Matching Text and Audio Embeddings: Exploring Transfer-learning Strategies for Language-based Audio Retrieval

Benno Weck ♯♭

Miguel Perez ♯♭

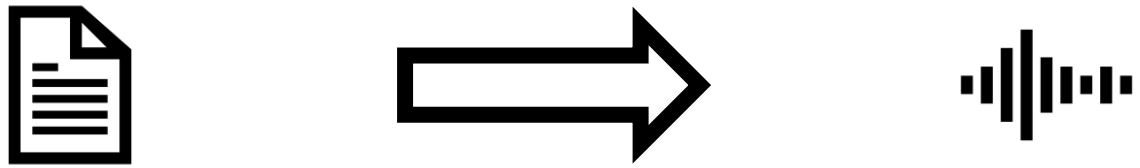
Holger Kirchhoff ♯

Xavier Serra ♭

HUAWEI ♯ upf. Universitat Pompeu Fabra Barcelona ♭

Language-based Audio Retrieval

Task: Rank recordings according to a given text query



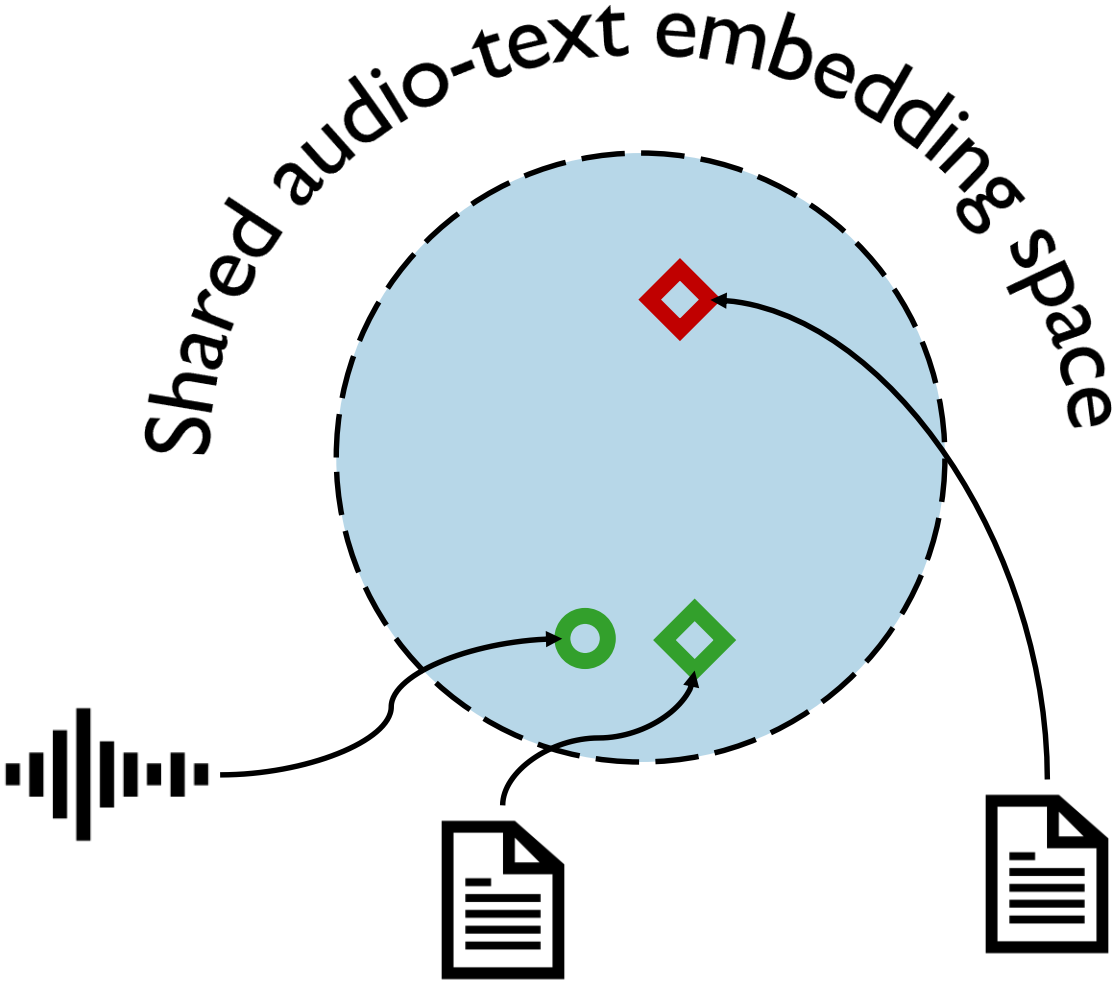
„A dog barks and then a baby cries“

1. *dog_barks.wav*
2. *cat_meow.wav*
3. *train_pass_by.wav*
4. *...*

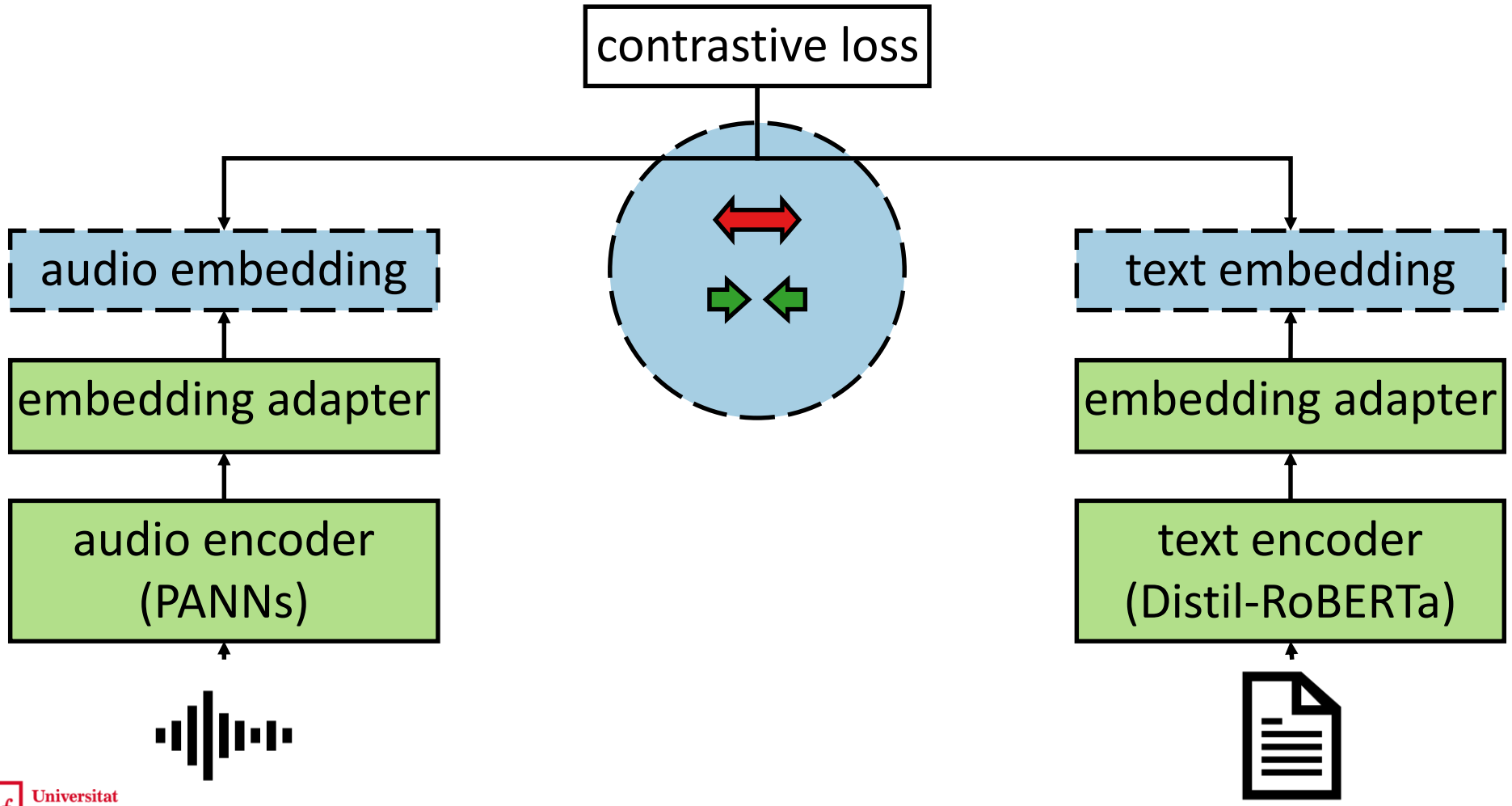
HUAWEI

upf. Universitat Pompeu Fabra Barcelona

Language-based Audio Retrieval



Challenge submissions



HUAWEI

upf. Universitat Pompeu Fabra Barcelona

Challenge submissions

	Dataset		mAP @ 10	
	Pretraining	**Training**	**Development test**	**Challenge test**
ATAE	-	Clotho	0.136	0.114
ATAE-ET	-	FSD50K & Clotho	0.121	0.113
ATAE-EP-F	FSD50K & Clotho	Clotho	0.127	0.121
ATAE-NP-F	FSD50K	Clotho	**0.139**	**0.128**

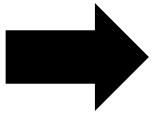
HUAWEI upf. Universitat Pompeu Fabra Barcelona

Challenge submissions

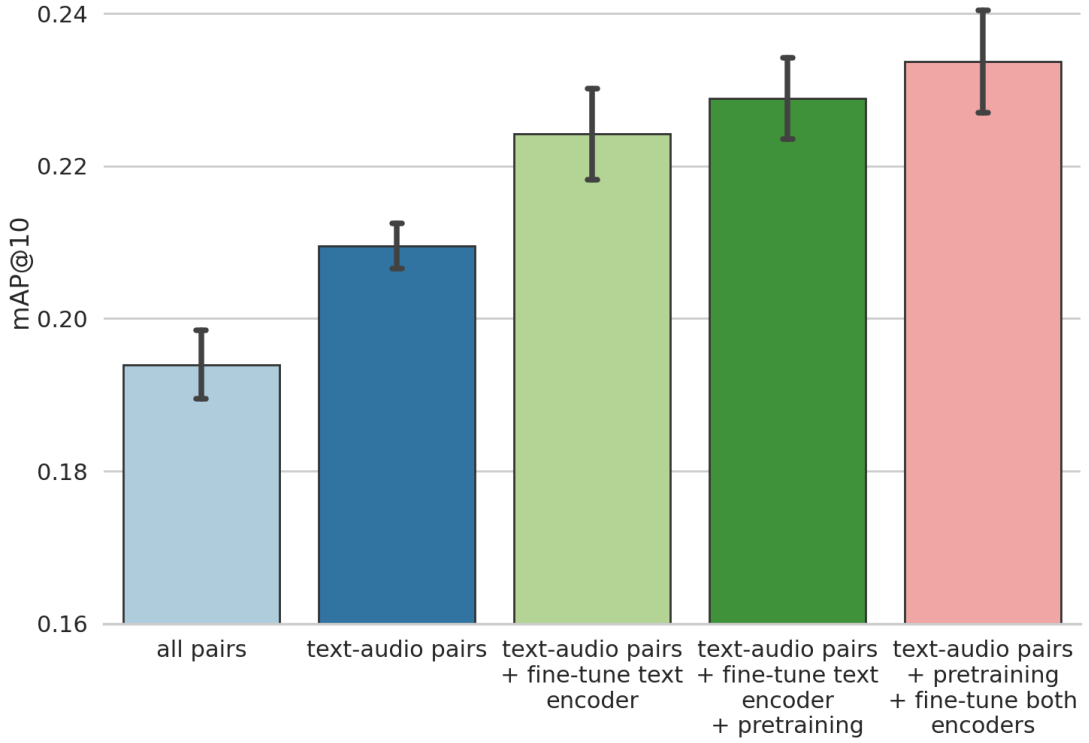
	mAP @ 10	
	Development test	**Challenge test**
Challenge baseline	0.07	0.061
Ours	0.139	0.128
Mei_Surrey_1 [1]	**0.260**	**0.251**

[1] X. Mei, X. Liu, H. Liu, et al., “Language-Based Audio Retrieval with Pre-trained Models,” DCASE2022 Challenge, Tech. Rep., July 2022.

Additional experiments

Contrastive loss  **NT-Xent loss**

Additional experiments



Improvements with:

- NT-Xent loss
- text-audio pairs only
- fine-tuning (both) pretrained encoders
- additional pretraining

HUAWEI

upf. Universitat Pompeu Fabra Barcelona

Thanks!

See you at the poster session!

HUAWEI

upf. Universitat Pompeu Fabra Barcelona