Tampere University

# A summarization approach to evaluate audio captioning

*Irene Martín-Morató, Manu Harju, Annamaria Mesaros*


DCASE 2022 WORKSHOP

# Motivation

- Current audio captioning metrics originate from:
  - Machine translation: **BLEU**, **METEOR** and **ROUGE**
  - Image captioning: **SPICE**, **CIDEr** and **SPIDEr**

- Problems:
  - Some are based on **n-gram** matching, which is **not sufficient** for the task of simulating human judgement in caption evaluation.
  - **Fluency** is **not** considered in any metric.
  - Only *objects*, *attribute* and *relations* are considered (concepts related to image processing)

- Proposal:
  - **Cross-modal summarization**: description including the **most relevant content**, subject to their own judgement.

Tampere University

| Candidate captions |
| --- |
| C1. **Children** are **talking** outside. |
| C2. A **dog** is **barking** at a **car passing by.** |
| C3. A **car** is **passing by** a group of **children** that are **playing** and **laughing**. |

# CB-score

## Reference Captions

$$rel_i = \frac{N_i}{\sum_{j=1}^{M}(N_j)}$$

Summary = caption

Content Units = sound events.

$N_i$ = Number of times sound event $i$ is mentioned in the captions.

$M$ = total number of sound events mentioned in all captions.

## Candidate Captions

$$\text{CB-score} = \frac{\sum_{j=1}^{K} rel_j}{\sum_{k=1}^{K} rel_k}$$

$K$= Number of sound events present in candidate caption

| | Sound events | Relevance | Ideal caption content | CB-score |
| --- | --- | --- | --- | --- |
| C1 | Children talking | 0.36 | Children laughing | 0.36/0.40=**0.90** |
| C2 | Dog barking, car passing by | 0.08 | Children laughing, Children talking | 0.08/(0.40+0.36)=**0.11** |
| C3 | Car passing by, Children laughing | 0.08+0.40 | Children laughing, Children talking | 0.48/0.76=**0.63** |

Tampere University

# Results

| | CLOTHO | | | AudioCaps | | |
|---|---|---|---|---|---|---|
| **System** | SPIDEr | FENSE | CB-score | SPIDEr | FENSE | CB-score |
| Baseline | **0.22** (0.21, 0.24) | **0.46** (0.45, 0.47) | **0.49*** (0.46, 0.51) | **0.34*** (0.32, 0.37) | **0.57** (0.56, 0.58) | **0.63*** (0.60, 0.65) |
| ED-RNN | 0.15 (0.14, 0.16) | 0.41* (0.40, 0.43) | 0.40 (0.38, 0.43) | 0.30 (0.28, 0.33) | 0.54* (0.53, 0.55) | **0.62*** (0.59, 0.64) |
| AACTransformer | 0.19 (0.18, 0.21) | 0.40* (0.39, 0.41) | **0.47***(0.45, 0.50) | **0.35*** (0.32, 0.37) | 0.54* (0.53, 0.55) | **0.65*** (0.63, 0.68) |
| Reference caption | 0.58 (0.55, 0.61) | 0.58 (0.57, 0.59) | 0.64 (0.62, 0.66) | 0.56 (0.52, 0.59) | 0.68 (0.68, 0.69) | 0.76 (0.75, 0.78) |

**SPIDEr**: concentrated, closed to 1.
**FENSE**: normally distributed between 0 and 1.
**CB-score**: many values at the extremes.

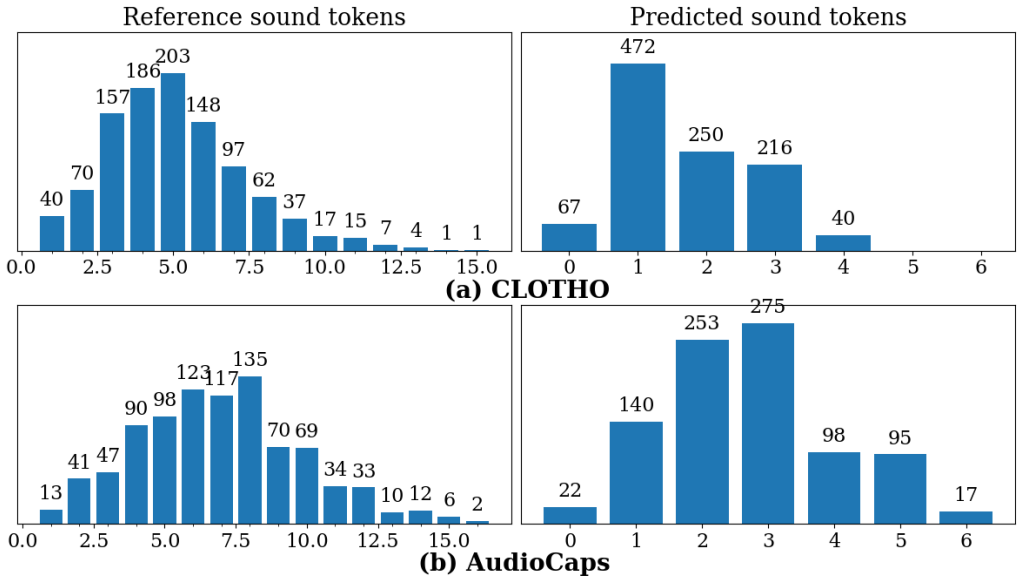
**Number of sound tokens extracted from the caption.**

**CLOTHO**

- 45% of the predicted captions have only one token.
- 6% do not match any sound

**AudioCaps**

- 54% of the predicted captions have 2 or 3 tokens.



# Discussion and Conclusions

- **CB-score**
  - Precision-type metric.
  - Penalizes the presence of sounds if the more relevant ones are not included in the captions.
  - Inserted sounds are not penalized.
  - Sound events instead of n-grams.

- **Future work**
  - To include the ability to measure lexical or grammar structure.
  - Penalize the extra information if it is not present in the reference captions.
  - More automatized correspondence to the AudioSet vocabulary.