Tampere University

Unsupervised Audio-Caption Aligning Learns Correspondences between Individual Sound Events and Textual Phrases

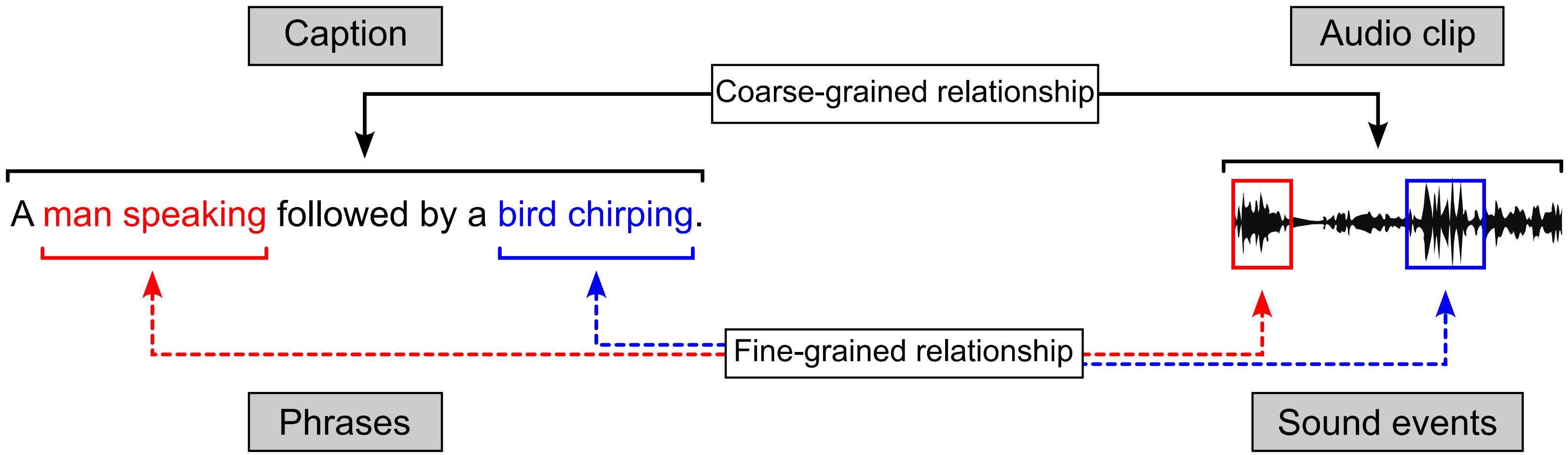
Huang Xie, Okko Räsänen, Konstantinos Drossos, Tuomas Virtanen

Tampere University

presented by Huang Xie at ICASSP 2022

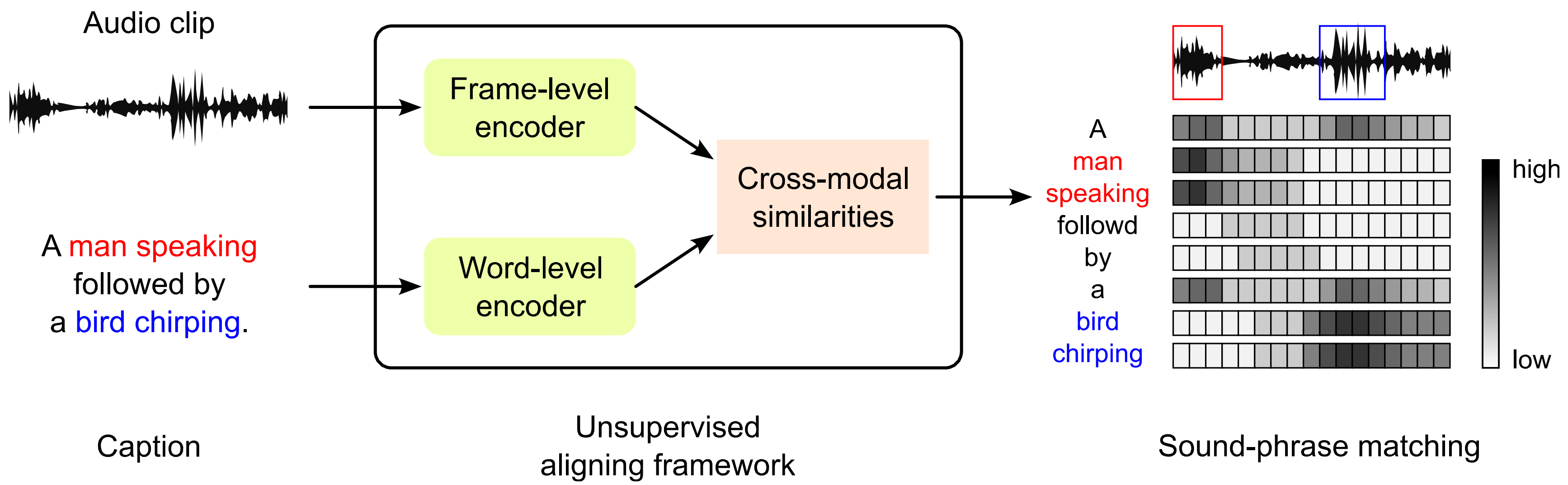
Motivation

- Audio-text cross-modal learning
  - Aims at processing and relating information across **audio clips** and **natural language sentences**.
    - E.g., language-based audio retrieval, automated audio captioning (AAC), audio question answering (AQA), etc.
  - Focuses mainly on modeling **coarse-grained** relationships.
  - Investigates rarely **fine-grained** relationships.
    - Key of interpreting audio with natural language descriptions.



Unsupervised Audio-Caption Aligning

- We infer **sound-phrase** correspondences via **unsupervised cross-modal aligning** [1]:
  - ➢ Deal with **unaligned**, **unannotated** audio-caption pairs.
  - ➢ Encode audio clips into frame-level representations.
  - ➢ Encode captions into word-level representations.
  - ➢ Extract matching audio and lexical concepts via measuring **multi-level cross-modal similarities**.
    - ✓ E.g., frame-word, frame-phrase, audio-caption.



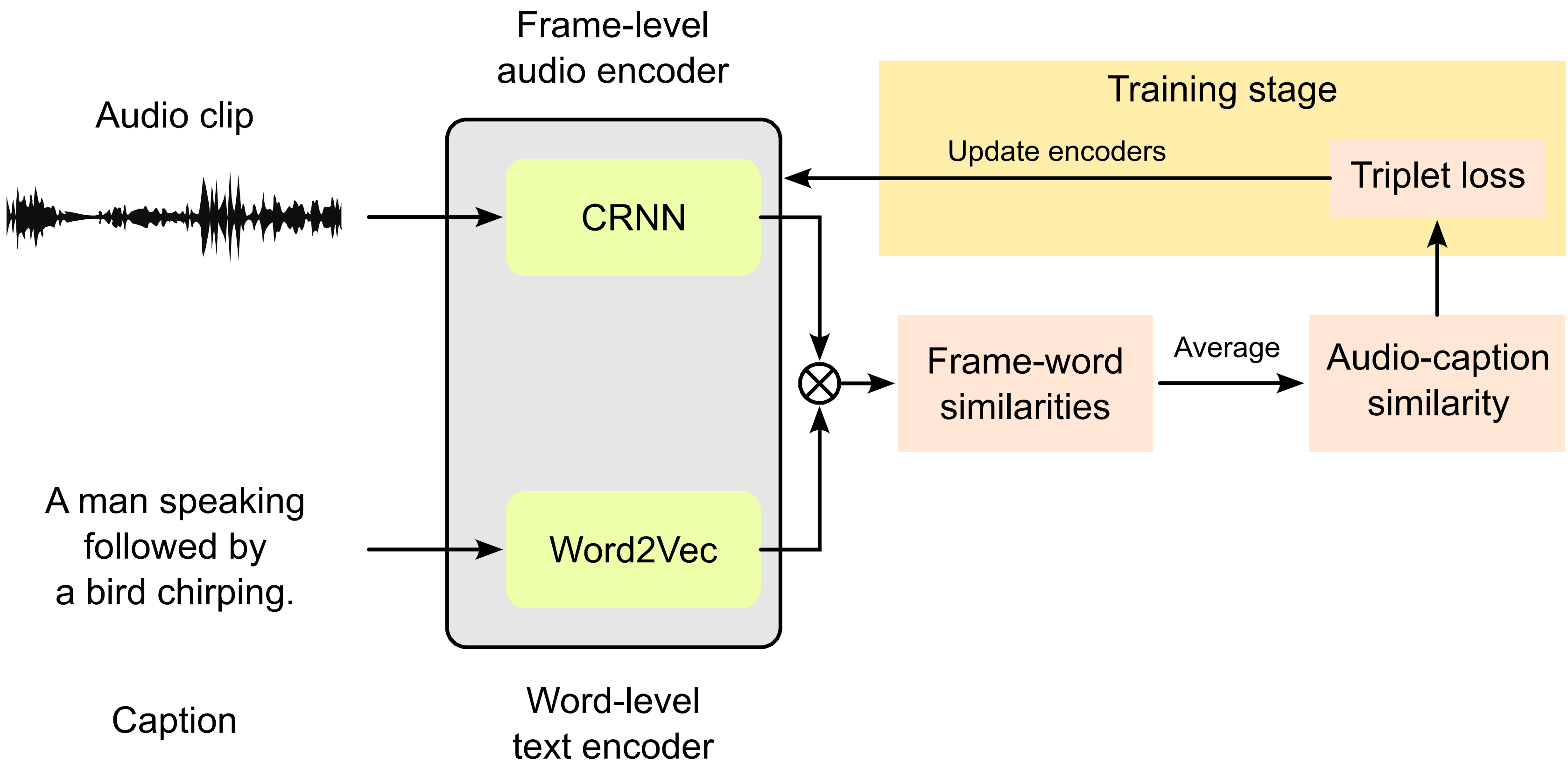
[1] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input,” *IJCV*, pp. 620–641, 2020.

Proxy Tasks

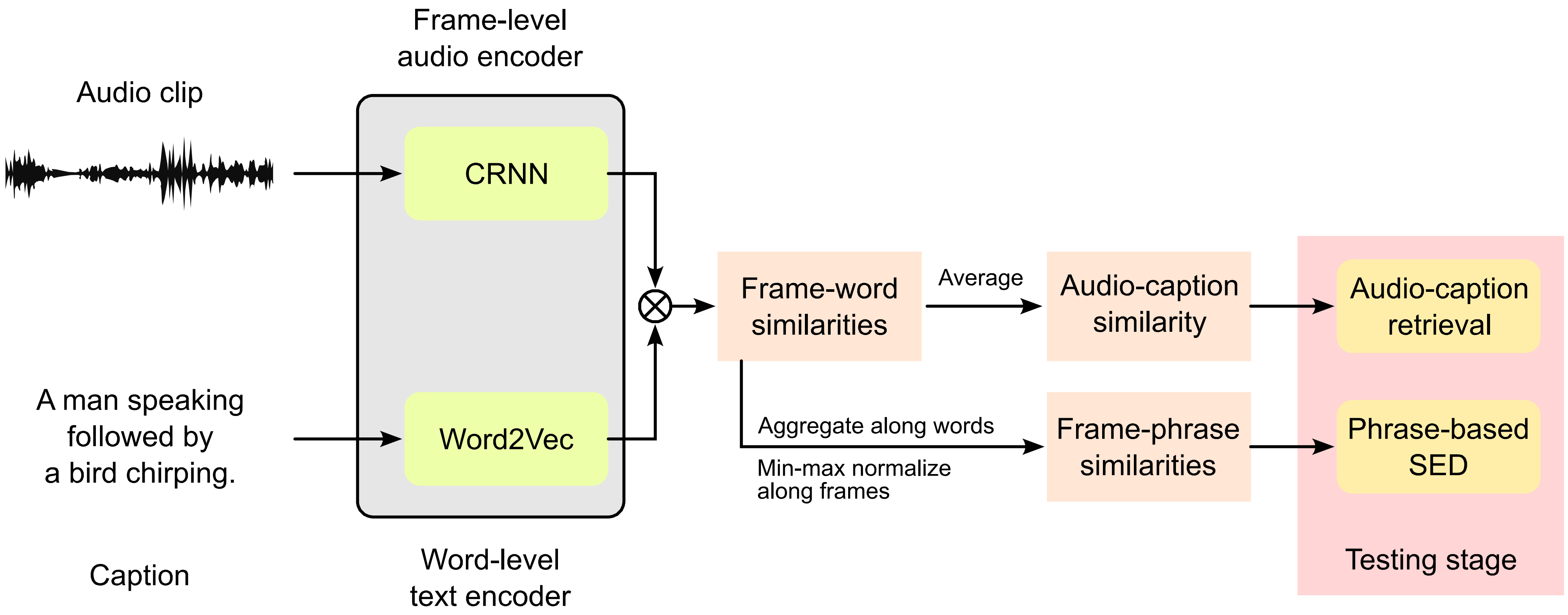
- Audio-caption retrieval:
  - Given a written caption, retrieve relevant audio clips, or vice versa.
  - Measures the ability for global pairing of audio clips and captions.

- Phrase-based **S**ound **E**vent **D**etection (SED):
  - Given a textual phrase (part of a caption), extract timestamps of the corresponding audio event.
  - Measures the ability of exploring sound-phrase correspondences.

Proposed Framework – Training Stage



Proposed Framework – Testing Stage



Audio Encoder

- **C**onvolutional **R**ecurrent **N**eural **N**etwork (**CRNN**) [2]:
  - ➢ Five convolutional blocks + one bidirectional gated recurrent unit (BiGRU).
  - ➢ Be able to process variable-length audio signals.

- Input: 64-dimensional log-mel energies (40 ms frame shift).
- Output: 300-dimensional frame-level acoustic embeddings.

- Final acoustic embeddings are L2-normalized.

[2] X. Xu, H. Dinkel, M. Wu, and K. Yu, “Text-to-audio grounding: Building correspondence between captions and sound events,” in ICASSP, 2021, pp. 606–610.

Text Encoder

- Word2Vec:
  - Two-layer fully-connected neural network with the skip-gram architecture.
  - Produces word embeddings that are good at predicting surrounding words in sentences or documents.
  - Pre-trained with Google News dataset (about 100 billion words).

- Output: 300-dimensional word embeddings.

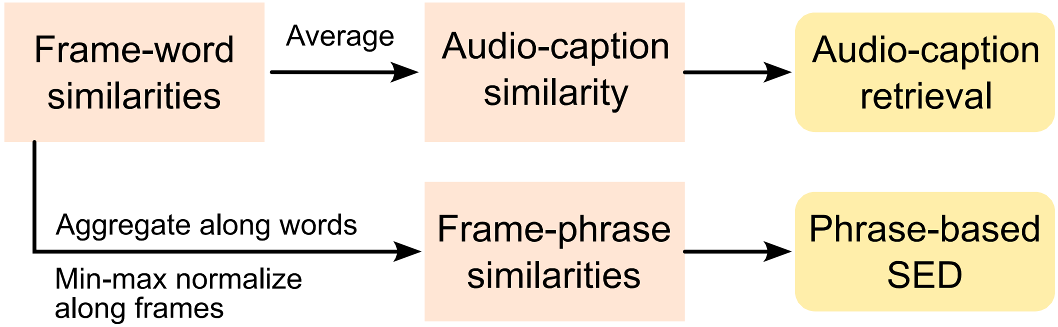
- Final word embeddings are L2-normalized.

Multi-Level Cross-Modal Similarities

- Frame-word similarities:
  - Dot products of word and acoustic embeddings.

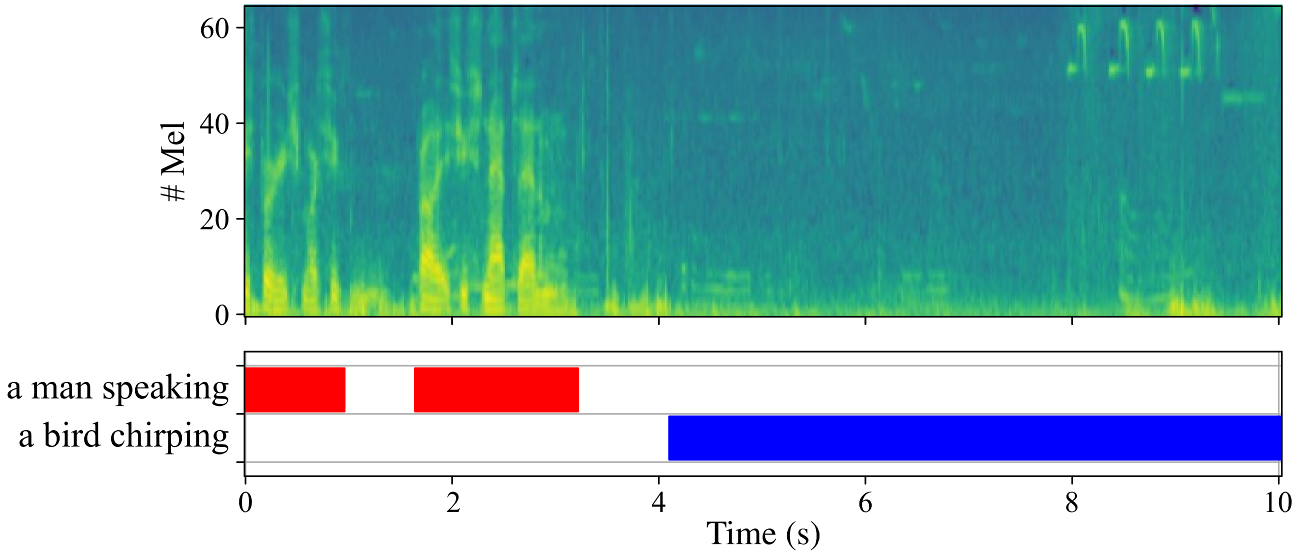
- Audio-caption similarity:
  - Average across all frame-word similarities.
  - Used for global pairing of audio clips and captions (training criterion for a triplet loss).
  - High values for matched audio-caption pairs.

- Frame-phrase similarities:
  - Min-max normalized aggregations of frame-word similarities.
  - A detection threshold applied for predicting sound activities, i.e., SED.
  - High values denote semantic correspondence between text phrases and audio frames.



Dataset for Experiments

- AudioGrounding [3]:
  - ➢ Paired audio clips and captions.
    - ✓ 4,590 10-second audio clips from AudioSet.
    - ✓ 4,994 audio captions from Audiocaps.
  - ➢ Annotated phrases and relevant sounds.
    - ✓ 13,985 human-annotated phrases, along with on- and off-sets of sounds.



Caption: a man speaking followed by a bird chirping close by

- Downloaded version:

Split	#Clips	#Captions	#Event phrases
Training	4,253	4,253	11,732
Validation	30	150	439
Test	67	335	1,118

[3] X. Xu, H. Dinkel, M. Wu, and K. Yu, “Text-to-audio grounding: Building correspondence between captions and sound events,” in ICASSP, 2021, pp. 606–610.

Task 1 – Audio-Caption Retrieval

- Task setup:
  - ➢ Given an instance in one modality (audio or caption), retrieve its paired instance from 30 candidates in another modality.
    - ✓ One positive + 29 negatives.
  - ➢ Repeat evaluation twenty times with randomly sampled negatives.

- Evaluation metrics:
  - ➢ Recall at rank K.
  - ➢ K = {1, 5}.

- Results:
  - ➢ The proposed method can match audio clips with captions, and vice versa.

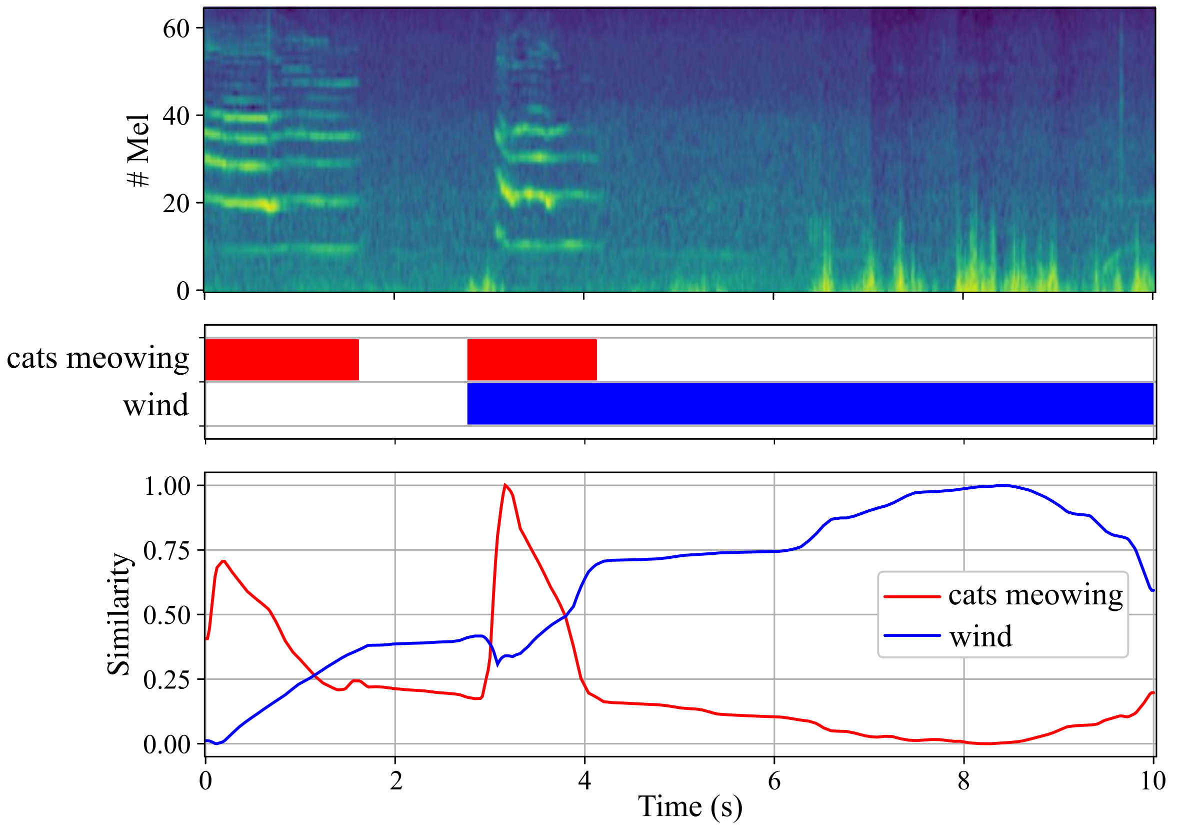
Retrieval		Chance Levels	Proposed
Audio2Caption	R@1	0.03	0.21 ± 0.04
	R@5	0.17	0.65 ± 0.05
Caption2Audio	R@1	0.03	0.23 ± 0.04
	R@5	0.17	0.71 ± 0.04

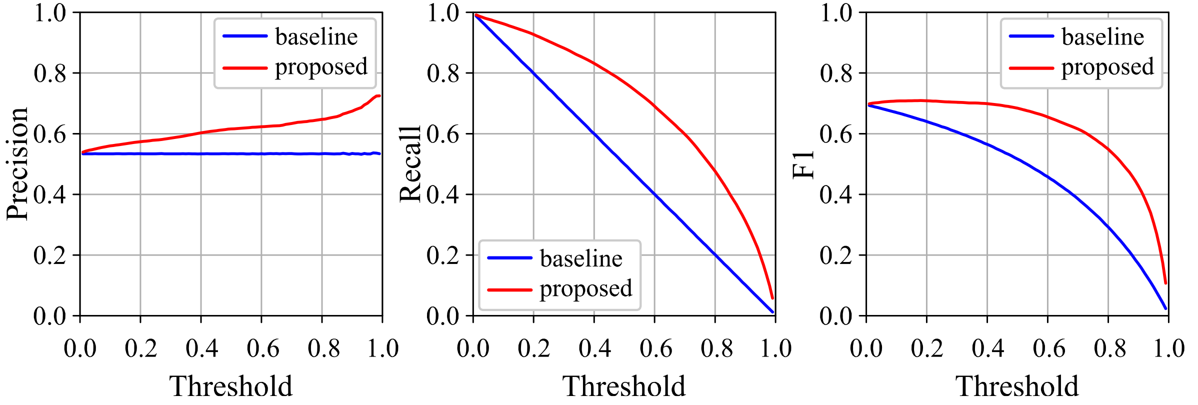
Task 2 – Phrase-based SED

- Task setup:
  - ➢ Given an ***audio clip*** and an ***event phrase*** from the corresponding caption, predict temporal position of the sound(s).
  - ➢ Baseline: random guessing.

- Evaluation metrics:
  - ➢ Global frame-based Precision, Recall, and F1.

- Results:
  - ➢ The proposed method can associate sound events with phrases, i.e., learning sound-phrase correspondences.
  - ➢ An example of learned frame-phrase similarities.





Conclusion

- We propose an unsupervised audio-text aligning method:
  - ➢ Learns semantic correspondences between audio clips and captions by aggregating frame-word similarities.
  - ➢ Learns to align individual sound events to text phrases without alignment information during training.

- We evaluated the proposed method in two cross-modal tasks:
  - ➢ Audio-caption retrieval.
  - ➢ Phrase-based SED.

Tampere University

Thank You!
For Watching!

Huang Xie