Evaluating the Suitability of SoundFingerprinting for Music Identification in Movies

Researcher: Tim Huisman
t.huisman-1@student.tudelft.nl
Supervisors: Dr. Cynthia Liem
Dr. Jaehun Kim
Date: 02-07-2021

1 - SoundFingerprinting

- Open-source audio fingerprinting framework implemented in C#
- Fingerprints overlapping fragments of songs, robust to position uncertainty in time
- Uses 'Haar-wavelets' to extract characterizing frequencies from spectrograms
- Only selecting wavelets with highest magnitude makes the framework robust to small changes in sound

2 - Research Questions

- How does SoundFingerprinting perform according to a benchmark established for this application?
- What configurable parameters influence the performance of the framework for music identification in movies?

3 - Configurable parameters

Default config

- Fingerprints 318 Hz-2000Hz
- Minimum 'votes' of 4

HighPrecision config

- Fingerprints 1500 Hz-2500Hz
- Minimum 'votes' of 3

Match confidence threshold (0-1)

- Discards low confidence matches, avoiding false positives
- Developer suggests 0.15, likely not suitable for this field

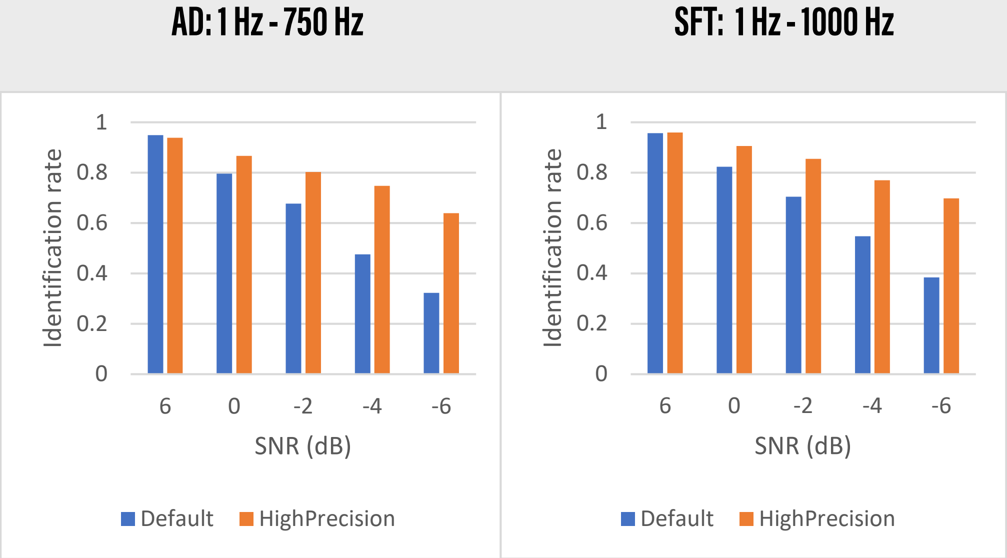
4 - Method

- Evaluate both configurations on each category in the benchmark separately to identify problematic areas and determine most suitable configuration
- Relativize performance to overlap in noise frequencies and config frequency range to explore correlation
- Calculate metrics for various thresholds in order to explore a more suitable confidence threshold

5 - Results

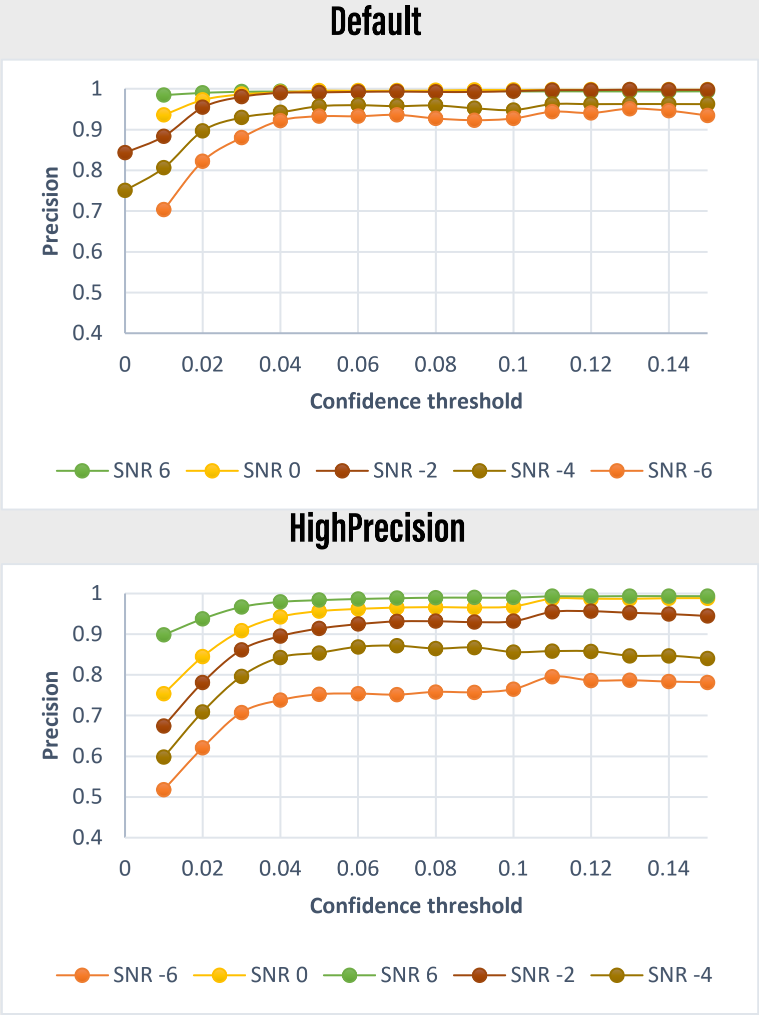
Performance analysis based on dominant frequencies

- In general, noise categories with dominant frequencies having large overlap with config frequency range proved more challenging
- An example for which the differing frequency ranges showed a clear performance difference is AD (Ambient Dining) and SFT (Speech Female Talking)



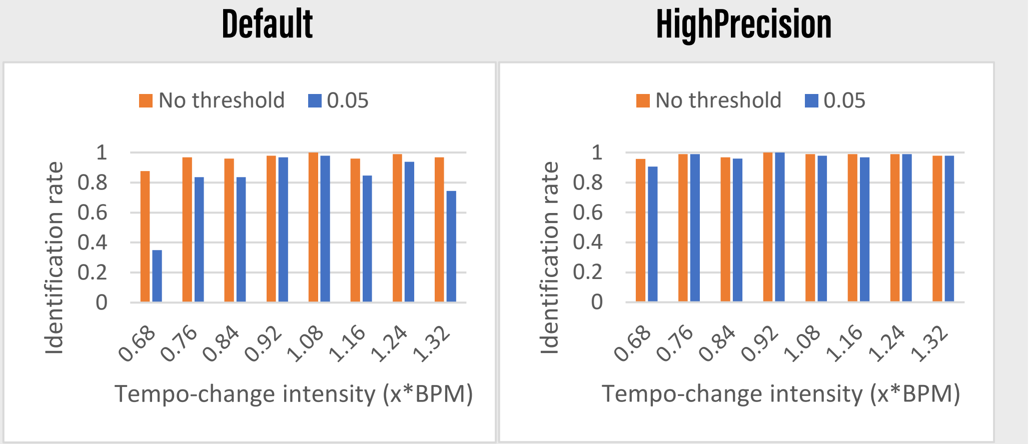
Establishing a suitable threshold

- Aim: suggest threshold for which performance only decreases at higher thresholds
- For both configurations, this held for all SNRs at 0.05

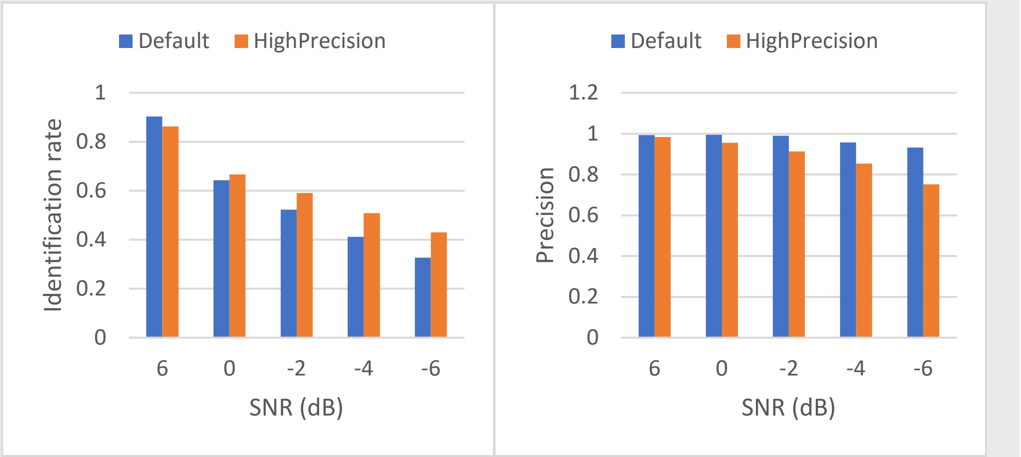


Performance on structural alterations

- Only 5% of pitch-shifted audio queries were identified correctly.
- Framework is robust to tempo-changes



Performance with 0.05 threshold



6 - Conclusions

- Correlation between identification rate, noise frequencies and frequency ranges indicate possible 'ideal frequency range', but to establish this, the evaluation data set should be expanded to cover more noise categories
- Suggested threshold by developer is not suitable, new threshold improves identification rate on average by 0.13
- As tempo-changes are not influenced by frequency range, lower minimum 'votes' appears to cause robustness
- As parameters are not evaluated individually, we do not know the effect of the individual parameters on identification for sure
- HighPrecision most suitable for this evaluation due to better performance on negative SNRs and tempo-changed audio.
- However, this data is not representative of real movie audio, as movie audio contains combinations of degradation categories. On actual movie data, SoundFingerprinting identified 5% of query clips. Therefore, it is currently unfit for this task.