CS4347

Sound and Music Computing

L8: Singing voice processing and transcription

Wang Ye

www.comp.nus.edu.sg/~wangye

wangye@comp.nus.edu.sg

Office: AS6-04-08

Topics to Cover (*selective approach*)

Part A: The Core

- Introduction
- Review of DFT, Audio Representation, and Machine Learning
- Music Representation, Analysis and Transcription
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)
- Generative Models for Text-to-Speech (TTS) & Singing Voice Synthesis (SVS)

Midterm break

Part B: The Breadth

- Spoken Language Assessment
- Singing Voice Processing
- Nonnegative Autoencoders with Applications to Music Audio Decomposing
- Automatic Music Generation
- Music Production & Audio Effects
- Synthesis of Sound & Music – a DSP Approach
- Project presentations/demo/concert

Topics Today

⇨ Part A: Physiology & Physics of the Singing Voice

Part B: Singing voice transcription/evaluation

Part C: Singing voice synthesis/generation

Why Singing is Interesting

- All popular music cultures around the world use singing
- The singing voice is the most expressive of all musical instruments
- **“Of all musical instruments the human voice is the most worthy because it produces both sound and words, while the others are of use only for sound" (Summa Musice, 13th century)**
- Our representations (e.g. MIDI, Western notation) are inadequate for expressive singing
- Knowledge about singing from other disciplines (e.g. physiology, psychology, pedagogy) is rarely exploited in MIR
- Many MIR tasks involving singing have never been attempted

The Theme of CS4347/CS5647

- An introduction of the exciting world of singing styles, **the mechanisms of the singing voice** and provide a guide to **representations**, engineering **tools and methods for analyzing and leveraging it**.

- **Sparking a passion for singing and ideas** of how to use our knowledge of singing, and singing information processing, to create new, exciting research and applications.

- **Bridging speech and music technologies**

Singing Styles

- The voice is a versatile instrument
- It is universal: everyone has one, can use it, and it is suitable for music of all cultures
- It is portable, affordable and expressive
- Use cases: entertainment, art, worship, communication, social, and language learning & speech rehab (SLIONS)
- We observe a great diversity of styles of singing*
- Aesthetics (taste, appreciation of beauty) vary by style, and sometimes within styles

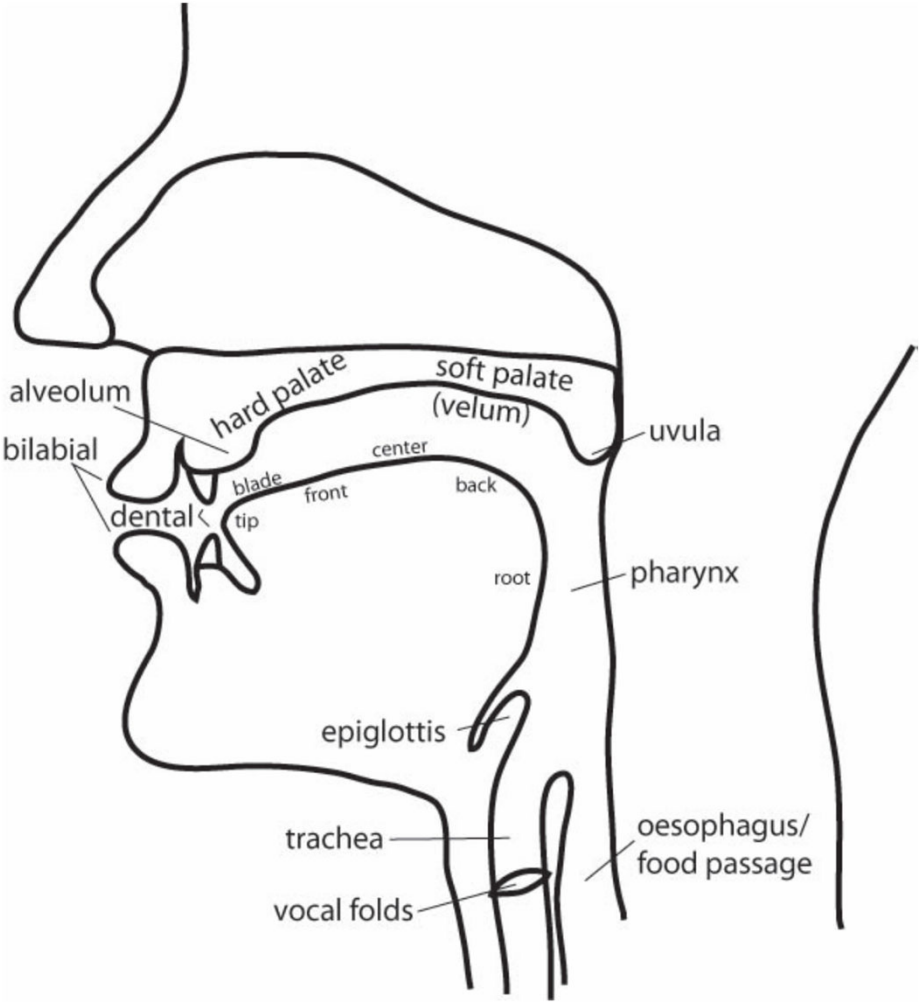
*J. Potter, ed. (2000). The Cambridge Companion to Singing. Cambridge, UK: Cambridge University Press.

Aesthetics: Natural or Artificial?

- Natural
  - Authenticity of expression (e.g. rock, pop, folk styles)
  - Speech-like quality (e.g. Broadway), directness
  - Clarity of lyrics: rap (lyrics foremost) vs opera (intelligibility sacrificed for volume)
- Artificial
  - Purity of tone, effortless (e.g. Western classical: “objectifying control”)
  - Training, discipline (\high" vs \low" culture)
  - Technical prowess (e.g. classical, jazz)
  - Performance, acting (e.g. rock, opera, musicals)
  - Microphone technique
  - **Audio effects**

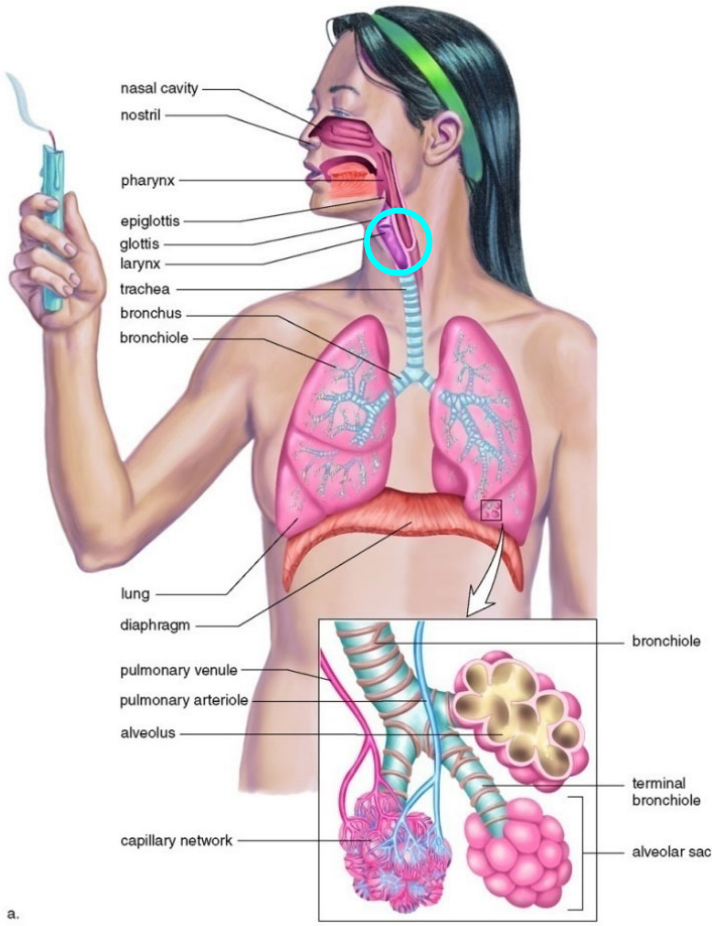
How Does the Human Voice Work?

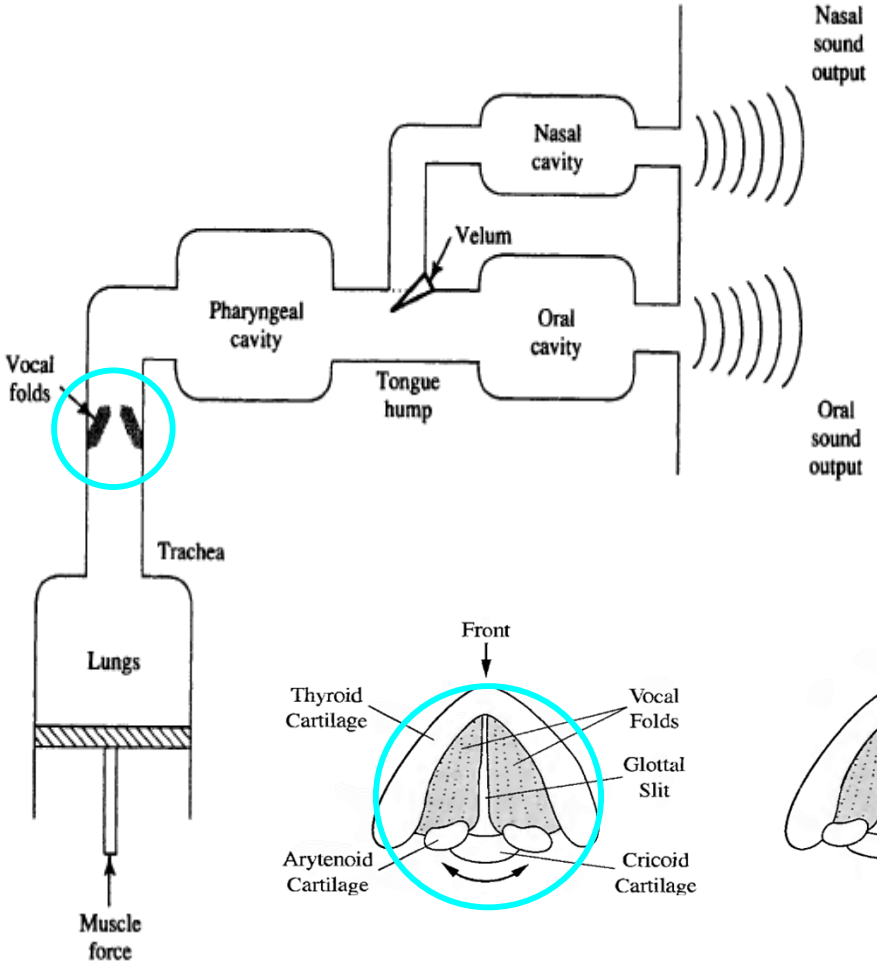
- **Respiratory system**:
  - compresses lungs to create airflow
- **Vocal folds**:
  - Chop airstream into a periodic pulsation
- **Vocal tract**:
  - Filters source waveform according to resonances (formants)*

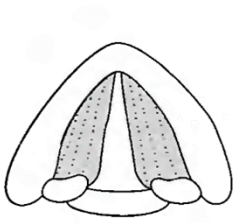


*J. Sundberg (1987). The Science of the Singing Voice. DeKalb IL: Northern Illinois University Press.

Physiology of the Human Voice

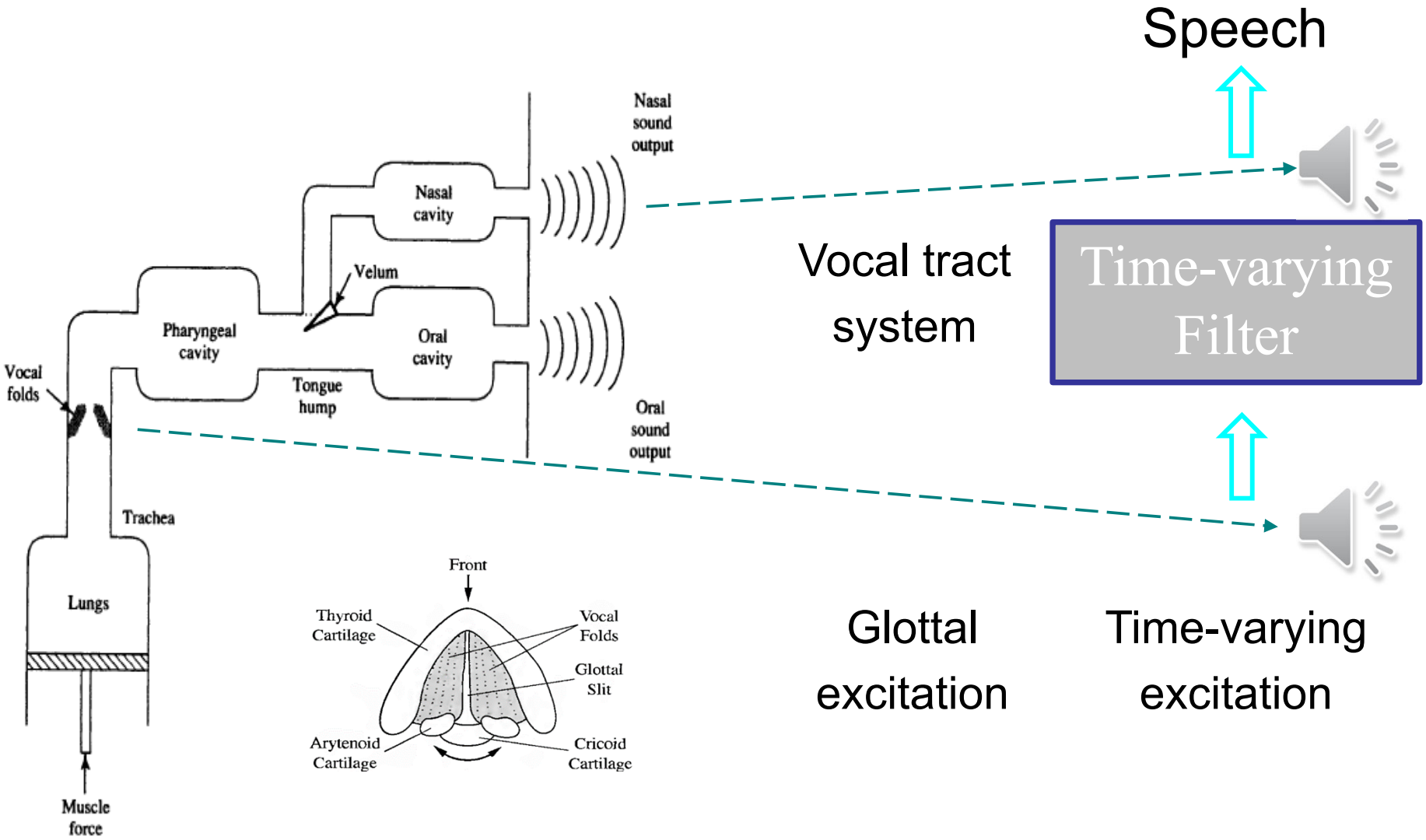






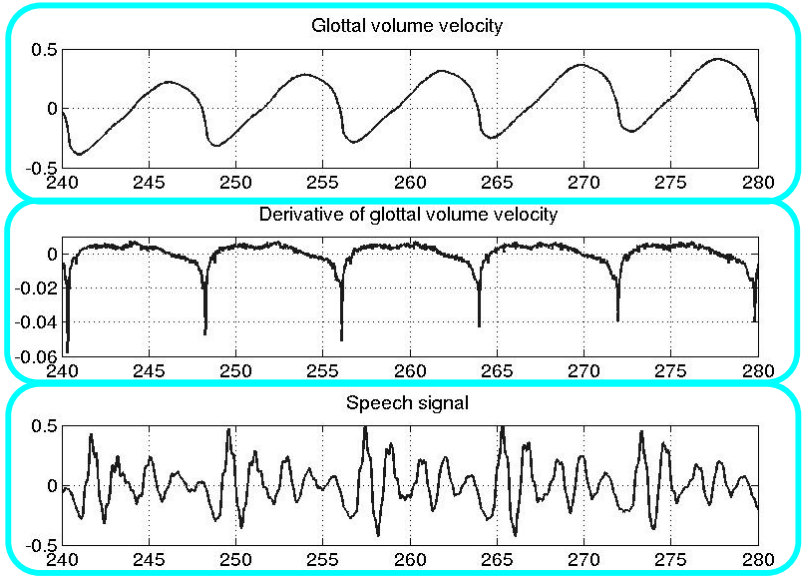
Modeling of the Human Voice

Source-filter model

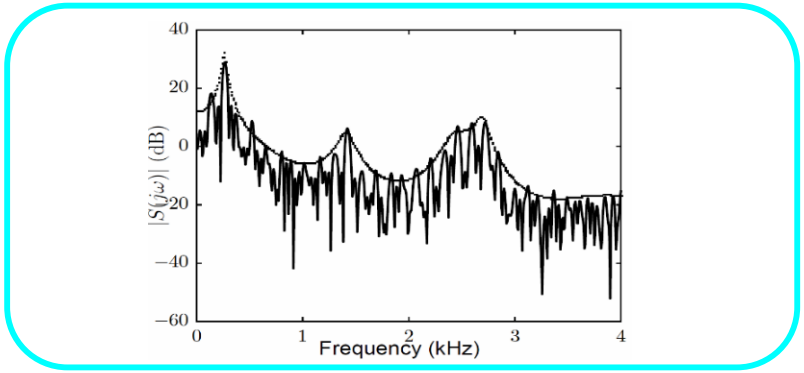


From Excitation to Sound

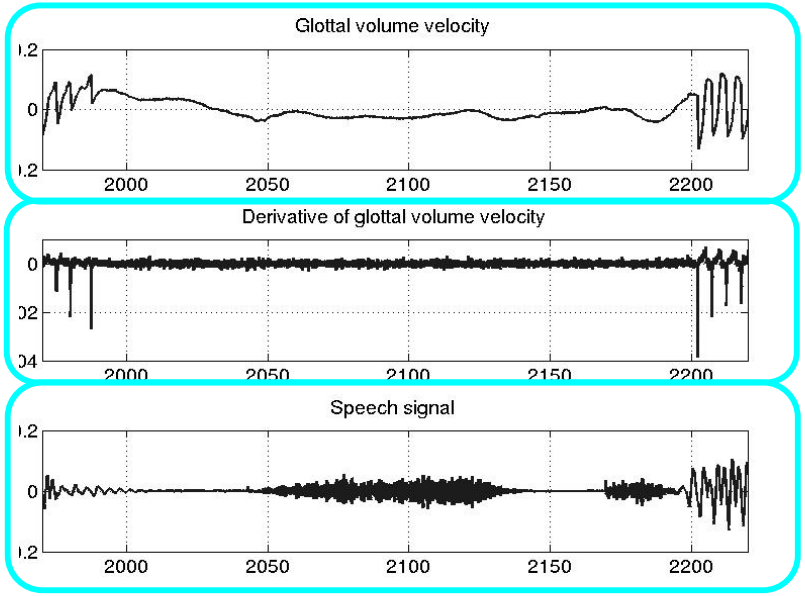
Voiced sound



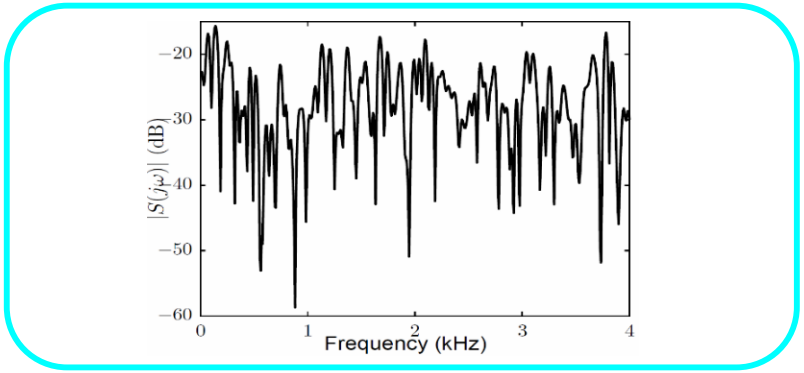
DFT



Unvoiced sound

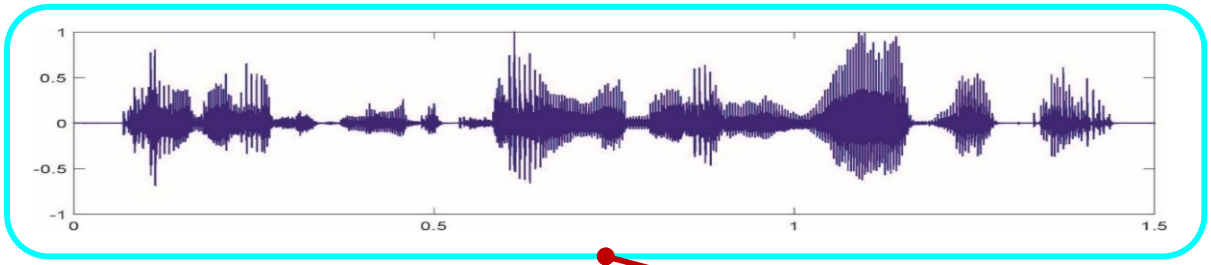


DFT

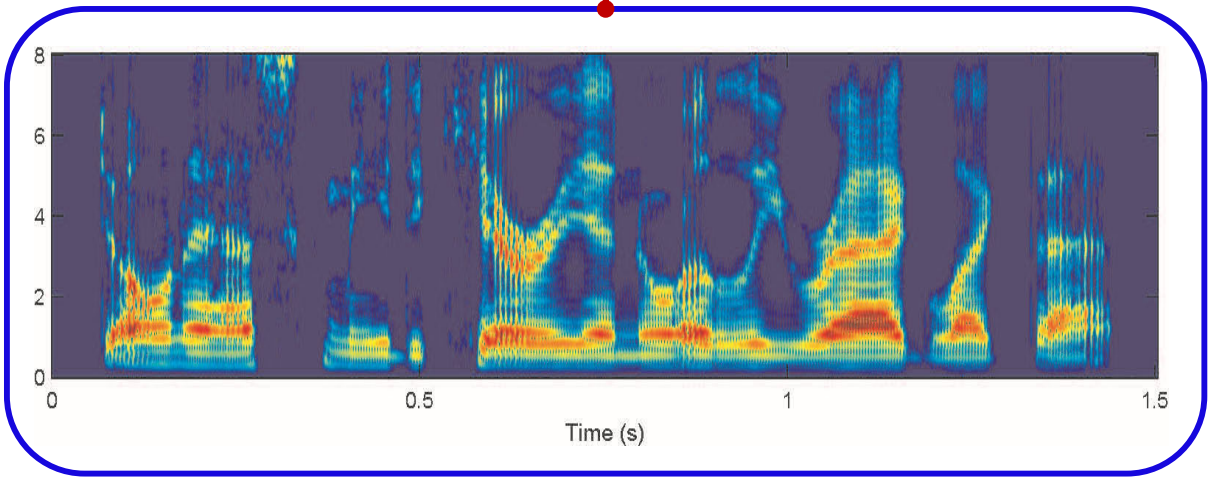


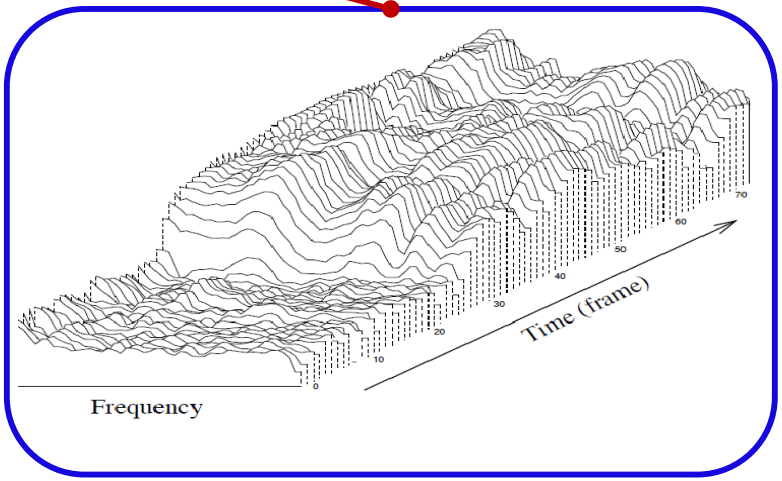
Speech Representation for Analysis & Synthesis

From Time Serie Data to Spectrogram



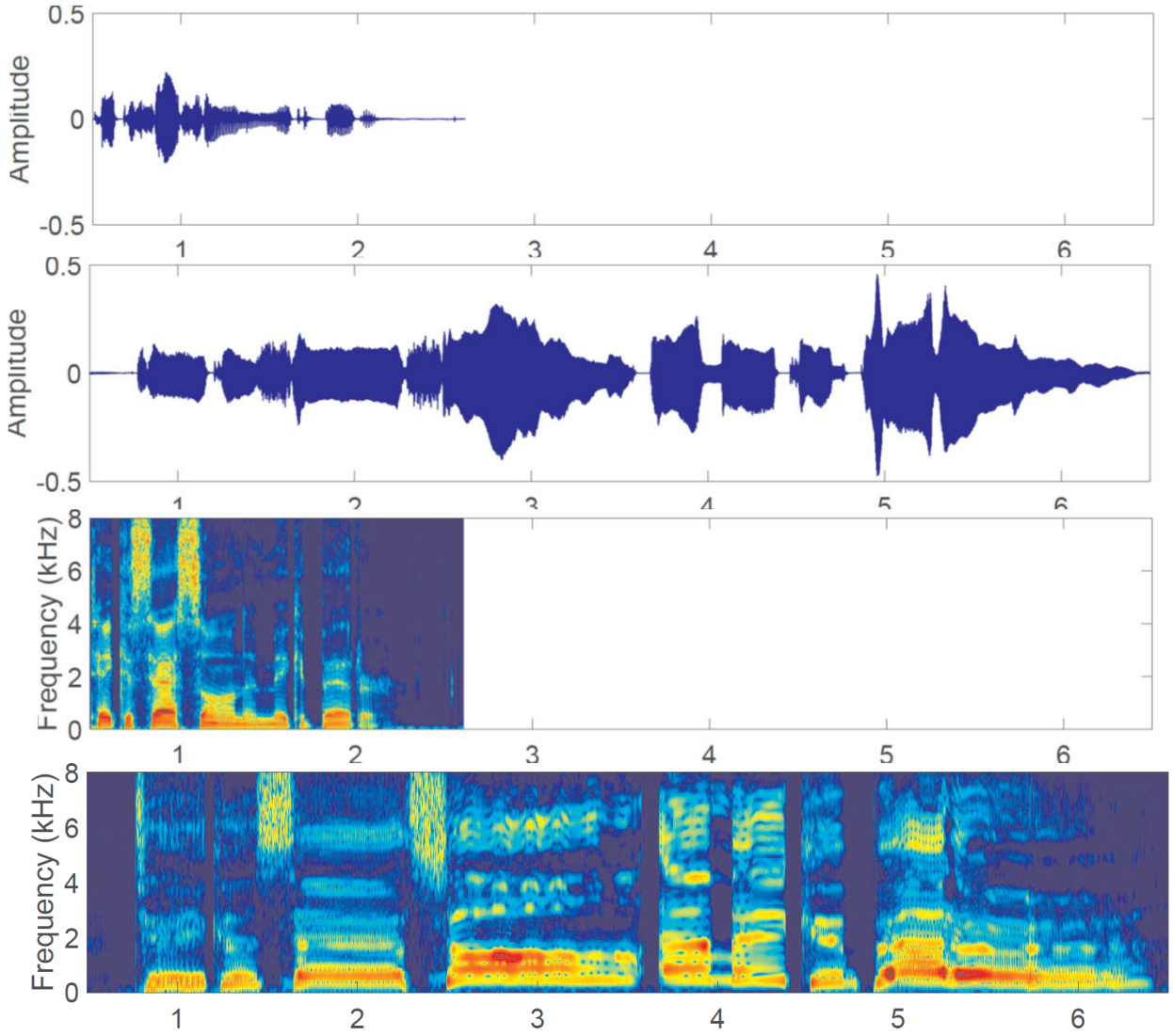
DFT

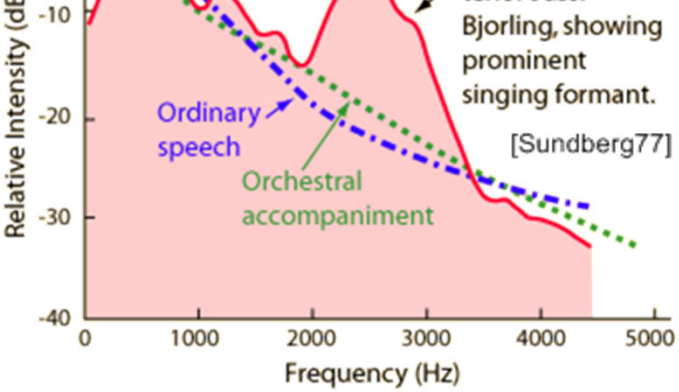




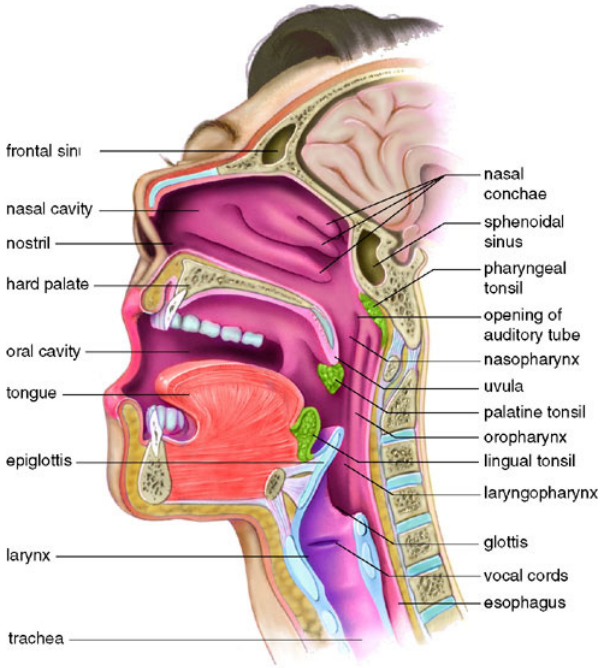
Why is spectrogram a better speech/singing representation?

From Speech to Singing Voice





What kind of music instrument does the vocal track resemble?

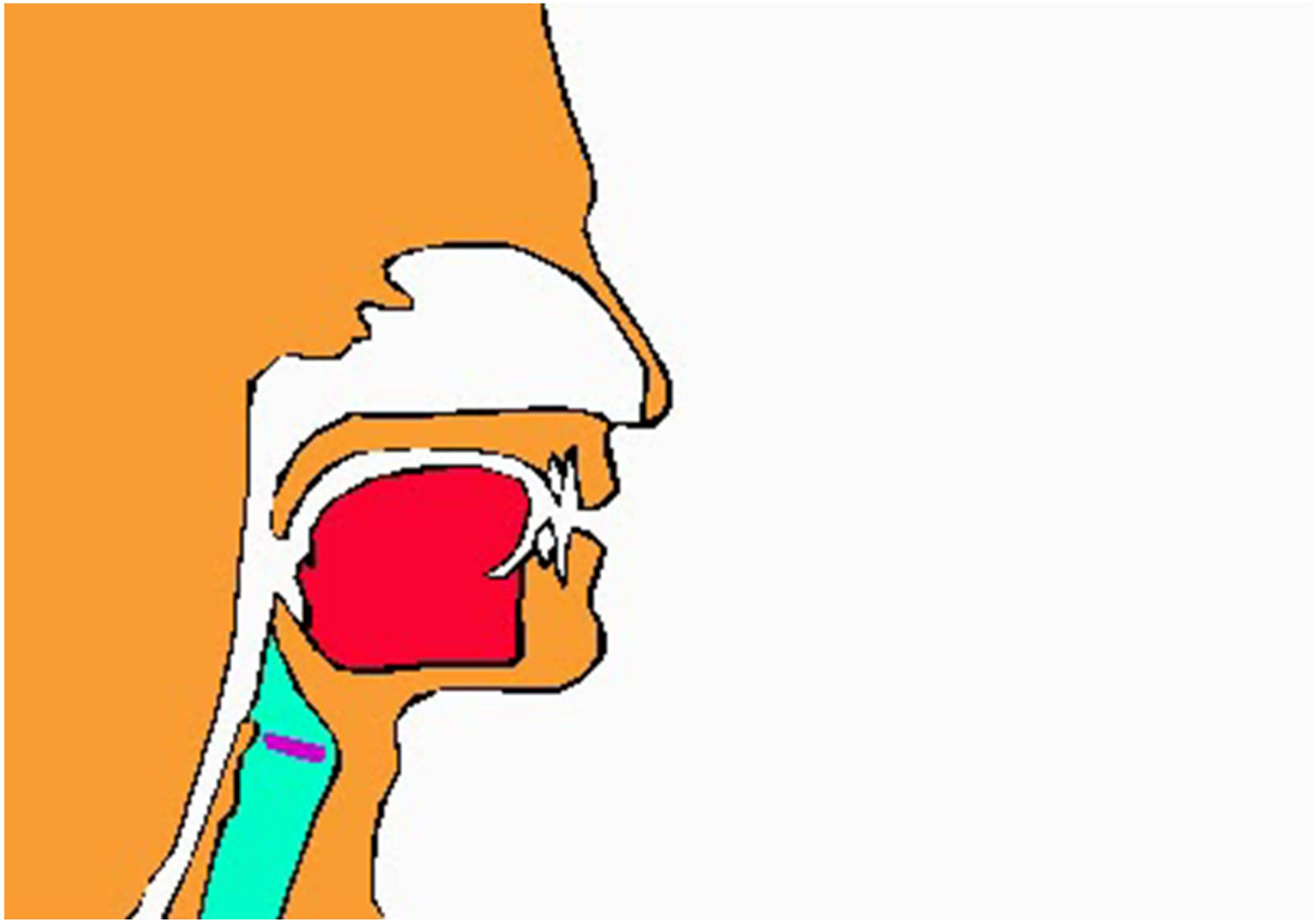


What makes the human vocal track so special/*vulnerable*? How to use our voice correctly when speaking or singing?

Wang, Y., Wei, W., and Wang, Y., (2023, June). Phonation Mode Detection in Singing: A Singer Adapted Model, ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)

Wang, Y., Wei W., Gu, X., Guan, X., and Wang, Y., (2023). Disentangled Adversarial Domain Adaptation for Phonation Mode Detection in Singing and Speech, IEEE/ACM Transactions on Audio, Speech, and Language Processing (DOI: 10.1109/TASLP.2023.3317568)

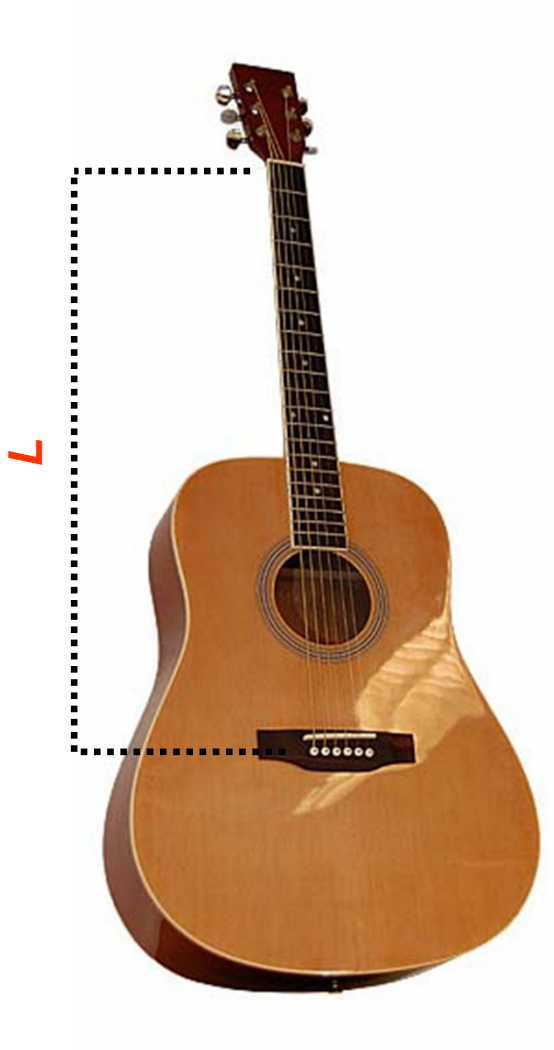
Mechanism of voice production

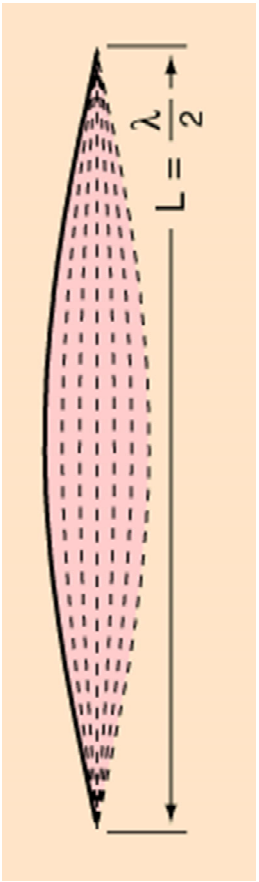


Breathing patterns in *whispering, yelling*, speaking and singing: the power of visualization

Fundamental Frequency & Pitch revisited

When plucked, the string fixed at both ends vibrates at a rate directly proportional to its length.





Wavelength λ is twice the length of the string.

The **fundamental frequency** is a **physical attribute**.

Pitch is a **perceptual attribute**.
Pitch is the perceived fundamental frequency of sounds that allows their ordering on a frequency-related scale.

$$f_1 = \frac{\sqrt{T / m/L}}{2L}$$

T = string tension
m = string mass
L = string length

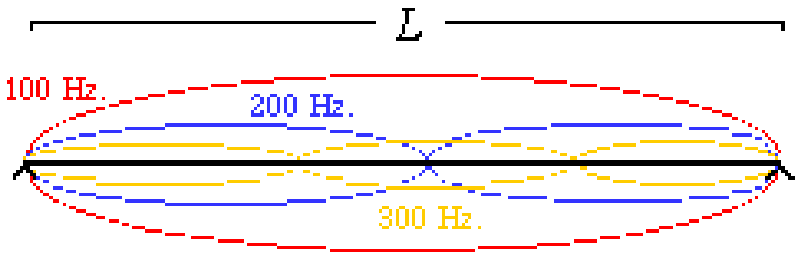
Physics revisited!

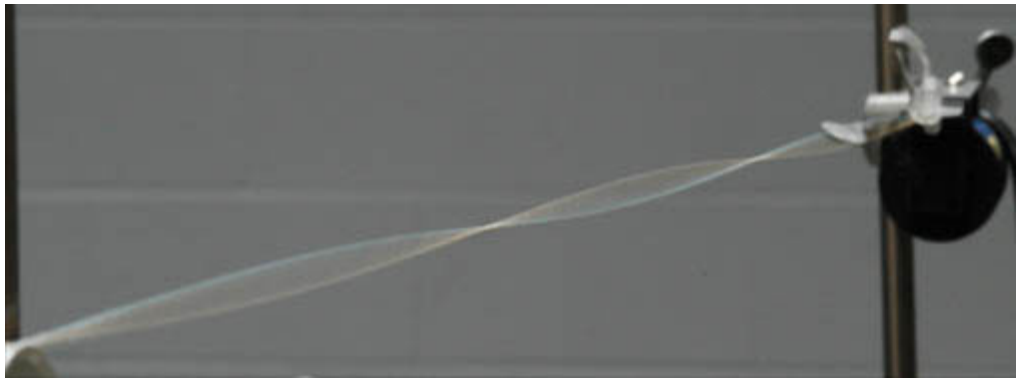
Fundamental Frequency & Harmonics (1)

- In addition to vibrating over its entire length, a string simultaneously vibrates over fractional divisions of its length (1/2, 1/2, 1/3, 1/4, 1/5, 1/6, etc).

- This produces a series of "HARMONICS" whose frequencies are inversely proportional (2x, 3x, 4x, 5x, 6x, etc., where x is the fundamental frequency of the string) to those fractional divisions.

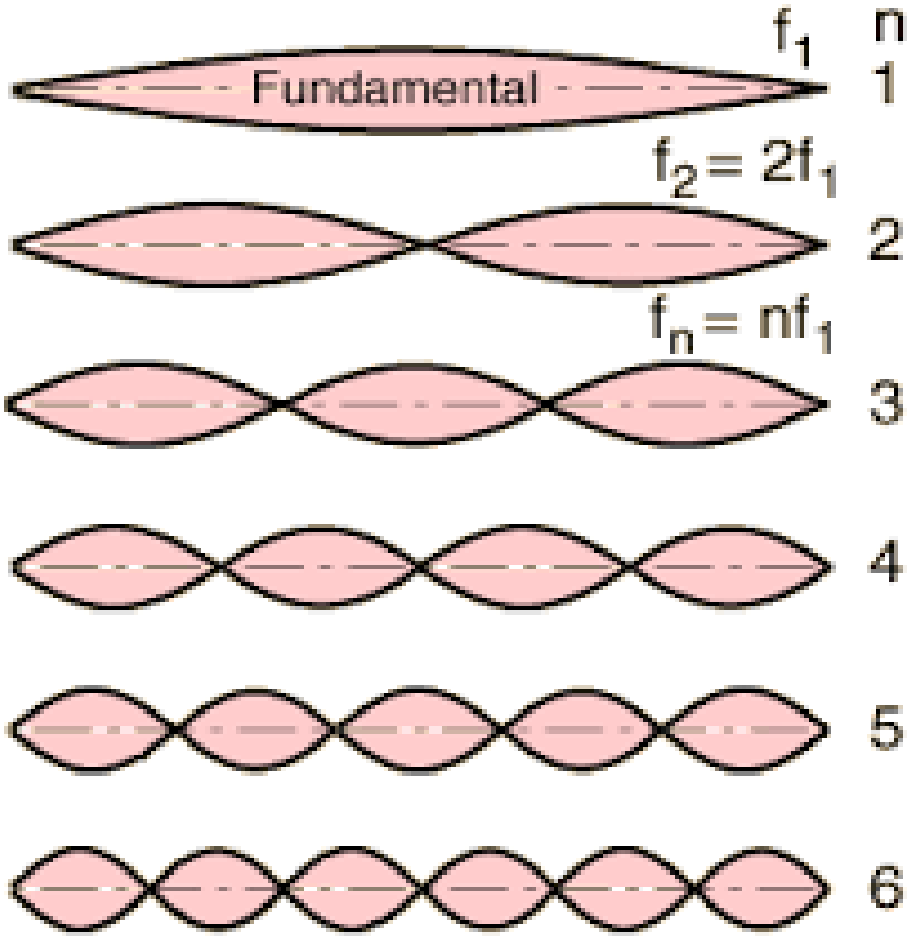
- F0 = 100 Hz



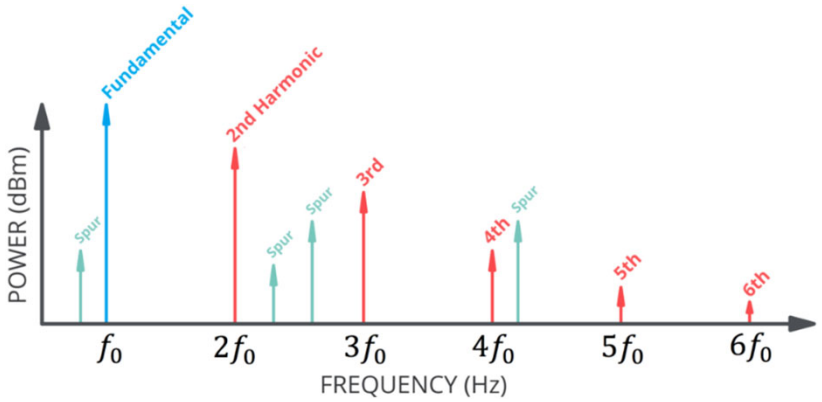


Physics revisited!

Fundamental Frequency & Harmonics (2)



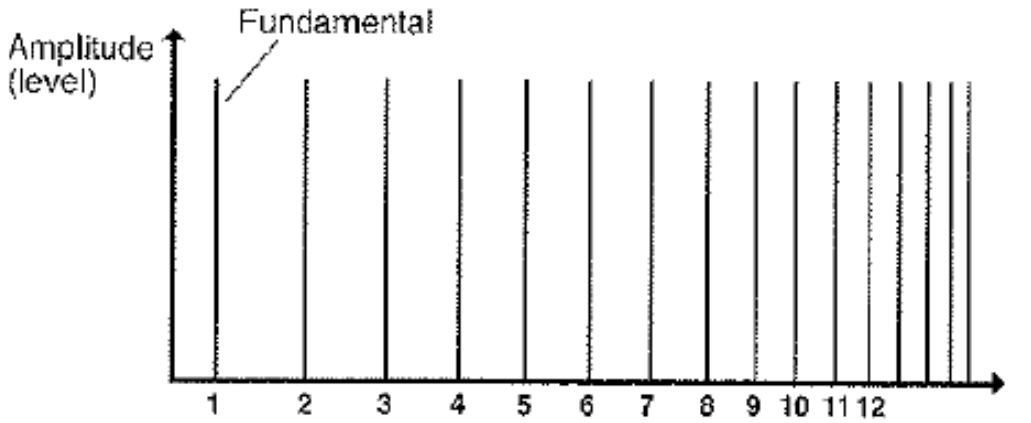
The fundamental frequency, often referred to simply as the fundamental, is defined as the lowest frequency of a periodic waveform.



Physics revisited!

Fourier Transform in Our Ear!

When a musical note arrives in our ear, what we hear is not just the fundamental frequency (e.g., 110 Hz for note A). Our ear/brain system tends to fuse harmonically-related frequency components into a single sensation we call PITCH!



Harmonic series for the note A 110Hz as far as the tenth harmonic. Note that all of the harmonics are shown as being of equal amplitude. This would not be the case for naturally generated sounds

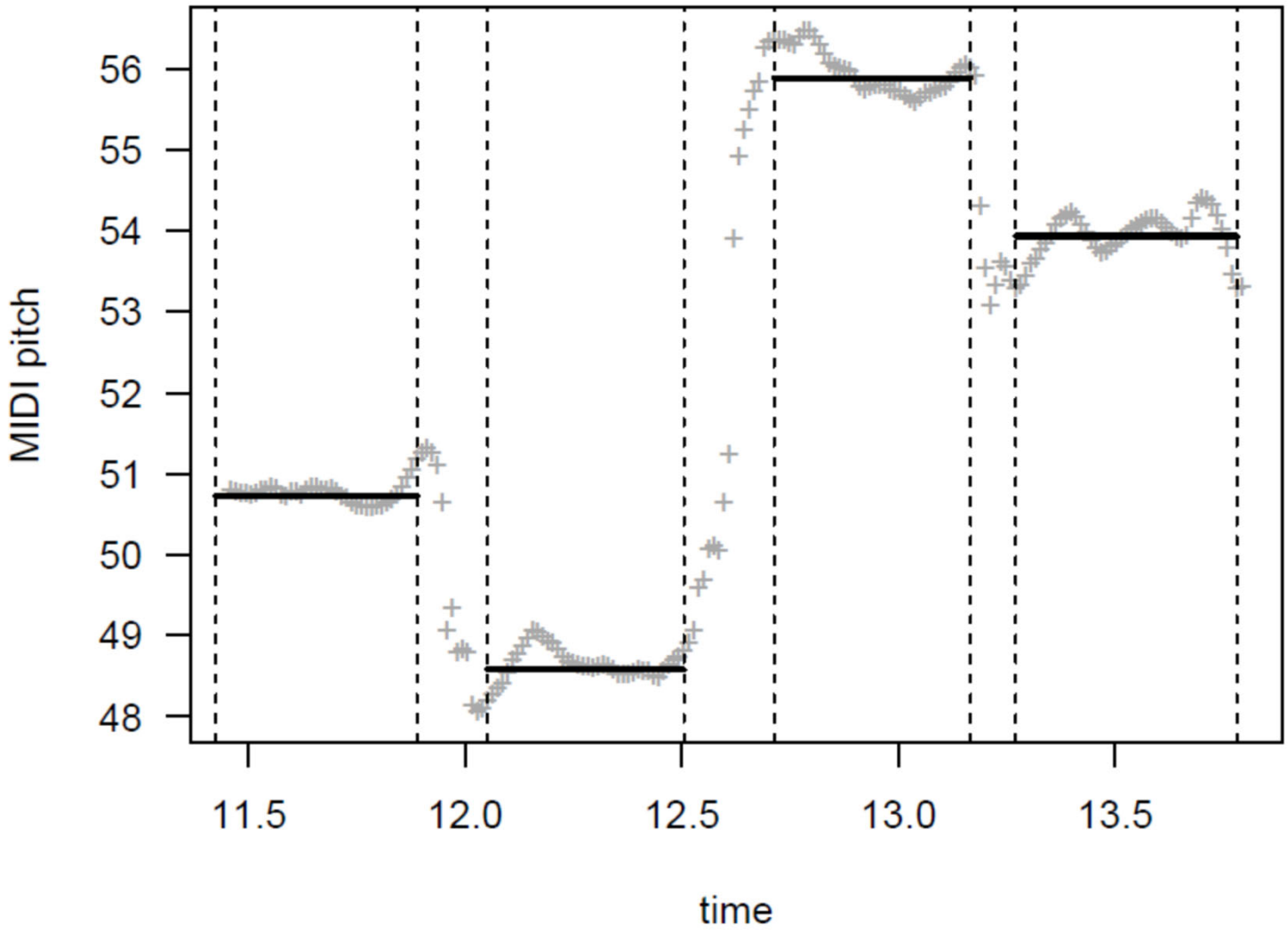
Harmonic	Frequency
Fundamental (A)	110Hz
Second	220Hz
Third	330Hz
Fourth	440Hz
Fifth	550Hz
Sixth	660Hz
Seventh	770Hz
Eighth	880Hz
Ninth	990Hz
Tenth	1,100Hz

Each harmonic in the series is a whole-number multiple of the fundamental

Pitch of Singing Voice

- Singer's pitch range is determined by length and mass of vocal folds
- Classical voices
  - Soprano: 260-1050 Hz (C4-C6)
  - Alto: 175-700 Hz (F3-F5)
  - Tenor: 130-520 Hz (C3-C5)
  - Bass: 80-330 Hz (E2-E4)
- Vibrato (classical)
  - Pitch modulation via pulsations in cricothyroid muscle
  - Rate: 5-7 Hz
  - Depth (pitch variation): ±0.5-1.5 semitones
- Vibrato (pop)
  - Amplitude modulation via variations in subglottal pressure

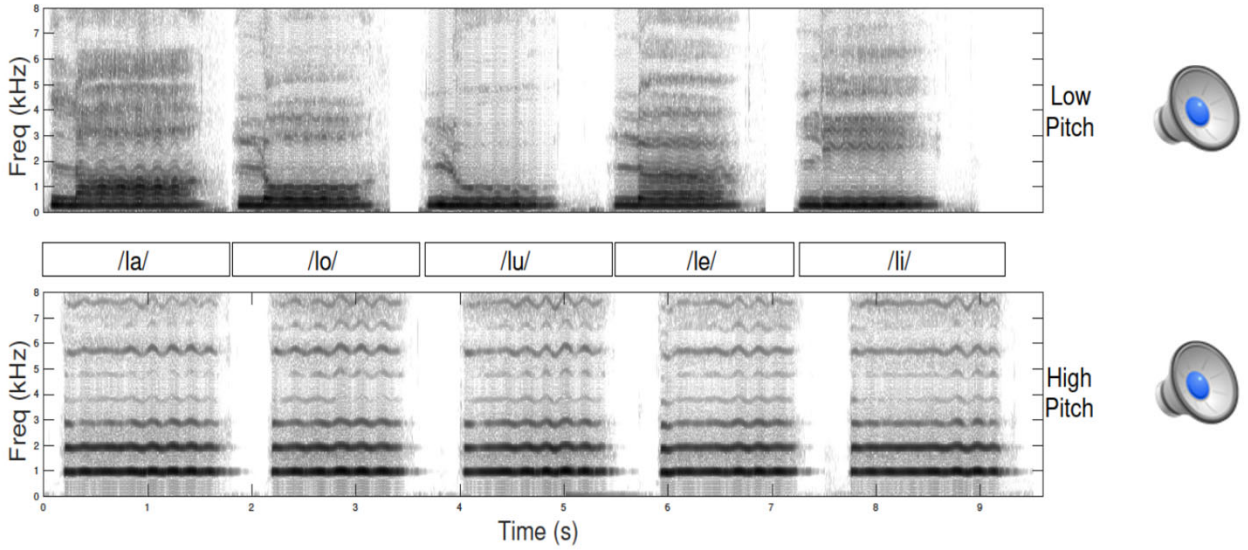
Note Segmentation & Framewise/Notewise Pitch Estimates



Vocal Tract & Brilliance of Singing Voice

- Resonances occur in the vocal tract according to its configuration
- Up to 5 formants are relevant for singing
- Vowel quality: mainly determined by first 2 formants
- Voice quality: determined by individual factors (size, shape, etc)
- Singer's formant
  - strong peak in spectral envelope of classical singers
  - clustering of the 3rd, 4th and 5th formants
  - bass (2.2 kHz), tenor (2.9 kHz), alto (3-3.5 kHz)
  - contributes to brilliance of sound and audibility over an orchestra without excessive effort

Intelligibility of Singing Voice



Vibrato is an important attribute of professional classic singing.

Can you think of a method for automatic vibrato detection?

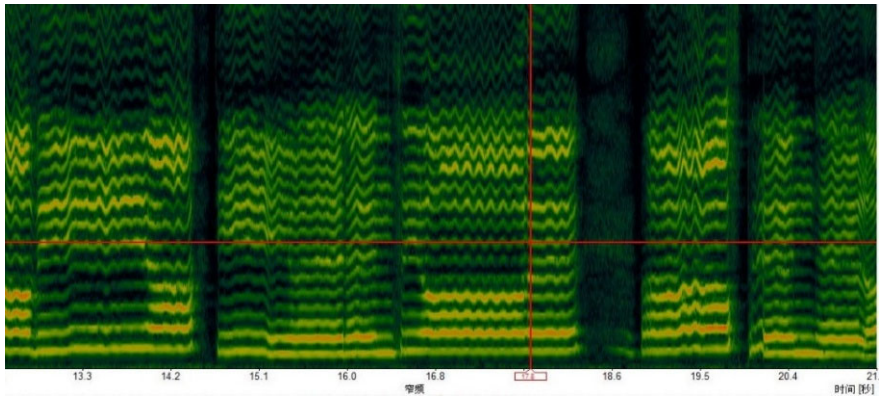
Sharma, B., & Wang, Y. (2019). Automatic Evaluation of Song Intelligibility using Singing Adapted STOI and Vocal-specific Features. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Ibrahim, K. M., Grunberg, D., Agres, K., Gupta, C., & Wang, Y. (2017). Intelligibility of sung lyrics: A pilot study. ISMIR.

Two Examples of Wrong Vibrato

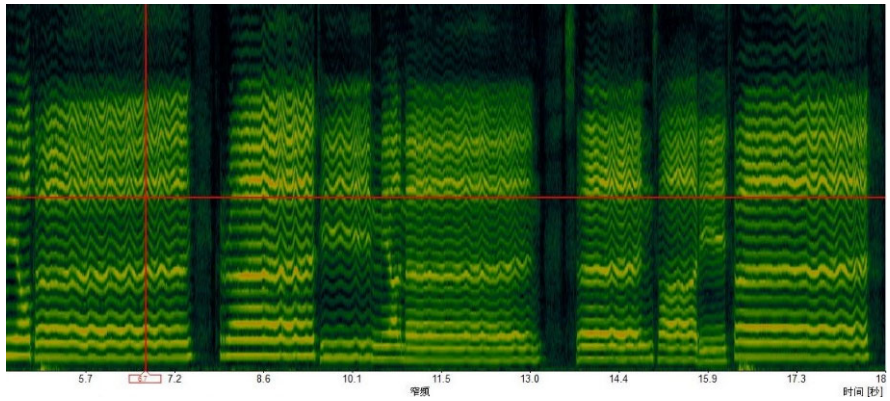
N>7

tremolo



N<4

wobble



How to generate a spectrogram which can visualize vibrato?

Which singer is better? Why?

Which song do you prefer? Why?

Sample 1

Sample 2

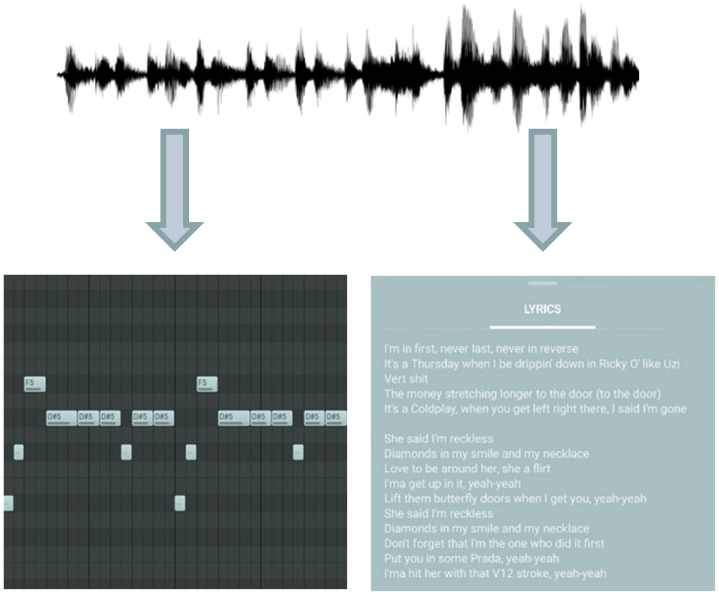
Topics Today

Part A: Physiology & Physics of the Singing Voice

⇨ Part B: Singing voice transcription / evaluation

Part C: Singing voice synthesis / generation

Singing Voice Transcription



R zxmhfg$ Styjx

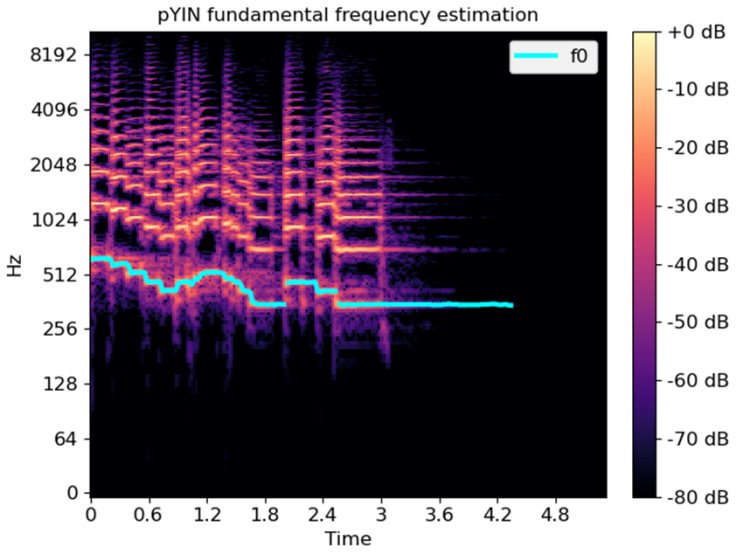
Q~whx

- **Singing Melody Transcription:**
  - F0 Estimation
  - Note-level Melody Transcription
- **Singing Lyrics Transcription:**
  - Lyrics Transcription
  - Lyrics Alignment

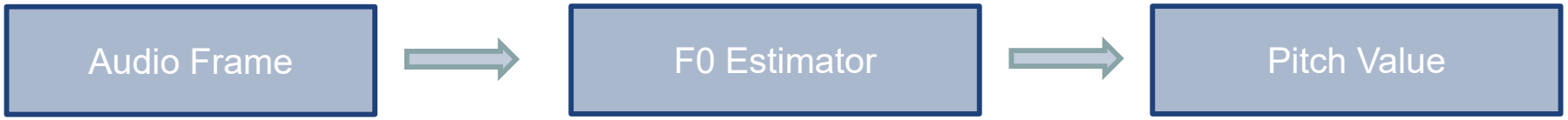
Applications: music and language education, music platforms, karaoke apps, etc.

Singing Melody Transcription (1)

F0 Estimation



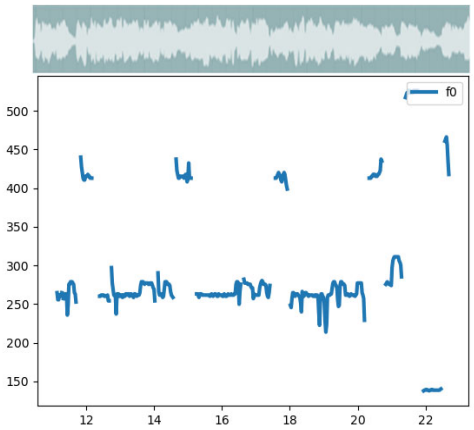
- **F0 estimation:** extract f0 from vocal audio signals
- **Applications:** autotune, singing audio synthesis, etc.
- **Methods:**
  - Digital signal processing models: Yin, DIO, etc.
  - Probabilistic models: pYin, etc.
  - Neural networks: CREPE, Patch-CNN, etc.

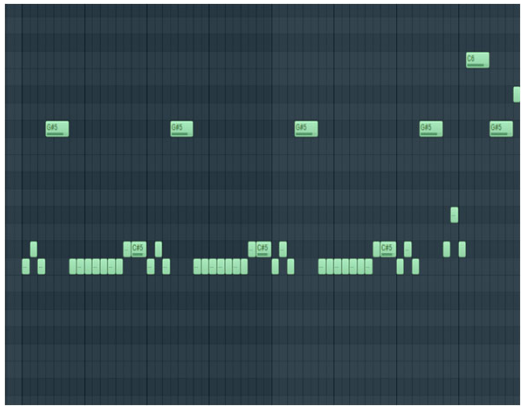


Kim, Jong Wook, et al. "Crepe: A convolutional representation for pitch estimation."
2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Singing Melody Transcription (2)

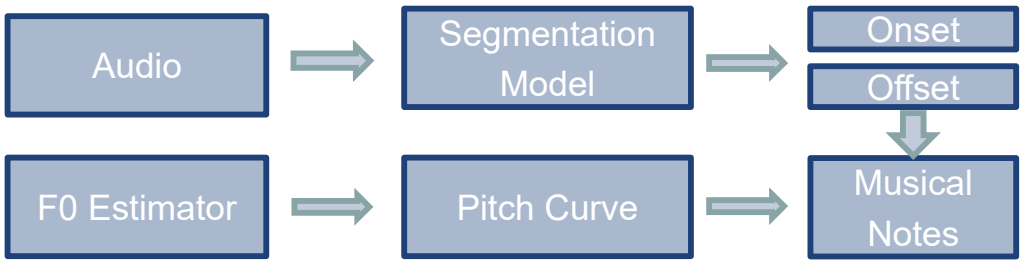
Note-level Melody Transcription





Pitch Contour

Musical Notes

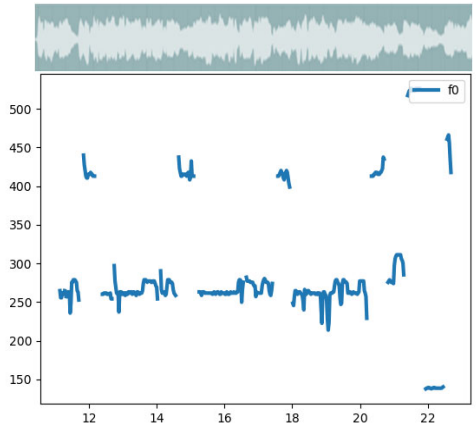


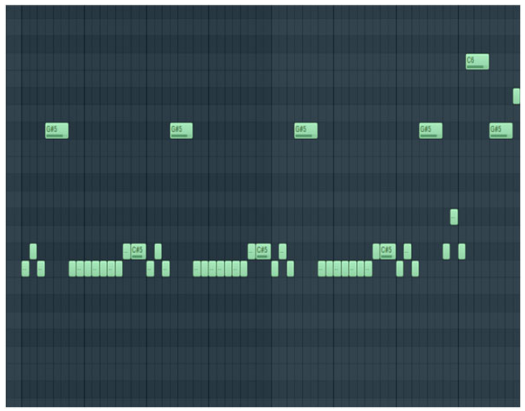
- **Note-level melody transcription:** the automatic extraction of a parametric representation of the singing performance
- **Applications:** singing data collection, pitch correction and time manipulation of recordings
- **Methods:**
  - Probabilistic models: HMMs and GMMs
  - Neural networks: VOCANO

Jui-Yang Hsu and Li Su. “VOCANO: A note transcription framework for singing voice in polyphonic music.” International Society of Music Information Retrieval Conference (ISMIR), 2021.

Singing Melody Transcription (3)

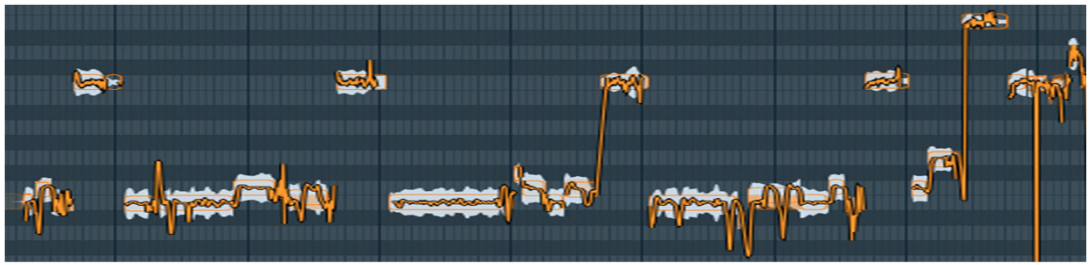
Note-level Melody Transcription





Pitch Contour

Musical Notes

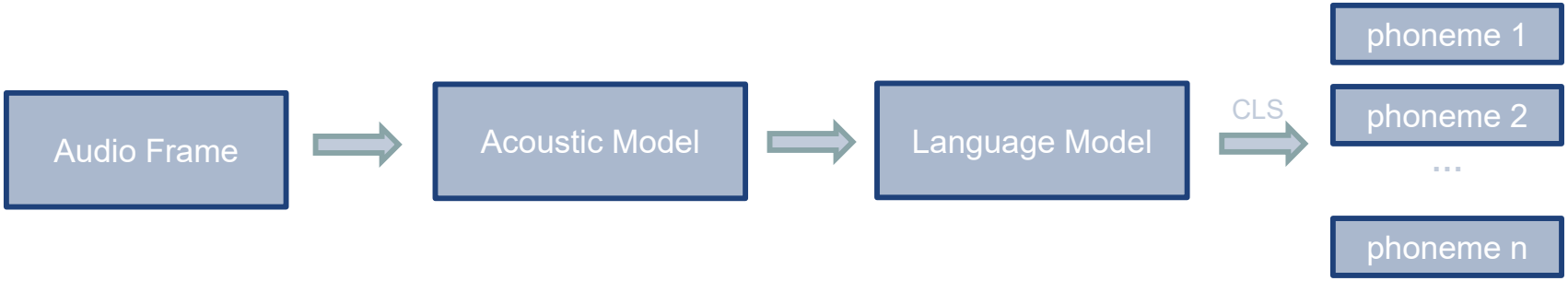


Pitch Correction Tool

Jui-Yang Hsu and Li Su. “VOCANO: A note transcription framework for singing voice in polyphonic music.” International Society of Music Information Retrieval Conference (ISMIR), 2021.

Singing Lyrics Transcription

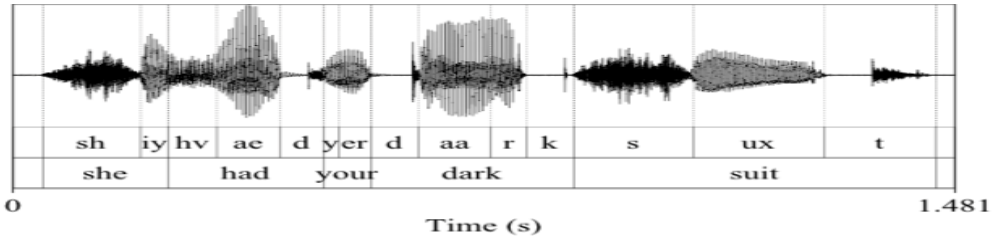
- **Lyrics transcription:** a variant of speech recognition task
- **Applications:** language education, subtitle generation
- **Challenges:** missing phonemes, irregular pronunciation, etc.
- **Methods:** DSP based HMMs, Neural networks (TDNNs)

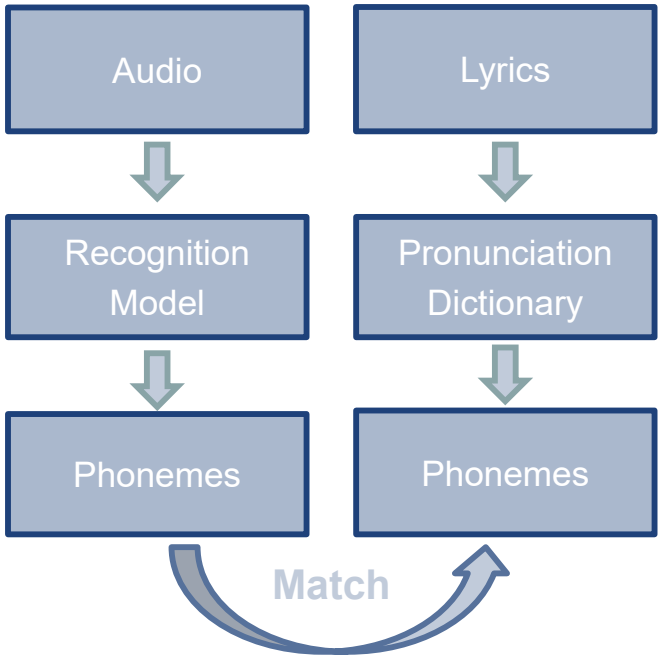


Sven Ahlbäck. “Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription.” International Society for Music Information Retrieval, 2021.

Singing Lyrics Alignment

- **Lyrics alignment:** a variant of the forced alignment task
- **Applications:** timestamps annotation, singing data collection
- **Possible methods:**
  - HMM-GMM models: P2FA, Kaldi, etc.
  - Neural blocks





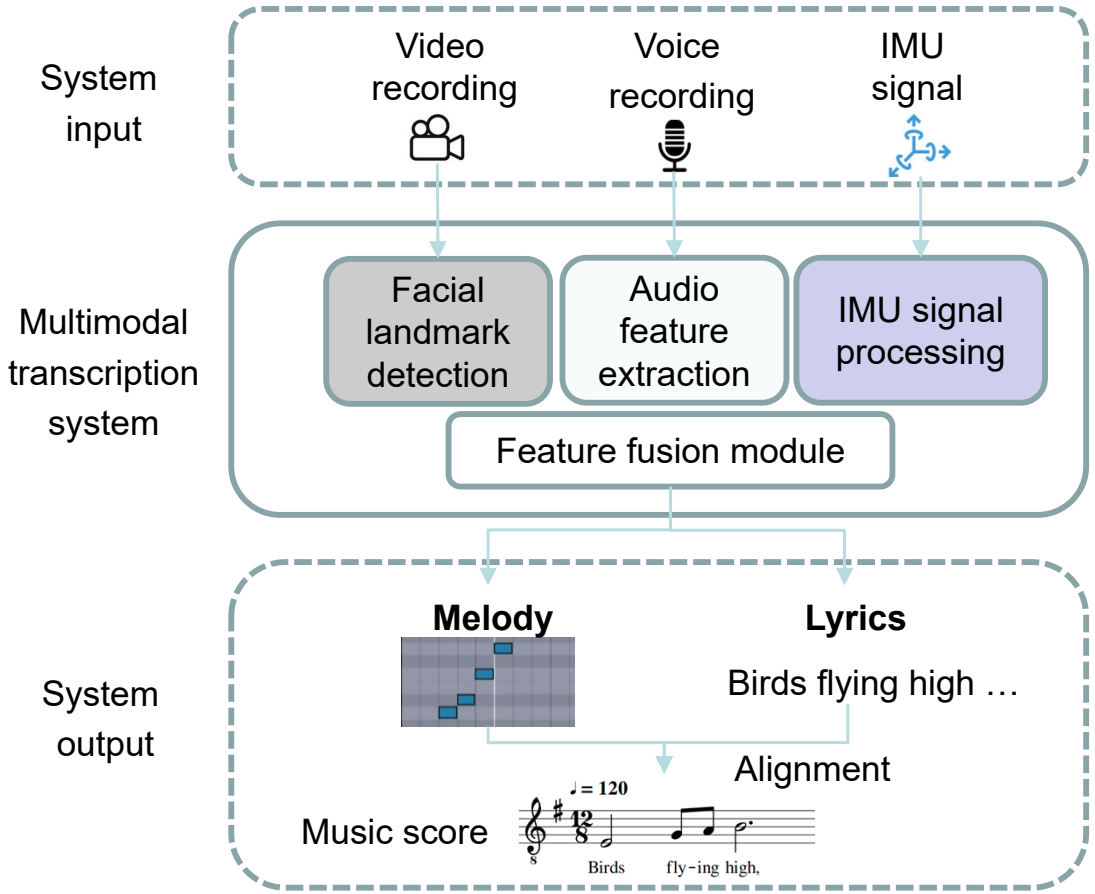
Demirel, Emir, Sven Ahlbäck, and Simon Dixon. “Low Resource Audio-to-Lyrics Alignment From Polyphonic Music Recordings.” IEEE ICASSP 2021.

Y. Wang, M. Kan, T. Nwe, A. Shenoy, and J. Yin, “LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics,” in Proceedings of ACM Multimedia 2004, pp.212-219.

M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe and A. Shenoy, “LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals,” IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no. 2, pp.338-349, 2008.

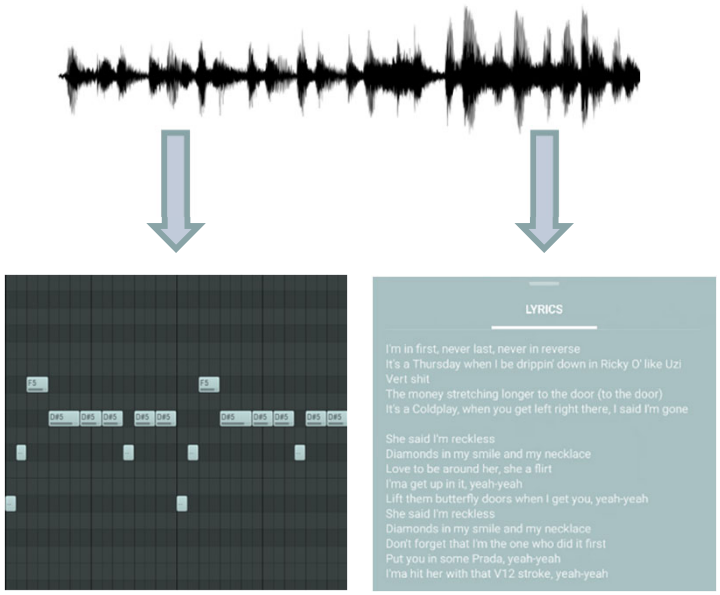
D. Iskandar, Y. Wang, M.-Y. Kan and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in Proc. ACM Multimedia 2006, pp.659-662.

Multimodal Singing Voice Transcription



Gu, X., Ou, L., Ong, D., and Wang, Y., (2022, October). MM-ALT: A Multimodal Automatic Lyric Transcription System, 2022 ACM Multimedia Conference (MM'22), October 10-14, 2022, Lisbon, Portugal. ACM. (Top Paper Award)

Current State of Singing Transcription



F0 Estimation	Lyrics Recognition
well developed	to be improved
Note Segmentation	**Lyrics Alignment**
to be improved	to be improved

Summary:

- publicly available **datasets** for singing melody transcription and lyrics recognition
- developing high-performed **neural singing transcription models** and **source separation models** that can handle polyphonic music

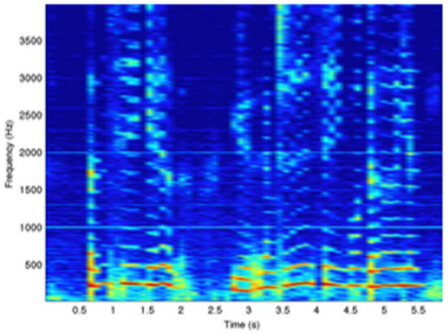
To do:

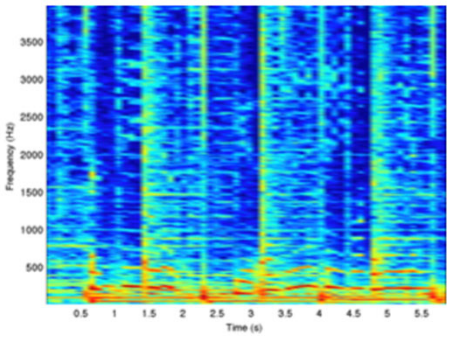
- **computational efficiency** of neural transcription models
- multi-notes transcription and multi-lyrics transcription of **singing signals with harmonies**
- **Fusion of multimodal data** for singing voice transcription

Monophonic and Polyphonic Music

Source separation

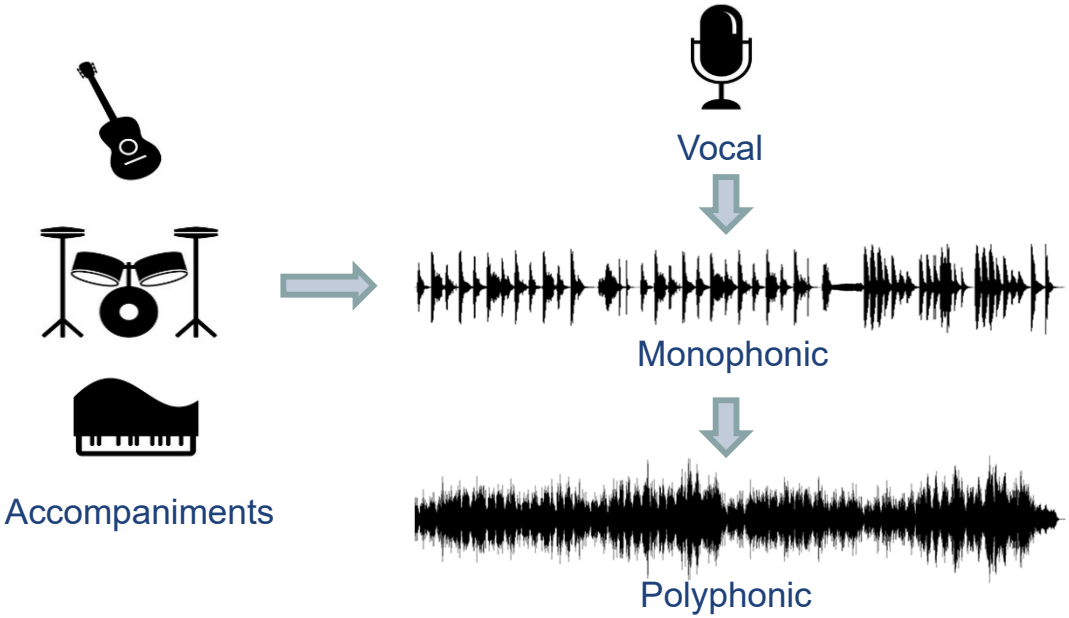
Vocal





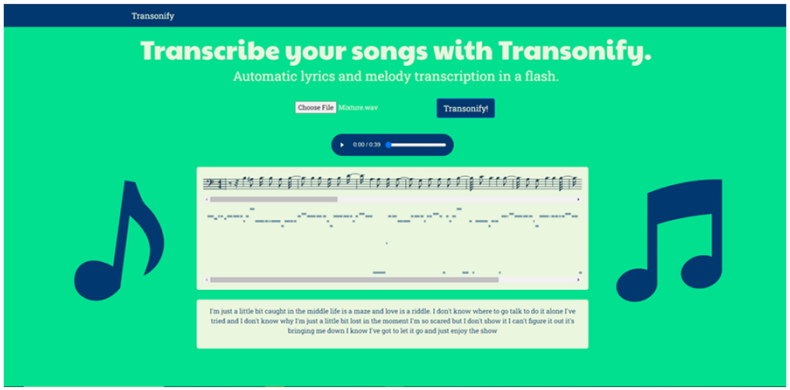
Mixture

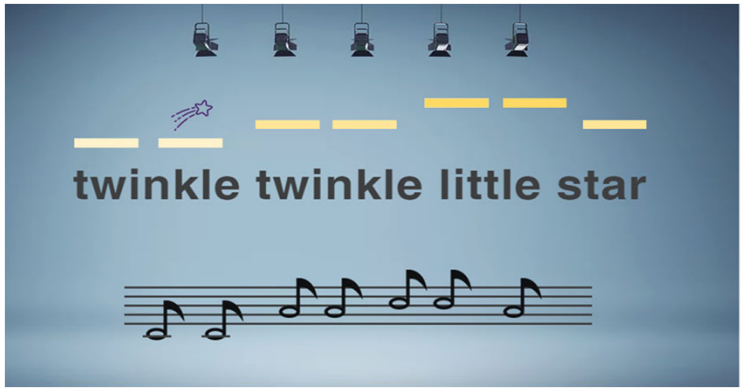
- **Monophonic Audio:** a single unaccompanied singing line
- **Polyphonic Audio:** singing audio with accompaniments
- **Challenges:** polyphonic music is more common, vocal and accompaniments overlap in the frequency domain, etc.

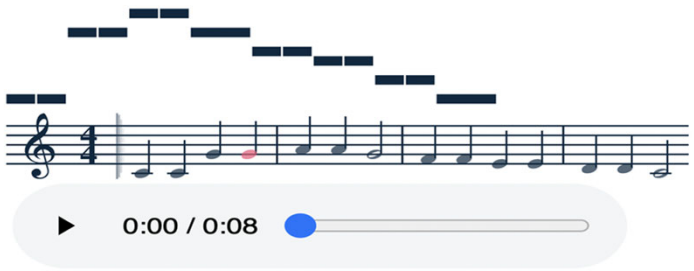


Comments on Mid-Project Presentations

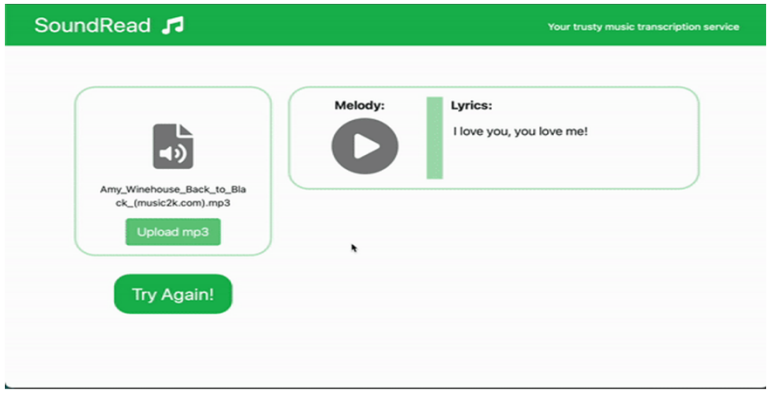
- Team work!
- Flexibility in the project scope
- Creativity in the interface design



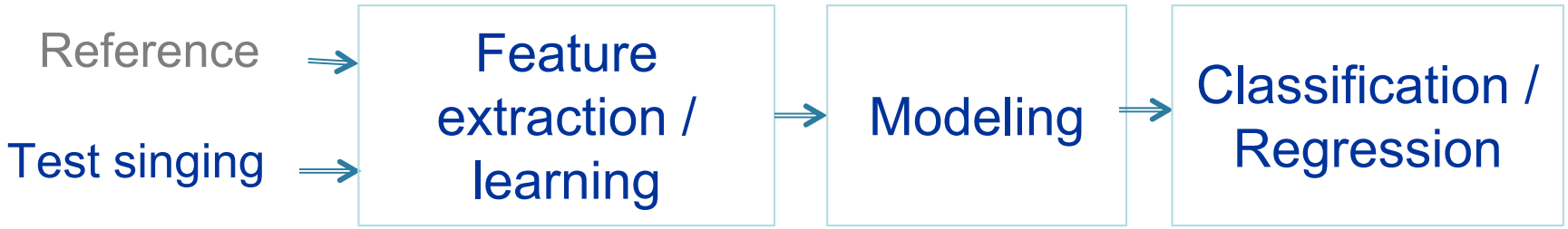




[Press to play the original song]



Automatic Singing Evaluation



Screening of singing talents / problems!

Murad, D., Wang, R., Turnbull, D., & Wang, Y. (2018, October). SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In 2018 ACM ACM International Conference on Multimedia (pp. 1679-1687). ACM.

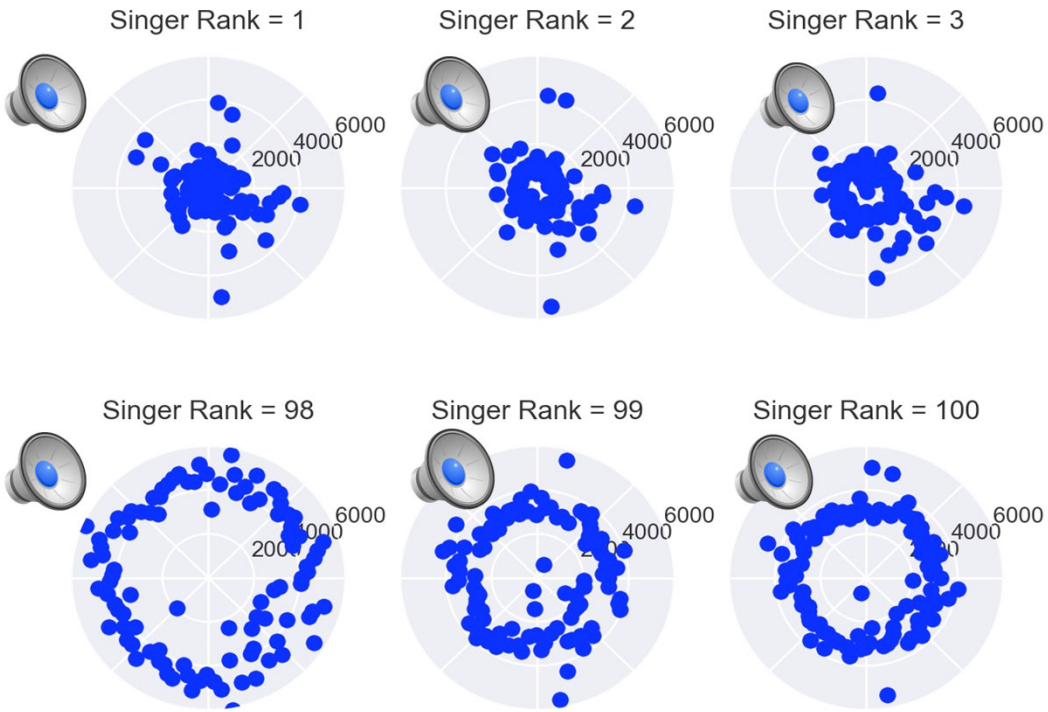
Gupta, C., Li, H., & Wang, Y. (2019). Automatic Leaderboard: Evaluation of Singing Quality without a Standard Reference. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1–1

Gupta, C., Li, H., & Wang, Y. (2018, November). Automatic Evaluation of Singing Quality without a Reference. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 990-997).

Gupta, C., Li, H., & Wang, Y. (2018). Automatic Pronunciation Evaluation of Singing. In Interspeech (pp. 1507-1511).

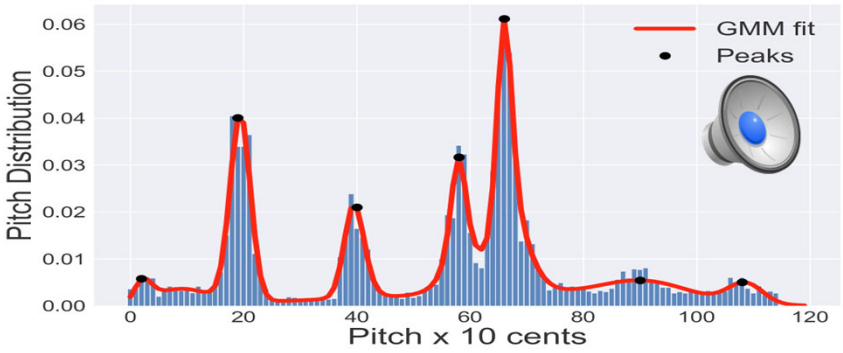
Gupta, C., Li, H., & Wang, Y. (2018). A technical framework for automatic perceptual evaluation of singing quality. APSIPA Transactions on Signal and Information Processing, 7.

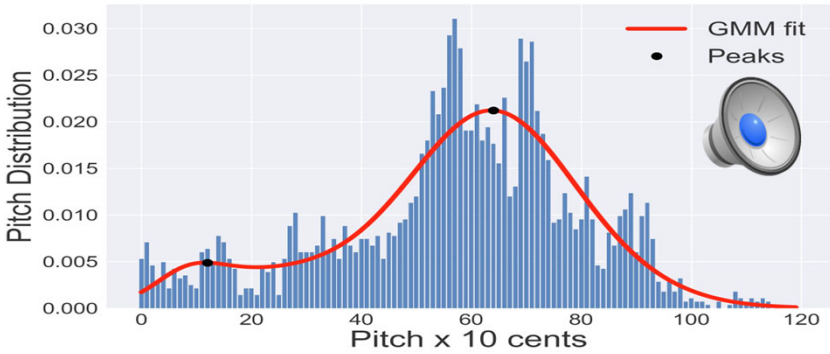
Key Ideas and Critiques



LEADERBOARD

Singer Names	Rank
Singer A	1
Singer B	2
Singer C	3
Singer D	4
Singer E	5
	...
	...
	...
Singer n	100





Gupta, C., Li, H., & Wang, Y. (2019). Automatic Leaderboard: Evaluation of Singing Quality without a Standard Reference. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1–1

Topics Today

Part A: Physiology & Physics of the Singing Voice

Part B: Singing voice transcription/evaluation

⇨ Part C: Singing voice synthesis/generation

Statistical method for singing synthesis

Virtual singer (a cartoon girl) VOCALOID

Concatenative synthesis

- Statistical unit selection
- Post-processing to nullify boundary discontinuity

Hatsune Miku Phenomenon

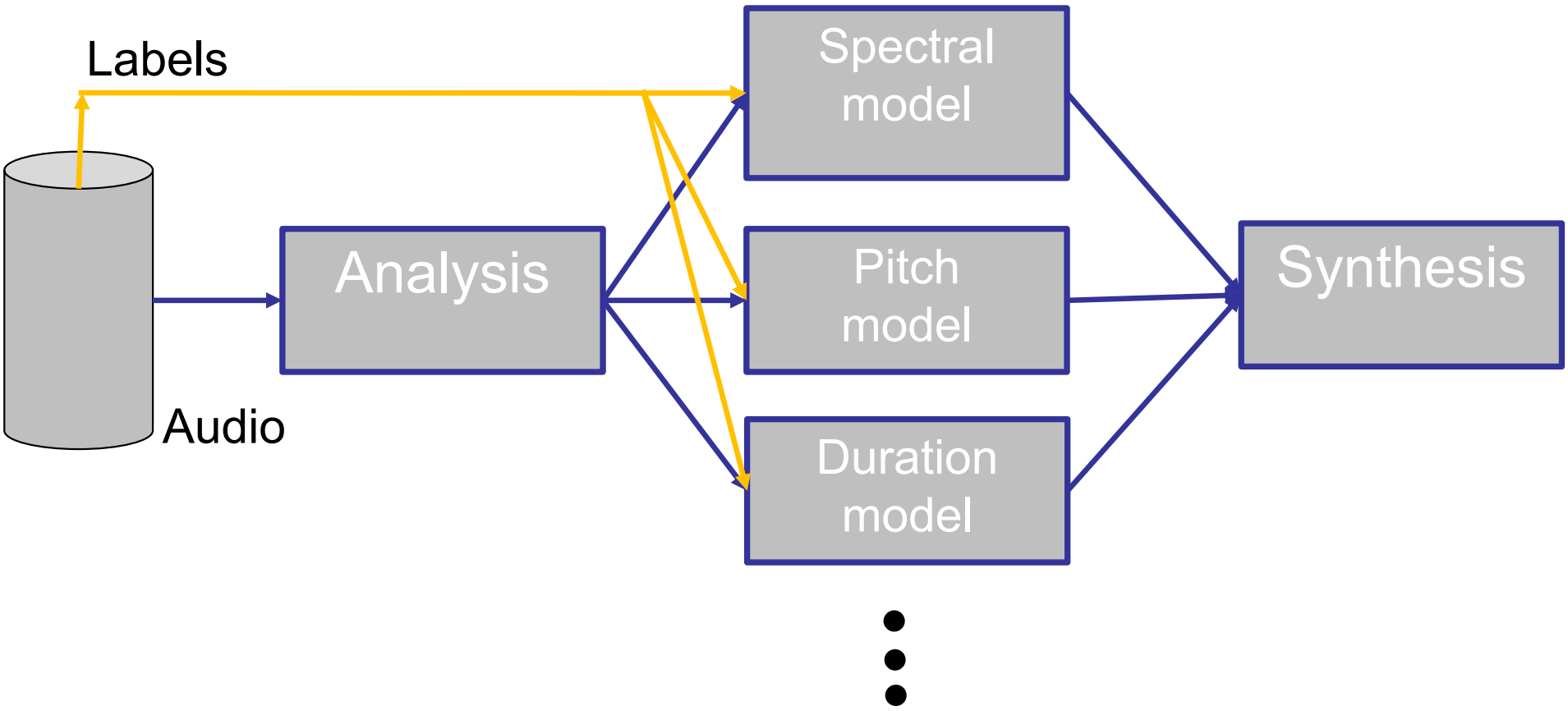
The most surprising change of human behavior:

Singing synthesis breaks down the long-cherished view that "listening to a non-human singing voice is worthless", emerging the "culture in which people actively enjoy songs with synthesized singing voices as the main vocals"

Singing synthesis: from Classic to DL

- Singing - what to cater to?
  - Pitch
  - Duration
  - Loudness
  - Spectra – singing formants
- Feature extraction
  - MFCC, PLP, pitch/energy contours, histograms, etc.
- Methods for singing synthesis
  - Statistical methods
  - Deep learning methods

Pipeline for Singing Analysis & Synthesis



Practical Issues

- Data availability
  - Large amount of data from single person
  - Labelled data
  - Quality of recording
- Average voice modelling
- Adaptation
  - Singer independent training
  - Singer dependent training (*mimicking*)

Eternal Voice?

How can we synthesize natural singing voices by analyzing and imitating human singing?

- Imitate the pitch, dynamics, phoneme timing, and breath of the singer's voice
- Estimate parameters of singing synthesizer

MIR Tasks Related to Singing

- Singing transcription and analysis
  - Predominant melody extraction
  - F0 estimation (monophonic, polyphonic)
  - Note segmentation
  - Representation issues
- Vocal activity detection (VAD)
- Singer identification (timbre analysis)
- Singing skill evaluation
- Vocal timbre analysis
- Lyric transcription and synchronization
- Singing synthesis

Open Problems - Challenges for MIR

- Representation of singing
  - Event based representations (scores, MIDI) are insufficient
  - Continuous pitch tracks capture detail of intonation (ornaments, glides, vibrato, kobushi*), but segmentation into notes is difficult
  - Integration of timbral information (phonation, spectral characteristics, phonemes/lyrics) into singing representations
- Algorithms to compare and assess pitch tracks
- Holistic similarity (or skill) estimation that includes pitch, timing and timbre
- New MIREX tasks: assess a singer's naturalness or purity of tone

*Y. Ikemiya, K. Itoyama, and H.G. Okuno (2014). “Transcribing Vocal Expression from Polyphonic Music”. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3127-3131.

Pitch tracking implementations I

- YIN Java implementation in Tarsos: https://github.com/JorenSix/TarsosDSP, Matlab implementation see http://www.auditory.org/postings/2002/26.html, Vamp implementation in pYIN: https://code.soundsoftware.ac.uk/projects/pyin
- MELODIA (Vamp plugin) http://mtg.upf.edu/technologies/melodia
- pYIN Vamp plugin and source code: https://code.soundsoftware.ac.uk/projects/pyin or Python implementation: https://github.com/ronggong/pypYIN
- STRAIGHT (Matlab, available upon request) http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html

Pitch tracking implementations II

- SWIPE Matlab: http://www.cise.ufl.edu/~acamacho/publications/swipep.m
  SPTK/Python: http://pysptk.readthedocs.org/en/latest/sptk.html#f0-analysis
- Tartini for Supercollider http://doc.sccode.org/Classes/Tartini.html or standalone http://miracle.otago.ac.nz/tartini/
- Aubio http://aubio.org/ or in Vamp: http://aubio.org/vamp-aubio-plugins/
- RAPT (Matlab) http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/fxrapt.html
- mirtoolbox https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox
- Cepstral Pitch Tracker https://code.soundsoftware.ac.uk/projects/cepstral-pitchtracker

Why Singing is Interesting

- Inherent reasons
  - people love to sing
  - people love to listen to other people singing
  - *people might love to listen to computers singing*

- Scientific reasons
  - scientific discovery in music psychology: how people sing, and how people perceive singing
  - scope for historical and cultural analysis: how people's singing differs and changes

- MIR reasons
  - many MIR tasks relating to singing can be improved, and new ones explored!
  - there's more data out there (even annotated), which we can exploit
  - Bridging speech and music technology research communities together

Is Singing a unique ability for human being?

Ye Wang, “Singing Voice Modelling for Language Learning” (Dagstuhl Seminar 19052). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Germany, 2019

When a Child Is Born

https://www.youtube.com/watch?v=ZCqhX89WV_0