Yandex

SCHOOL OF DATA ANALYSIS

Speech-to-Speech

Sergey Dukanov

May 23rd 2022

Recap and Lecture Plan

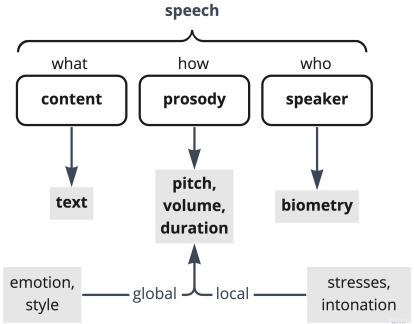
We discussed transformations from speech to text domain and vice versa.

But speech domain is richer, so let's consider tasks when we transform speech to another speech.

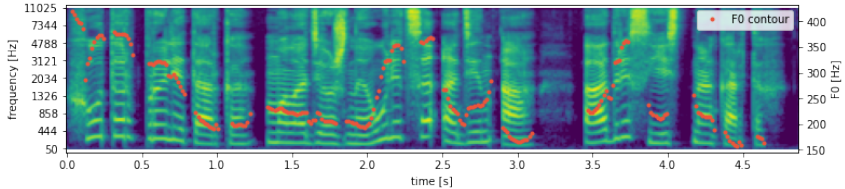
In this lecture:

1. what components speech consists of?
2. speech denoising and speech separation
3. voice cloning and voice conversion
4. speech translation

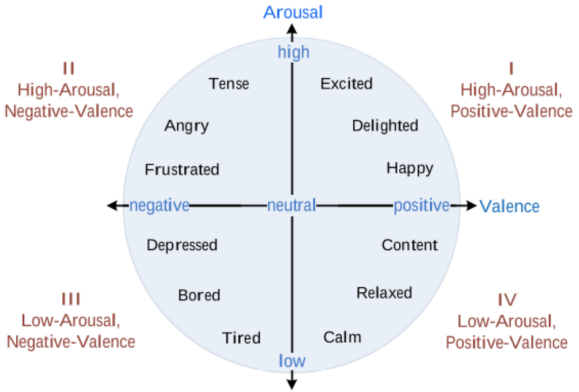
Speech components

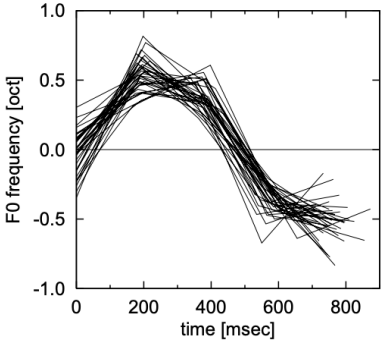


- **pitch** – perceptive value, **F0** (**fundamental frequency**) – physical value
- we are interested in **prosody modeling** in modern speech synthesis, **what approaches do you already know?**
- but it's hard to formalize it, **can you try?**

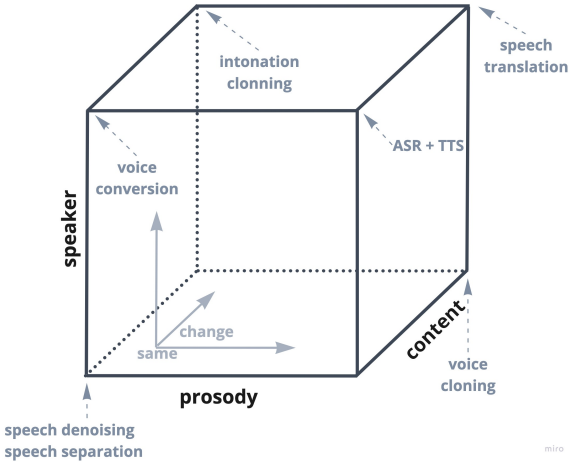


High-level and Low-level representation of prosody





Speech-to-Speech Tasks

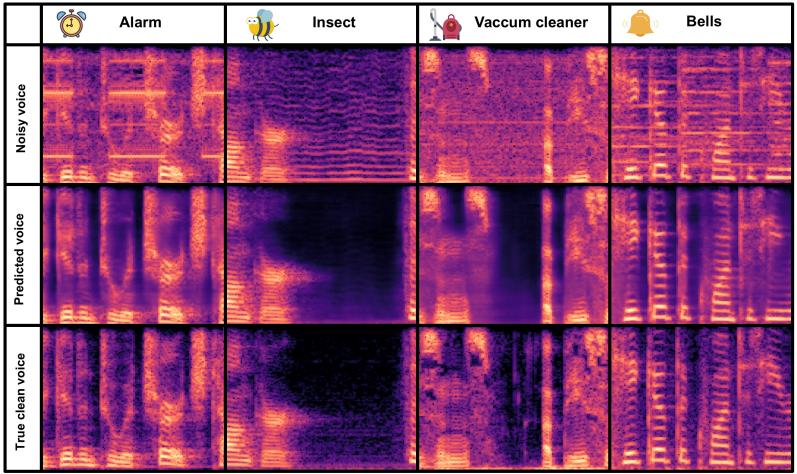


Consider the problem of speech-to-speech conversion.

We produce new audio signal by modifying (or not) components of the original speech.

What task from computer vision are similar to them?

Speech Enhancement (Denoising)



Speech improvement can mean the following tasks:

- Background noise removal
- Reverberations removal
- Quality improvement

How to obtain data for given tasks?

Objective

- Perceptual Evaluation of Speech Quality (**PESQ**)
- Short-Time Objective Intelligibility (**STOI**)
- Speech-to-reverberation Modulation Energy Ratio (**SRMR**)
- Frequency-weighted Segmental SNR (**FWSegSNR**)

Subjective

- Mean Opinion Score (**MOS**)
- MUltiple Stimuli with Hidden Reference and Anchor (**MUSHRA**)

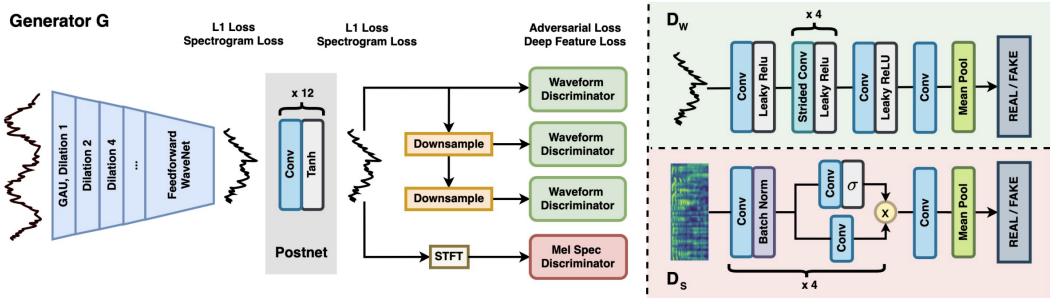
HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks

Jiaqi Su[1,2], *Zeyu Jin*[2], *Adam Finkelstein*[1]

[1]Princeton University [2]Adobe Research

- uses a **feed-forward** (e.g. without auto-regression) WaveNet architecture
- multi-scale adversarial training in both time domain and time-frequency domain
- relies on the deep feature matching losses of the discriminators
- generalizes well to new speakers, new speech content, new environments

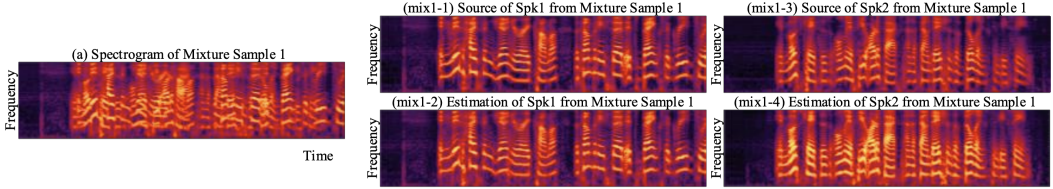
Speech Enhancement: HiFi-GAN Denoiser



Further investigations:

- solving the discriminator-evaluation mismatch problem → MetricGAN
- unsupervised speech enhancement (on noise data) → MetricGAN-U
- real-time inference on mobile CPU → DEMUCS

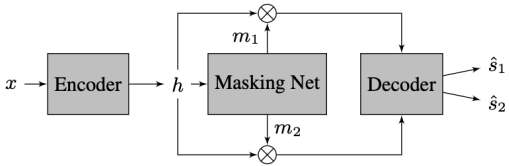
Speech Separation



ATTENTION IS ALL YOU NEED IN SPEECH SEPARATION

Cem Subakan[1], *Mirco Ravanelli*[1], *Samuele Cornell*[2], *Mirko Bronzi*[1], *Jianyuan Zhong*[3]

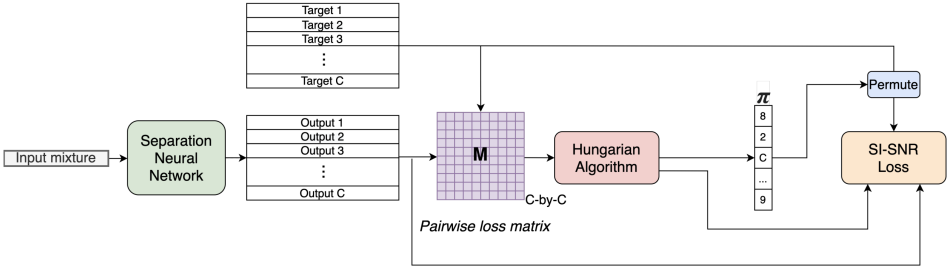
- fully convolutional encoder (takes time-domain signal)
- transformer-based masking network (estimates mask for each of Ns speakers in the mixture
- transposed convolution layer in decoder
- permutation-invariant loss
- **what is the problem?**



Many-Speakers Single Channel Speech Separation with Optimal Permutation Training

Shaked Dovrat[1]*, *Eliya Nachmani*[1,2]*, *Lior Wolf*[1]

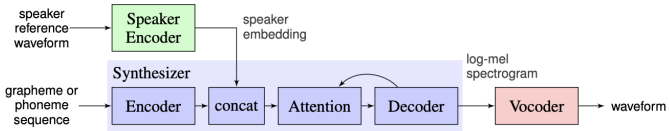
A method for single channel sound separation for a large number of sources.



Voice Cloning from Speaker Verification

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

Ye Jia* Yu Zhang* Ron J. Weiss* Quan Wang Jonathan Shen Fei Ren
Zhifeng Chen Patrick Nguyen Ruoming Pang Ignacio Lopez Moreno Yonghui Wu



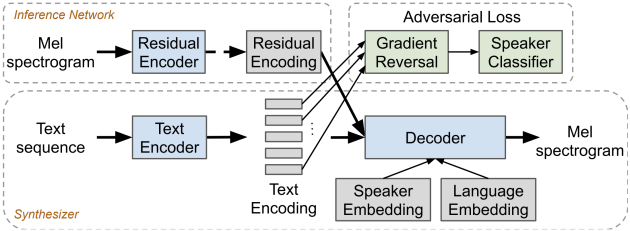
Speaker encoder is trained on speaker verification task and frozen during Tacotron training.

The representation must capture the characteristics of **different speakers**, be able to identify these characteristics using only a **short adaptation signal**, be **independent** of its phonetic content and background noise.

Cross-Language Voice Cloning

Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, Bhuvana Ramabhadran



We want same voice for different languages, but recordings from bilingual speakers are expensive to collect.

Incorporating an adversarial loss term to encourage the model to disentangle its representation of speaker identity (which is perfectly correlated with language in the training data) from the speech content.

Voice Conversion

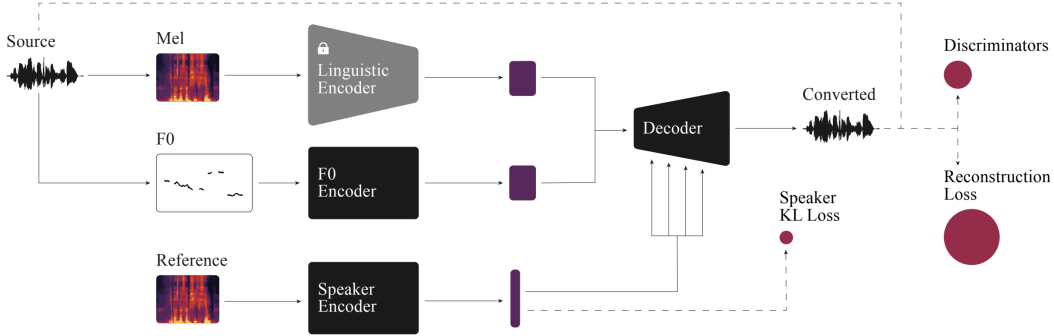
Voice conversion architectures classification by internal representation:

- architectures with some latent space (VAE-like)
- mel-based architectures (CycleGAN-like models)
- phoneme-based features

HiFi-VC: High Quality ASR-Based Voice Conversion

Anton Kashkin[1], *Ivan Karpukhin*[1], *Svyatoslav Shishkin*[1]

HiFi-GAN based voice conversion using lingustic features from pretrained ASR, F0 estimation and speaker embedding.

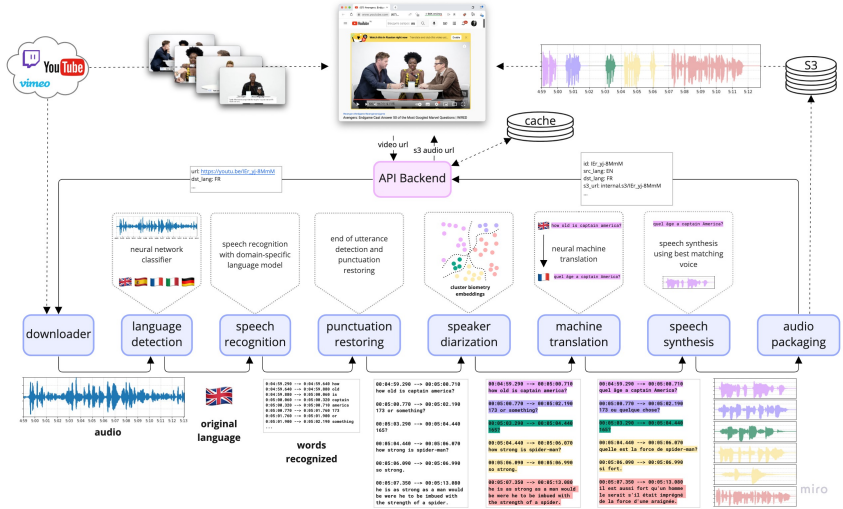


Neural Voiceover

Fully automated voiceover based on neural networks.

Current achievements in the field of machine translation and speech technologies make it possible to do human-level transcribing and voiceover in cost of few cents.

Video Translation Scheme



Conclusion