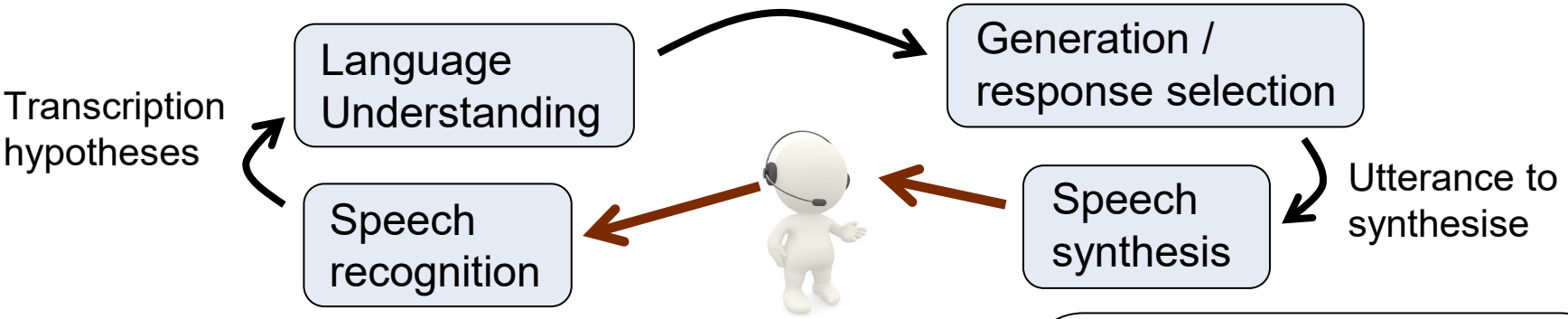
NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

Speech processing

Pierre Lison
plison@nr.no

IN4080, autumn semester 2025

Spoken dialogue systems



Speech interfaces add a layer of complexity:

- Need to handle *uncertainties*, ASR errors etc.
- Speech communicates *more than just words* (intonation, emotions in voice, etc.)
- Need to handle *turn-taking*

Note: An alternative is to rely on *speech language models*, which work directly at the speech level, without using text transcriptions as intermediate

[Cui et al. ACL 2025.]

A difficult problem!



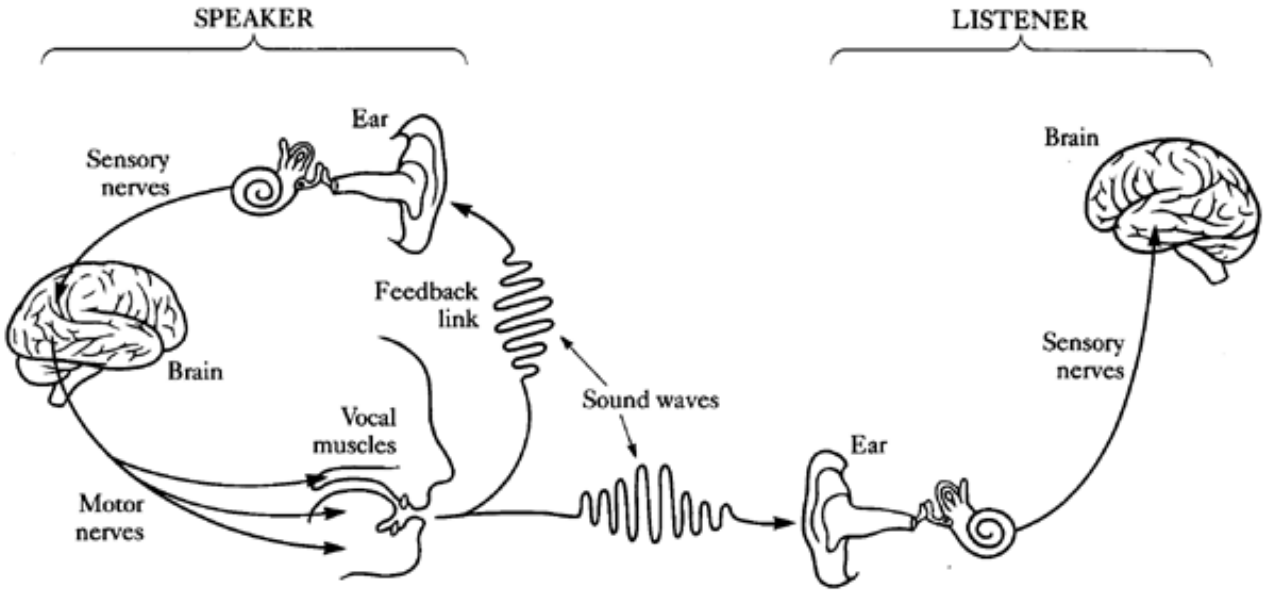
Outline

- Short intro to phonetics and acoustics
- Speech recognition
- Speech synthesis
- Disfluencies

Outline

- **Short intro to phonetics and acoustics**
  - Speech production
  - Speech perception
- Speech recognition
- Speech synthesis
- Disfluencies

The speech chain



[Denes and Pinson (1993), «The speech chain»]

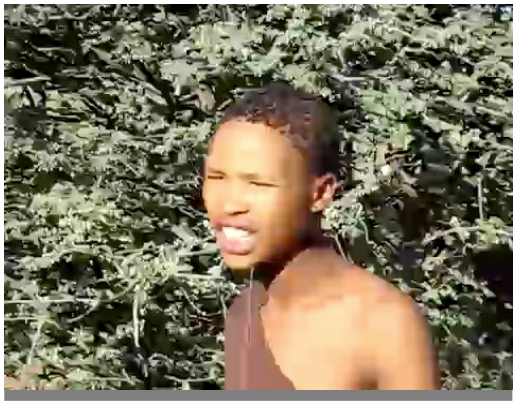
Speech production

- Sounds are *variations in air pressure*
- How are they produced?
  - An **air supply**: the *lungs* (we usually speak by breathing out)
  - A **sound source** setting the air in motion (e.g. vibrating) in ways relevant to speech production: the *larynx*, in which the *vocal folds* are located
  - A set of 3 **filters** modulating the sound: the *pharynx*, the *oral tract* (teeth, tongue, palate,lips, etc.) & the *nasal tract*

Speech production

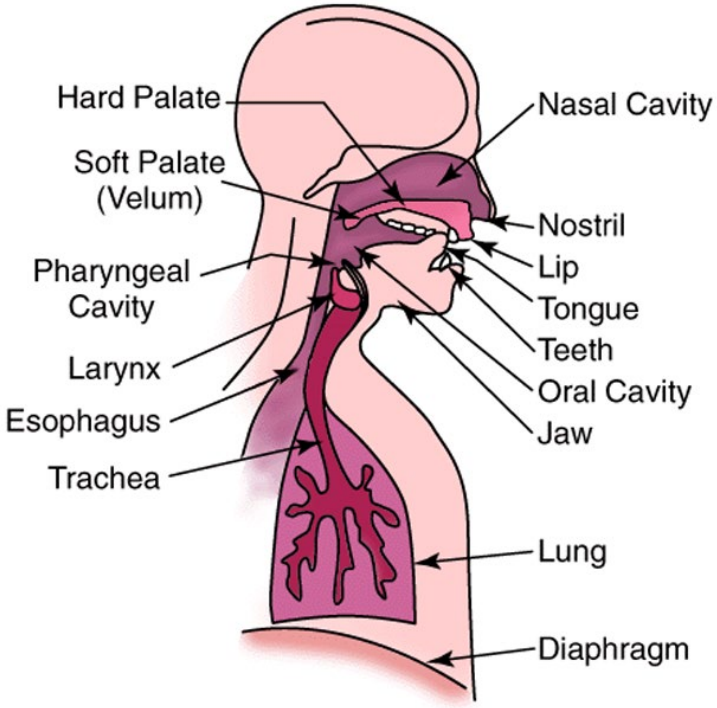
Visualisation of the vocal tract via *magnetic resonance imaging* [MRI]:





NB: A few languages also rely on sounds not produced by vibration of vocal folds, such as *click languages* (e.g. Khoisan family in south-east Africa):

Speech production



Key differences between speech sounds:

1. **Voiced** vs **voiceless**: are the vocal cords vibrating?
   - Diff between e.g. [p] and [b]
   - *Note*: all vowels are voiced
2. **How** does the restriction of airflow occur?
   - Stop, nasal, fricatives, etc.
3. **Where** does the restriction occur?
   - Labial, coronal, guttural

Speech production

Words can be pronounced very differently:

- *Phonetic processes*: aspirations, assimilations, deletions
- *Co-articulation*: anticipation of next phone by articulators, or perseverance of previous move
- Influence of various *contextual factors*: age, gender, environment, rate of speech, dialect, register

IPA transcription

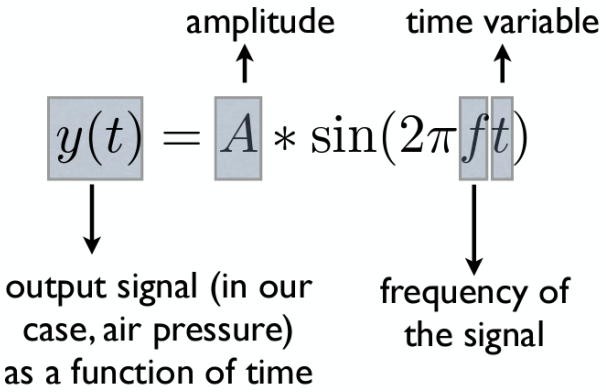
[Nordavinden og sola, en norsk dialektprøvedatabase på nettet, https://www.hf.ntnu.no/nos/list.php]

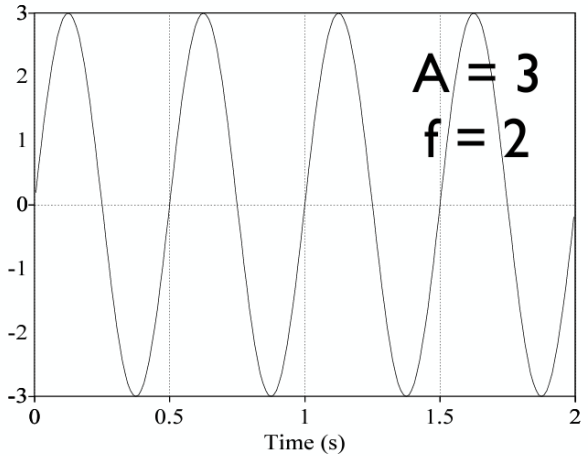
Bokmål	Nordavinden og sola kranglet om hvem av dem som var den sterkeste. Da kom det en mann gående med en varm frakk på seg. De blei enige om at den som først kunne få mannen til å ta av seg frakken skulle gjelde for sterkere enn den andre. Så blåste nordavinden av all si makt, men dess mer han blåste, dess tettere trakk mannen frakken rundt seg, og til sist gav nordavinden opp. Da skinte sola fram så godt og varmt, og straks tok mannen av seg frakken. Og så måtte nordavinden innrømme at sola var den sterkeste av dem.
IPA (Oslo)	[ˈˈnuːrɑˌʊinːˌɳ ɔ ˈsuːlɳ ˈˈkrɑŋlət ɔm ˈʊem ɑ dem smˌ ˈʊɑː ɖɳ ˈˈstærkəstə ˌdɑː ˈkʰɔmː de n ˈmɑnː ˌgɔːənə me n ˈʊɑrm ˈfrɑkː pɔ ˌsæ di ble ˈˈeːnjə ɔm ɑt ˈdenː smˌ ˈføʂ̲t ˌkʰʉnːə fɔ ˈmɑnːɳ tɔ ˈˈtʰɑː ɑ sæ ˈfrɑkːən ˌskʉlːə ˈˈjelːə fɔ ɖɳ ˈˈstærkəstə ɑ ˌdemː ˈsoː ˈˈbloːstə ˈˈnuːrɑˌʊinːˌɳ ɑ ˈʔalː sin ˈmɑkʰtʰ men ju ˈmeːr hɑm ˈˈbloːstə ju ˈˈtʰetːərə ˌtrɑkː ˈmɑnːˌɳ ˈfrɑkːən ˈrʉnt sæ ɔ tə ˈsist ˌmɔtːə ˈˈnuːrɑˌʊinːˌɳ ˈˈjiː ˌɔpʰː ˌdɑː ˈˈʂintə ˈsuːlɳ ˈfrem ˌsɔː ˈgɔtʰː ɔ ˈʊɑrmtʰ ɑt̄ ˈmɑnːən ˈstrɑks ˌmɔtːə ˈˈtɑː ɑ sæ ˈfrɑkːən ɔ ˈsoː ˌmɔtːə ˈˈnuːrɑˌʊinːˌɳ ˈinːˌrømːə ɑt ˈsuːlɳ ˈʊɑːr ɳ ˈˈstærk̚əstə ʔɑ ˈdemː]
IPA (Fyresdal)	[ˈˈnuːrɑˌʊinːˌɳ ɔ ˈsuːla ˈˈkrɑŋla um ˈkʊenː ɑʊ ˌd̪æi sɔmˌ ˈʊɑː d̪ən ˈˈstærkɑstə ˈdoː ˈkɔmː də ɛn ˈmɑnː ˈˈgɑŋgandə mə æn ˈʊɑrmə ˈfrɑkːə pɔ ˌseːg dæi blæɪ ˈˈʔeːnigə um ɑt dæn sɔmˌ ˈfysː ˌkʰunːə fɔ ˈmɑnːən tə ɔ ˈˈtʰɑːk ɑʊ sə ˈfrɑçːən ˌskʉlːə ˈˈreknɑs fer ɛ̃n ˈˈstærkɑst ɑʊ ˌd̪ɛɪ ˈsoː ˈbluːs ˈˈnuːrɑˌʊinːˌɳ ɑʊ ˈɑlː si ˈmɑkʰtʰ mən tɪ ˈmæi hɑn ˈbluːs tɪ ˈˈtʰetːɑrə ˈdruːg ˈmɑnːən ˈfrɑçːən ˈrʊnt seɯ ɔ tɪ ˈs̪l̚ʉtː ˈˈmɔtːə ˈˈnuːrɑˌʊinːˌɳ? ˌjeː ˈupːʰ ˌd̪oː ˈʂæin ˈsuːla ˈfrɑmː sɔ ˈgøʰtː ɔ ˈʊɑrmt ɑt ˈmɑnːˌɳ m̊e ˈʔæi ˈgɔŋ ˈˈmɔtːə ˈˈtʰɑːkə ˈɑːʊ sə ˈfrɑçːən ɔ ˈsoː ˈˈmɔtːə ˈˈnuːrɑˌʊinːˌɳ ˈʊeːˌçɛn̪ːə ʔɑt ˈsuːla ˈʊɑː dən ˈˈstærk̚ɑ̰st ɑʊ ˈðæ̰ɪ]

Speech perception

A (speech) sound is a variation of air pressure

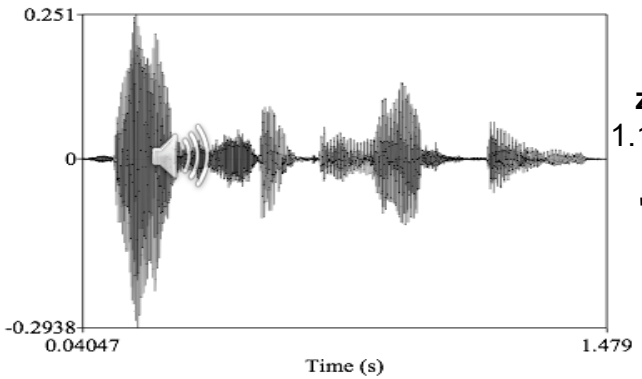
- This variation originates from the speaker's speech organs
- We can plot a **wave** showing the changes in air pressure over time (zero value = normal air pressure)



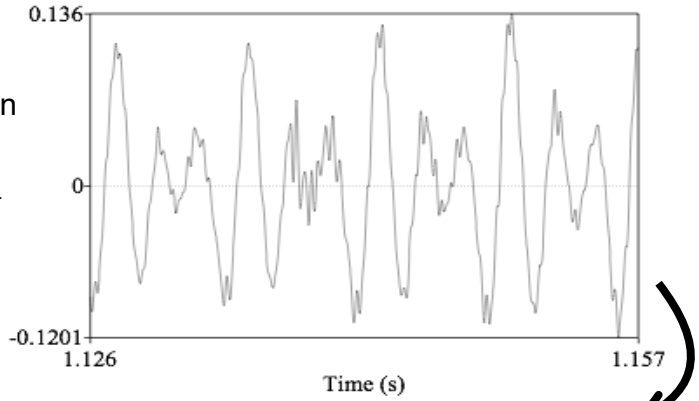


Speech perception

- Of course, speech is more complex than a simple sine function
- But it can still be described using the same mathematical apparatus, in terms of frequency, amplitude etc.



zoom on the part between 1.126 and 1.157 s.



About 4 cycles in the waveform, which means a frequency of about 4/0.03 ≈129 Hz

Important measures

1. The **fundamental frequency F_0**: lowest frequency of the sound wave, corresponding to the speed of vibration of the vocal folds (between 85-180 Hz for male voices and 165-255 Hz for female voices)

2. The **intensity**: the signal power normalised to the human auditory threshold, measured in **dB** (decibels):

$$\text{Intensity} = 10\ \log_{10}\frac{\text{Power}}{P_0} = 10\ \log_{10}\frac{1}{NP_0}\sum_{i=1}^{N} y(t_i)^2$$

Total energy of signal

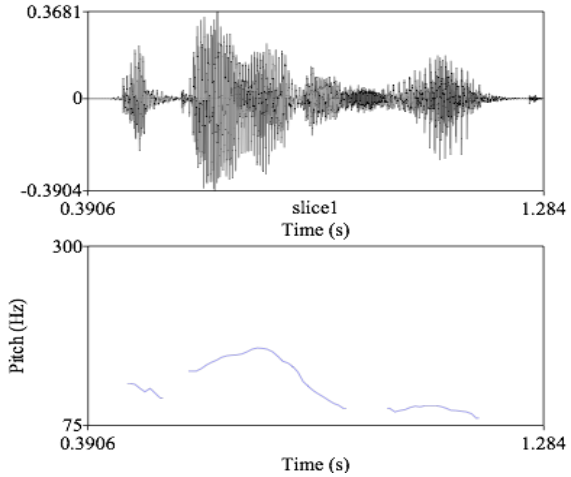
for a sample of N time points t_1,... t_N
P_0 is the human auditory threshold, = 2 x 10^{-5} Pa

Note: dB scale is logarithmic, not linear!

Why are F_0 and the intensity important?

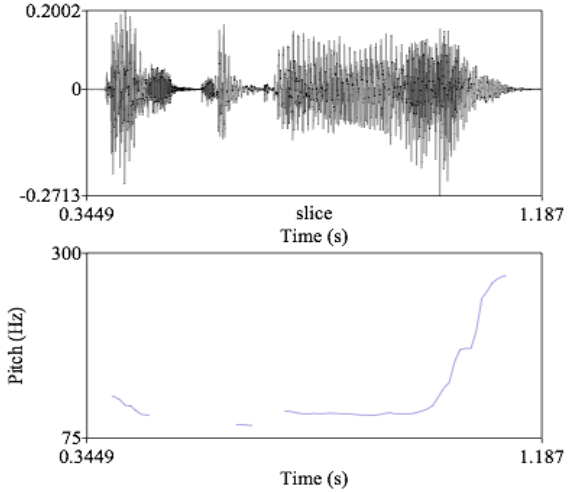
F_0 correlates with the *pitch* of the voice, and the pitch movement for an utterance will give us its *intonation*

"The ball is red"



Interrogative utterance
= rising intonation at the end

"Is the ball red?"

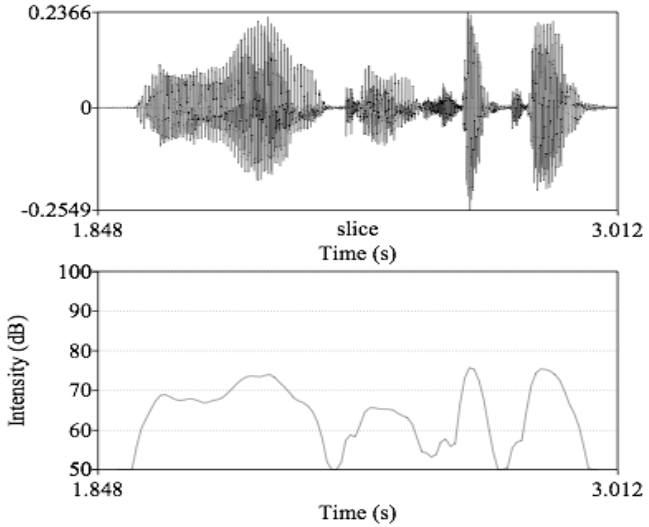


Why are F_0 and the intensity important?

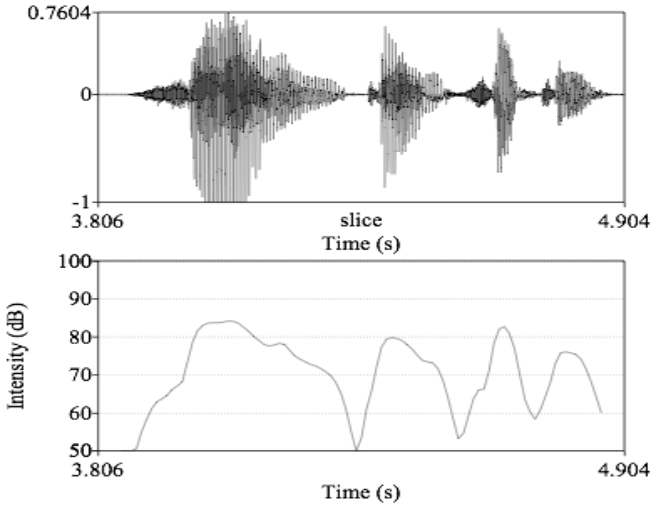
F_0 correlates with the *pitch* of the voice, and the pitch movement for an utterance will give us its *intonation*

The signal intensity corresponds to the *loudness* of the speech sound

low intensity



high intensity



Outline

- Short intro to phonetics and acoustics
- **Speech recognition**
- Speech synthesis
- Disfluencies

Why is ASR difficult?

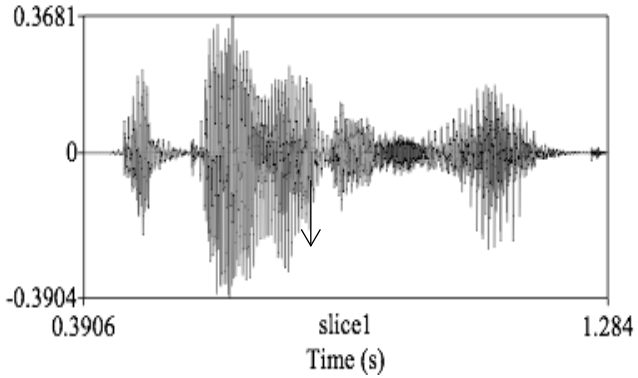
Many sources of variation:

- speaker voice (and style),
- accents,
- ambient noise
- etc.



The speech recognition task

Input: Audio data



Sequence **O** of acoustic observations (i.e. every 20 ms)

Output: Transcription

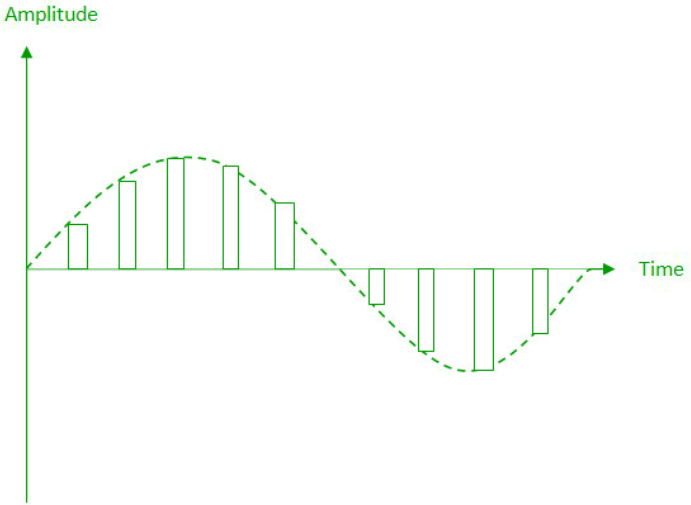
"The ball is red"

Goal: Map speech signal **O** into sequence of linguistic symbols $\hat{W}$ (typically characters):

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|O)$$

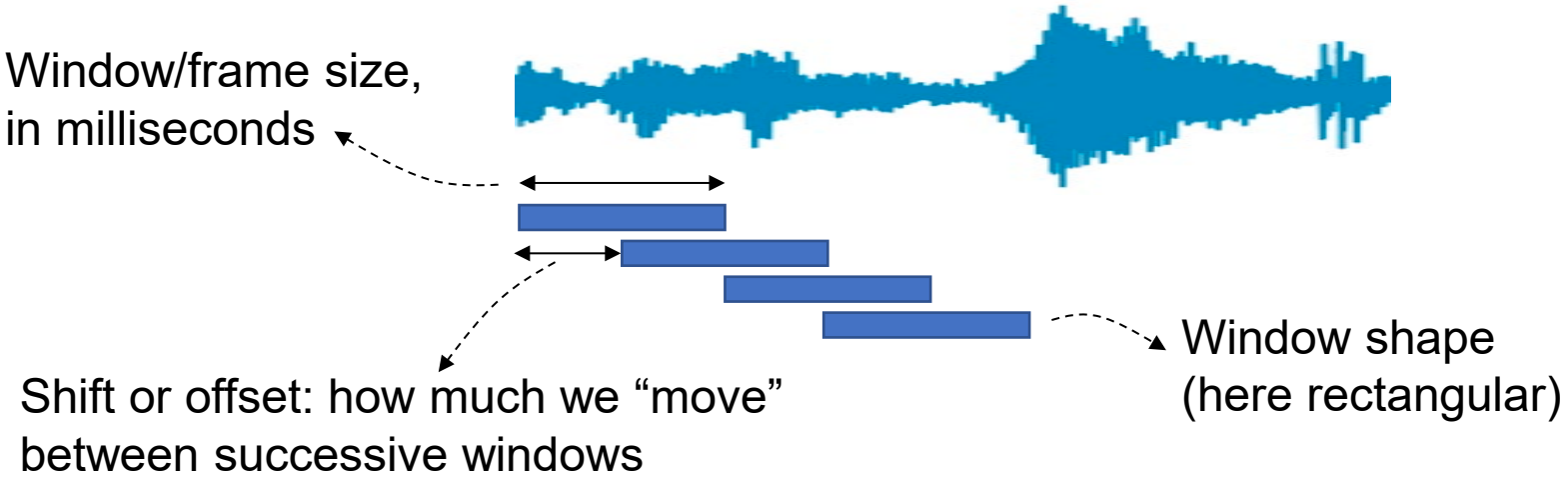
Feature extraction

- Most ASR models do not operate on the raw waveform, but on (long!) sequences of *acoustic feature vectors*
- **Step 1**: Conversion from analog signals (from microphone) to digital signals
  - **Sampling**: Measures of signal amplitude (x times per second)
  - **Quantization**: Storage in integers of given precision (8 or 16 bit)



Feature extraction

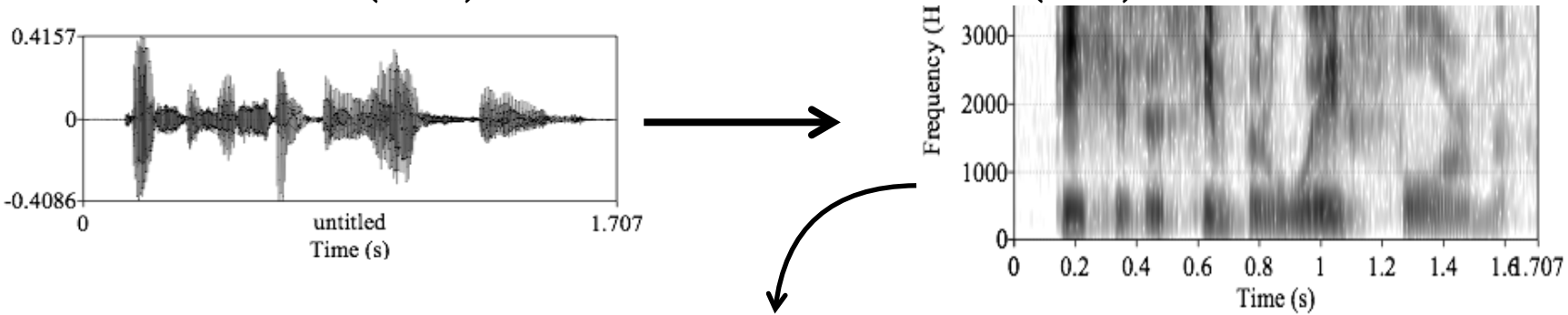
- **Step 2**: Windowing = Division of the signal into frames
  - NB: The frames may overlap (to ensure no information is lost)



Feature extraction

Step 3: spectral analysis

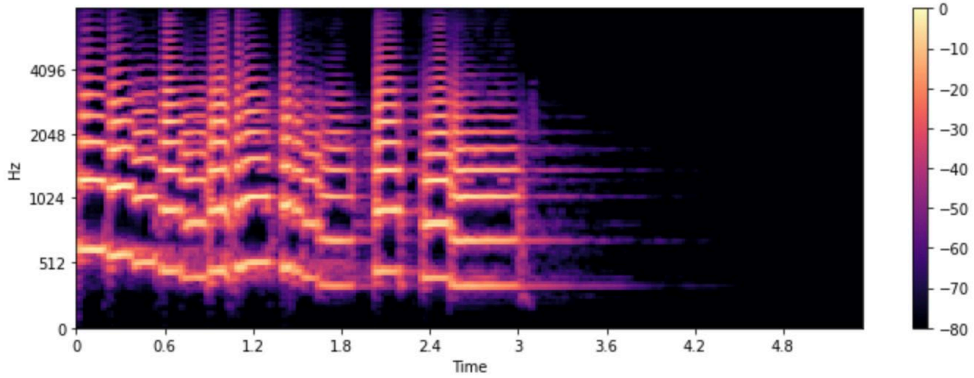
- = Convert the signal to a representation of the signal's *component frequencies*
- Based on the (Fast) Discrete Fourier Transform (DFT)



spectrogram showing which frequencies are most active at a given time

Feature extraction

- **Step 4**: Log Mel-filter bank: conversion of the spectral values to get better resolution at low frequencies
    - Mimic the sensitivity of the human ear

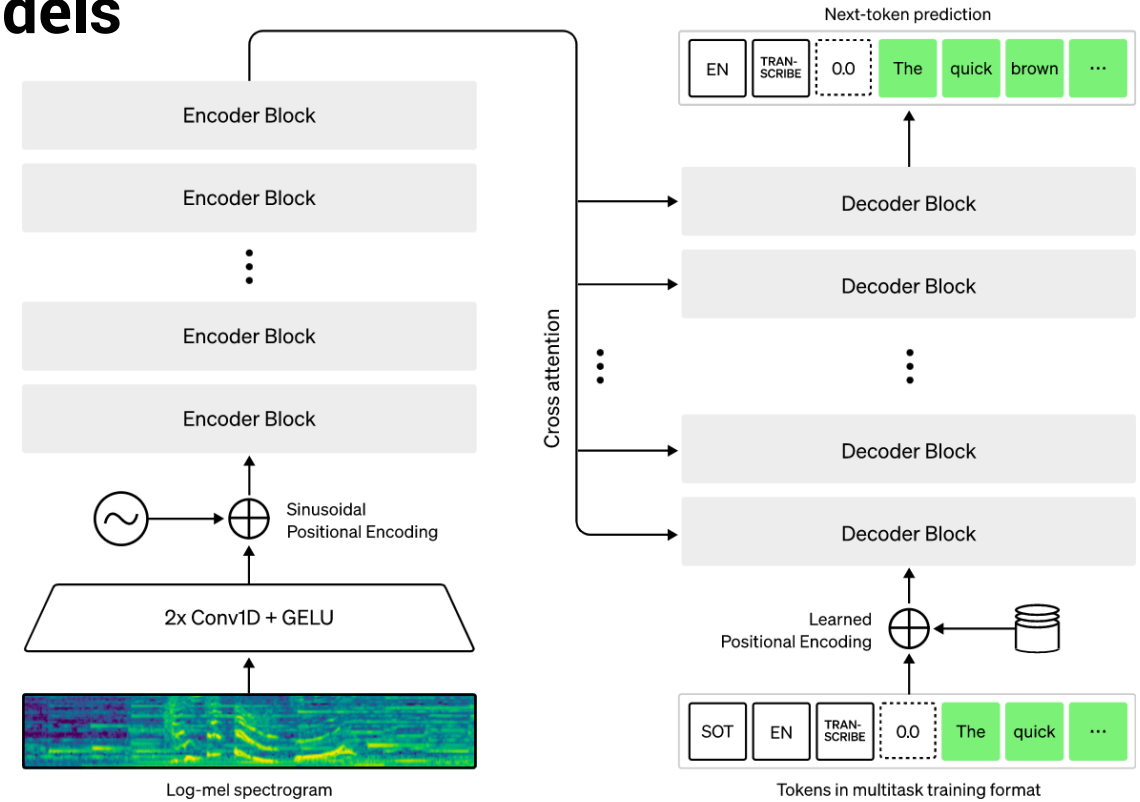


- Once those preprocessing steps are complete, we end up with a *long sequence of frames*, each frame being expressed by a *feature vector*

Encoder-decoder models

ASR Models like Whisper rely on an *encoder-decoder* transformer

- **Input**: Sequence of feature vectors (one per frame)
- **Output**: sequence of tokens or letter
- Input *much longer* than output !



Encoder-decoder training

- The ASR model is trained **end-to-end** on <audio, transcription> examples
- As usual, the loss function is the **cross-entropy loss**

  Measure of the error between

  - the next-token distribution predicted by the model

  - the actual token (from gold standard transcript)

- The model is applied to each token w_i of the transcription
  - **Teacher forcing**: when predicting the token at position *i,* the input is the audio and the previous tokens w_1, … w_{i-1} *from the actual transcription*

Language model rescoring

- A classical ASR model is only trained from (audio, transcription) pairs
- If there are enough transcriptions, the model will have learned a language model (conditioned on audio input) from the transcriptions
- But it will have seen drastically less text data than a “full” LLM
- We can use an LLM to rescore the recognition hypotheses:

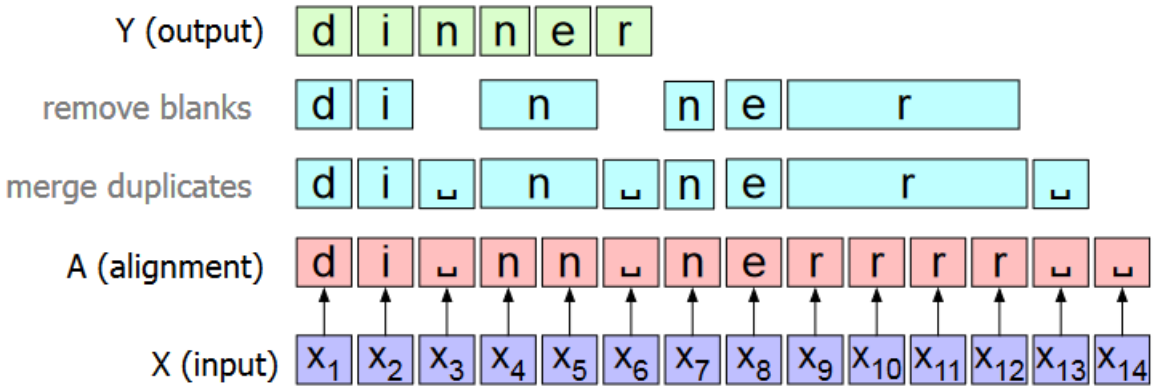
$$score(Y|X) = \frac{1}{|Y|_c} \log P(Y|X) + \lambda \log P_{LM}(Y)$$

From the ASR

Weight of the LLM rescoring

CTC speech recognition

- Once can also do speech recognition with a pure encoder model
- Idea: we predict a letter/symbol for each audio frame
- As there is no one-to-done mapping, we merge duplicate letters

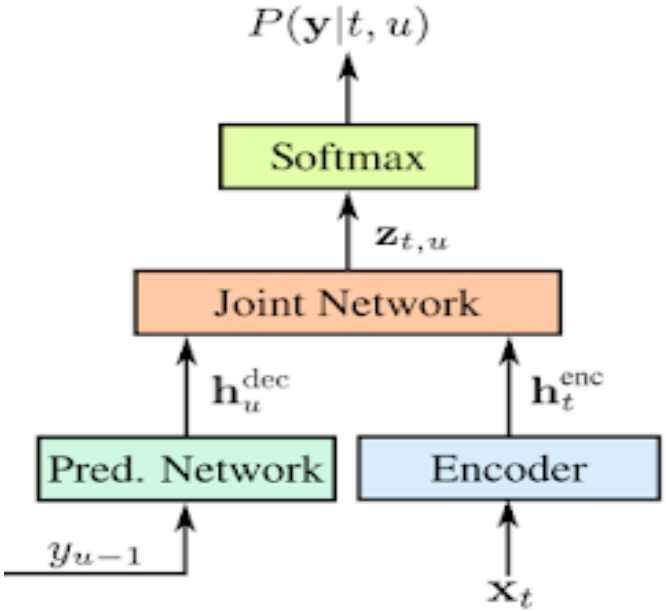


We use a special “blank“ symbol to prevent the merge of double letters
(and for frames where we don’t want to transcribe anything)

CTC speech recognition

https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html

- CTC models do not reach the same performance as encoder-decoders, but they are easier to apply to *streaming audio*
- Example of a relatively simple system: Google's 2019 on-device ASR
  - *Encoder* maps audio signal $\mathbf{x}_t$ to hidden representations with CTC
  - *Pred. Network* is a language model
  - The two representations are then merged, and the output is predicted character-by-character



ASR evaluation

Common metric: **Word Error Rate (WER)**

- Measures how much the utterance hypothesis h differs from the «gold standard» transcription t*
- = Minimum edit distance between h and t*, counting the number of word substitutions, insertions and deletions:

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Number of words in transcription}}$$

ASR evaluation

Gold standard transcription	*yes can you now rotate this triangle*
ASR hypothesis	*yes can you not rotate this triangle there*

$$\text{WER} = 100 \times \frac{1\ \text{Sub} + 1\ \text{Ins}}{7} = 28.6\%$$

Gold standard transcription	*there is five and*
ASR hypothesis	*the size and*

$$\text{WER} = 100 \times \frac{2\ \text{Sub} + 1\ \text{Del}}{4} = 75\%$$

Outline

- Short intro to phonetics and acoustics
- Speech recognition
- **Speech synthesis**
- Disfluencies

Text-to-Speech

- Modern TTS based on either:
  - Encoder-decoder models (but this time from string → waveforms)
  - Decoder-only models (with so-called “speech tokens”)
  - Diffusion-models (like in image generation)
- Three steps:
  - **Text normalization**: verbalization of non-standard words
  - **Spectogram prediction**: from input strings to mel spectrograms
  - **Vocoding**: from mel spectrograms to waveforms

Text normalisation

semiotic class	examples	verbalization
abbreviations	**gov't**, *N.Y., mph*	government
acronyms read as letters	**GPU**, *D.C., PC, UN, IBM*	G P U
cardinal numbers	**12**, *45, 1/2, 0.6*	twelve
ordinal numbers	*May* **7**, *3rd, Bill Gates III*	seventh
numbers read as digits	*Room* **101**	one oh one
times	*3.20*, **11:45**	eleven forty five
dates	**28/02** *(or in US, 2/28)*	February twenty eighth
years	**1999**, *80s, 1900s, 2045*	nineteen ninety nine
money	**$3.45**, *€250, $200K*	three dollars forty five
money in tr/m/billions	**$3.45 billion**	three point four five billion dollars
percentage	**75%** *3.4%*	seventy five percent

Figure 16.15 Some types of non-standard words in text normalization; see Sproat et al. (2001) and (van Esch and Sproat, 2018) for many more.

Text normalisation

semiotic class	examples	verbalization
abbreviations	**gov't**, *N.Y., mph*	government
acronyms read as letters	**GPU**, *D.C., PC, UN, IBM*	G P U
cardinal numbers	**12**, *45, 1/2, 0.6*	twelve
ordinal numbers	*May* **7**, *3rd, Bill Gates III*	seventh
numbers read as digits	*Room* **101**	one oh one
times	*3.20*, **11:45**	eleven forty five
dates	**28/02** *(or in US, 2/28)*	February twenty eighth
years	**1999**, *80s, 1900s, 2045*	nineteen ninety nine
money	**$3.45**, *€250, $200K*	three dollars forty five
money in tr/m/billions	**$3.45 billion**	three point four five billion dollars
percentage	**75%** *3.4%*	seventy five percent

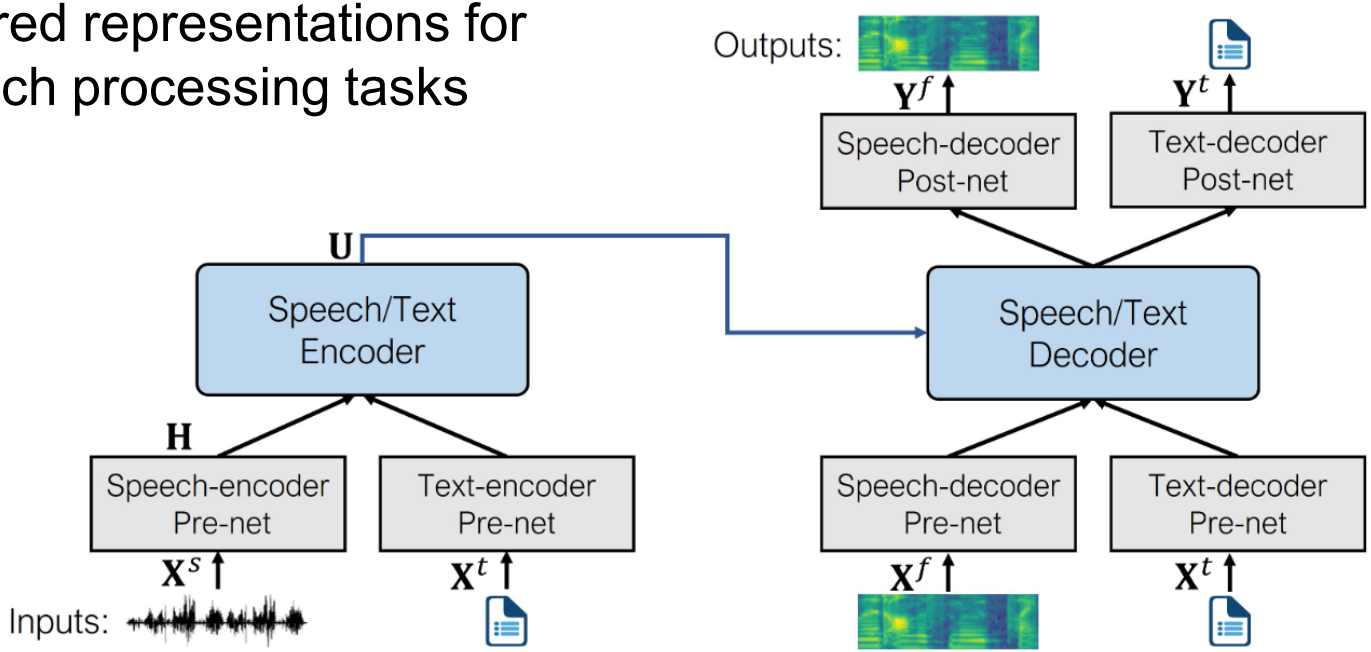
Figure 16.15 Some types of non-standard words in text normalization; see Sproat et al. (2001) and (van Esch and Sproat, 2018) for many more.

Can be implemented with

- handcrafted rules
- fine-tuned LLMs,
- or hybrid systems combining the two (using rules to constrain or steer the LLM outputs)

SpeechT5

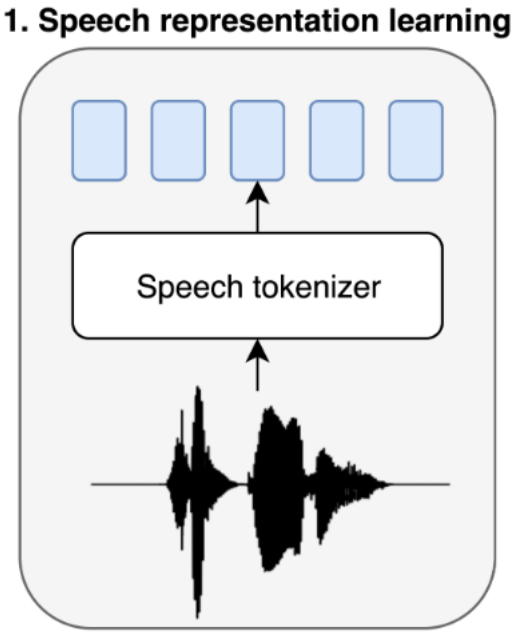
→ Using shared representations for multiple speech processing tasks



(a) The model architecture of SpeechT5

Decoder-only TTS

- One can also build a TTS model using a pure decoder-only approach
- To do this, we must discretise our speech signal, to express the full input (= text input and speech already synthesized) as a sequence of tokens
- We can train a speech tokenizer from audio data using an autoencoding approach
- The output of the speech tokenizer is a sequence of speech tokens, also called **speechcodes**

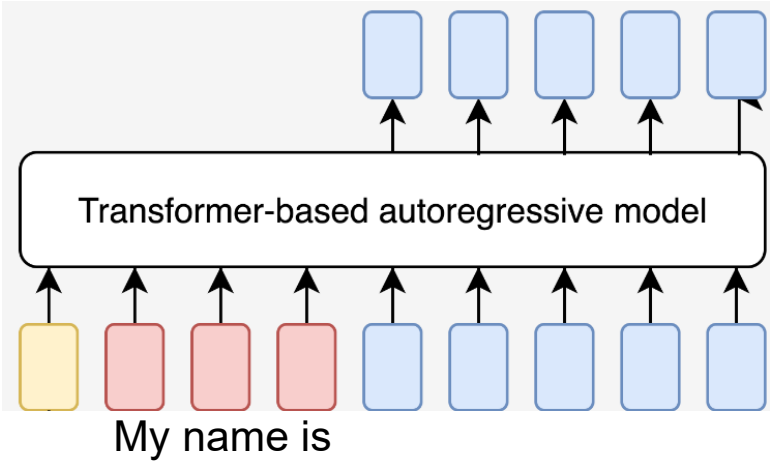


[Łajszczak, M. et al (2024). BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data. *arXiv preprint arXiv:2402.08093*.]

Decoder-only TTS

When we have a speech tokenizer, we can formalize TTS as an autoregressive task where we generate speechcodes one by one, conditioned on:

- The text tokens corresponding to the string to synthesise
- The speechcodes generated until now
- Context information about the speaker and/or speaking style



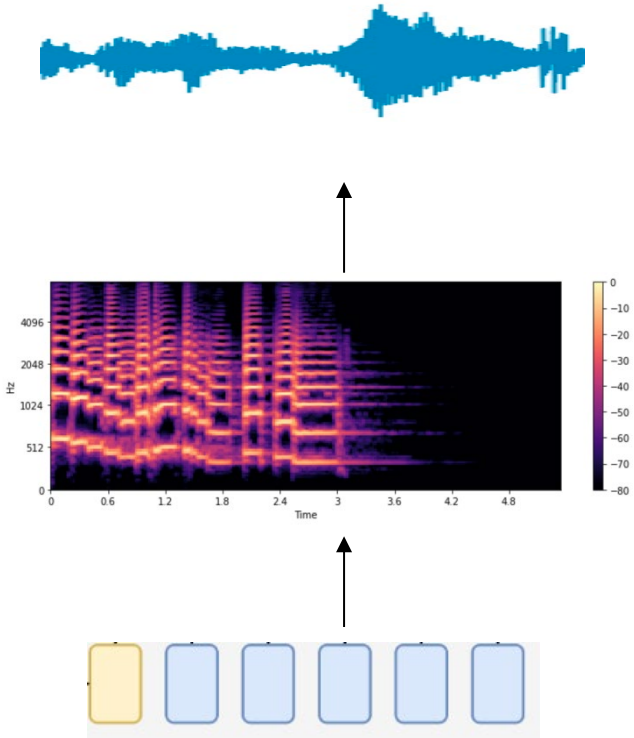
Yellow: context embedding
Red: text tokens
Blue: speech tokens

Can be a speech encoding extracted from another utterance from the same speaker (like in BASE TTS)

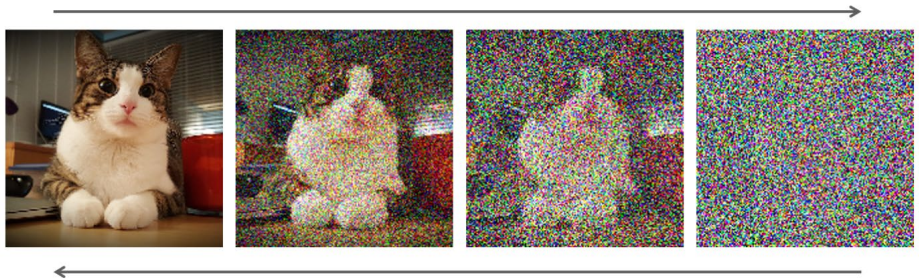
Decoder-only TTS

One the sequence of speechcodes is generated, it can be mapped to a waveform in two final steps:

- Converting the speechcodes (and a speech encoding that captures the user voice) to a mel spectrogram
- A *vocoder* converting the spectrogram to a waveform



Diffusion-based TTS



- Diffusion models used in many generative ML tasks, notably *image generation*
  - Intuition: start with random noise and gradually «denoise it»
- Diffusion can also be applied for TTS
  - Gradual «denoising» of the mel-spectogram, based on the text prompt
- A recent diffusion-based TTS: **NaturalSpeech 3** from Microsoft

Outline

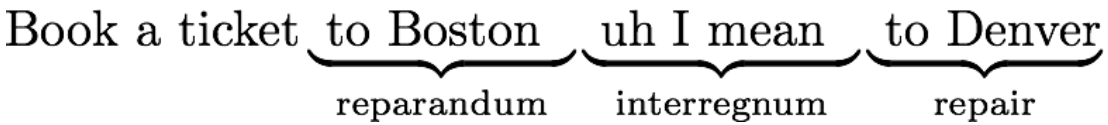
- Short intro to phonetics and acoustics
- Speech recognition
- Speech synthesis
- **Disfluencies**

Disfluencies

- Speakers construct their utterances «as they go», incrementally
  - = Production leaves a *trace* in the speech stream
- Presence of multiple disfluencies
  - Pauses, fillers («øh», «um», «liksom»)
  - Repetitions («the the ball»)
  - Corrections («the ball err mug»)
  - Repairs («the bu/ ball»)

Disfluencies

Internal structure of a disfluency:



- reparandum: part of the utterance which is edited out
- interregnum: (optional) filler
- repair: part meant to replace the reparandum

[Shriberg (1994), «Preliminaries to a Theory of Speech Disfluencies», Ph.D thesis]

Some disfluencies

[«*Norske talespråkskorpus - Oslo delen*» (NoTa), collected and annotated by the Tekstlaboratoriet]

så gikk jeg e flytta vi til Nesøya da begynte jeg på barneskolen der
og så har jeg gått på Landøya ungdomsskole # som ligger ## rett over broa nesten # rett med Holmen

jeg gikk på Bryn e skole som lå rett ved der vi bodde den gangen e barneskole
videre på Hauger ungdomsskole

da hadde alle hele på skolen skulle liksom # spise julegrøt og det va- det var bare en mandel
og da var jeg som fikk den da ble skikkelig sånn " wow # jeg har fått den " ble så glad