Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 5: Vocal Messages

Louis-Philippe Morency
Jeffrey Girard

Originally developed with help from
Stefan Scherer and Tadas Baltrušaitis

Outline of this week's lectures

- Multiple Layers of Vocal Messages
  - What we can convey with speech
- Fundamentals of speech production and hearing
  - Anatomy of the vocal tract and the physiology of hearing
  - Fundamental speech measures (direct vs. perceptual measures)
- Prosodic manipulation and its meaning
- Use and detection of varying voice quality
- Nonverbal vocal expressions
  - Laughter, pause filler (e.g. uh, um), and moans
- Practical tools for speech signal processing
- Automatic Techniques for visual processing

Upcoming Schedule

- Week 5
  - Tuesday 2/12: Lecture on Vocal Messages
  - Thursday 2/14: Discussion (visual & vocal messages)
- Week 6:
  - Tuesday 2/19: Lecture on Verbal Messages
  - Thursday 2/21: Proposal presentations
  - Sunday 2/24: Due date for proposal reports
- Week 7:
  - Tuesday 2/26: Lecture on Statistical Analysis
  - Thursday 2/28: Discussion (verbal messages)

A story is told as much by silence as by speech.

Susann Griffin

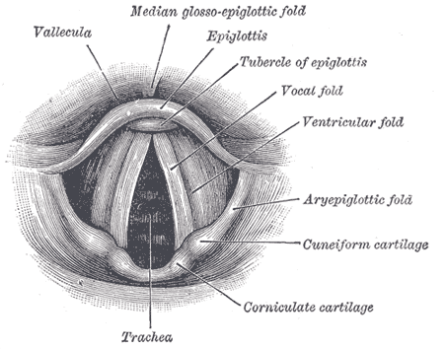
Multiple Layers of Vocal Messages

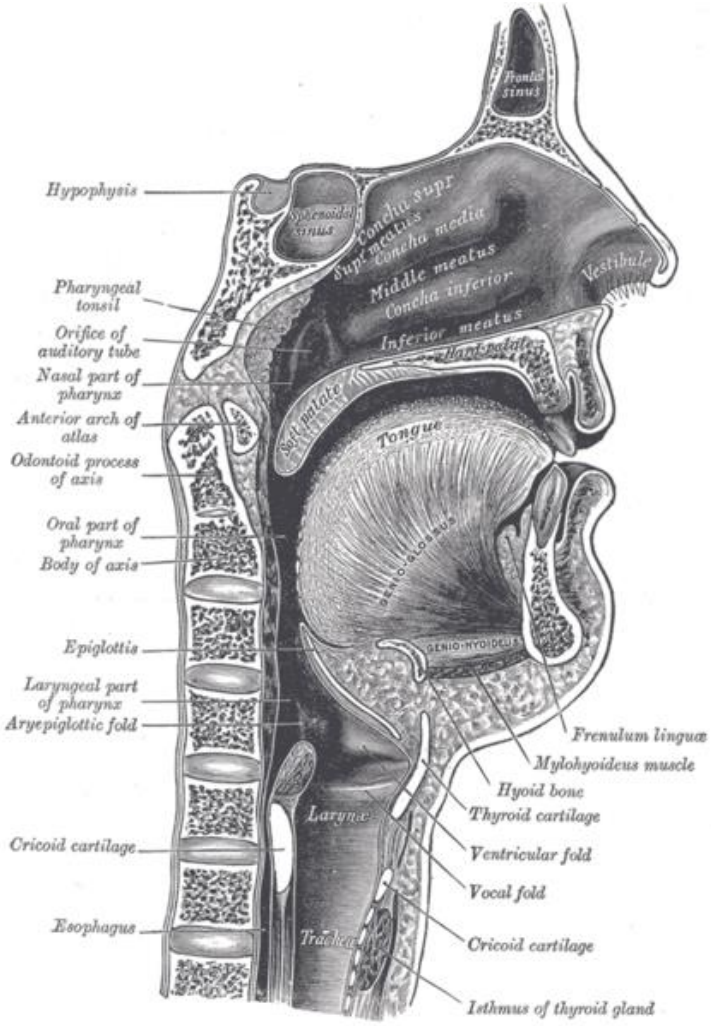
- **Linguistic layer**
  - Carries the semantic information
  - Language, grammar, phonemes, parts of prosody
- **Paralinguistic layer**
  - Non-linguistic and nonverbal layer
  - Conveys information about current affective state, mood, attitude etc.
  - Voice quality, prosody, nonverbal vocal expressions (e.g. laughter, moans, sighs)
  - Main topic of this lecture!
- **Extralinguistic layer**
  - Identifies the speaker (e.g. age, gender, pitch range, habitual characteristics)

Fundamentals of Speech

Anatomy of the vocal tract

- Speech production involves multiple organs that shape the sound
  - Lungs
  - Vocal folds
  - Mouth/tongue
  - Nasal cavity
  - Lips





Pictures from Gray's anatomy 1918

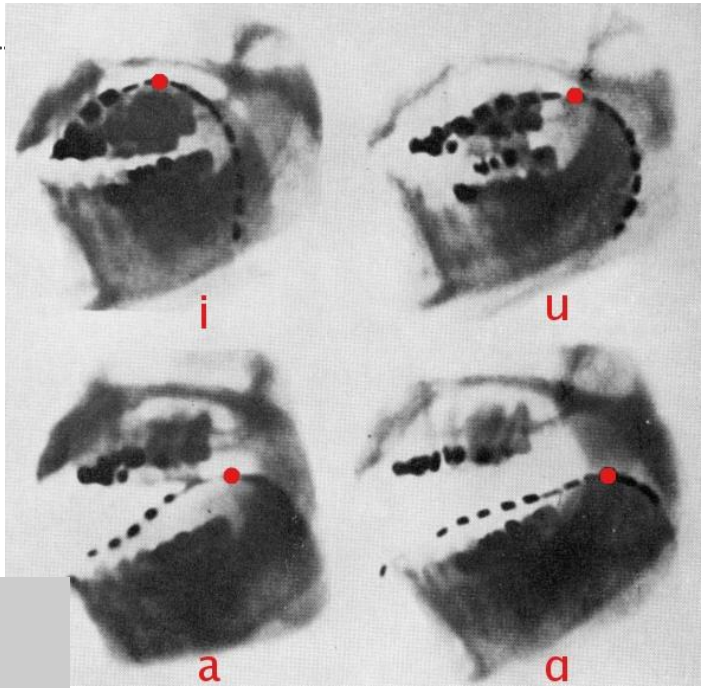
Anatomy of the vocal tract (2)

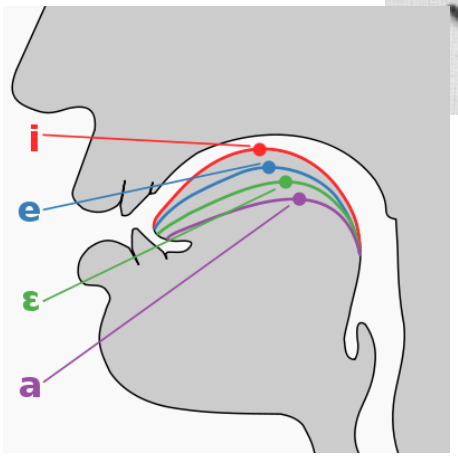
- Vocal folds vibrate and produce **fundamental frequency** (f0)
- Vibration is based on **muscular tension** and **air pressure**

Source: youtube.com

Anatomy of the vocal tract (3)

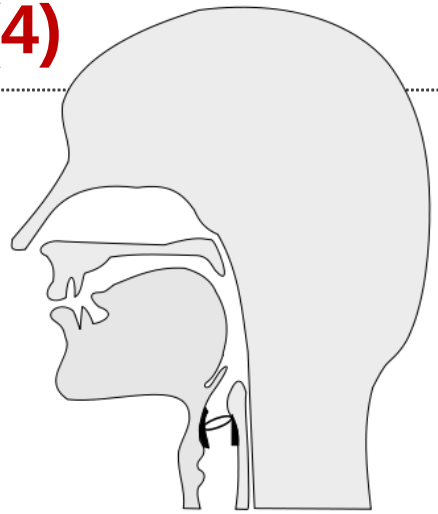
- **Vowels**
- **Tongue position** influences the tone
- Tongue modulates the length of the cavities in the mouth





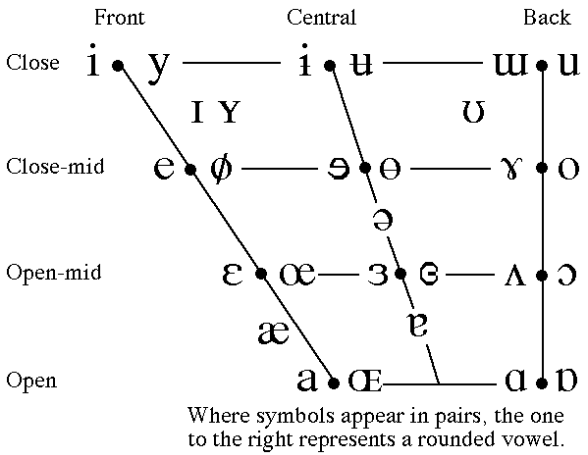
Source: wikipedia.org

Anatomy of the vocal tract (4)



- Mouth opening, lip rounding and teeth influence the sound

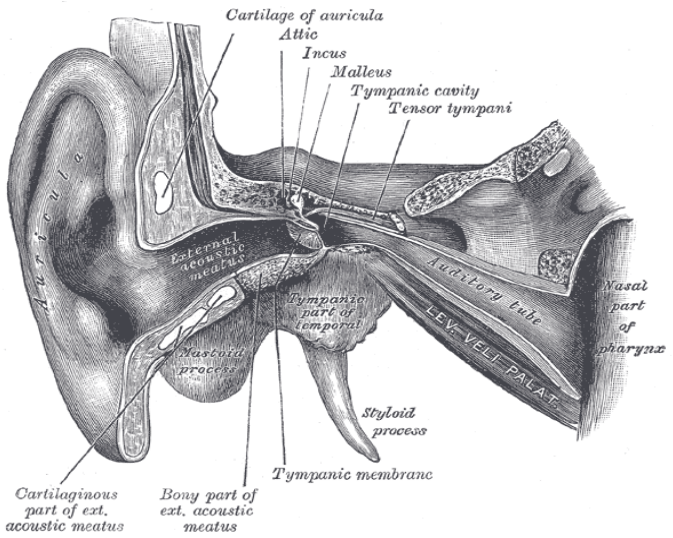
- Other Sounds:
  - Nasal sounds (e.g. [m], [n])
  - Stops (e.g. [k],[t],…)
  - …



Ladefoged, A course in phonetics, 2004

Anatomy of the hearing organ

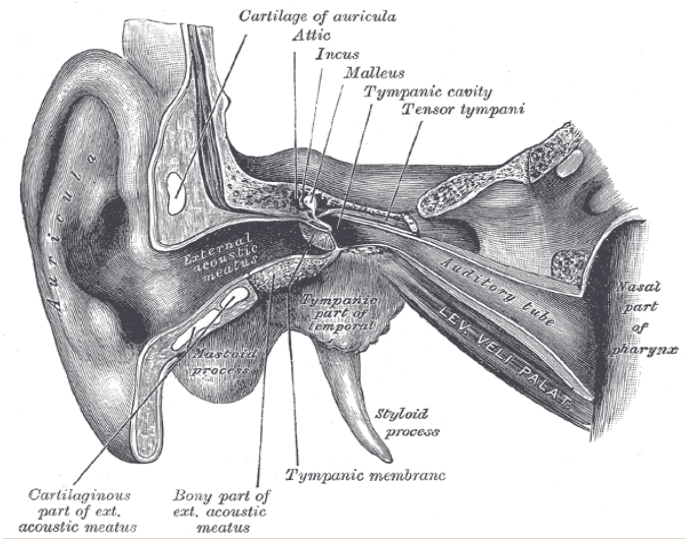
- Outer ear collects sound waves
  - Filters and allows directionality detection
- In the middle ear the sound waves are transferred via drum and three little bones called ossicles
- The inner ear transfers physical waves into nervous signals

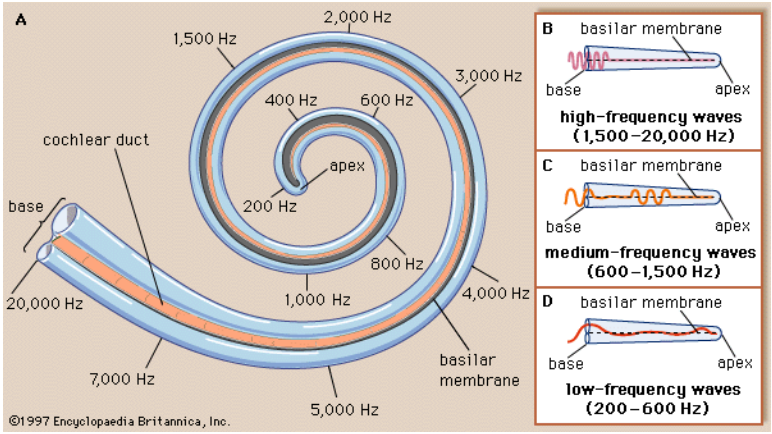


Pictures from Gray's anatomy 1918

Anatomy of the hearing organ (2)

- Transformation between sound waves and nervous signals sent to the brain is not linear
- There are perceptual differences:
  - Not every frequency is perceived with the same intensity or accuracy

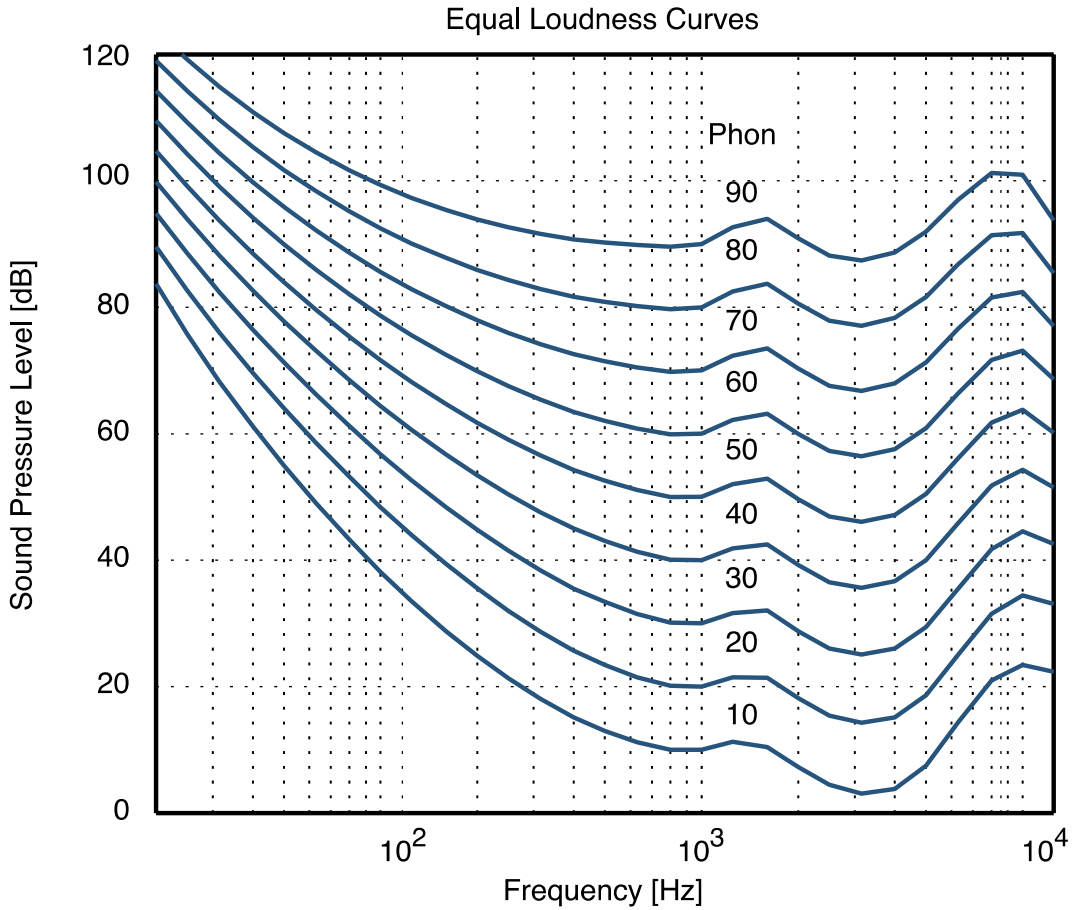




Pictures from Gray's anatomy 1918

Anatomy of the hearing organ (3)

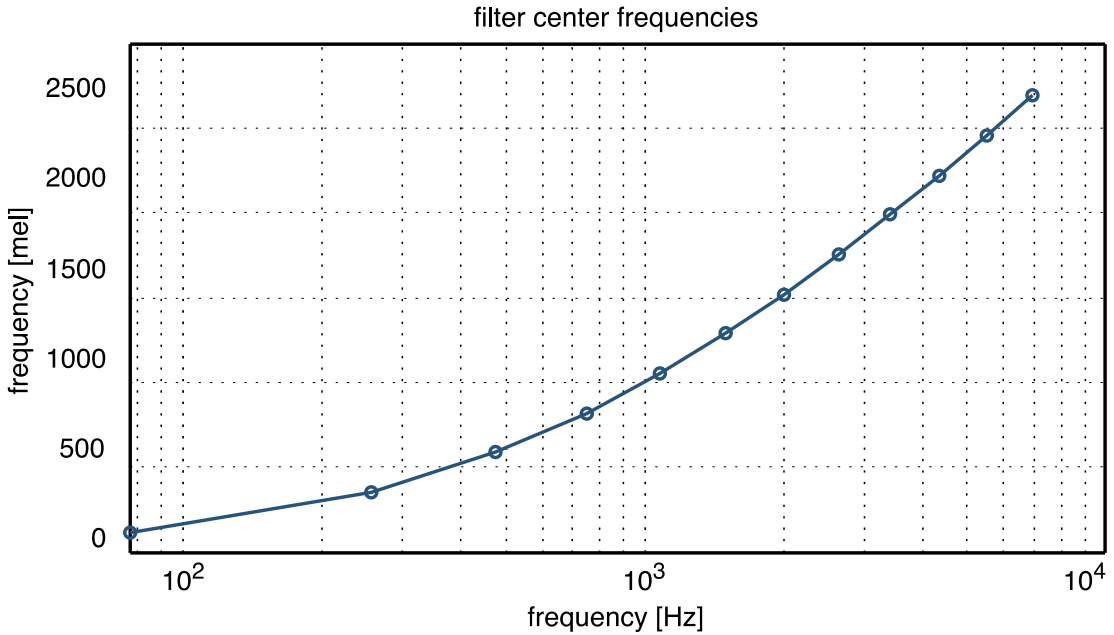
- Perceptual measures
  - Loudness vs. intensity



Try it out: http://tinyurl.com/px5g7cf

Anatomy of the hearing organ (4)

- Perceptual measures
  - Loudness vs. intensity
  - Pitch vs. fundamental frequency



Try it out: http://tinyurl.com/px5g7cf

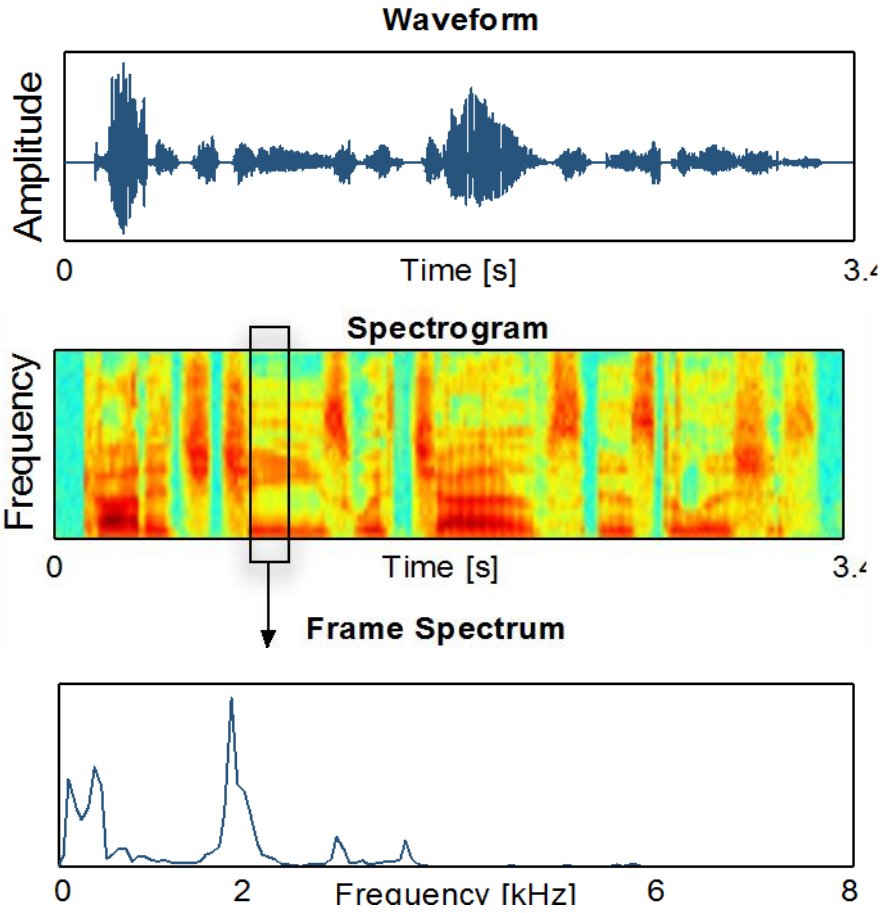
Representation Of Speech

Representations of speech

Different domain representations:

- Time
- frequency

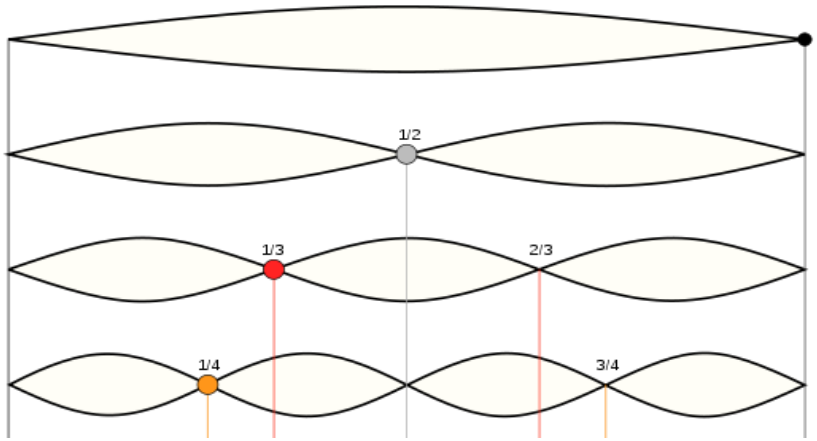
- Signal can be transferred from one domain to the other using **Fourier transformation**
  - Measures the amount of "presence" of a sine wave with a certain frequency in the original signal
- Two types: **broadband** and **narrowband** spectrogram
  - Bandwidth is defined by width of analysis window

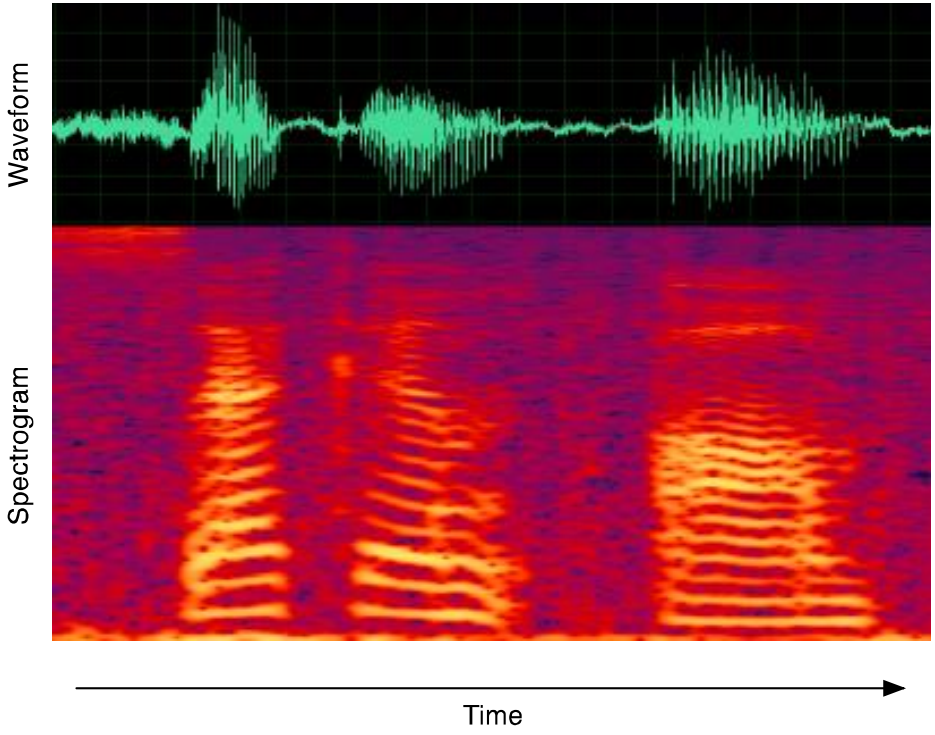


Fundamental Frequency

Fundamental frequency (f0) is the basic resonance of the vocal folds

- Harmonics are multiples of f0



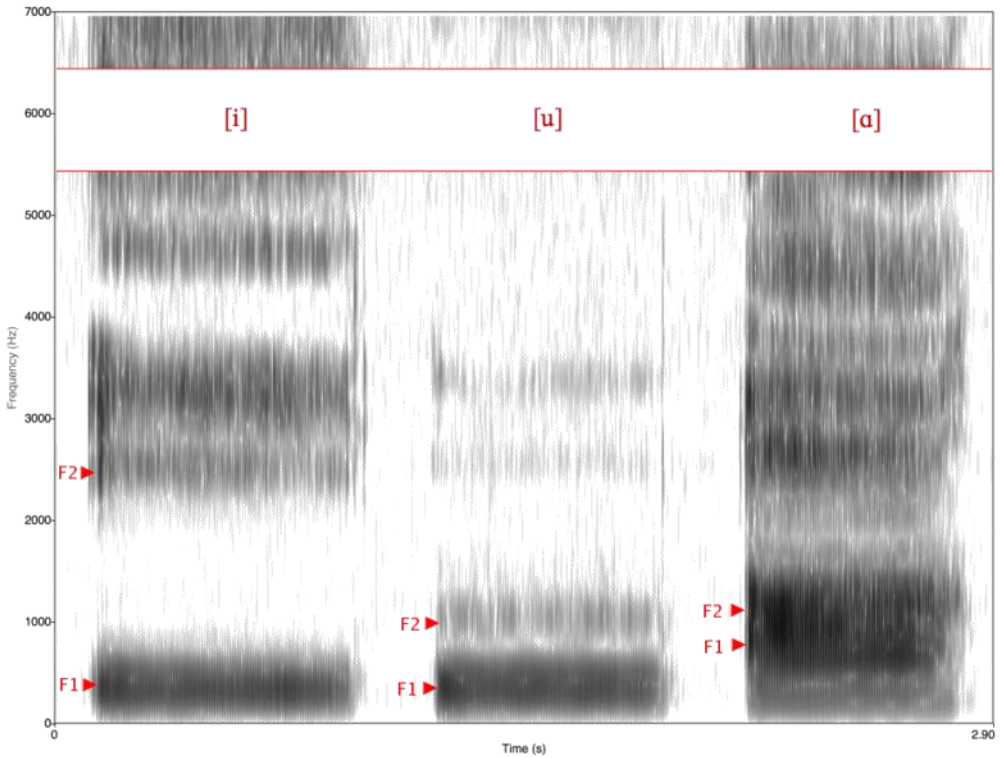


Source: wikipedia.org

Formant Frequencies

Formant frequencies are resonance frequencies dependent on the length of the vocal tract cavities

- The vocal tract is manipulated by the tongue etc.



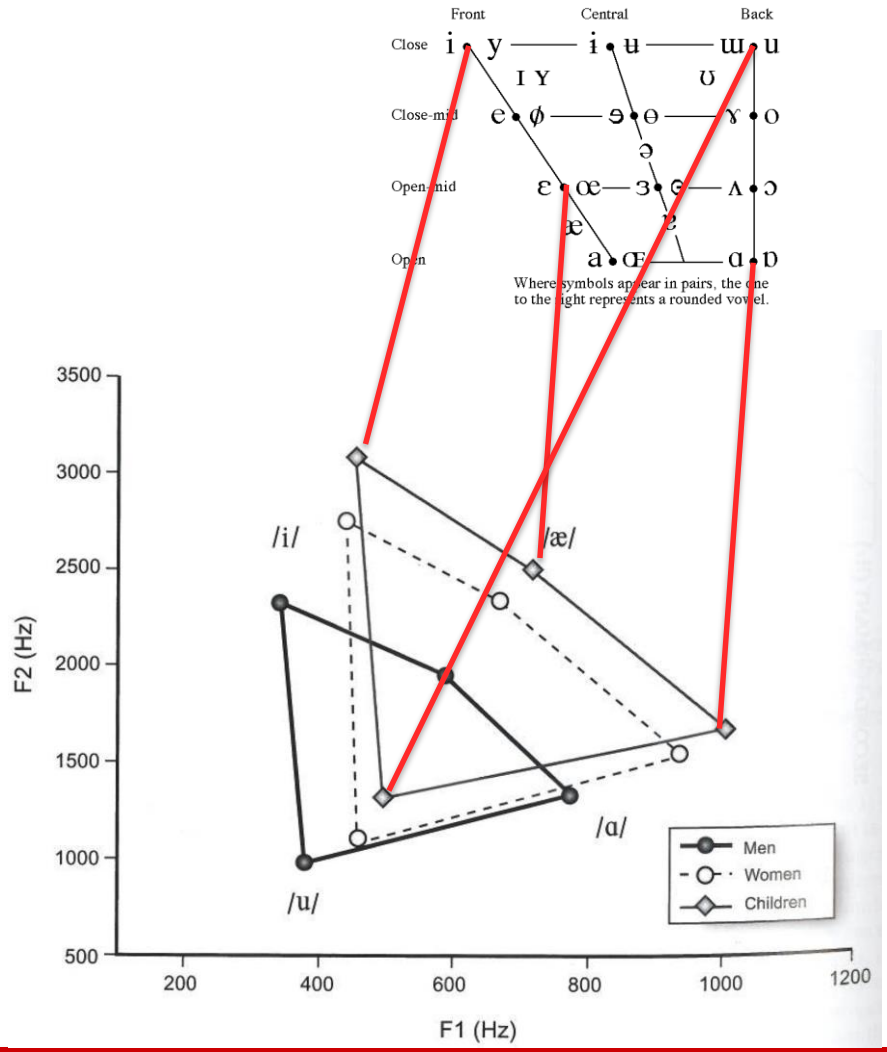
Source: wikipedia.org

Example Use of Formant Frequencies

Vowel space

- Gender and age define vowel space size.

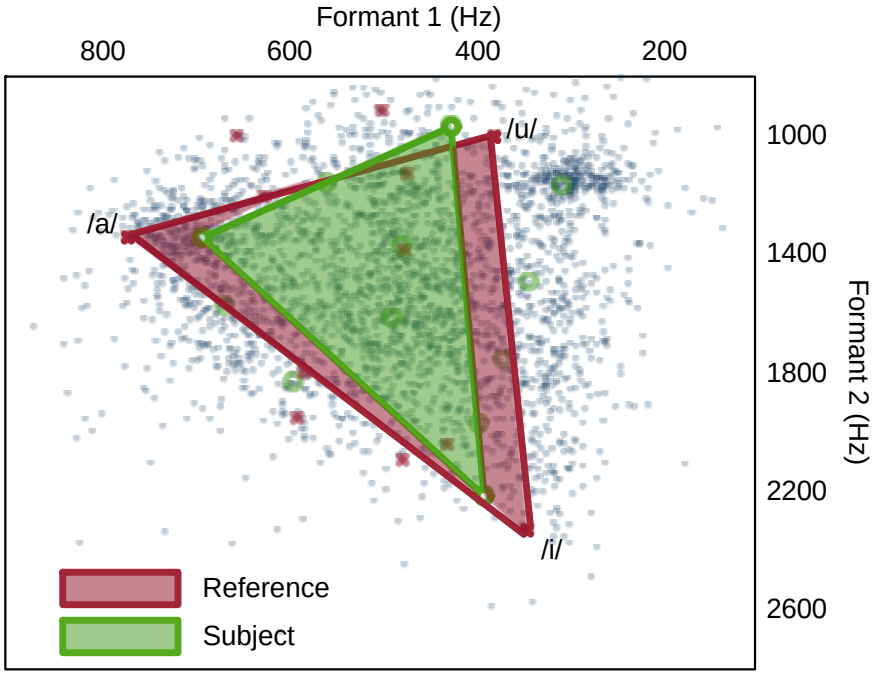
What could be a reason for reduced vowel space?



Example Use of Formant Frequencies

Vowel space

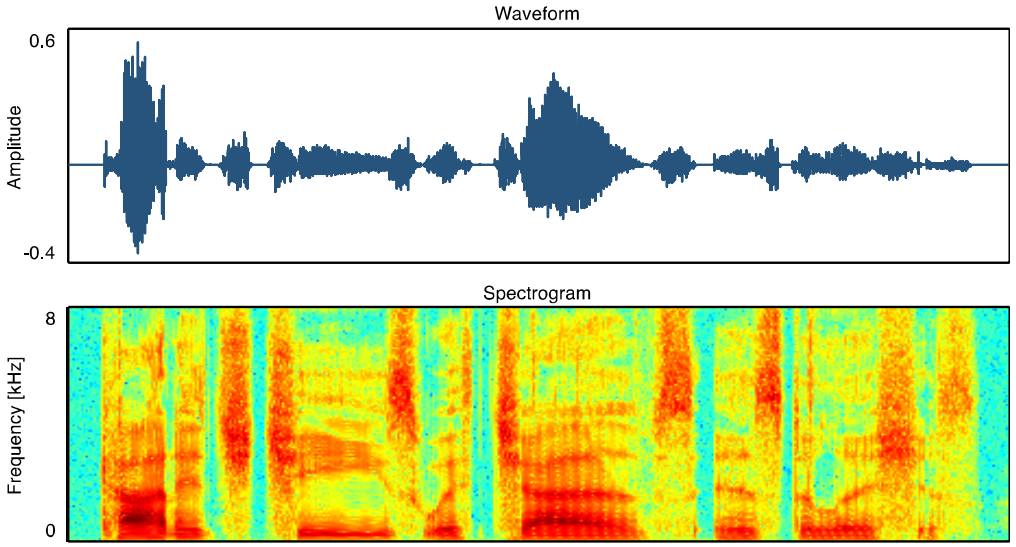
- **Medical relevance:**
  Parkinson's Disease
  ALS
  Depression



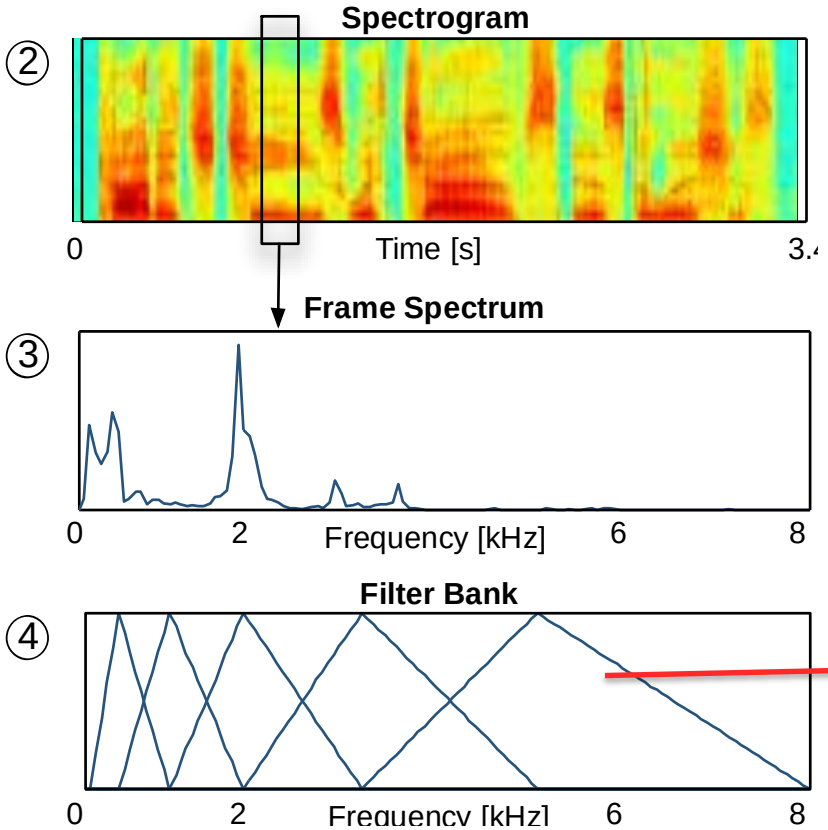
Mel frequency cepstral coefficients (MFCC)

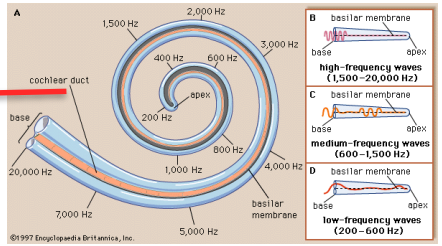
Mel frequency cepstral coefficients (MFCC)

- Compact representation of the spectrum
- Emulates human hearing
- Popular for speech recognition

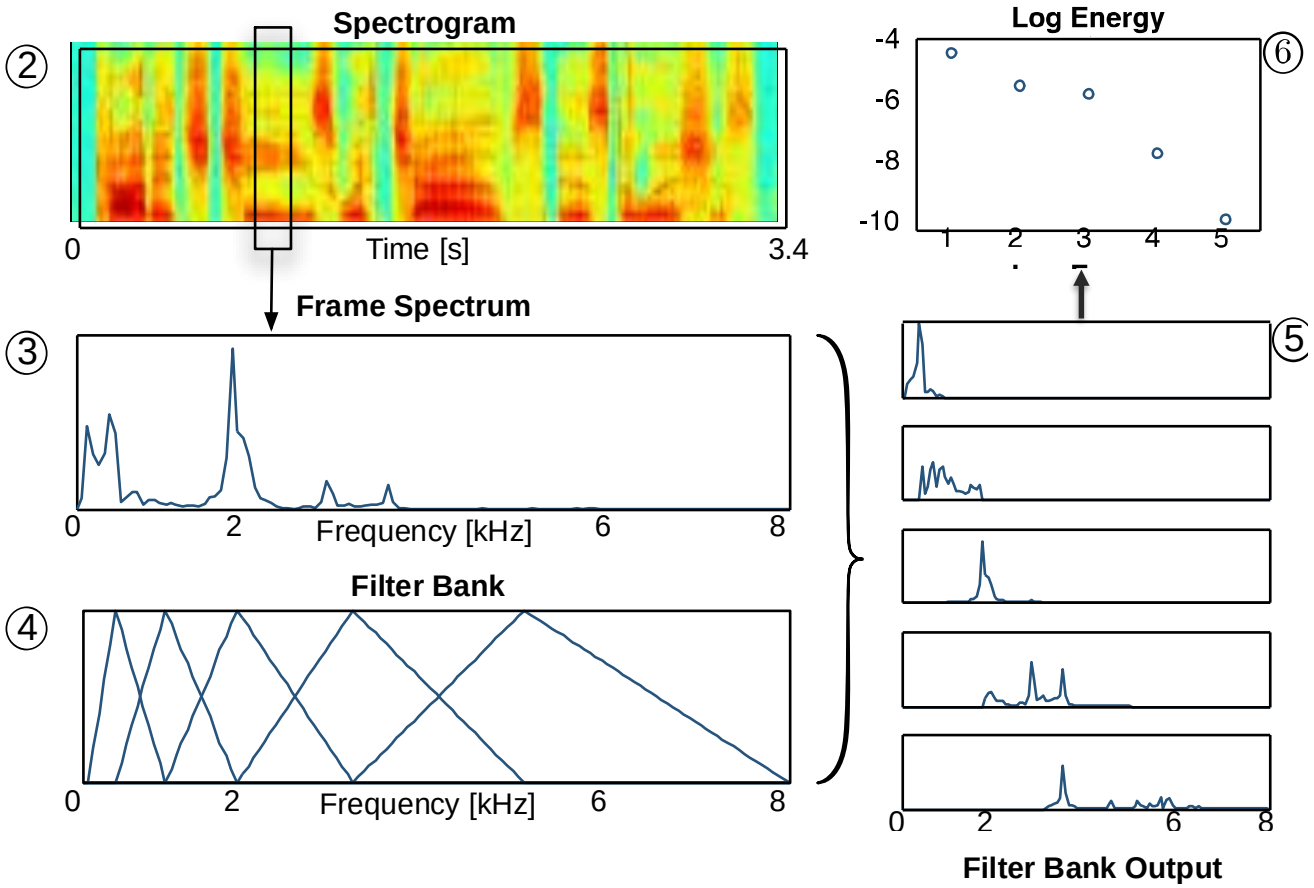


Mel frequency cepstral coefficients (MFCC)





Mel frequency cepstral coefficients (MFCC)



Prosody

Prosody can strongly influence the meaning

Elements of prosody

- Prosody, the suprasegmental envelope of an utterance, is composed by:
  - Syllable length
  - Loudness
  - Pitch
  - Pauses
- Prosody influences and defines:
  - Prosodic boundaries
  - Question or statement
  - Sarcasm
  - Emotional state
  - Meaning of words (e.g. in Chinese)

50 shades of "*yeah*"

Prosodic Boundaries

Can you hear the difference?

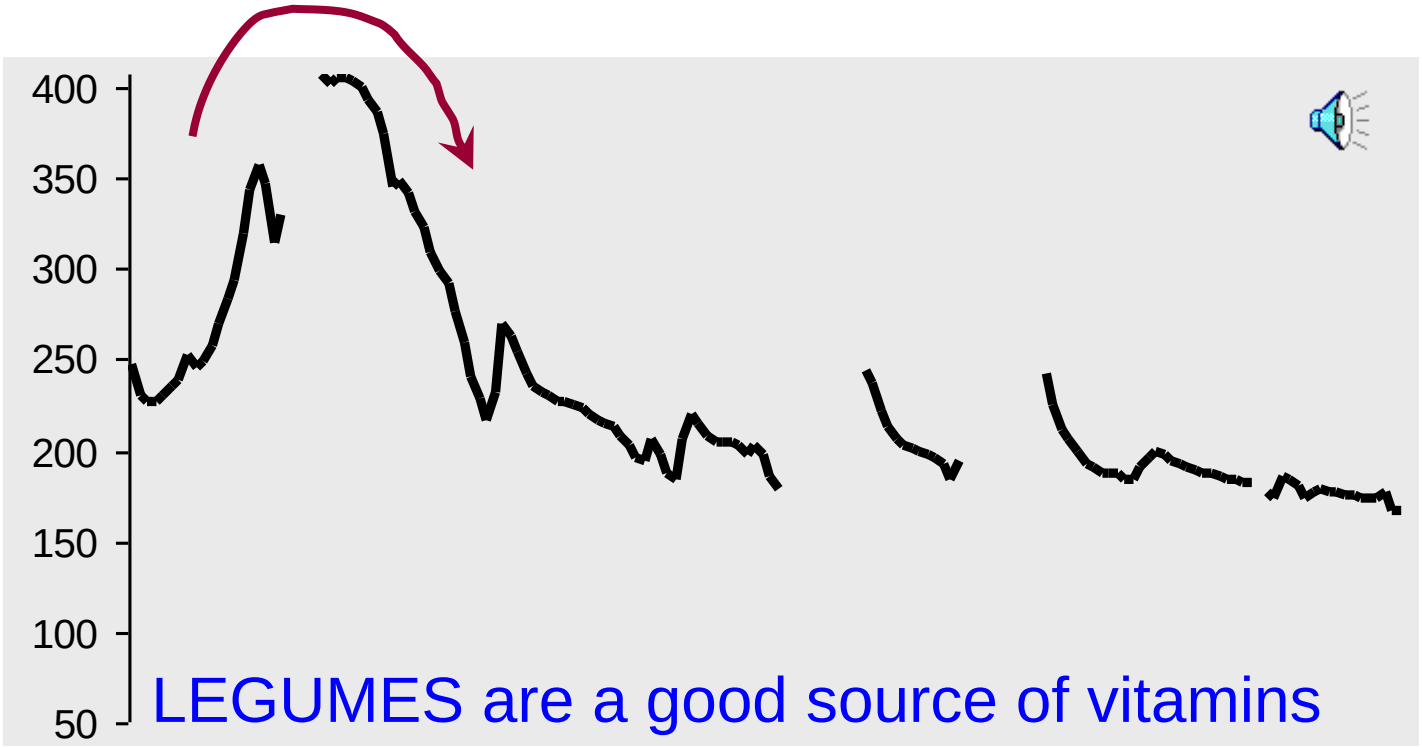
I met Mary and Elena's mother at the mall yesterday.

Sally saw % the man with the binoculars.
Sally saw the man % with the binoculars.

When Madonna sings % the song is a hit.
When Madonna sings the song % it's a hit.

Same ‘tune’, different alignment

What is a good source of vitamins?

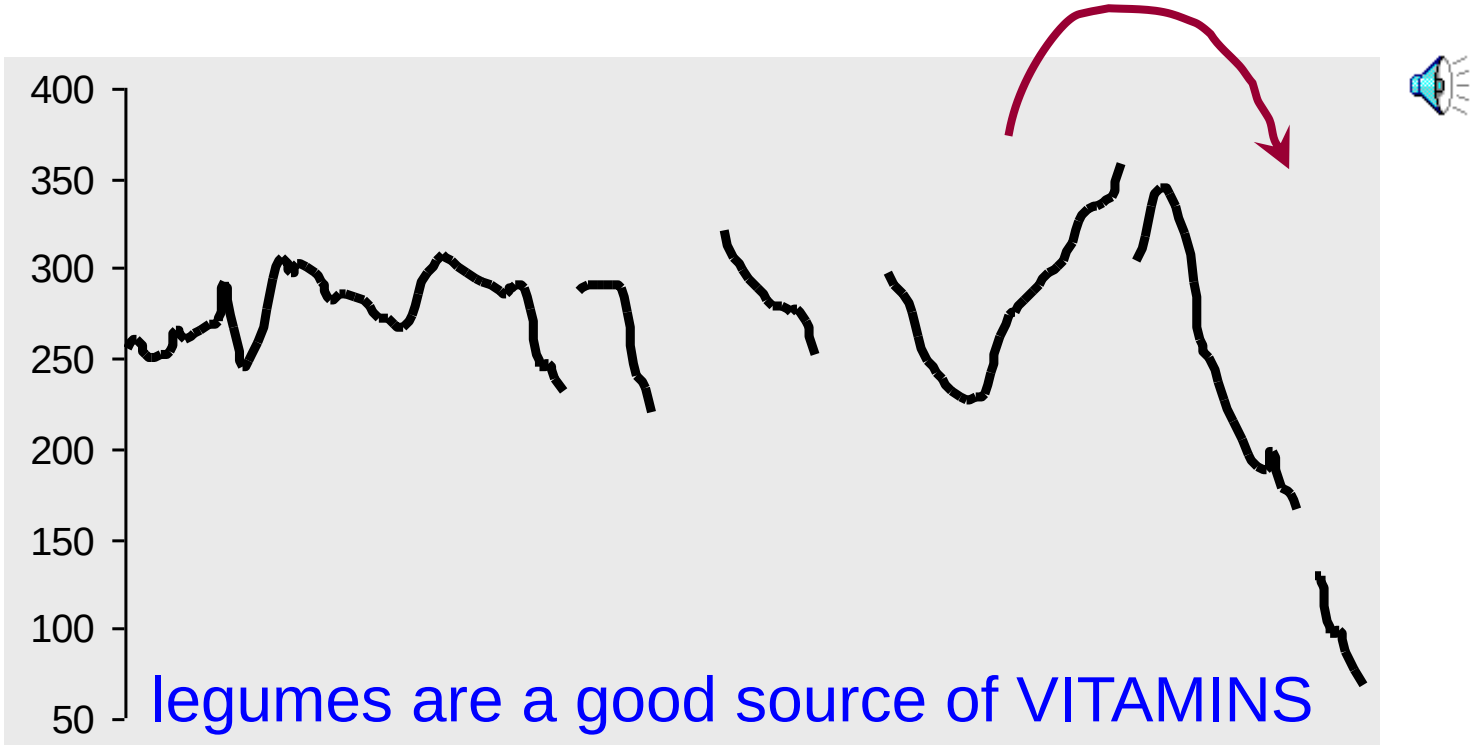


Same ‘tune’, different alignment

Are legumes a source of vitamins?

Same ‘tune’, different alignment

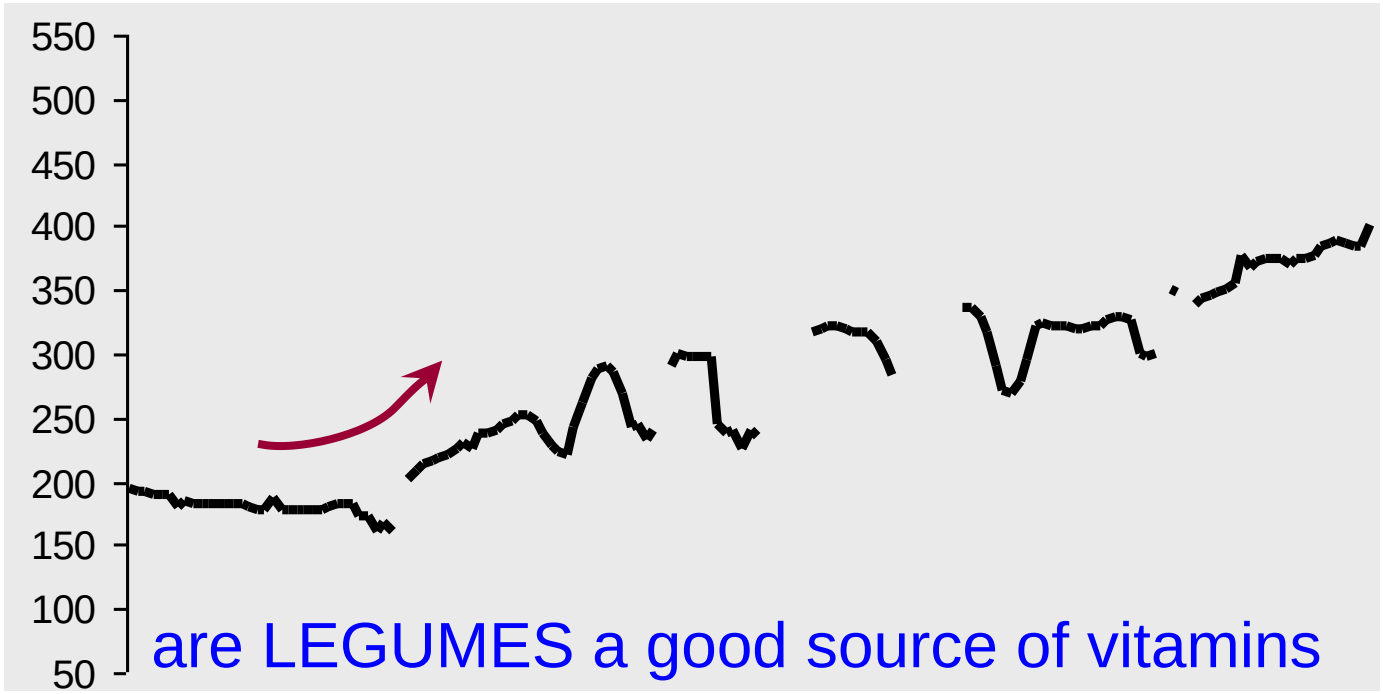
What are legumes a good source of?



Other Uses of Pitch Contour Analysis

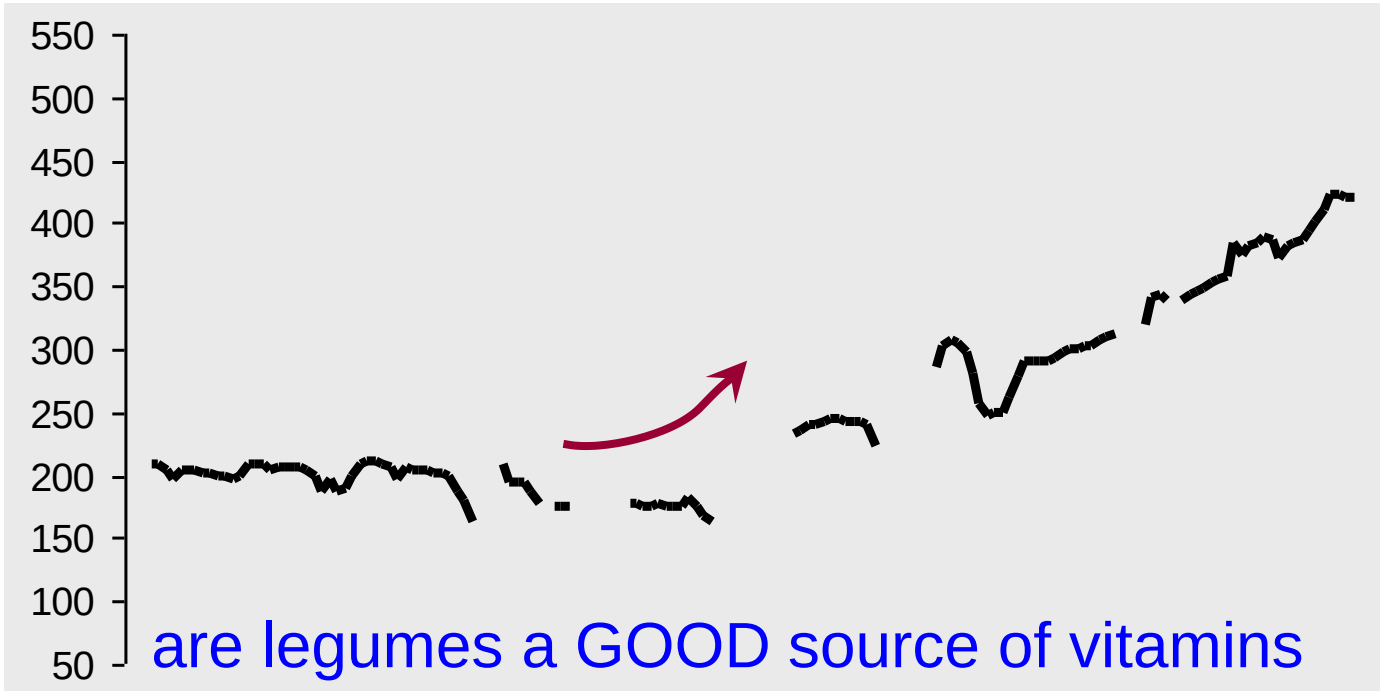
- Rising pitch contour towards the end of an utterance indicates a yes-no question
- WH-questions have a falling pitch contour
- Signaling doubt or uncertainty can be expressed by rising contour in the end of an utterance

Yes-No question tune



Rise from the main accent to the end of the sentence.

Yes-No question tune

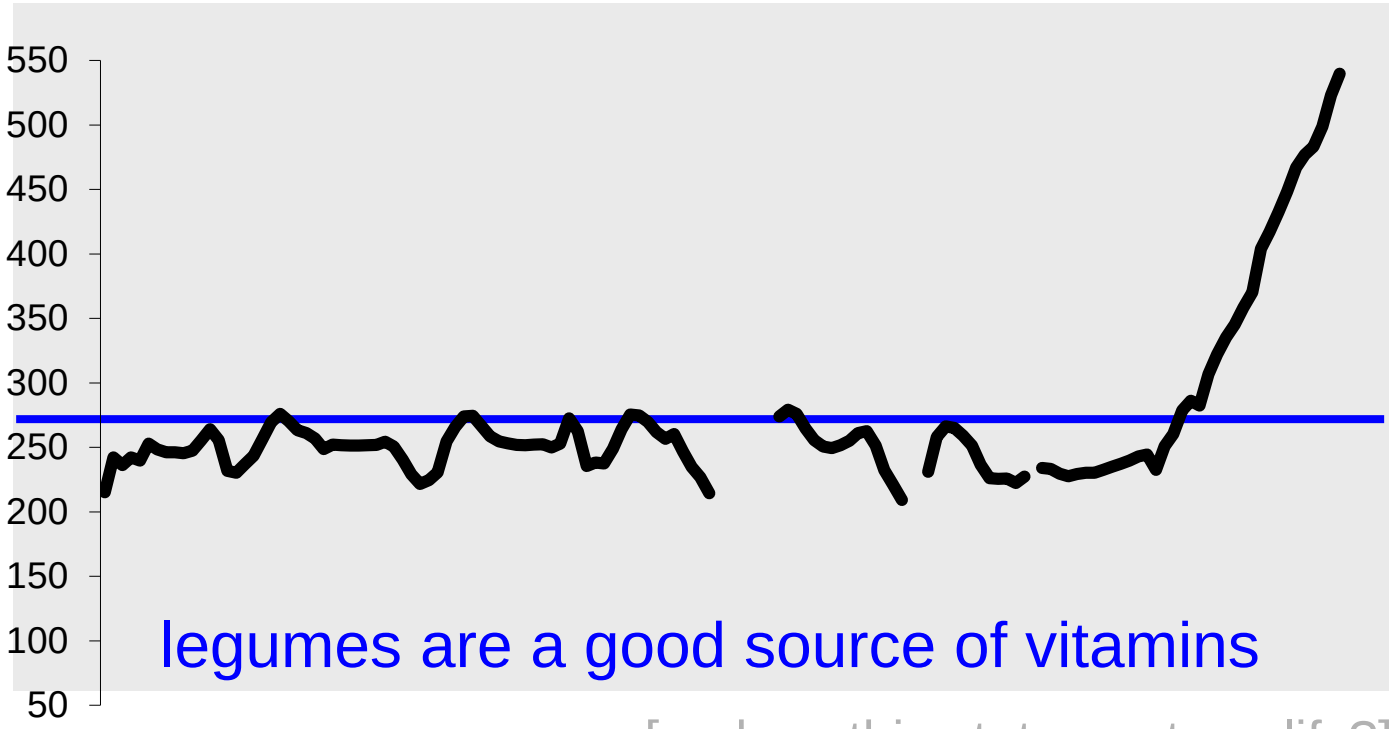


Rise from the main accent to the end of the sentence.

Yes-No question tune

Rise from the main accent to the end of the sentence.

Rising statements



[... does this statement qualify?]

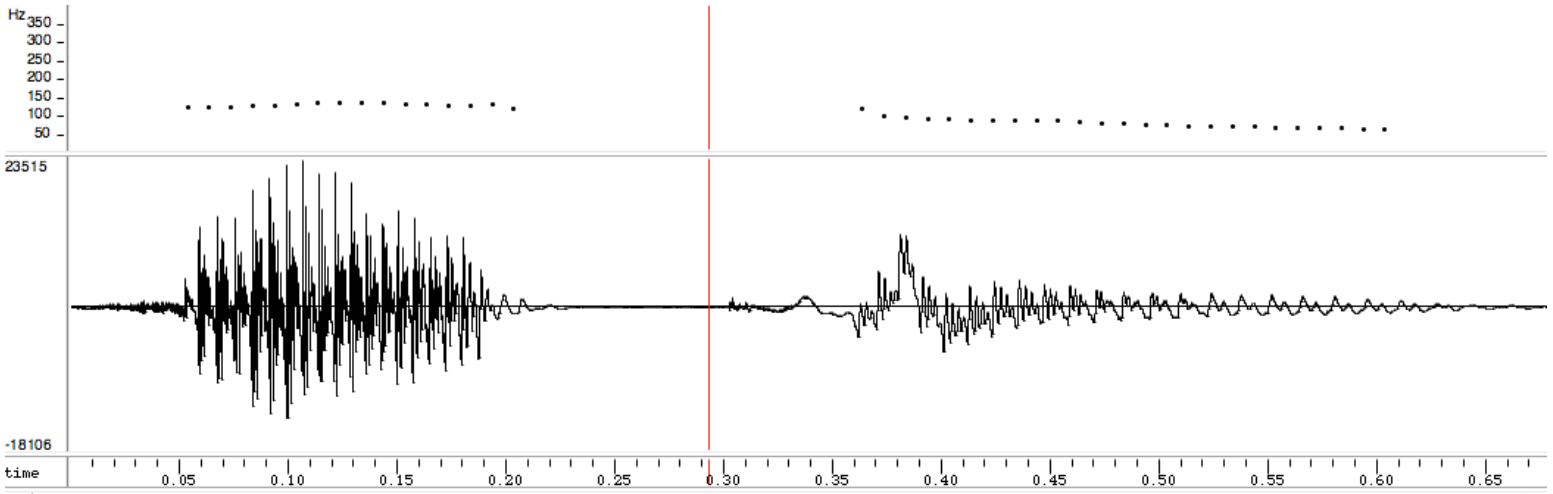
High-rising statements can signal that the speaker is seeking approval.

Emotional state

- Prosody can be used to express emotion

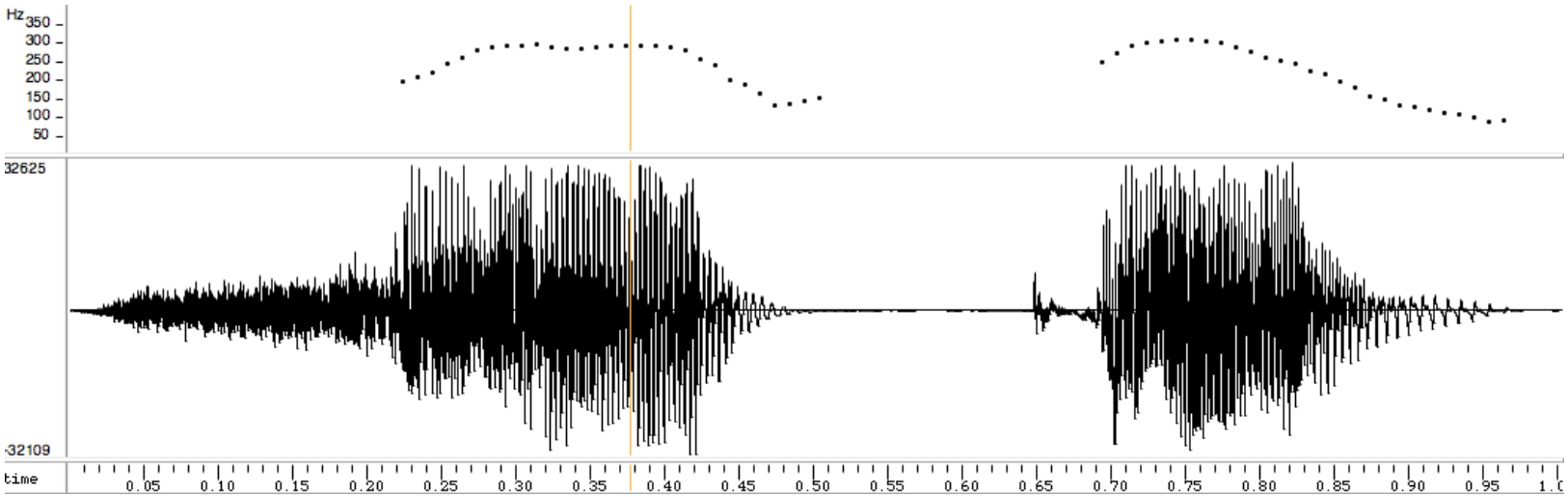
Acoustic Parameters	Arousal/Stress	Happiness	Anger	Sadness	Fear	Boredom
Speech rate and fluency						
Number of syllables	>	>=	<>	<	>	<
Syllable duration	<	<=	<>	>	<	>
Number/duration of pauses	<	<	<	>	<>	>
f_0 and prosody						
f_0 mean	>	>	>	<	>	<=
f_0 deviation	>	>	>	<	>	<
f_0 range	>	>	>	<	<>	<=
Gradient of f_0	>	>	>	<	<>	<=
Vocal Effort/phonation						
Intensity (dB) mean	>	>=	>	<=		<=
Intensity (dB) deviation	>	>	>	<		<
Jitter		>=	>=		>	=
Shimmer		>=	>=		>	=

Emotional state: Neutral



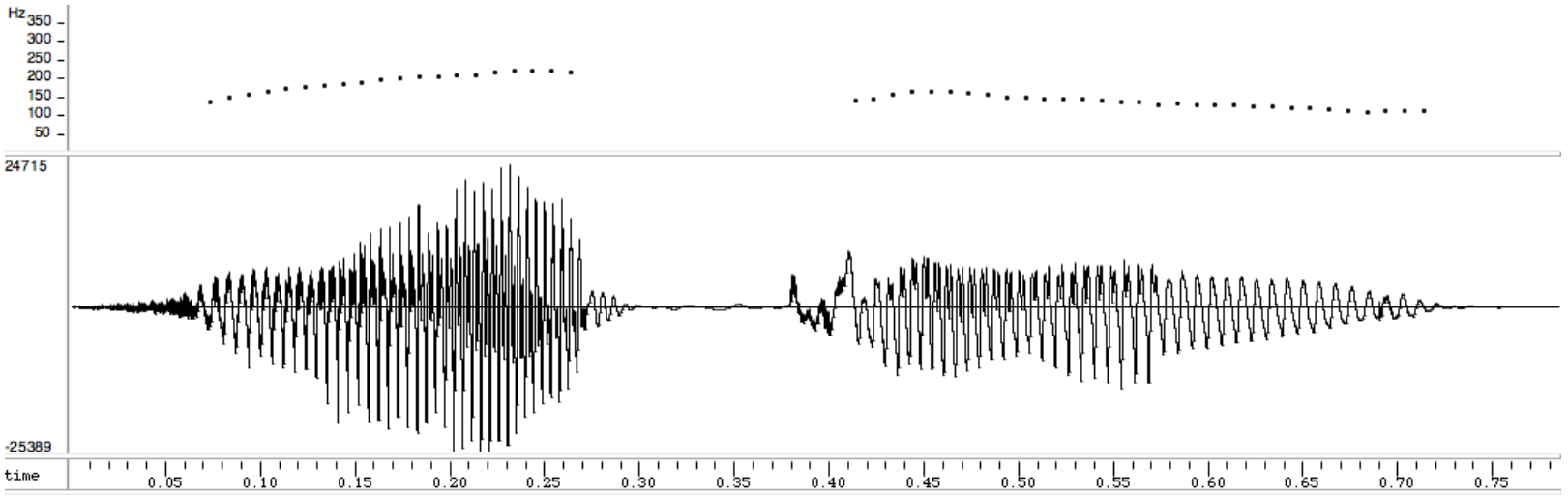
- Low pitch
- Low pitch variation
- Moderate loudness

Emotional state: Angry



- High pitch
- High pitch variation
- High intensity

Emotional state: Fear



- Average pitch
- Average pitch variation
- Average intensity

Emotional state (5)

- QUIZ (*Who knows the Germans best?*)

- Allowed answers:
  - Happiness
  - Anger
  - Sadness
  - Disgust
  - Fear
  - Neutral

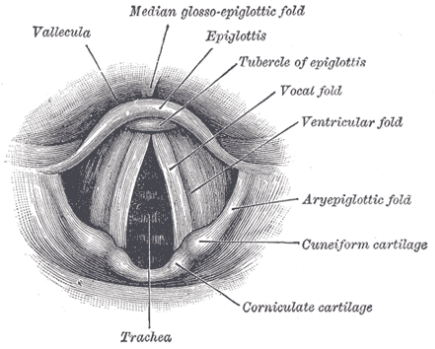
Voice Quality

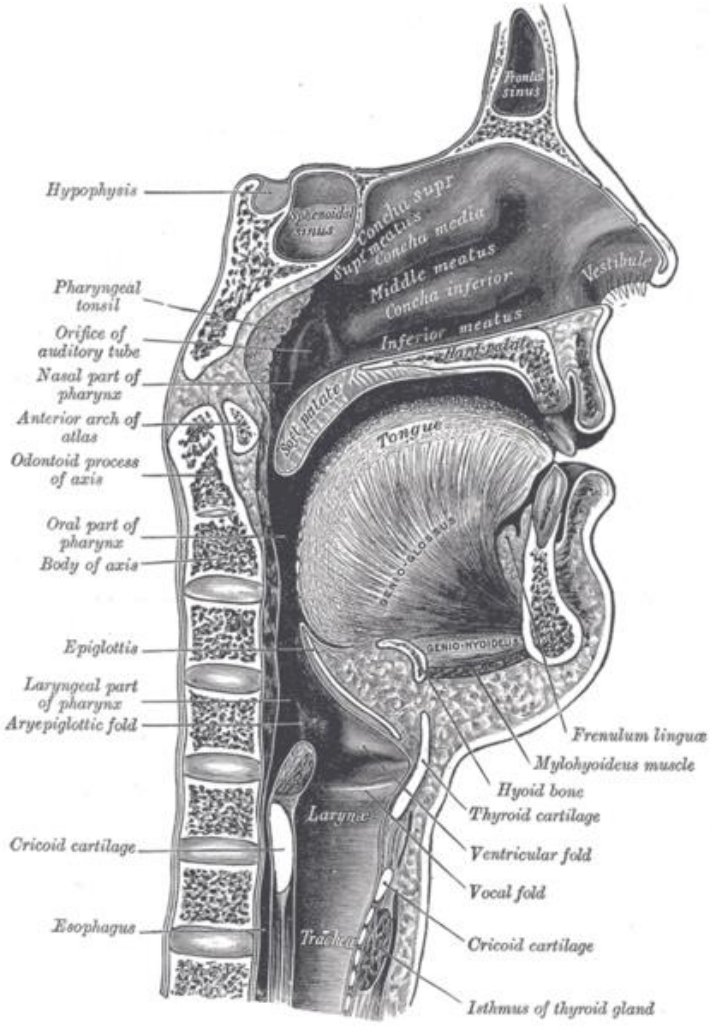
Voice Quality

- **Does not** refer to a term concerning fidelity or "goodness"
- Refers to the **timbre** or **coloring** of a voice
- Functions:
  - Meaning or disambiguation of words (e.g. in Gujarati) (http://www.phonetics.ucla.edu/vowels/chapter12/gujarati.html)
  - Paralinguistic signal
    - Attitude
    - Mood
    - Social factors (e.g. standing, (*inter*-)personality)
    - Affective state
    - Turn-management (e.g. in Finnish)

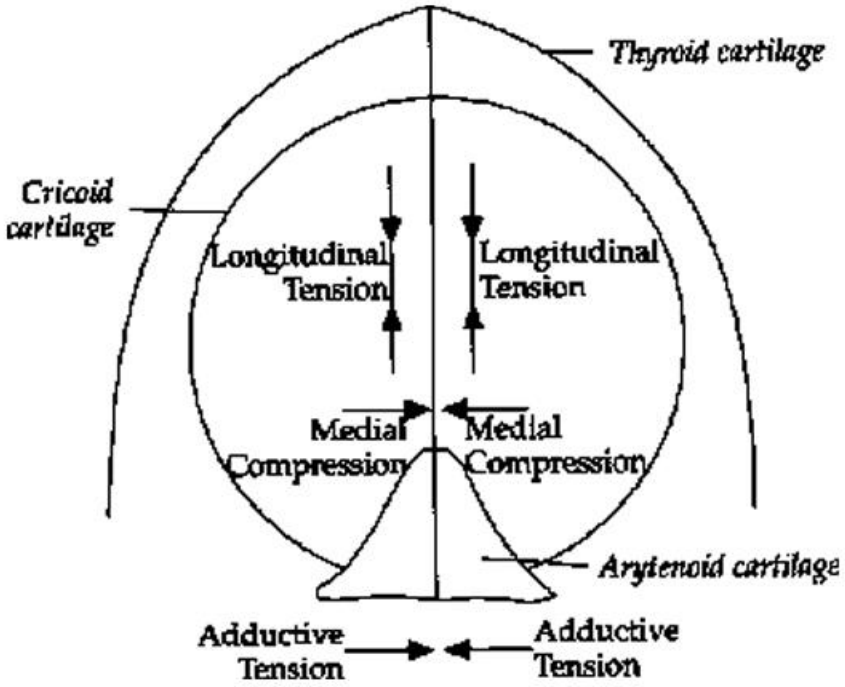
Voice Quality

- Can be seen as the *signal residue* after removing effects of the vocal tract filter
- The phonation and manner of the **vocal folds** vibrate play a major role (exceptions: e.g. whisper)





Vocal Folds: Phonation Gestures



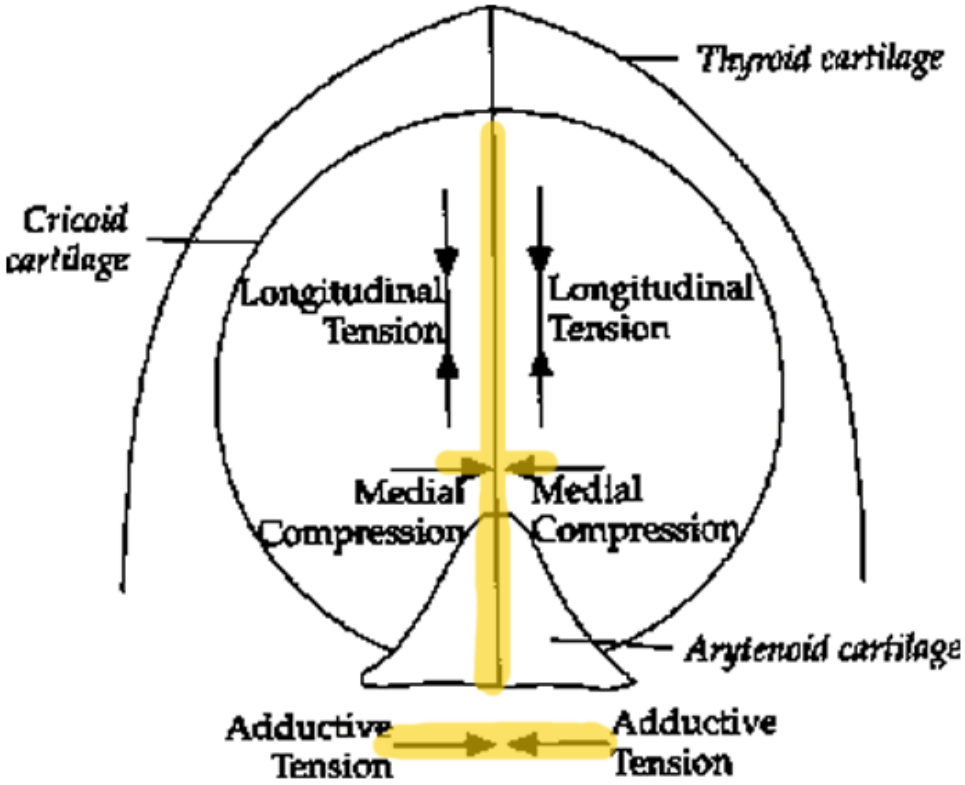
- **Adductive tension:** Movement toward the mi dline of the vocal fold.
- **Medial compression:** adductive force on vocal processes
- **Longitudinal pressure:** tension of vocal folds

Modal Voice

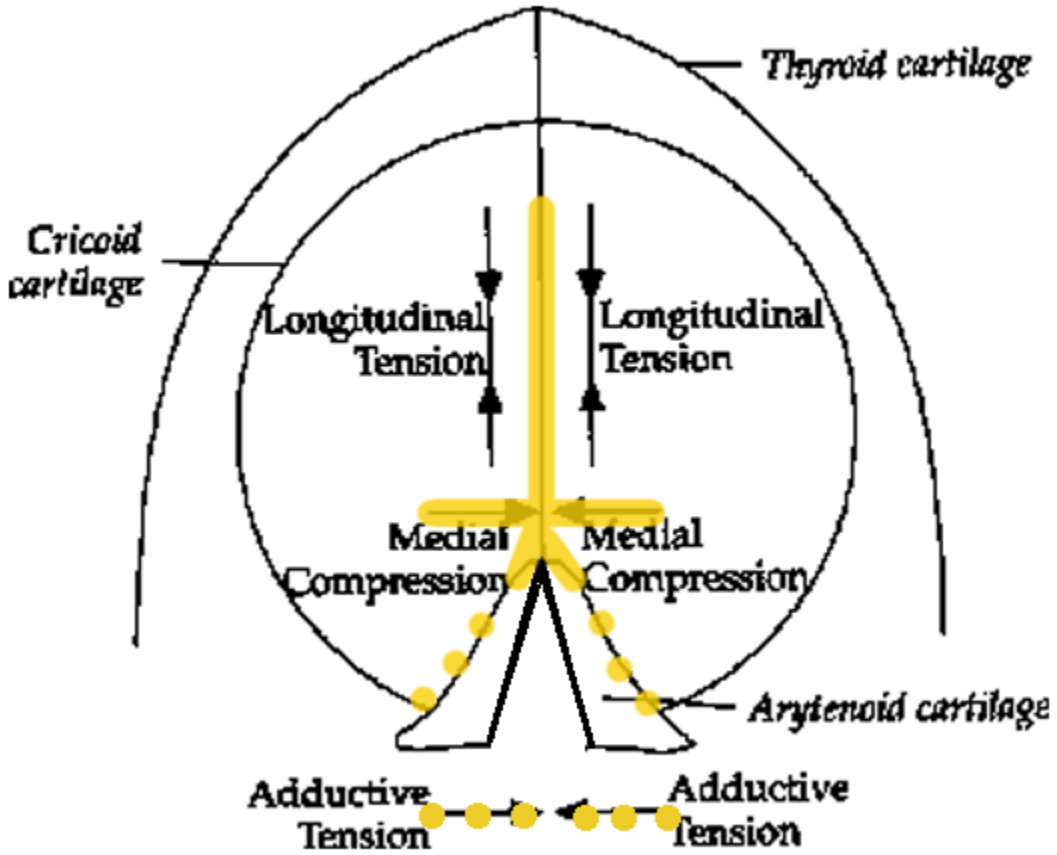
- “Neutral” mode
- Muscular adjustments moderate
- Vibration of vocal folds periodic, full closing of glottis, no audible friction
- Frequency of vibration and loudness in low to mid range for conversational speech

Harsh/Tense Voice

- Very strong tension of vocal folds, very high tension in vocal tract



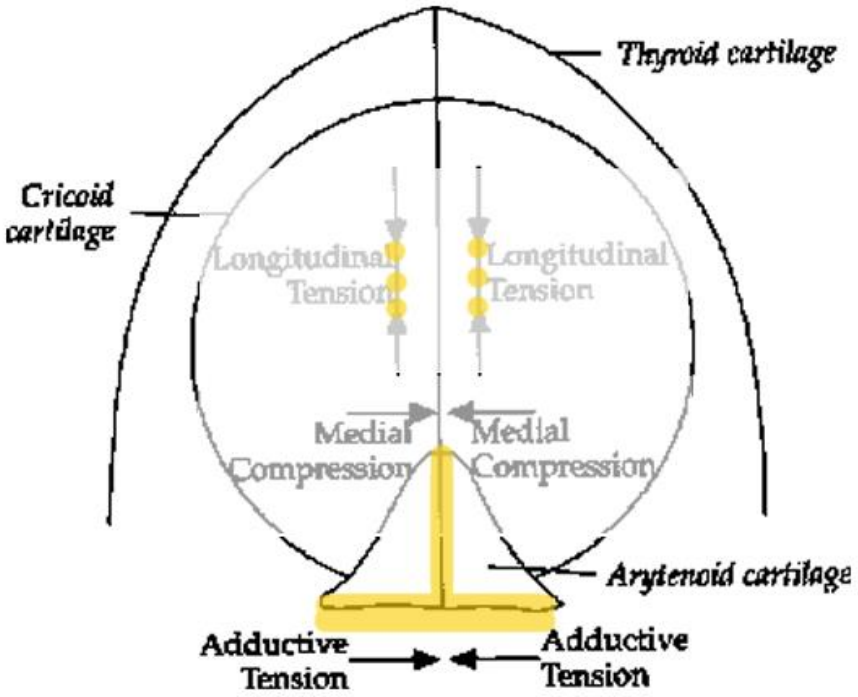
Whispery Voice



- Little or no vocal fold vibration
- Very low adductive tension
- Medial compression moderately high
- Longitudinal tension moderately high
- Turbulence generated by friction of air in and above larynx, with vocal folds not vibrating

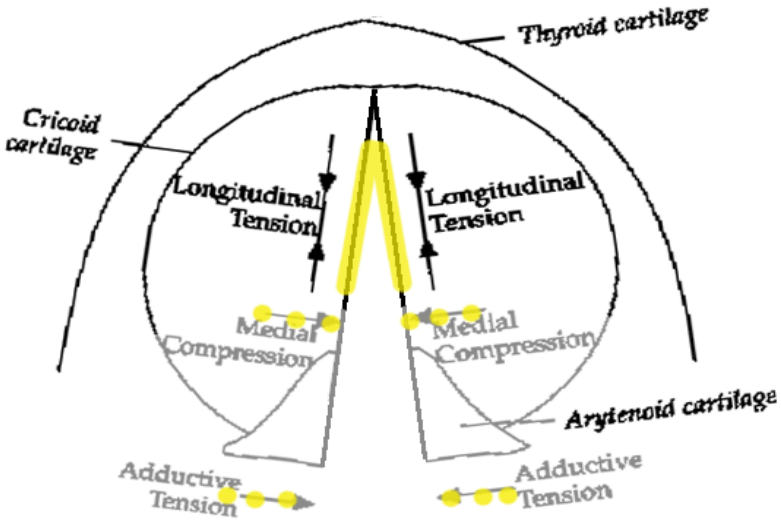
Creaky Voice (vocal Fry)

- Vocal fold vibration at low frequency, irregular
- Low tension
- The vocal folds strongly adducted
- Longitudinal tension weak
- Moderately high medial compression

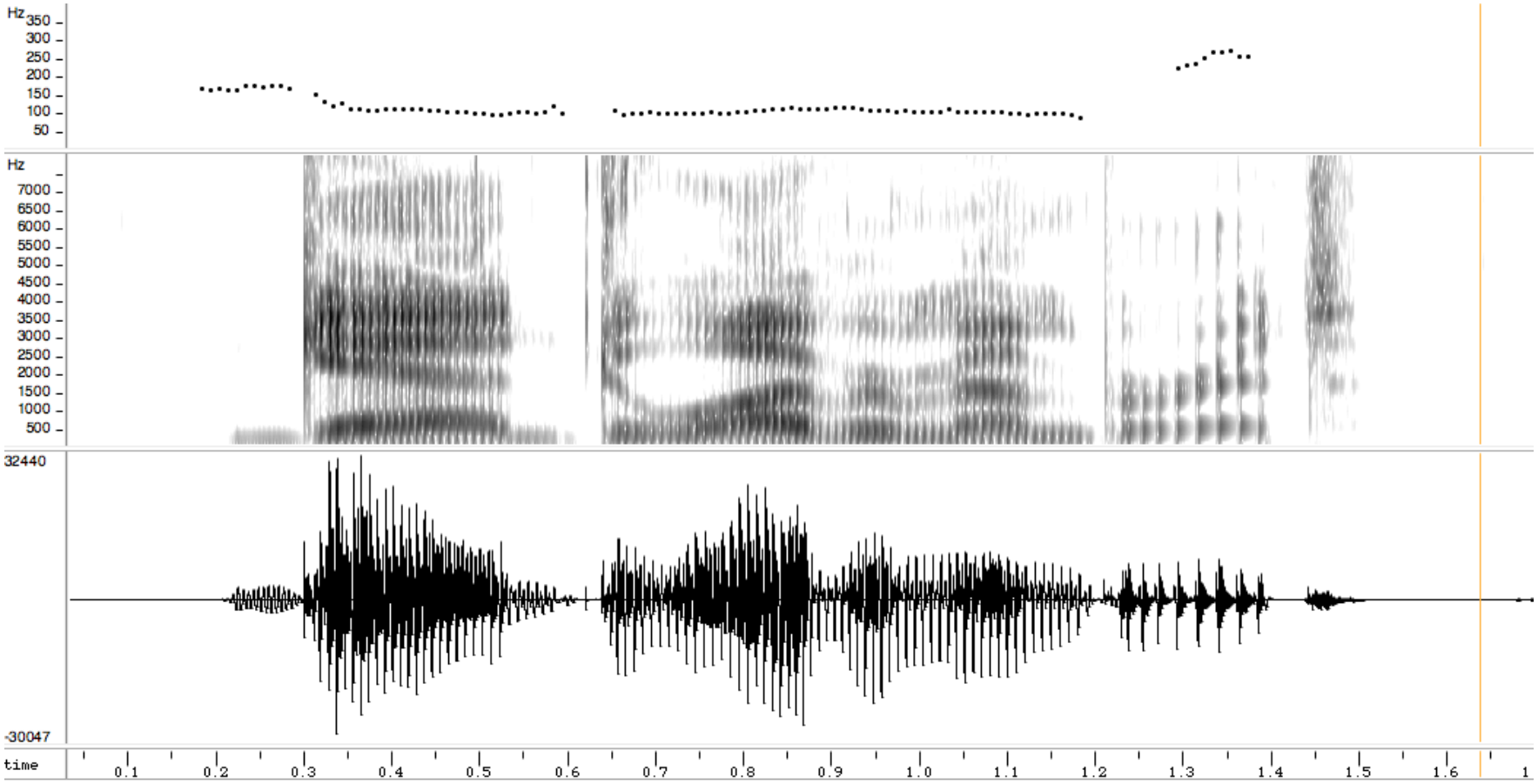


Breathy Voice

- Tension low
  - Minimal adductive tension
  - Weak medial compression
- Medium longitudinal vocal fold tension
- Vocal folds do not come together completely, leading to frication



Creaky voice (Vocal fry)



Examples of Voice Quality Measures

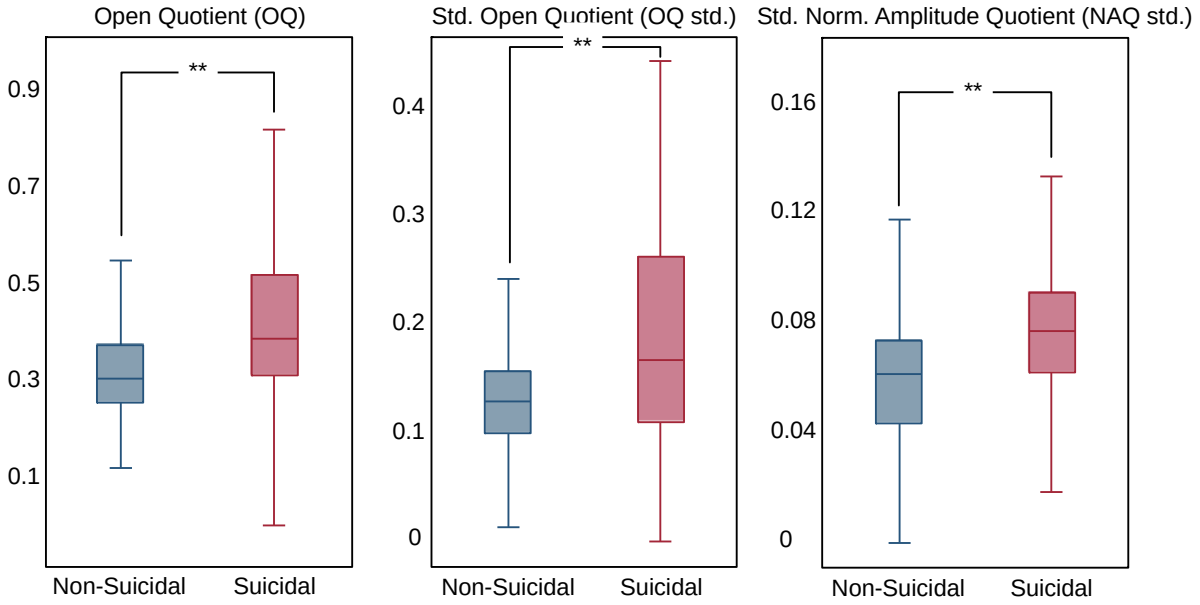
- Open Quotient (OQ)
- Normalized Amplitude Quotient (NAQ)
- Peak Slope

Suicide Prevention [ICASSP 2013]

- Nonverbal indicators of suicidal ideations
- Dataset: 30 suicidal adolescents/30 non-suicidal adolescents
- Suicidal teenagers use **more breathy tones**

Source: CDC

	Age Groups
Rank	11-17
1	Unintentional Injury 3,924
2	Homicide 1,101
3	Suicide 823
4	Malignant Neoplasms 745
	Heart Disease



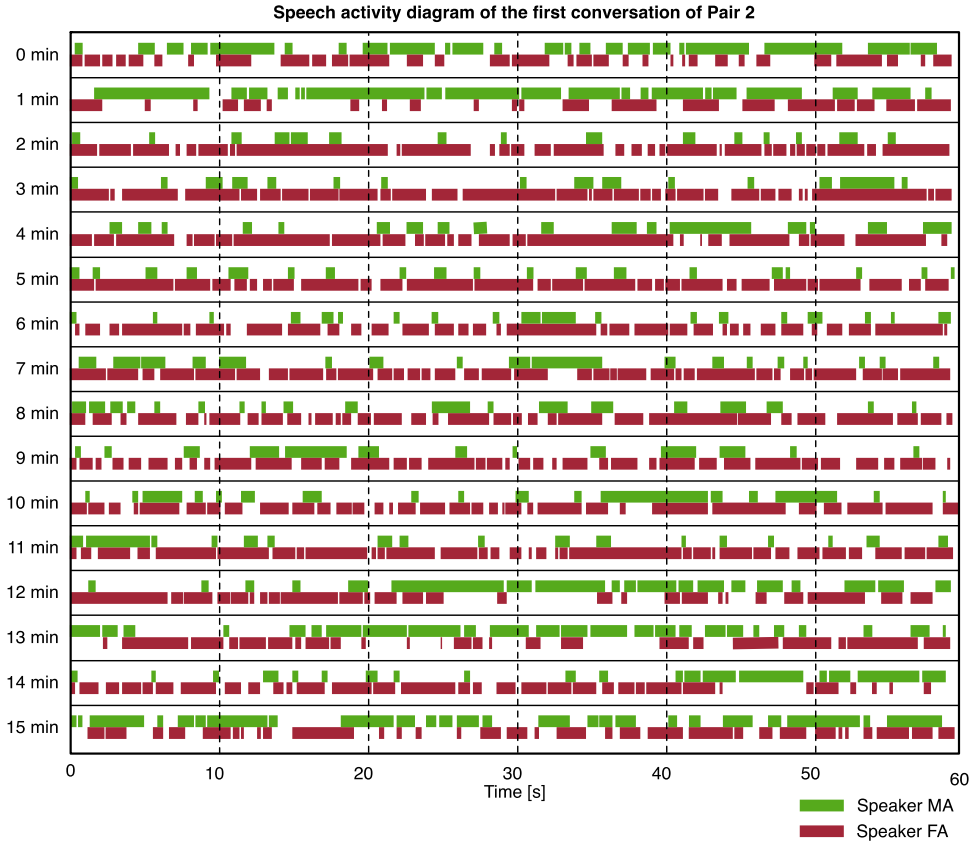
Nonverbal Vocal Expressions

Nonverbal vocal expressions

- Nonverbal vocal expressions are paralinguistic utterances
  - Backchannel (e.g. uhu, hm, yeh)
  - Laughter
  - Moans/Sighs
  - Pause fillers (e.g. um, uh)
- Varying functions and possible interpretations
- **Distinct prosodic and vocal characteristics**
- Often accompanied with distinct facial expressions (e.g. laughing)

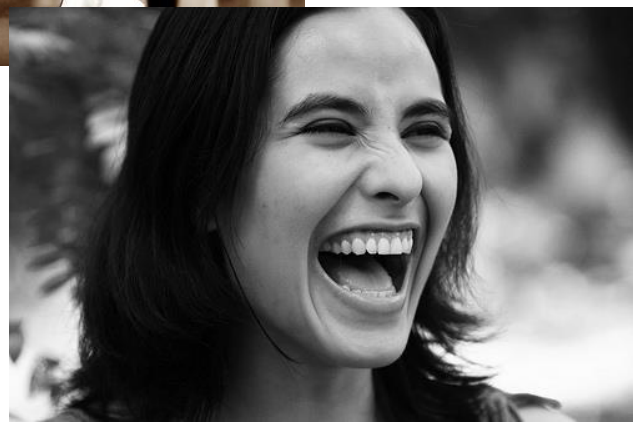
Backchannels

- Backchannels are important fragments of speech
- Functions:
  - Turn-taking management
  - Signals agreement and attention
  - …



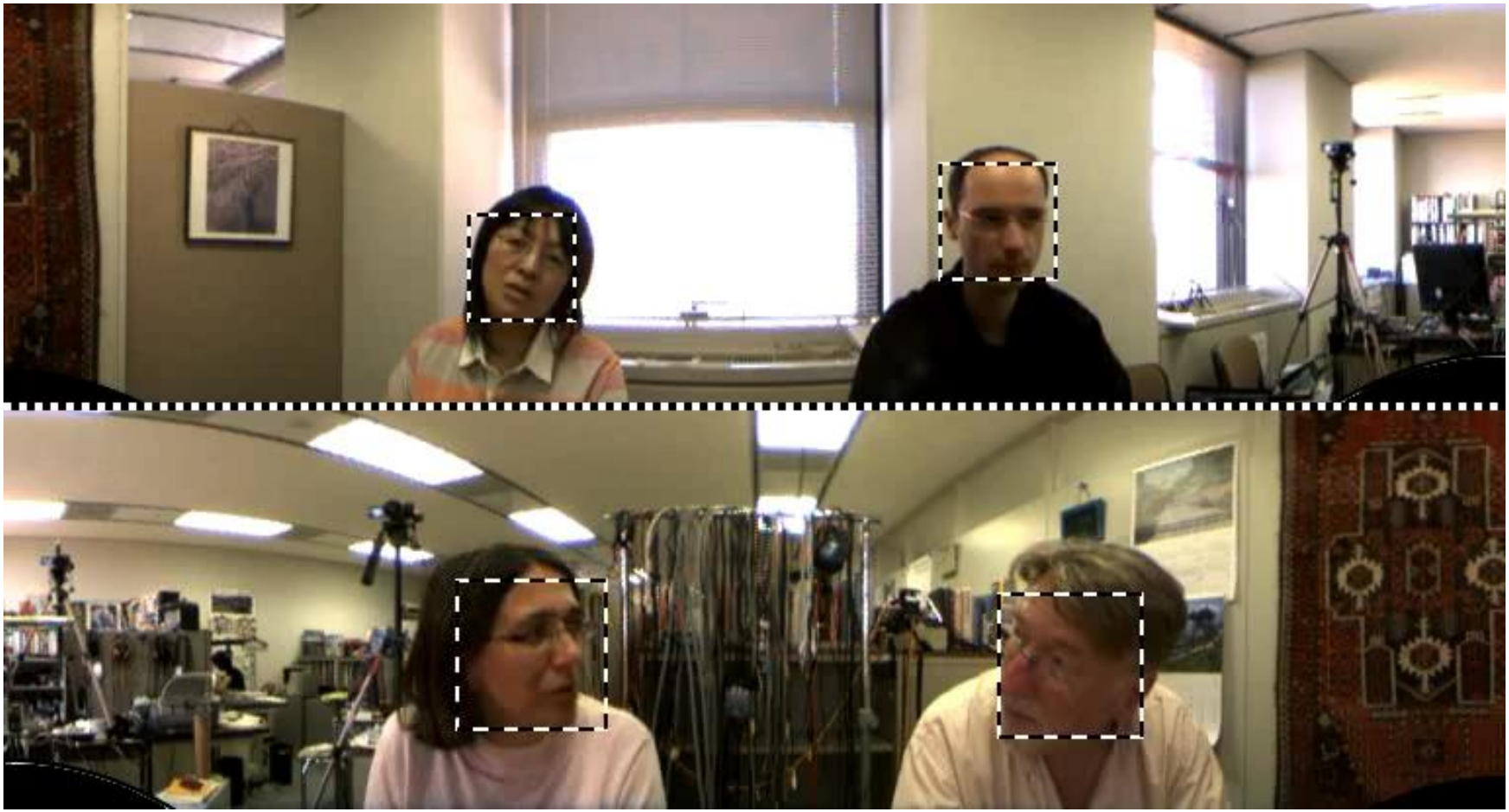
Laughter

- Important social communicational expression
- Understood by all cultures
- Varying meanings:
  - Humorous laughter
  - Signals agreement
  - Uncertainty (e.g. social laughter, nervous laughter)
- **Multimodal!**

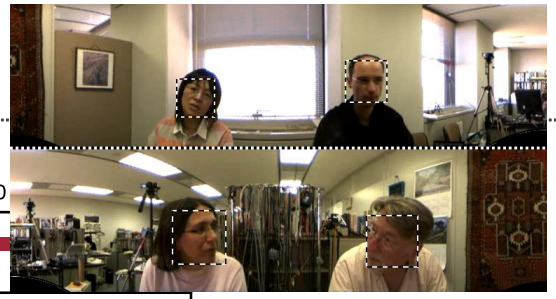


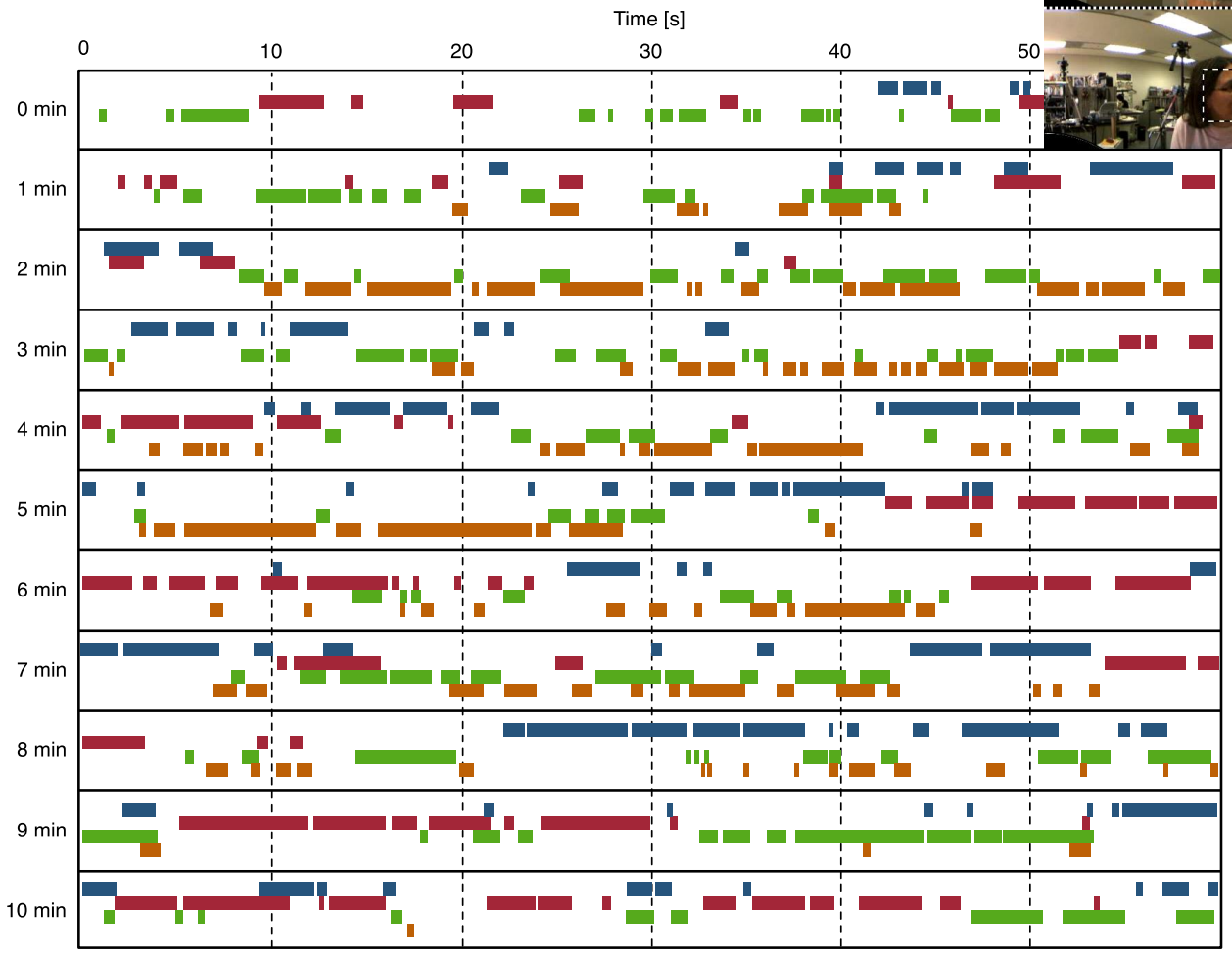
"Laughter […] is performed almost exclusively during social encounters; solitary laughter seldom occurs except in response to media, a source of vicarious social stimulation." – Provine and Yong

Laughter



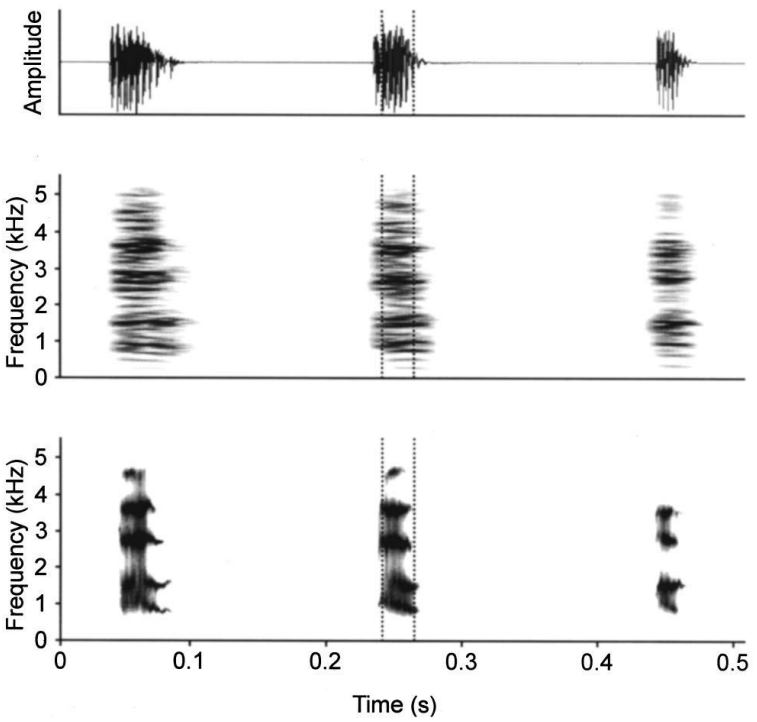
Laughter




Time [s]
0
10
20
30
40
50
0 min
1 min
2 min
3 min
4 min
5 min
6 min
7 min
8 min
9 min
10 min

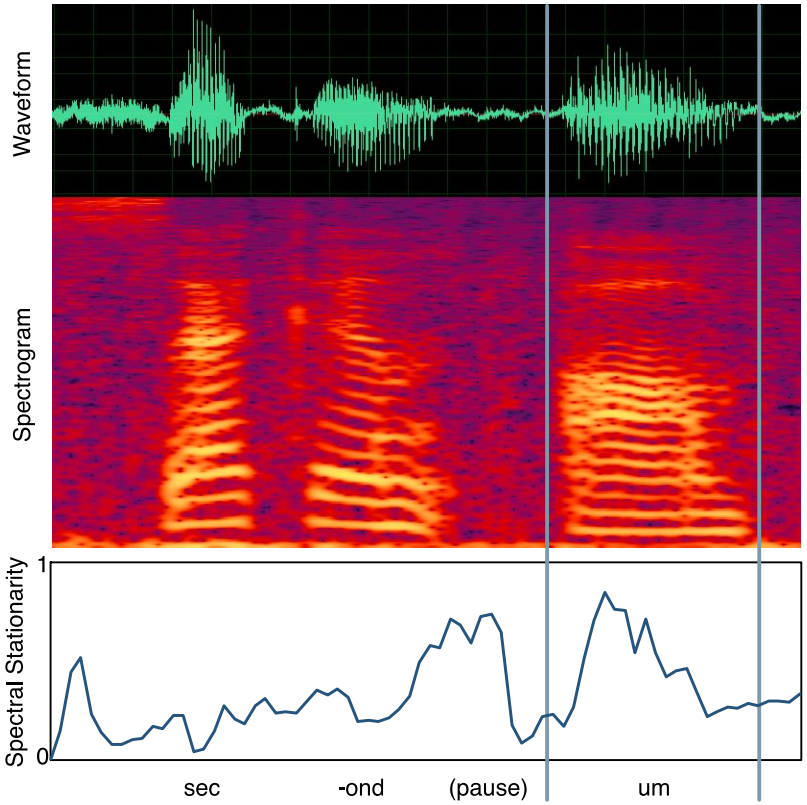
Laughter

- Acoustically very distinct but variable
  - Snort-like
  - Inhaled
  - Exhaled
- Often arranged in *bouts* consisting of single *calls*
  - Often about 200ms in length, but quite variable in length



Hesitations/Pause fillers

- Hesitations (e.g. pause fillers: um, uh)
- Multiple functions
  - Signal anxiety or nervousness
  - Proficiency in speaking
  - …
- Characterized by prolonged vowels and static spectrum



Practical Tools for Speech Processing

Useful Software

COVAREP - A Cooperative Voice Analysis Repository for Speech Technologies (Matlab and Octave)
https://github.com/covarep/covarep/

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S., COVAREP - A collaborative voice analysis repository for speech technologies, under review at International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Useful Software

openSMILE - Speech & Music Interpretation by Large Space Extraction

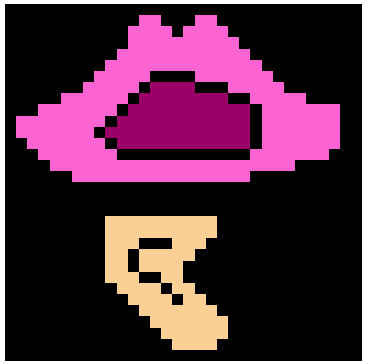
- openSMILE is a fast, real-time (audio) feature extraction utility (C++)

http://opensmile.sourceforge.net/

TECHNISCHE UNIVERSITÄT MÜNCHEN

Useful Software

- Praat
- http://www.fon.hum.uva.nl/praat/
- Useful for feature extraction/visualization and more!



Upcoming Conferences

- Affective Computing & Intelligent Interaction (ACII)
  - http://acii-conf.org/2019/
  - Deadline: April 12, 2019
- International Conference on Multimodal Interaction (ICMI)
  - https://icmi.acm.org/2019/
  - Deadline: May 7th, 2019
- Affective Computing Pre-Conference at ISRE (International Society of Research on Emotion)
  - https://www.isre2019.org/program/pre-conferences/affective-computing
  - Deadline: May 7th, 2019
- Empirical Methods in Natural Language Processing
  - https://www.emnlp-ijcnlp2019.org/
  - Deadline: May 21th, 2019