Perceptual Evaluation of Singing Quality (PESnQ)

Chitralekha Gupta[1,2], Haizhou Li[3], and Ye Wang[1]

chitralekha@u.nus.edu, haizhou.li@nus.edu.sg, wangye@comp.nus.edu.sg

[1]School of Computing, [2]NUS Graduate School for Integrative Sciences and Engineering, [3]Department of Electrical and Computer Engineering, National University of Singapore

1. Introduction

- **Singing pedagogy** is dependent on human music experts, and is not always accessible to the masses
- A **perceptually-valid automatic singing evaluation score** could serve as a complement to singing lessons, and make singing training more accessible to learners

2. How do experts perceptually evaluate singing quality?

- Rhythm Consistency
- Intonation Accuracy
- Appropriate Vibrato
- Voice Quality
- Pitch Dynamic Range

Reference 

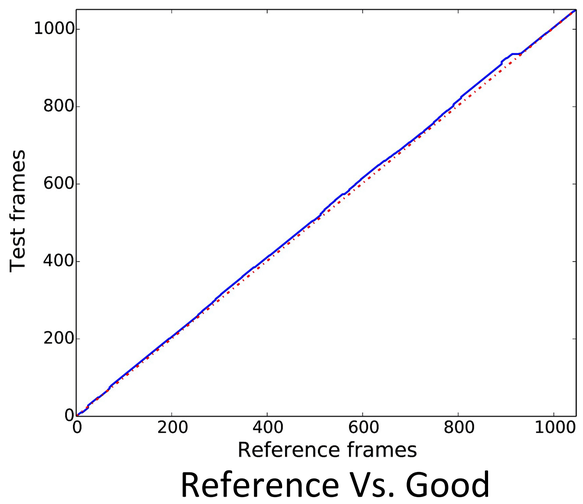
Good 

Poor 

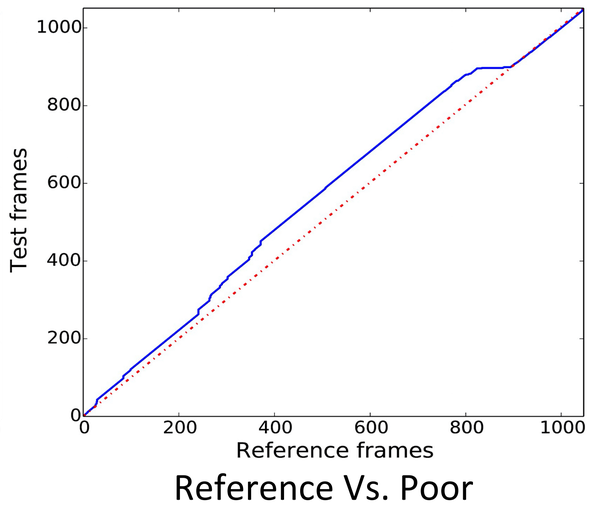
3. Objective Characterization of Singing Quality

Rhythm Consistency

- Use **DTW** of **MFCC vectors** between **frame-equalized** reference and test. Uniformly faster or slower tempo shouldn't be penalized



Reference Vs. Good



Reference Vs. Poor

Intonation Accuracy

- Compare post-processed **pitch contours** from rhythm-aligned reference and test
- **Key transposition** should be allowed → **pitch derivative**, and **median-subtracted pitch**

Appropriate Vibrato

- Vibrato oscillations: **Rate**: 5-8 Hz; **Extent**: 30-150 cents
- Features: vibrato likeliness, rate, extent

Voice Quality and Pronunciation

- DTW distance between MFCC feature vectors

Pitch Dynamic Range

- Comparison of difference between min and max pitch values

Disturbance Features

- **Frame-level deviation** of the optimal path from the diagonal in DTW for rhythm and intonation features

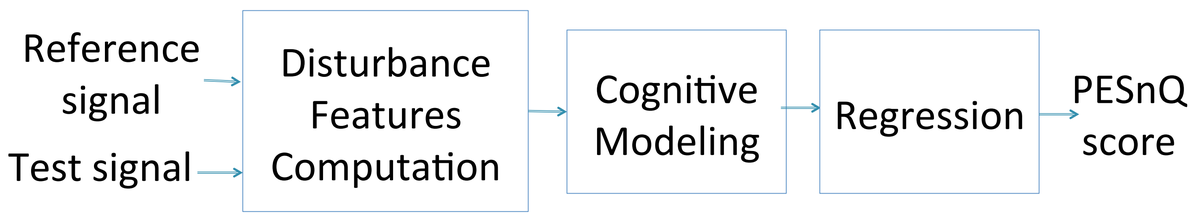
4. PESQ-based Feature Modeling

Combine frame-disturbances of these features with cognitive modeling inspired by telecommunication standard PESQ [Rix2001]:

a localized error in time has a larger subjective impact than a distributed error

- *Localized error*: L6-norm over split second intervals (320ms)
- *Distributed error*: L2-norm over all split second intervals

5. PESnQ Formulation

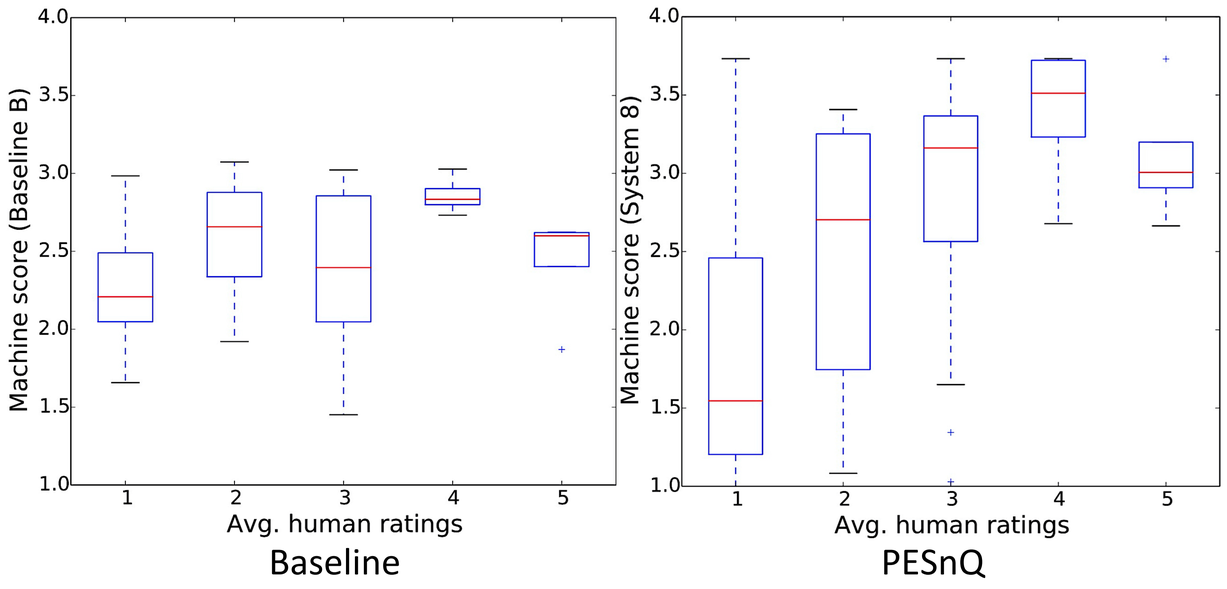


Experimental Dataset

- 20 audio recordings collected from 20 singers with varied singing abilities – professional to poor
- Subjective evaluation for singing quality by 5 professionally trained musicians – inter-judge agreement was 0.82

System	Description
Baselines	Pitch distance [Tsai2012], pitch-aligned rhythm distance [Molina2013], volume distance [Chang2007, Tsai2012]
PESnQ systems	Combinations of L2-norm, L6+L2-norm and distance features for the various MFCC-aligned perceptual features

6. Results



Baseline

PESnQ

System	Correlation objective score with avg. overall human score	Leave-one-judge-out avg. correlation score
Human Judge	–	0.87
Baseline	0.30	0.38
PESnQ	0.59	0.66

7. Conclusions

- We propose **perceptually relevant features** to objectively **evaluate singing quality**
- We adopt **the cognitive modeling theory** of **PESQ** to design a **PESnQ** score which performs better than distance features
- PESnQ shows **96%** improvement over baseline scores in correlating with the music-expert human judges

Sound & Music Computing Lab
http://www.smcnus.org/

NUS National University of Singapore | School of Computing

NUS Graduate School for Integrative Sciences and Engineering
National University of Singapore