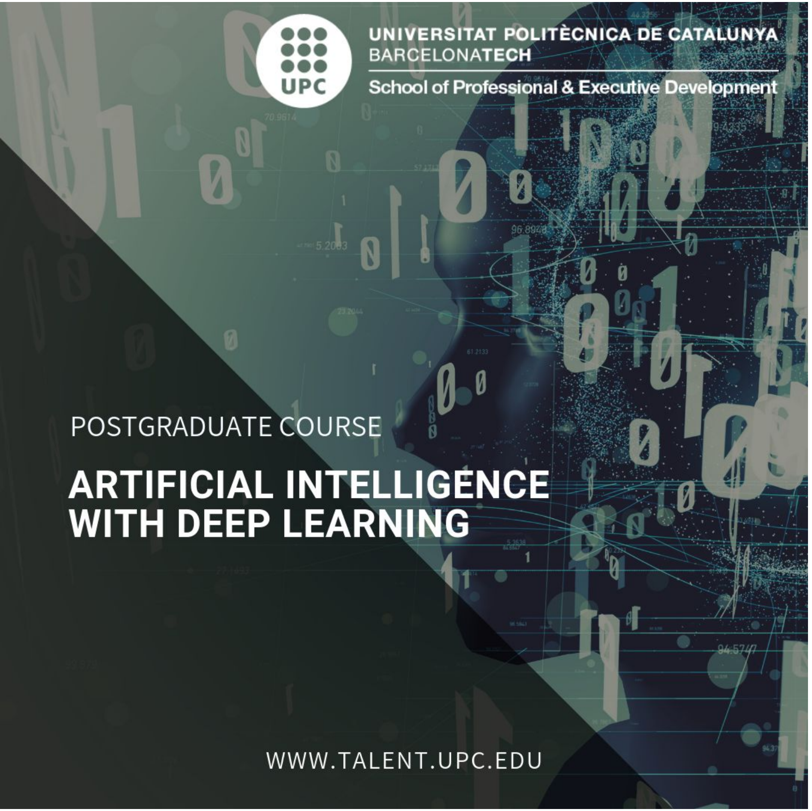
#DLUPC



Speech to speech paradigms

Carlos Segura Perales
carlos.seguraperales@telefonica.com

Scientific Research
Telefónica Research
Telefónica I+D

Telefónica
Research

Acknowledgments

Santiago Pascual, Universitat Politècnica de Catalunya, slides

Outline

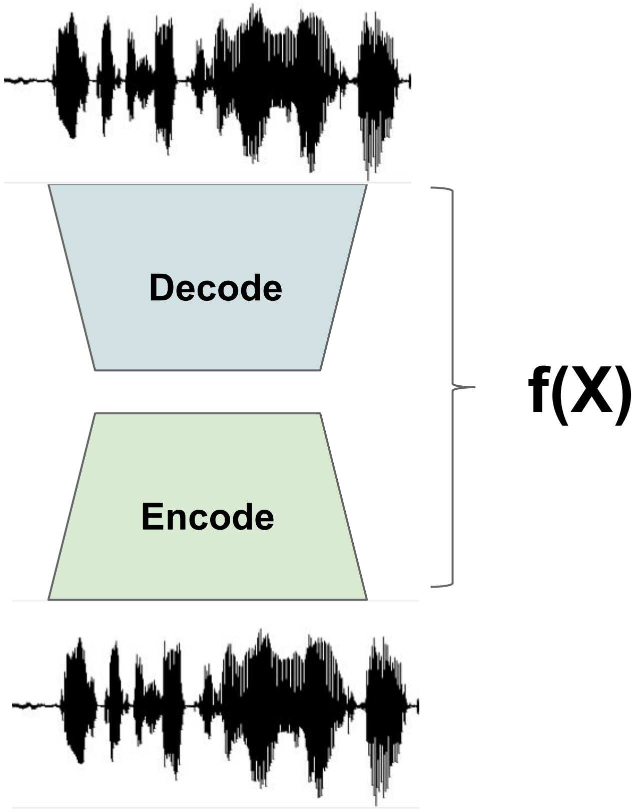
1. Introduction
2. Encoder-Decoder Paradigms
    a. Generative modeling
3. Speech Enhancement
    a. Discriminative Procedure
    b. SEGAN/FSEGAN
4. Voice Conversion

Introduction

Speech to speech

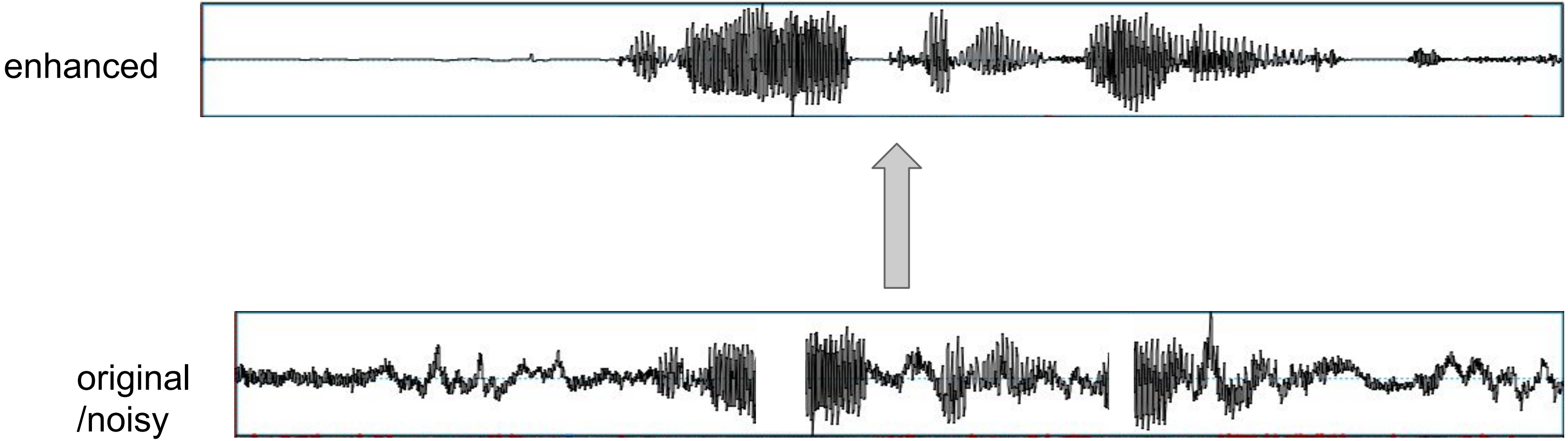
Speech is transformed through a non-linear function Y = f(X):

- Enhance/Denoise signal
- Convert content respecting identity
  - Translation
- Convert identity respecting content
  - Voice Conversion



Speech Enhancement/Denoising

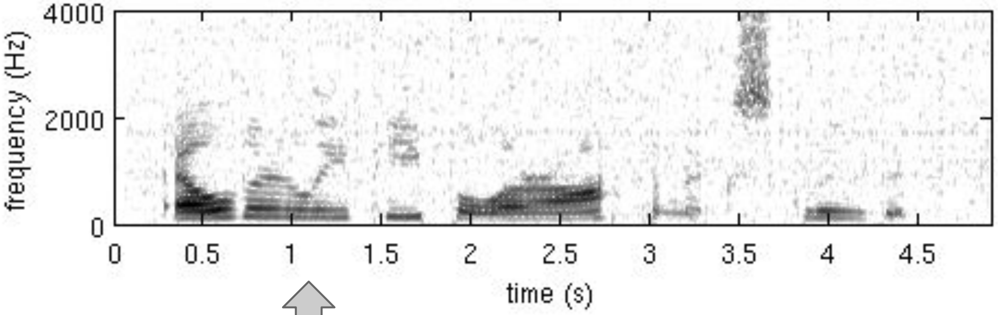
Recover lost information or add enhancing details by learning the natural distribution of audio samples.



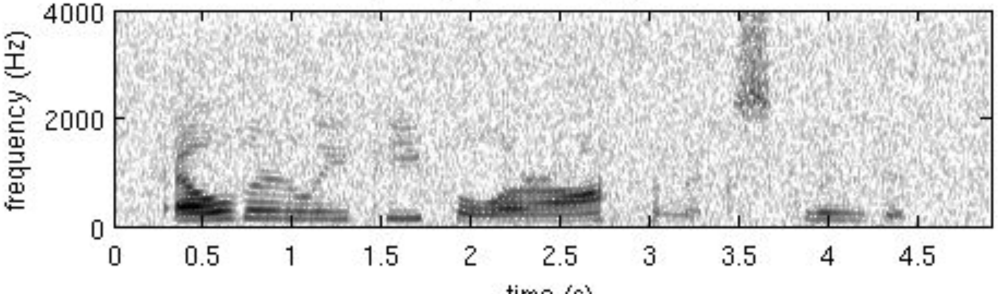
Speech Enhancement/Denoising

Denoise spectral features

enhanced



original
/noisy



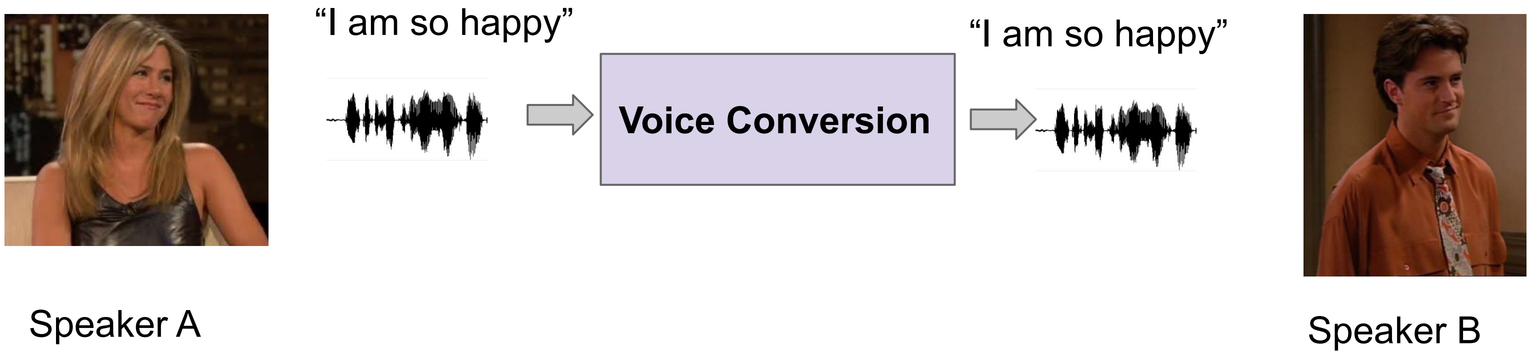
Speech Enhancement

Applicable to many scenarios:

- Improving automatic speech recognition (ASR).
- Improve intelligibility in complex communication scenarios (like airplanes).
- For hearing aid implants.
- Enhance low quality recordings in speech synthesis data to train a system.

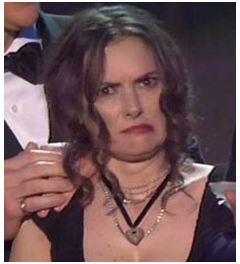
Voice Conversion

Transfer the spoken contents and style from one speaker A to another speaker B.



Voice Conversion

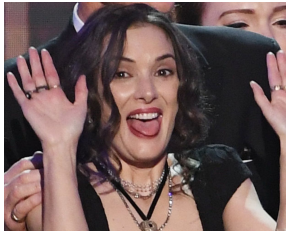
Also: transfer the spoken contents and style from within same speaker identity.



“We won...”

Voice Conversion

“We won!”



Speaker A

Speaker A

Voice Conversion

Potential Applications:

- Technologies to help people with motor speech disorders like dysarthria.
- Additional flexible block to speech synthesis systems, where we can enforce emotions and prosody changes.
- Dubbing industry. Human speech contains a set of expressive and natural patterns that are hard to obtain directly from text like in TTS.

Risks:

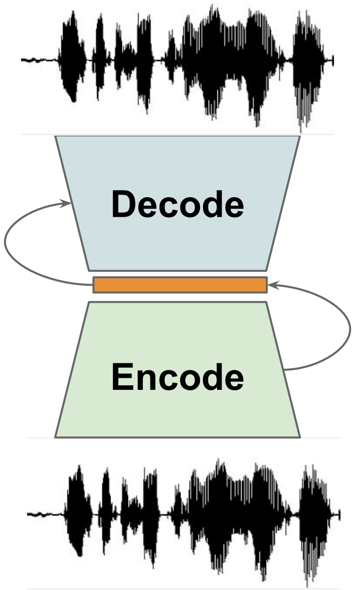
- Identity Spoofing

Encoder-Decoder Paradigms

Encoder-Decoder paradigm

These speech2speech systems typically work under an encoder-decoder framework:

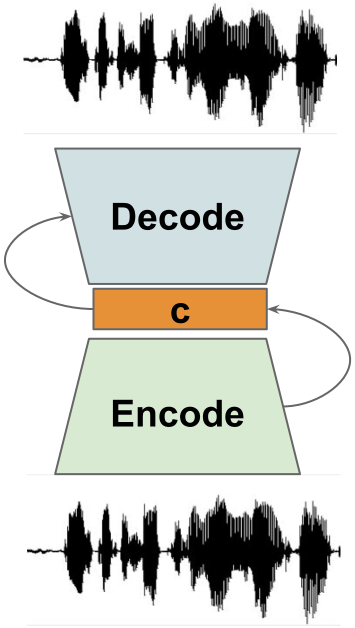
- Build an intermediate representation that captures latent characteristics of the spoken utterance.
- Reconstruct the signal with the proper new features.



Vanilla AutoEncoders

- Encoder mapping **c = E(x)** is deterministic, as well as code vector **c**.

- Decoder mapping reconstructs **x** into a plausible version **x^** deterministically.

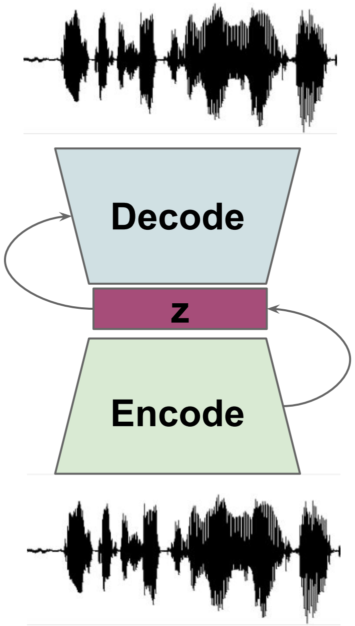


Variational AutoEncoders

(Kingma and Welling, 2014)

- Encoder mapping **z = E(x)** is deterministic, but we apply restrictions on **Z** space, so that it follows a prior probability density, like isotropic Normal one: N(0, I).

- Decoder mapping reconstructs a sampled **z** into a plausible version **x^** deterministically.



NOTE: Working directly with waveforms is a very recent thing (2 year at most), and one of the most challenging parts of deep speech2speech systems.

VQ-VAE

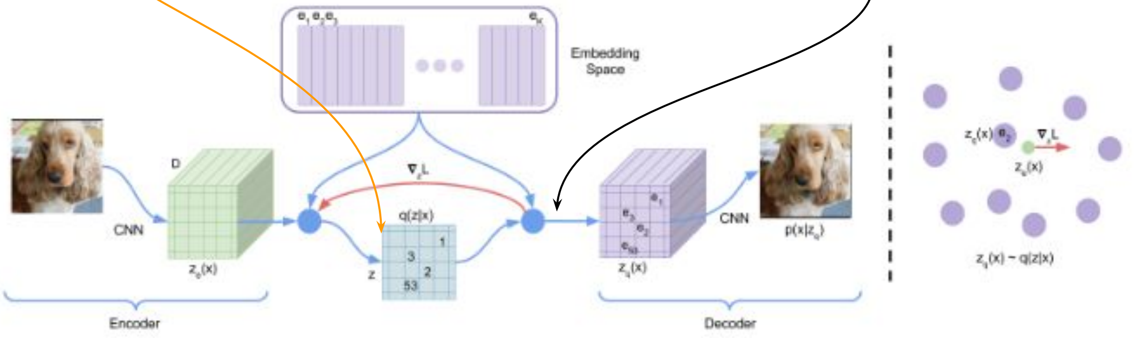
([Van den Oord et al. 2017](#))

- **Z** space is a discretized embedding space, so every encoded point z(**x**) is mapped to nearest embedding **e**, which is the information given to decode the sample.

$$q(z=k|x)=\begin{cases}1 & \text{for } \mathrm{k}=\mathrm{argmin}_j\|z_e(x)-e_j\|_2,\\ 0 & \text{otherwise}\end{cases},$$

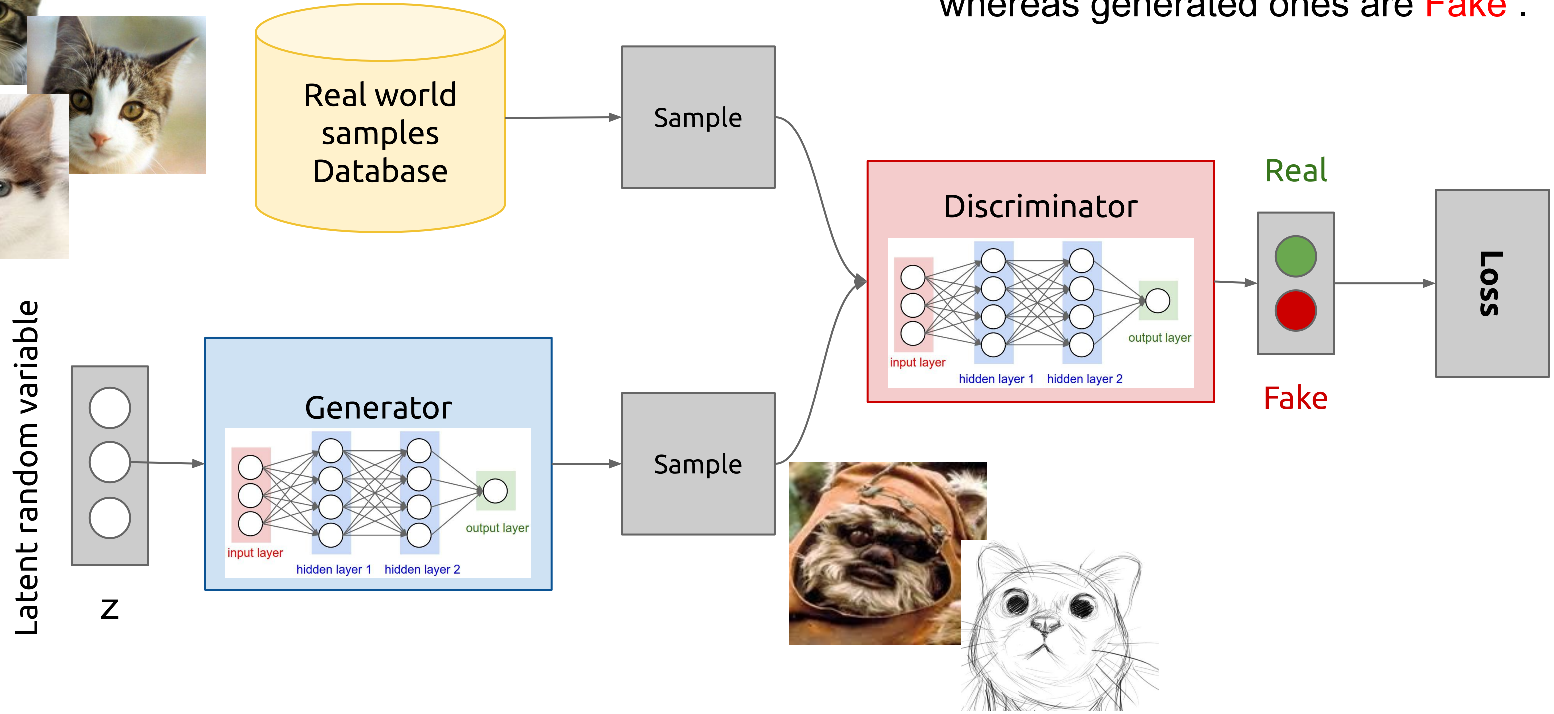
$$e \in R^{K\times D}$$

$$z_q(x)=e_k, \quad \text{where} \quad k=\mathrm{argmin}_j\|z_e(x)-e_j\|_2$$



Generative Adversarial Networks

D determines database images are Real, whereas generated ones are Fake .

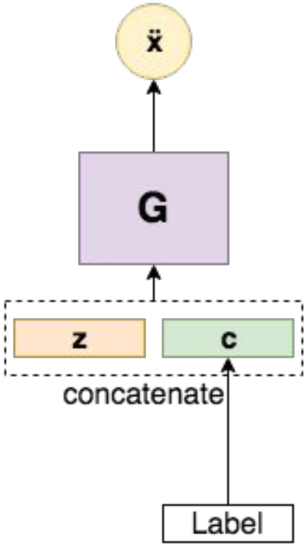


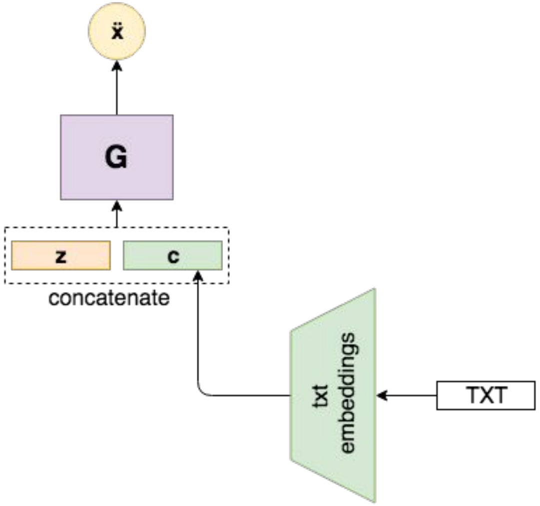
Conditional GANs

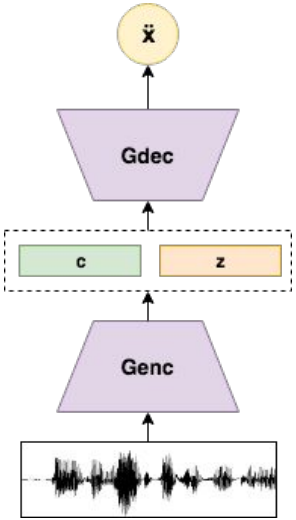
For details on ways to condition GANs:
[Ways of Conditioning Generative Adversarial Networks (Wack et al.)]

GANs can be conditioned on other info extra to **z**: text, labels, speech, etc..

z might capture random characteristics of the data (variabilities of plausible futures), whilst **c** would condition the deterministic parts !





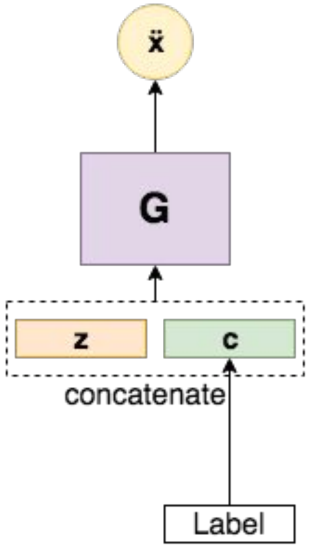


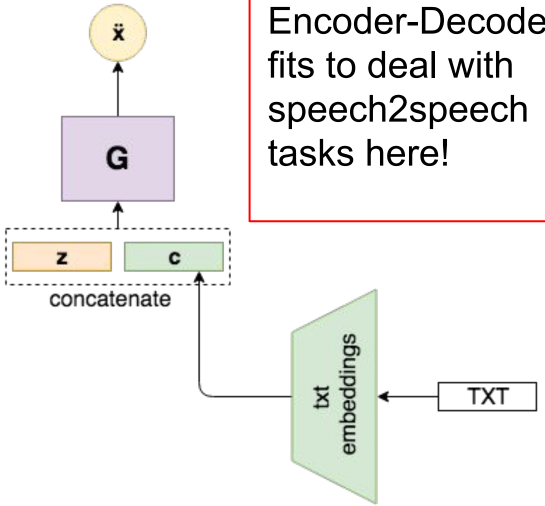
Conditional GANs

For details on ways to condition GANs:
[Ways of Conditioning Generative Adversarial Networks (Wack et al.)]

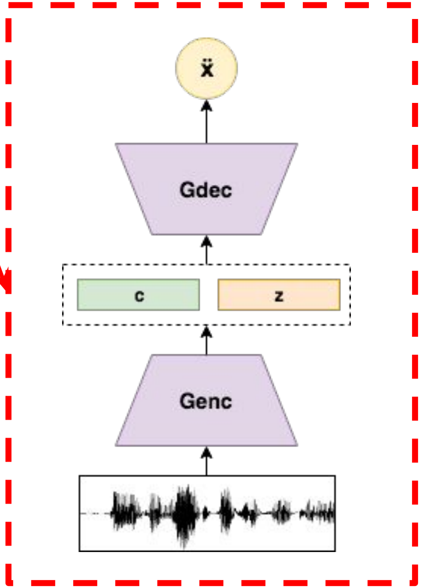
GANs can be conditioned on other info extra to **z**: text, labels, speech, etc..

z might capture random characteristics of the data (variabilities of plausible futures), whilst **c** would condition the deterministic parts !





Encoder-Decoder fits to deal with speech2speech tasks here!



CycleGAN

[Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, Jun-Yan Zhu et al. 2017](#)

[Samples](#)

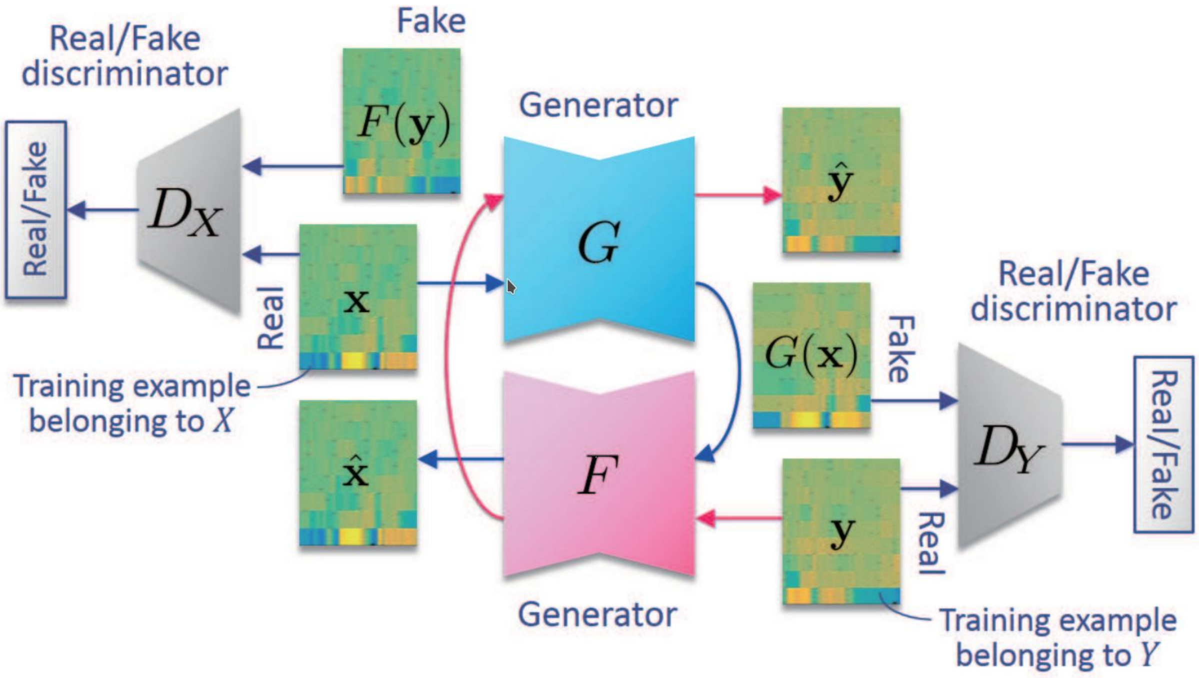


Figure credit: Hirokazu Kameoka, [source](#)

StarGAN

[StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation, Yunjey Choi et al. 2018](#)

[Samples](#)

- Cycle loss + Classification loss + Discrimination loss

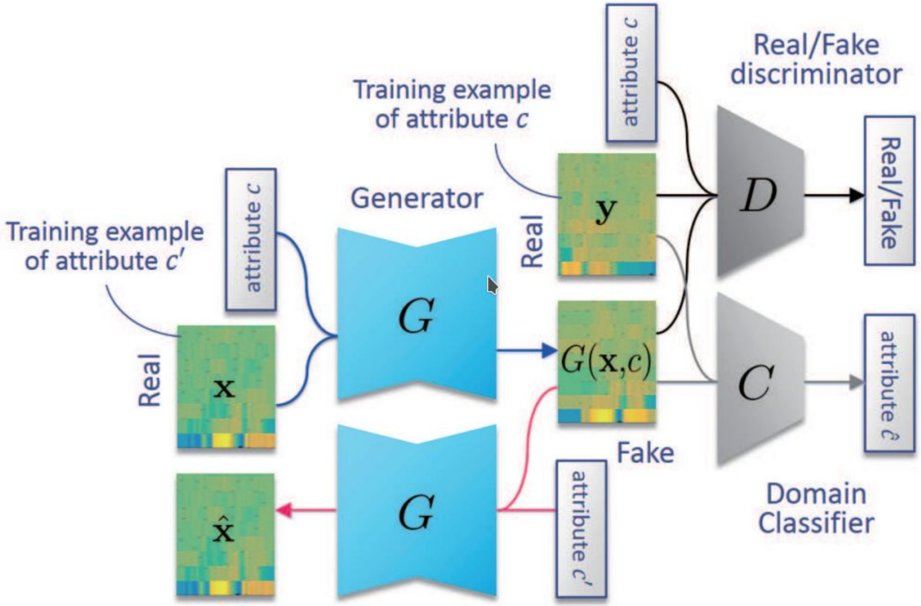


Figure credit: Hirokazu Kameoka, [source](#)

Voice Conversion

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

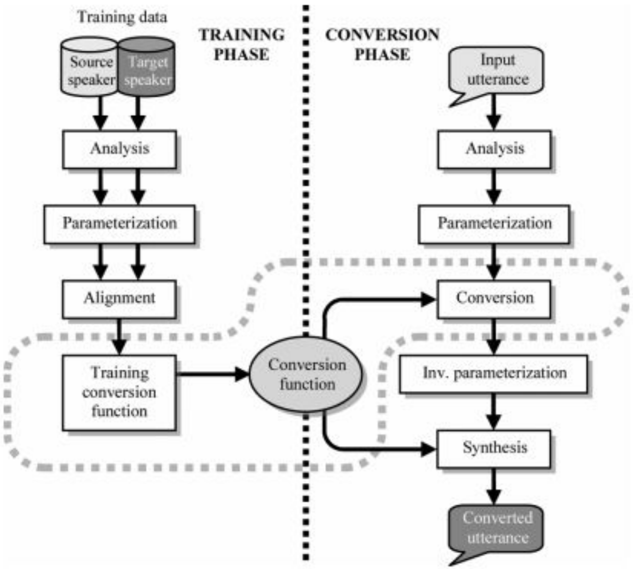


Figure credit: Daniel Erro

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(1) Spectral features are extracted

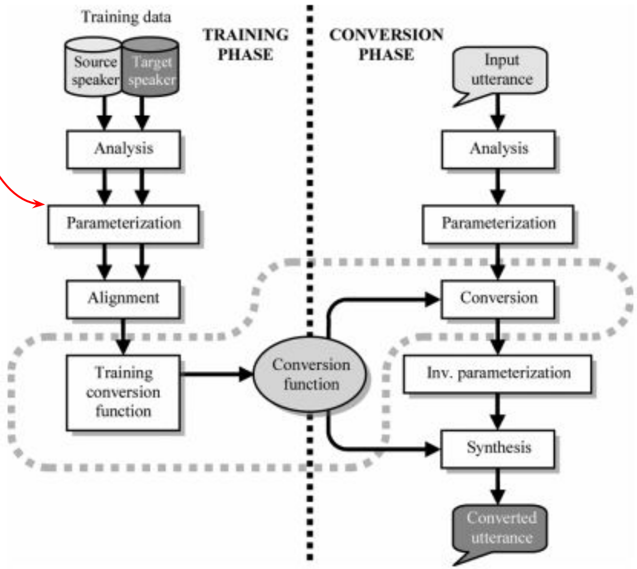


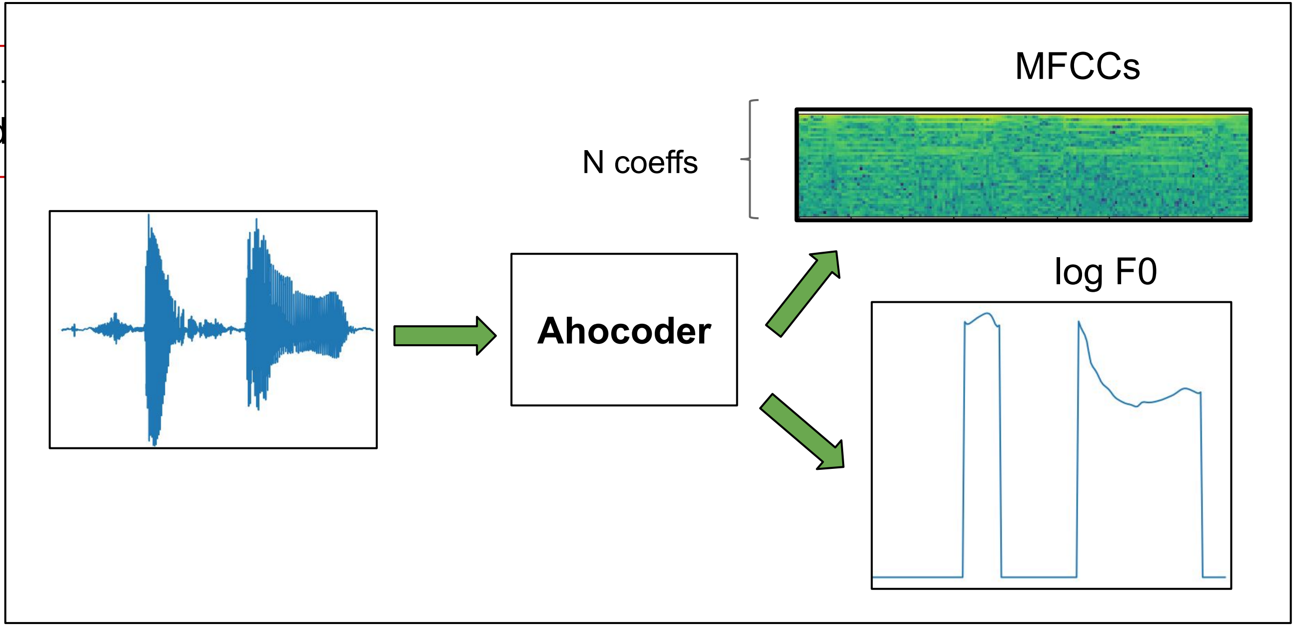
Figure credit: Daniel Erro

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(1) Spectral
extracted



TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(2) Alignment process in training data: Dynamic Time Warping

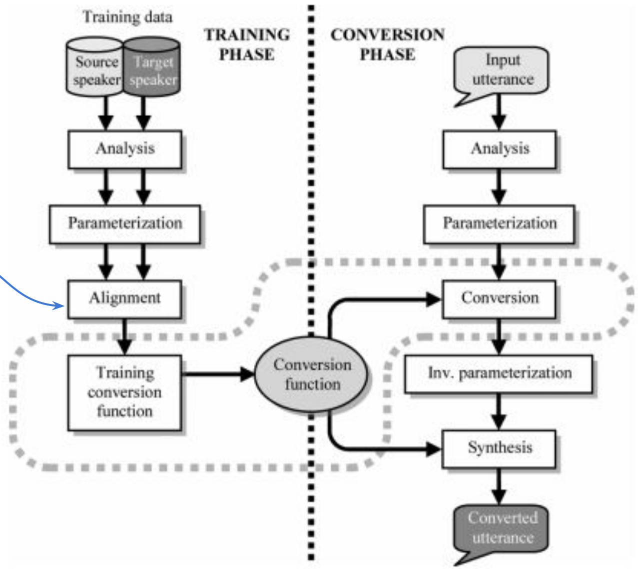


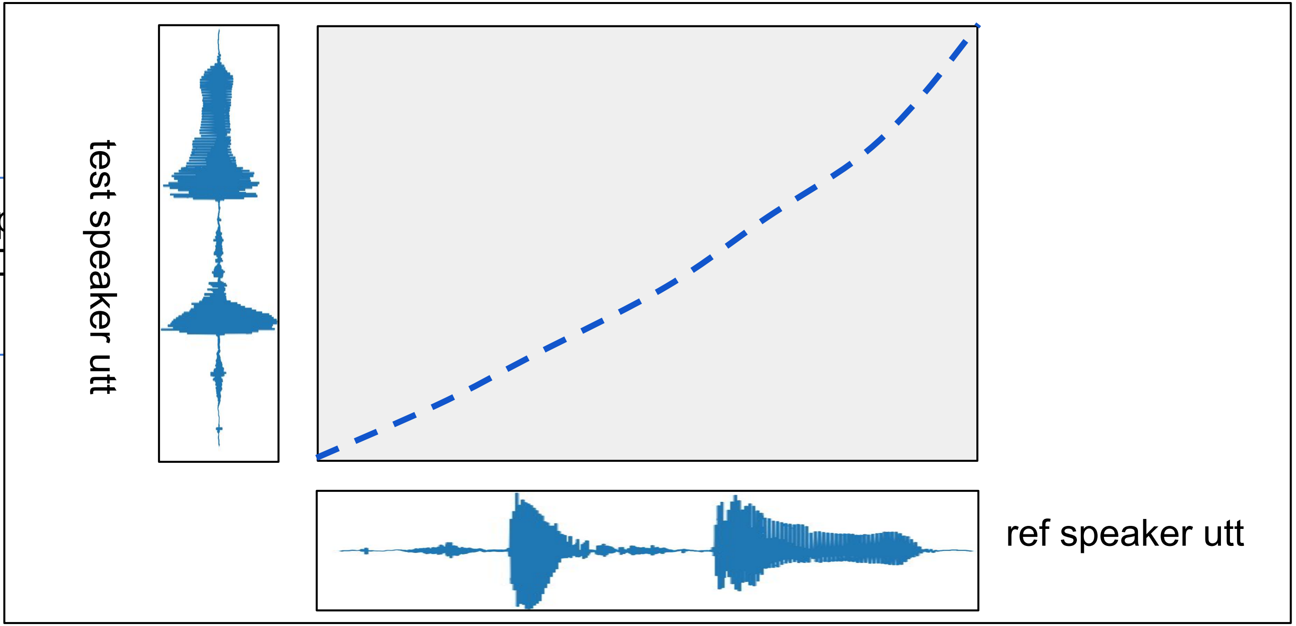
Figure credit: Daniel Erro

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(2) Alignment p
training data: D
Time Warping



TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a system to perform the mapping

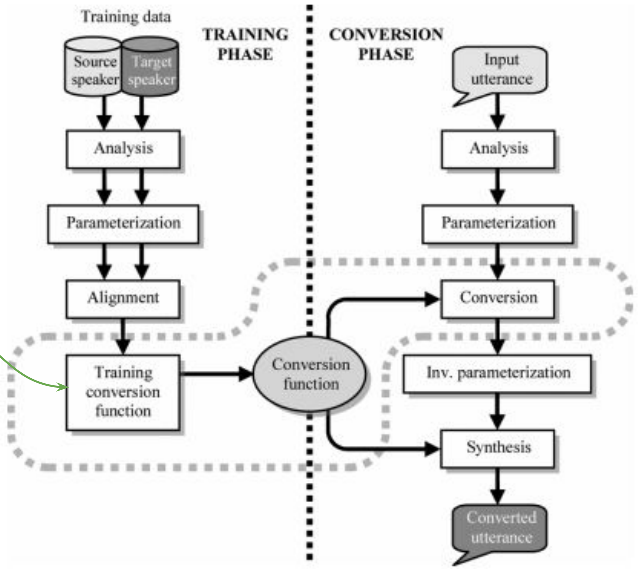


Figure credit: Daniel Erro

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a
perform th

Pitch can be linearly converted, pre-calculating both speakers' (source and target) statistical moments (mean and variance) among sliding window frames in training set:

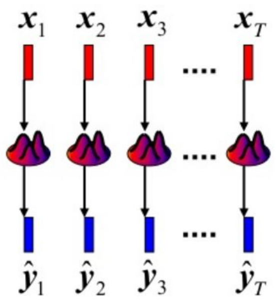
$$log(f0_{conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(log(f0_{src}) - \mu_{src})$$

Figure credit: Daniel Erro

TRAIN

Parallel corpora and frame-wise VC

Convert speech features frame by frame independently



TRAIN

Frame-based conversion function

$$\hat{y}_t = f_\lambda(x_t)$$

Figure credit: Tomoki Toda

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a
perform th

Spectral mapping: pick a powerful frame by frame non-linear model:

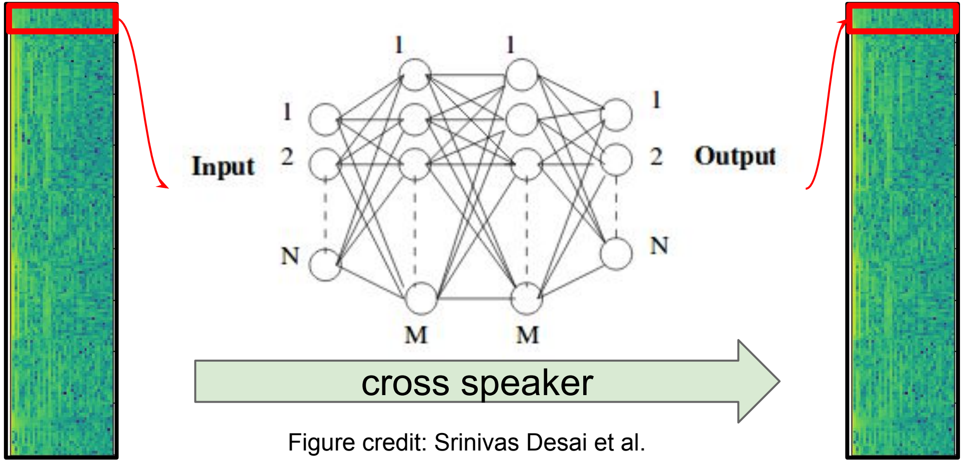


Figure credit: Srinivas Desai et al.

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a
perform th

Spectral mapping: pick a powerful frame by frame non-linear model:

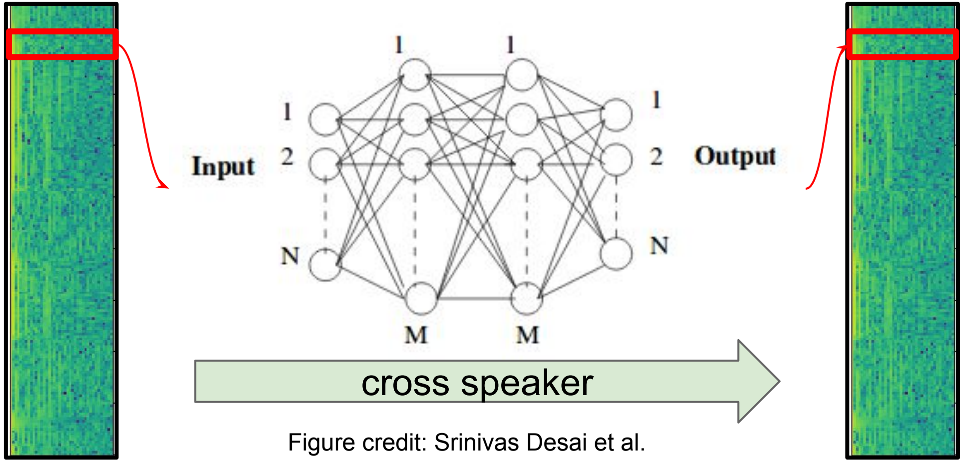


Figure credit: Srinivas Desai et al.

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a
perform th

Spectral mapping: pick a powerful frame by frame non-linear model:

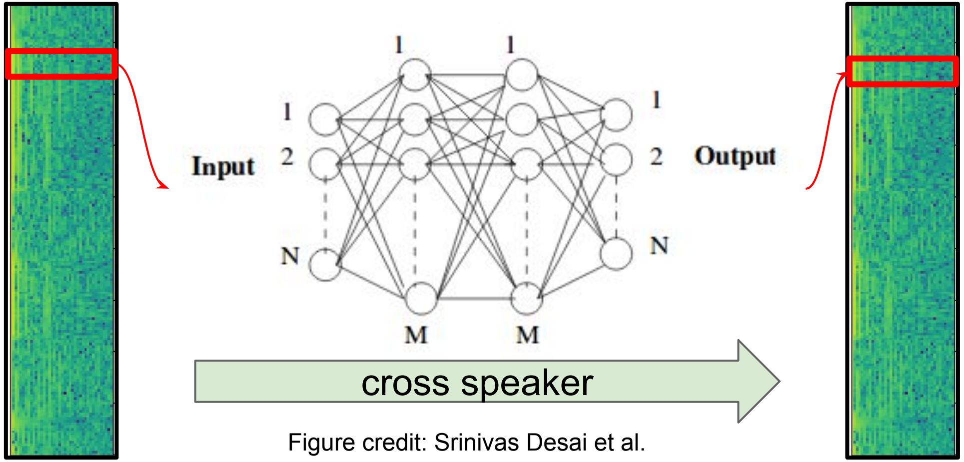


Figure credit: Srinivas Desai et al.

TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:

(3) Train a
perform th

Spectral mapping: pick a powerful frame by frame no ear model:

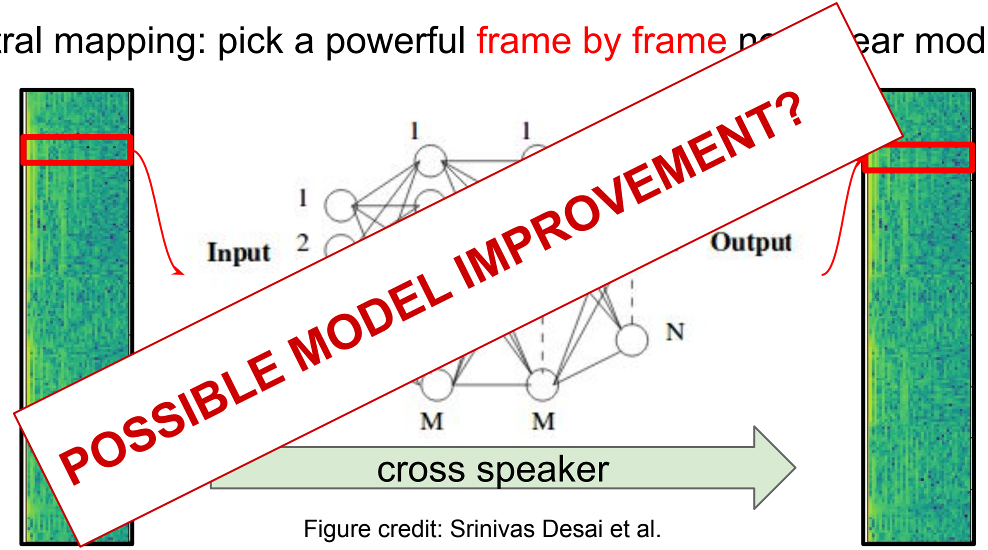


Figure credit: Srinivas Desai et al.

TRAIN

Parallel corpora and frame-wise VC

Conversion considering inter-frame correlation over an utterance to properly model speech dynamics

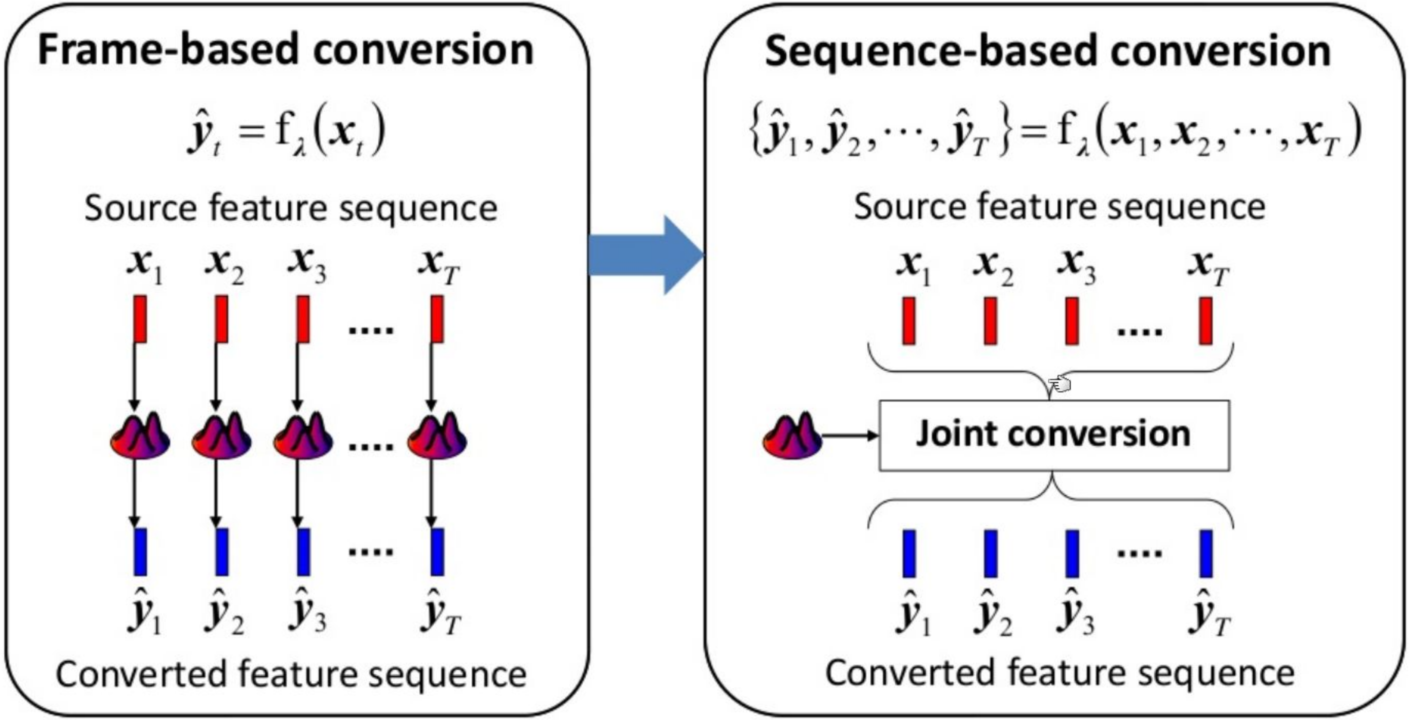
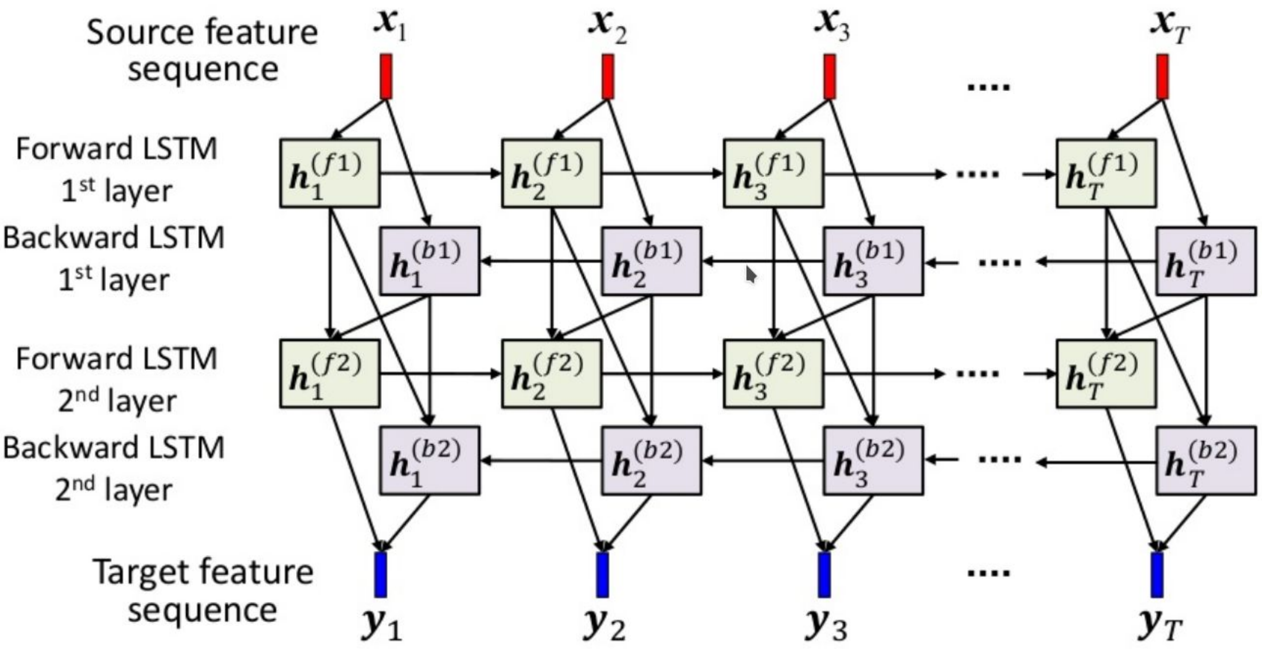


Figure credit: Tomoki Toda

Parallel corpora and frame-wise VC (Sun et al. 2015)

Spectral mapping: pick a powerful frame by frame non-linear model with memory



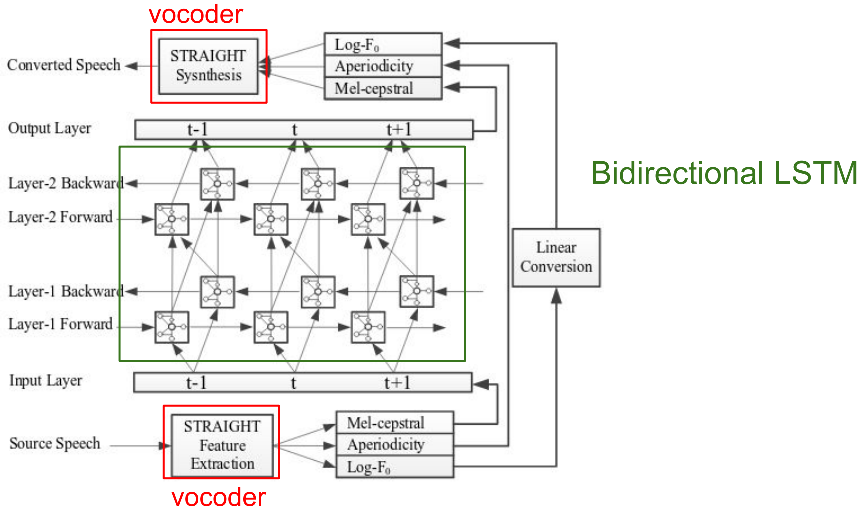
TRAIN

Parallel corpora and frame-wise VC [(Sun et al. 2015)]

General VC pipeline with Discriminative model:

(3) Train a
perform th

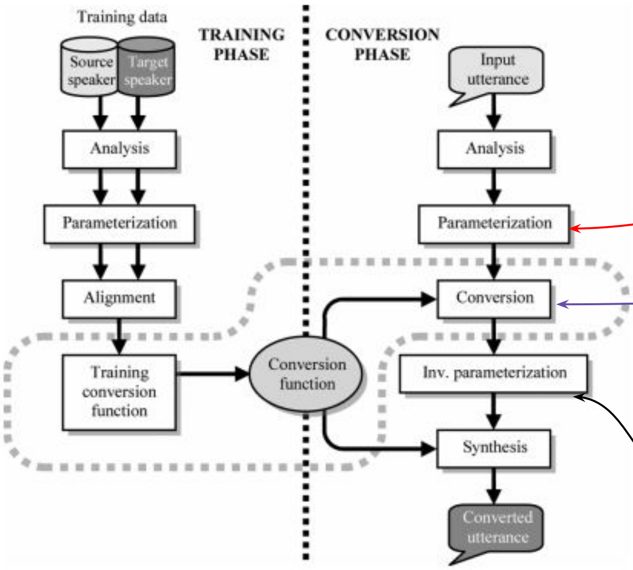
Use bi-directional LSTMs to convert a source feature sequence to a target sequence



TRAIN

Parallel corpora and frame-wise VC

General VC pipeline with Discriminative model:



(1) Spectral features are extracted

(2) Infer output features from input features

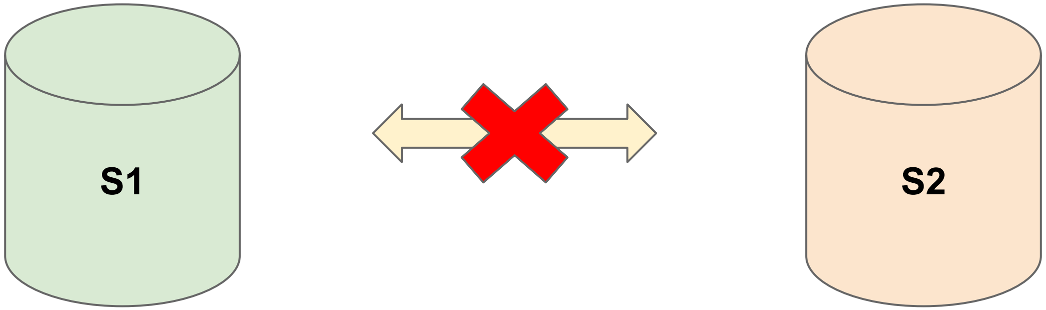
(3) Invert parameterization (decode features)

INFERENCE

Figure credit: Daniel Erro

Unaligned corpora

- Speakers do NOT say the same, so there's no content to align.
- Speakers can even speak in different languages!



Challenging transferability problem: no supervised discriminative approach

VAE based VC

(Hsu et al. 2016)

We can take advantage of Variational Auto-Encoder training procedure to learn latent representations of speakers, and a deterministic identity code will map all back to destination acoustic space.

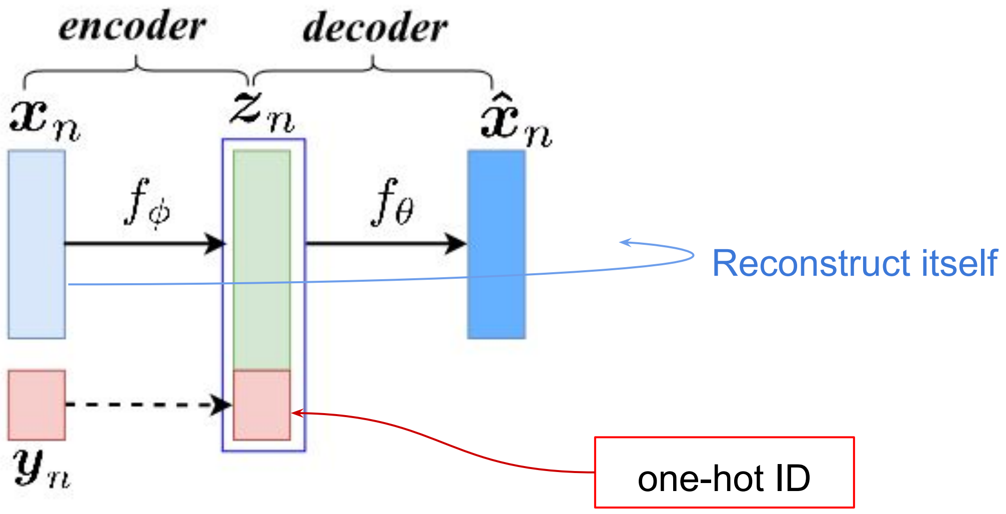
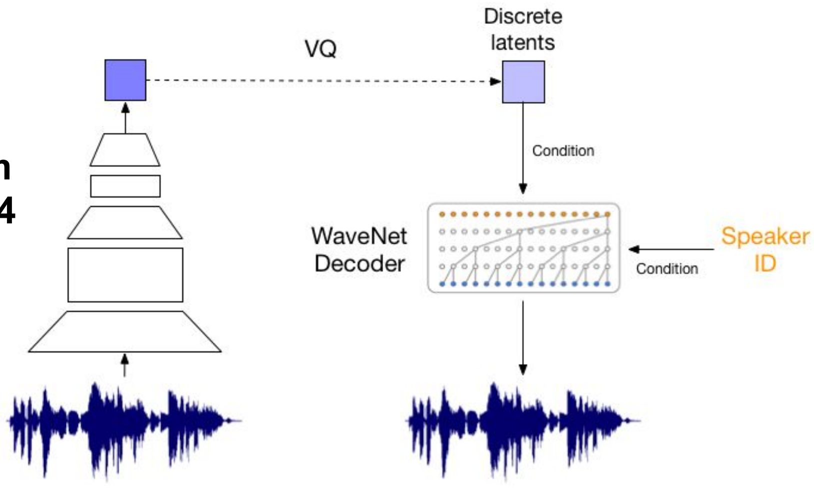


Figure credit: Hsu et al.

Vector Quantised-VAE (end to end) [samples]

Latest most successful and natural sounding approach has been VQ-VAE by Google DeepMind. They build a discrete latent space that resembles a phoneset unsupervisedly! A **Wavenet** decodes the latent codes **conditioned on one-hot ID**.

Encode waveform decimating by x64 factor



Auto-regressive decoding with WaveNet

Figure credit: Äaron van den Oord

StarGAN-VC

[StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks, Hirokazu Kameoka et al. 2018](#)

[Samples](#)

Speakers correspond to different domains

Works on spectral features

Downsample:

- Conv2d
- Batchnorm
- Gated Linear Units

Upsample:

- Deconv2d
- Batchnorm
- GLU

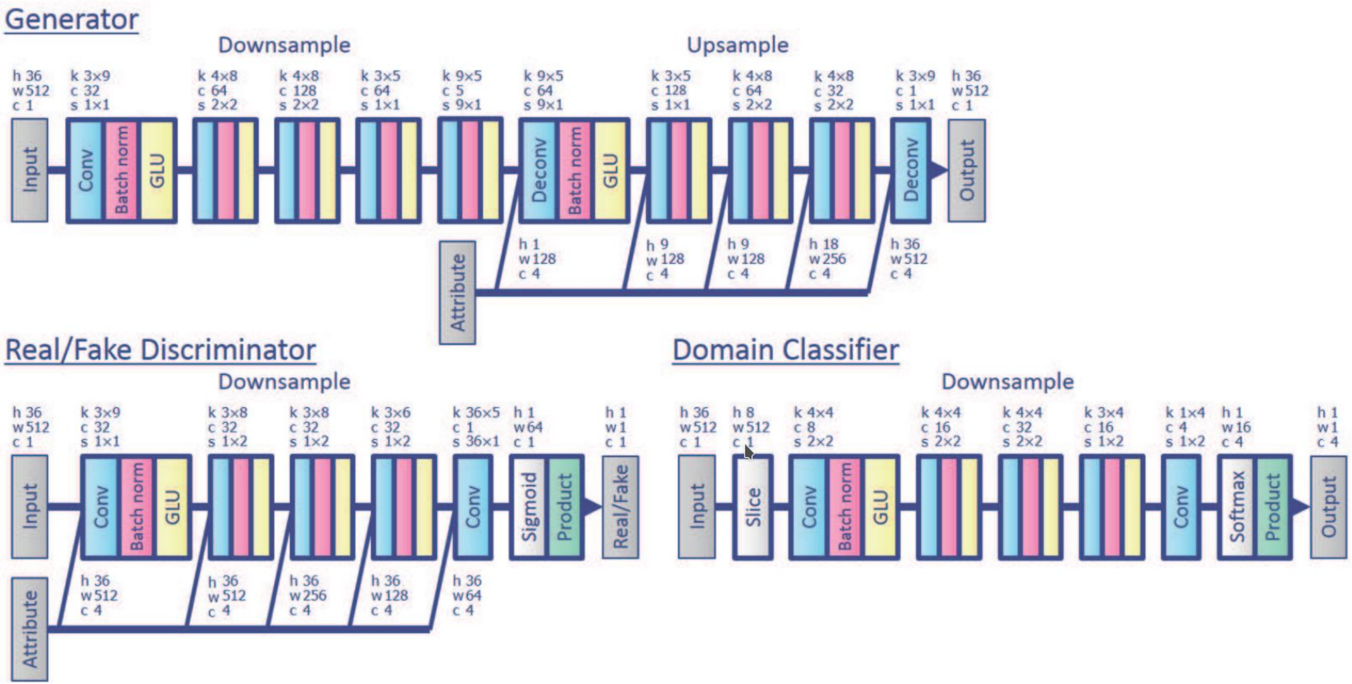


Figure credit: Hirokazu Kameoka, [source](#)

VC Evaluation

- Typically subjective evaluation: like Mean Opinion Score (MOS) [1, 5] pooling a group of listeners opinions’ in terms of (1) naturalness and (2) similarity to target.
- Objective metrics for specific features (e.g. Mel Cepstral Distortion [dB] for MFCCs, or RMSE [Hz] for pitch can serve as a guidance, but not as a final decision).

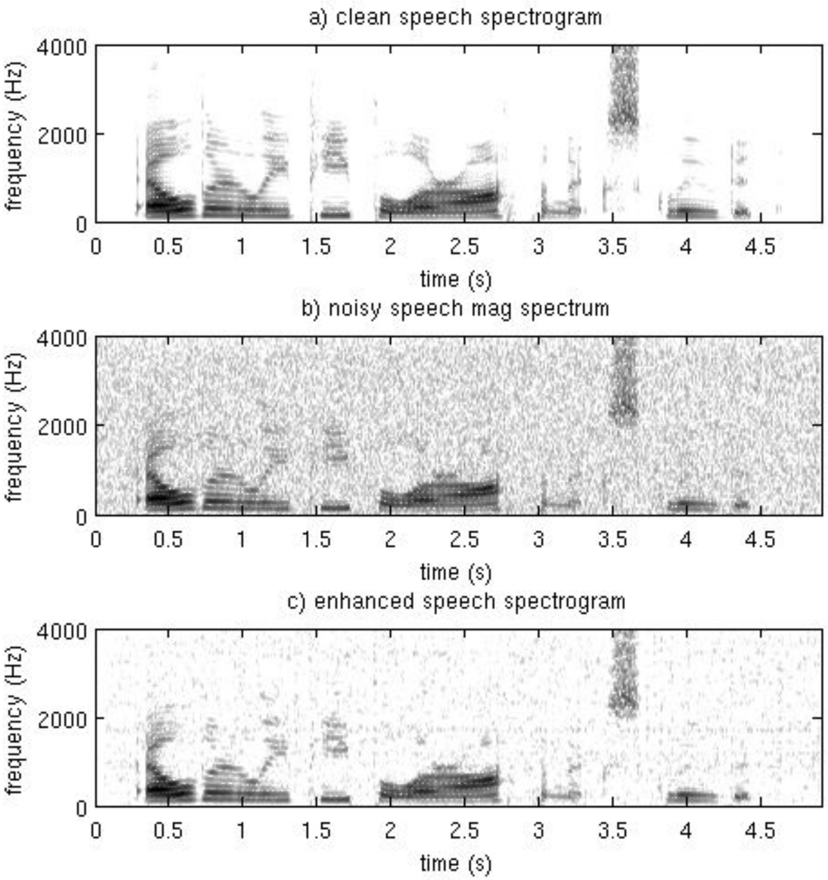
Speech Enhancement

SE Approaches

- Spectral subtraction: estimate noise activity during non-speech regions and subtract it.

SE Approaches

- Spectral subtract ... ch regions and subtract it.



SE Approaches

- Spectral subtraction: estimate noise activity during non-speech regions and subtract it.
- Subspace algorithms: decompose the higher dimensional noisy signal into a lower dimensional one where clean version lays.

$$\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}$$

$\hat{\mathbf{x}}$ Enhanced signal

$\mathbf{y}$ Noisy signal

$$\boldsymbol{\varepsilon} = \hat{\mathbf{x}} - \mathbf{x} = \mathbf{H}\mathbf{y} - \mathbf{x} = (\mathbf{H} - \mathbf{I})\mathbf{x} + \mathbf{H}\mathbf{d}$$

$$= \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_d$$

speech distortion and residual noise

$\longrightarrow$ Singular Value Decomposition

SE Approaches

- Spectral subtraction: estimate noise activity during non-speech regions and subtract it.
- Subspace algorithms: decompose the higher dimensional noisy signal into a lower dimensional one where clean version lays.
- Spectral masking: predict a binary freq-time mask that can cancel out noisy bins.

SE Approaches

- Spe ... nd
  sub
- Sub ... to a
  low
- Spe ... isy
  bin

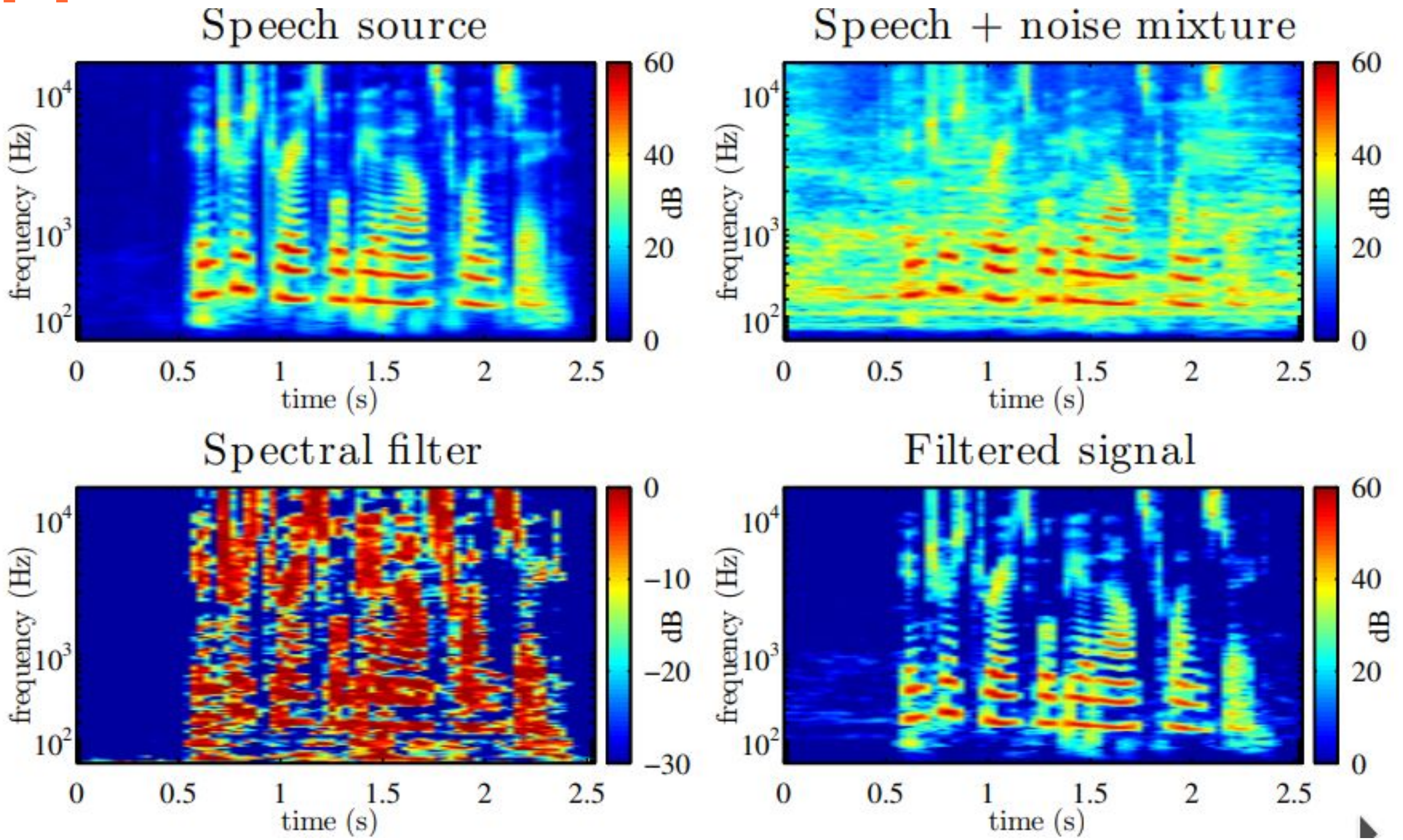


Image credit: Jonathan Le Roux

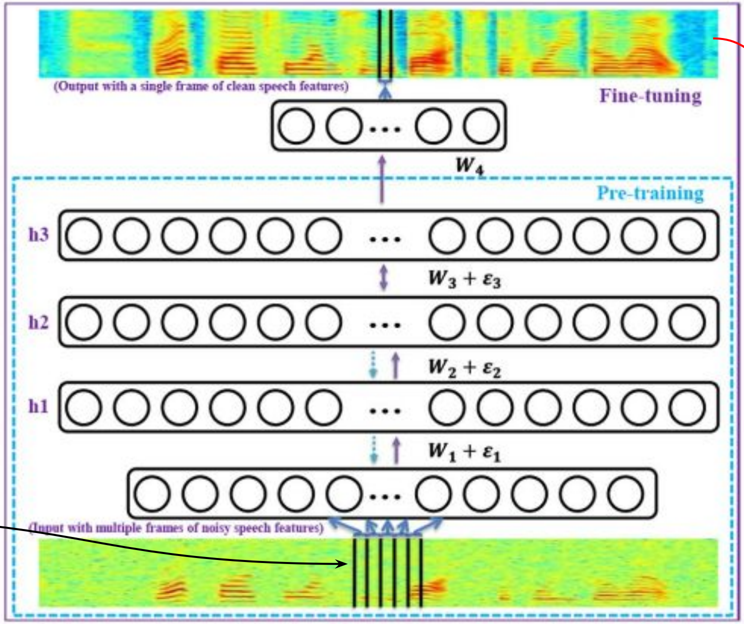
SE Approaches

- Spectral subtraction: estimate noise activity during non-speech regions and subtract it.
- Subspace algorithms: decompose the higher dimensional noisy signal into a lower dimensional one where clean version lays.
- Spectral masking: predict a binary freq-time mask that can cancel out noisy bins.
- **Statistical model based: predict the clean features/signal as a statistical regression problem.**

Discriminative regression

(Xu et al. 2015)

A DNN is used to map noisy parameterized speech (features) into the clean version as a regression problem (MSE estimation).



$$Er = \frac{1}{N}\sum_{n=1}^{N}\|\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau}, \mathbf{W}, \mathbf{b}) - \mathbf{X}_n\|_2^2.$$

Pre-training can be performed stacking RBMs as in paper's proposal.

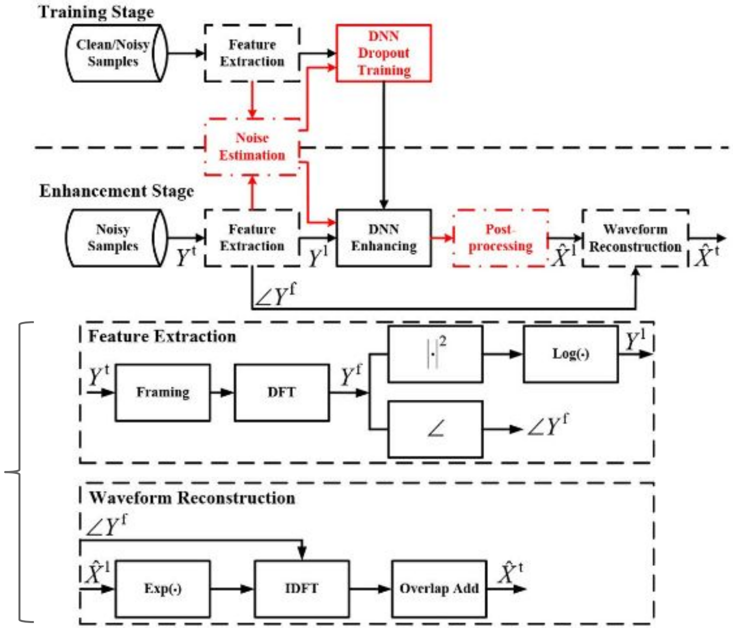
Many input frames consider time correlation

Figure credit: Xu et al.

Discriminative regression

([Xu et al. 2015](#))

A DNN is used to map noisy parameterized speech (features) into the clean version as a regression problem (MSE estimation).



The log power of spectral module is enhanced (predicted). Phase remains the same and ISTFT recorvers signal back.

Figure credit: Xu et al.

SEGAN

Samples

(Pascal et al. 2017)

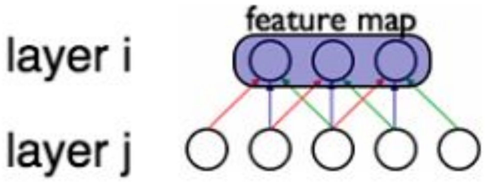
Two stages in Generator (fully convolutional) network:

1. Encoder (Downconv): Project noisy signal into a deterministic representation **c** and concatenate to latent variable **z** ~ N(0, I)

2. Decoder (Deconv): Interpolate the intermediate hidden features w/ learnable params. until re-generation of clean speech.

SEGAN: underlying structures

- 1D convolutional neural networks

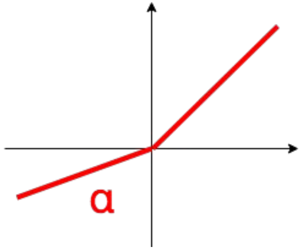


$$h_i^k = \tanh(W_{ij}^k * h_j + b_i^k)$$

x

- Virtual Batch Normalization: normalize layer responses with statistics from (reference_batch + current_batch) → less intra dependent statistics to avoid GAN instability.

- LeakyReLU/ParametricReLU:
    - α fixed (0.3) or learnable

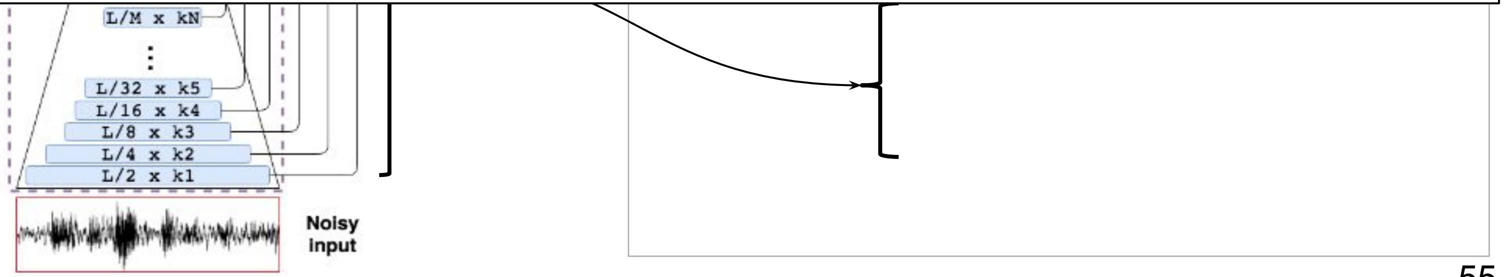


SEGAN end to end training

- Show pairs of signals to “learn” a reconstruction loss.
- Use of L1 regularization to guide the GAN training.

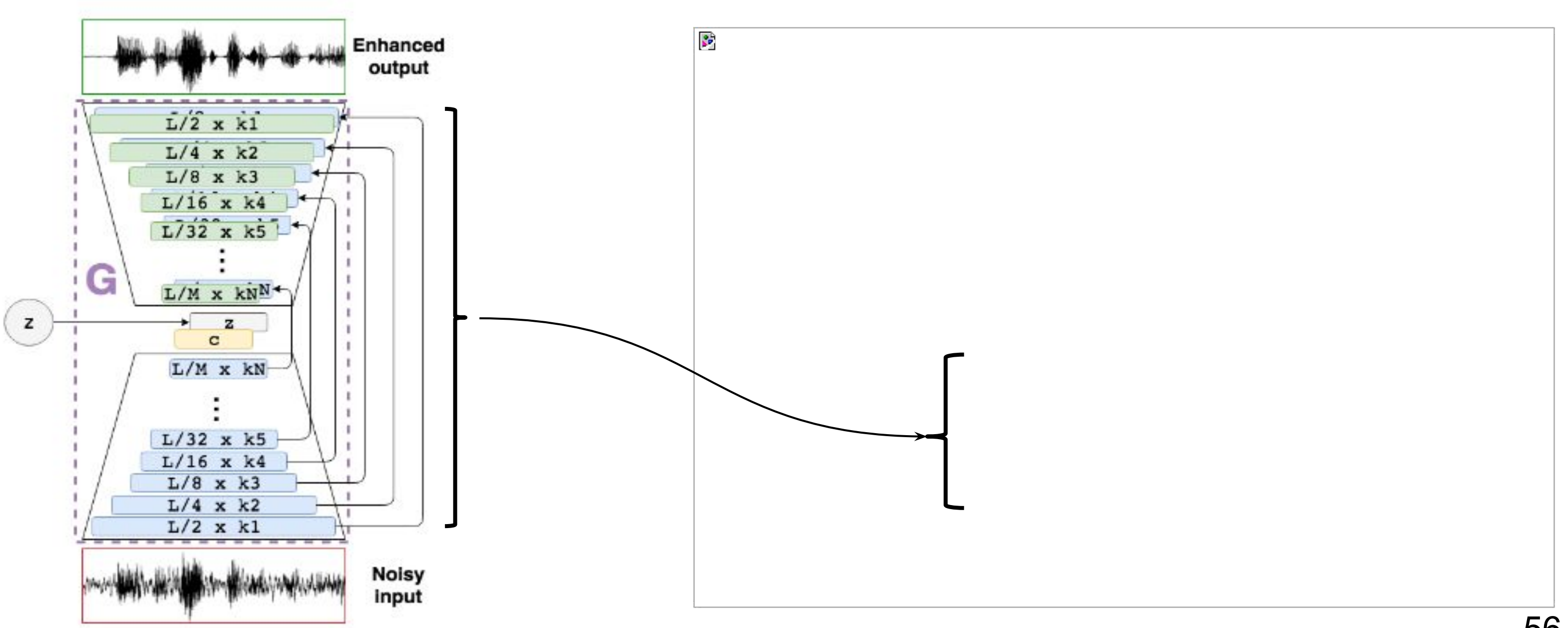
Final G loss: LSGAN Adversarial + weighted L1 regularization/regression

$$\min_{G} V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x}_c),\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[(D(G(\mathbf{z},\mathbf{x}_c))-1)^2] + \lambda\,\|G(\mathbf{z},\tilde{\mathbf{x}})-\mathbf{x}\|_1 .$$



SEGAN end to end training

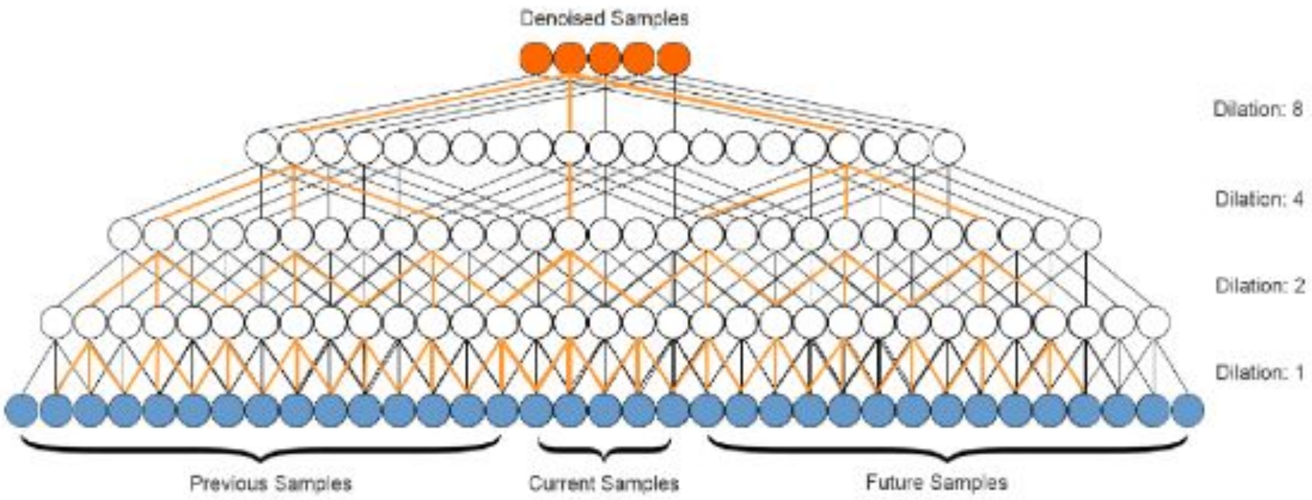
- Show pairs of signals to “learn” a reconstruction loss.
- Use of L1 regularization to guide the GAN training.



Wavenet for Speech Denoising

(Rethage et al. 2017)

Wavenet proved to be effective as a generative model for raw speech and audio. A modified version of it was applied to speech denoising too, getting rid of the original autoregressive behavior, and dealing with a regression problem!



Current SE research

Other active research focus on using **perceptually weighted losses**, or **using enhancement as an internal stage** within another task, like Text-to-Speech (TTS) or Automatic Speech Recognition (ASR):

- RNN-based SE for noise-robust TTS (Valentini et al. 2016)
- Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement (Gurunath and Georgiou 2016)
- Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition (Donahue et al. 2017)

SE Evaluation

Typical objective metrics:

- PESQ: Perceptual Evaluation of Speech Quality [-0.5, 4.5]: designed for telephonic compression assessment.
- COVL: MOS prediction of the overall effect [1, 5]
  - CSIG: Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal [1, 5].
  - CBAK: MOS prediction of the intrusiveness of background noise [1, 5].
- SSNR: Segmental SNR [0, inf).

Nonetheless, subjective eval is always preferable (in any speech synthesis task)!

Summary

- Speech2speech paradigms have been discussed, emphasizing the two salient ones at the moment: enhancement and conversion. All these methods are converging to **end-to-end** approaches.
- Voice Conversion parallel and non-parallel approaches have been reviewed, from classic frame-by-frame analysis to end-to-end VQ-VAE.
- Speech Enhancement methods have been reviewed, specially end-to-end ones, like SEGAN and Denoising Wavenet.
- Speech Enhancement is being included as an inherent end-to-end component for ASR and TTS, among others.
- Speech2speech paradigms are gaining momentum, specially the end-to-end embedded versions to process speech signals in real time in our handset devices.

References

- Auto-Encoding Variational Bayes (Kingma and Welling 2014)
- Generative Adversarial Networks (Goodfellow et al. 2014)
- Voice Conversion Using Artificial Neural Networks (Desai et al. 2009)
- Voice Conversion Using Deep Bidirectional Long-Short Term Memory Based Recurrent Neural Networks (Sun et al. 2015)
- Voice Conversion from Non-Parallel Corpora Using Variational Auto-encoder (Hsu et al. 2016)
- Neural Discrete Representation Learning (van den Oord et al. 2017)
- A Regression approach to speech enhancement based on deep neural networks (Xu et al. 2015)
- Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement (Gurunath and Georgiou 2016)
- RNN-based SE for noise-robust TTS (Valentini et al. 2016)
- SEGAN: Speech Enhancement Generative Adversarial Network (Pascual et al. 2017)
- Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition (Donahue et al. 2017)
- A Wavenet for Speech Denoising (Rethage et al. 2017)