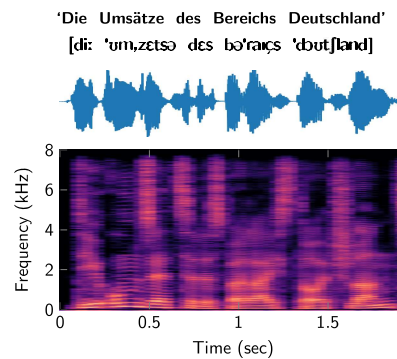


Winter Term 2026/27
Selected Topics in Deep Learning for
Audio, Speech, and Music Processing

Lecture
Introduction to Neural Speech Synthesis

Stefan Balke, Andreas Brendel, Ben Maman, Meinard Müller

Speech Signal Representations



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

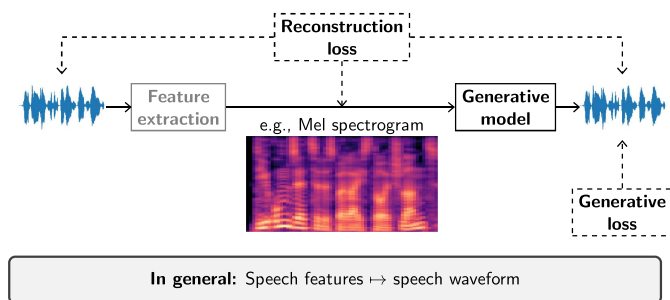
Waveform:

- ▶ Reflects air pressure variations
- ▶ Digitization needed

Spectrogram:

- ▶ STFT magnitude in dB
- ▶ Windows determine resolution_{2/16}

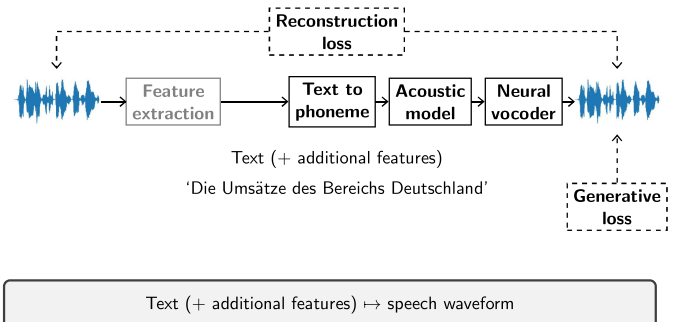
Speech Synthesis - Vocoding



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

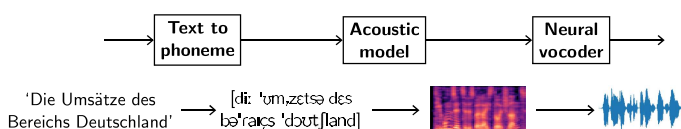
3 / 16

Text-To-Speech (TTS) Synthesis



4 / 16

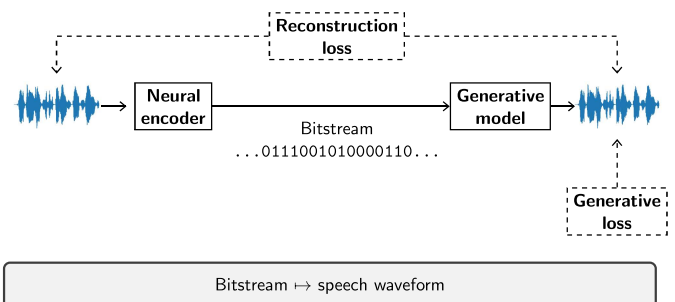
TTS Systems



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

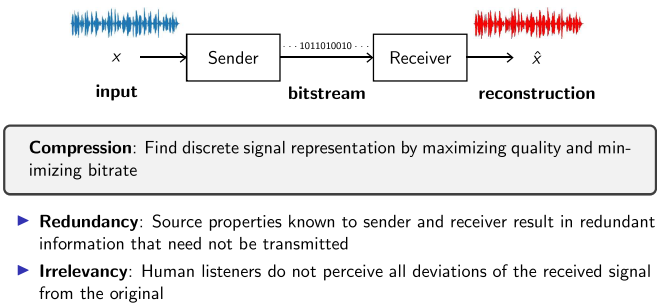
5 / 16

Neural Speech Coding

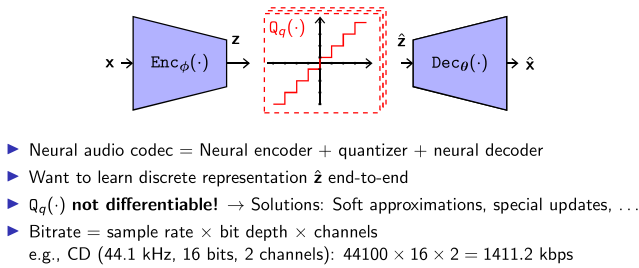


6 / 16

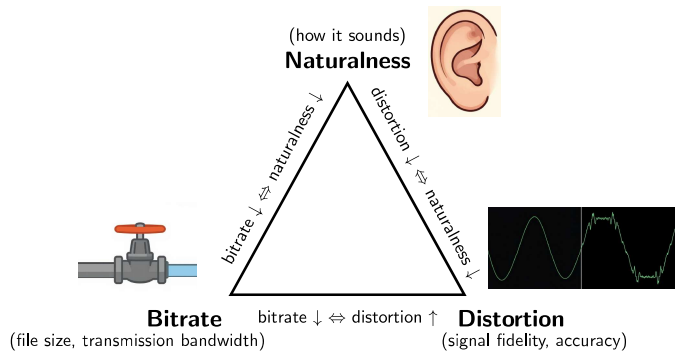
Audio Signal Compression



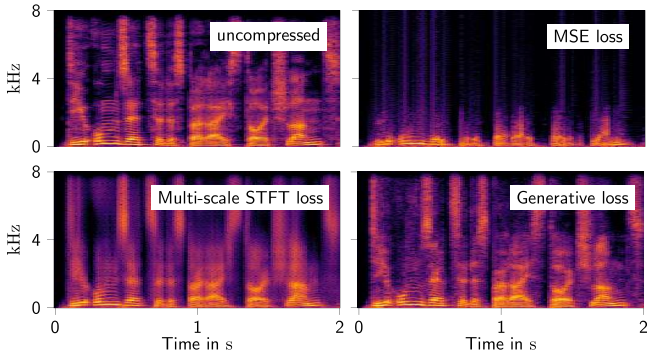
Quantization in Neural Audio Codecs



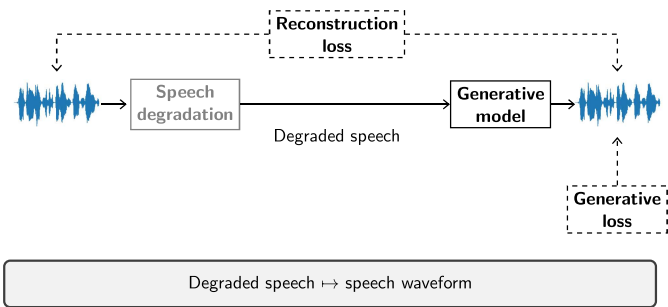
Rate-Distortion-Naturalness Trade Off



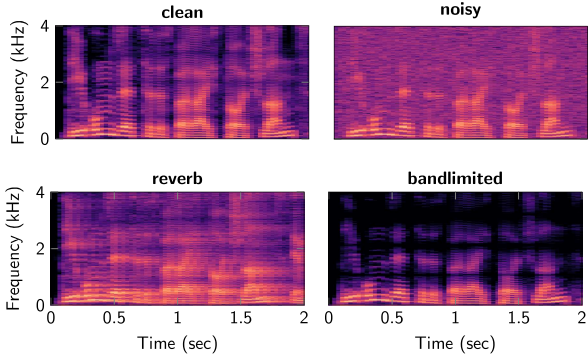
Results for Neural Speech Coding at 1.5 kbps



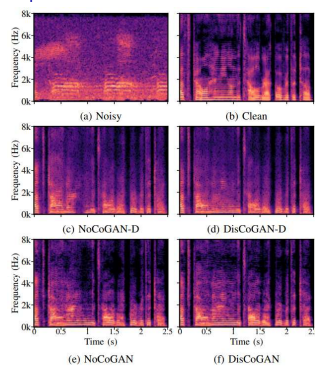
Generative Speech Enhancement



Speech distortions



Generative Speech Enhancement Results



only reconstruction loss

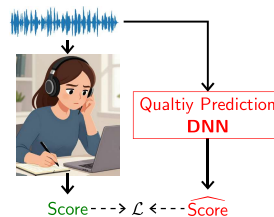
with generative loss

Image: Shetu, Habets, Brendel, "Leveraging Discriminative Latent Representations for Conditioning GAN-Based Speech Enhancement", TASLP 2026

13 / 16

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
- ▶ ...

14 / 16

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

15 / 16

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?

16 / 16