

Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching

Colin Raffel
May 16, 2016

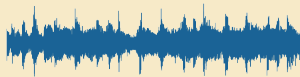
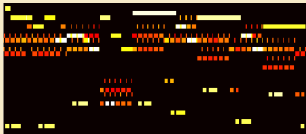


IGERT Integrative Graduate
Education and Research Traineeship

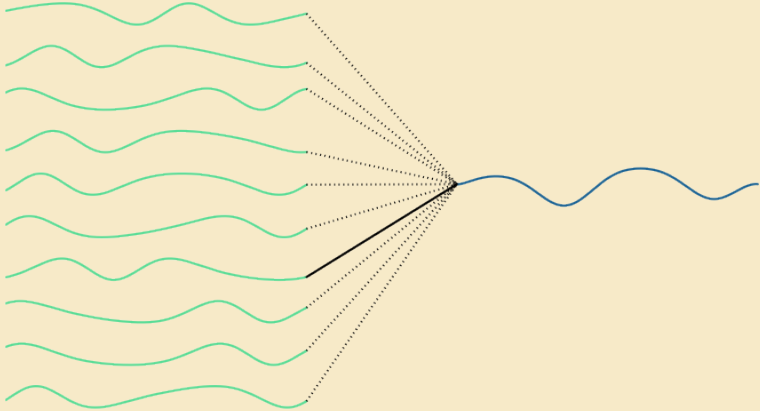


The Goal

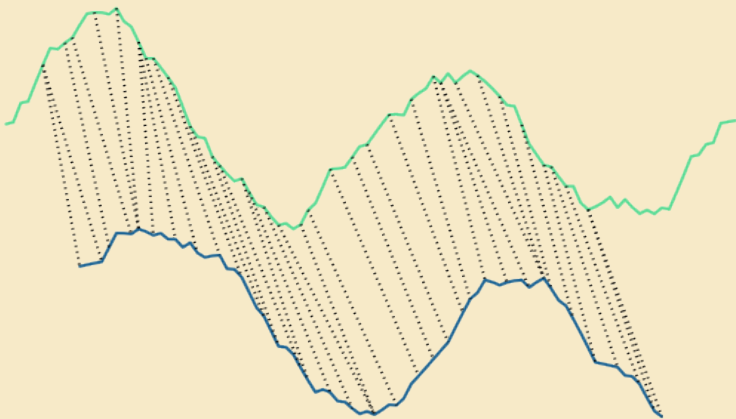
```
artist: 'Tori Amos'  
release: 'LIVE AT MONTREUX'  
title: 'Smells Like Teen Spirit'  
id: 'TRKUYPW128F92E1FC0'  
duration: 216.4502  
sample_rate: 22050  
audio_md5: '8'  
7digitalid: 5764727  
year: 1992
```



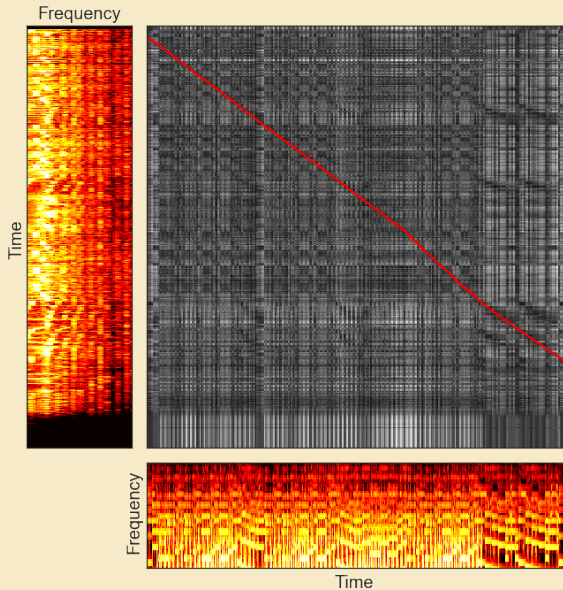
Sequence Matching



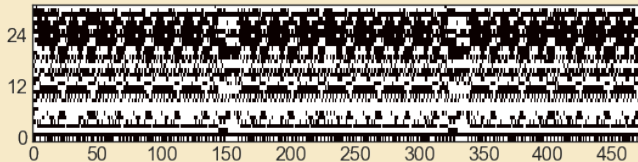
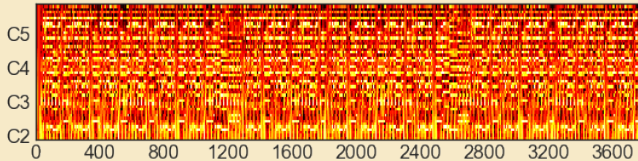
Dynamic Time Warping



Comparing MIDIIs with DTW

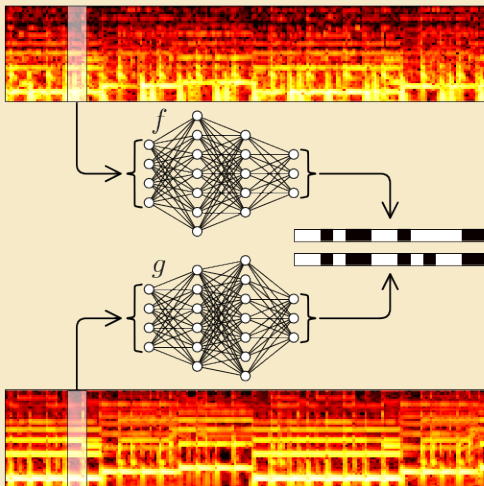


Downsampled Hash Sequences

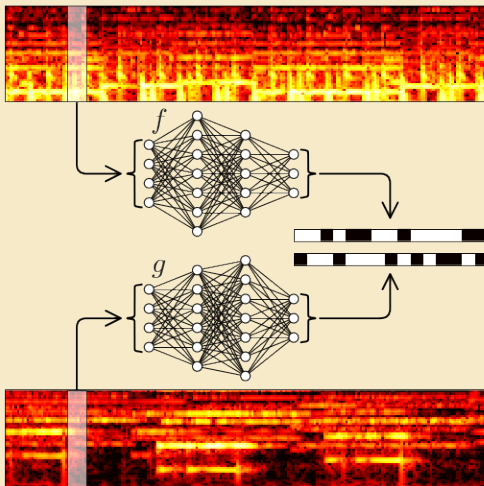


A diagram illustrating the POPCNT (Population Count) operation. It shows three horizontal bars representing binary sequences. The top bar has 10 segments, with 5 segments being black (representing 1s) and 5 being white (representing 0s). The middle bar has 10 segments, with 4 segments being black and 6 being white. The bottom bar has 10 segments, with 9 segments being black and 1 being white. A circled plus sign (\oplus) is positioned to the left of the middle bar, indicating a bitwise XOR operation between the top and middle bars. The equation below the bars is:
$$\text{POPCNT}(\text{resulting sequence}) = 9$$

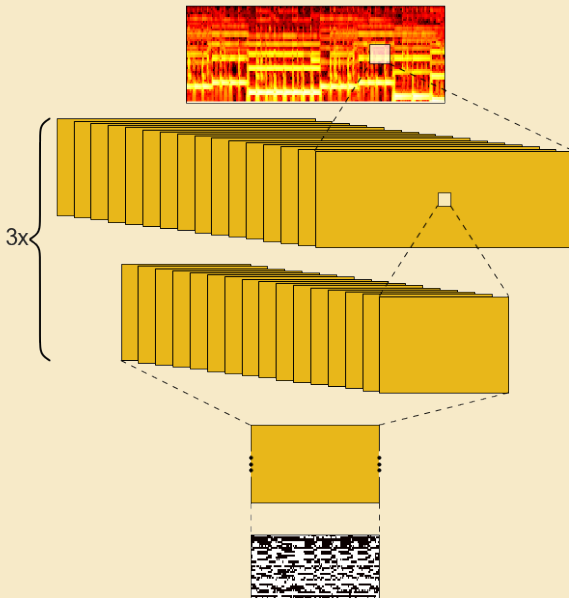
Similarity-Preserving Hashing



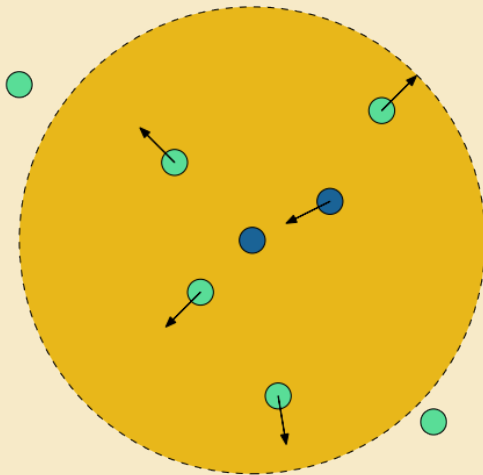
Similarity-Preserving Hashing



Network Structure

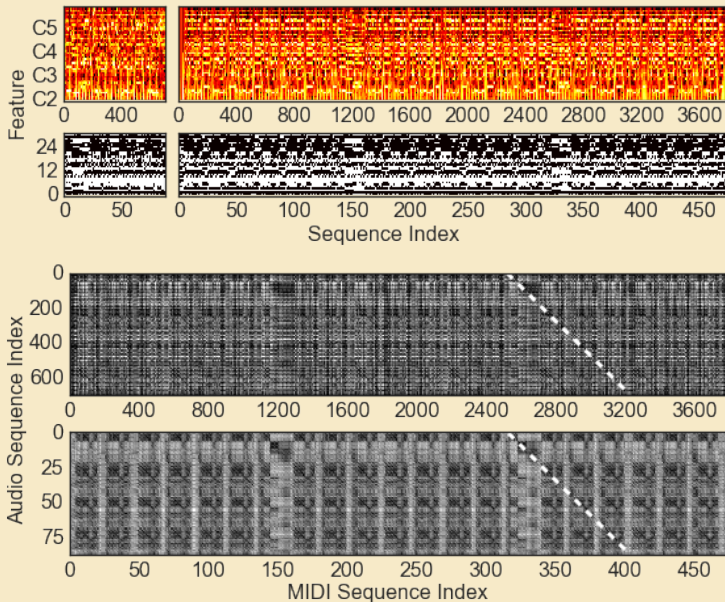


Loss Function

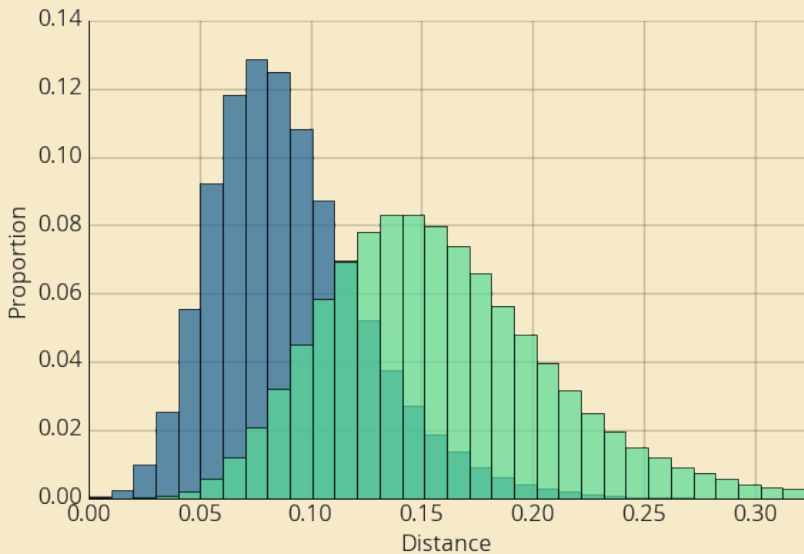


$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y) \in \mathcal{P}} \|f(x) - g(y)\|_2^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} \max(0, m - \|f(x) - g(y)\|_2)^2$$

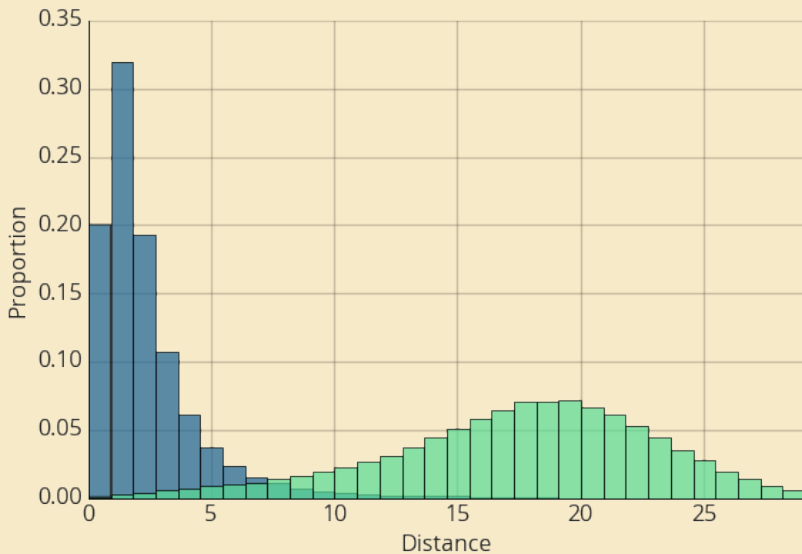
Example Output



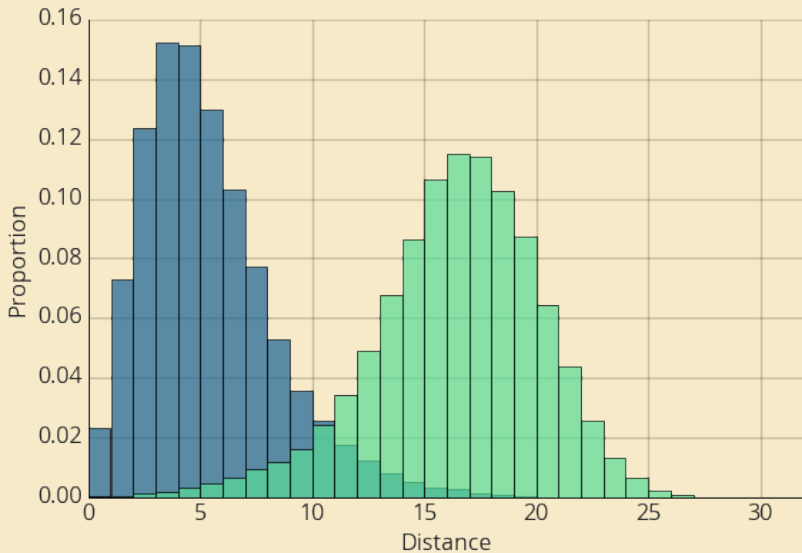
Raw Distance Distributions



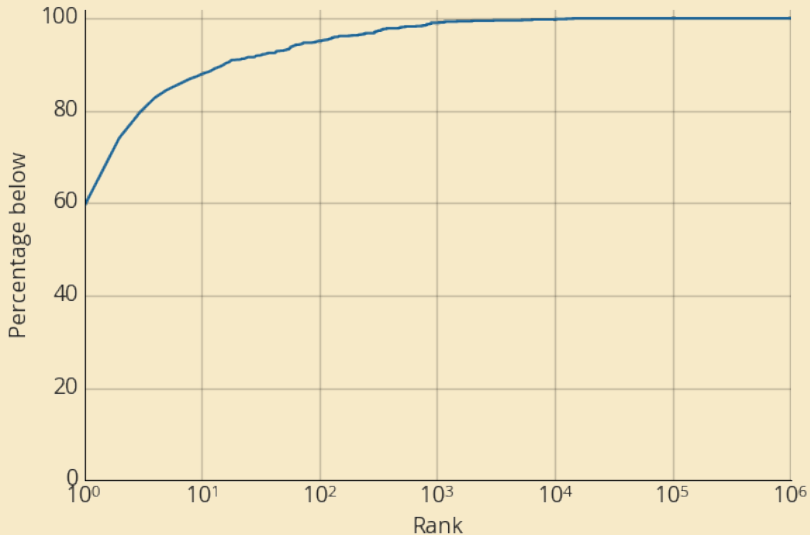
Output Distance Distributions



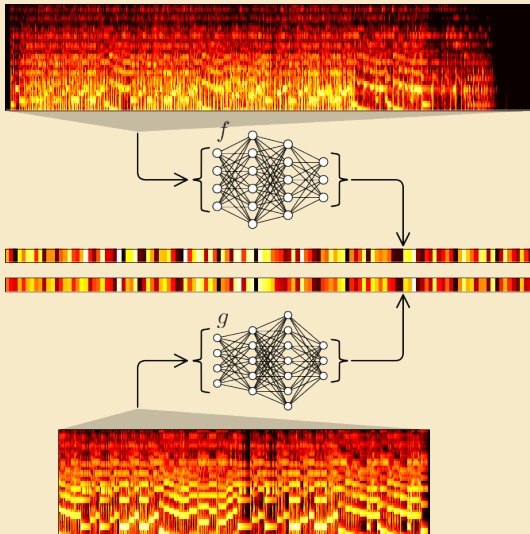
Hash Distance Distributions



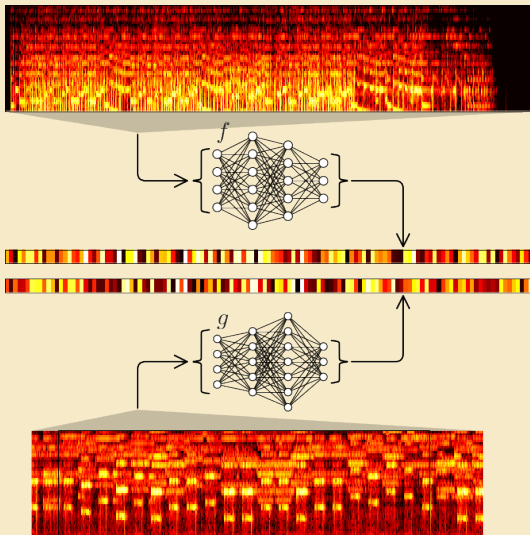
Match Ranks



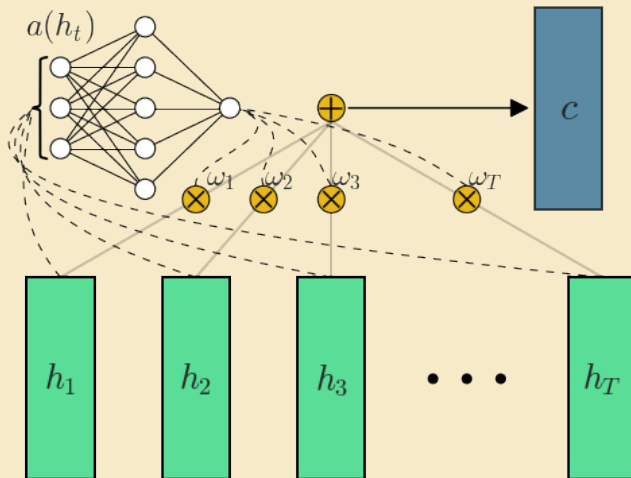
Pairwise Sequence Embedding



Pairwise Sequence Embedding

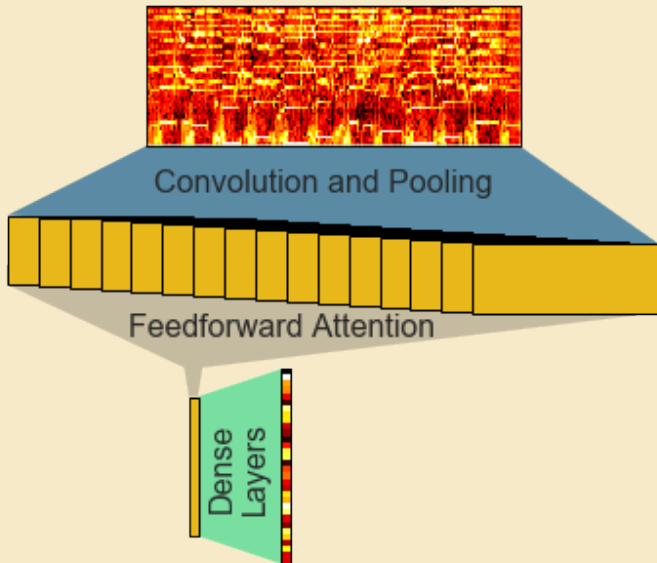


Feed-Forward Attention

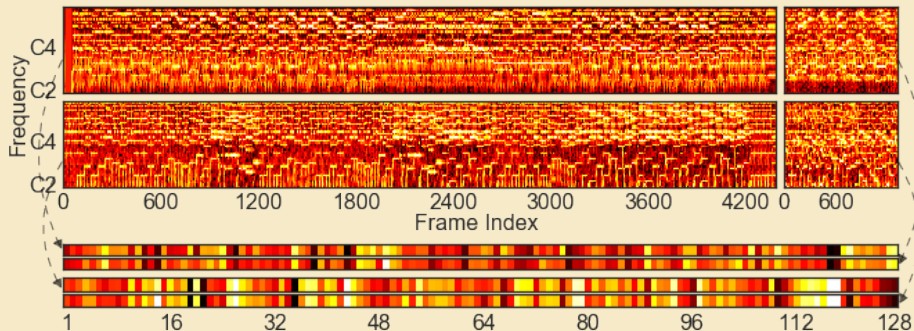


Raffel & Ellis, "Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems", ICLR 2016

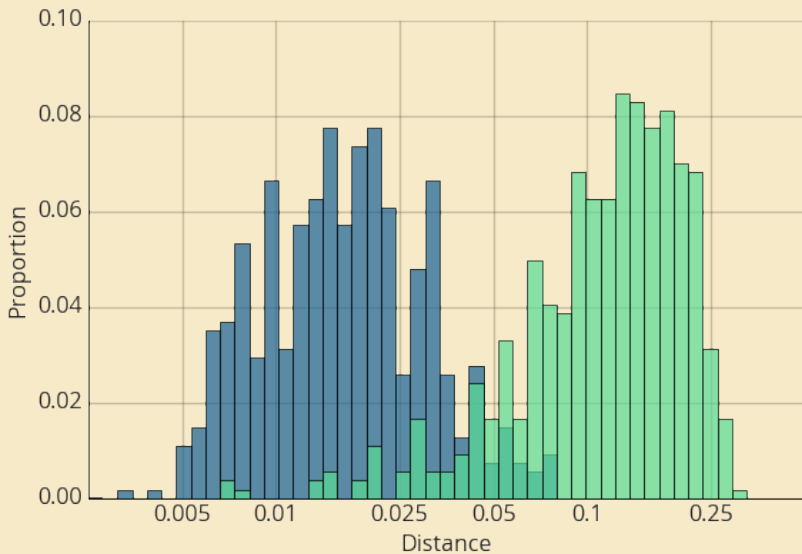
Embedding Network



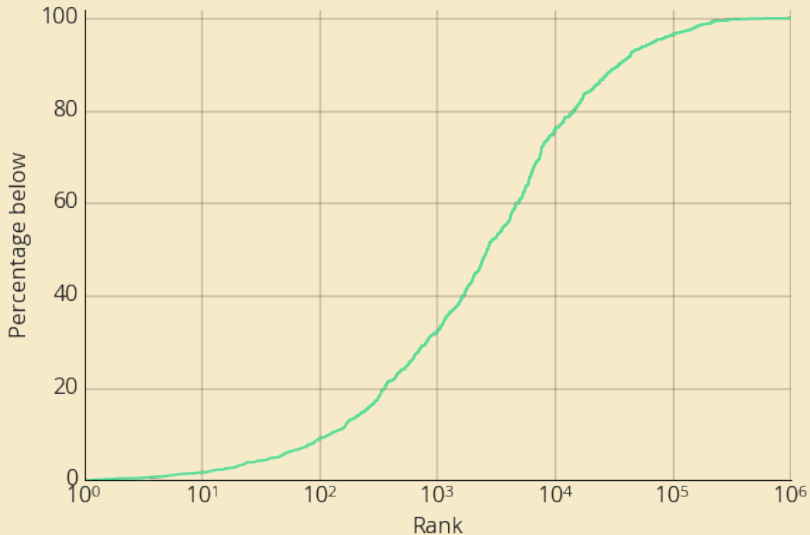
Example Embeddings



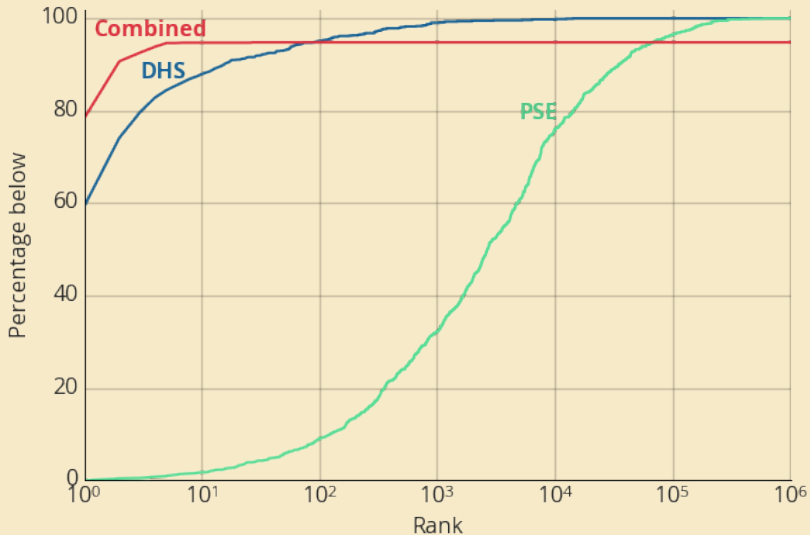
Embedding Distances



Match Ranks



Combined Match Ranks



References

- [1] Raffel, “Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching”, PhD Thesis
- [2] Raffel & Ellis, “Large-Scale Content-Based Matching of MIDI and Audio Files”, ISMIR 2015
- [3] Raffel & Ellis, “Pruning Subsequence Search with Attention-Based Embedding”, ICASSP 2016