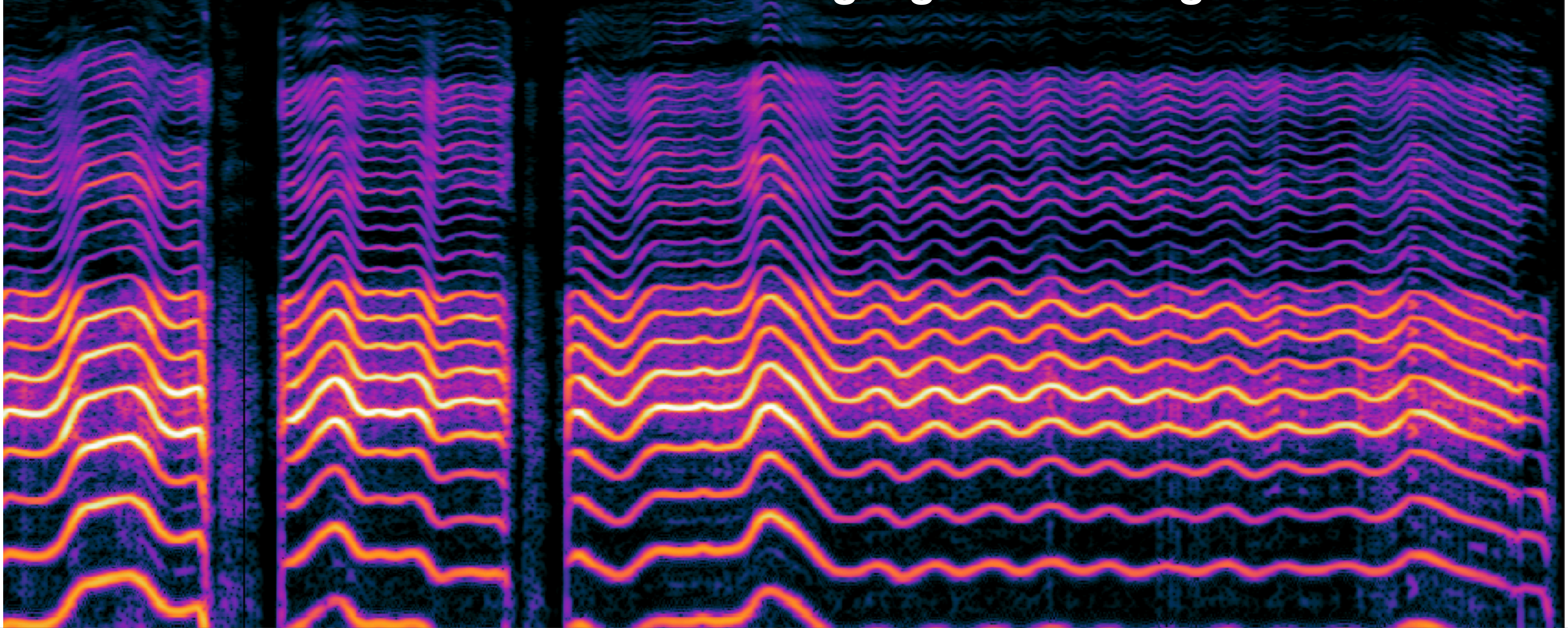
CSC3100 - Fundamentals of Speech and Language Processing
MDS6002 - Natural Language Processing



Lecture 4: Understanding human speech

Zhizheng Wu

Outline

- Information in human speech
- Timbre
  - Speaker identity
  - Content
- Prosody

Text version of a speech

Trask: Sir, you are out of order!

Slade: Outta order? I'll show you outta order! You don't know what outta order is, Mr. Trask! I'd show you but I'm too old; I'm too tired; I'm too fuckin' blind. If I were the man I was five years ago I'd take a FLAME-THROWER to this place! Outta order. Who the hell you think you're talkin' to? I've been around, you know? There was a time I could see. And I have seen boys like these, younger than these, their arms torn out, their legs ripped off. But there isn't nothin' like the sight of an amputated spirit; there is no prosthetic for that. You think you're merely sendin' this splendid foot-soldier back home to Oregon with his tail between his legs, but I say you are executin' his SOUL!! And why?! Because he's not a Baird man! Baird men, ya hurt this boy, you're going to be Baird Bums, the lot of ya. And Harry, Jimmy, Trent, wherever you are out there, FUCK YOU, too!


Spoken version

Ways to say mom

Text version

Mom

妈妈

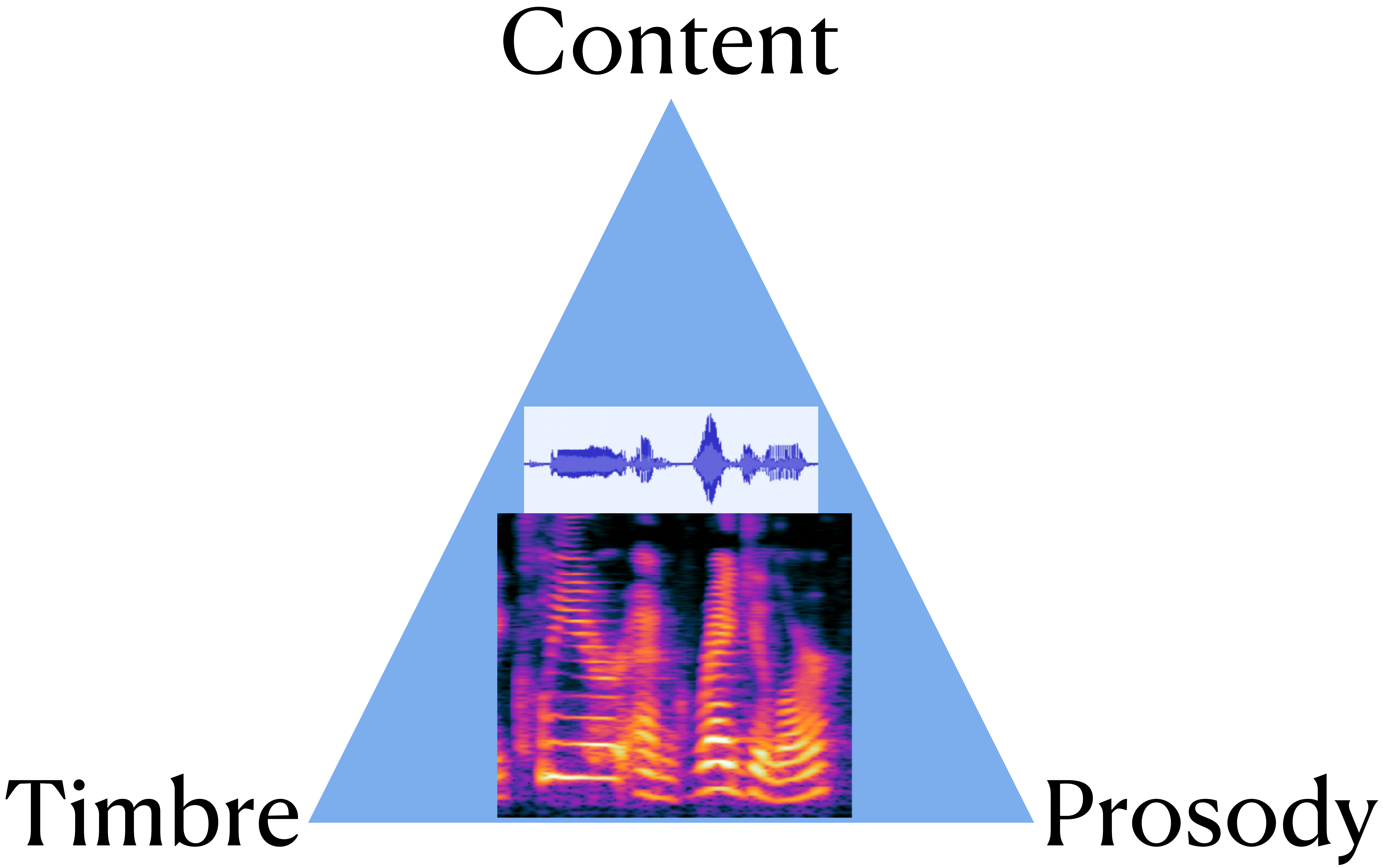
Spoken version

Content

Speaker

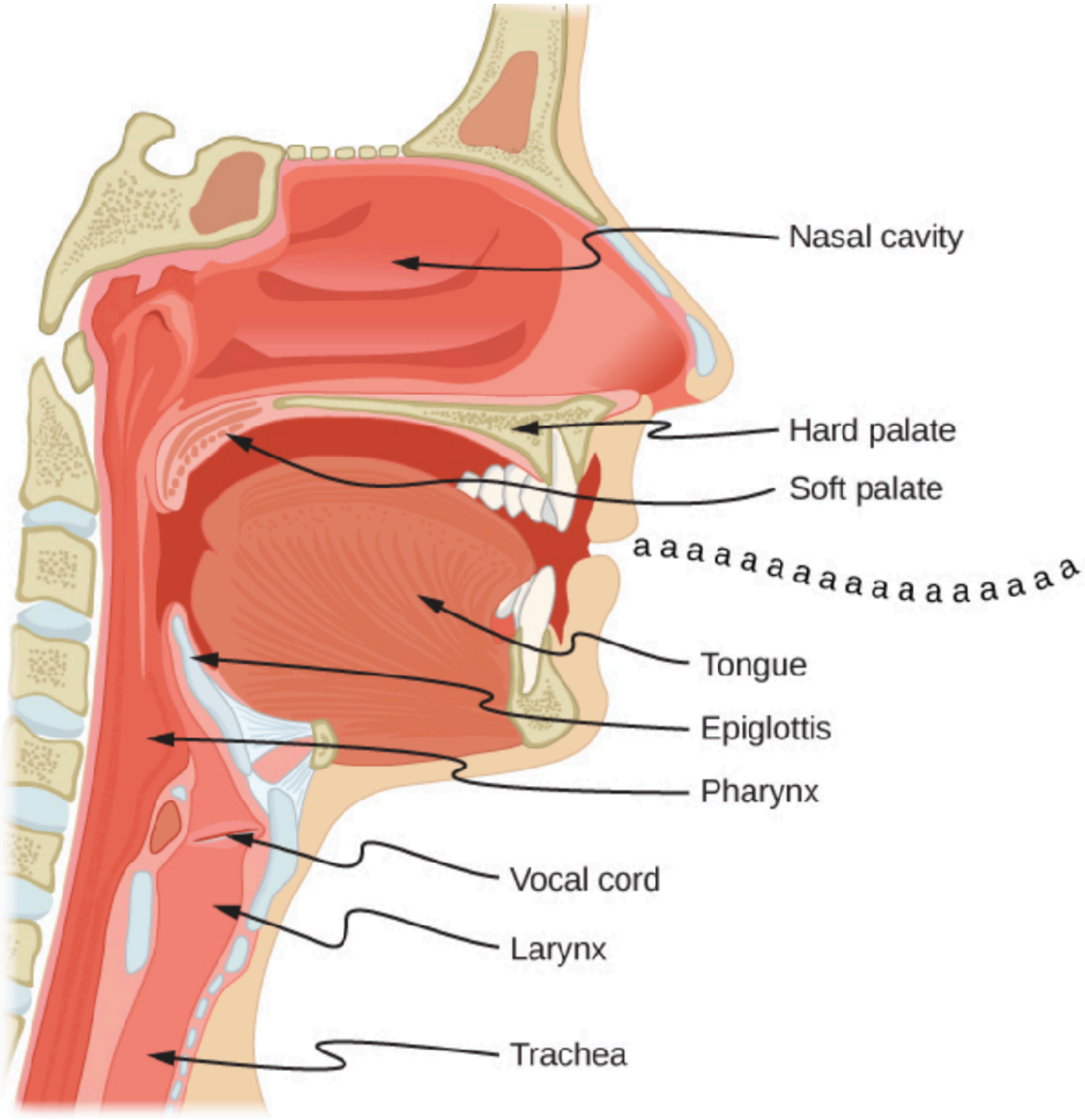
Emotion

Age, etc


Content
Timbre
Prosody

Speech production

- Source-filter model
  - Source produces an initial sound
  - Vocal tract filter modifies it
- Source
  - An input of acoustic energy into the speech production system
- Vocal tract filter
  - Articulators: tongue, teeth, lips, velum etc



Source-filter model

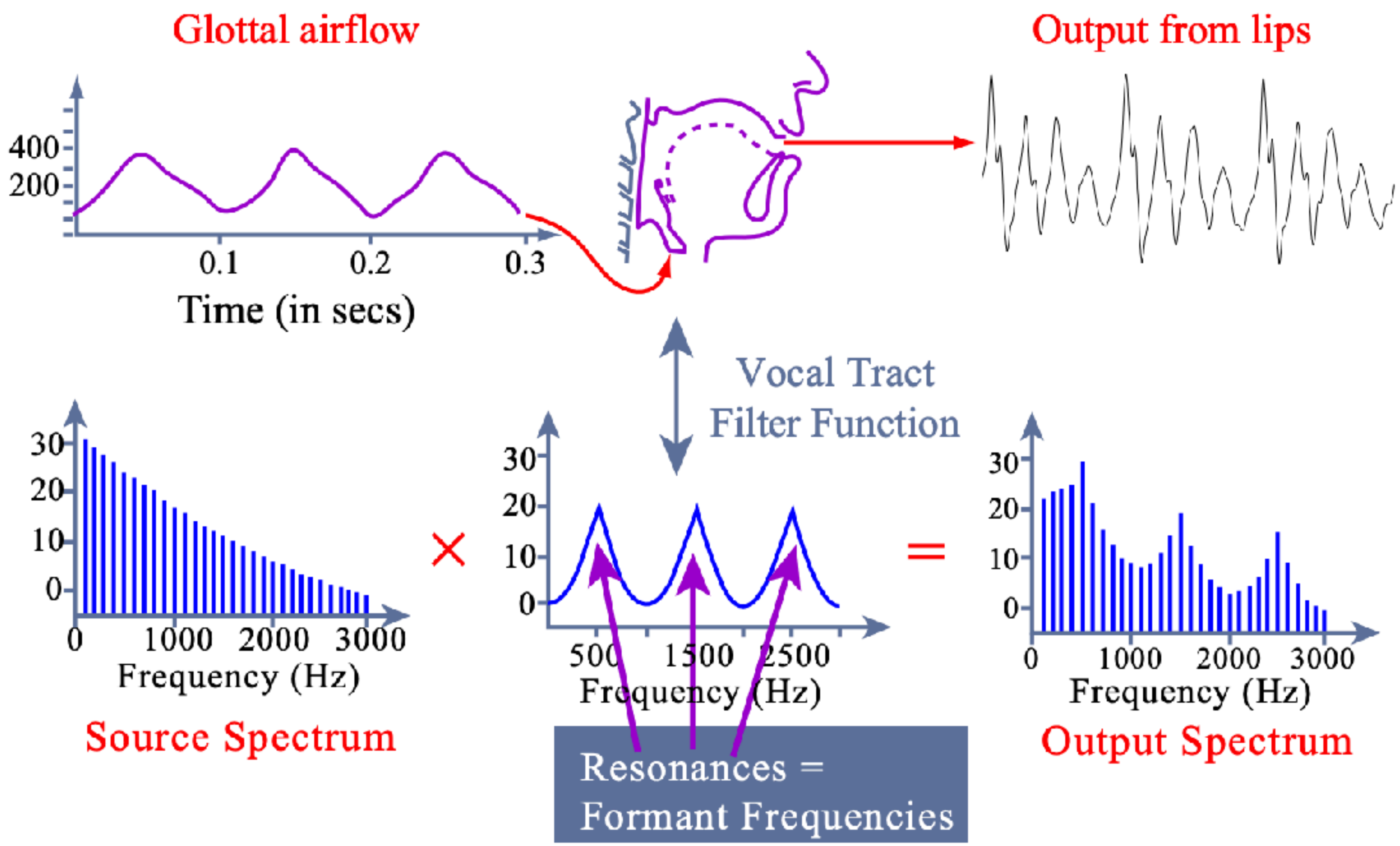
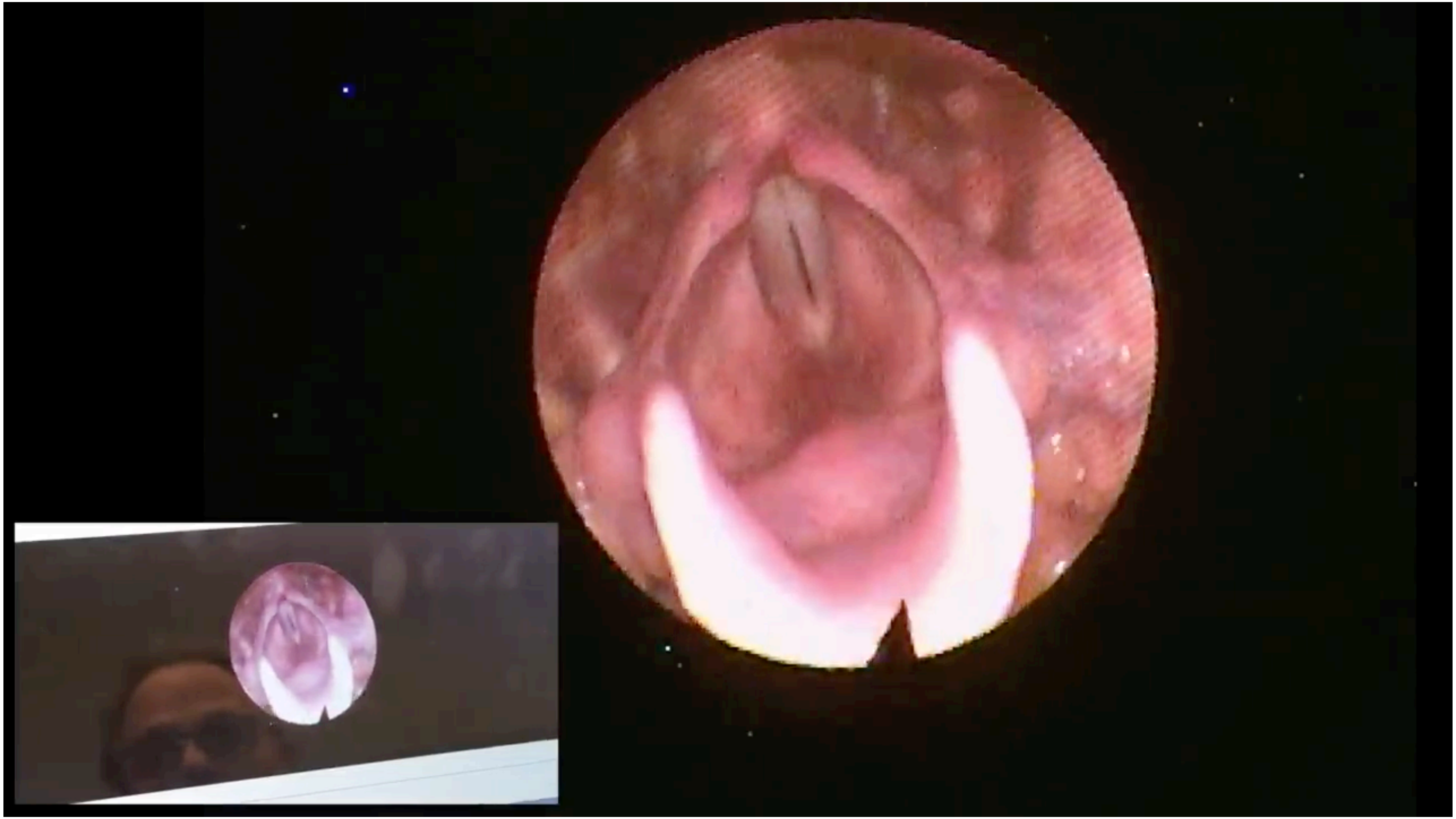


Image credit: MIT OCW

Source

- Voicing source: Vocal folds vibrating
  - A periodic source produced by modulation of the airflow from the lungs by the vocal folds
    - The **vocal folds** are muscular folds located in the **larynx**
    - The space between the vocal folds is the **glottis**
  - If the vocal folds are close together, then air pressure from the lungs can cause them to vibrate periodically, generating voicing.
- Unvoicing source: vocals fold holds close but not vibrating

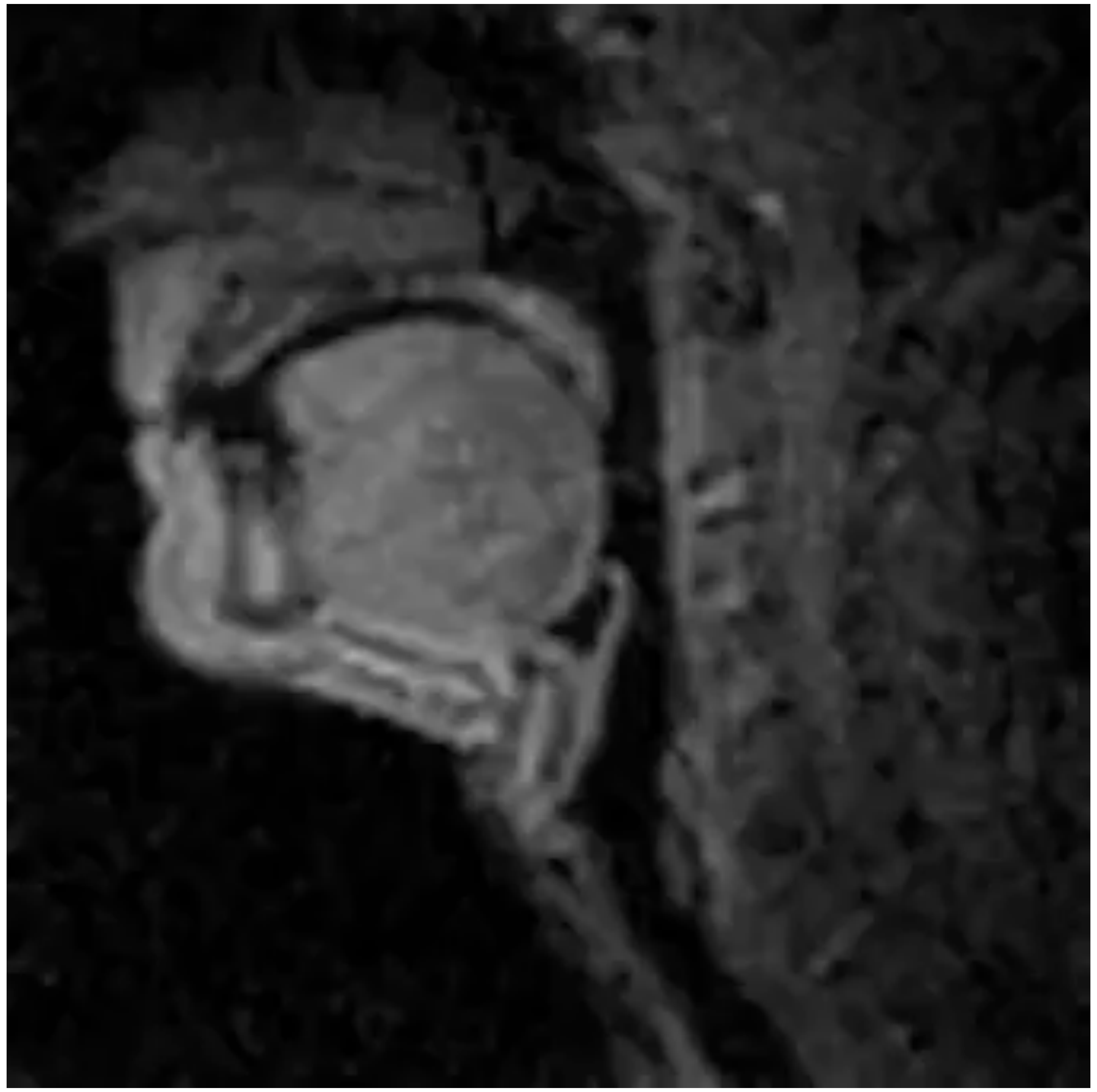
Source

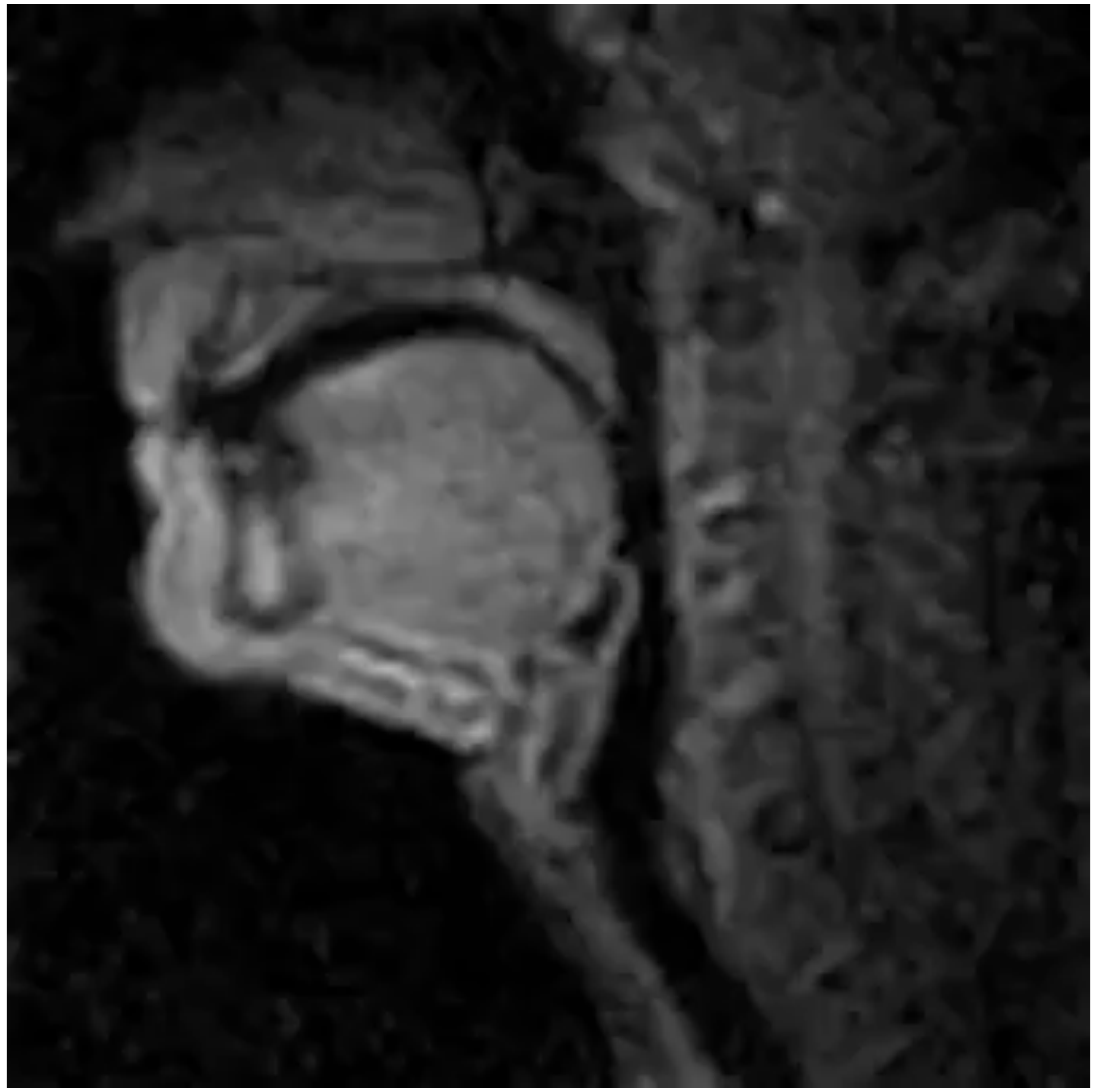


Filter

- The vocal tract acts as a filter, modifying the source waveform
- The sound wave at some distance from the speaker is the result of filtering the source with the vocal tract filter, plus the radiation characteristics of the lips/nose.

Head





Had

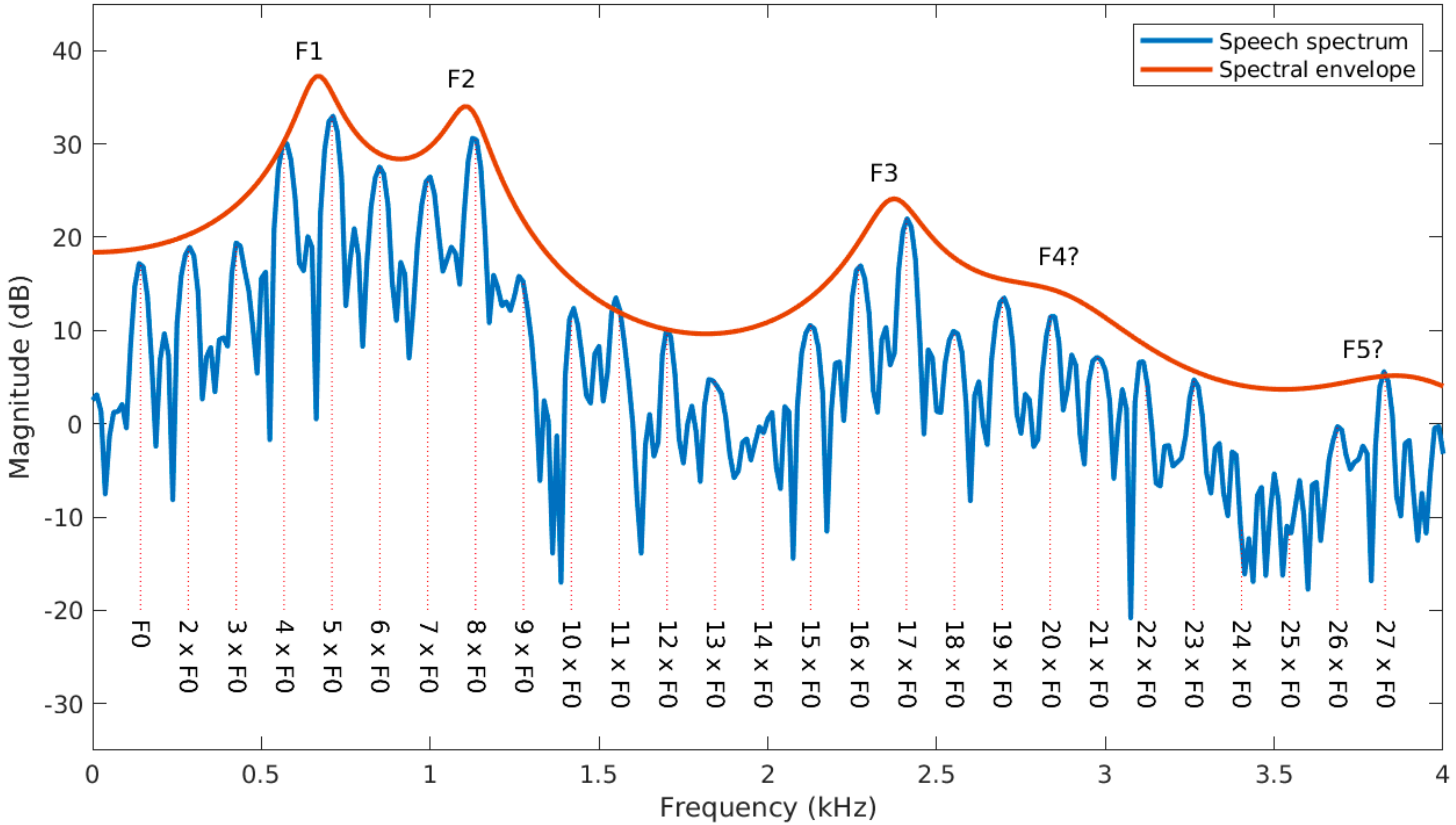
Video credit: https://sail.usc.edu/~lgoldste/General_Phonetics/

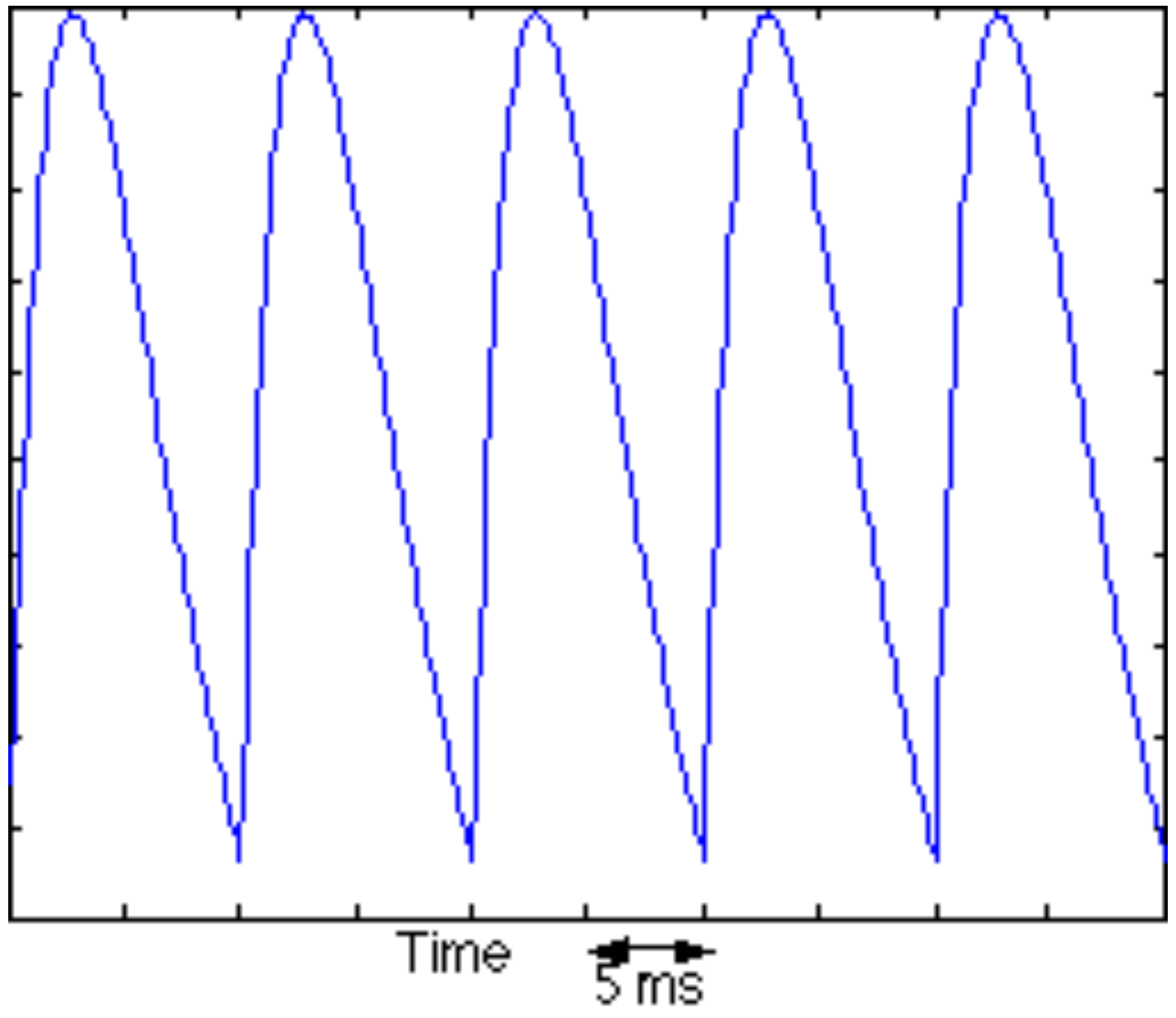
Resonance

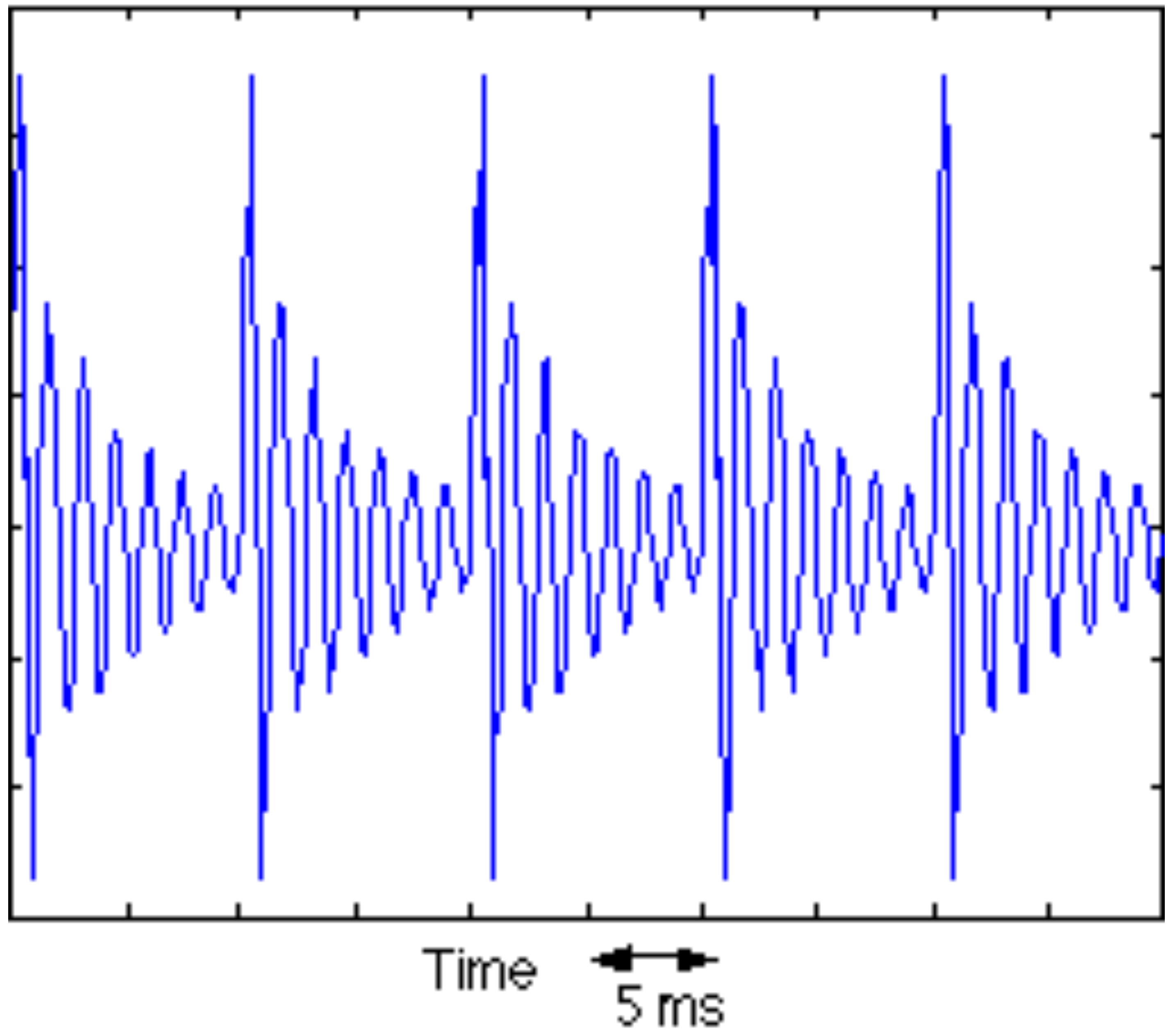
- A resonant frequency is a natural frequency of vibration determined by the physical parameters of the vibrating object.

Resonance

- The resonances of the vocal tract are called **formants**







Source

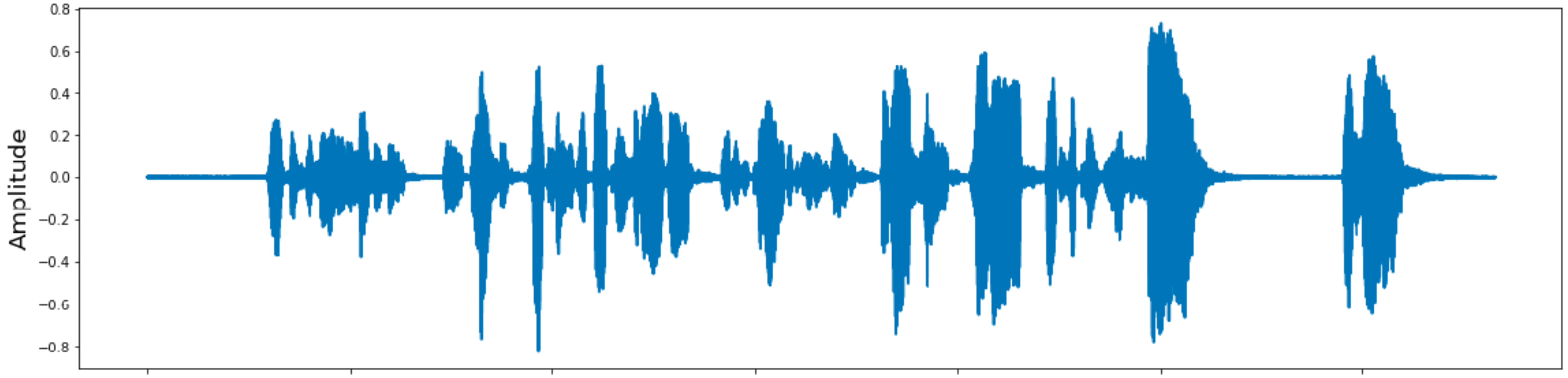
Filter

https://sail.usc.edu/~lgoldste/General_Phonetics/Source_Filter/MATLAB_demo/source-filter.html

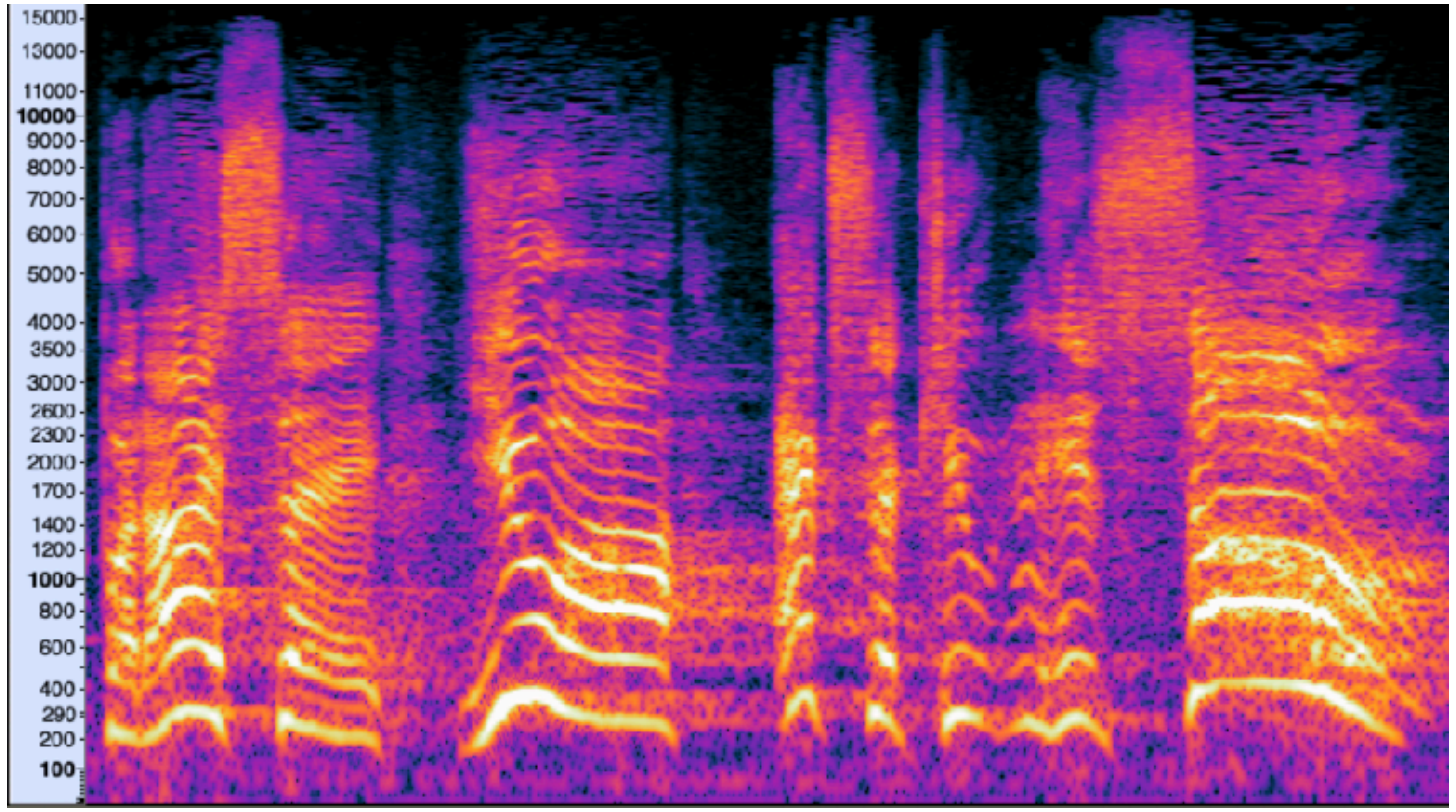
Reflection of speech production in speech representation

Speech representation

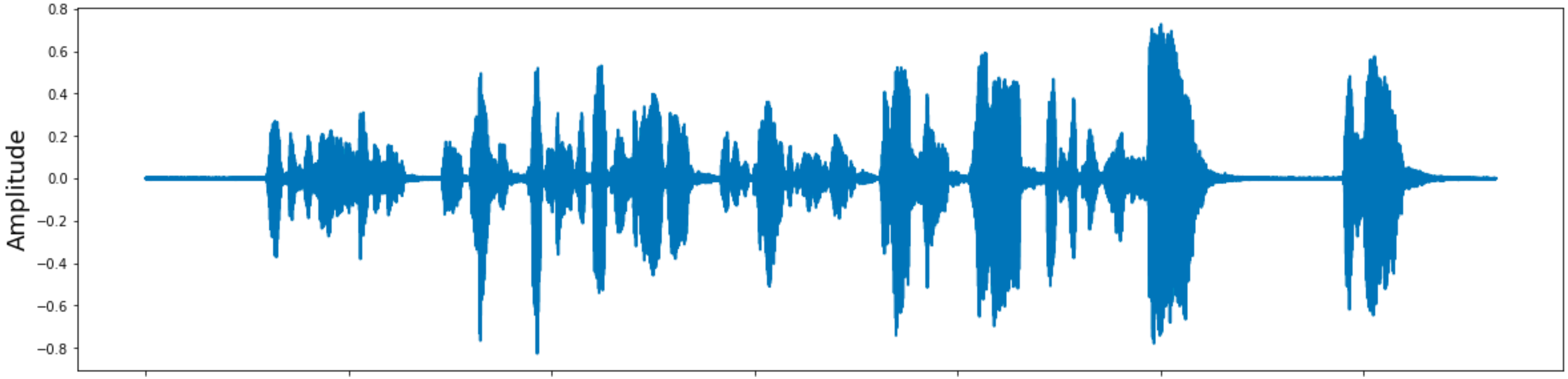
- Time domain

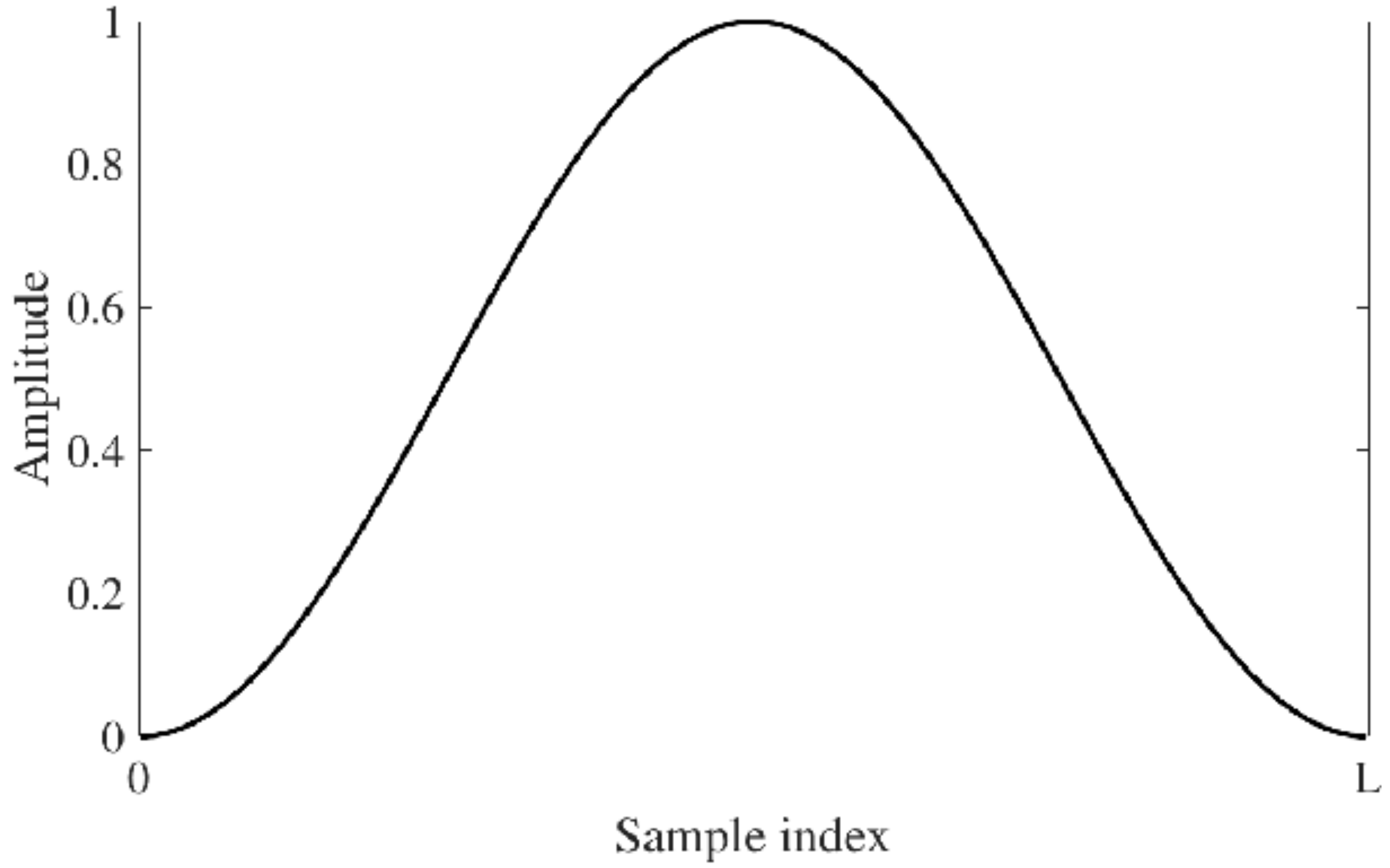


- Frequency domain



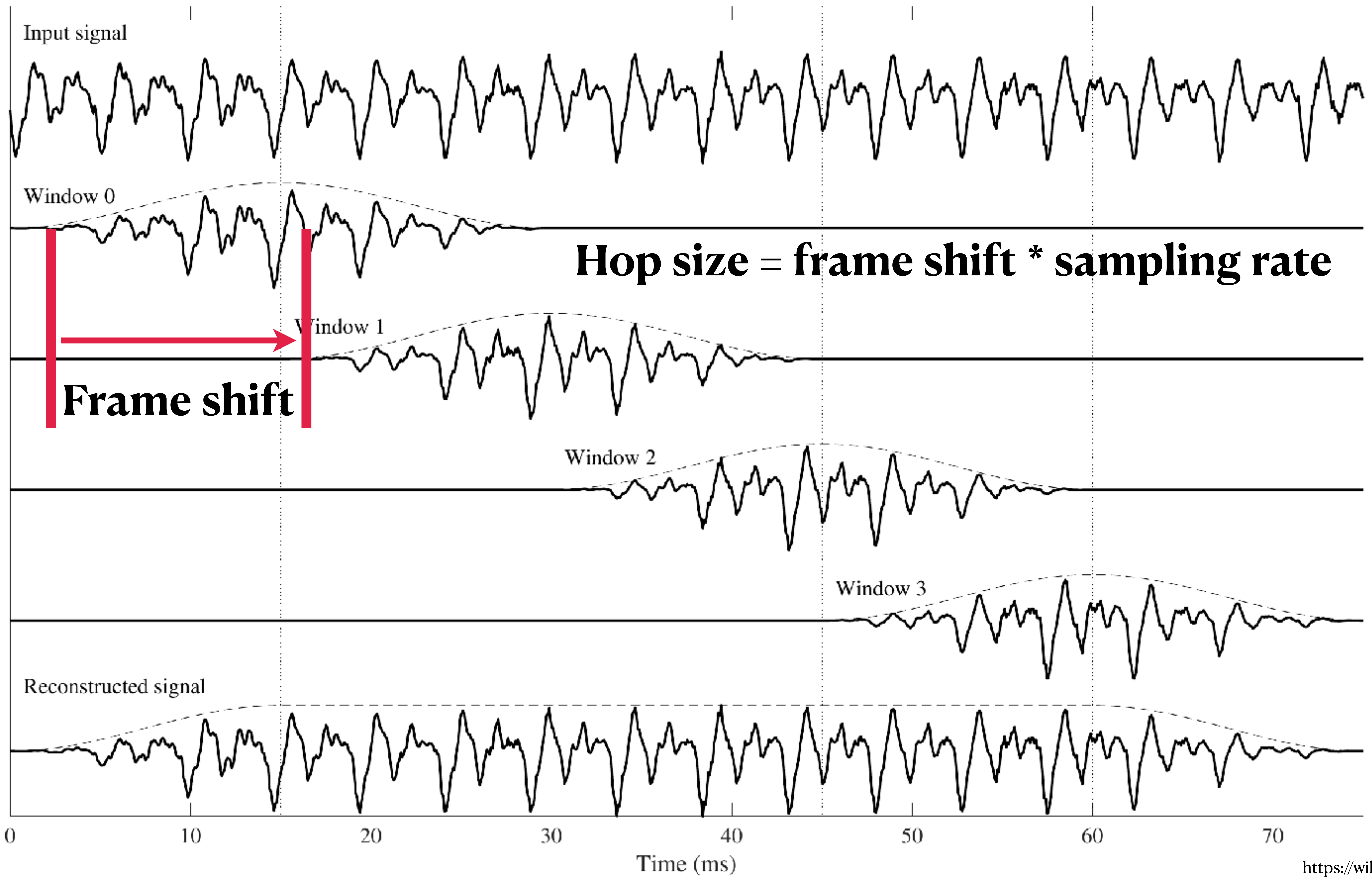
Windowing





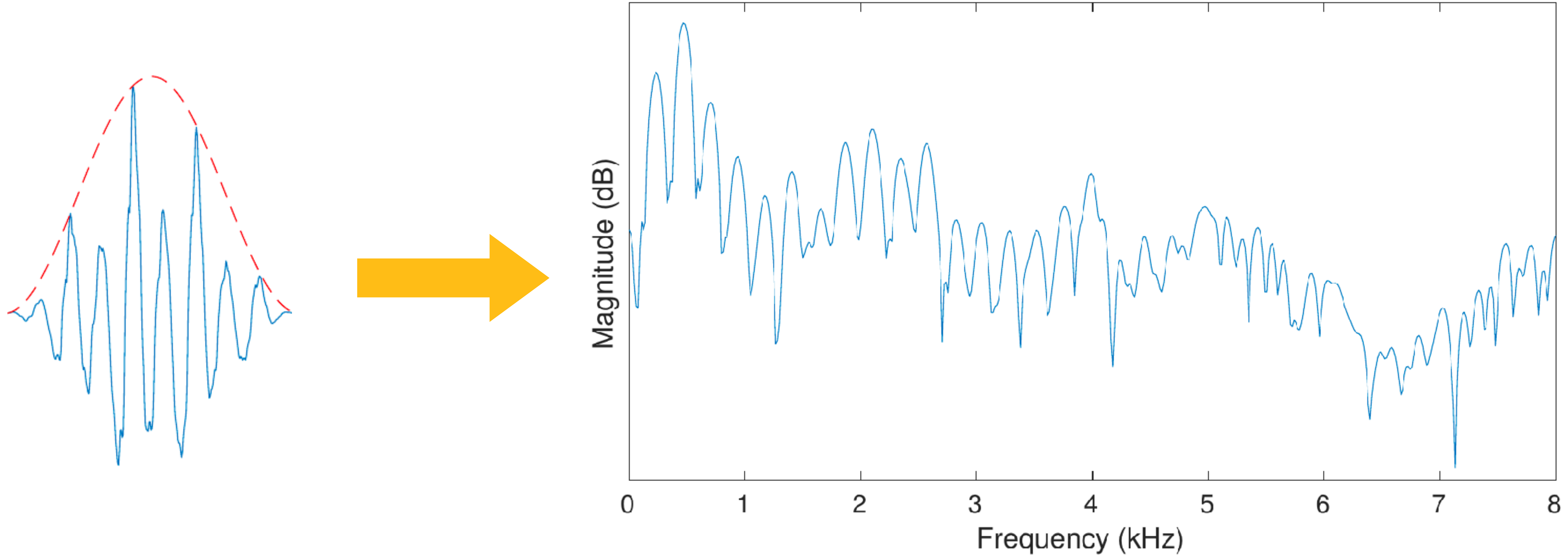
L = Window size

Windowing for analysis

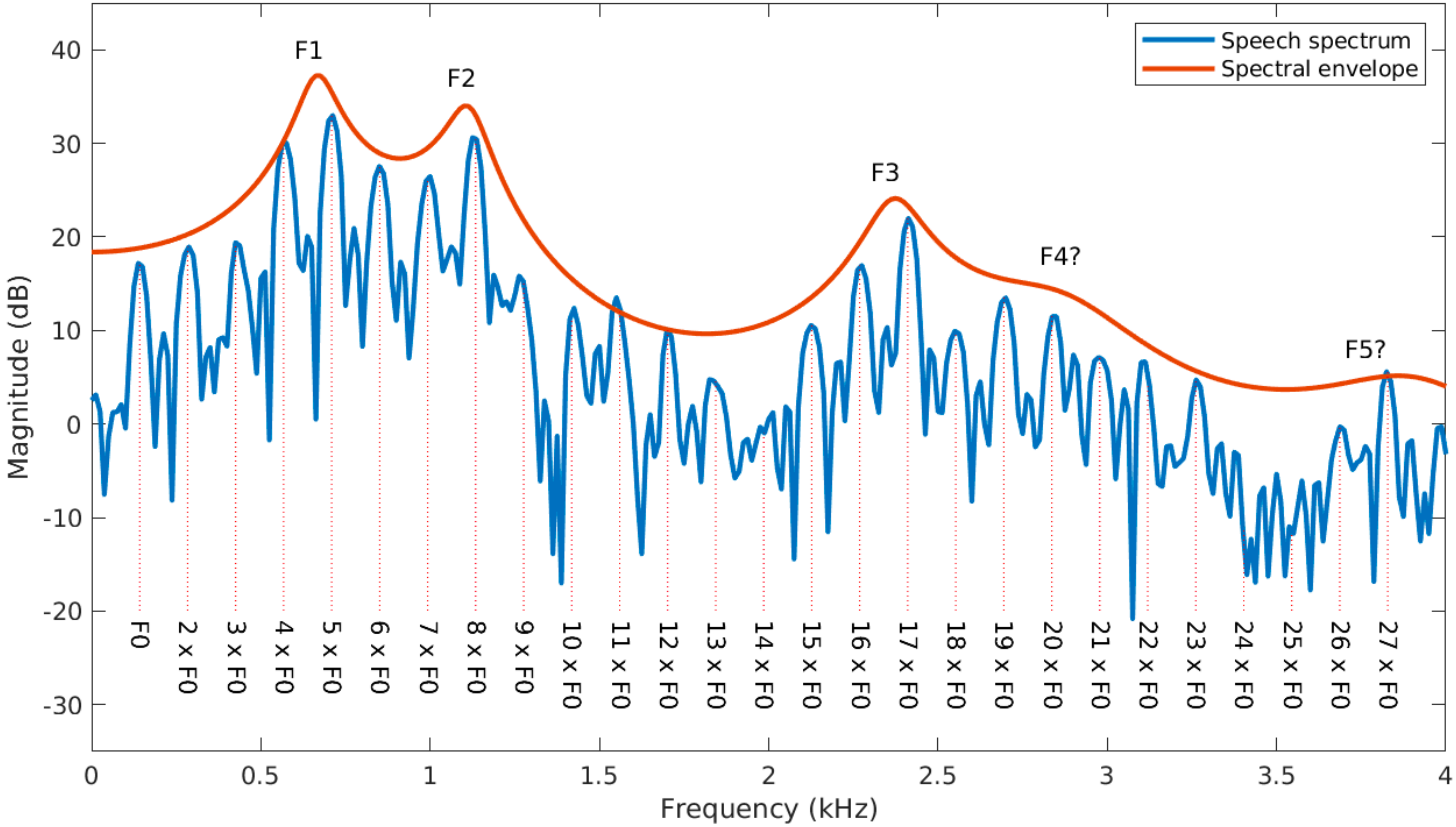


https://wiki.aalto.fi/display/ITSP/Windowing

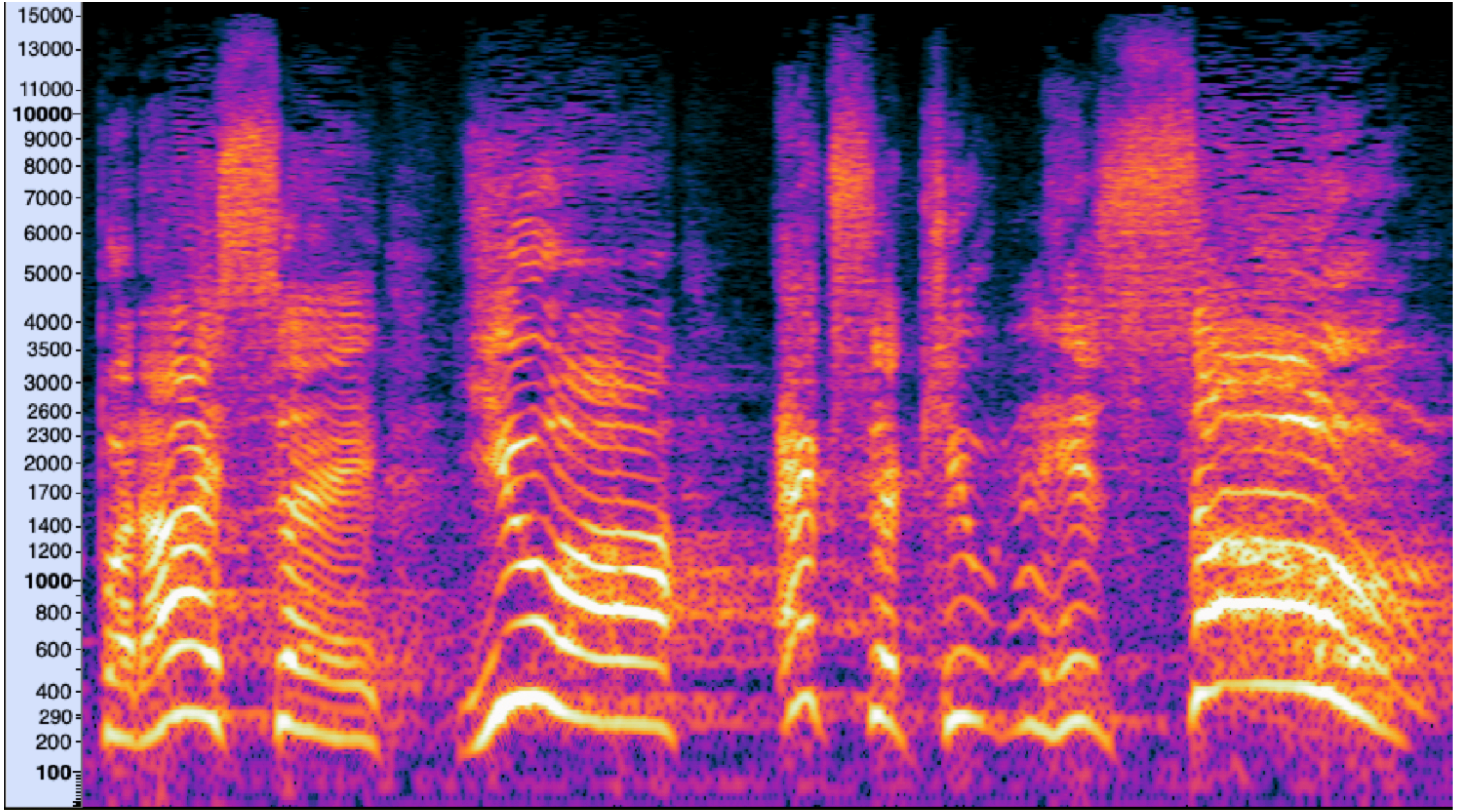
From time domain to spectrum



Spectral envelop and formants



Spectrogram: time-frequency representation



Source-filter in time domain

- Convolution in time domain

$$s[t] = e[t] + \sum_{k=1}^{K} a[k]s[t-k]$$

Source **Filter**

Source-filter: Multiplication in frequency domain

- Convolution in time domain equivalent to multiplication in frequency domain

$$S(m) = H(m) \cdot X(m)$$

Filter

Source spectrum

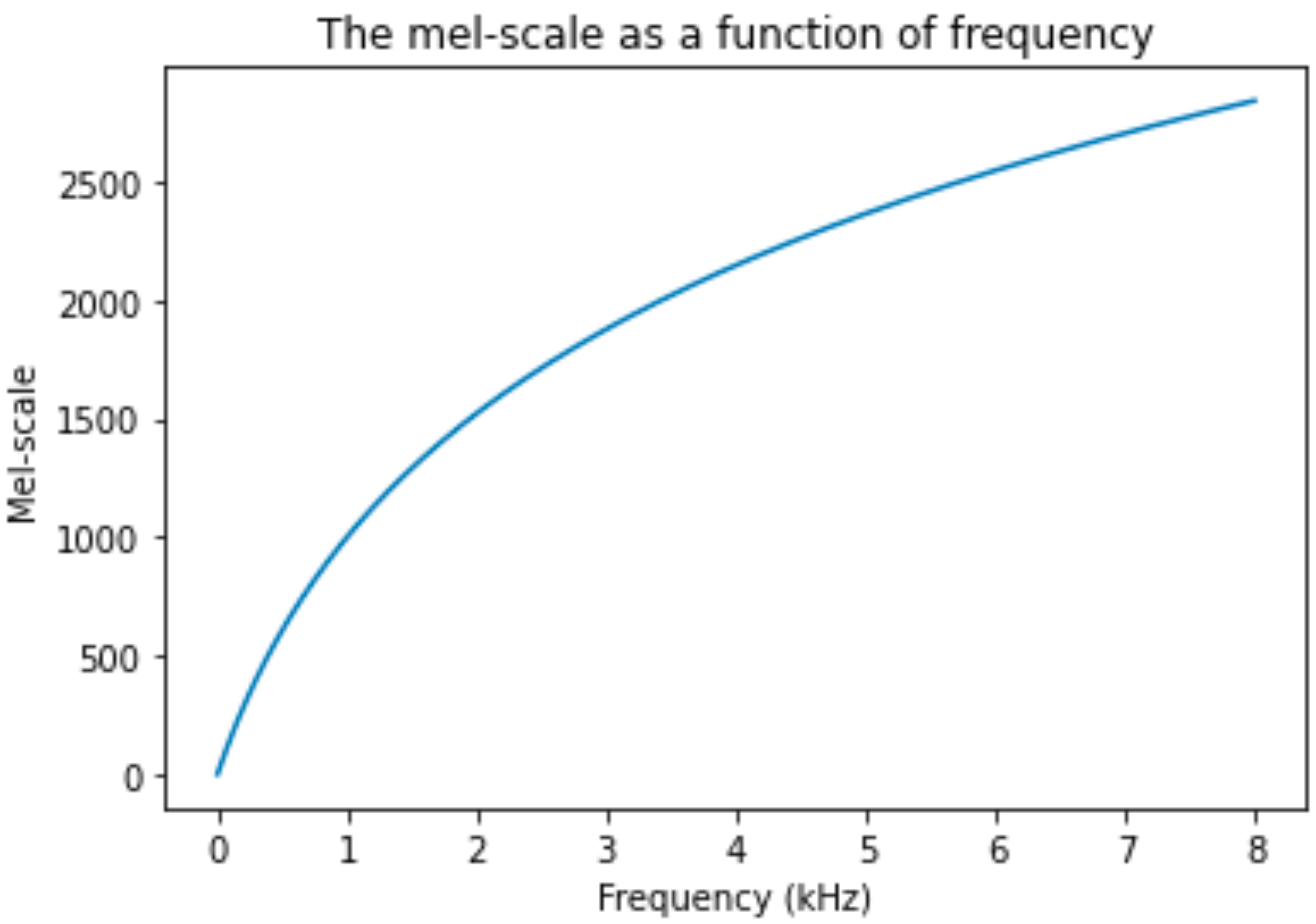
https://sail.usc.edu/~lgoldste/General_Phonetics/Source_Filter/MATLAB_demo/source-filter.html

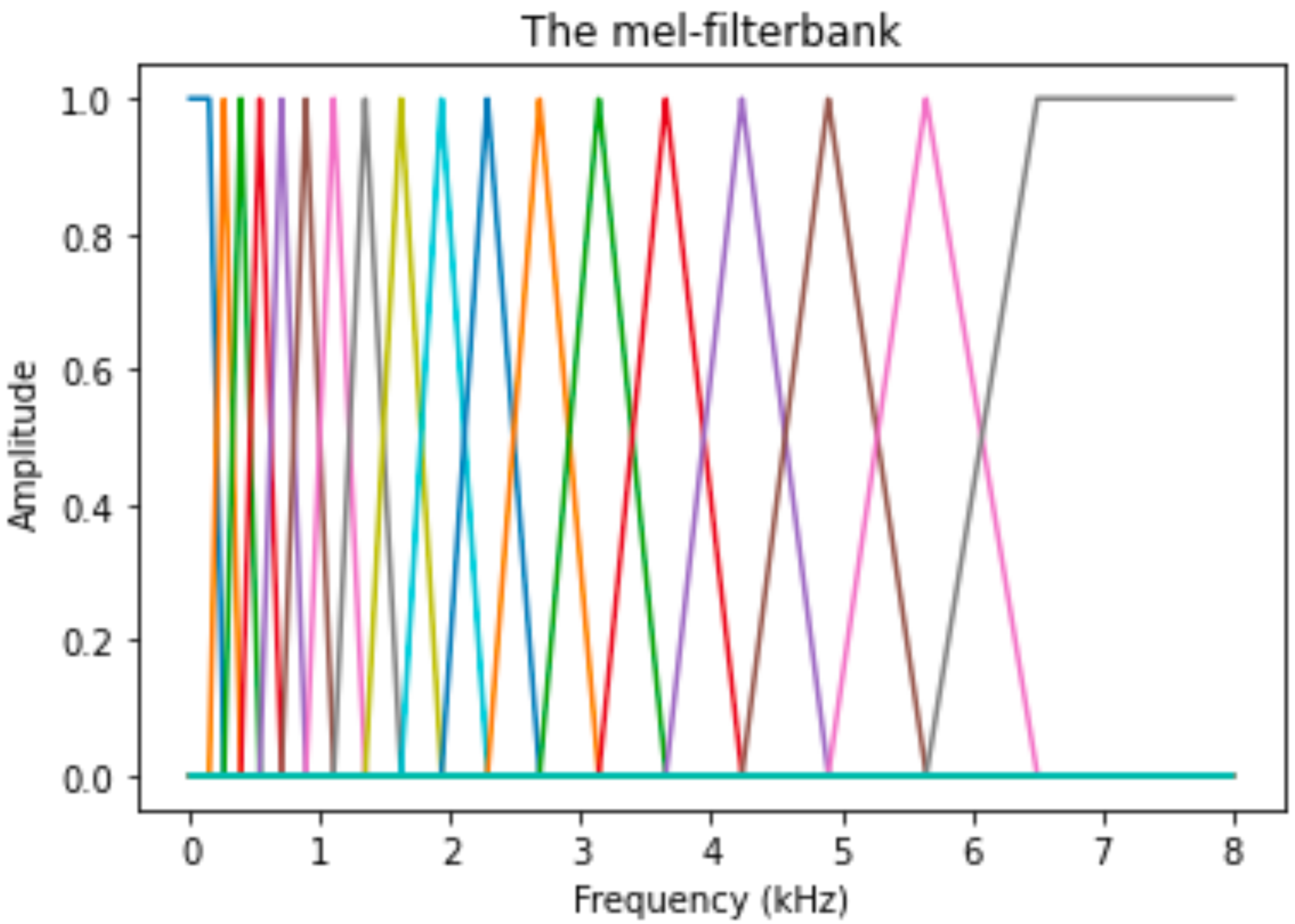
Independence of source and filter

- Source
  - Fundamental frequency (F0) is driven by the frequency of vocal fold vibrations
  - Harmonics are multiples of F0
- Filter
  - Resonances are driven by the shape of the vocal tract (physical property)
  - Formants are peaks in the spectral envelope that correspond to resonances (acoustic property)
- Independence of source and filter
  - You can change F0 without changing the vowel you are saying: harmonics change, formants stay the same

Mel-scale

- A scale that maps frequencies such that steps between tones align with our perception of steps





Timbre

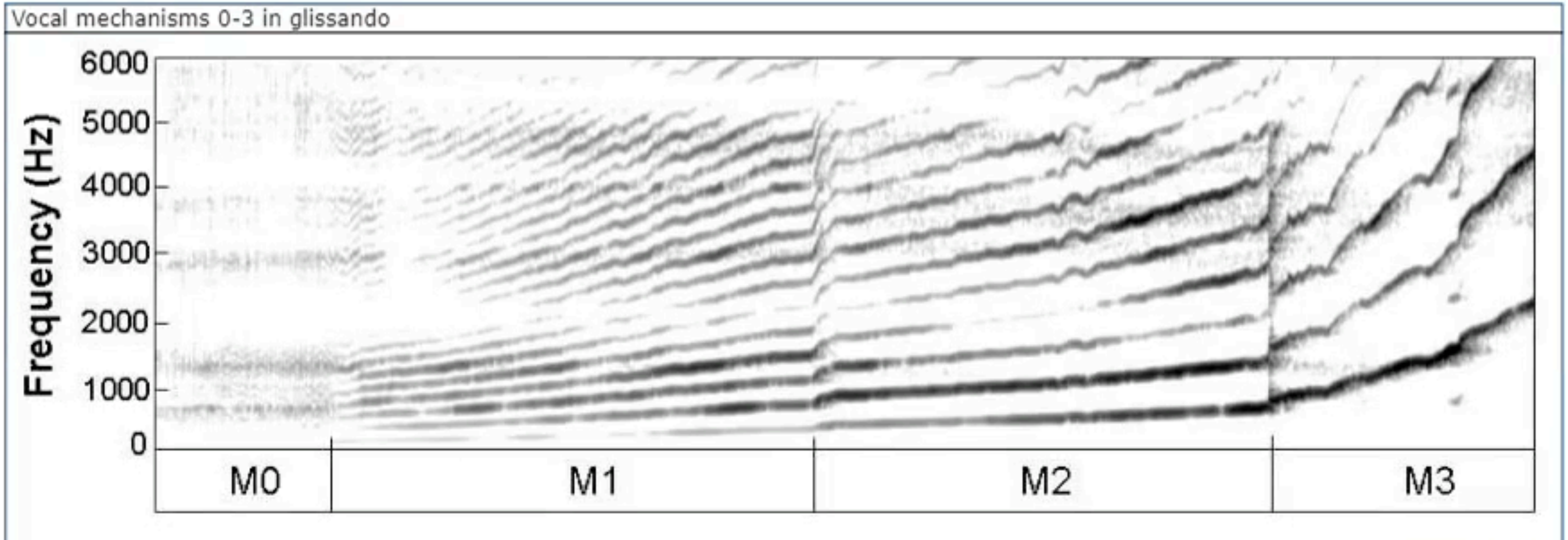
- The characteristic quality of a sound, independent of pitch and loudness
- Spectral envelope and its time variation can represent timbre
- The independence of source and filter explains
    - why vowels of the same timbre can be produced on different pitches
    - why vowels of the same pitch can have different timbres

Prosody: Melody of speech

- Same word can have different prosody
- Prosody includes pitch, duration, stress

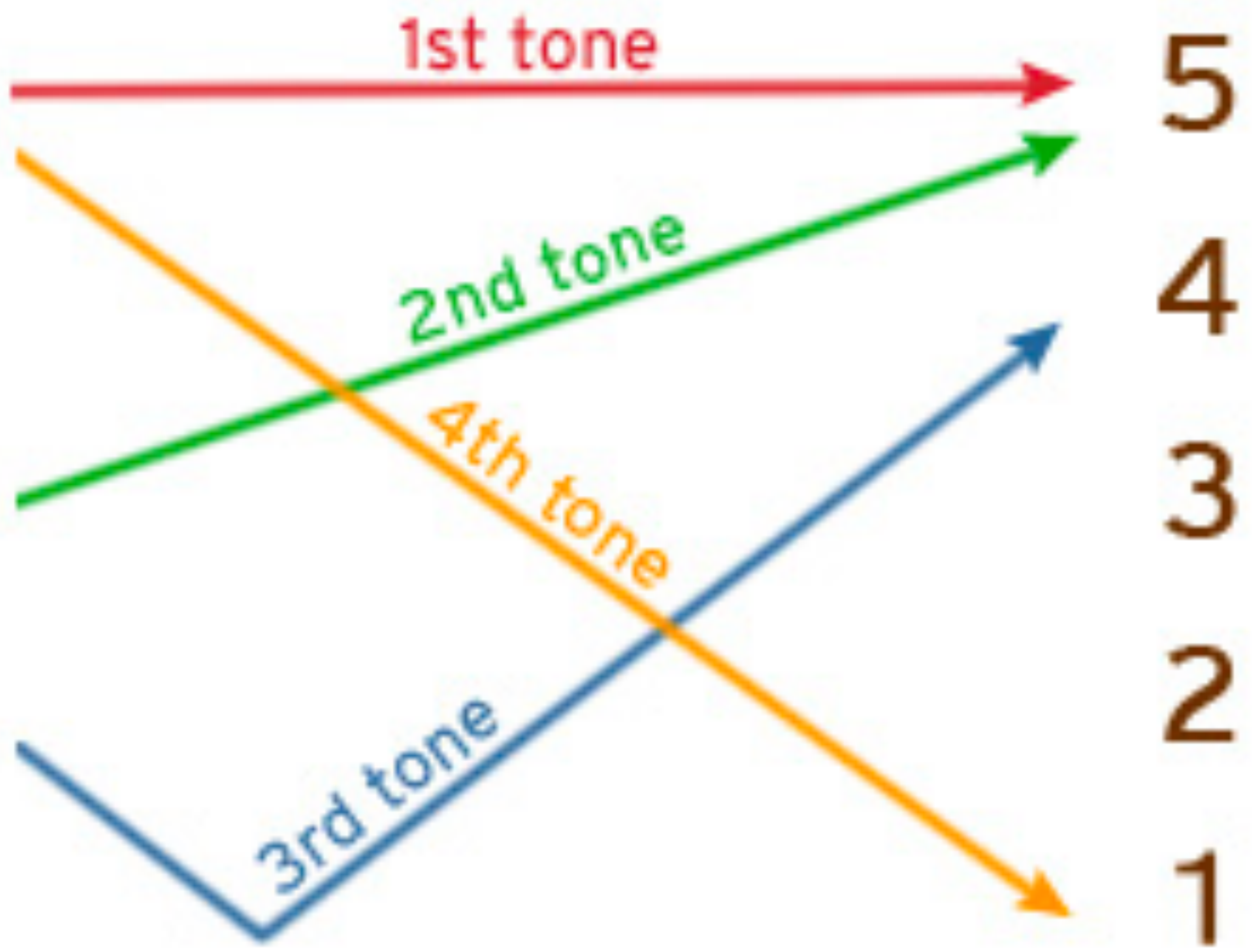
Perceived Pitch

- Fundamental frequency (F0)
  - the lowest frequency of a periodic waveform
  - F0 is driven by the frequency of vocal fold vibrations, not vocal tract resonances
- In a speech segment, F0 is semi-continuous

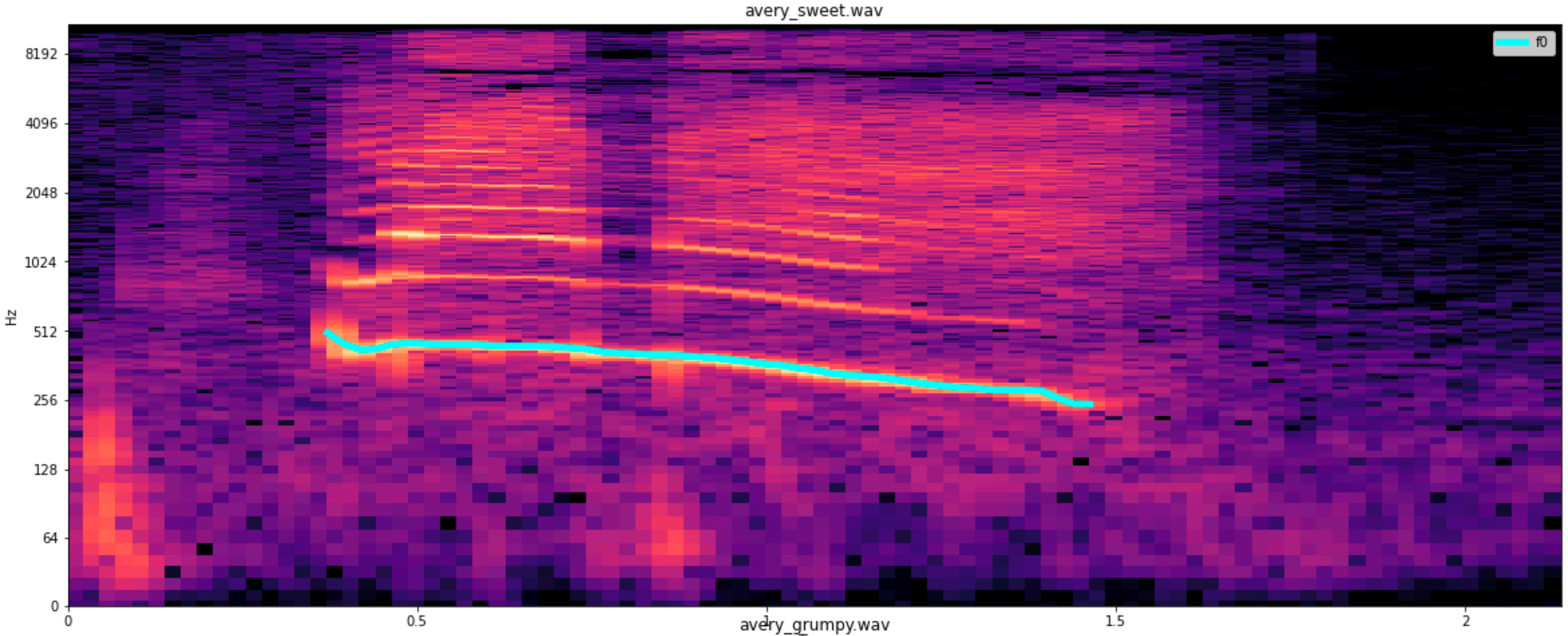


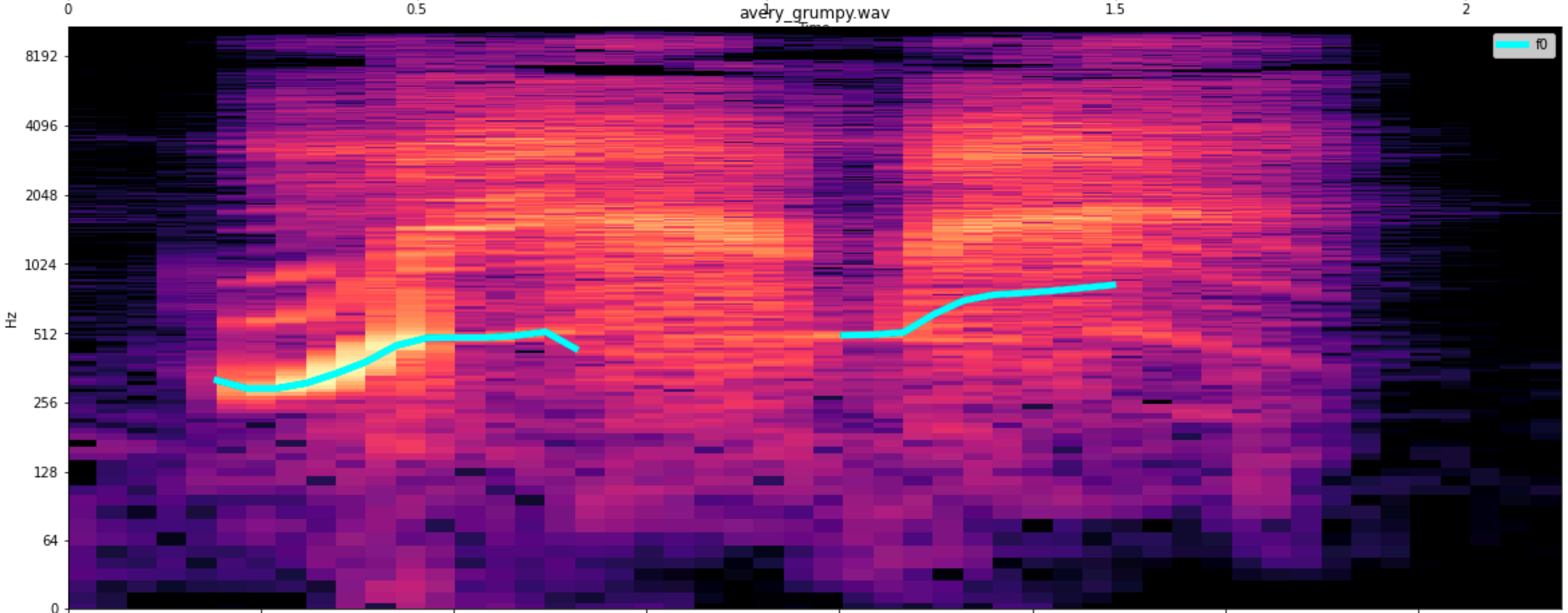
Tone

- Tonal language: different tonal inflections will convey different meanings

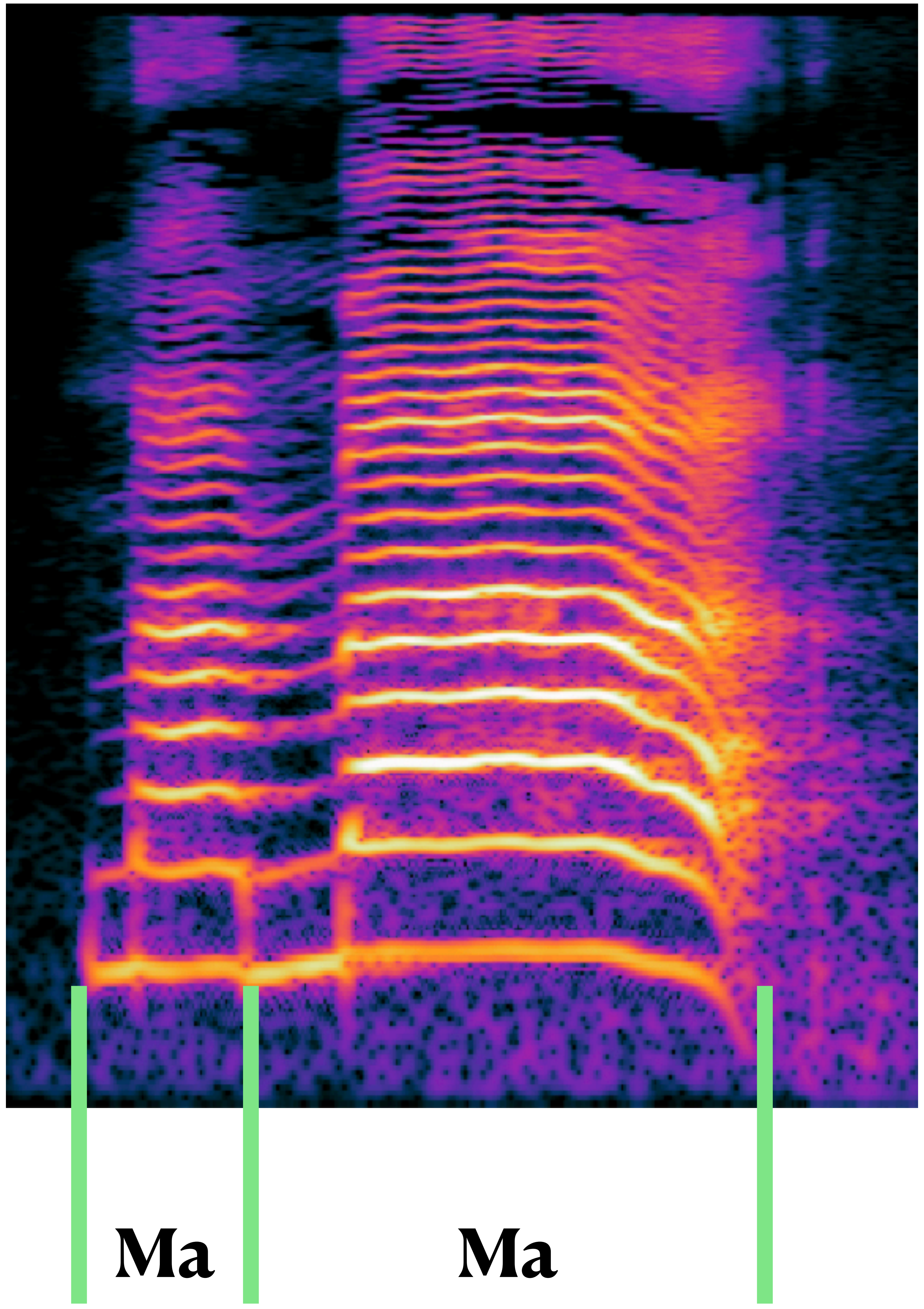


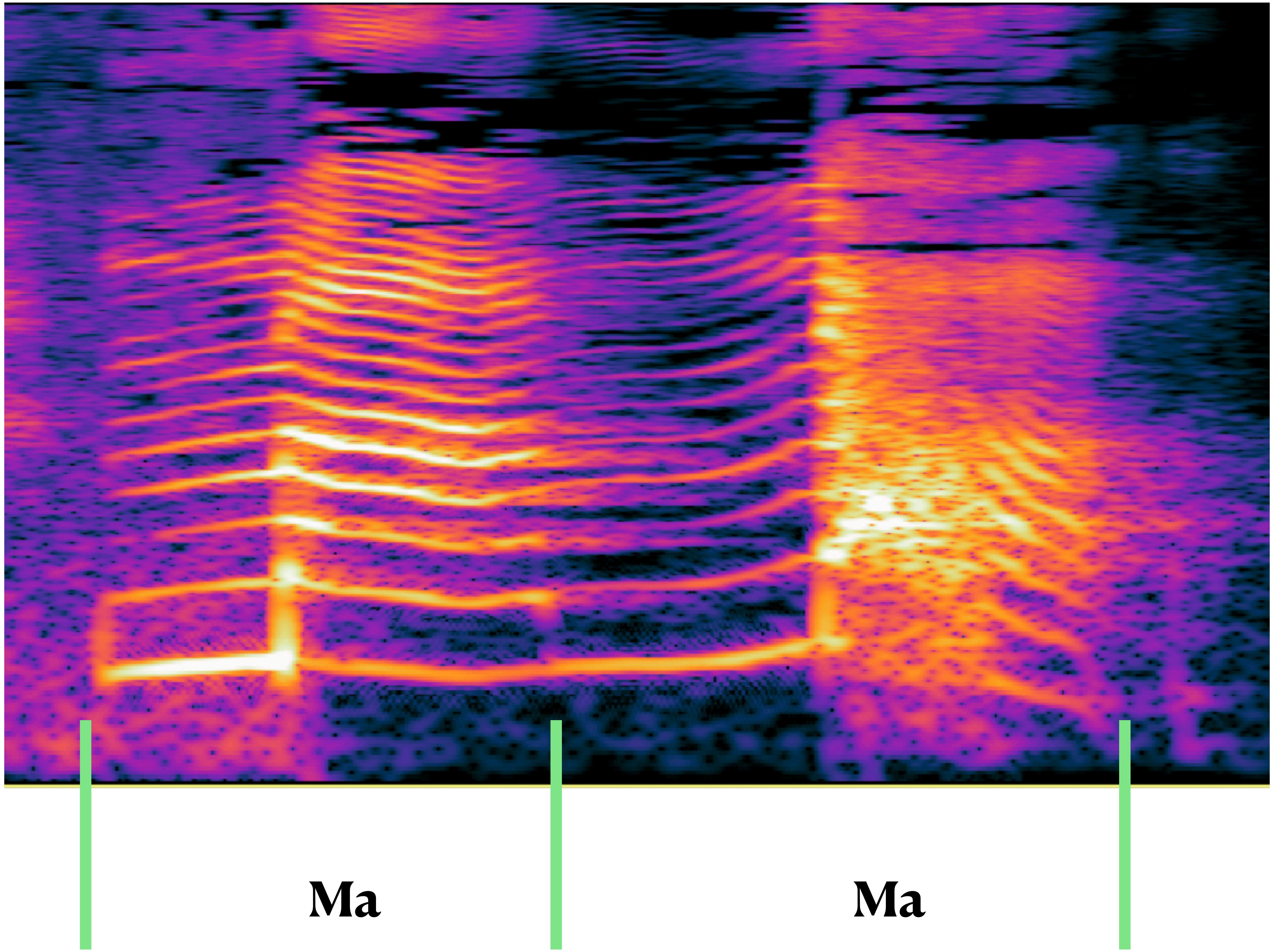
Intonation


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

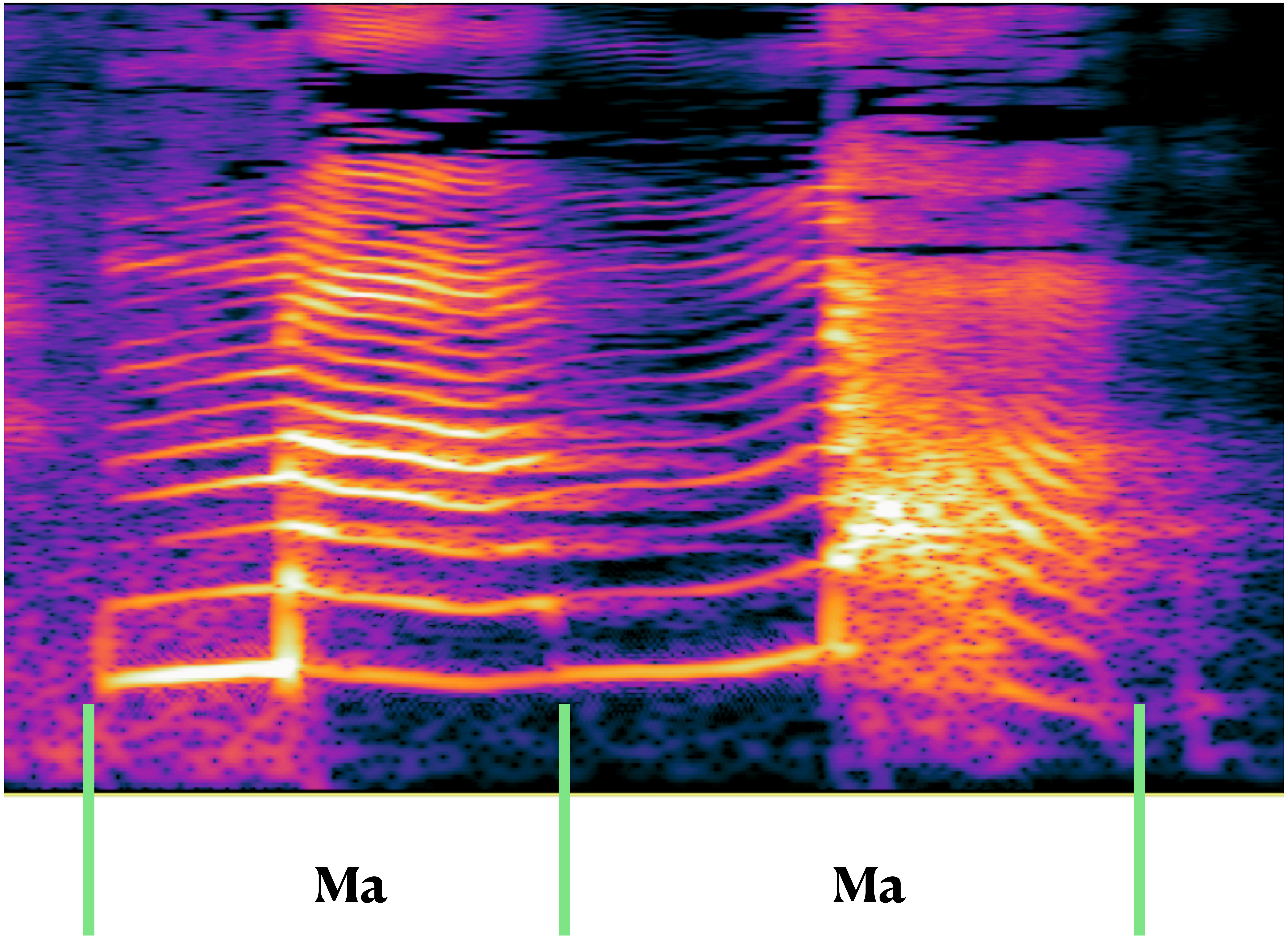

avery_grumpy.wav
f0
8192
4096
2048
1024
512
256
128
64
0
Hz

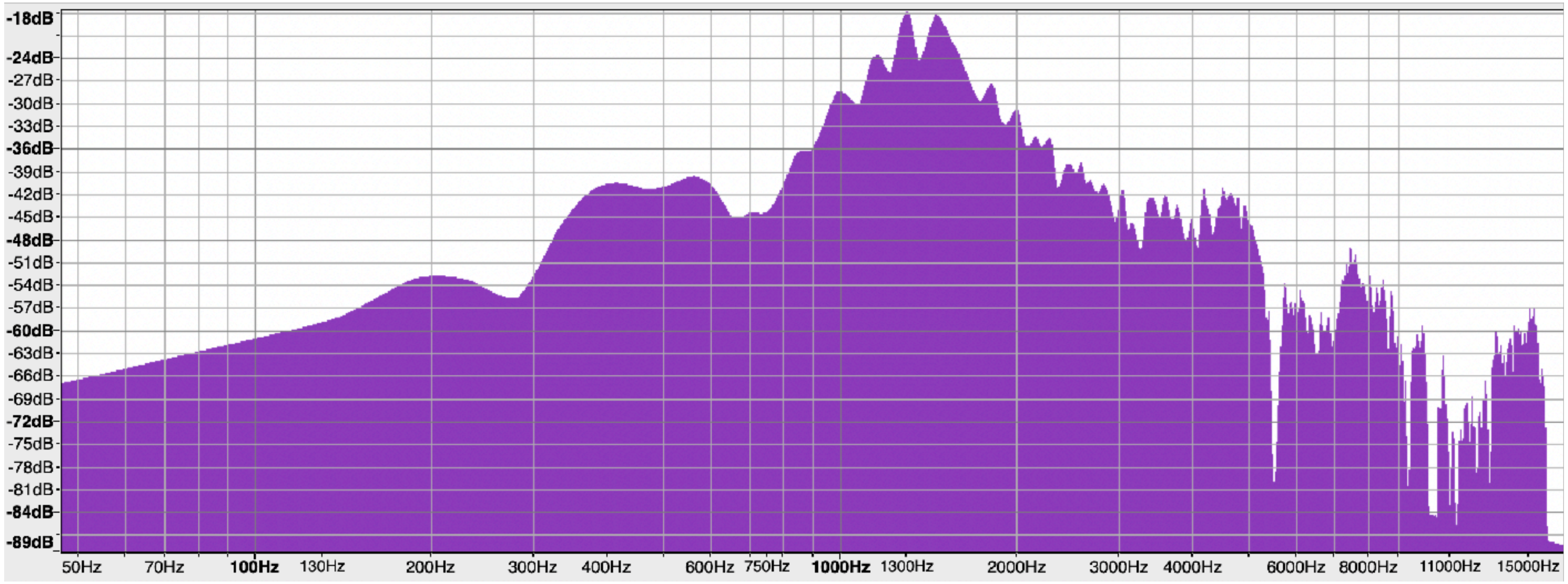
Duration

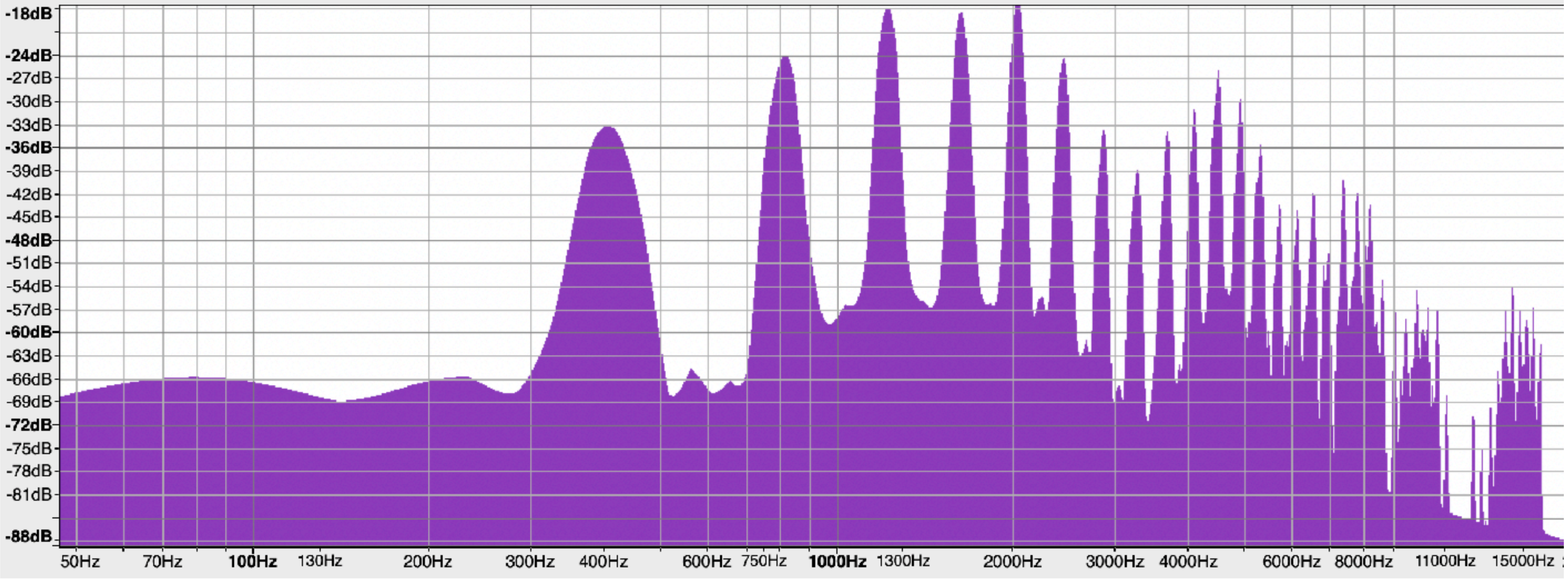




Stress/Energy




-18dB
-24dB
-27dB
-30dB
-33dB
-36dB
-39dB
-42dB
-45dB
-48dB
-51dB
-54dB
-57dB
-60dB
-63dB
-66dB
-69dB
-72dB
-75dB
-78dB
-81dB
-84dB
-89dB
50Hz
70Hz
100Hz
130Hz
200Hz
300Hz
400Hz
600Hz
750Hz
1000Hz
1300Hz
2000Hz
3000Hz
4000Hz
6000Hz
8000Hz
11000Hz
15000Hz


-18dB
-24dB
-27dB
-30dB
-33dB
-36dB
-39dB
-42dB
-45dB
-48dB
-51dB
-54dB
-57dB
-60dB
-63dB
-66dB
-69dB
-72dB
-75dB
-78dB
-81dB
-88dB
50Hz
70Hz
100Hz
130Hz
200Hz
300Hz
400Hz
600Hz
750Hz
1000Hz
1300Hz
2000Hz
3000Hz
4000Hz
6000Hz
8000Hz
11000Hz
15000Hz

Summary

- Source-filter model
  - Independence of source and filter
    - vowels of the same timbre can have different pitches
    - vowels of the same pitches can have different timbre

- Prosody
  - Pitch: related to fundamental frequency (F0)
  - Duration: how long a word/phoneme pronounced
  - Stress: whether a phonetic unit is emphasized

Thanks

Enjoy your break!

Backup: convolution vs cross-correction vs autocorrelation