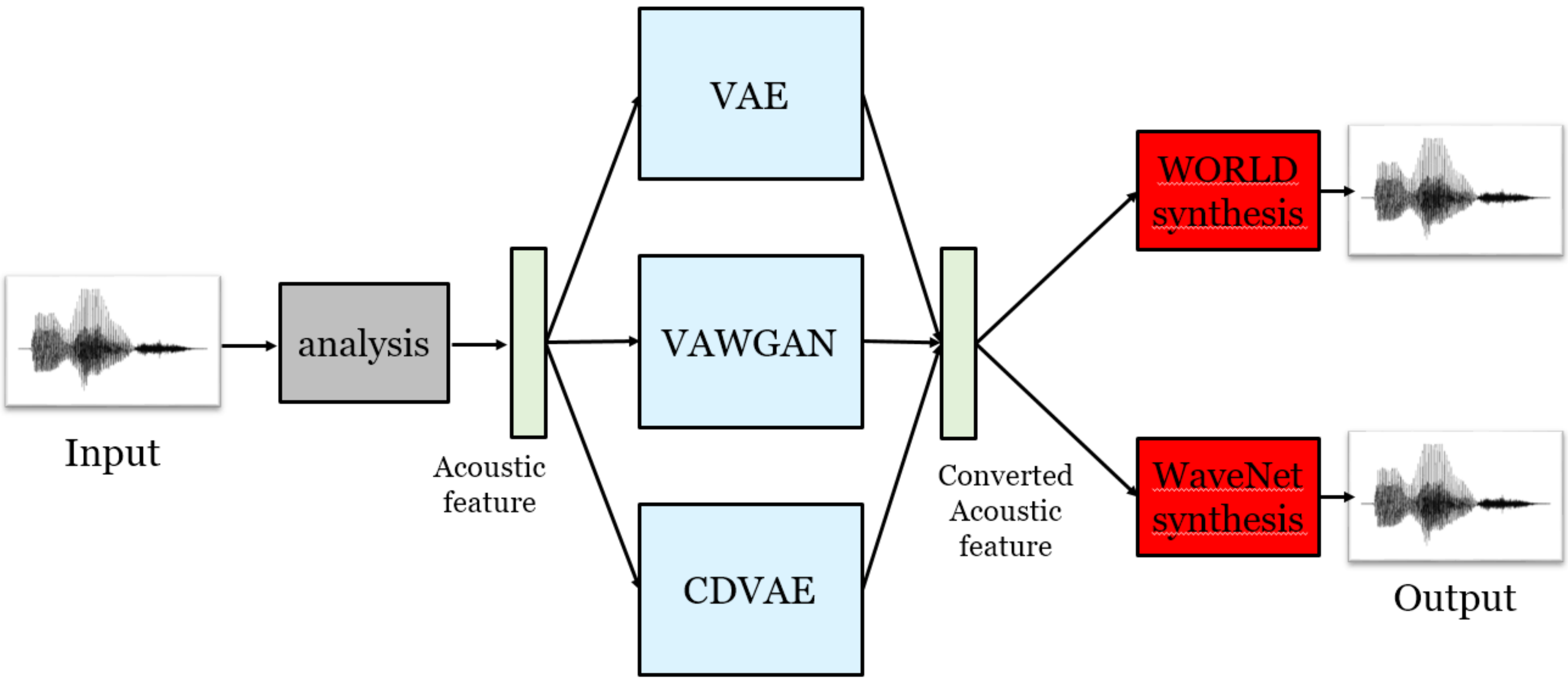
(a) *Conventional Vocoder [17]*



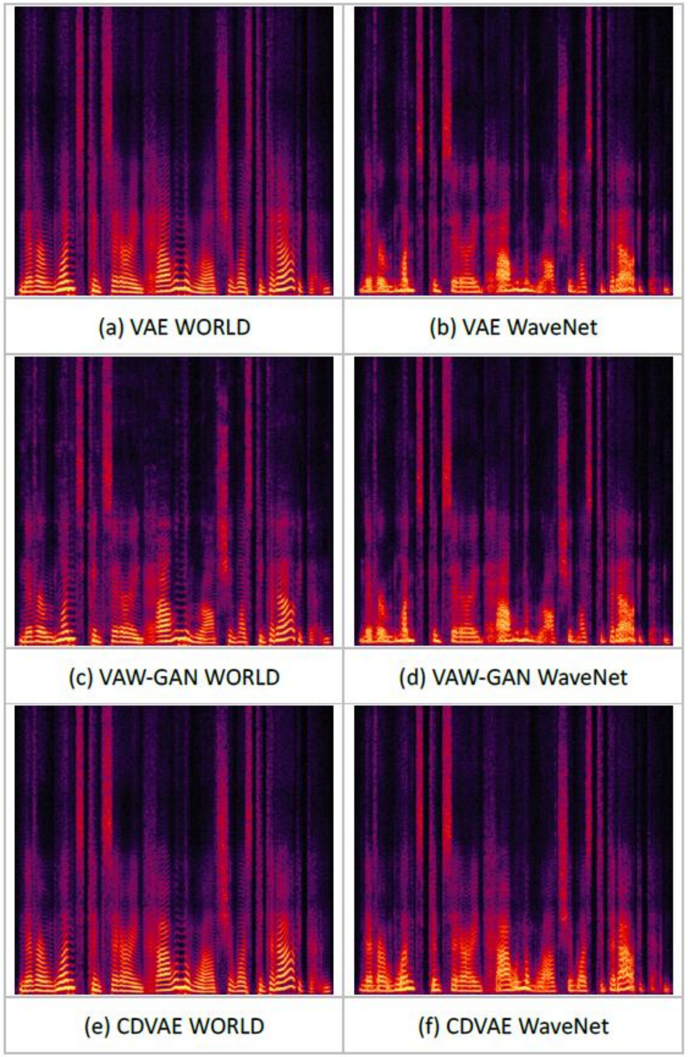
(b) *Proposed*

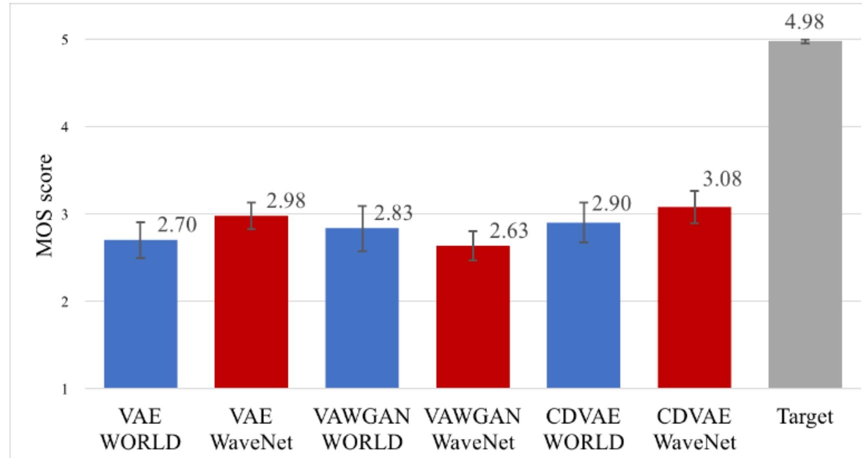
Tamamori et al. Interspeech2017

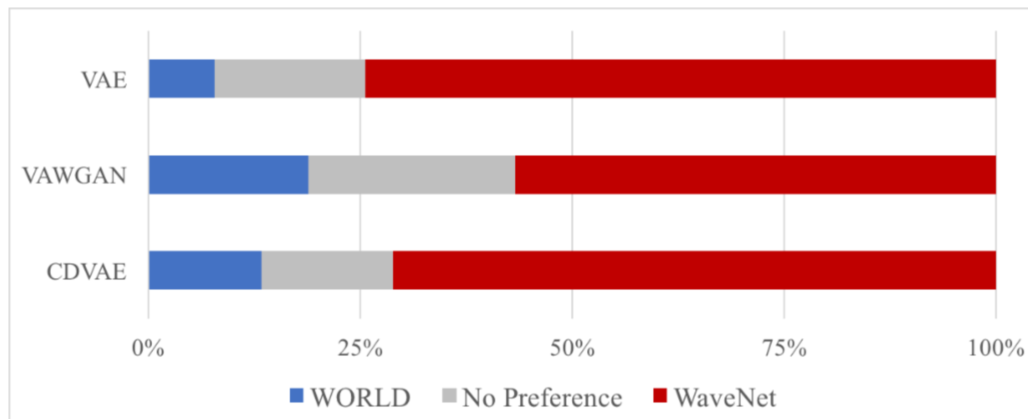
Comparison of VC Systems and Vocoders



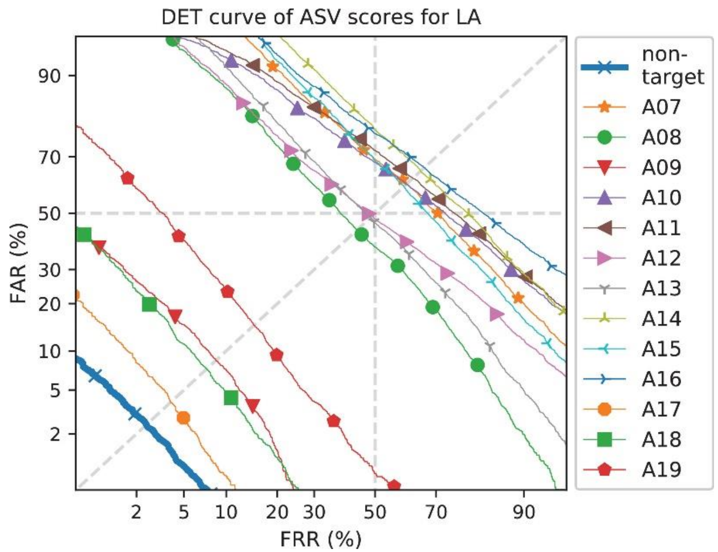
Results


(a) VAE WORLD
(b) VAE WaveNet
(c) VAW-GAN WORLD
(d) VAW-GAN WaveNet
(e) CDVAE WORLD
(f) CDVAE WaveNet



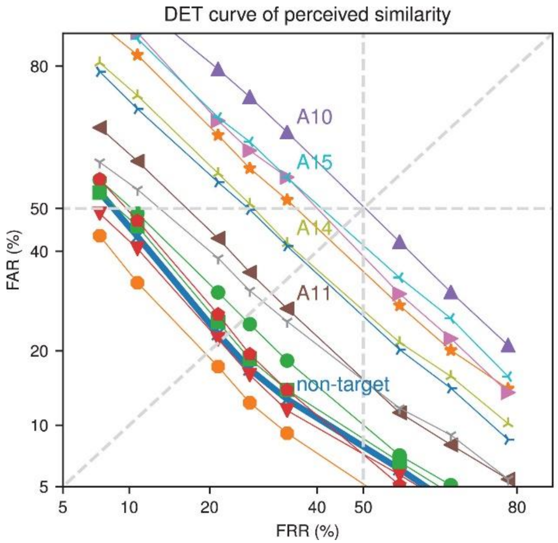


Results

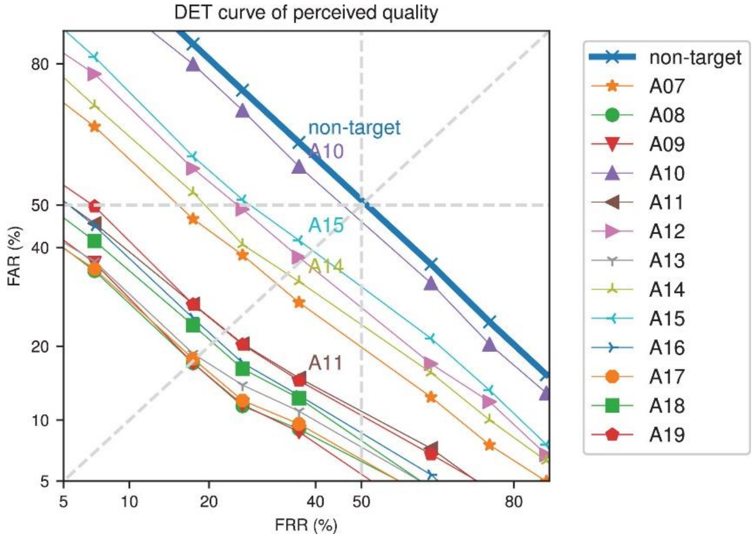


ASV results

Neural Vocoder: A10, A12, A15

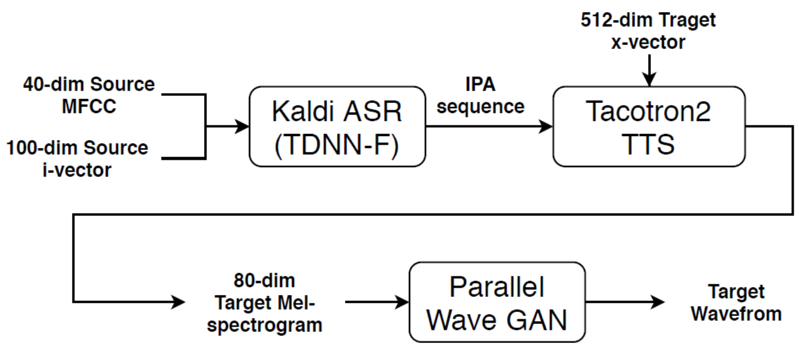


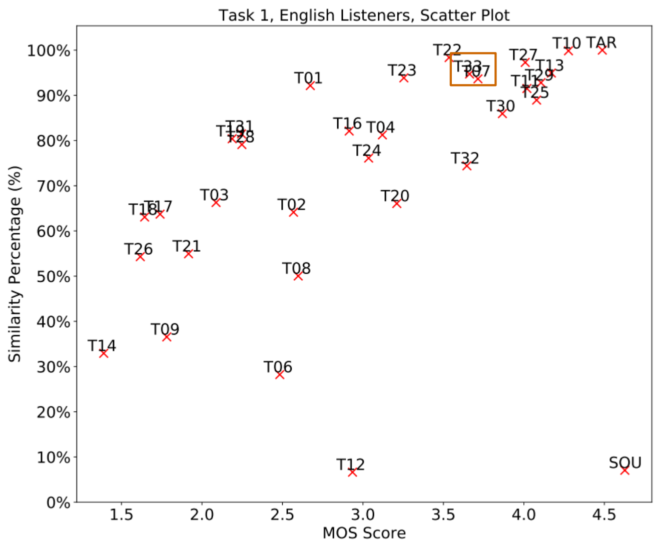
Human listening results (Similarity)



Human listening results (Quality)

Voice Conversion (ASR + TTS)





VC Challenge 2020

T26: One shot VC Griffin-Lim; One shot VC Griffin-Lim
T10: ASR-TTS (Transformer) / PPG-VC (LSTM) WaveNet; PPG-VC (LSTM) WaveNet

Outline

- Spoofing Attack and Automatic Speaker Verification
- Spoofing Attacks Methods
  - Replay attack
  - Impersonation (twins and siblings)
  - Cut and paste
  - Voice conversion (text-to-speech)
  - Acoustic scene conversion