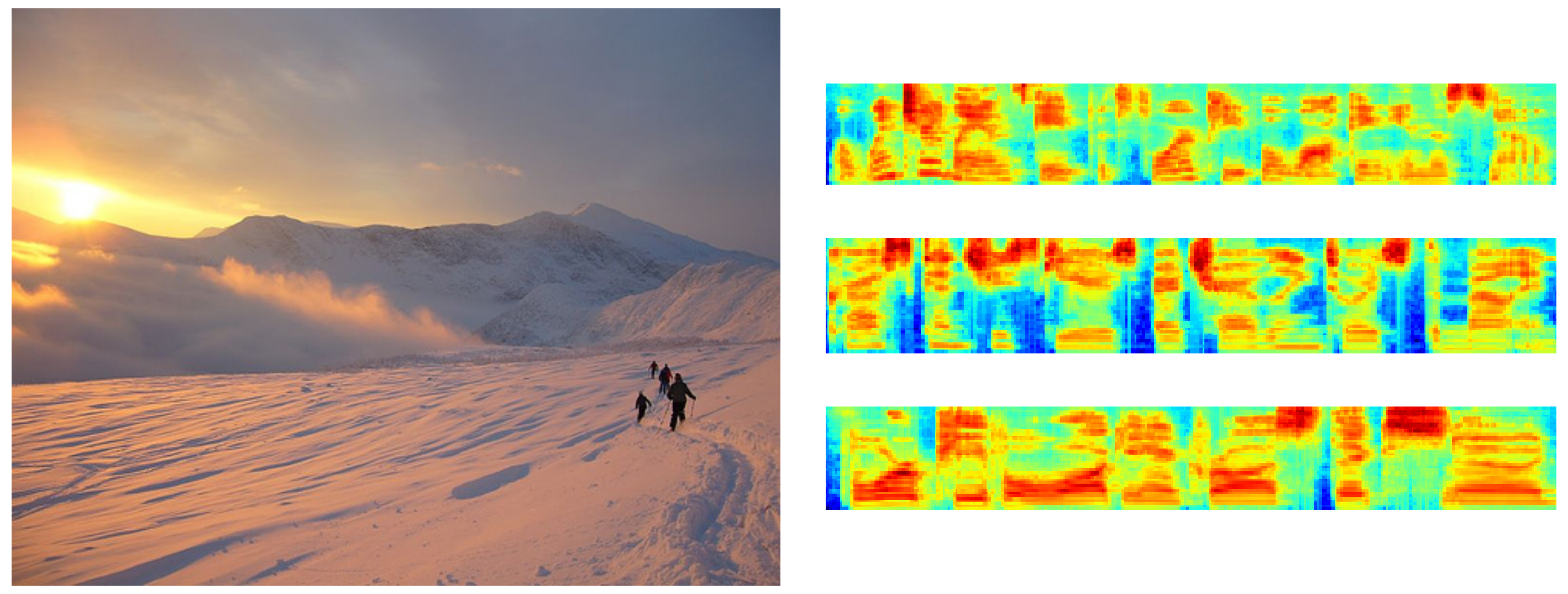
Semantic query-by-example speech search using visual grounding

Herman Kamper[1] Aristotelis Anastassiou[1] Karen Livescu[2]

[1]E&E Engineering, Stellenbosch University, South Africa & [2]Toyota Technological Institute at Chicago, USA

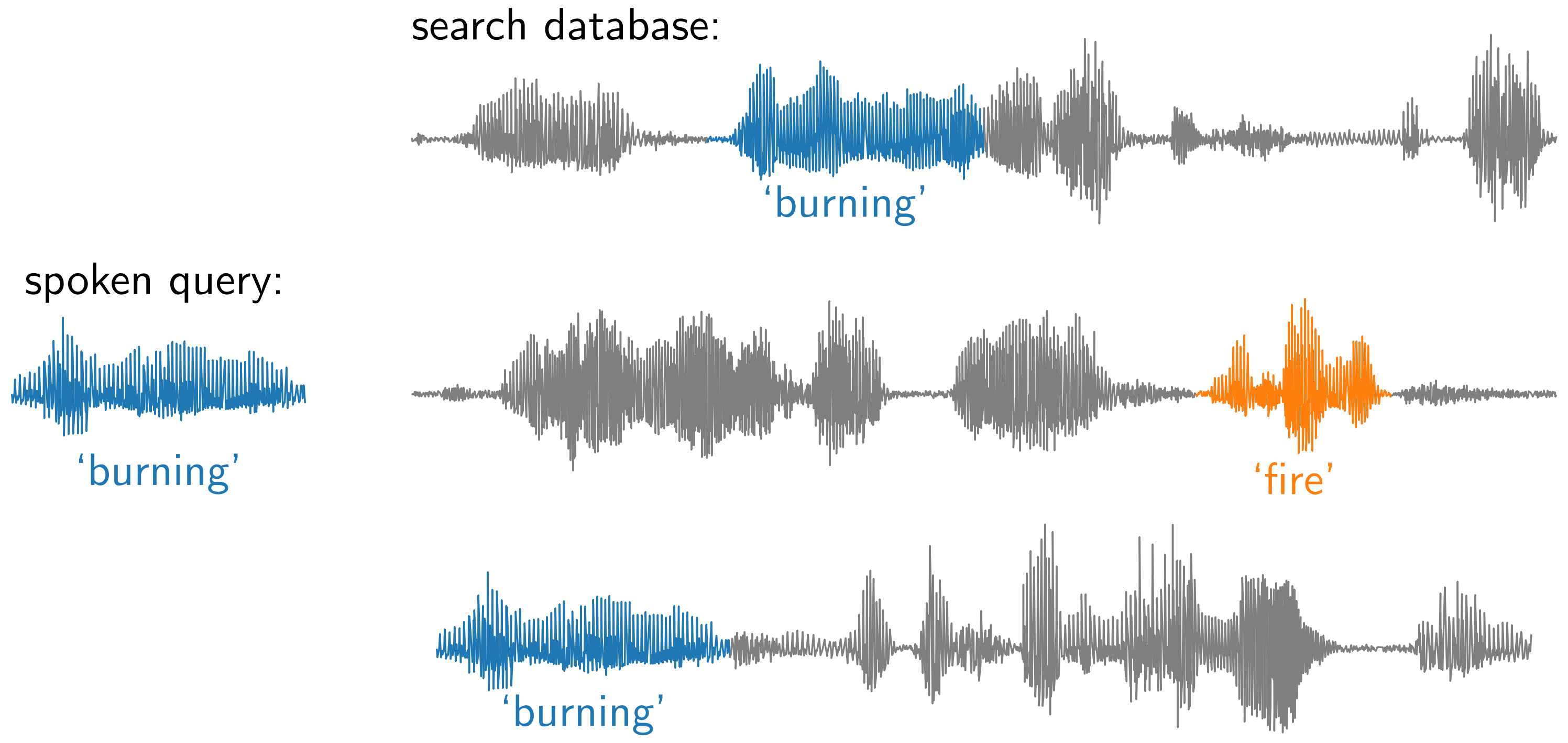
Background

- Current speech recognition methods require large labelled data sets.
- Annotated data is not always available, e.g., for unwritten languages.
- Can we use images as weak labels in low-resource settings?



- Although full ASR might be difficult, other tasks might be possible?
- **Goal:** Use this type of visual supervision for training a (semantic) query-by-example (QbE) speech search model.

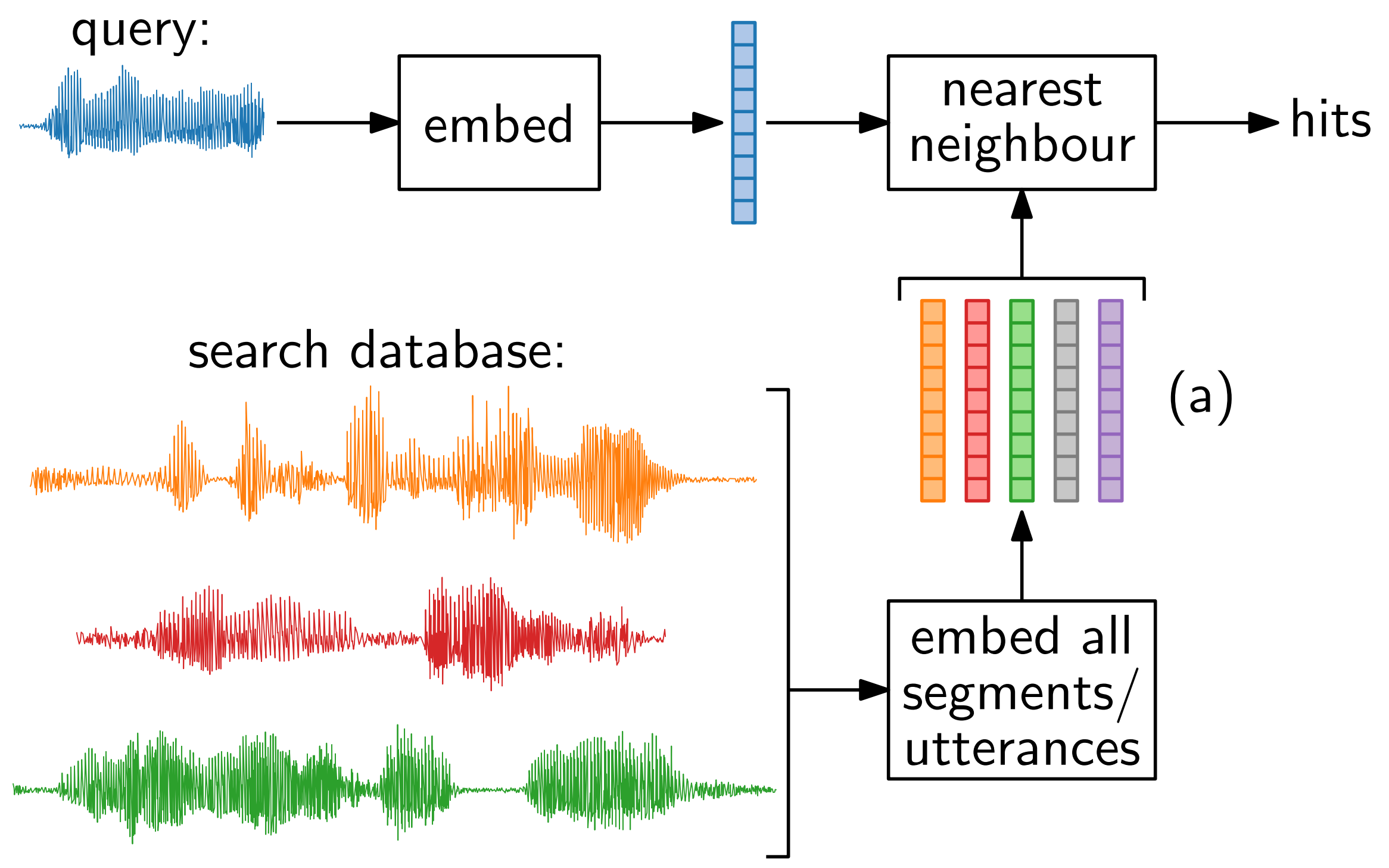
Exact and semantic QbE speech search



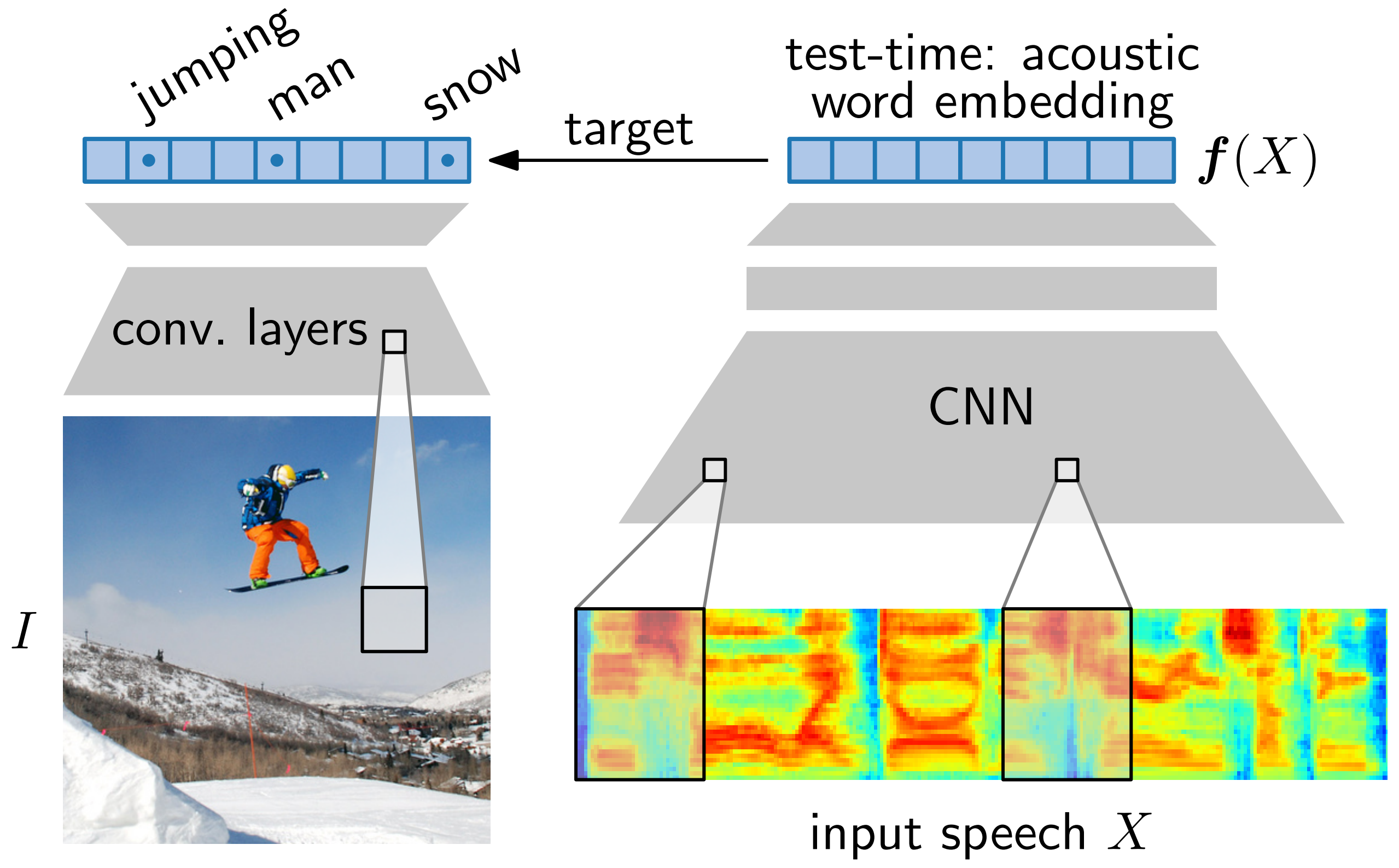
Main idea

- Powerful multi-label visual taggers are available.
- Tag training images with text labels using external visual tagger.
- Use as targets for a speech convolutional neural network (CNN).
- The output of the CNN is an acoustic embedding, which can be used for embedding-based QbE.
- Does not require any transcriptions: Low-resource speech technology.
- Here we simulate low-resource setting using unlabelled English data.

Embedding-based query-by-example (QbE)

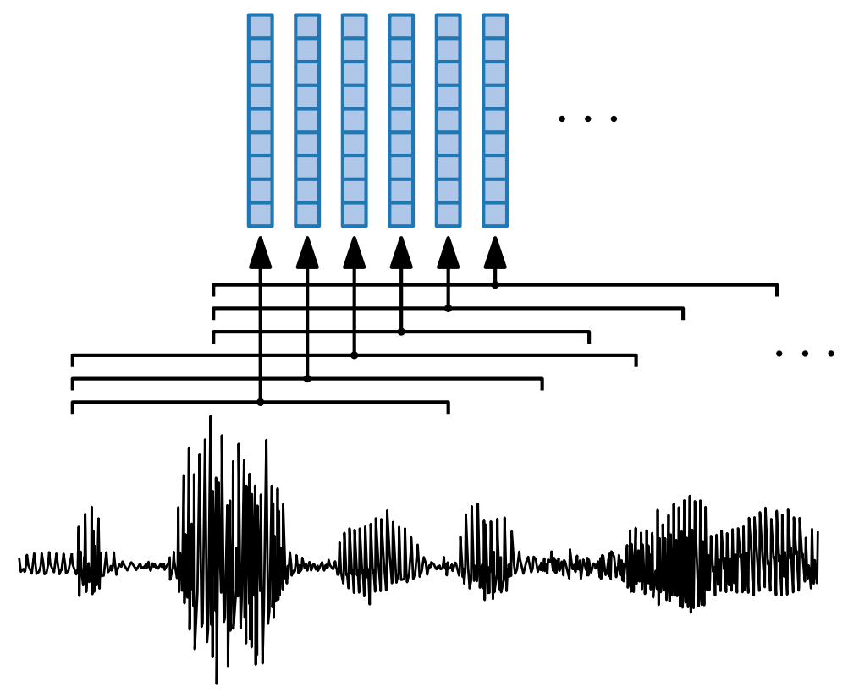


Visually grounded acoustic embedding



Embedding search utterances: Two options

- FAST: Embed and compare query and search utterances as single vectors.
- DENSE: Embed and compare queries to sub-segments within search utterances (shown on right).



Experimental details

- **Data:** 8000 images tagged with 5 English spoken captions (~37 h).
- **Weak labels:** Visual tagger trained on Flickr30k and MSCOCO.

Keyword spotting results

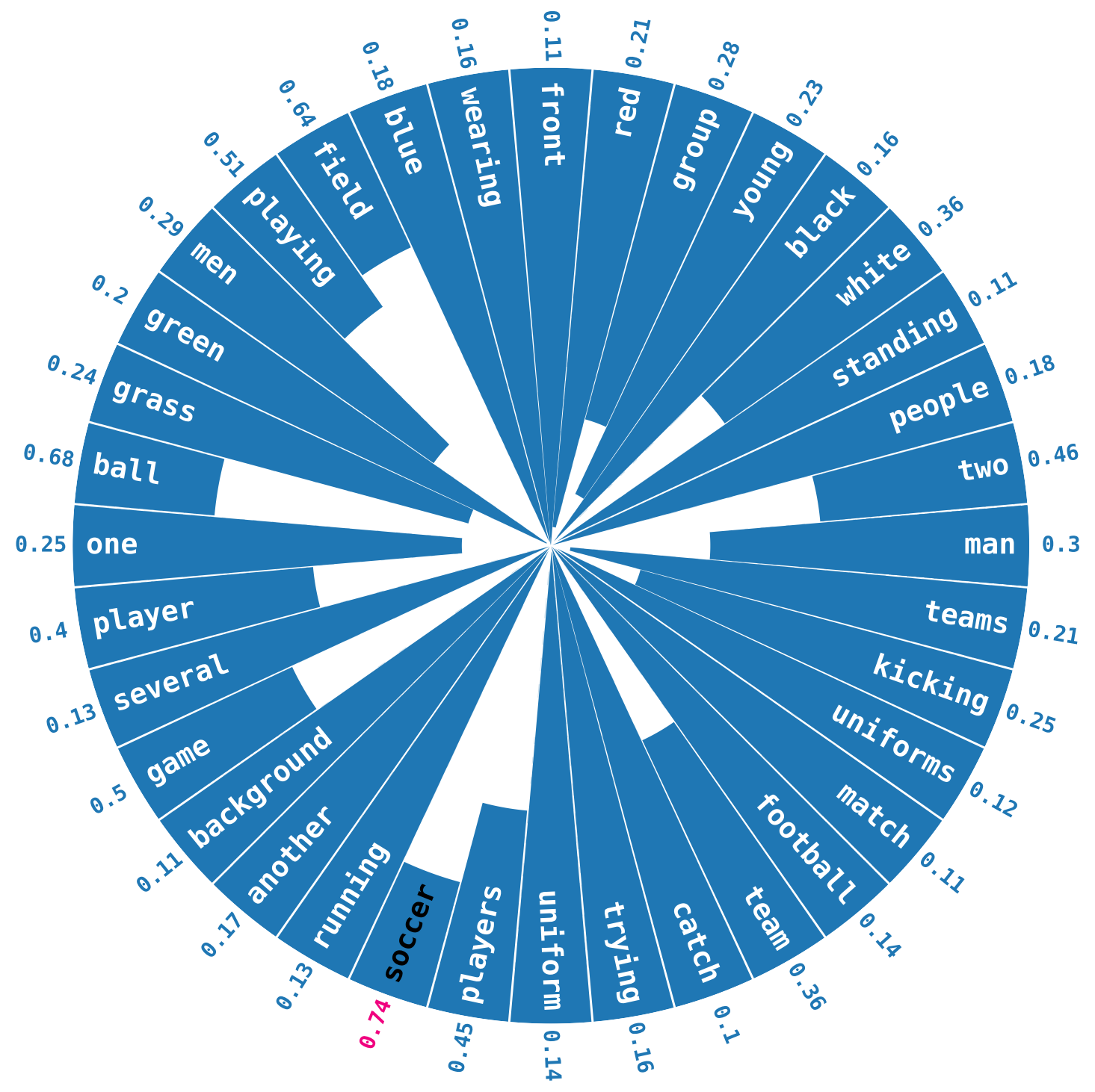
Exact QbE results (%):

	Model	$P@10$	$P@N$	EER	Run-time (min)
Baselines:	RANDOM	4.5	4.5	50	-
	DTW	54.6	24.9	32.1	4080
Our systems:	FASTGROUNDED	27.5	17.9	38.9	< 1
	DENSEGROUNDED	56.0	37.3	21.7	621
Supervised:	FASTSUPERVISED	60.7	41.3	27.2	< 1
	DENSESUPERVISED	72.0	55.7	12.0	568

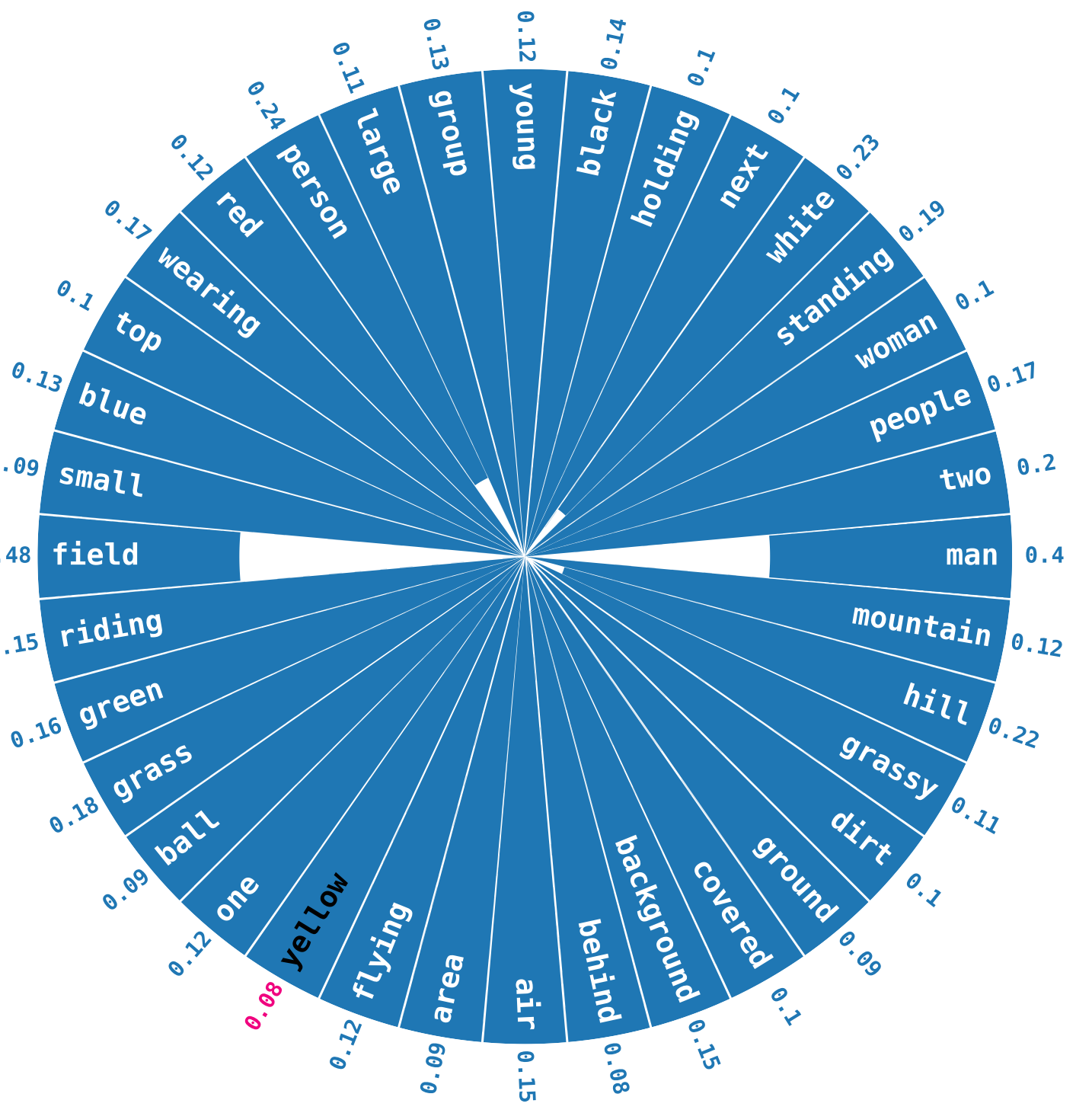
Semantic QbE results (%):

	Model	$P@10$	$P@N$	EER	Spearman's ρ
Baselines:	RANDOM	9.5	9.1	50	5.9
	DTW	44.3	24.3	38.7	13.7
Our systems:	FASTGROUNDED	32.6	23.2	41.4	12.8
	DENSEGROUNDED	55.5	37.3	30.0	14.9
Supervised:	FASTSUPERVISED	56.6	30.9	39.8	8.5
	DENSESUPERVISED	71.2	46.4	27.4	13.5

Examples of query acoustic embeddings



Embedding of 'soccer'.



Embedding of 'yellow'.

Conclusions

- Visual grounding makes it possible to perform semantic QbE without any transcribed speech data.
- **Future:** Apply approach to a truly low-resource language.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

https://www.kamperh.com