# TG-Critic: A Timbre-Guided Model for Reference-Independent Singing Evaluation

Xiaoheng Sun, Yuejie Gao, Hanyao Lin and Huaping Liu
NetEase Cloud Music | Fudan University

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

## 1. Introduction

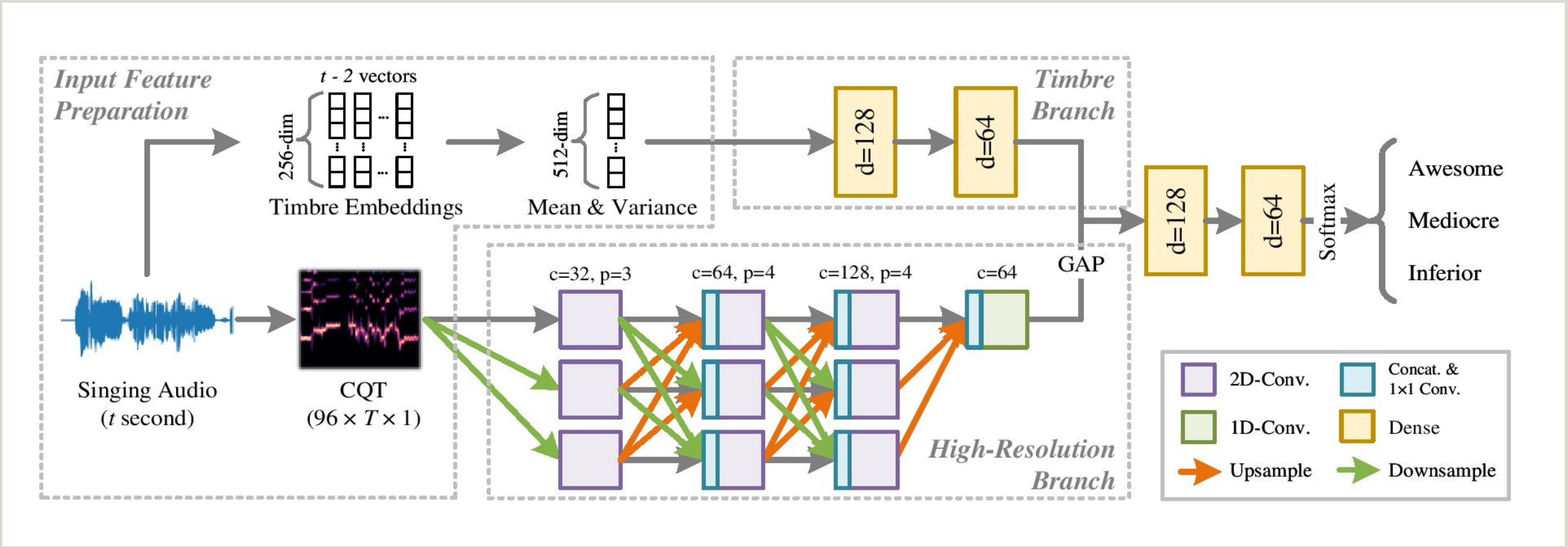
***What is Automatic Singing Evaluation?***

Automatic singing evaluation aims to assess the quality of singing performances without the participation of music experts, thus reducing manpower costs. Depending on whether a reference melody is required, the existing automatic singing evaluation systems can be roughly divided into two types:

- Reference-dependent approaches
- Reference-independent approaches

***Challenges in Automatic Singing Evaluation?***

Automatic singing evaluation independent of reference melody is a challenging task as the criteria are subjective and multi-dimensional. As an essential attribute of singing voices, vocal timbre has a non-negligible effect and influence on human perception of singing quality. But so far, no research has been done to include timbre information explicitly in singing evaluation models.



▲ The overall architecture of the proposed TG-Critic.

## 2. Approach

In this paper, we explore adding timbre embeddings as the model inputs and propose a **timbre-guided** singing evaluation model named TG-Critic:

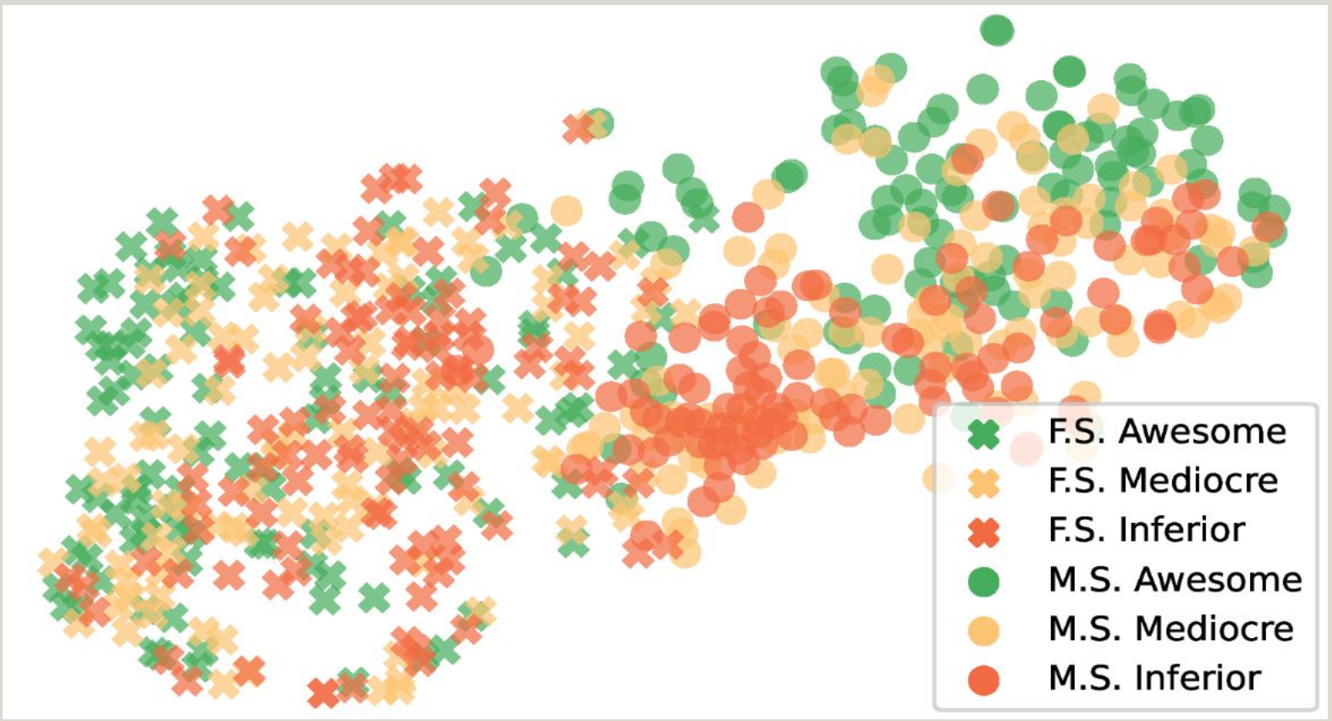
- In addition to CQT, timbre embeddings are introduced as one of the model inputs.
- The trunk structure of TG-Critic is designed as a multi-scale CNN-based network.
- An automatic annotation method is designed to construct a large three-class singing evaluation dataset with low manual cost.

### 2.1. Timbre Branch

A metric learning based embedding model designed for singer-relevant tasks (e.g., singer identification) is adopted to produce **timbre embeddings**. These embeddings are then further processed into 64 dimensions vectors by TG-Critic's Timbre Branch.

***Really helpful?***

- T-SNE shows that vectors of the same quality level are closer to each other than those of different levels.
- Even if only timbre embeddings are used as model inputs, an accuracy of 62% can still be achieved.

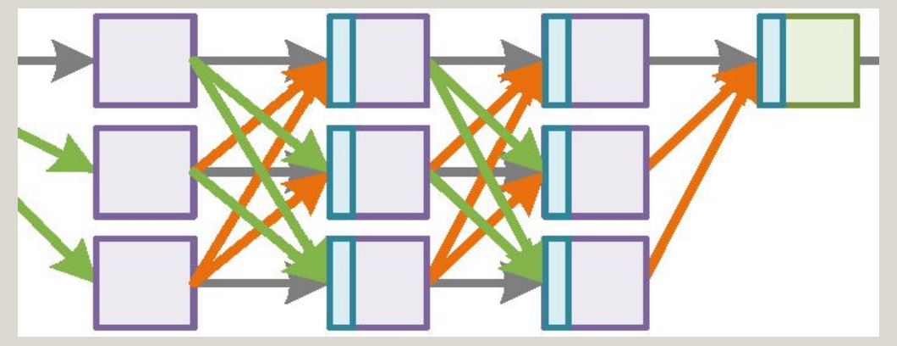


◂ The t-SNE visualization of timbre vectors derived from samples with different singing qualities.

### 2.2. High-Resolution Branch

We use CQT as the input mid-level feature. To better detect local patterns, a CNN- based structure is designed as the backbone of High- Resolution Branch. We also introduce a multi-scale structure to summarize the contextual information from features in a high-resolution way:

- Downsampling to expand the context range
- Retain high-resolution features to ensure detail patterns
- Rescaling & Merging to exchange information from different scales



◂ The multi-scale structure of YG-Critc's High-Resolution Branch

***And finally ...***

- Concatenate output vectors from two branches
- Produce classification results for singing quality: Awesome (A) - Mediocre(M) - Inferior(I)

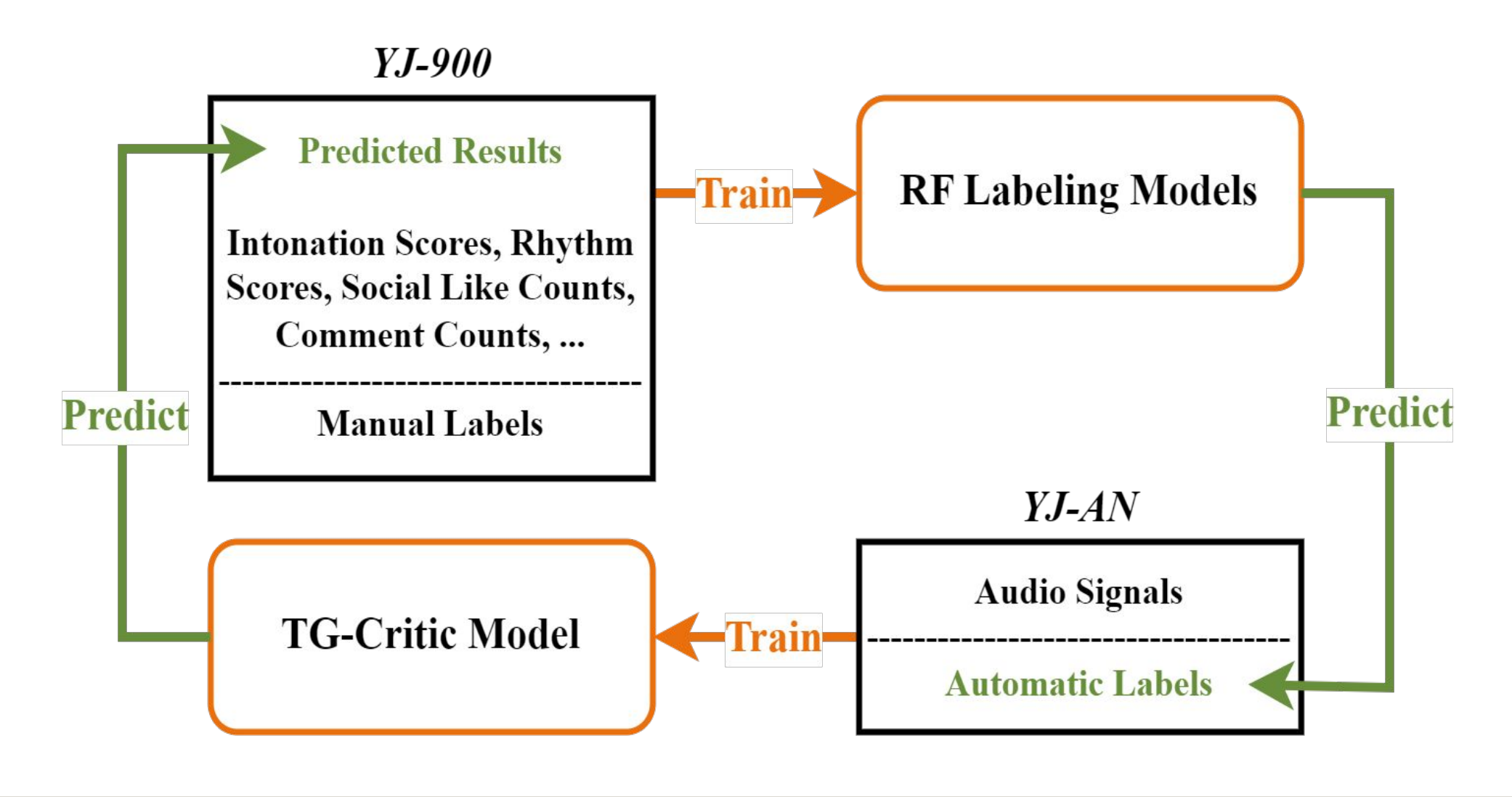
### 2.3. Automatic Annotation

***Dataset YJ-16K***

- Totally 32,623 unaccompanied singing pieces.
- YJ-900: 894 manually annotated samples.
- YJ-AN: 31729 automatically annotated samples.

***Iterative Automatic Annotation***

To alleviate the problem of insufficient data, we propose an iterative automatic annotation method using metadata and predicted results from last iteration. ▼



## 3. Evaluation

| Model | Precision (%) | | | Recall (%) | | | Acc. |
|---|---|---|---|---|---|---|---|
| | A | M | I | A | M | I | (%) |
| *TG-Critic-1S* | 83.5 | 71.6 | 84.0 | 90.5 | 69.2 | 79.7 | 79.8 |
| *TG-Critic-2S* | **87.2** | **73.6** | 86.7 | 89.9 | **75.5** | 81.8 | **82.3** |
| *CQT-Only* | 84.3 | 69.8 | 79.5 | 88.9 | 63.6 | **82.4** | 78.2 |
| *TG-Simple* | 82.1 | 68.4 | **88.7** | **92.9** | 72.5 | 71.6 | 79.0 |

***◂ Ablation Study***

- For the proposed TG-Critic, two models are trained by different training strategies:
  - ○ **TG-Critic-1S**: The High-Resolution Branch and the Timbre-Branch are trained together in one step;
  - ○ **TG-Critic-2S**: The High-Resolution Branch is first trained & frozen, and then the Timbre Branch is trained.
- **CQT-Only**: Remove the Timbre Branch.
- **TG-Simple**: Replace the High-Resolution Branch with a simple CNN structure

***Comparison with Previous Works ▸***

- To compare with previous works, we reproduce three baseline models. All three models are reference-independent singing evaluation models.
- In addition to YJ-900 (894 samples), we use two public datasets PESnQ-DS (20 samples) and NUS48E (48 samples) for tests.
- To make a comprehensive comparison, we obtain a weighted score for each prediction using the output probability distribution.

| Model | Param. | YJ-900 | PESnQ-DS | | NUS48E | |
|---|---|---|---|---|---|---|
| | | Acc. | Acc. | Corr. | Acc. | Corr. |
| *Kuaishou* [15] | 1.97M | 68.3 | 85.0 | 0.858 | 68.8 | 0.497 |
| *NUS20* [17] | 0.72M | 76.3 | 85.0 | 0.930 | 68.8 | 0.552 |
| *NUS21* [18] | 1.45M | 78.4 | 85.0 | 0.925 | 72.9 | 0.548 |
| *TG-Critic-1S* | 0.82M | 79.8 | 80.0 | 0.927 | 72.9 | **0.671** |
| *TG-Critic-2S* | | **82.3** | **95.0** | **0.933** | **77.1** | 0.631 |

## 4.Conclusion

In this paper, we have proposed TG-Critic, a timbre-guided singing evaluation model independent of the reference melody. The proposed model includes timbre information explicitly by using timbre embedding as one of the model inputs. A multi-scale structure is introduced to process the CQT features in a high-resolution way. We also construct a large singing dataset YJ-16K with annotations labeled by an iterative automatic annotation method. Experimental results show the proposed model outperforms the existing state-of-the-art models in most cases.

Our team ▼

For further results ▼