# Translating speech with just images

Dan Oneață[1], Herman Kamper[2]
[1]Politehnica Bucharest, Romania [2]Stellenbosch University, South Africa

INTERSPEECH 2024

audio (Yorùbá) → audio-to-text model → text (English): *Two people riding motorcycles on the road in the daytime.*

image → image captioner → text (English)

Goal · Speech translation: audio (Yorùbá) → text (English).
Given · Images paired with audio: Yorùbá Flirckr Audio Captions Corpus (YFACC).
Idea · Use pretrained image captioner to generate targets for audio-to-text model.

## Image captioning with the GIT model

input image

decoding: beam search
*A young boy standing on a dirt road next to a field.*
*A young boy standing on a dirt road in a field.*
*A little boy standing on a dirt road in a field.*

decoding: multinomial sample
*Many small kids on a path running through a field.*
*A little boy using a camera to look for watermelons.*
*A little boy standing in a very narrow dirt road.*

decoding: diverse beam search
*A young boy standing on a dirt road in a field.*
*A small child standing in the middle of a dirt road.*
*The boy is looking at something in the distance.*

## Audio-to-text model

audio → encoder (wav2vec → project) → decoder

text: *Two people . . .* → embed. → self attention → cross attention → mlp (×12) → head → output: *riding*

Architecture:
- Encoder: wav2vec2 XLS-R 2B (frozen).
- Decoder: GPT-2 (frozen).
- Projection and cross-attention layers (learnable).

- Num. parameters: 29M learnable; 2.3B in total.
- Loss: Cross-entropy on next-token prediction.
- Train on audios paired with one of the five generated captions by the image captioning model.

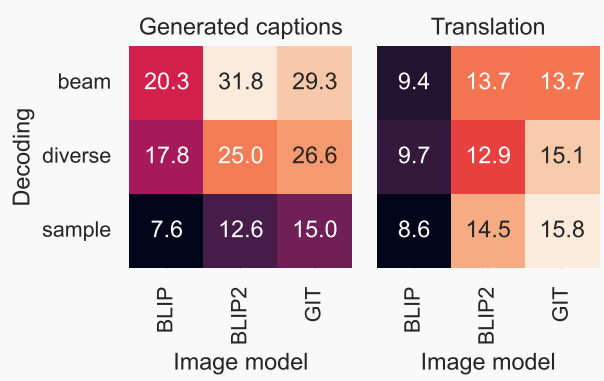
## Experimental results: BLEU scores on FACC and YFACC

| | method | input language | targets language | targets decoding | num. references 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|
| | *Toplines* | | | | | | | | |
| 1 | annotator | N/A | N/A | N/A | 8.32±0.5 | 13.95±1.0 | 17.84±0.9 | 21.59±0.7 | N/A |
| 2 | translation | Yorùbá | → English | annotations | 15.23±0.0 | 18.25±0.3 | 19.87±0.4 | 21.07±0.3 | 22.01±0.0 |
| 3 | generated captions | N/A | English | beam search | 9.62±0.9 | 17.07±1.0 | 22.16±0.8 | 25.88±0.6 | 29.37±0.6 |
| | *Visually grounded speech models* | | | | | | | | |
| 4 | translation | Yorùbá | → English | beam search | 6.65±0.0 | 9.37±0.5 | 11.32±0.5 | 12.72±0.2 | 13.71±0.0 |
| 5 | translation | Yorùbá | → English | diverse | 6.10±0.0 | 9.54±0.6 | 12.28±0.9 | 14.22±0.4 | 15.82±0.0 |
| 6 | paraphrasing | English | → English | diverse | 6.56±0.5 | 10.45±0.8 | 13.10±0.7 | 15.45±0.4 | 17.46±0.9 |

- Row 1: Inter-annotator performance is moderate.
- Row 2: Audio-to-text model trained on groundtruth captions can perform better than humans.
  Model has access to the audio (can infer exact words), while humans only to images (semantics).
- Row 3: The captions generated by the image captioner are well aligned to the human annotations.
  BLEU score is a precision metric and favours simpler texts.
- Rows 4–5: Translation is modest, but intelligible (see qualitative results); diverse decoding helps.
- Row 6: Paraphrasing yields better results. FACC has five times more audio samples than YFACC.

## Impact of image captioning

- Image captioning systems: BLIP, BLIP2, GIT.
- Decoding techniques: beam search, multinomial sampling, diverse beam search.




- Beam search captions are most accurate, . . .
- but diverse captions are better for translation.

## Qualitative results

input audio (Yorùbá)

groundtruth transcript (Yorùbá)

| | | |
|---|---|---|
| *Ọkùnrin kan dúró leti omi nitosi àpáta.* | *Eniyan kan fò ninu afẹ́fẹ́.* | *Ọmọkùnrin kan ninu ṣokoto penpe pupa ti nmu bọ́ọ̀lù inu agbọn bọ́ọ̀lù lori pápá.* |

groundtruth translation (English)

| | | |
|---|---|---|
| *A man stands at the edge of the water near the rocks.* | *A snowboarder flies in the air.* | *A boy with red shorts is holding a basketball in a basketball court.* |

model prediction (English)

| | | |
|---|---|---|
| *A man standing on a rock with a camera.* | *A person jumping in the air on a skateboard.* | *A young boy is playing soccer on a field.* |

input audio (English)

groundtruth transcript (English)

| | | |
|---|---|---|
| *The brown dog is walking through a river surrounded by bushes.* | *Two women in white shirts talking.* | *A woman holding a small ball chasing after a small boy.* |

model prediction (English)

| | | |
|---|---|---|
| *A dog running through the water with its mouth open.* | *A couple of women standing next to each other.* | *A young boy holding a baseball bat on a field.* |

Code · https://github.com/danoneata/strim
Contact · dan.oneata@gmail.com
Paper · https://arxiv.org/pdf/2406.07133