Task 6b
Language-based Audio Retrieval

Huang Xie, Félix Gontier, Samuel Lipping,

Konstantinos Drossos, Tuomas Virtanen, Romain Serizel

DETECTION AND
CLASSIFICATION OF
ACOUSTIC
SCENES AND
EVENTS

Tampere University

Inria

UNIVERSITÉ
DE LORRAINE

Motivation

➤ The "Big Data Era"

- Explosive growth of audio data on the web.

- Great demand for **content-based audio search** tools.

➤ Search Queries

- Natural language vs. Keywords & Phrases

- For example, "a baby yelling as a woman talks followed by a dog barking".

➤ Language-based Audio Retrieval

- Retrieving desired audio with free-form text.

➤ Real-world Applications

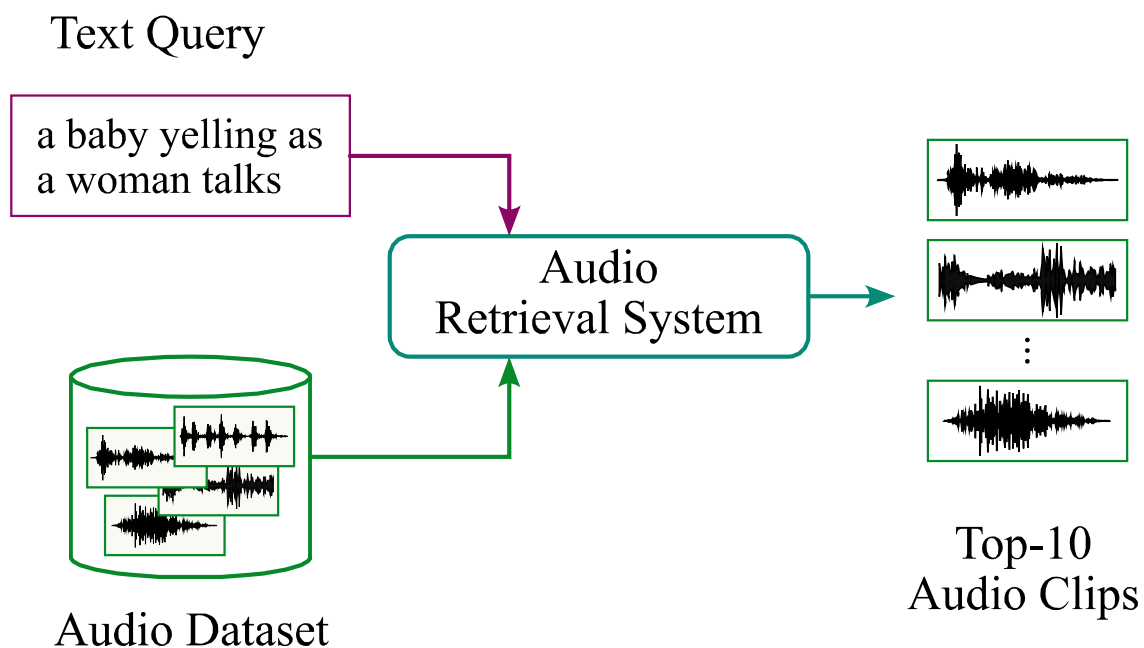- Search engines, multimedia databases, human-computer interactions, etc.

Task Setup

➤ Task Description

- To retrieve 10 audio files from a given dataset.

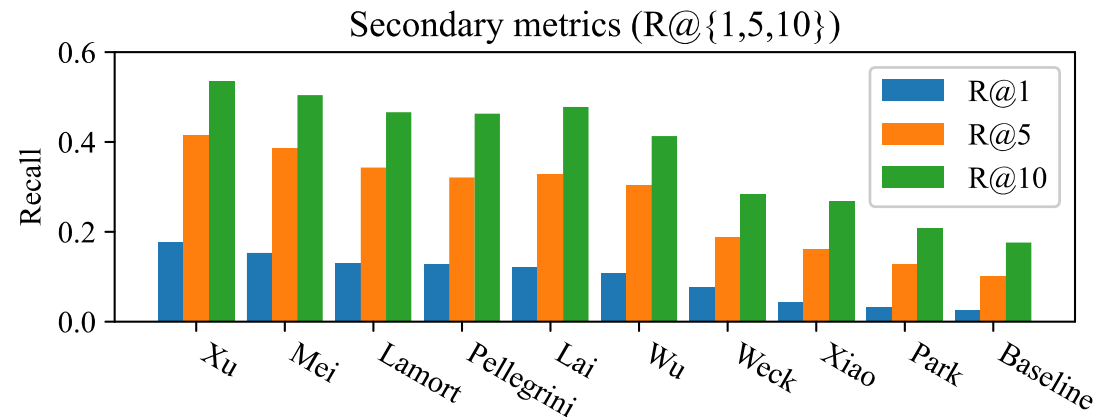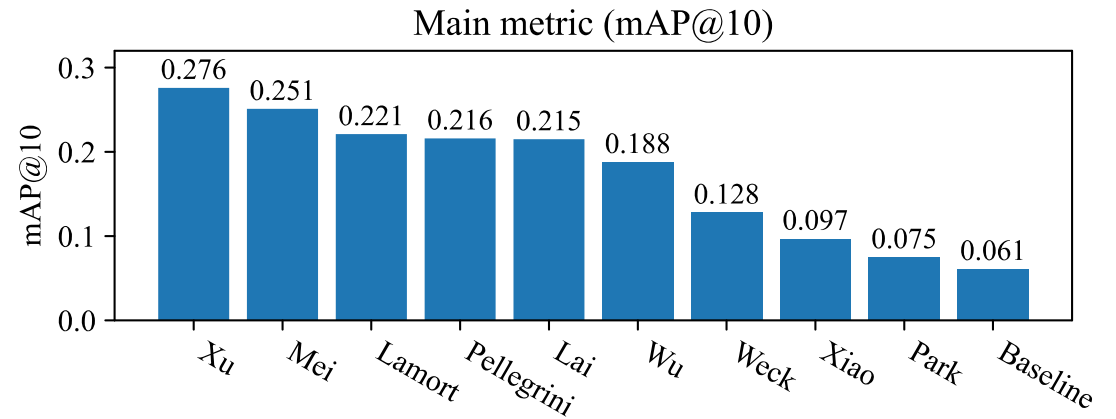- To sort them according to their semantic relevance to the query.

➤ Audio Dataset

- 1,000 audio-caption pairs.

Text Query

a baby yelling as a woman talks

Audio Retrieval System

Audio Dataset

Top-10 Audio Clips

Tampere University

Inria

UNIVERSITÉ DE LORRAINE

Submissions

➢ Total 31 systems from 9 teams.

  • Teams ranking (mAP@10)

  • Secondary metrics: R@{1, 5, 10}.

➢ System Summary

  • **Bi-encoder architecture** adopted in all systems.

  • **Pretrained audio** and **NLP models** as encoders.

  • **Contrastive learning** approaches for training.

➢ Further information present in our poster.

Main metric (mAP@10)

Secondary metrics (R@{1,5,10})

Thank You!