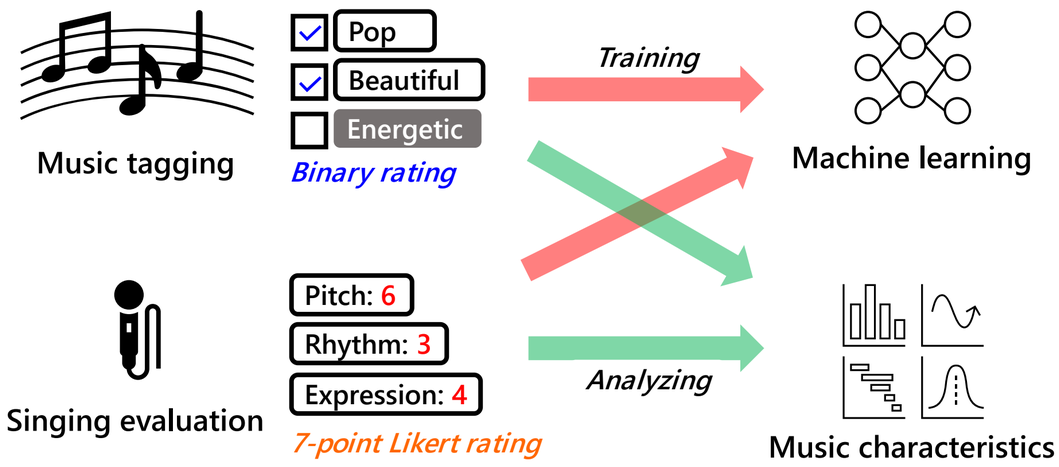
Using Item Response Theory to Aggregate Music Annotation Results of Multiple Annotators

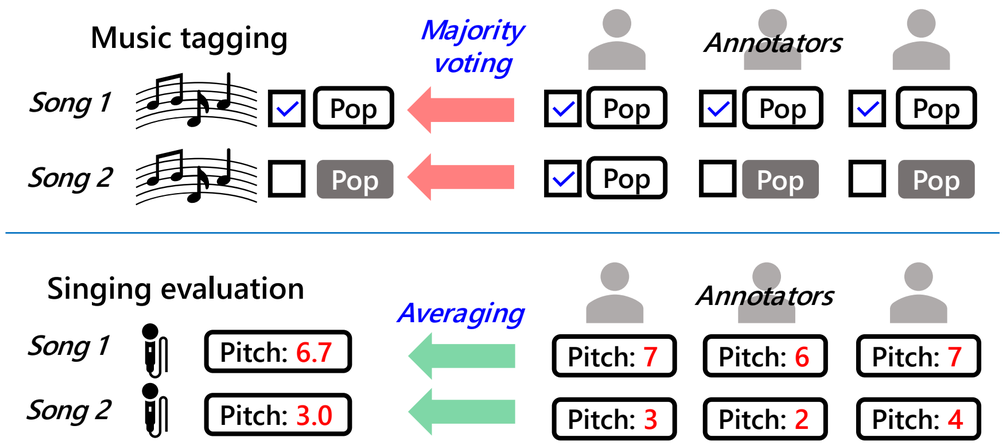
Tomoyasu Nakano Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan

Introduction

- Human music annotation is one of the most important tasks in music information retrieval (MIR)
  - ✓ For training machine learning models
  - ✓ For analyzing music characteristics

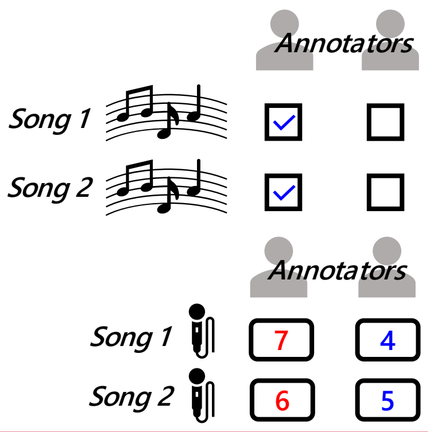


- A single target (e.g., a song or part of a song) is usually annotated by multiple human annotators
- The results are aggregated by majority voting or averaging in music annotation



Drawbacks

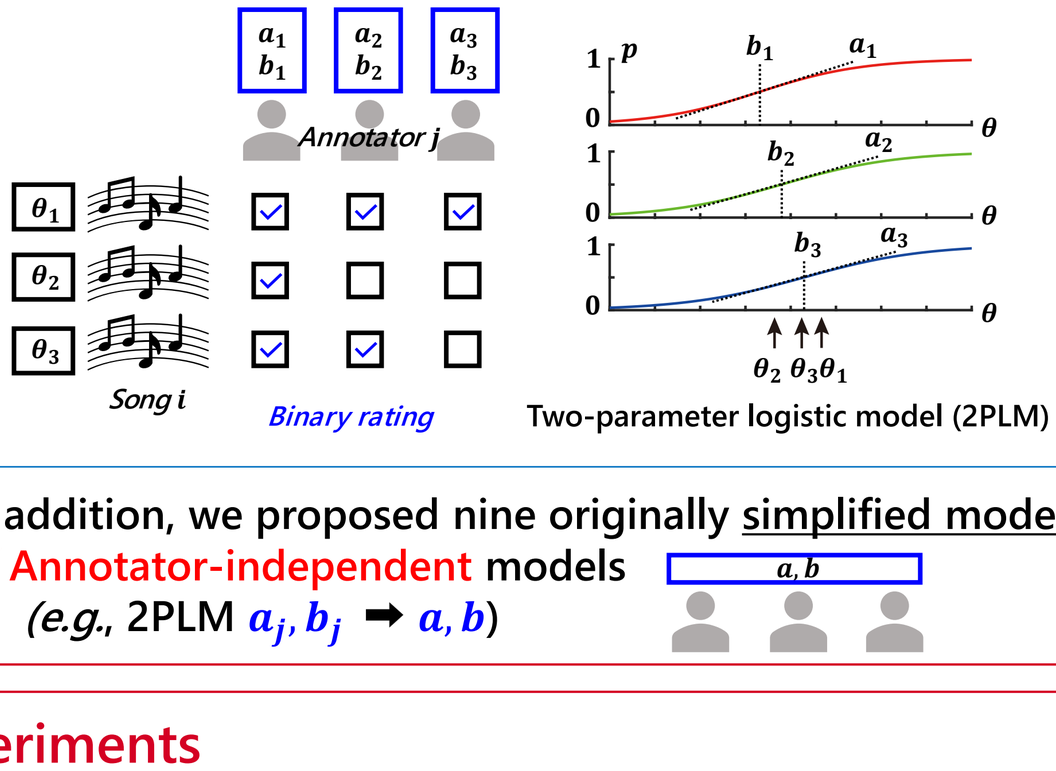
- Majority voting
  - ✓ Requires an odd number of annotators
  - ✓ The binarization loses information
- Averaging
  - ✓ Cannot be used for ordinal scale values
- Both methods
  - ✓ Cannot consider the differences in annotators' characteristics
  - ✓ There are differences in the thresholds for each annotator that determine whether a song is tagged or not, or which score is appropriate to rate the song.

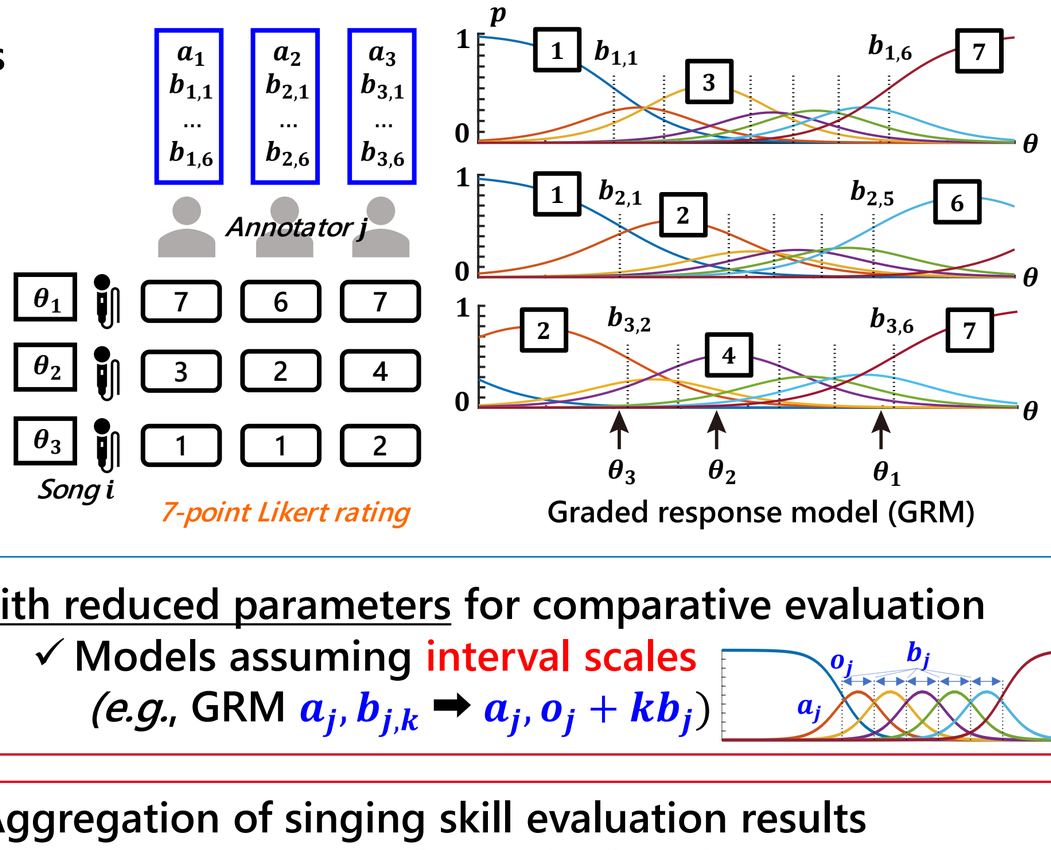


Proposed method

- Item Response Theory (IRT)-based music annotation aggregation [Lord, 1980]
  - ✓ It can model the annotators' characteristics
  - ✓ It can be used with any number of annotators
  - ✓ It can estimate latent continuous scores θ_i =aggregated annotation results
  - ✓ It can handle ordinal scale values

Well-known 3 IRT-based models
(1) Two-parameter logistic model (2PLM) [Birnbaum, 1968] [Hambleton, 1991]
(2) One-parameter logistic model (1PLM)
(3) Graded response model (GRM) [Samejima, 1969]





- In addition, we proposed nine originally simplified models with reduced parameters for comparative evaluation
  - ✓ Annotator-independent models (*e.g.*, 2PLM a_j, b_j ➡ a, b)

  - ✓ Models assuming interval scales (*e.g.*, GRM $a_j, b_{j,k}$ ➡ $a_j, o_j + kb_j$)

Experiments

- Aggregation of music tagging results
  - ✓ 6 annotators (3 males and 3 females)
  - ✓ 120 songs (each annotator tagged 60 songs)
  - ✓ 81 tags (15 genres, 38 subgenres, and 28 semantics)
  - ✓ 4 models
    - ✓ 2PLM a_j, b_j and 1PLM b_j
    - ✓ 2PLM' a, b and 1PLM' b (Annotator-independent models)
  - ✓ MCMC-based parameter estimation (NUTS) [Hoffman+, 2014]
  - ✓ ELPD-based model comparison (PSIS-LOO) [Vehtari+, 2017]
- Aggregation of singing skill evaluation results
  - ✓ 10 annotators (5 males and 5 females)
  - ✓ 140 songs
  - ✓ 7-point Likert scale from 6 evaluation perspectives (Pitch, Rhythm, Pronunciation, Expression, Vocal projection, and Overall performance)
  - ✓ 8 models
    - ✓ GRM $a_j, b_{j,k}$ and GRM-a $b_{j,k}$
    - ✓ GRM' a, b_k and GRM-a' b_k
    - ✓ GRMi $a_j, o_j + kb_j$ (GRM assuming interval scales)
    - ✓ GRMi-a $o_j + kb_j$, GRMi' $a, o + kb$, GRMi-a' $o + kb$
  - ✓ MCMC-based parameter estimation (NUTS)
  - ✓ ELPD-based model comparison (PSIS-LOO)
- Results
  - ✓ 1PLM > 1PLM' > 2PLM > 2PLM' (Annotator-dependent models)
  - ✓ GRMi > GRM' > GRM > GRMi' > GRM-a > GRM-a' > GRMi-a > GRMi-a' (Annotator-dependent models; Models assuming ordinal scale)

Contribution

- To the best of our knowledge, this is the first paper to introduce IRT in music annotation

Future direction

- We will verify the effectiveness of using θ as training data in machine learning