

Winter Term 2026/27

Selected Topics in Deep Learning for Audio, Speech, and Music Processing

Lecture

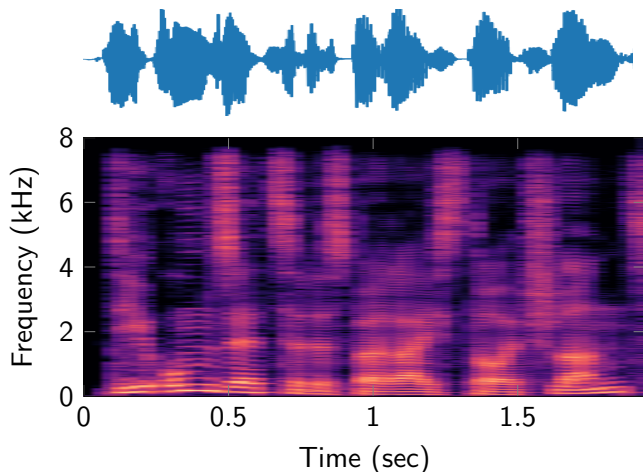
Introduction to Neural Speech Synthesis

Stefan Balke, Andreas Brendel, Ben Maman, Meinard Müller

Speech Signal Representations

‘Die Umsätze des Bereichs Deutschland’

[di: 'ʊm,zɛtsə dɐs bə'raɪçs 'dɔʊtʃlant]



Text:

- ▶ Atoms of writing systems
- ▶ Language dependent
- ▶ No direct mapping to sounds

Phoneme transcription:

- ▶ Atoms of speech sounds
- ▶ Different phoneme alphabets
- ▶ Direct mapping to sounds

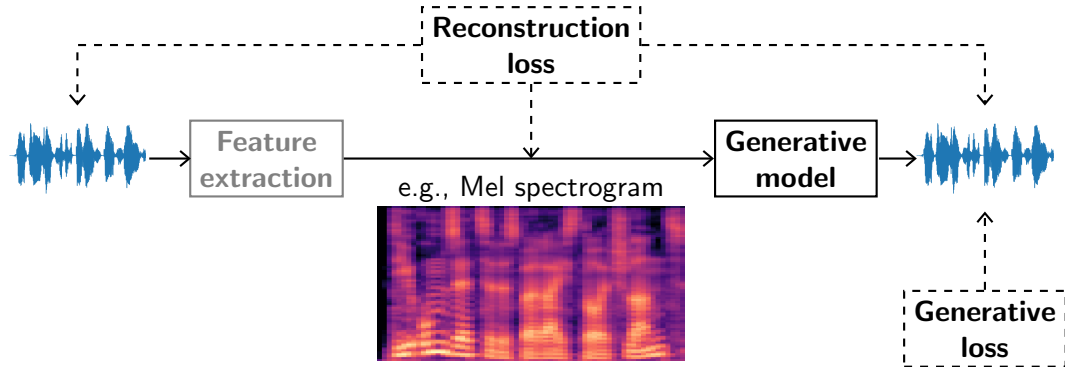
Waveform:

- ▶ Reflects air pressure variations
- ▶ Digitization needed

Spectrogram:

- ▶ STFT magnitude in dB
- ▶ Windows determine resolution

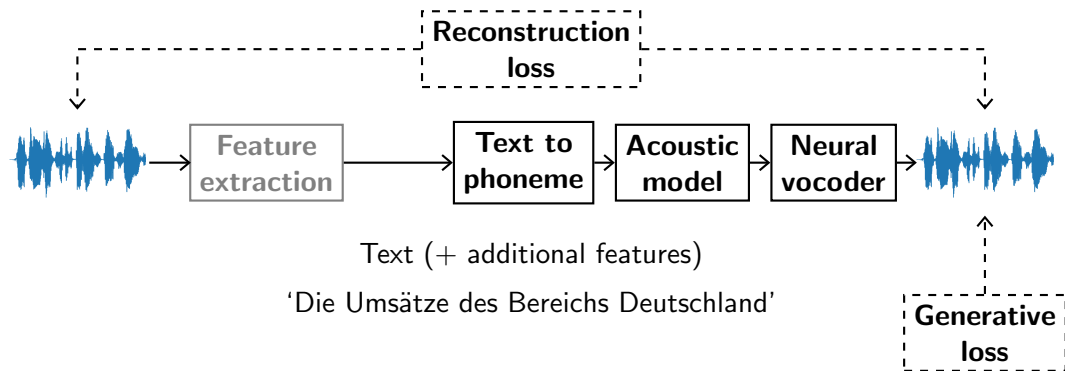
Speech Synthesis - Vocoder



In general: Speech features $\mapsto$ speech waveform

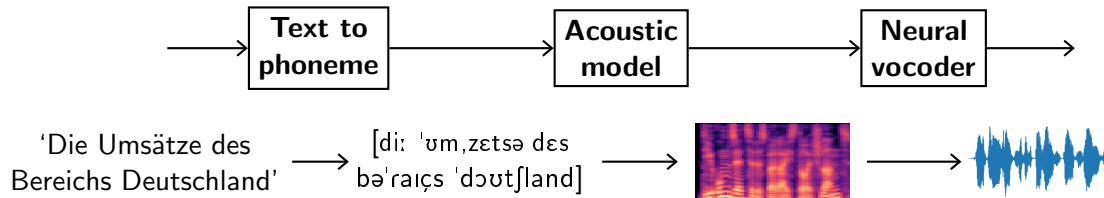
- ▶ Reconstruction loss: Adherence to input
- ▶ Generative loss: Naturalness

Text-To-Speech (TTS) Synthesis



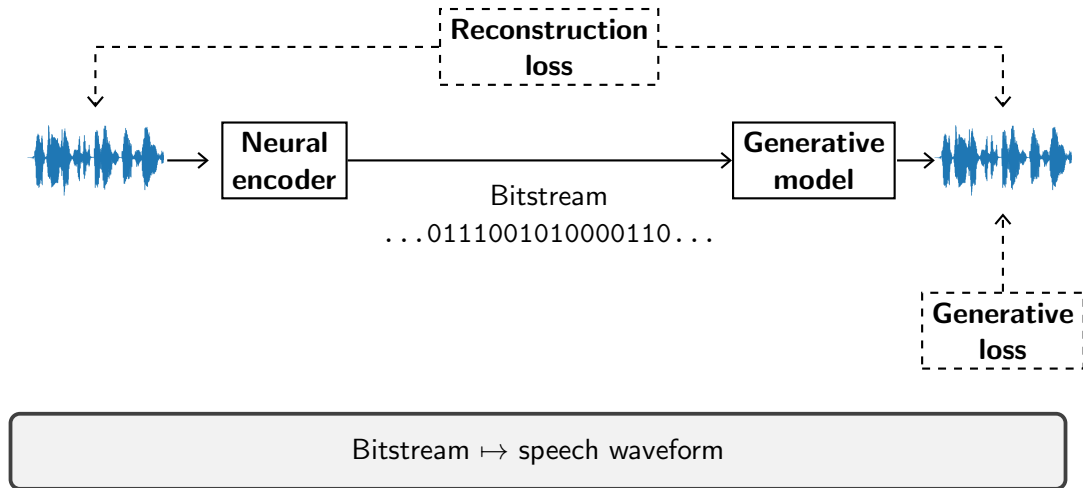
Text (+ additional features) $\mapsto$ speech waveform

TTS Systems

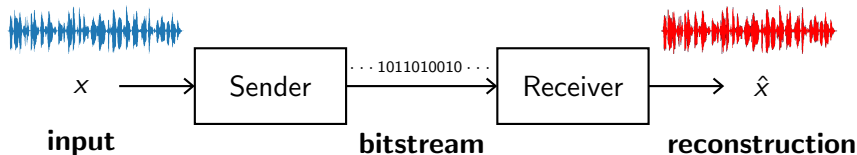


- ▶ Need paired training data
 - ▶ Text $\leftrightarrow$ phoneme annotations
 - ▶ Phoneme annotations $\leftrightarrow$ mel spectra
 - ▶ Mel spectra $\leftrightarrow$ speech recordings
- ▶ Separate training of three model components

Neural Speech Coding



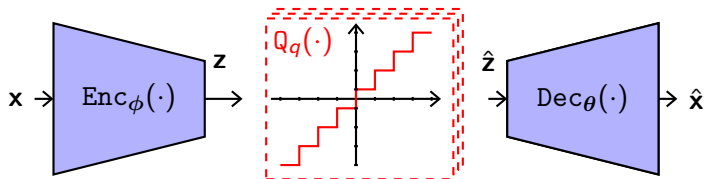
Audio Signal Compression



Compression: Find discrete signal representation by maximizing quality and minimizing bitrate

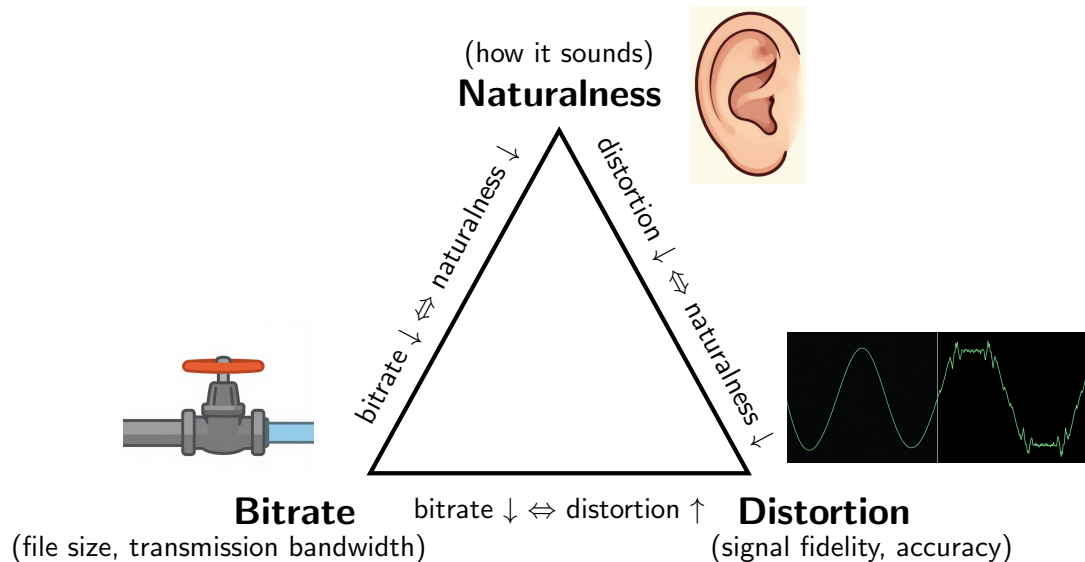
- ▶ **Redundancy:** Source properties known to sender and receiver result in redundant information that need not be transmitted
- ▶ **Irrelevancy:** Human listeners do not perceive all deviations of the received signal from the original

Quantization in Neural Audio Codecs

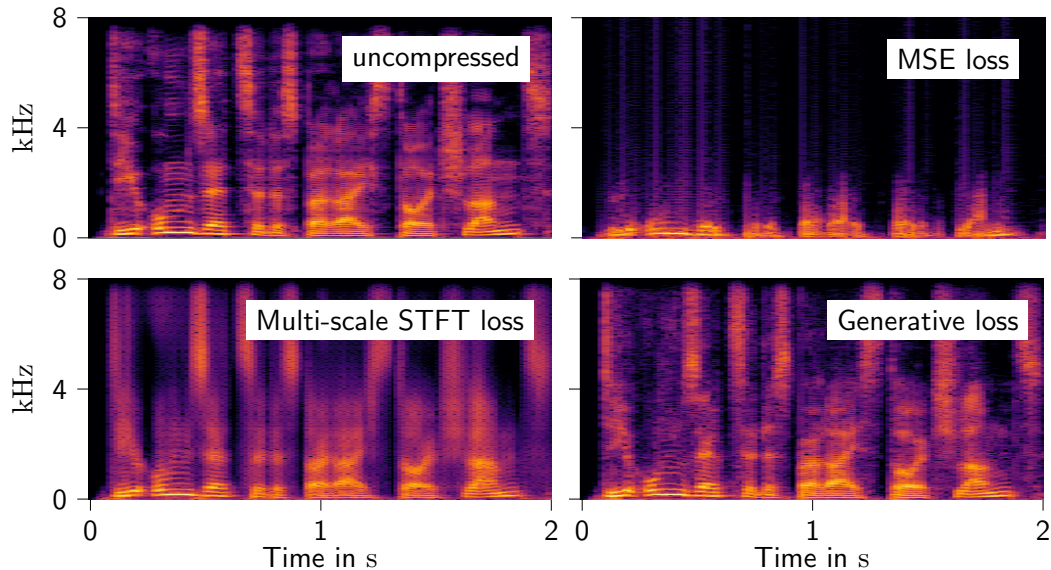


- ▶ Neural audio codec = Neural encoder + quantizer + neural decoder
- ▶ Want to learn discrete representation $\hat{z}$ end-to-end
- ▶ $Q_q(\cdot)$ **not differentiable!** → Solutions: Soft approximations, special updates, ...
- ▶ Bitrate = sample rate $\times$ bit depth $\times$ channels
e.g., CD (44.1 kHz, 16 bits, 2 channels): $44100 \times 16 \times 2 = 1411.2$ kbps

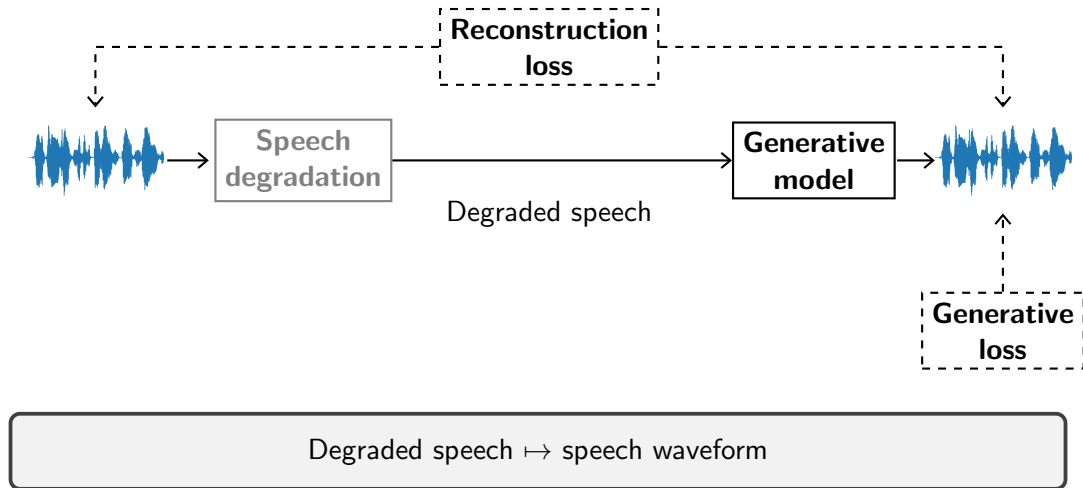
Rate-Distortion-Naturalness Trade Off



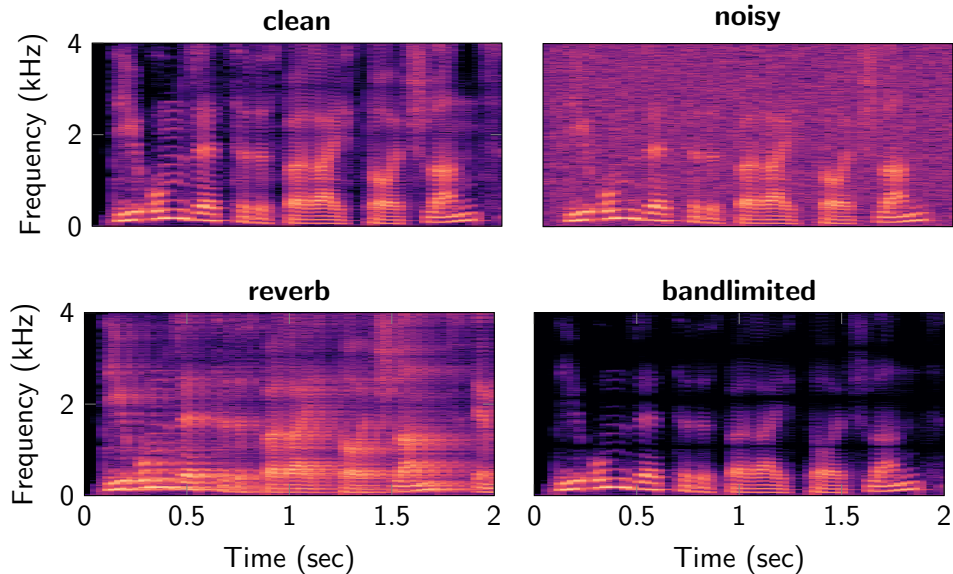
Results for Neural Speech Coding at 1.5 kbps



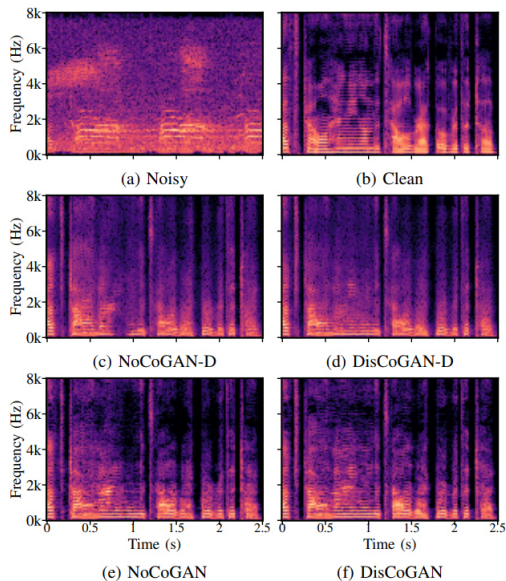
Generative Speech Enhancement



Speech distortions



Generative Speech Enhancement Results

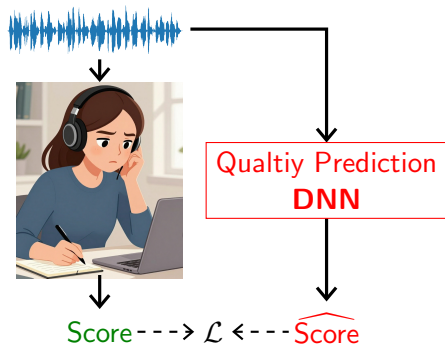


only reconstruction loss

with generative loss

Quality Evaluation of Speech Synthesis Systems

Subjective Eval Objective Eval



- ▶ Subjective eval (listening tests) is costly gold standard
 - ▶ Objective eval cheap and scalable but inaccurate
- ⇒ **Own field of research**

Quality dimensions:

- ▶ Naturalness
- ▶ Preservation of signal content
- ▶ Intelligibility
- ▶ ...

Real-world Application: Additional Desiderata

...depend on the specific application but include

- ▶ low algorithmic latency
- ▶ low computational complexity
- ▶ no demographic or biometric bias
- ▶ generalization and robustness to, e.g., different languages, background noises etc.
- ▶ modularity and compatibility
- ▶ data protection and adversarial robustness

Fundamental Questions in Speech Synthesis

- ▶ What are common speech representations? What are pros/cons of different representations?
- ▶ What is vocoding and for which applications is it useful?
- ▶ What is a Mel spectrogram?
- ▶ What is data compression and what principles can be used here?
- ▶ What influences the bitrate of an audio signal and how is it related to quantization?
- ▶ What is the rate-distortion-naturalness trade off?
- ▶ What is the role of a reconstruction loss and a generative loss in speech synthesis?
- ▶ What are examples of common speech distortions?
- ▶ What are requirements for a speech synthesis system?
- ▶ What is objective and subjective quality estimation?