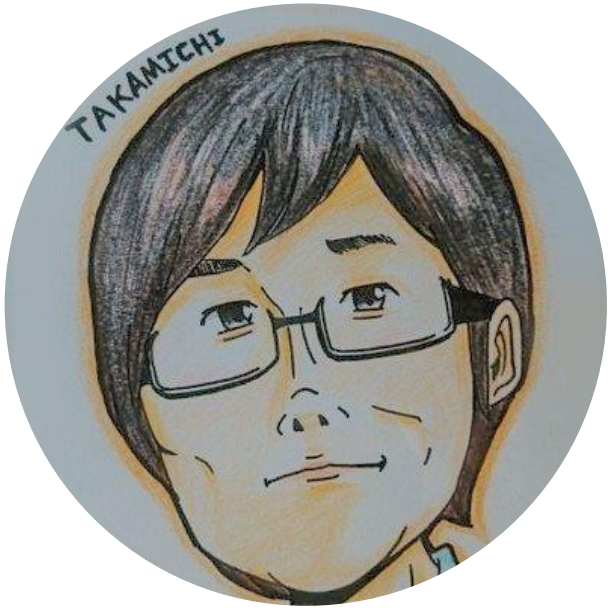
Invited talk at International Workshop on Symbolic-Neural Learning

How do audio foundation models understand sound?

Shinnosuke Takamichi (Keio University)

Takamichi Laboratory
慶應義塾大学 高道研究室

Self introduction



@forthshinji

Name

Shinnosuke Takamichi / 高道 慎之介

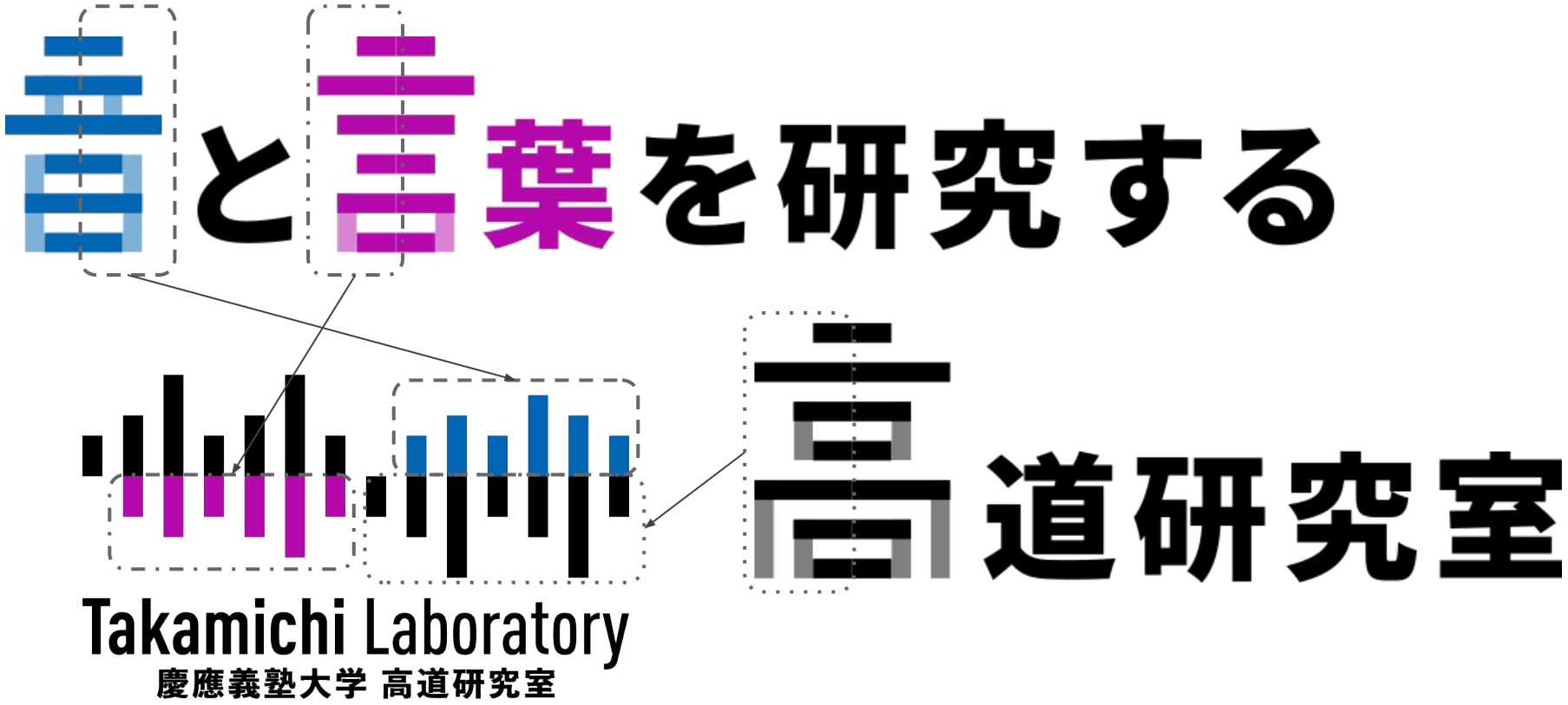
Affiliation

Associate Prof. of Keio University, Japan

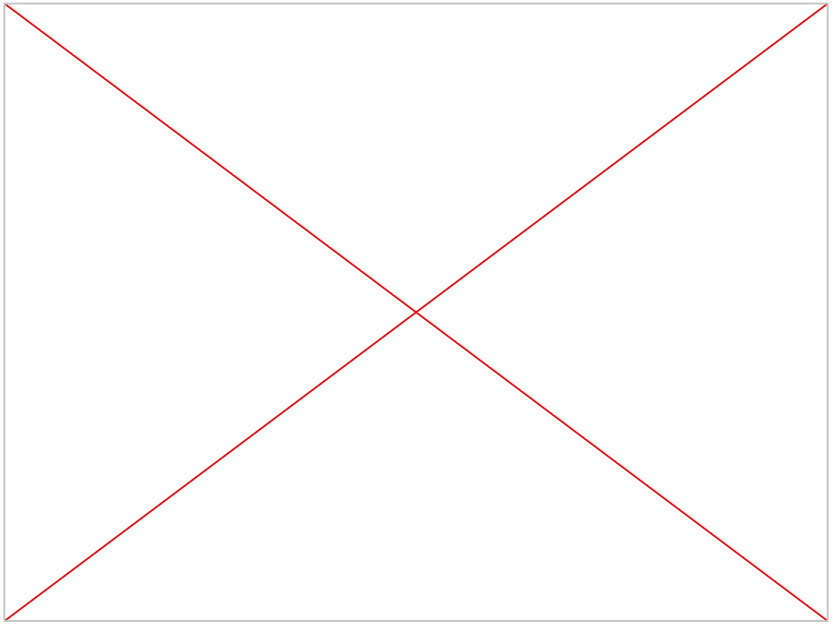
Major

Speech processing

Takamichi (高道) laboratory in Keio,
studying audio (音) and language (言葉).

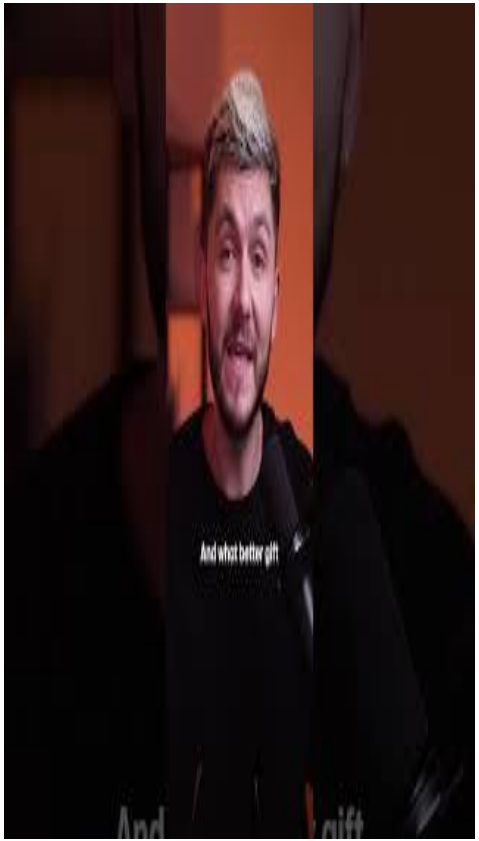


Foundation models for audio (from understanding to synthesis)



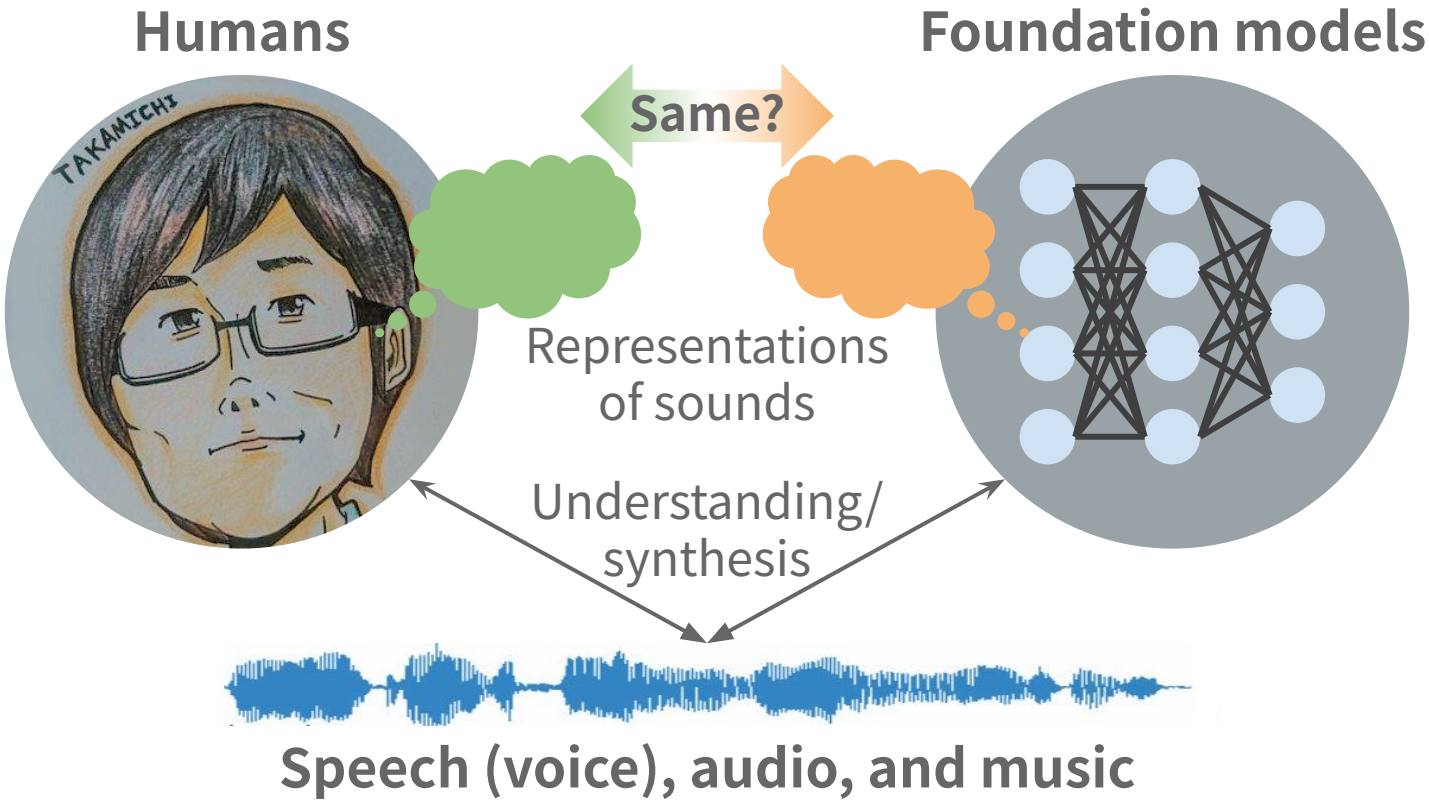
video-to-audio (MMAudio)

text-to-music (Suno)



text-to-speech (E labs)

What do the foundation models learn from data?



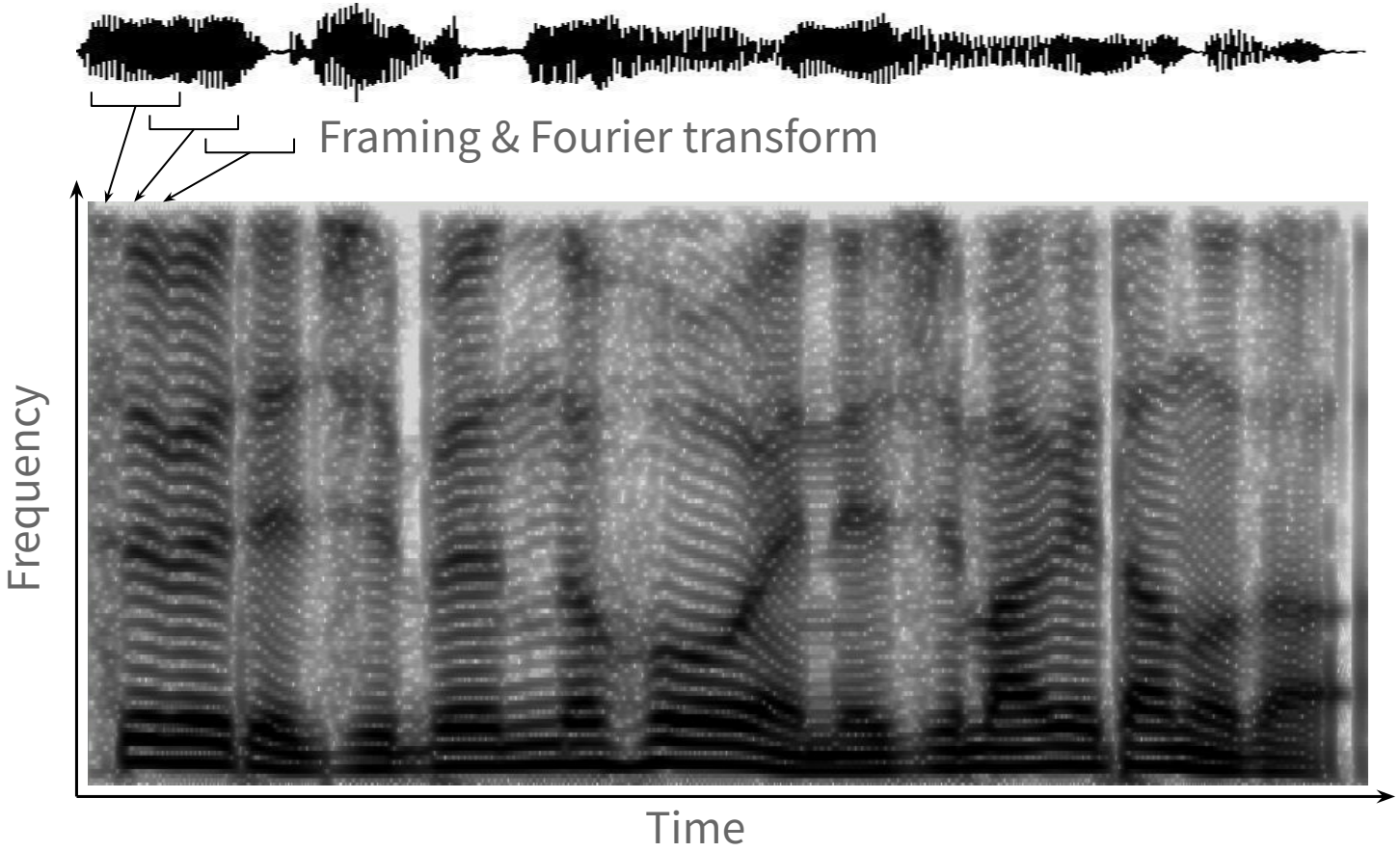
Topic 1: Language of learned audio symbols

S. Takamichi et al., “Do learned speech symbols follow Zipf's law?,” ICASSP, 2024.

S. Kando, S. Takamichi et al., “Exploring the effect of segmentation and vocabulary size on speech tokenization for speech language models,” Interspeech, 2025.

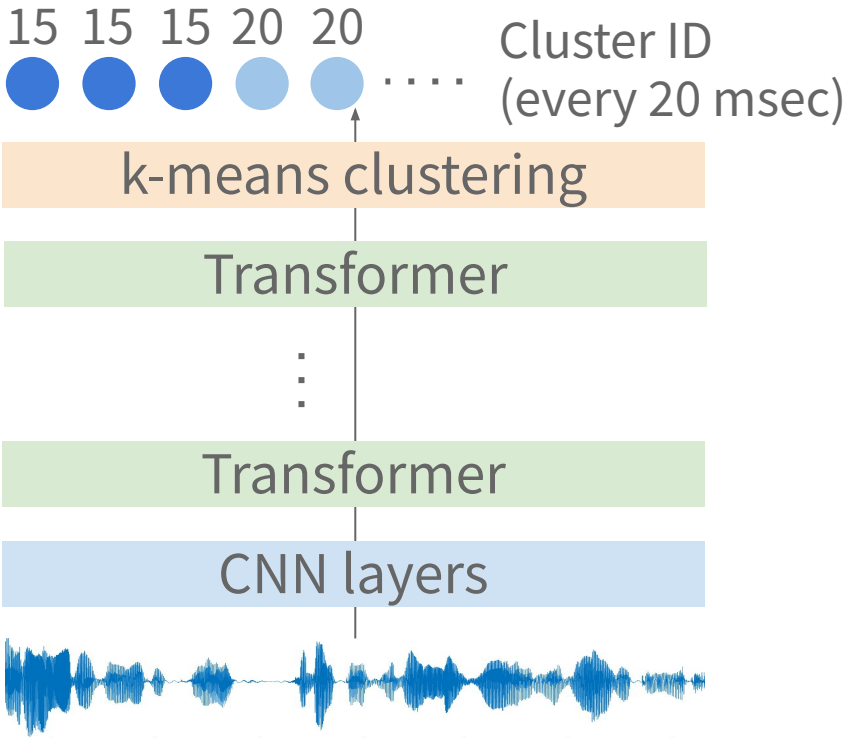
J. Park, S. Takamichi et al., “Analyzing the language of neural audio codecs,” ASRU, 2025.

Spectrogram: traditional representations of sound

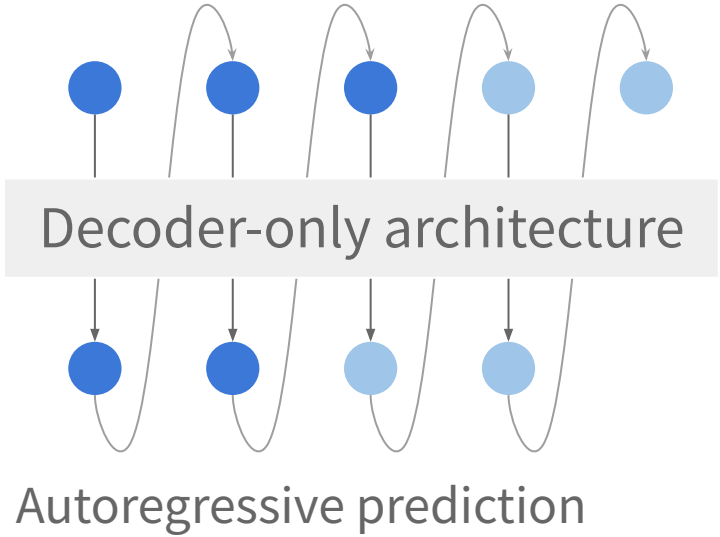


Speech foundation models based on learned audio symbols

Obtaining learned symbols



GPT-like modeling of learned symbols



(On a different note) Zipf's law for natural language

Zipf's law [Zipf49]

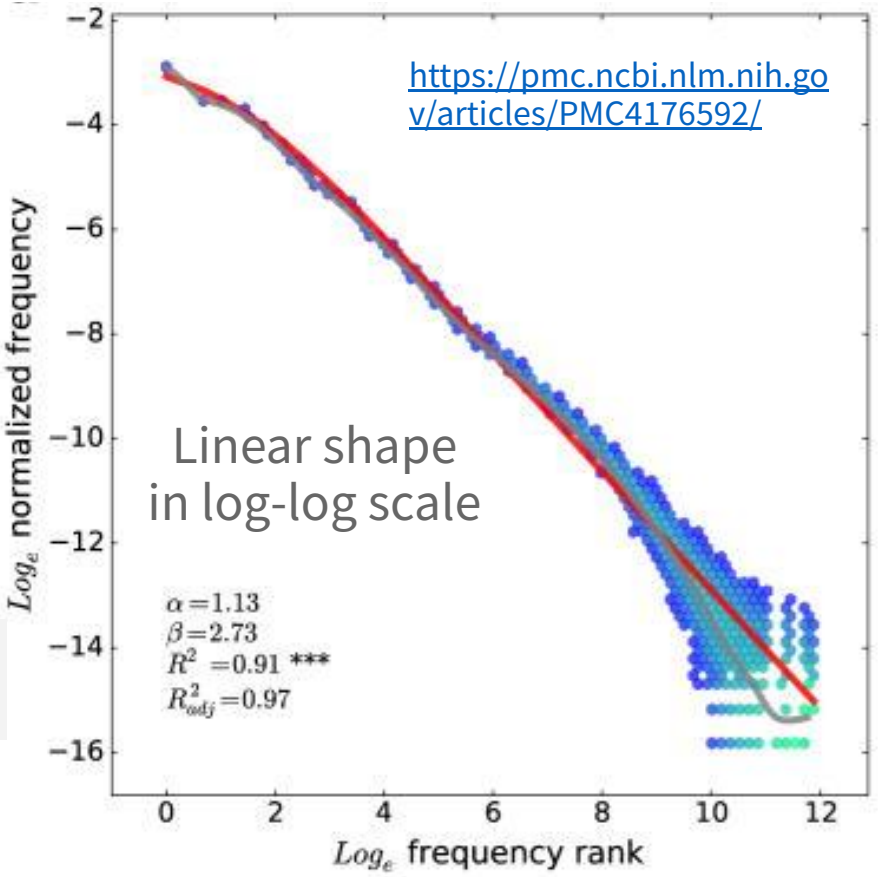
- Empirical principle that delineates the frequency of occurrence
- When the occurrence frequency of an element ranks as the *k*-th highest in a set, it equates to *1/k* of the frequency of the most occurring element.

Parameters

$$f_r = ar^{-\eta}$$

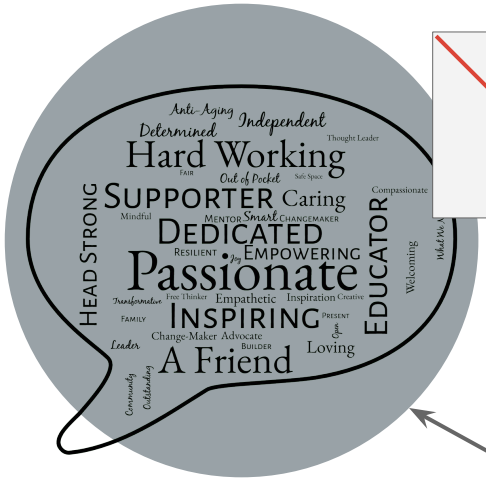
Frequency (count)

Rank

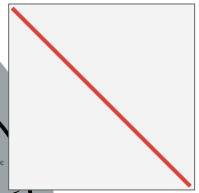


Natural language symbols vs. learned symbols

Natural language
invented by humans

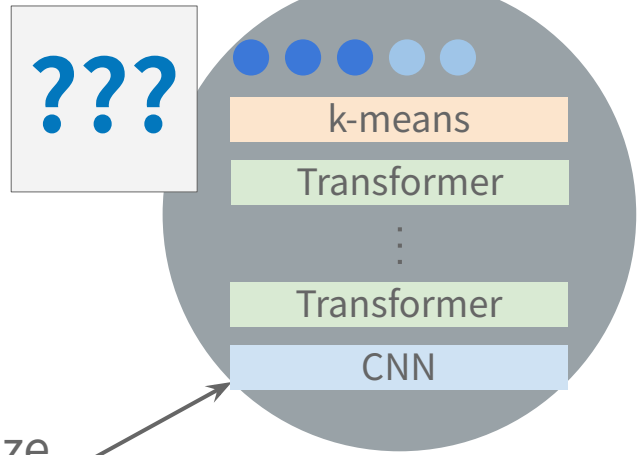


https://wordcloud.app/user-clouds/dedicated-inspiring-passionate-supporter-a-friend-2a2c6aa



Zipf's law

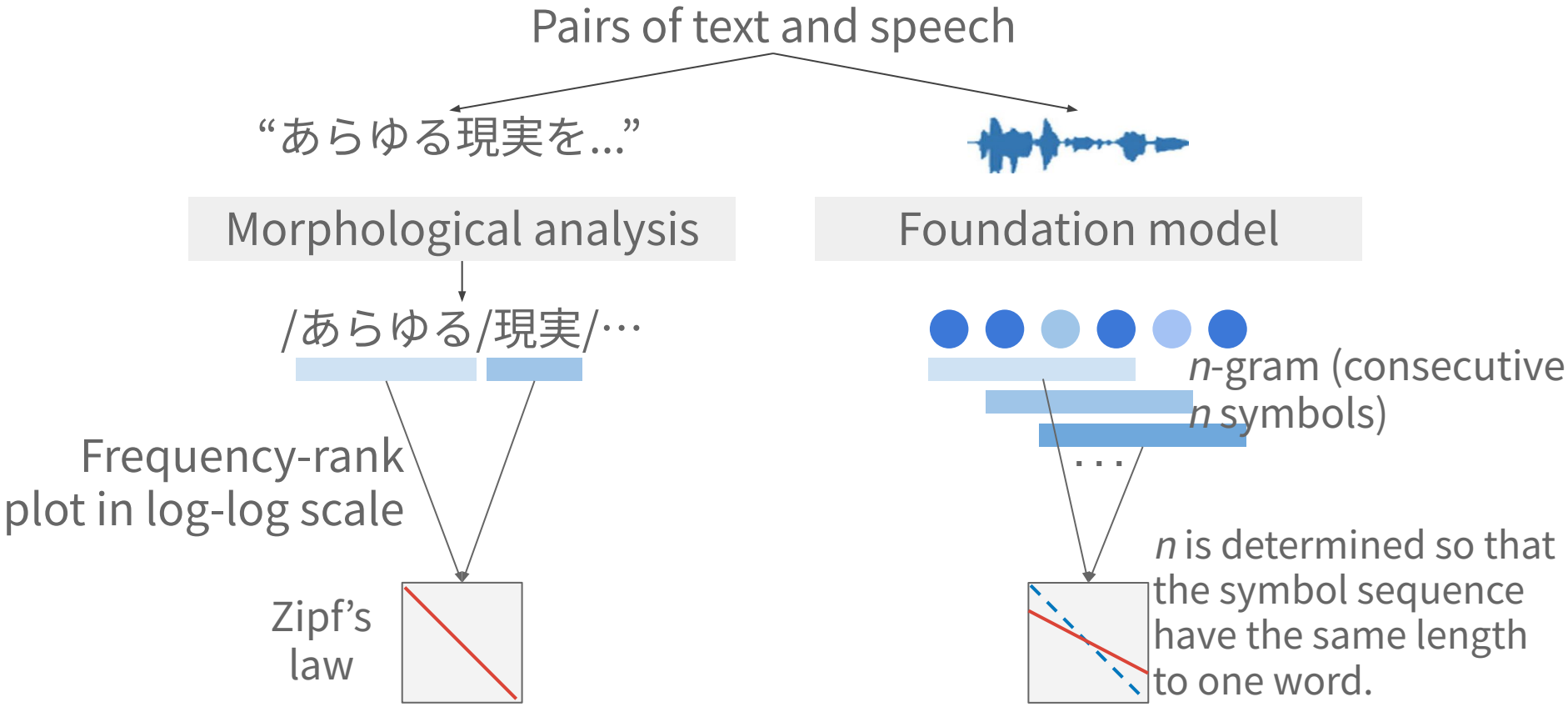
Symbol language
found by foundation models



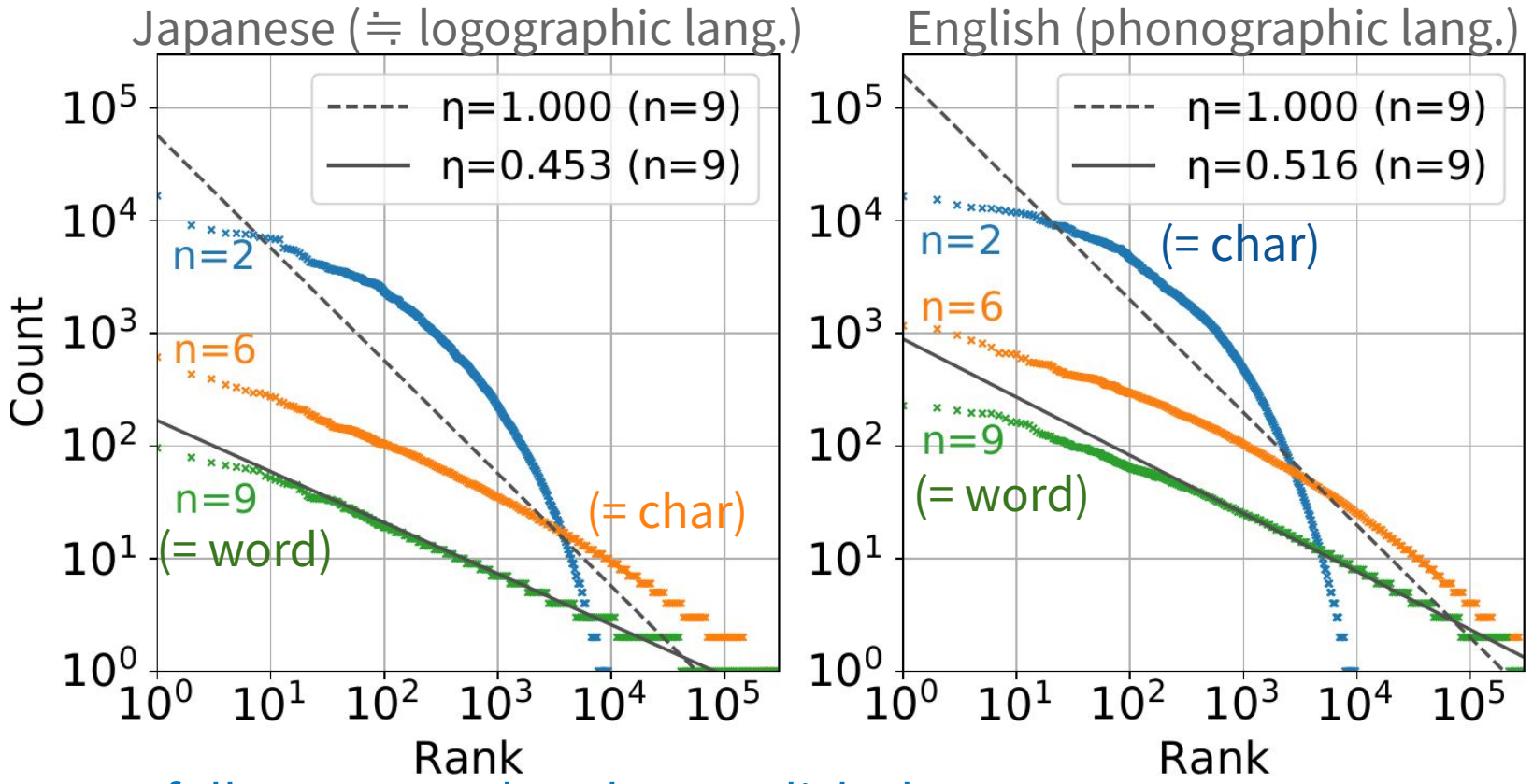
Symbolize

Speech

Methodology

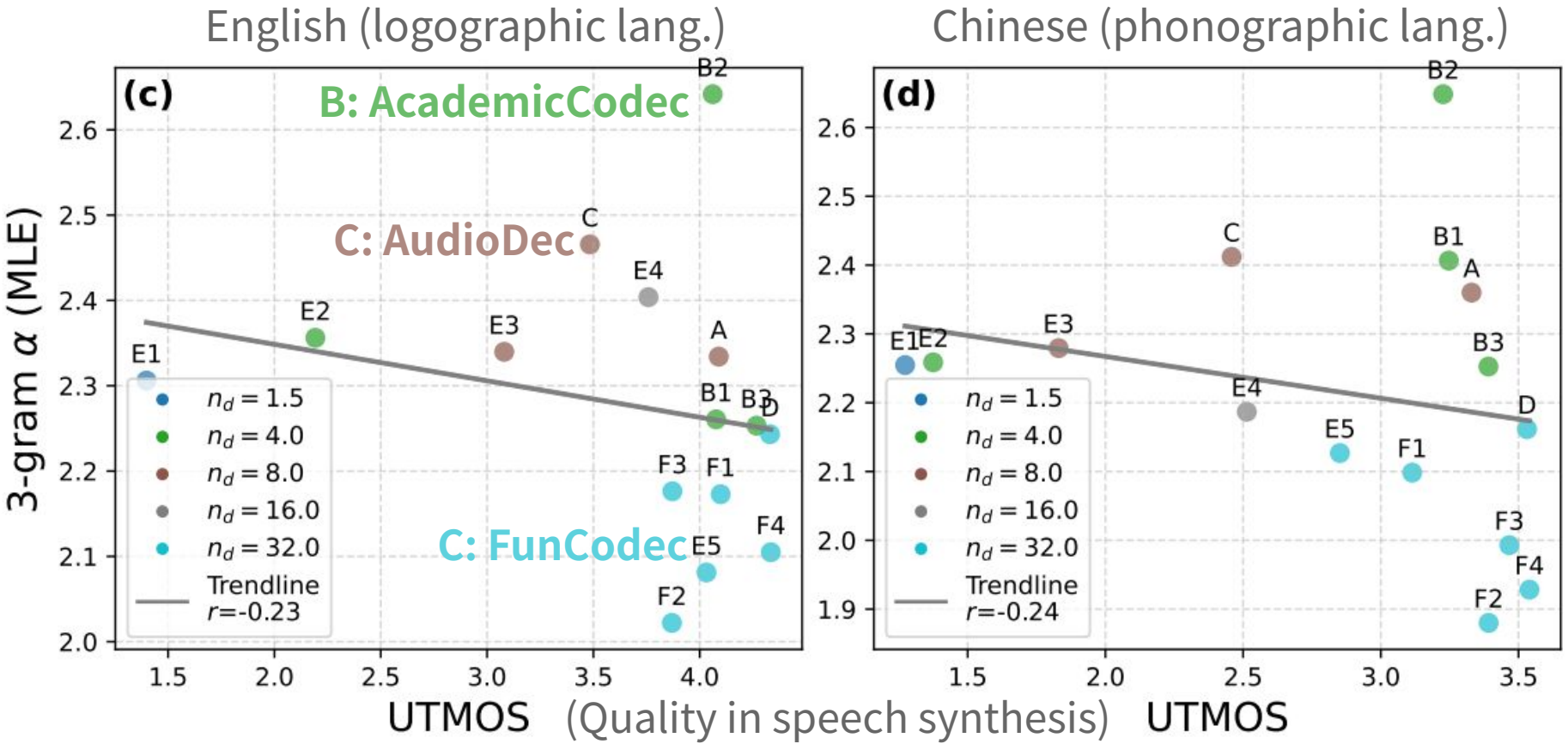


Results of learned speech symbols



Japanese follows power law, but English shapes convex.

Parameter ($\alpha = \eta + 1$) in Zipf's law correlates with synthetic speech quality.



Future work: spoken language is not just audible text, but…

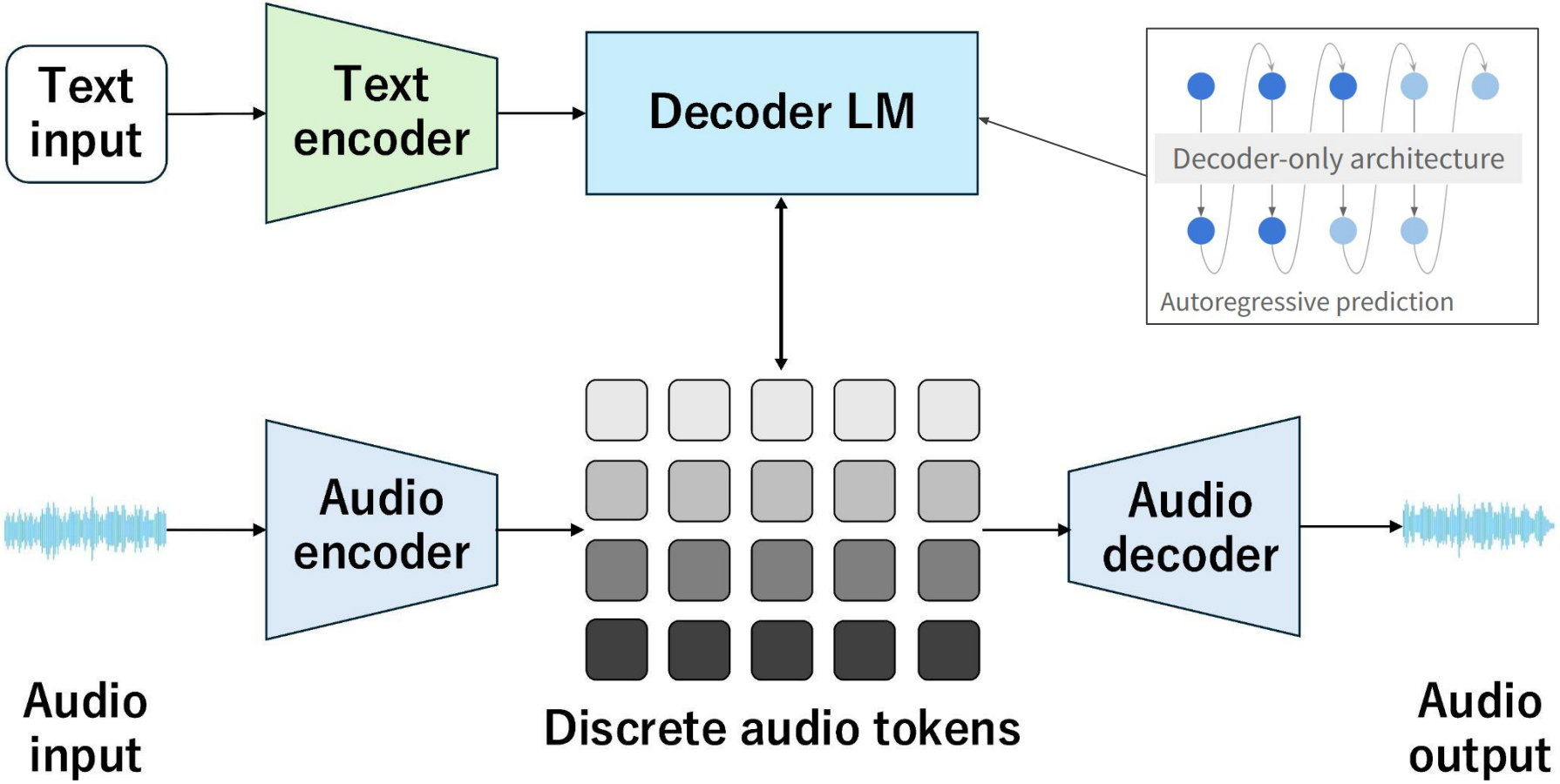
- "Spoken language is not just audible text (= natural language)"
  - Prof. Roger K. Moore (Sheffield Univ.) in Interspeech 2025
    - (not his original phrase)
  - Speech has not only textual information but also tone, pitch, disfluency. timing, etc.
  - Sometimes speech cannot be symbolized natural languages.

- Learned-symbol languages can symbolize any speech and be analyzed by NLP-inspired methods.
  - Zipf's law, Heap's law, etc.

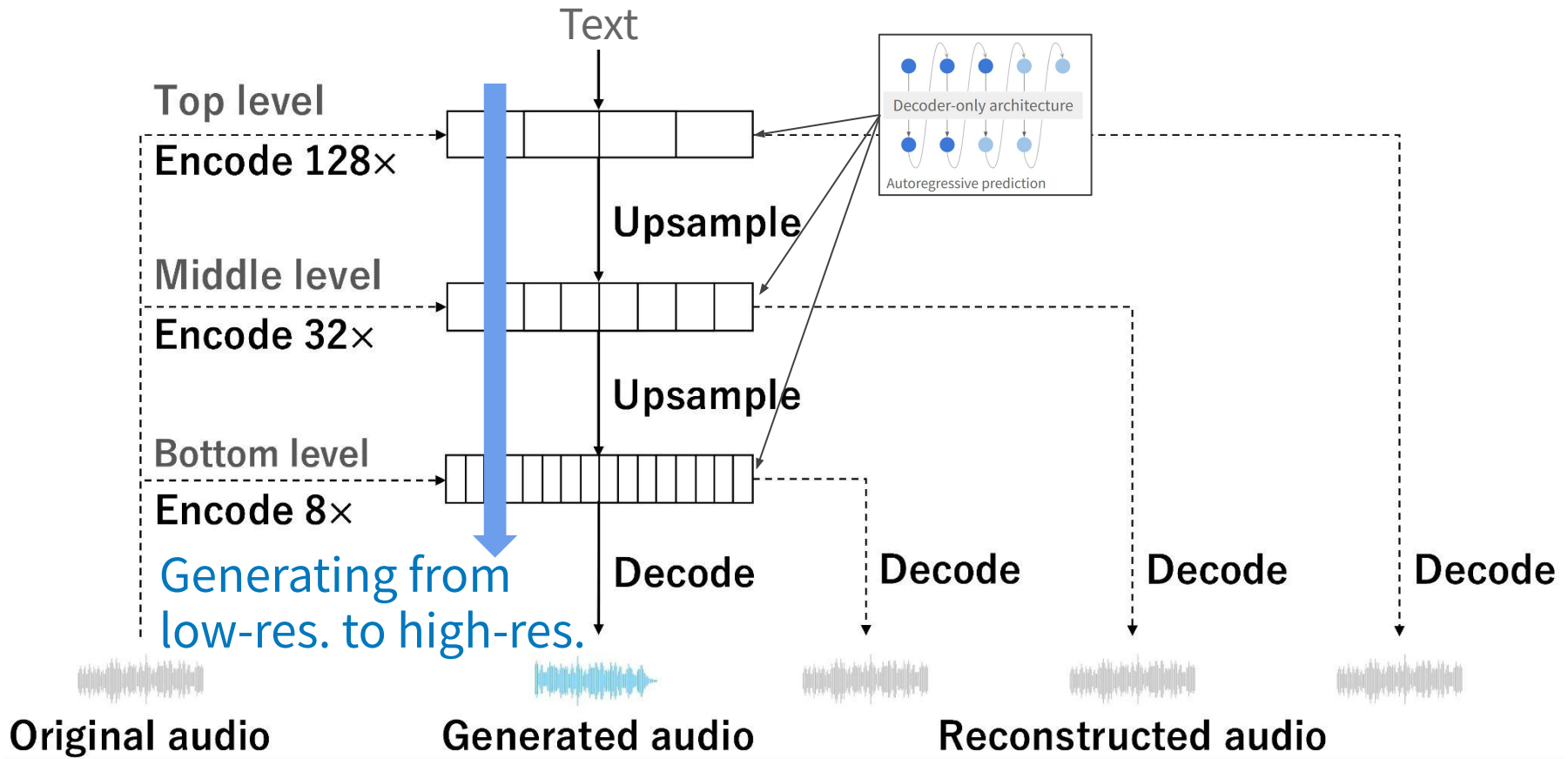
Topic 2: Geometry of music theory in music foundation models

八木 and 高道, “音楽基盤モデルは音高情報を螺旋構造に埋め込むか？,” MUS, 2025.
(as of now, Japanese-language paper only)

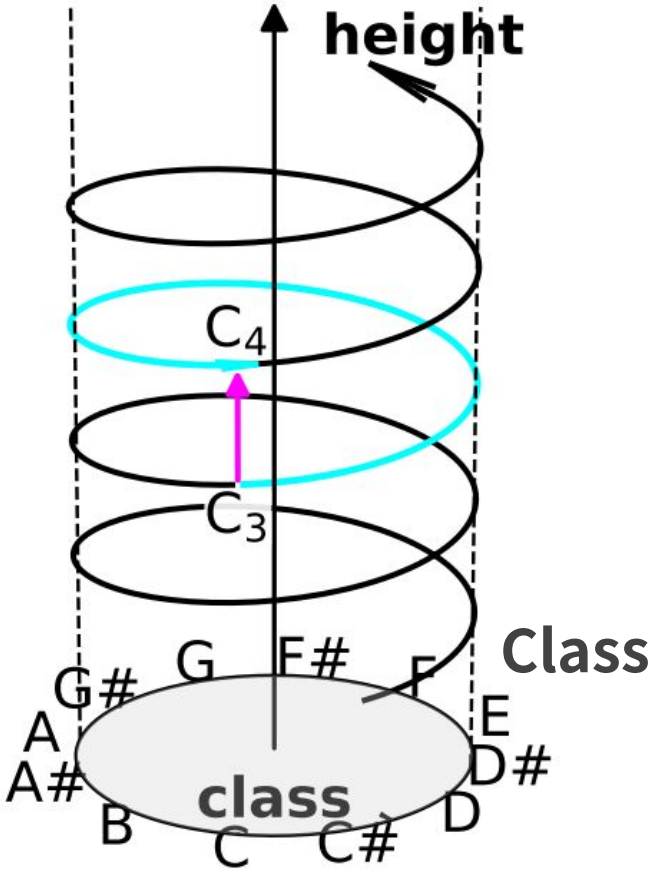
Music foundation models: non-hierarchical MusicGen[Copet23]



Music foundation models: hierarchical JukeBox[Dhariwal20]

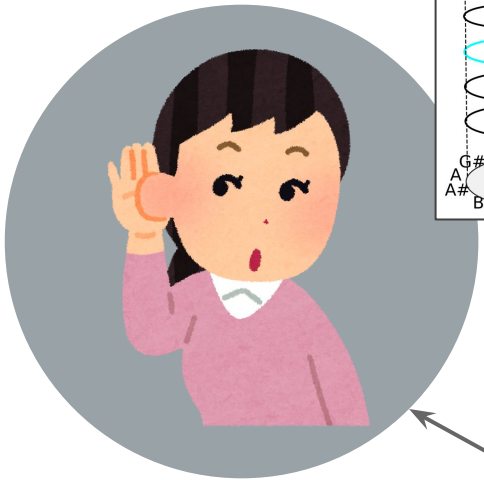


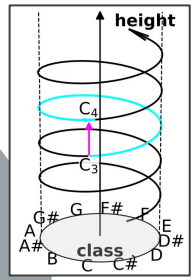
Pitch helix in human perception



Pitch helix vs. ???

Pitch perception
in humans

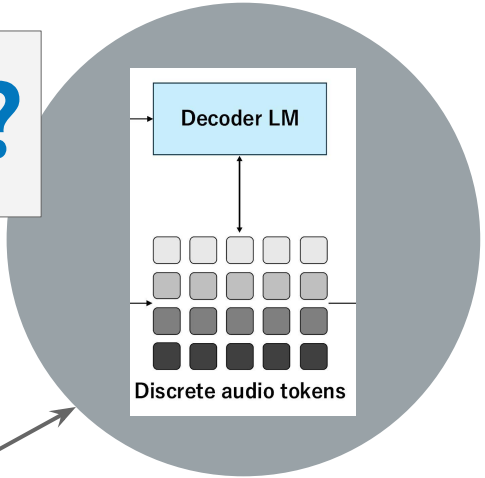




Pitch
helix

Pitch representation
in foundation models

???



Symbolize

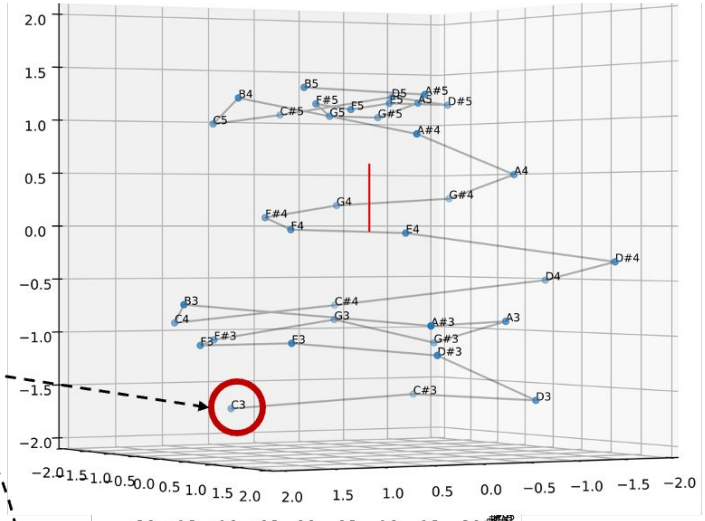
Music

Methodology 1: feature extraction

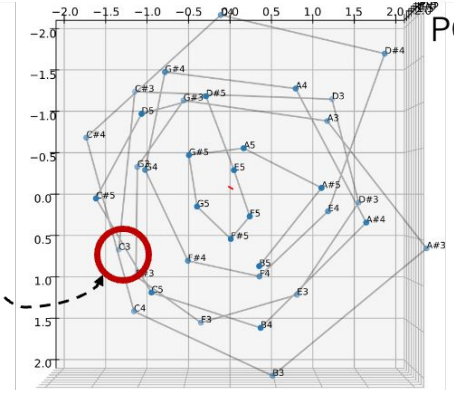
Input: Single Note（C3~B5）

(3-dim)

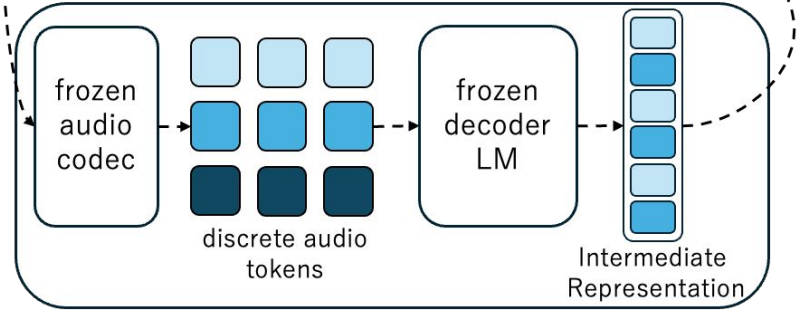
PCA



PC1-PC2-PC3



PC2-PC3



Foundation Model

Methodology 2: helix function

Height Rotation

$$\boldsymbol{y}(t) = h(t) \cdot \boldsymbol{c} + r(t)\{\cos\theta(t) \cdot \boldsymbol{u} + \sin\theta(t) \cdot \boldsymbol{v}\}$$

Height Radius Phase

$\{\boldsymbol{c}, \boldsymbol{u}, \boldsymbol{v}\}$: Orthonormal basis

t : pitch index

$$h(t) = h_{\text{pitch}} \cdot t + h_0$$

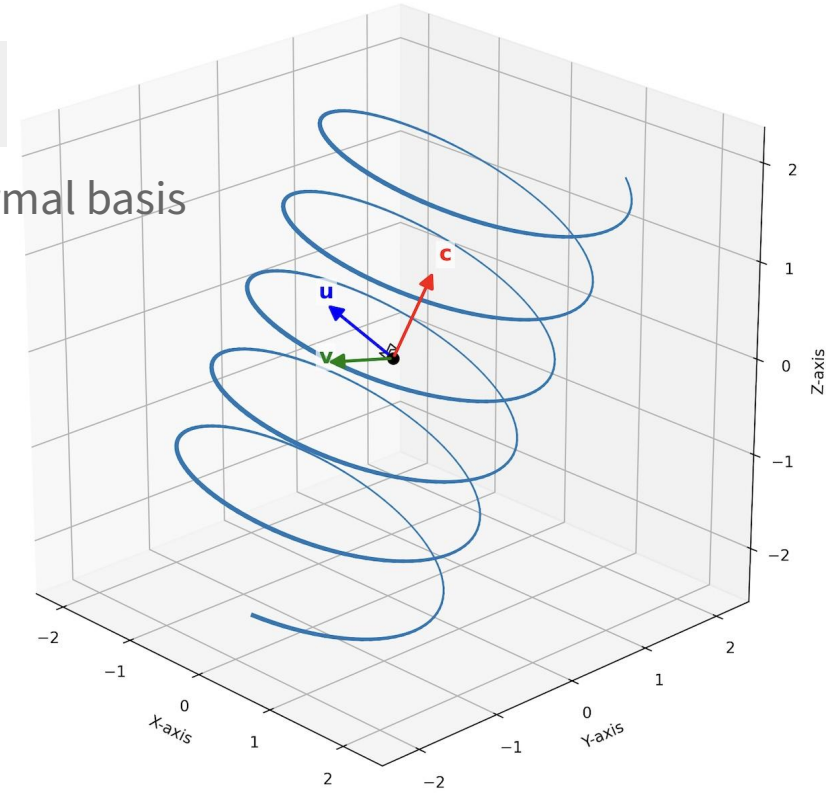
Height coef. Init. height

$$r(t) = r_{\text{slope}} \cdot t + r_0$$

Radius coef. Init. radius

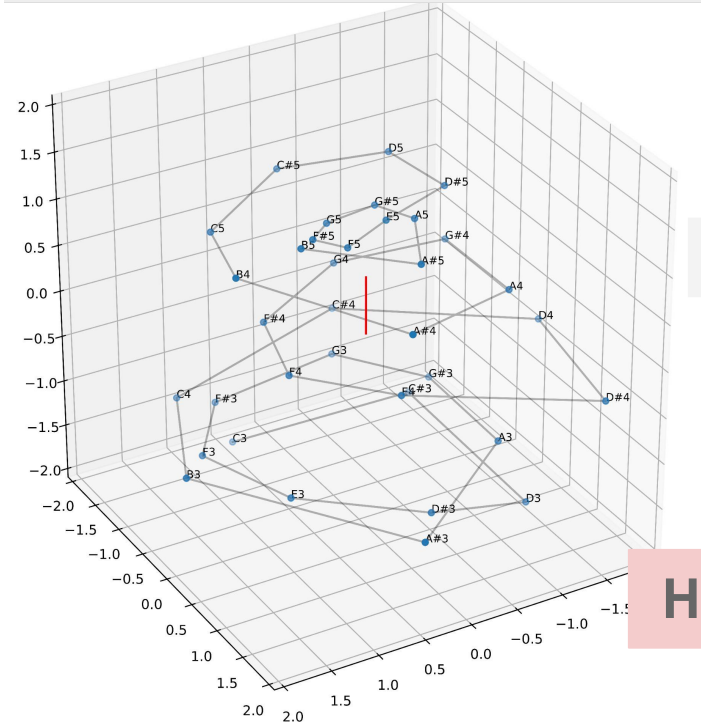
$$\theta(t) = \omega_{\text{chroma}}(t - t_0)$$

Angular freq. Init. position

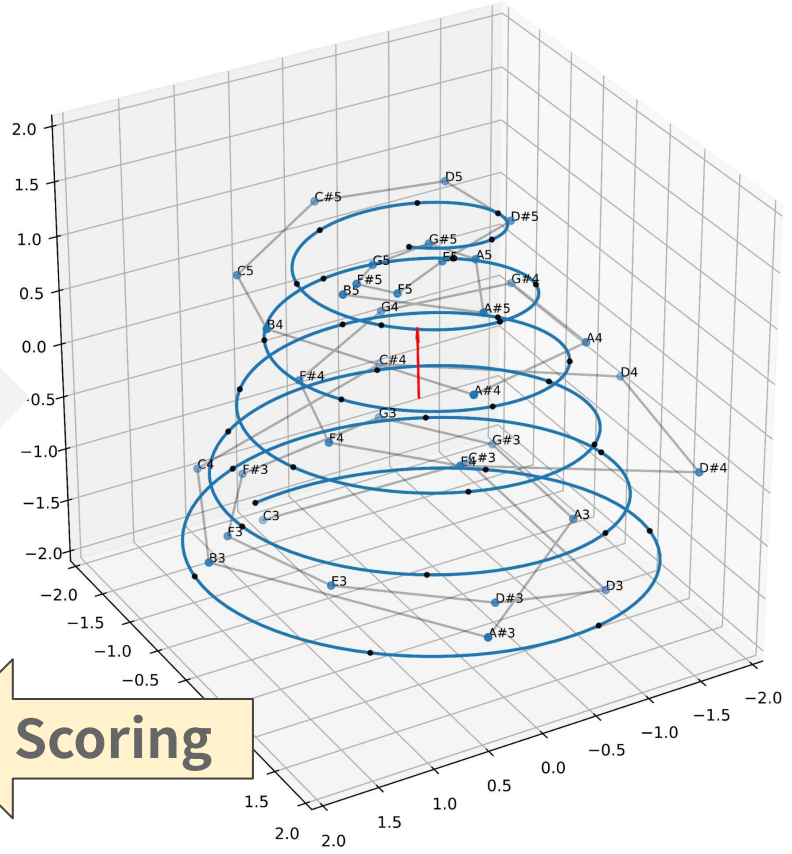


Methodology 3: fitting and scoring

$$\text{Helicality} = \left(\frac{1}{N_{\text{key}}}\sum_{t=1}^{N_{\text{key}}} |\boldsymbol{x}_t - \boldsymbol{y}(t)|^2\right)^{-1}$$



Fitting

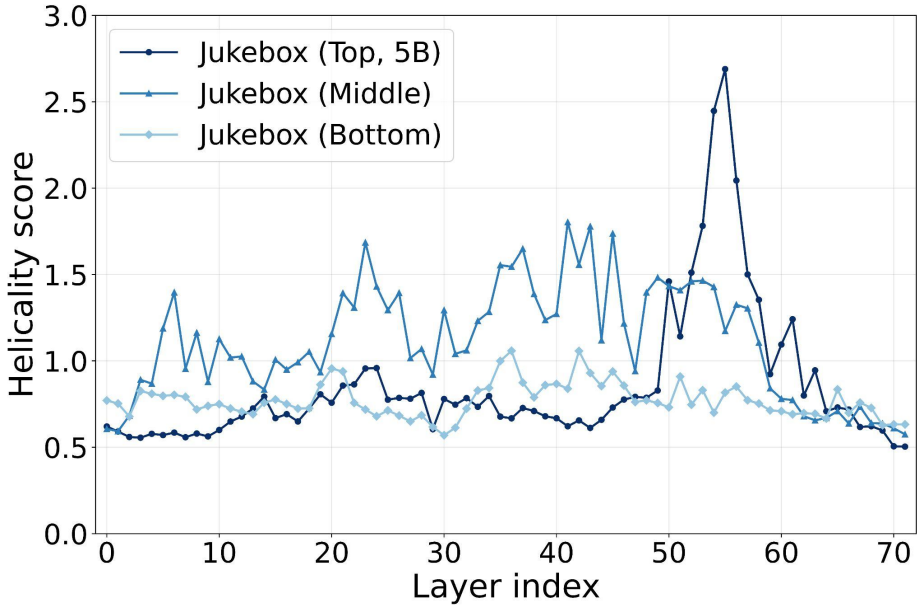


H = 2.60

Scoring

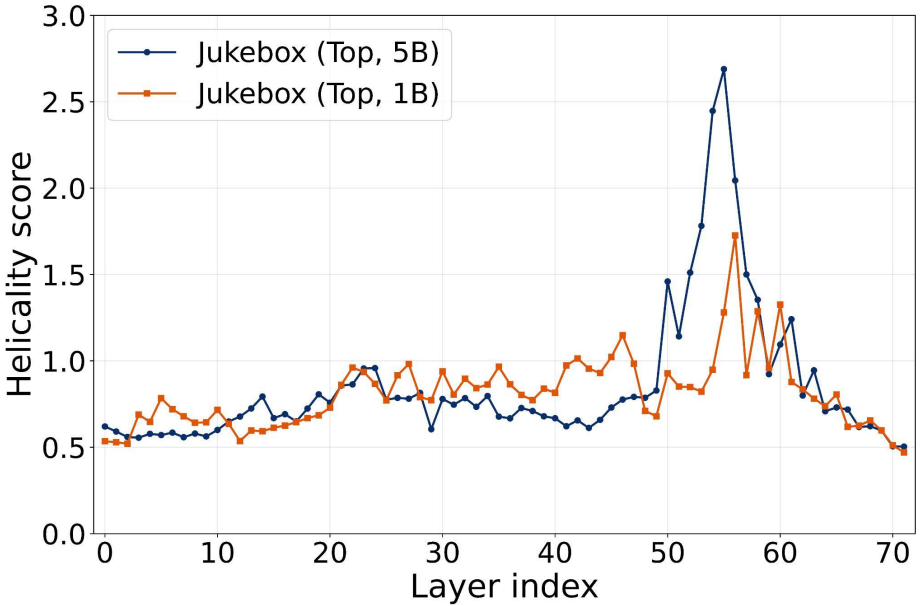
Results in Jukebox (hierarchical model)

Comparison in hierarchy



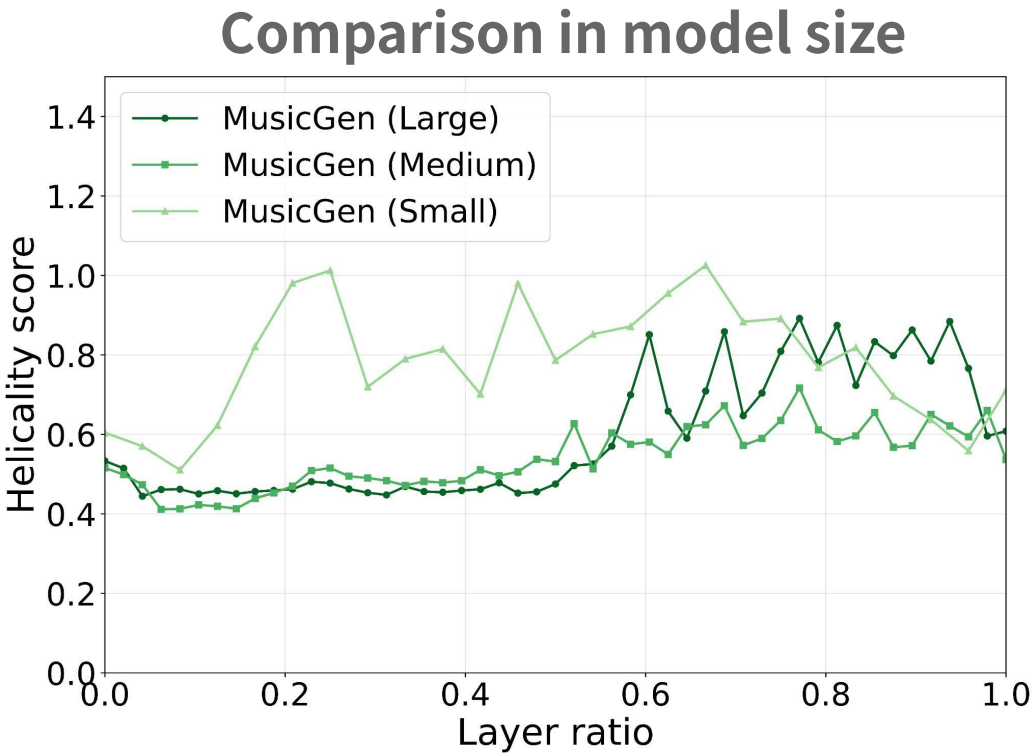
Low temporal model (“Top”) has the helix-specific layer.

Comparison in model size



Larger model (5B) develop layers in which pitch helix becomes clearer.

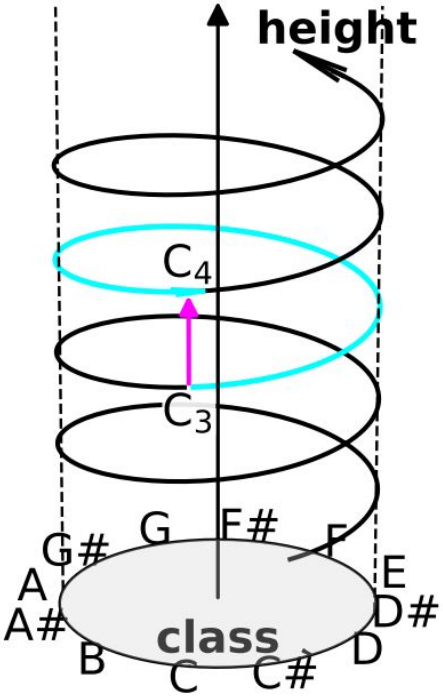
Results in MusicGen (non-hierarchical model)



“Larger model develop layers in which pitch helix becomes clearer” <- observed!

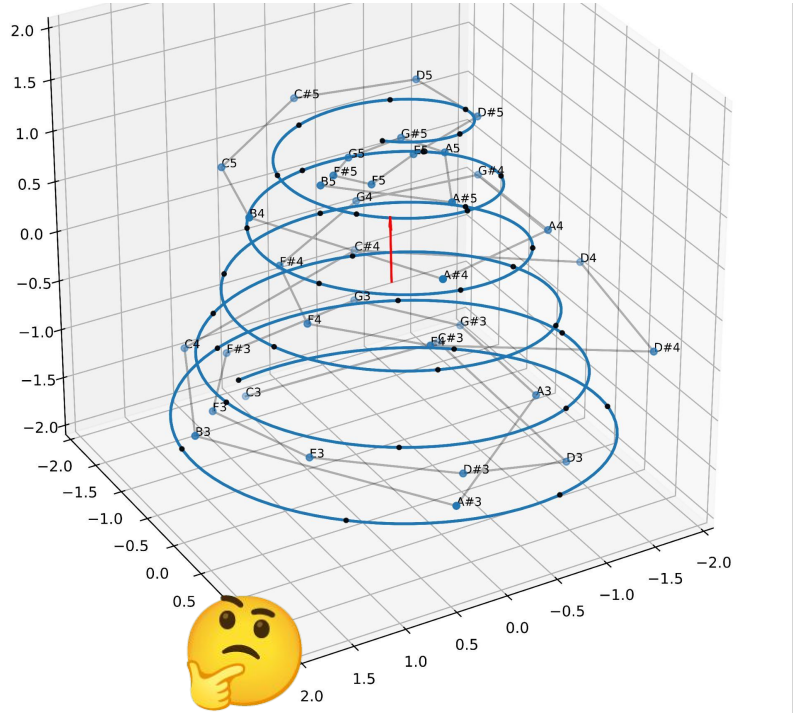
Compassion of perception models

Human (pitch helix)



Assuming **one** rotation per octave.

MusicGen



Observing **two** rotations per octave.

Future work

- No more “black box” in music processing.
  - Controllable/explainable music factors.
  - Handling unseen domain data.
    - Too high pitch, speed, tone, color, etc.

- Understanding human perception.
  - Pitch helix really holds true in pitch perception of human?
  - Aligning perception of humans and foundation models.

Conclusion

Conclusion

- Language of learned audio symbols
  - Learned speech symbols instead of hand-crafted features
  - The learned symbols follows the power law.
  - Parameter in the law correlates to speech quality.

- Geometry of music theory in music foundation models
  - Pitch helix, a pitch perception model for human.
  - Fitting helix-inspired function to data in foundation models.
  - The foundation models also follows the pitch helix.