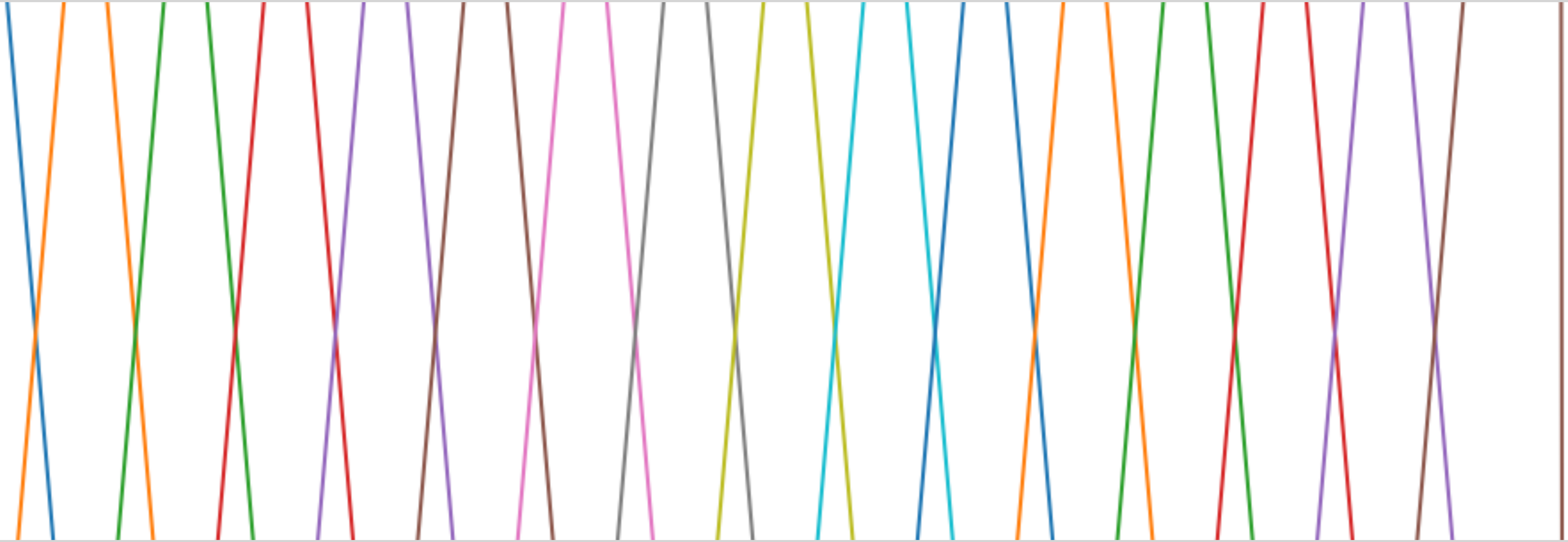
CSC3160 - Fundamentals of Speech and Language Processing



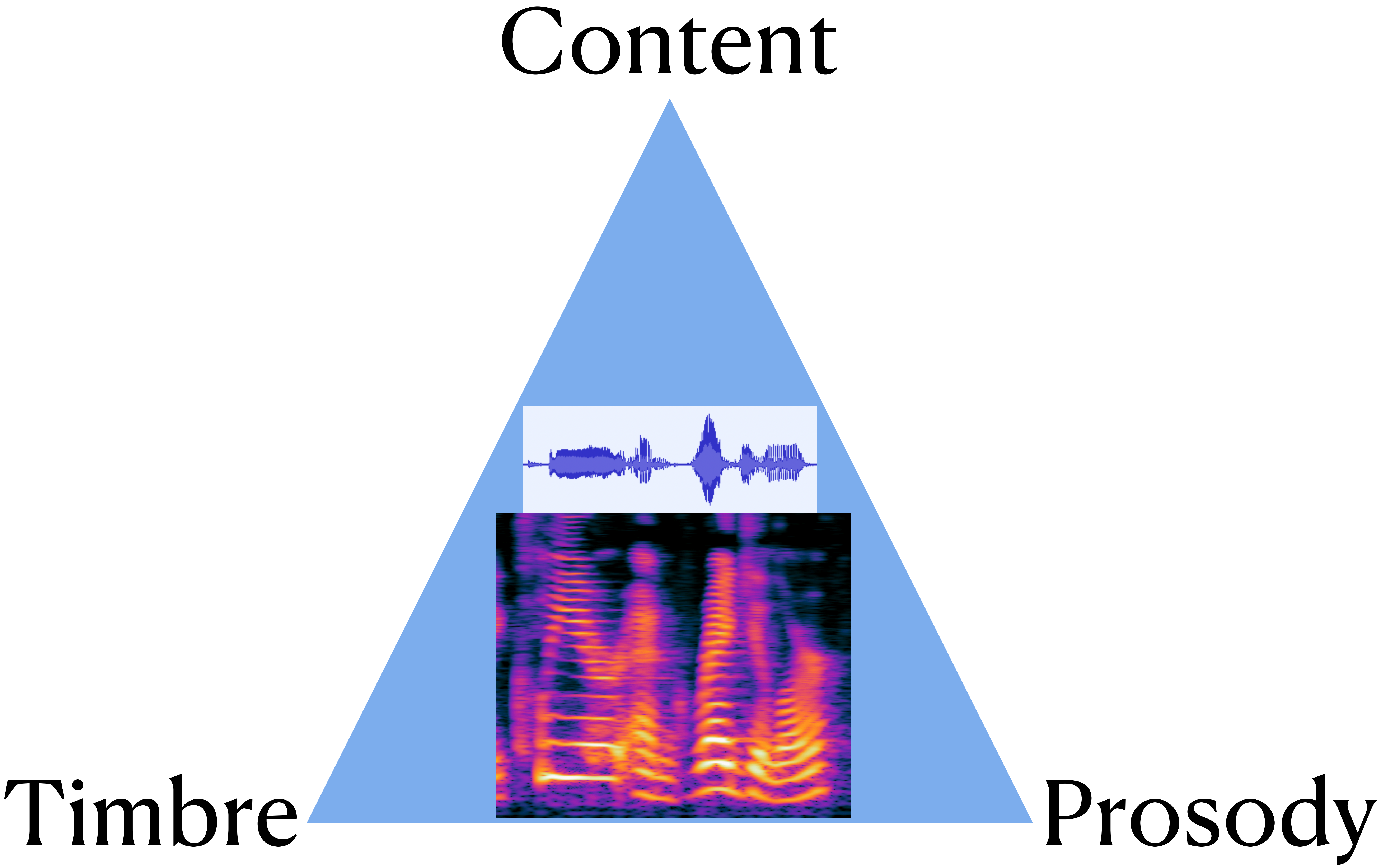
Lecture 5: Speech representation

Zhizheng Wu

https://drwuz.com/CSC3160/

Outline

- Information in human speech
- Speech production
- Source filter model
- Timbre
- Prosody

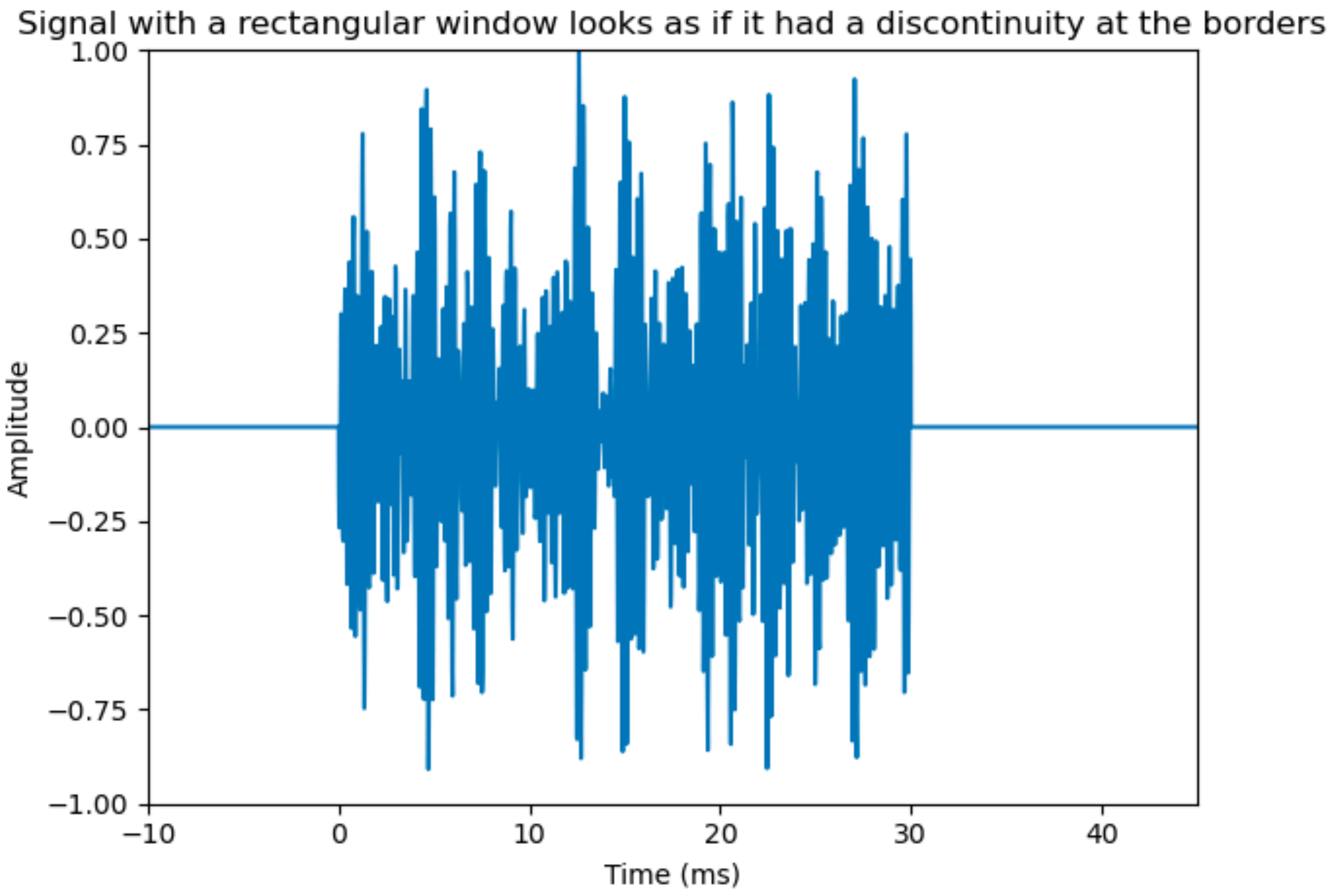


Content

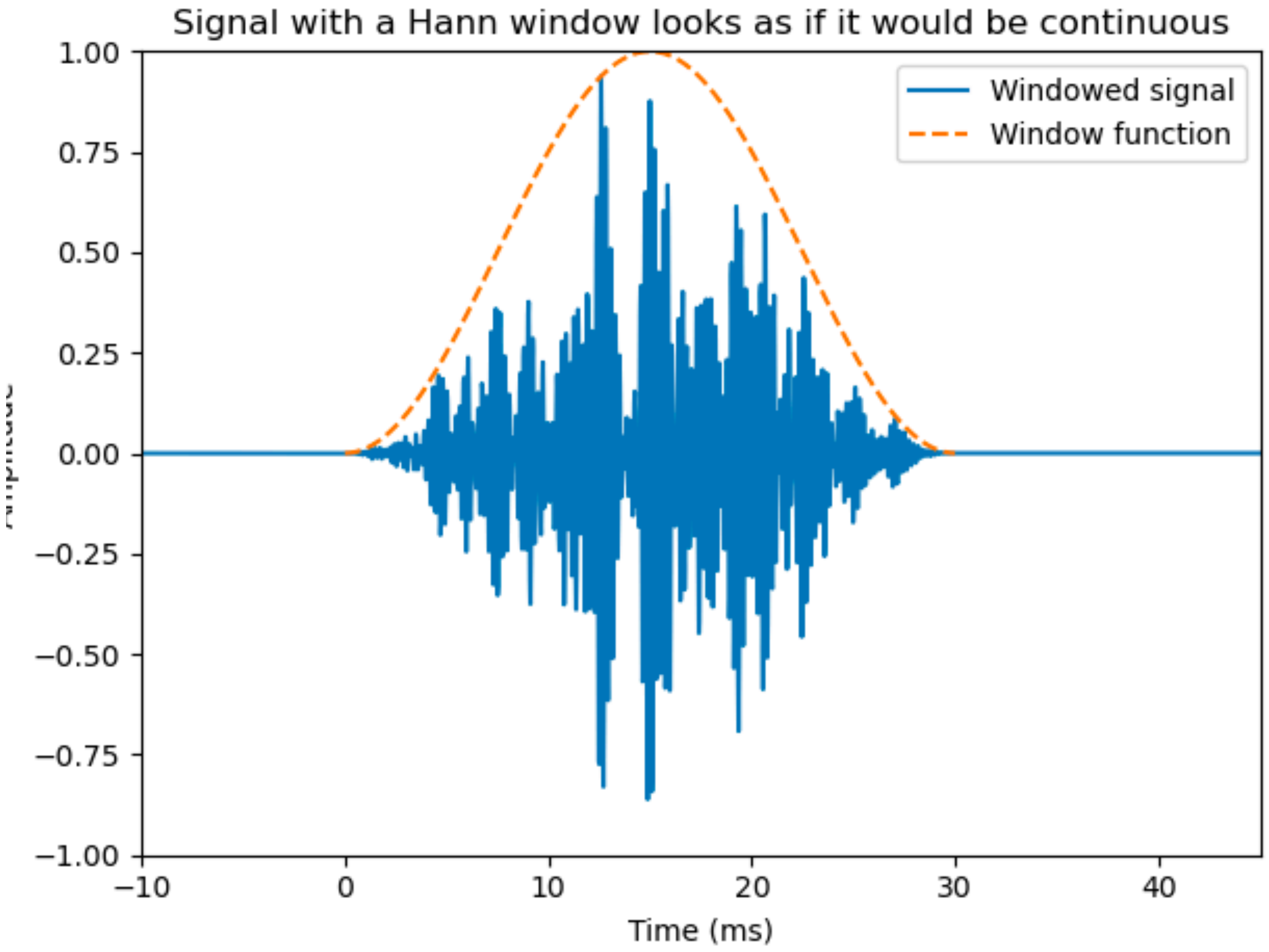
Timbre

Prosody

Waveform



Signal with a rectangular window looks as if it had a discontinuity at the borders



Signal with a Hann window looks as if it would be continuous

Prosody

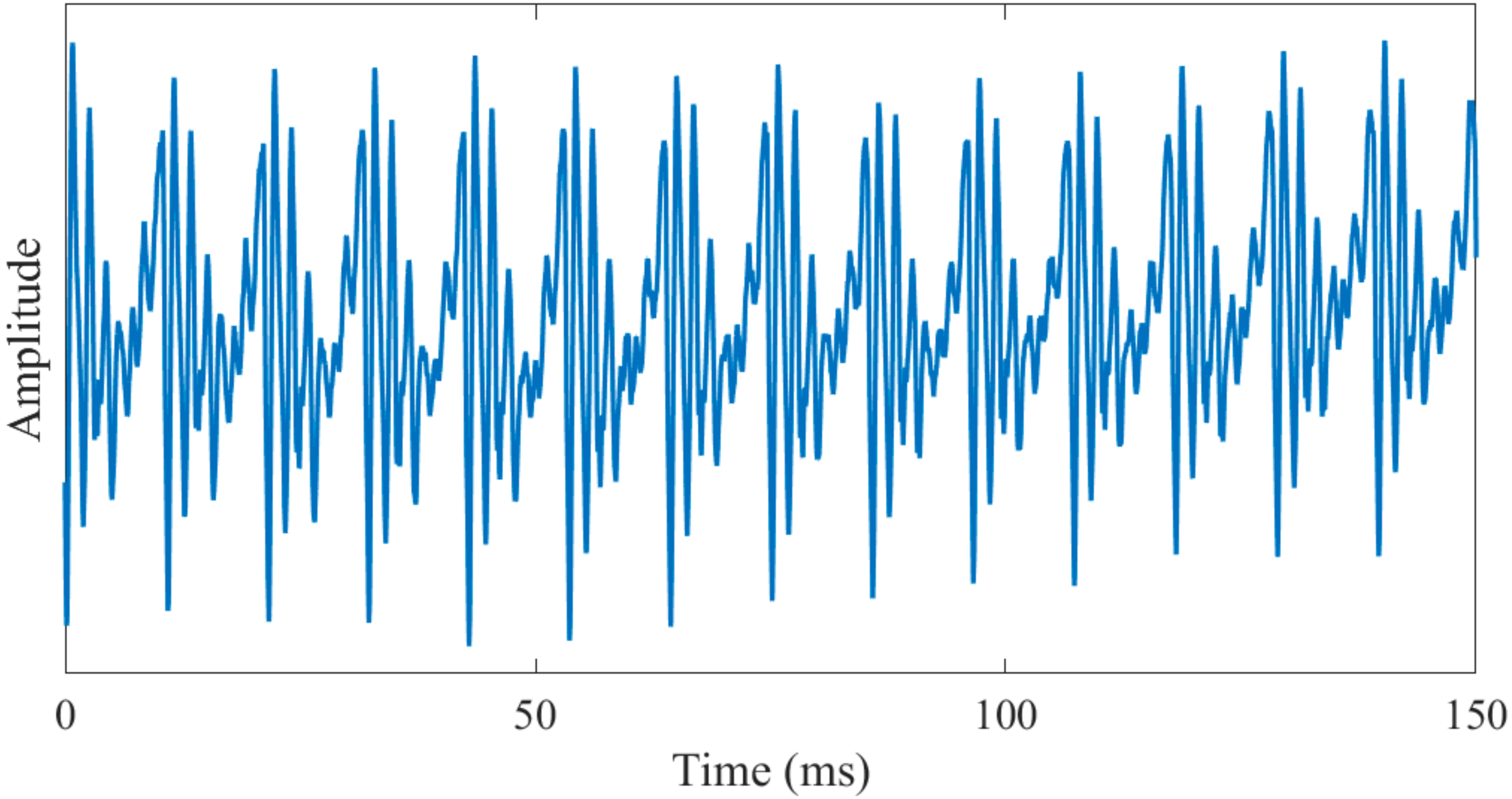
- Pitch
- Loudness
- Duration: Length of each segment (phone, syllable, word, phrase, etc)

Pitch

- Pitch is the perception of fundamental frequency
- Pitch describes how our ears and brains interpret the signal

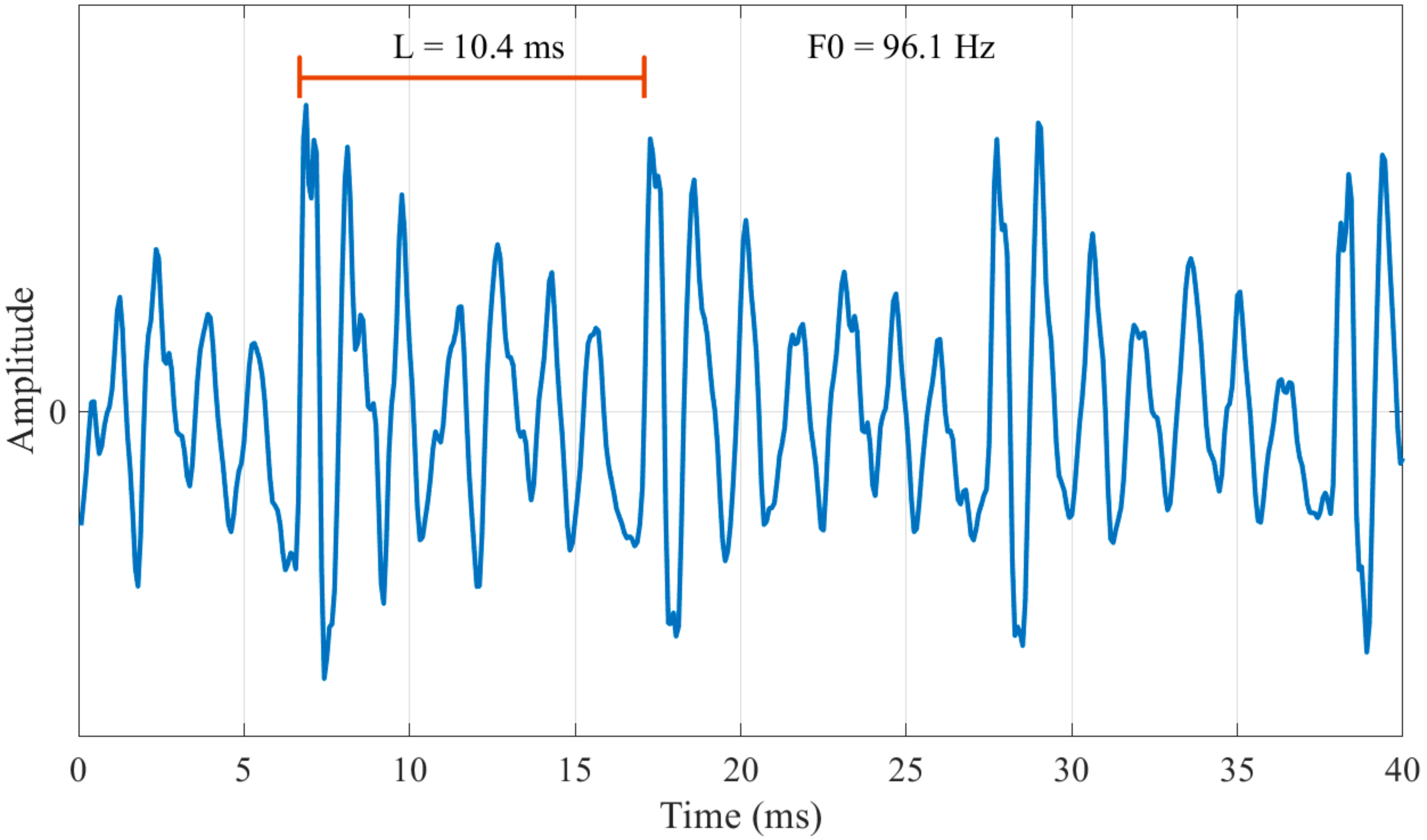
Fundamental frequency

- F0 of an individual speaker depends primarily on the length of the vocal folds
- F0 describes the actual physical phenomenon
- Typically F0 range 80 to 450 Hz



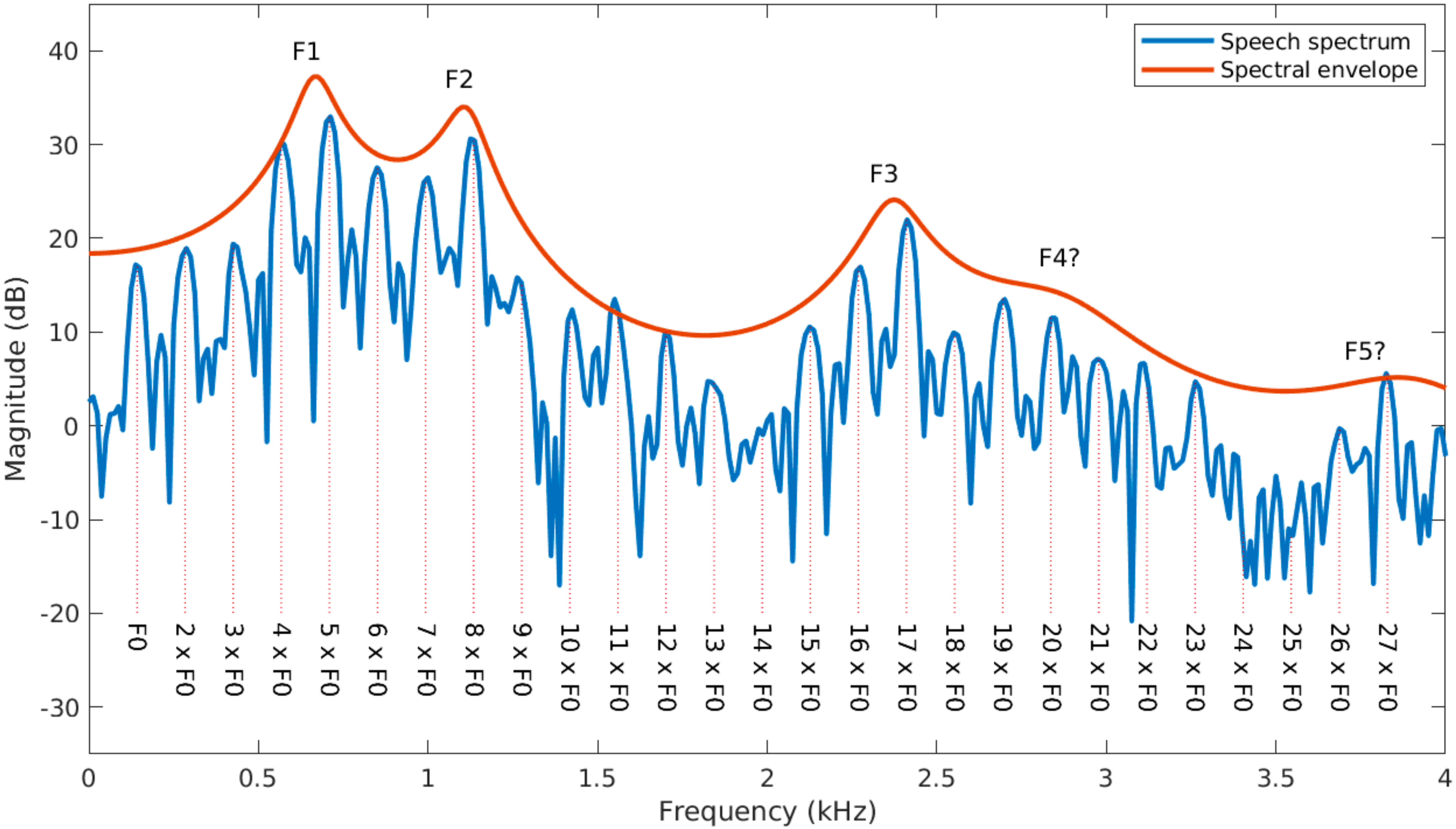
Fundamental frequency

- L: period length
- F0 = 1 / L

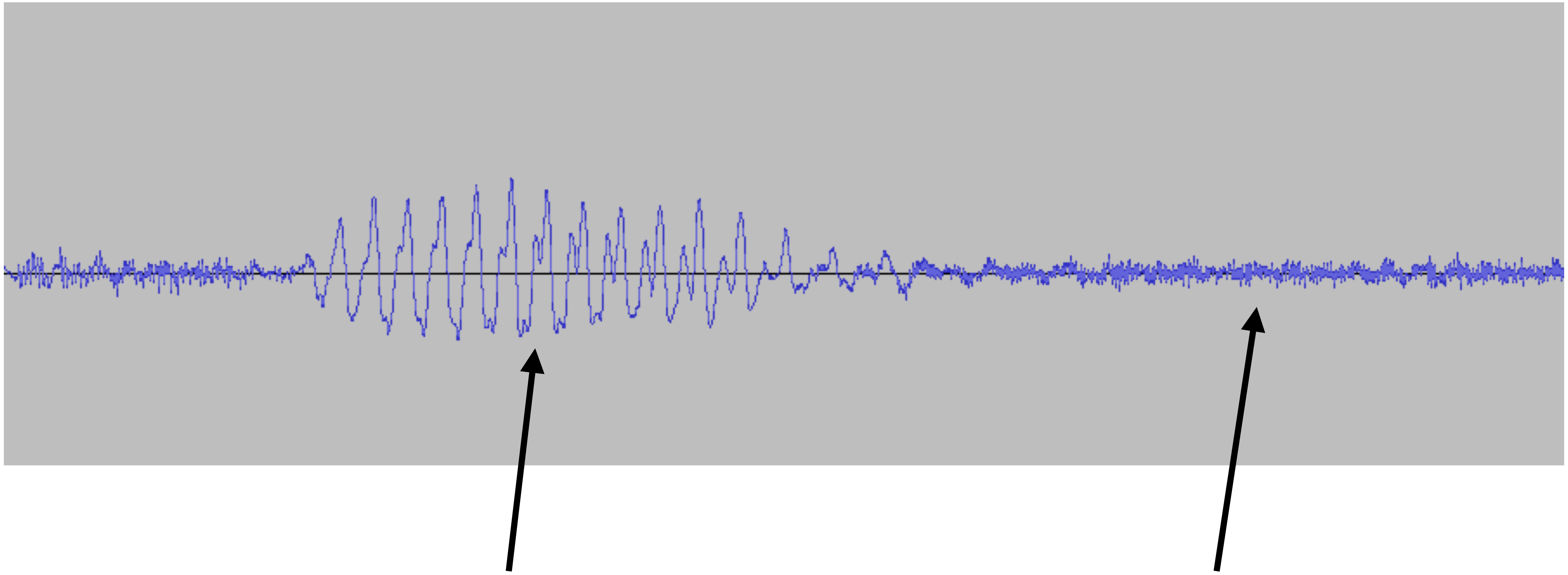


Fundamental frequency

- F_0 and harmonics kF_0



No F0 for unvoiced region



Voiced region

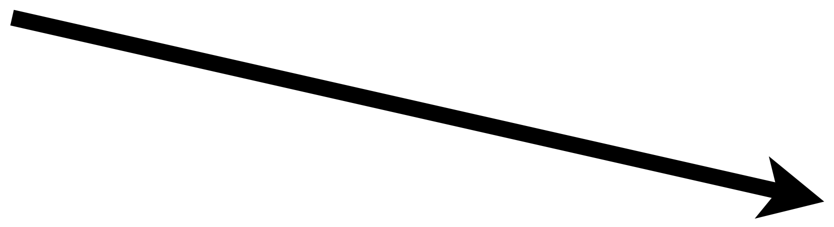
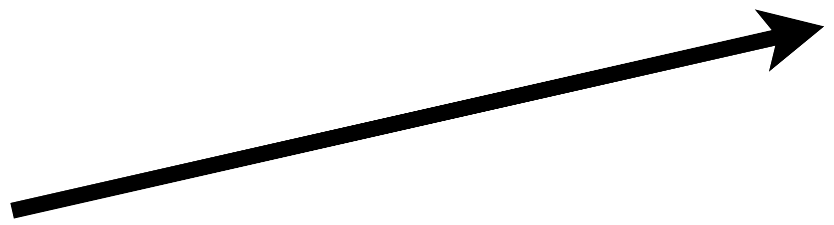
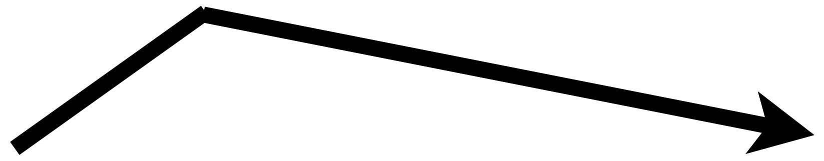
Unvoiced region

https://colab.research.google.com/drive/1j707gmlYED8r0AICb1Re-waULZ4DYWk6?usp=sharing

Intonation

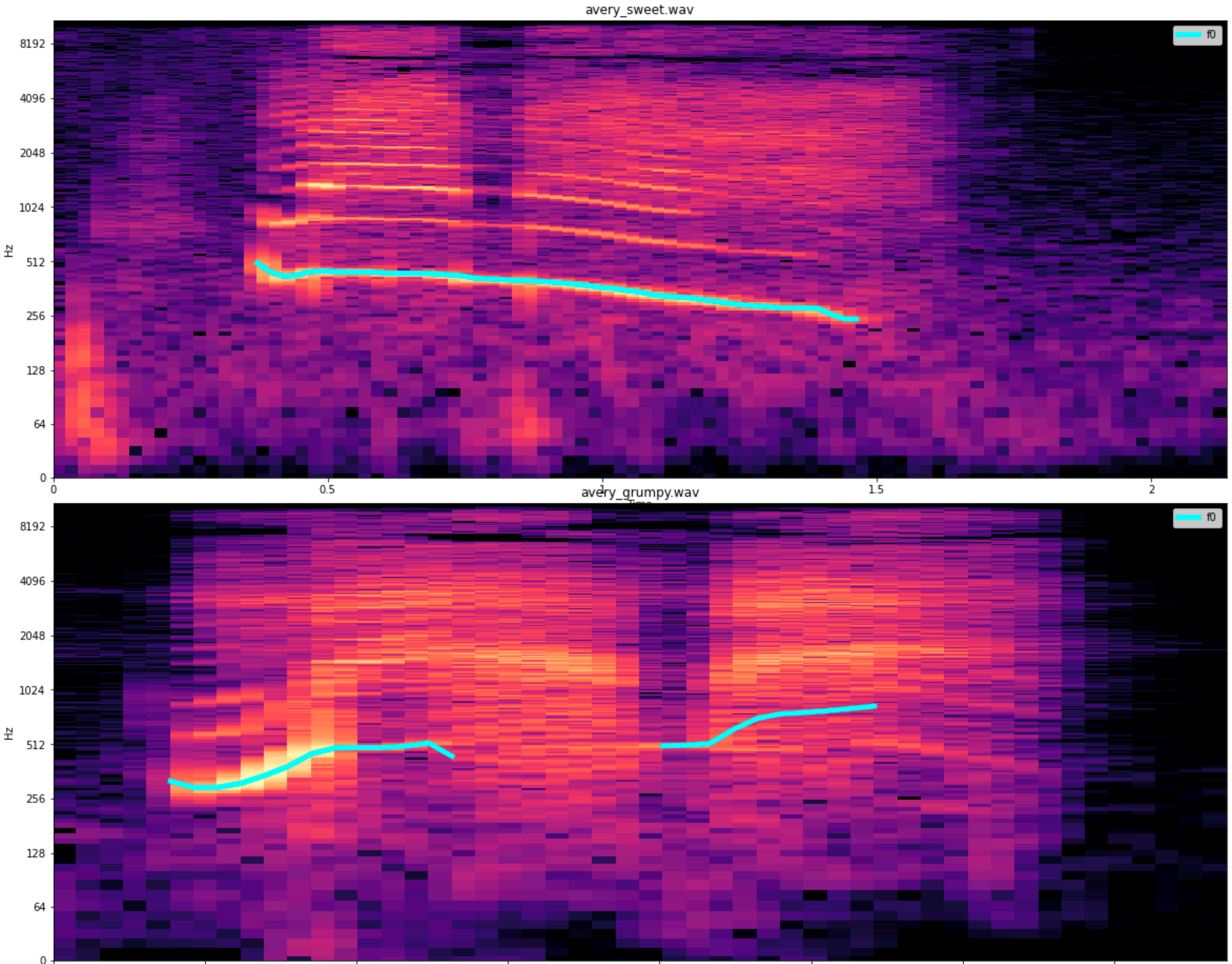
- Intonation is a complex system of meaning communicated through the rise and fall of a speaker's voice.
- Intonation can change the meaning of what a person says even when the same words are used.

Intonation

It was interesting	Communicative purpose and function	Audio
	You are giving information. You are certain and confident about the information.	
	This intonation could indicate that this is a question even though the grammar indicates a statement. It could also indicate that you aren't sure or that you haven't finished yet.	
	You want to emphasise this. Depending on the context, you may feel enthusiastic, happy or surprised. Or you may want to contrast this strongly with what someone else has said.	

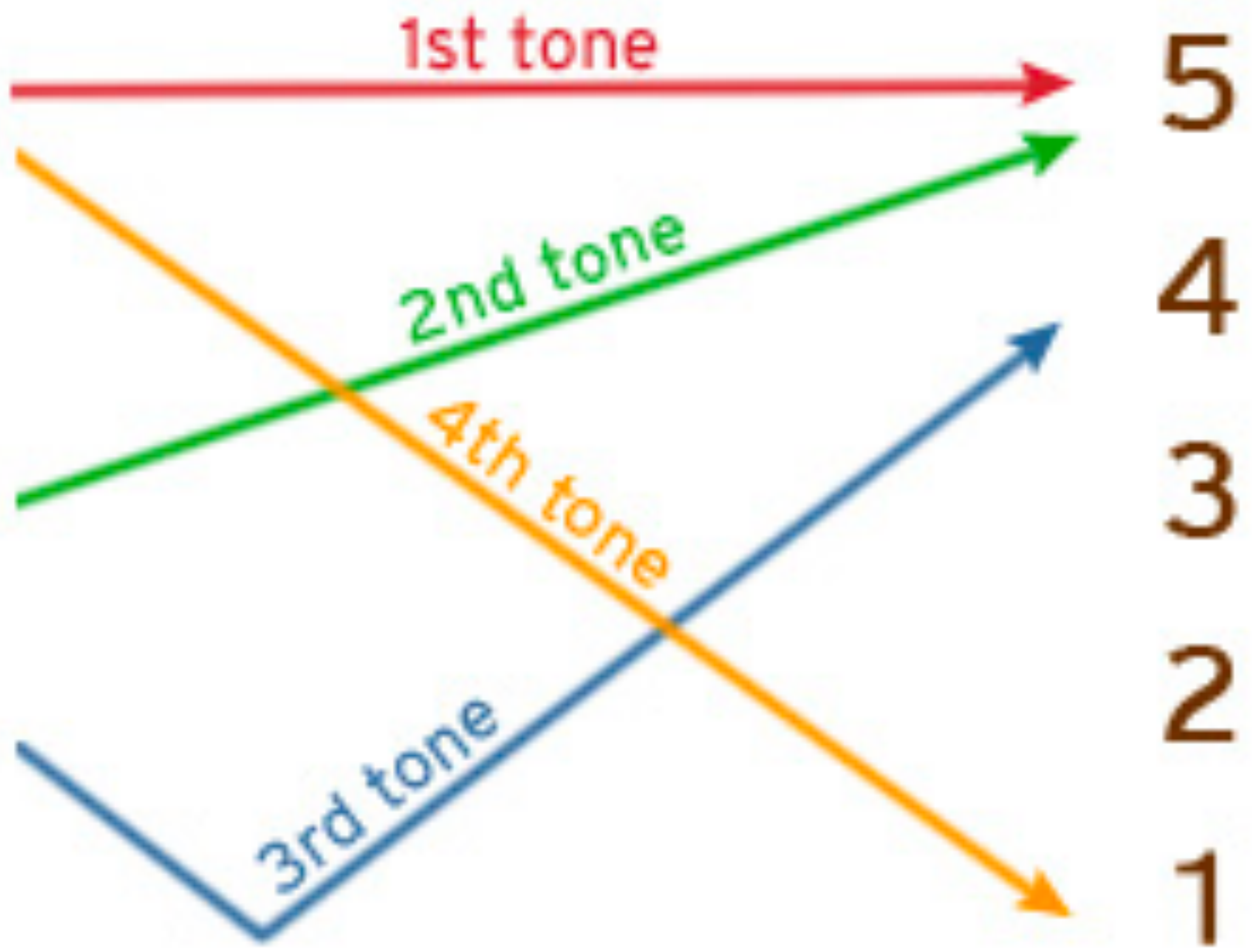
https://www.uts.edu.au/current-students/support/helps/self-help-resources/pronunciation/intonation

Intonation


avery_sweet.wav
f0
8192
4096
2048
1024
512
256
128
64
0
Hz
0
0.5
1
1.5
2
Time
avery_grumpy.wav
f0
8192
4096
2048
1024
512
256
128
64
0
Hz

Tone

- Tonal language: different tonal inflections will convey different meanings

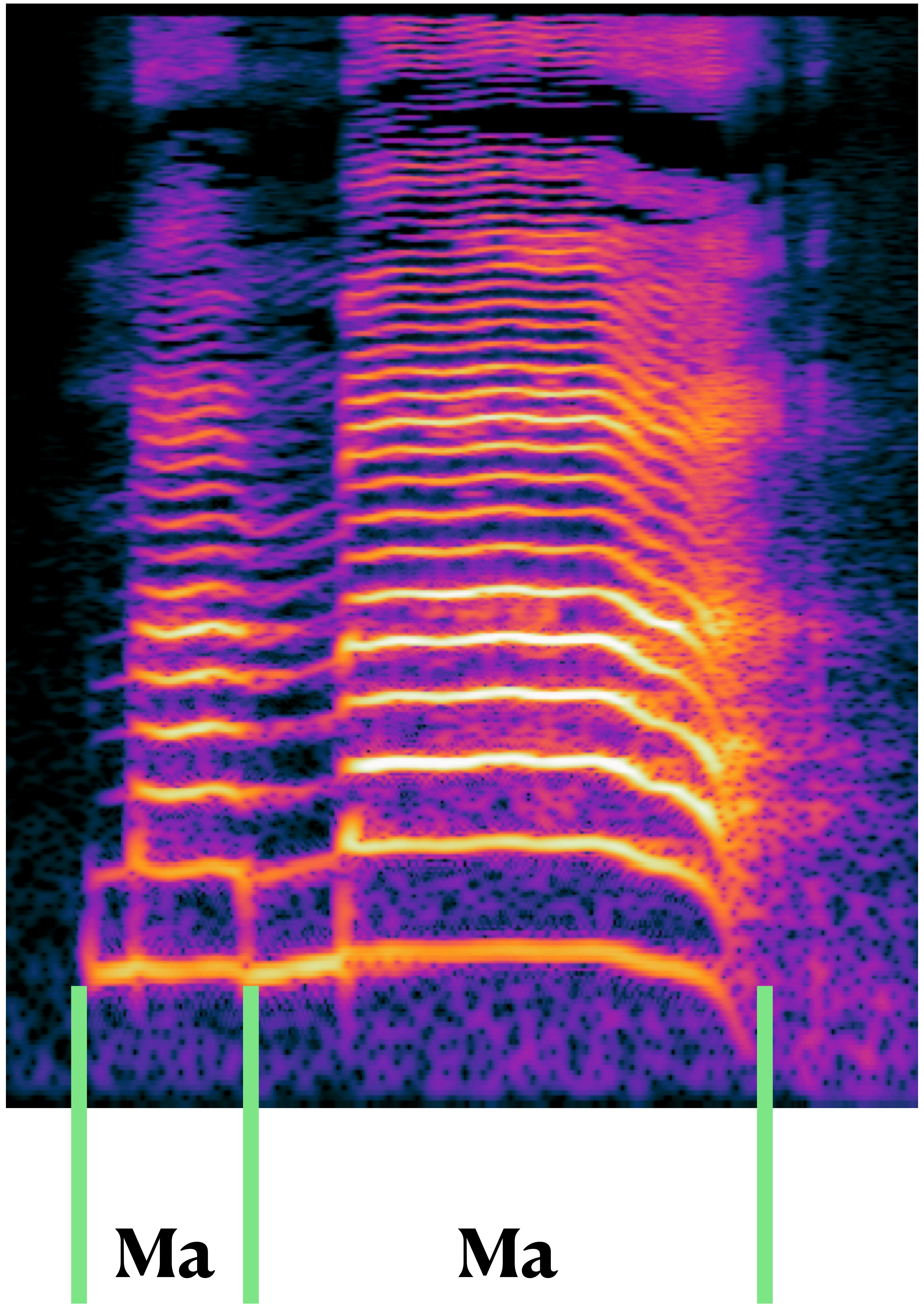


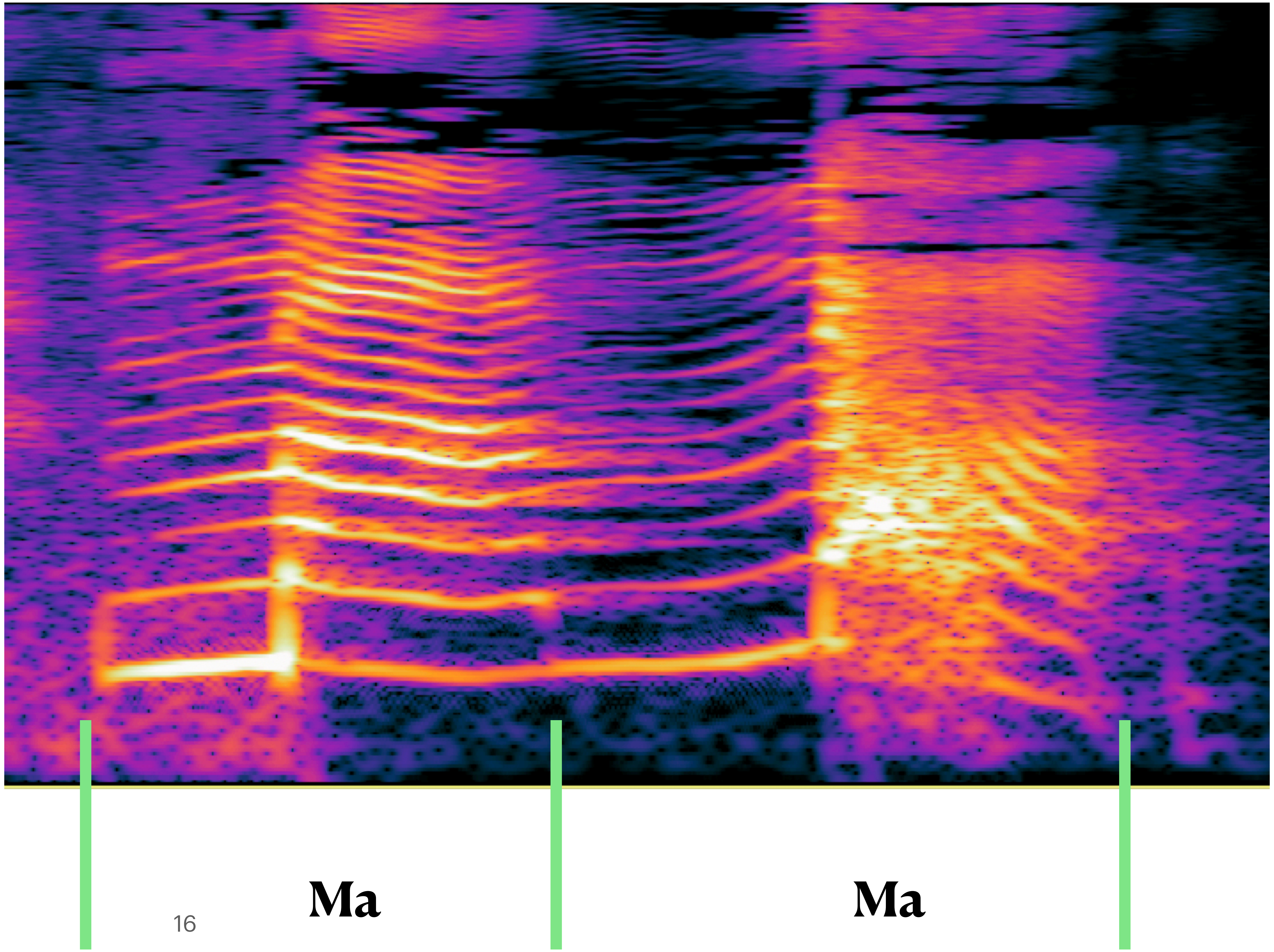
Duration

- Duration of speech sounds can help to convey meaning and differentiate between words

- Duration and boundaries of speech units are important feature for many downstream tasks
  - Speech recognition
  - Text-to-speech synthesis
  - etc

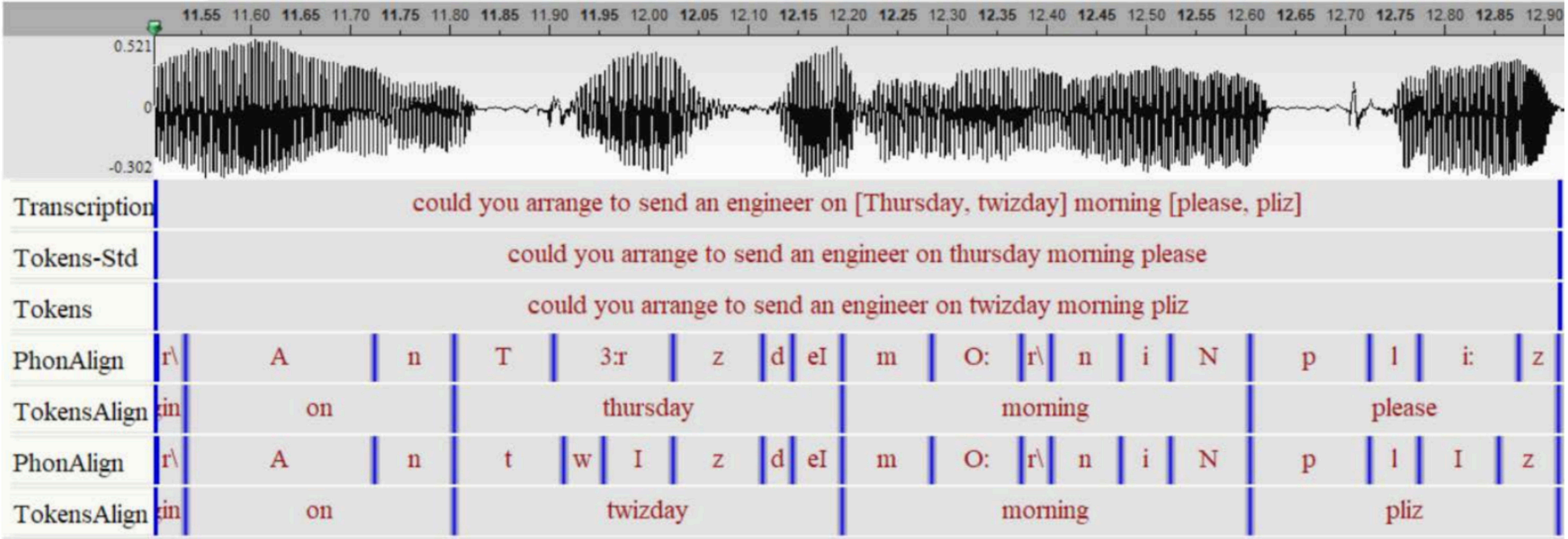
Duration





Duration

- Duration at different semantic levels



Transcription: could you arrange to send an engineer on [Thursday, twizday] morning [please, pliz]

Tokens-Std: could you arrange to send an engineer on thursday morning please

Tokens: could you arrange to send an engineer on twizday morning pliz

PhonAlign: r\ | A | n | T | 3:r | z | d | eI | m | O: | r\ | n | i | N | p | l | i: | z

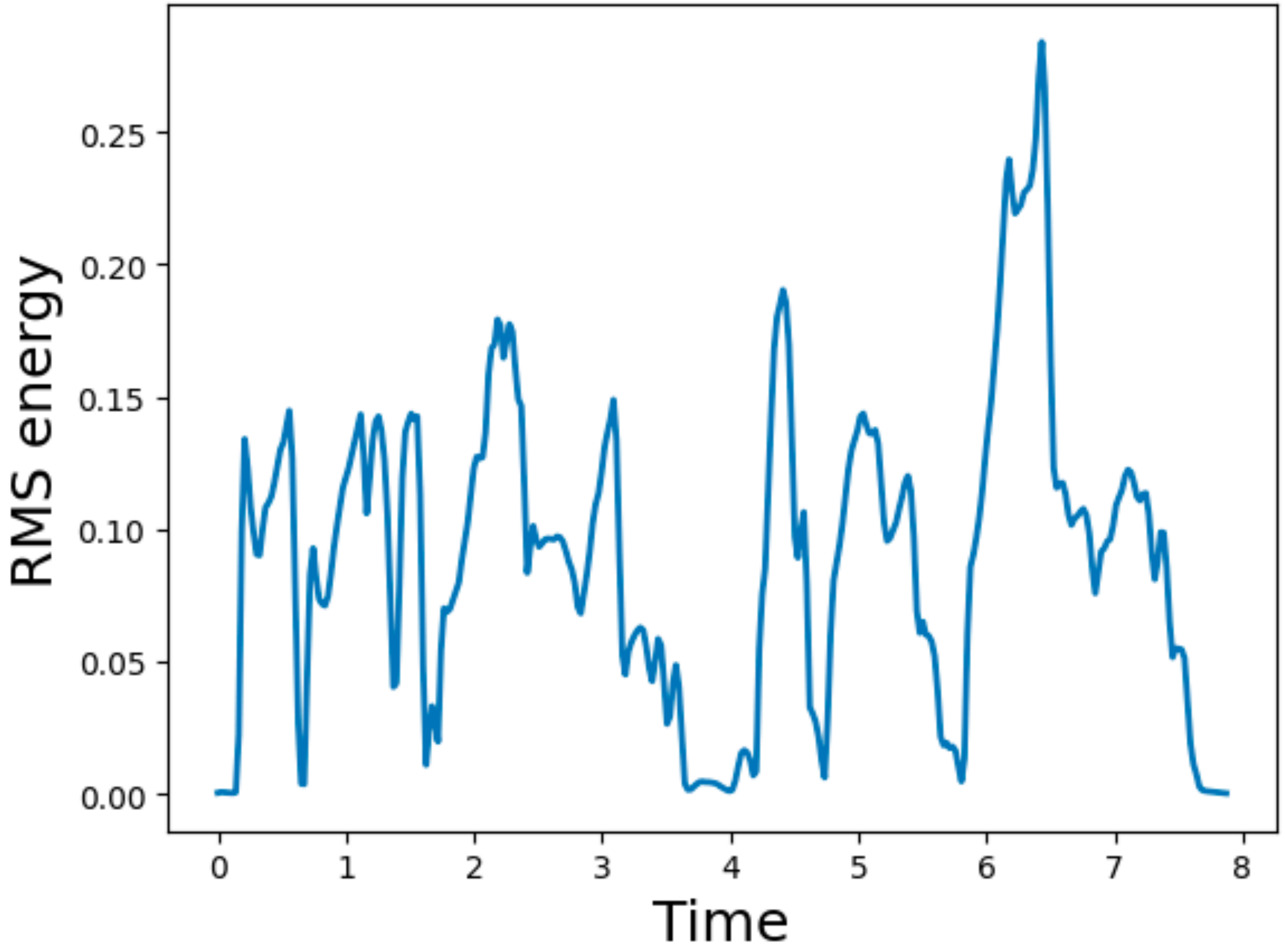
TokensAlign: in | on | thursday | morning | please

PhonAlign: r\ | A | n | t | w | I | z | d | eI | m | O: | r\ | n | i | N | p | l | I | z

TokensAlign: in | on | twizday | morning | pliz

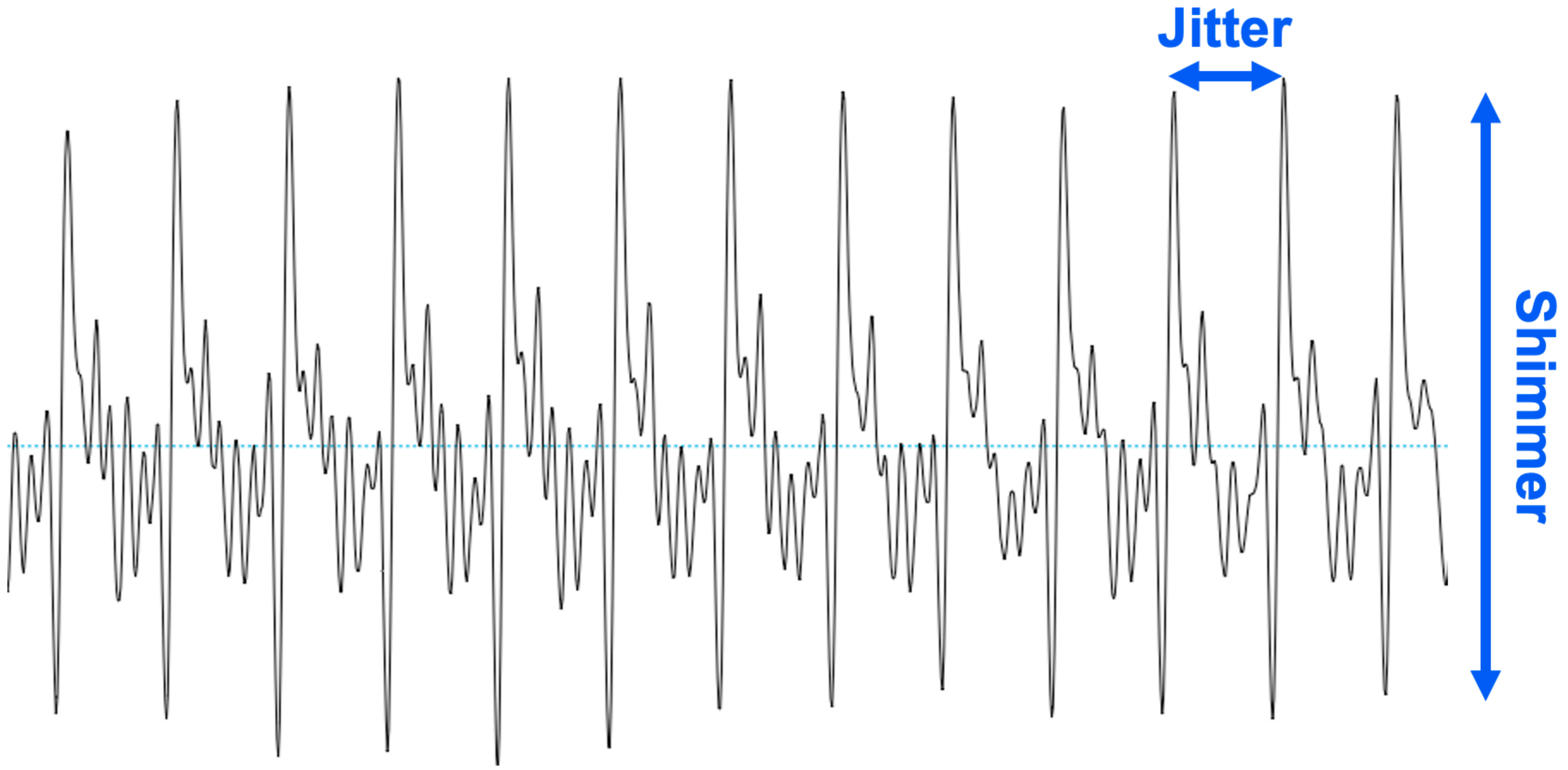
Energy

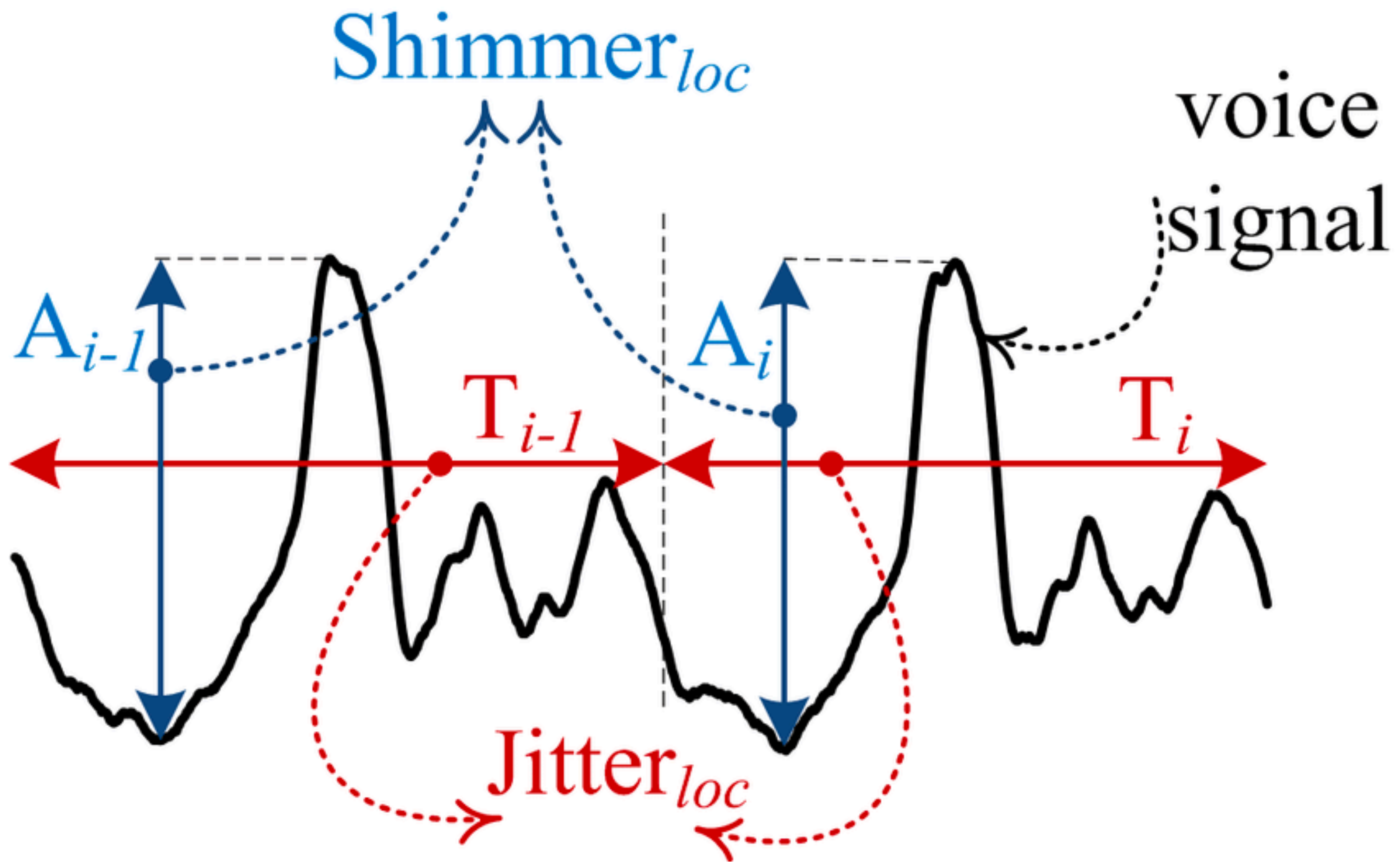
- Energy or intensity determines the loudness
- Loudness is perception of intensity or energy



Jitter and shimmer

- Jitter: Variations in signal frequency
- Shimmer: Variations in signal amplitude





Jitter and shimmer

- Jitter and shimmer are caused by irregular vocal fold vibration
  - Perceived as roughness, breathiness, or hoarseness in a speaker’s voice
  - Measuring them is a common way to detect voice pathologies

- Personal habits such as smoking or alcohol consumption might increase the level of jitter and shimmer in voice

Jitter

- A common way: Average absolute difference between consecutive periods

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} \|T_i - T_{i+1}\|$$

  - *Ti* are the extracted F0 period lengths and *N* is the number of extracted F0 periods

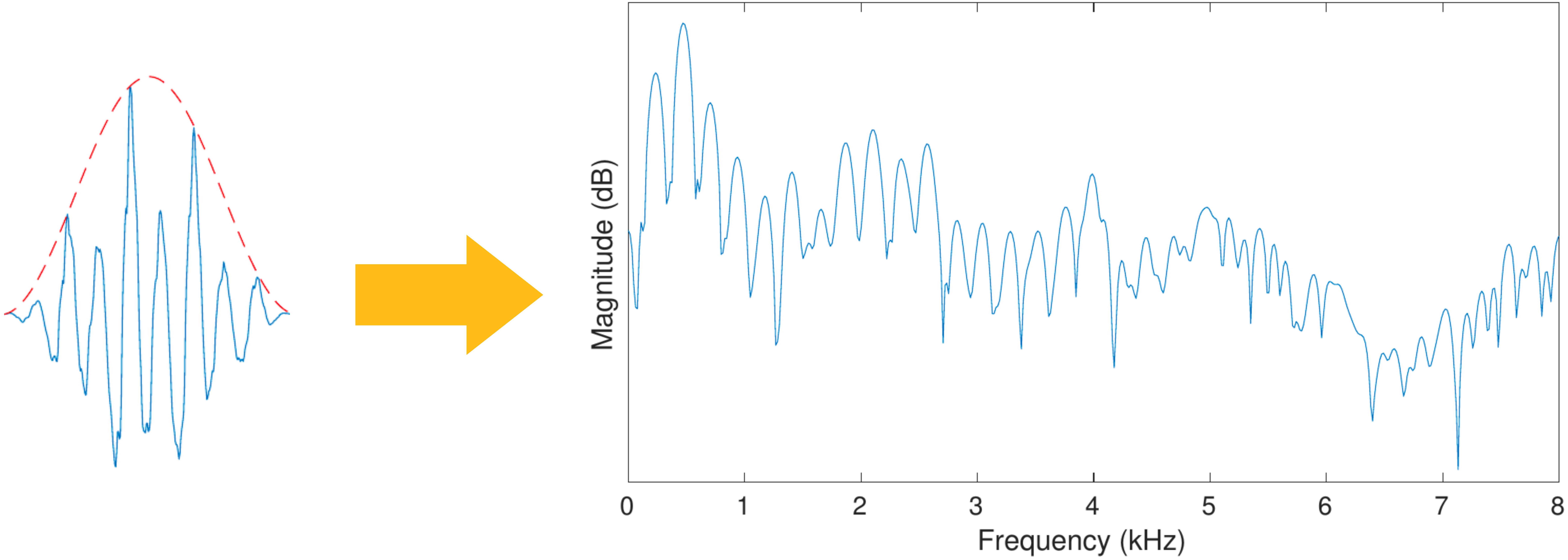
Shimmer

- The difference between the amplitudes of consecutive periods multiplied by 20

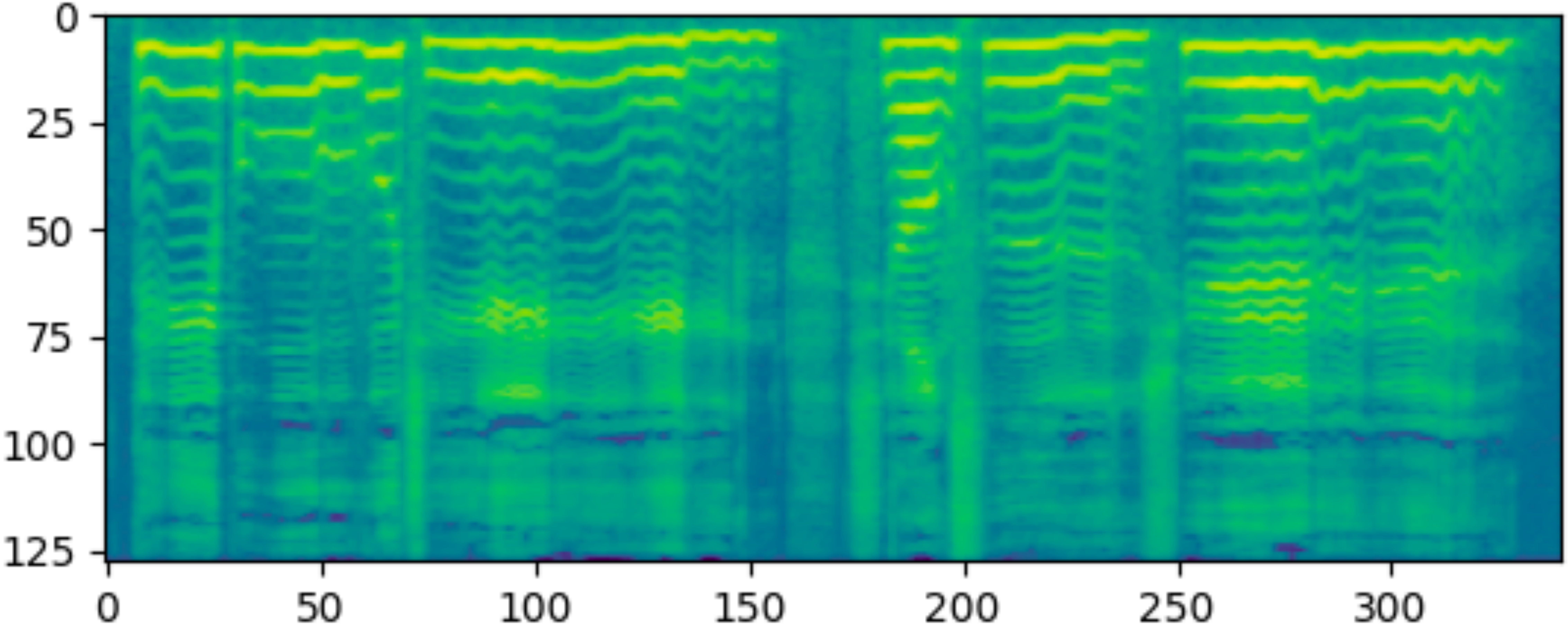
$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1} \|20\log(A_{i+1}/A_i)\|$$

  - *Ai* are the extracted peak-to-peak amplitude data and *N* is the number of extracted fundamental frequency periods

Spectrum



Spectrogram

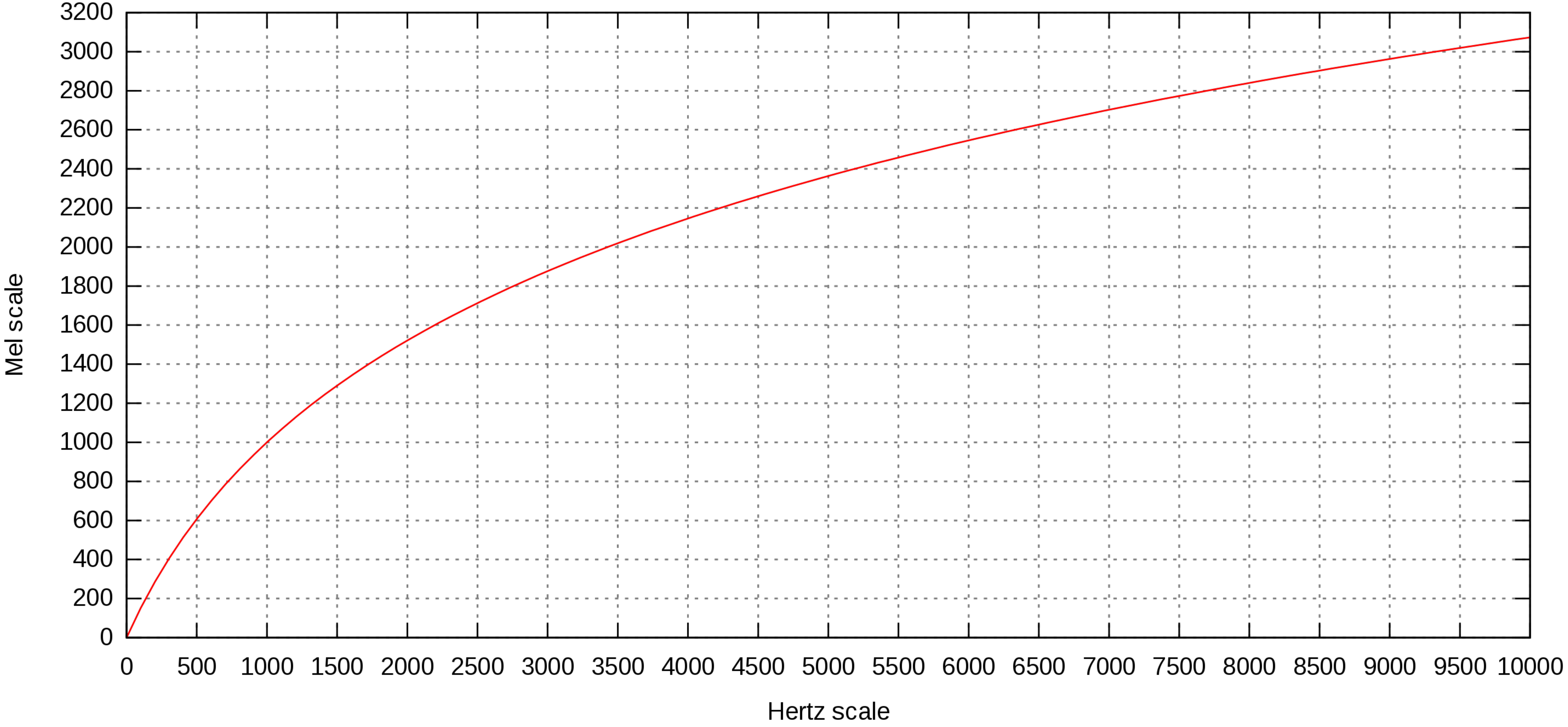


Mel scale

- Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another

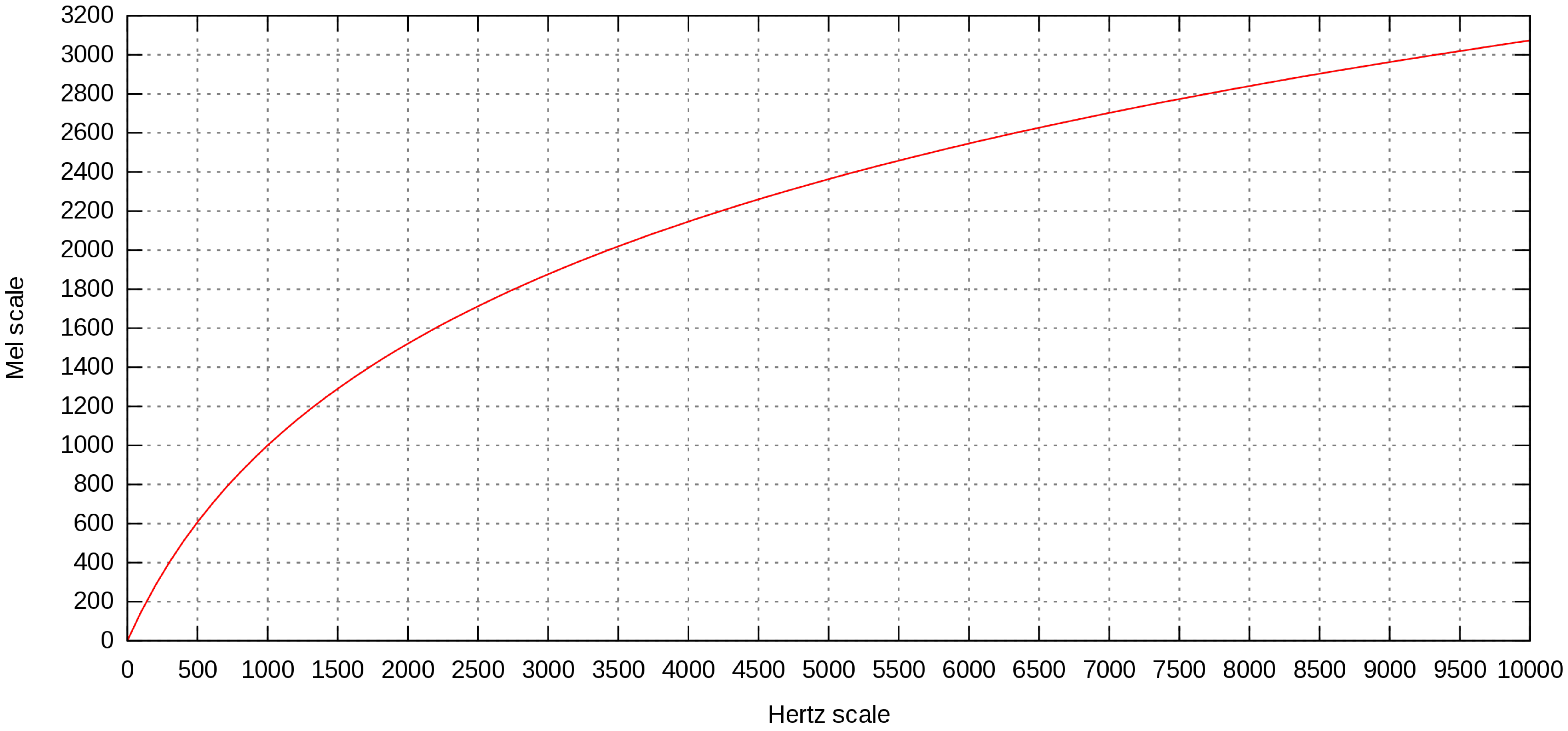
$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Mel scale


Mel scale
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
10000
Hertz scale

Mel scale

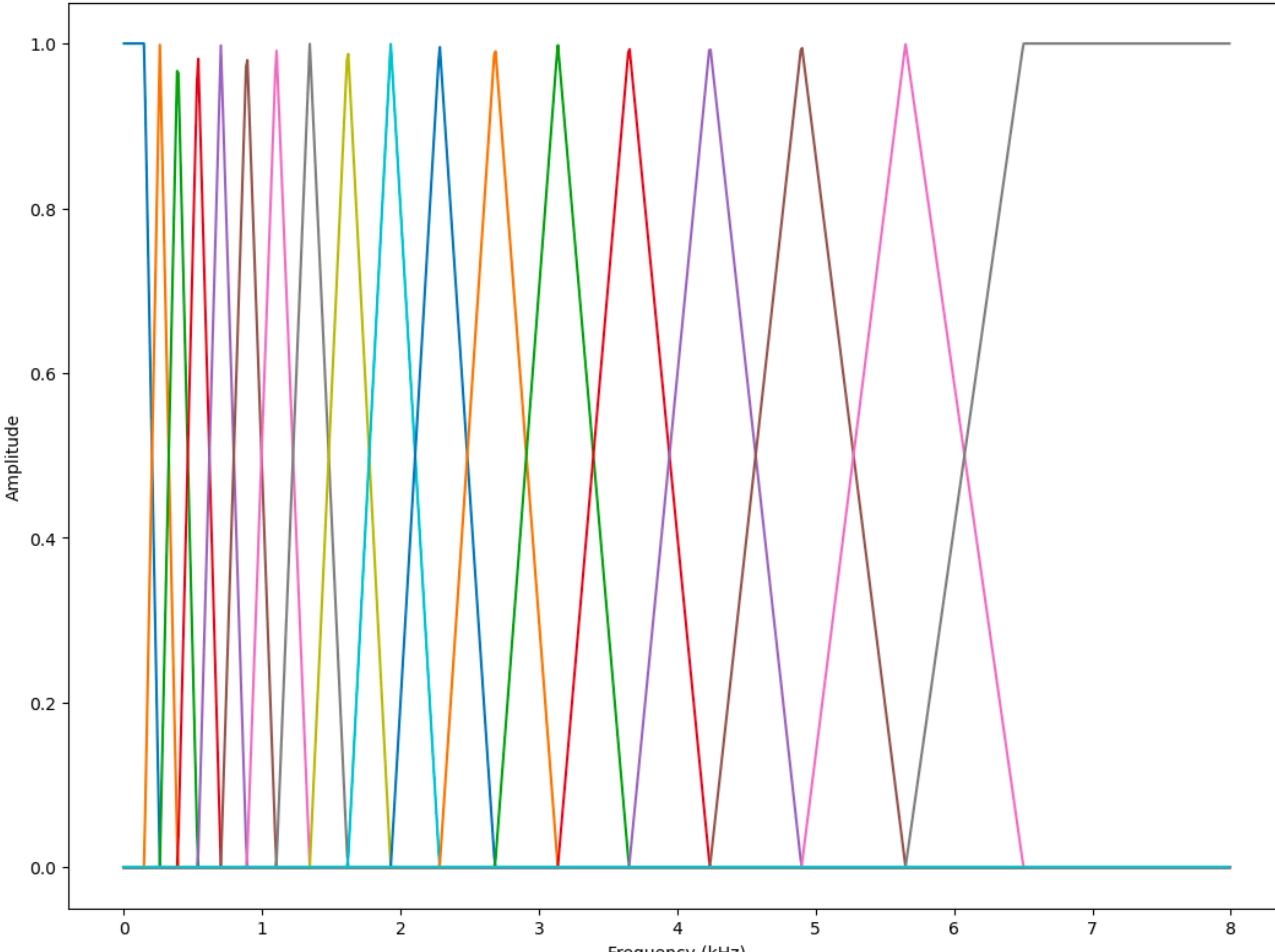
- Demo: Mel-scale from 200 to 1500, in intervals of 50



Mel filterbank

- Filterbank
  - triangle-centres are at the frequencies corresponding to equal distance steps on the mel scale

- Higher frequencies, above 6.5 kHz in particular, are poorly modelled

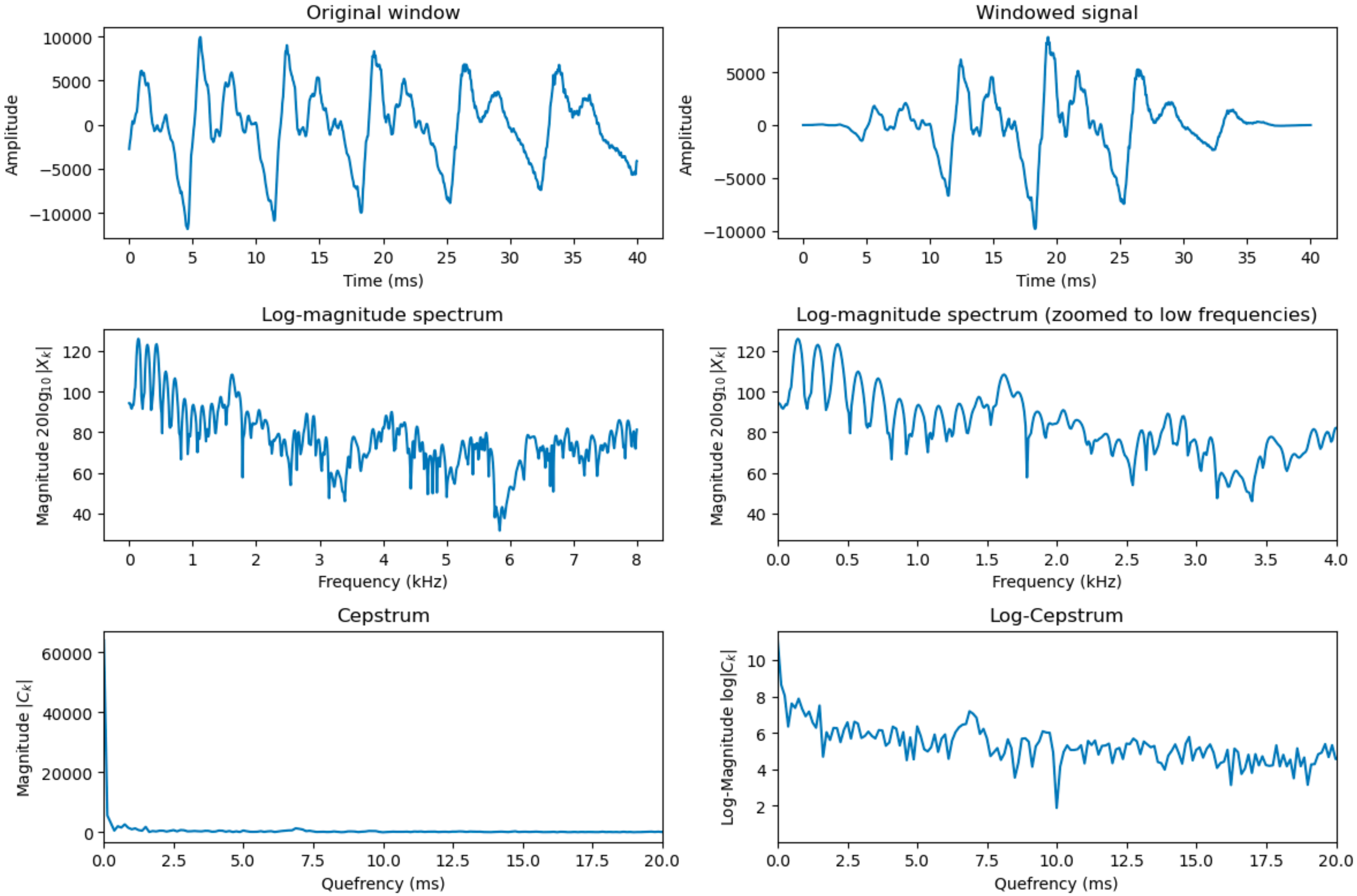


Cepstrum

- The output after the second time-frequency transform is known as the cepstrum

    - Apply analysis windowing to signal
    - Apply time-frequency transform (DFT or DCT)
    - Take the logarithm of the absolute value
    - Apply second time-frequency transform

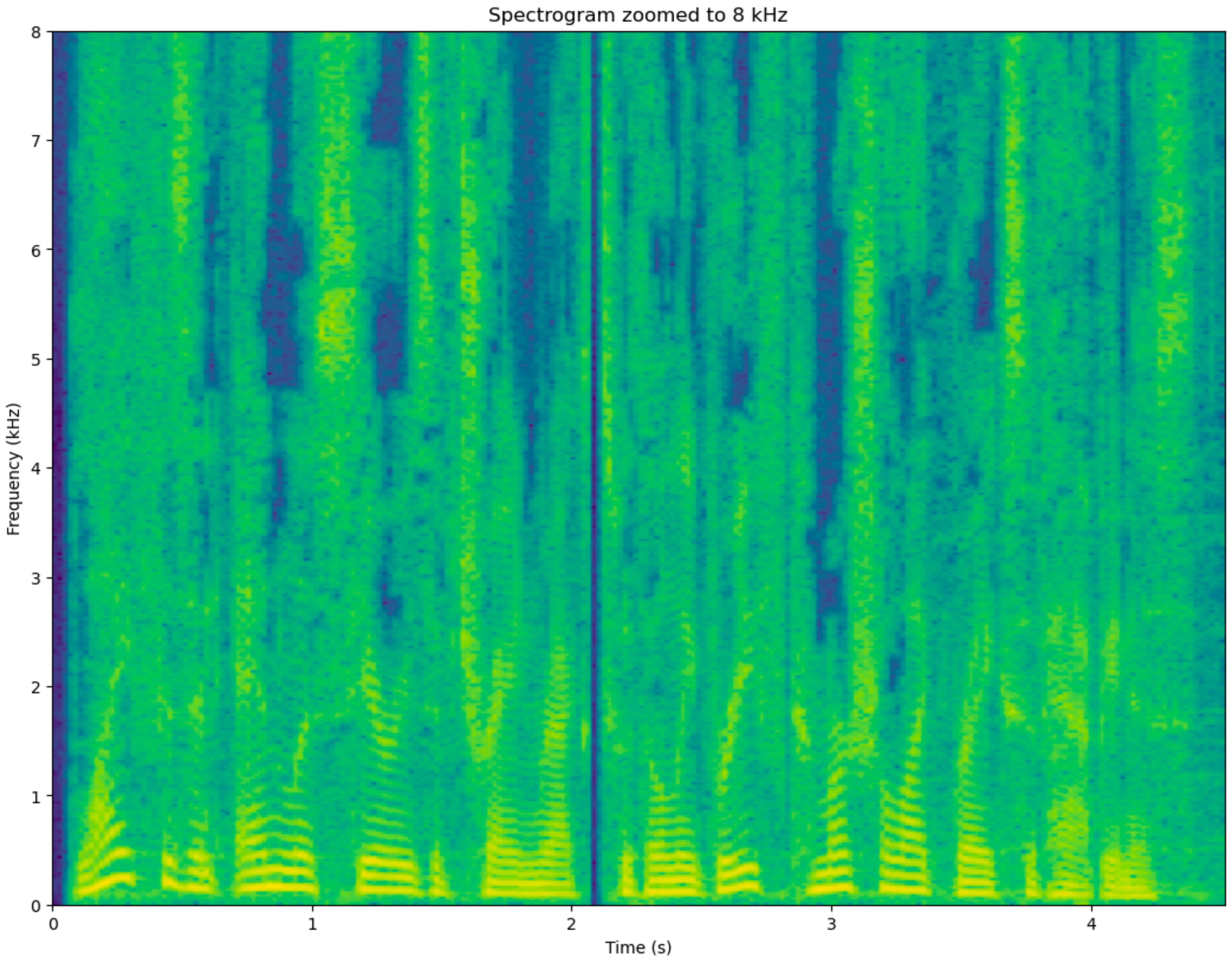
Cepstrum

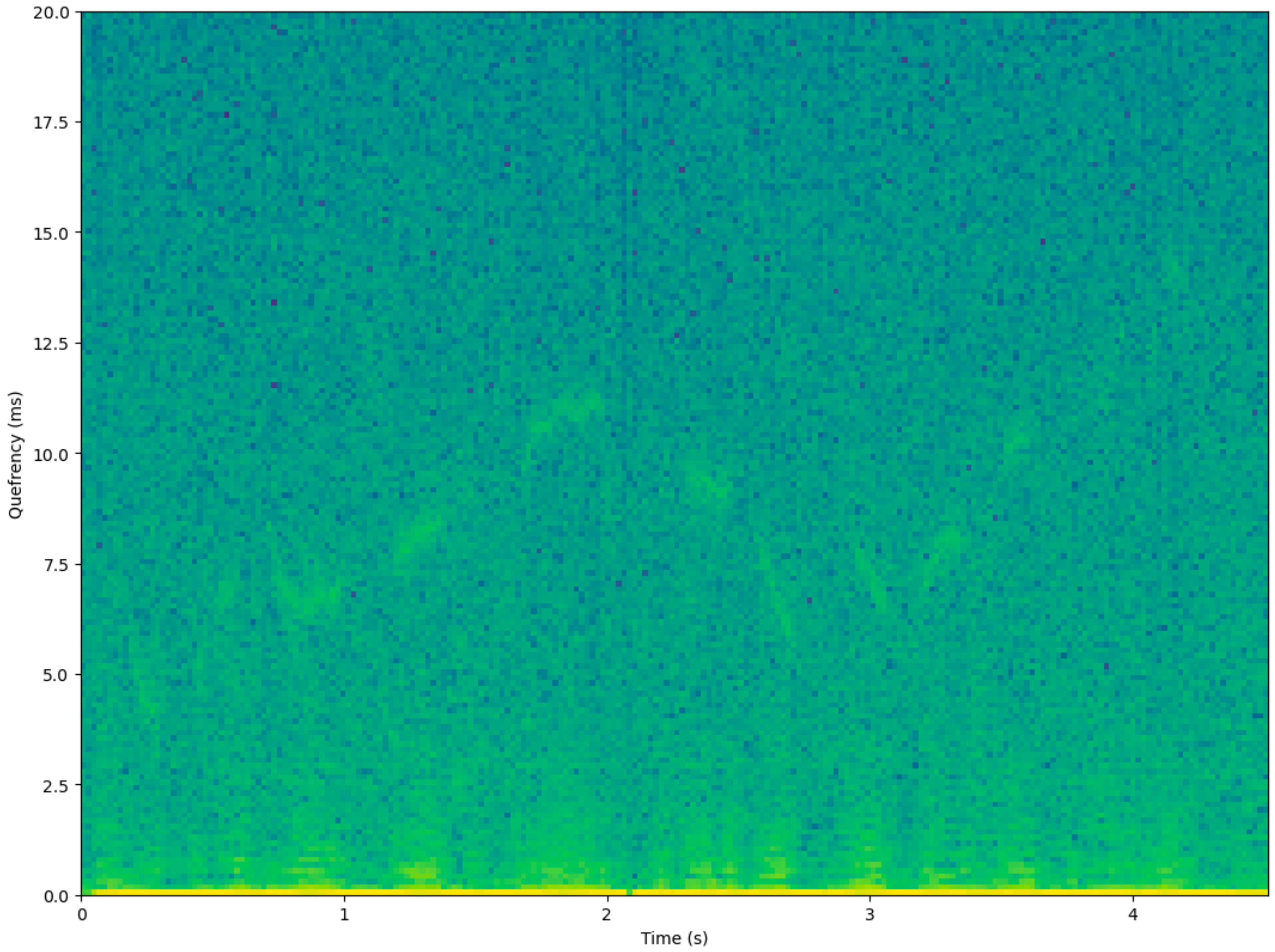


https://colab.research.google.com/drive/1j707gmlYED8r0AICb1Re-waULZ4DYWk6?usp=sharing

Cepstrum

- F0 is usually prominently visible as a peak in the cepstrum
- Quefrencies *q* can be easily converted to frequencies *f* by *f = 1 / q*





Summary

- Prosody
  - Pitch - Fundamental frequency
  - Loudness - Energy
  - Duration
- Jitter and Shimmer
- Spectrogram
- Cepstrum

Readings

- Chapter 3: Basic Representations
  - https://speechprocessingbook.aalto.fi/Representations/Representations.html