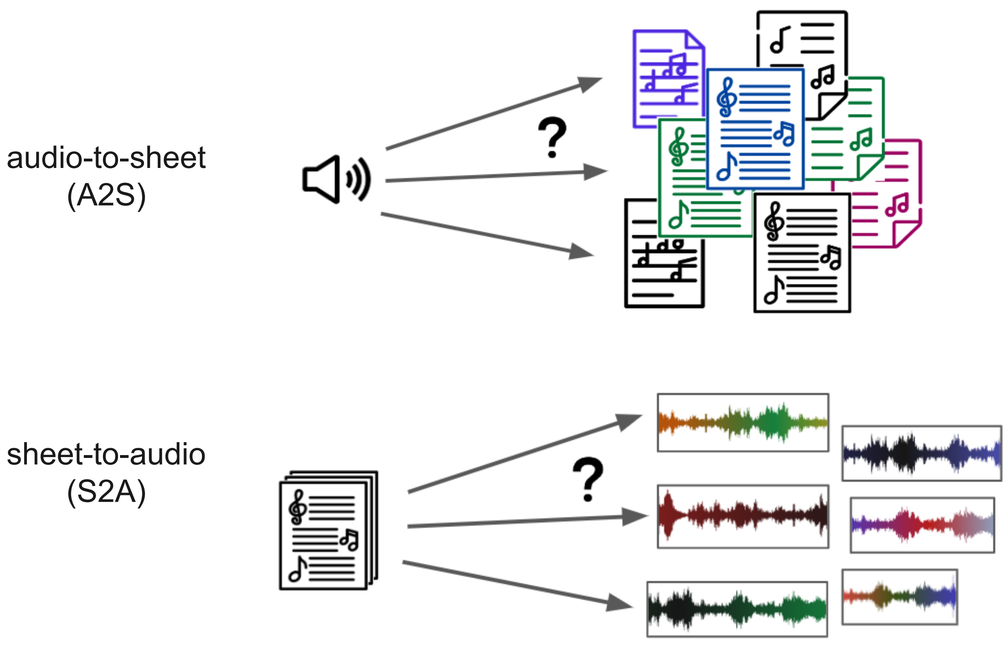
PASSAGE SUMMARIZATION WITH RECURRENT MODELS FOR AUDIO-SHEET MUSIC RETRIEVAL

Luís Carvalho[1] and Gerhard Widmer[1,2]
[1]Institute of Computational Perception (CP-JKU), [2]LIT Artificial Intelligence Lab
JOHANNES KEPLER UNIVERSITY LINZ
Institute of Computational Perception

Task and Motivation

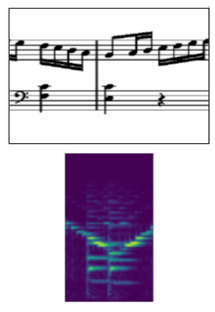
Audio-Sheet Music Retrieval (ASR):

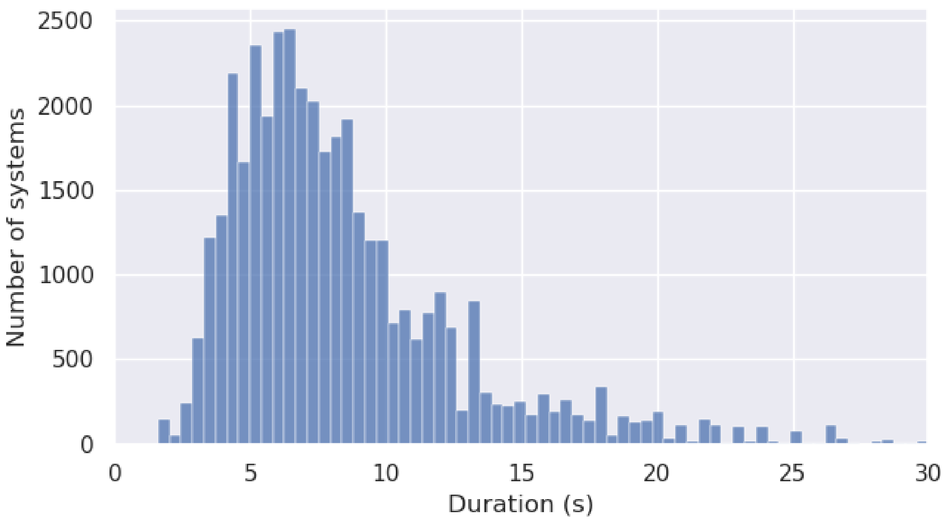
- find the appropriate recording(s) from a database given a sheet music query, and vice versa
- raw data: audio recordings and scanned sheet music
- target: classical piano music



Limitations of deep-learning ASR methods:

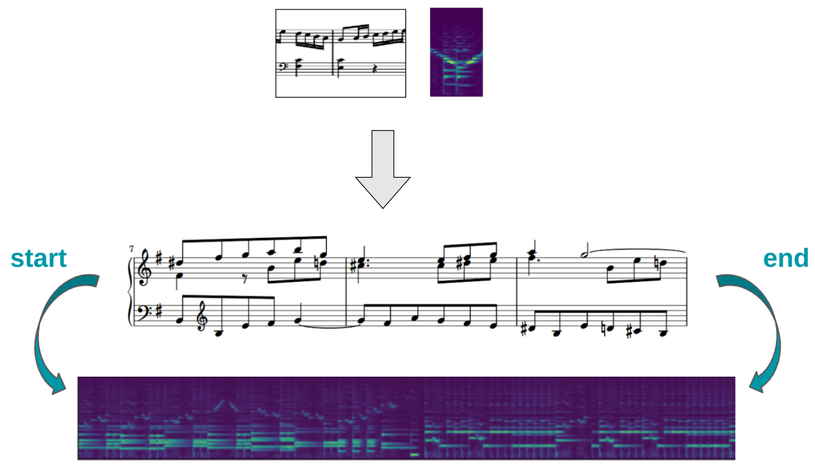
- fine-grained annotated data for training
- Noteheads aligned to onsets
- we rely on synthetic data
- training samples are fixed in size
- different musical content in sample pairs
- global and local tempo variations





Our Approach and Proposed RNN Model

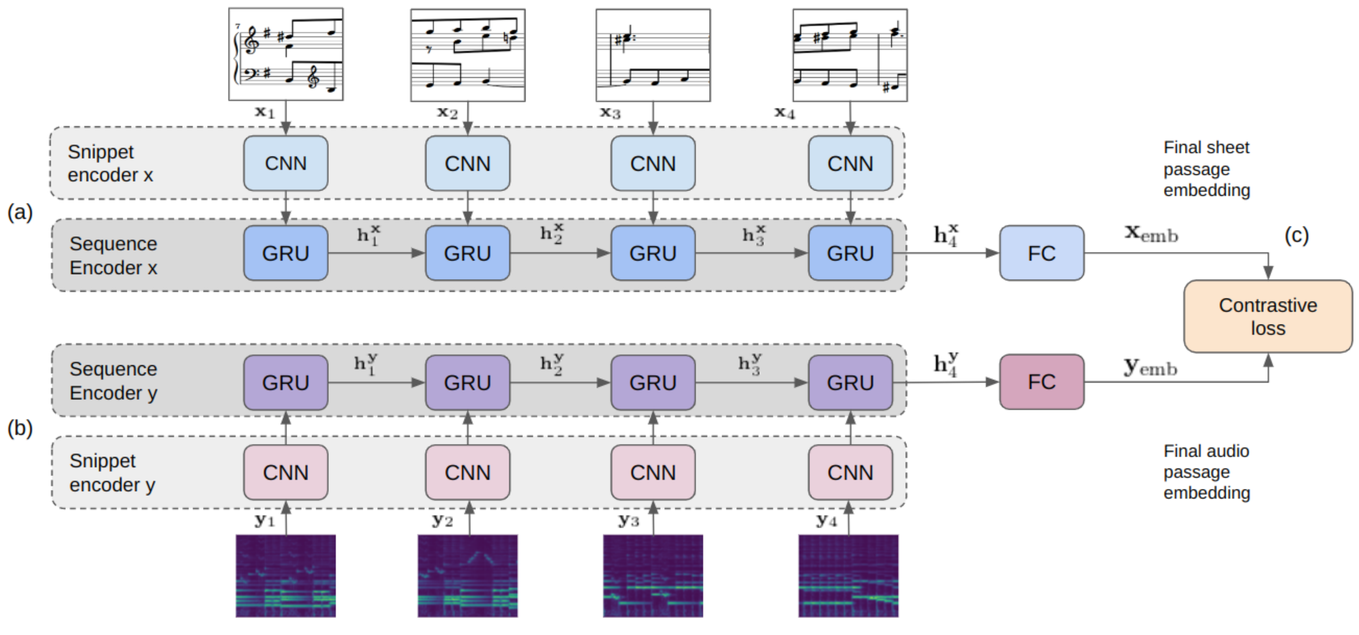
- learn cross-modal representations from longer and variable-length passages of music



Benefits:

- weaker-level mappings
- handle the temporal mismatches

- encode a pair of matching sequences of fixed-size snippets using GRUs

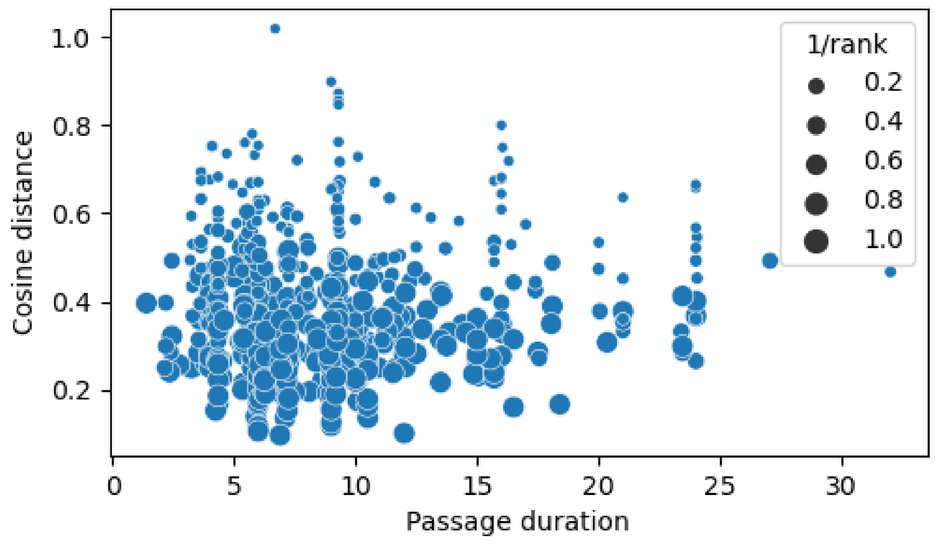


Experiments and Results

- passage retrieval evaluation

	Audio-to-Score (A2S)					Score-to-Audio (S2A)				
	R@1	R@10	R@25	MRR	MR	R@1	R@10	R@25	MRR	MR
I MSMD (Fully synthetic)										
BL [1]	47.56	81.68	90.80	0.592	1	51.37	83.51	92.59	0.628	1
RNN	51.12	84.46	92.88	0.627	1	54.30	85.95	94.94	0.670	1
RNN-FT	55.27	87.98	95.02	0.651	1	56.32	87.12	96.44	0.697	1
RNN-FT-CCA	**60.04**	**89.66**	**97.73**	**0.692**	**1**	**62.11**	**91.44**	**98.41**	**0.734**	**1**
II RealScores_Synth (Sheet music scans and synthetic recordings)										
BL [1]	20.19	55.47	74.99	0.343	7	25.15	70.27	83.11	0.391	5
RNN	25.09	61.24	78.27	0.374	5	30.15	72.47	86.89	0.439	3
RNN-FT	28.87	66.41	81.32	0.447	4	33.98	75.47	88.51	0.462	2
RNN-FT-CCA	**33.36**	**69.49**	**83.88**	**0.481**	**3**	**37.35**	**79.22**	**89.95**	**0.538**	**1**
III RealScores_Rec (Sheet music scans and real recordings)										
BL [1]	15.67	31.46	48.12	0.226	29	18.30	36.71	54.94	0.266	18
RNN	19.11	35.98	53.65	0.278	21	22.76	39.95	57.47	0.303	15
RNN-FT	22.39	39.53	57.19	0.338	18	26.76	42.77	59.38	0.371	7
RNN-FT-CCA	**26.62**	**44.81**	**60.01**	**0.362**	**7**	**29.84**	**46.71**	**60.88**	**0.435**	**4**

- embedding space analysis: pairwise cosine distances



[1] M. Dorfer, J. Hajic jr., A. Arzt, H. Frostel, and G. Widmer. Learning audio-sheet music correspondences for cross-modal retrieval and piece identification. TISMIR, 1(1), 2018.

contact email: luis.carvalho@jku.at

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No 101019375 (Whither Music?) and the Federal State of Upper Austria (LIT AI Lab).