AUTOMATED AUDIO CAPTIONING WITH WEAKLY SUPERVISED PRE-TRAINING AND WORD SELECTION METHODS

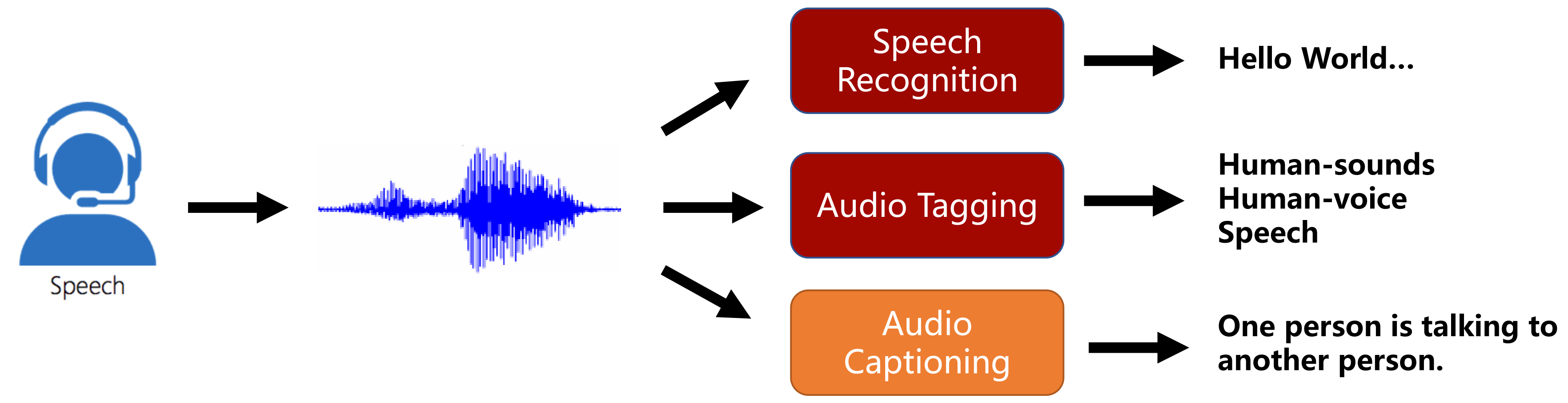
Qichen Han, Weiqiang Yuan, Dong Liu, Xiang Li, Zhen Yang

NetEase (Hangzhou) Network Co., Ltd., China

{hanqichen, yuanweiqiang, hzliudong, hzlixiang,yangzhen1}@corp.netease.com

Introduction

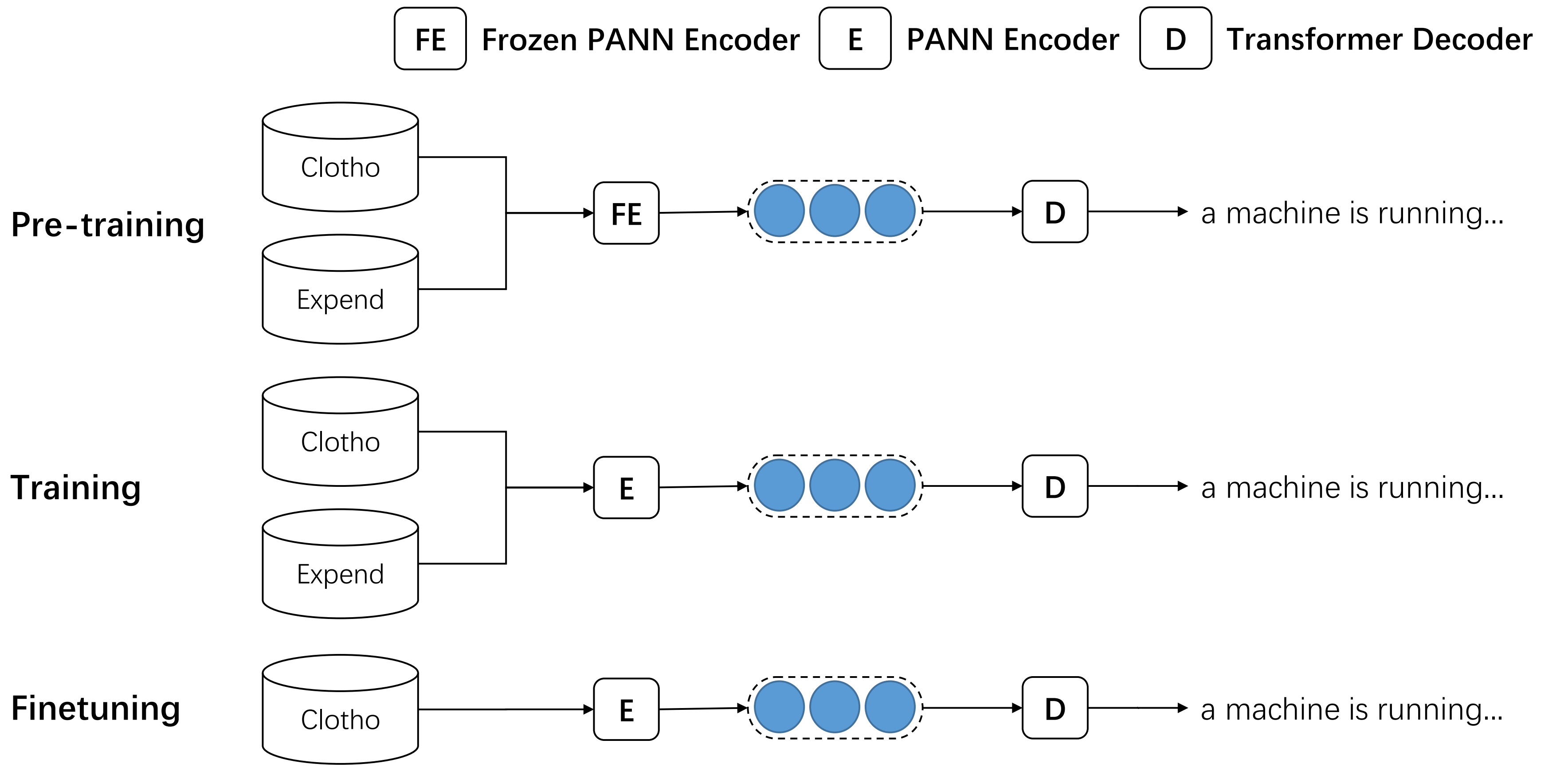
- Automated audio captioning is a **translation task** where the model outputs a textual description given an audio signal.



Motivation

- Pre-training on large weakly labeled datasets followed by fine-tuning on target datasets boosts performance.
- Similar audios may have similar captions
  - Tag-Level: we use **audio tag** generated by PANN to assist decoding
  - Caption-Level: we use **keywords** extracted from the **captions of similar audios** to assist decoding

Overview Diagram



Proposed Methods

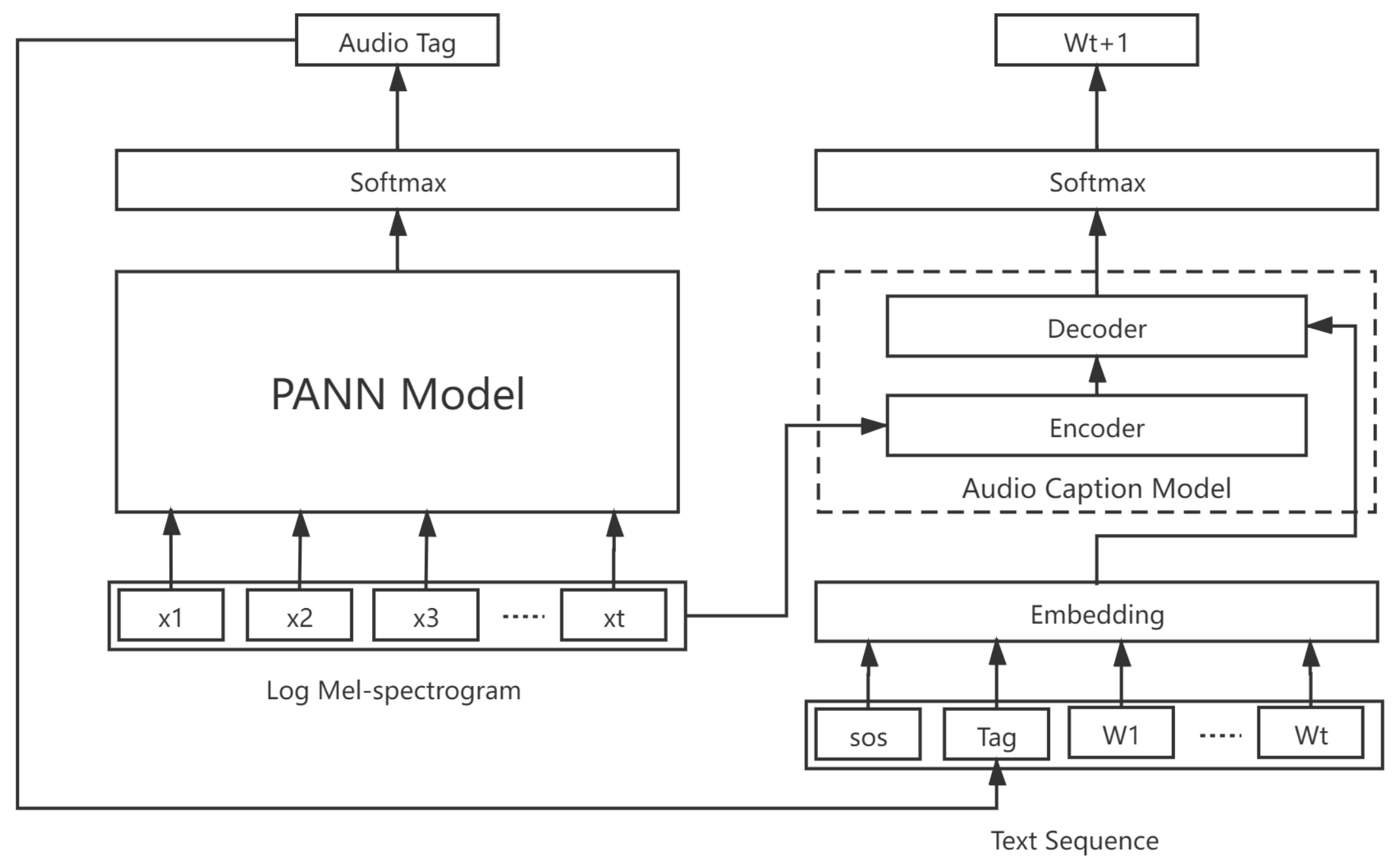
- **Weakly labeled dataset (65K Audio Clips and captions)**
  - Collect audios and captions from the free sound effects sites
  - Using heuristic rules to filter and clean captions
  - Remove Audios shorter than 5 seconds and randomly select 15 to 30 seconds from the long audios
- **Tag-Level: Tag-enhanced decoder**
  - Use the PANN model to predict the tag of each audio in the dataset
  - Training Stage: Use Tag as label and place it in front of caption
  - Test Stage: Add the tags as known information to the decoder
- **Caption-Level: Keyword-enhanced decoder**
  - **Training Data**
    - For each mini-batch random choice K anchor audios
    - For each anchor random choice M similar audios
    - For each (anchor, similar) pair random choice N dissimilar audios
  - **Model Training**
    - Using ESIM to calculate the similarity of audios **s(a, b)**
    - Using triplet dynamic margin loss

$$Loss(a,p,n) = \max(0, m(a,p,n) + s(a,n) - s(a,p))$$

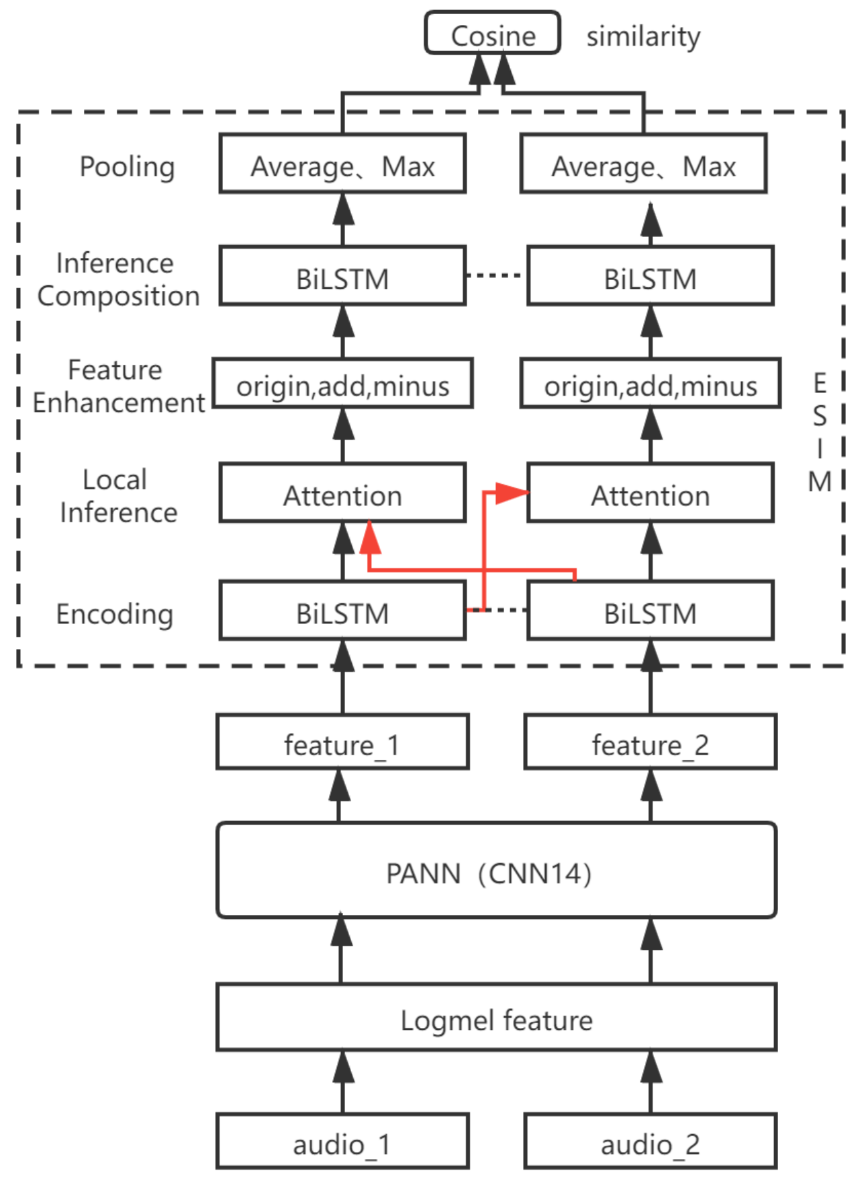
$$m(a,p,n) = \max(0.4, SPIDEr(a,p) - SPIDEr(a,n))$$

  - **Decoding Enhancement**
    - Get 10 most similar audios and their captions (50 captions)
    - Get 10 most important keywords and their weights (tf-idf)
    - Modify the vocabulary probability by the weights of these keywords

Tag-enhanced decoder



Similarity Model



Experiment & Case Study

PE: Pre-trained encoder. KD: Keyword enhanced decoder. TD: Tag enhanced decoder. PD: Perturbed audio data. AD: AudioCaps dataset. WD: Weak label dataset.

Model	BLUE1	BLUE2	BLUE3	BLUE4	METEOR	ROUGE-L	CIDEr	SPICE	SPIDEr
Baseline	0.521	0.328	0.216	0.139	0.153	0.353	0.326	0.102	0.2142
PE	0.541	0.348	0.228	0.149	0.162	0.362	0.386	0.112	0.2490
PE+KD	0.552	0.360	0.240	0.156	0.167	0.372	0.409	0.119	0.2641
PE+TD	0.537	0.341	0.225	0.148	0.163	0.359	0.371	0.114	0.2427
PE+PD	0.550	0.353	0.232	0.149	0.164	0.366	0.385	0.118	0.2514
PE+PD+AD	0.554	0.356	0.235	0.153	0.167	0.364	0.405	0.117	0.2609
PE+PD+AD+WD	0.578	0.381	0.258	0.171	0.176	0.384	0.444	0.123	0.2837
PE+PD+AD+WD+KD	0.583	0.391	0.267	0.177	0.179	0.388	0.456	0.128	**0.2920**

- Adding a pre-trained encoder can significantly improve SPIDEr scores[PE]
- The benefits of data augmentation are significant[PD AD WD]
- The keyword enhanced decoder can assist the generation of captions [KD]
- The tag enhanced decoder is not effective in this case [TD]

The case for Chopping pieces of mushrooms vigorously.wav

Item	Value
Reference	Vegetables are cut and chopped on a cutting board by someone.
w/o Keyword enhanced decoder	chopping vegetables with a knife.
Keyword enhanced decoder	chopping vegetables on a cutting board with a knife.
keyword	knives/knife chopping/chopped/chop/chops vegetable/vegetables woods/wood cutting/cuts/cut saw/saws/sawed/sawing/ boards/board wooden food slices/sliced/slicing

The case for SamyeLing_Pheasant121102.wav

Item	Value
Reference 1	A bird is chirping while another bird is calling for a mate.
Reference 2	A bird making a call and another bird that is chirping.
Tag	Animal
w/o Tag enhanced decoder	a person uses a tool to each other
Tag enhanced decoder	a bird is chirping and then another bird is chirping in the background

Conclusion & Future work

- Conclusion
  - Pre-trained PANN encoder and weakly labeled data can significantly improve SPIDEr scores
  - The keyword enhanced decoder can assist the generation of captions, which indicates that similar audio captions contain valuable information
- Future work
  - Explore the promotion of pre-training with larger-scale weakly label data
  - Try other effective methods to integrate similar audio captions information into AAC tasks