Visually grounded models of spoken language

LOT School 2024

Grzegorz Chrupała

What are the sources of information children use to learn a language?

- Language-internal co-occurrence patterns statistics
- Correlations with objects and actions present in scenes co-occurring with speech.
- Clues such as gaze and pointing
- Inferring intentions of speakers
- Questioning
- Other incompletely understood mechanisms

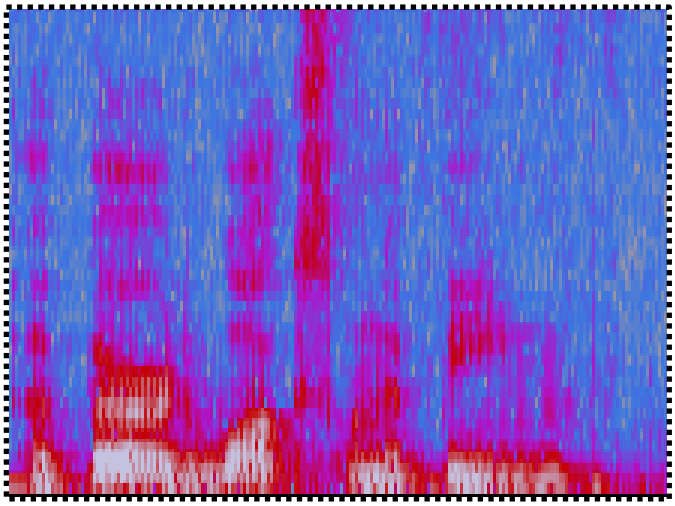
Clearly the first of these sources of information is not the only one.

Cognitive models are more plausible if they have exploit some of the non-linguistic information as well.

As AI systems mature they will increasingly act in the world and rely on multiple modalities their actions and interactions with humans.

Thus visual grounding, or perceptual grounding more in general could be useful from the AI engineering perspective also.

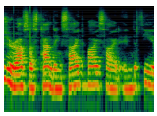
Image from http://rehabilitacionymedicinafisica.blogspot.com/2012/01/robot-y-frank-la-amistad-de-un-anciano.html



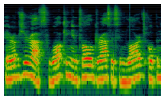
Today we will focus on one type of extralinguistic clue: grounding in the visual modality, and specifically on a very special scenario: models which learn representations of spoken language from datasets of utterances paired with still images depicting scenes these utterances describe.

Why?

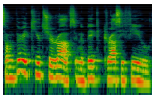
Visual modality is arguably the most important (compared to auditory, smell, touch)
Such datasets are relatively straightforward to collect.
Modeling is also relatively uncomplicated.
We have to start somewhere.



Two giraffes are standing side by side in the field.



Pair of giraffes walking in tall grasses in large open field.



Some very cute giraffes in a big grassy field.

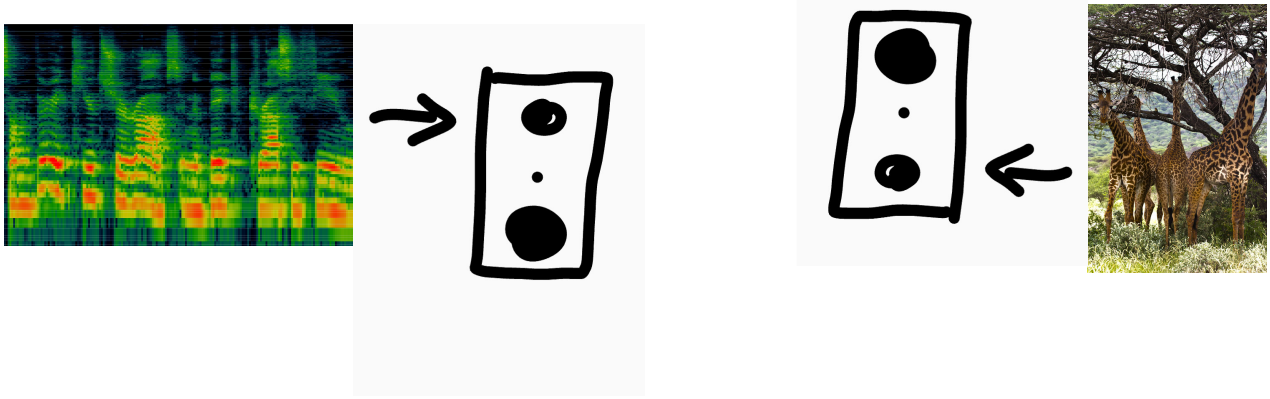
This is an example of from a dataset of images with spoken captions.

On the one hand, the supervision from the captions is **weak/noisy**:

- Relies on perceiving/understanding the image
- Number of giraffes is absent or incorrect
- Subjective judgments may be present (cute), which are not directly visible in the image
- Some aspects of the image may be mentioned (giraffe, grass), some not (trees)

On the other hand the supervision is **still unrealistically strong**:

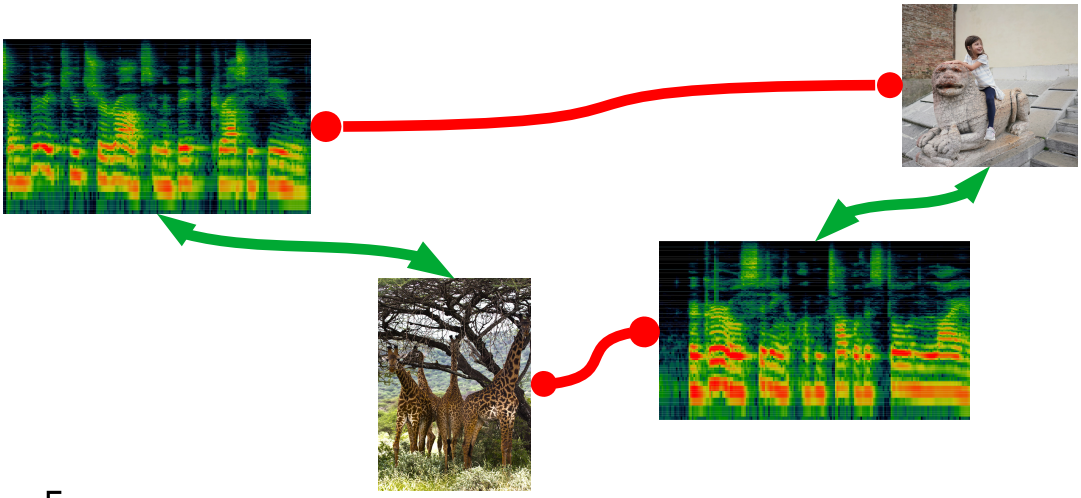
- In real life, speech may be loosely or not at all related to the visual scene
- In the data, speakers are randomly assigned, and speaker identity does not act as a confound.



The general architecture for visually grounded models has been to design two modality-specific encoders (possibly pretrained) which project the inputs into a shared representation space.

The models then learn to modify the projections such that the proximity of inputs make sense, that is it follows the matching of utterances and images in the training data.

$$\mathcal{L}(a, p, n) = \max\left(d(a, p) - d(a, n) + \alpha, 0\right)$$



$$\ell = \sum_{ui}\left[\sum_{u'}\max(0, S_{u'i} - S_{ui} + \alpha) + \sum_{i'}\max(0, S_{ui'} - S_{ui} + \alpha)\right]$$

Specifically, models use some kind contrastive, triplet-like loss function.

It can be similarity/distance based (such as the on the slide) or it can be a probabilistically formulated loss such as InfoNCE (https://paperswithcode.com/method/infonce).

In the simplest case (top) triplet loss is the difference between the distance between anchor and positive example and the distance between anchor and negative example (with a margin).

In the formulation below similarities are used as the reverse of distances, and the score is aggregated over multiple negatives (mimatched pairs).

Once we have a model which can learn to associate spoken utterances with images, how can we evaluate its language skills?

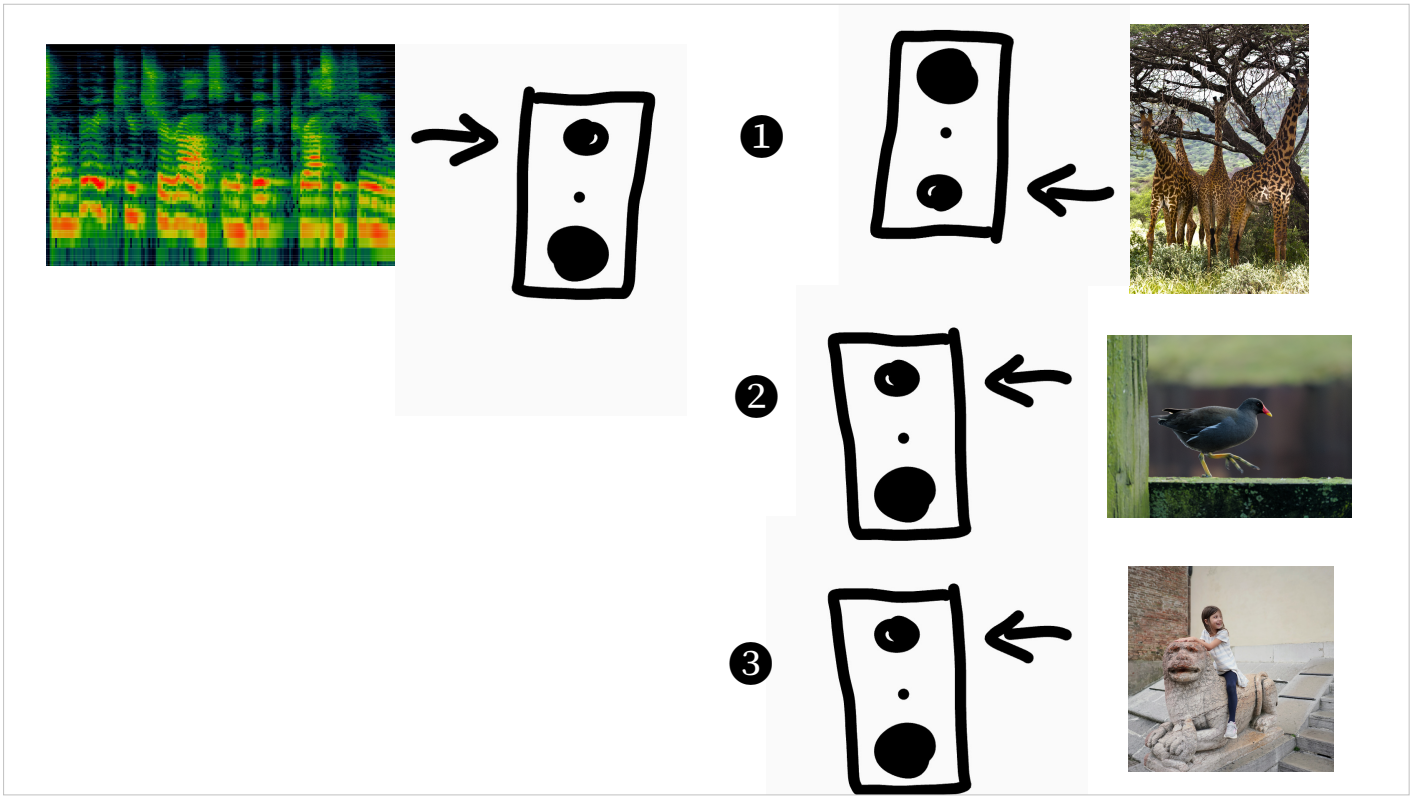
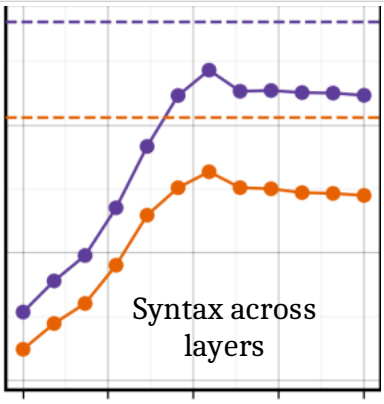
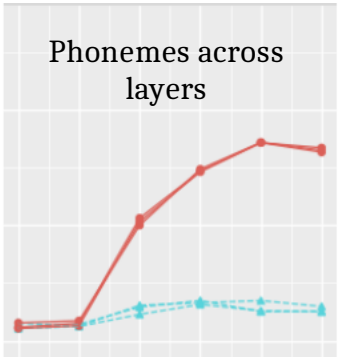


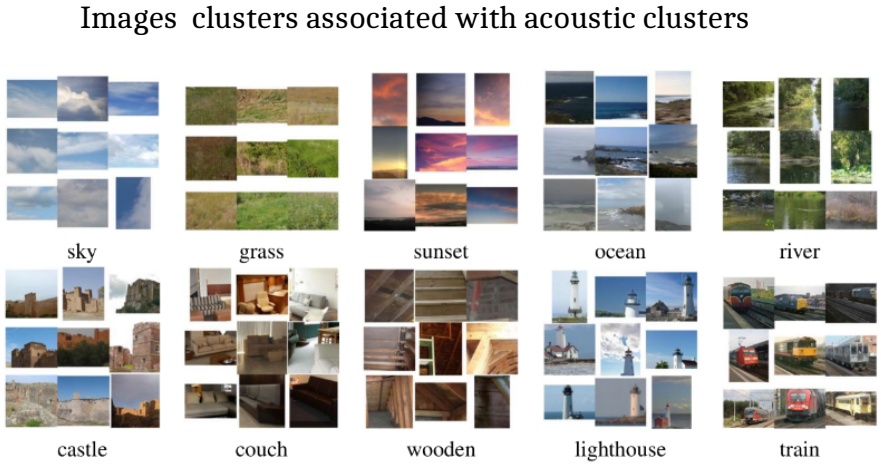
Image retrieval via captions (intrinsic evaluation).

We encode an utterance and a set of candidate images using the model, and rank the images in order or similarity to the utterance.

We then evaluate the quality of the ranking by measuring recall @ N, or median rank of the correct image.



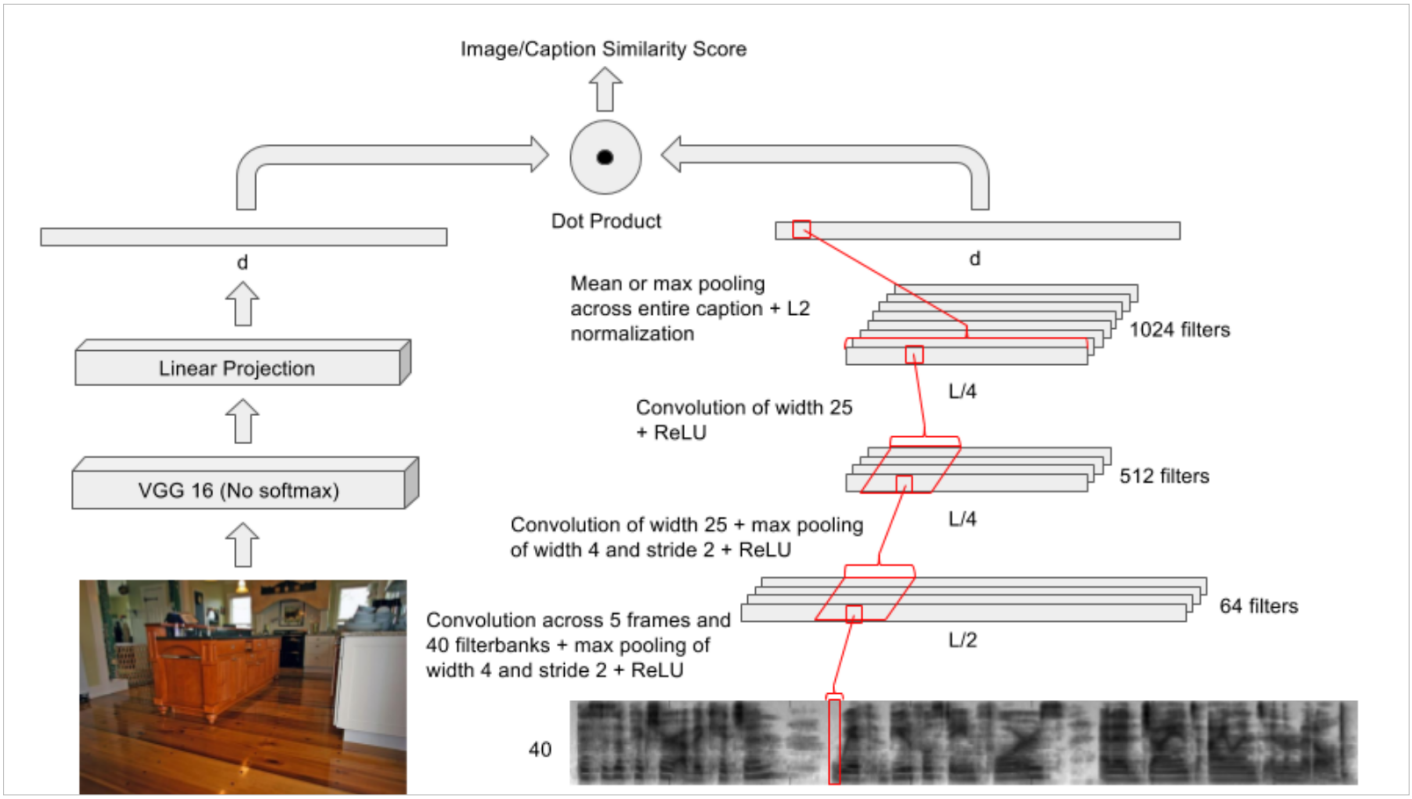




Intrinsic evaluation is often complemented with various types of analyses of the learned representations – this can be considered extrinsic evaluation.

Examples include:

- Encoding of syntax (https://doi.org/10.21437/Interspeech.2023-679)
- Encoding of phonemes (https://doi.org/10.18653/v1/2020.acl-main.381)
- Association of acoustic and visual patterns (https://doi.org/10.18653/v1/P17-1047, K-means clustering of modalities separately, then compute affinity between each pair of clusters by summing over the pairwise similarities for items in the pair of clusters)



Example encoder architecture, after https://proceedings.neurips.cc/paper/2016/hash/82b8a3434904411a9fdc43ca87cee70c-Abstract.html

Both the audio encoder and the image encoder use convolutional networks, with different settings.

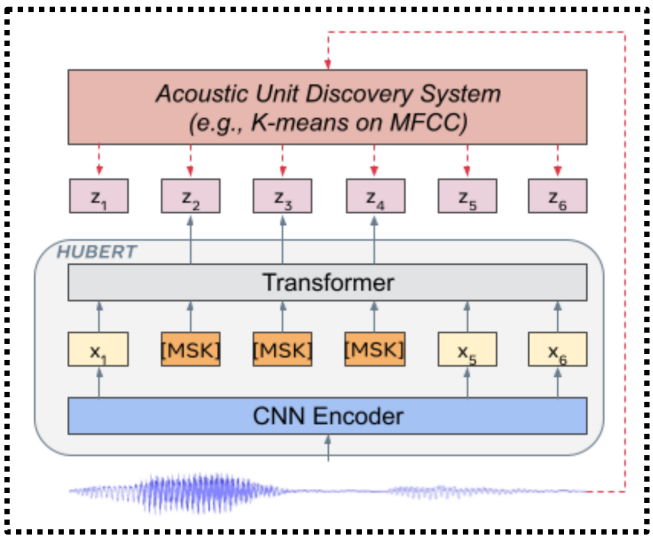
In other papers, such as https://doi.org/10.18653/v1/P17-1057 a recurrent audio encoder was used instead.

In this paper, and many other cases, the visual encoder is pre-trained in a supervised ways (on image classification).

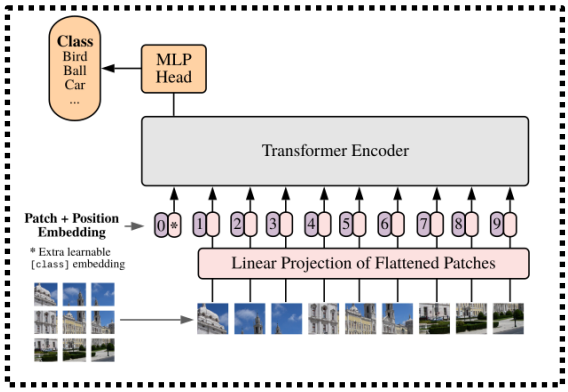
It has been subsequently shown that the general idea works also without pretraining, albeit performance is lower.

VG-HuBERT

HuBERT



ViT



More recent models use transformers in both encoders, notably in https://doi.org/10.21437/Interspeech.2022-10652 and in https://arxiv.org/abs/2305.11435

Otherwise the general approach is similar to what we have seen so far.

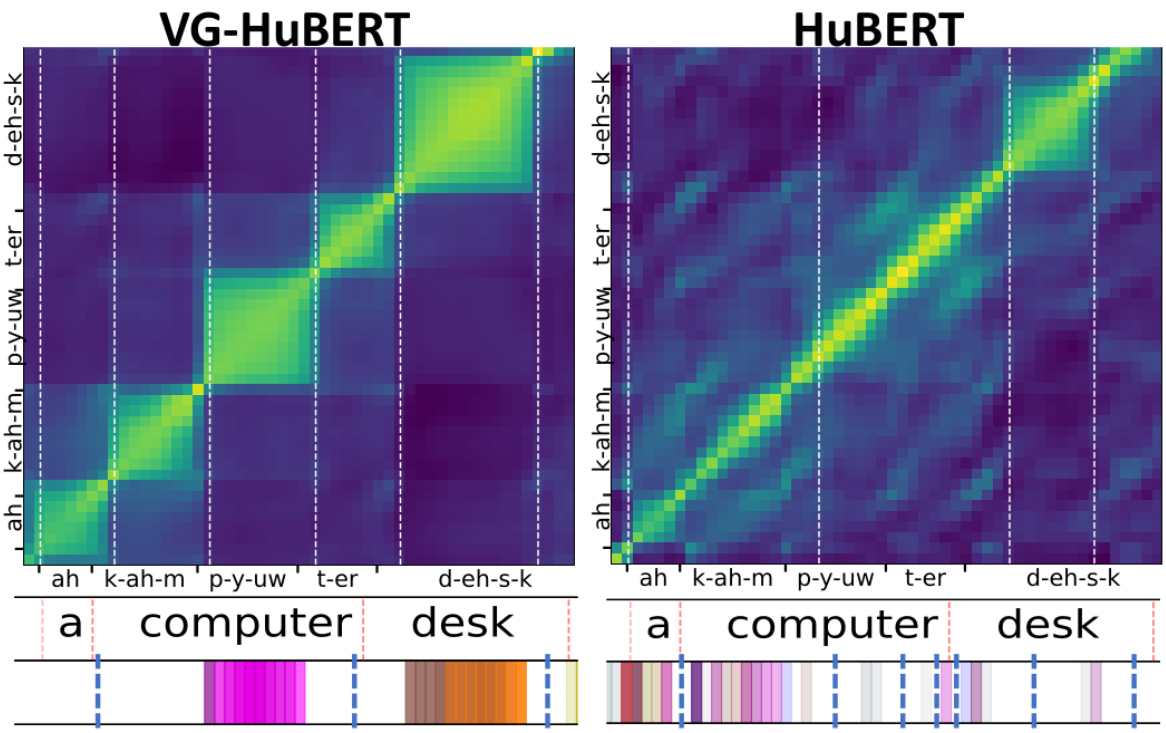
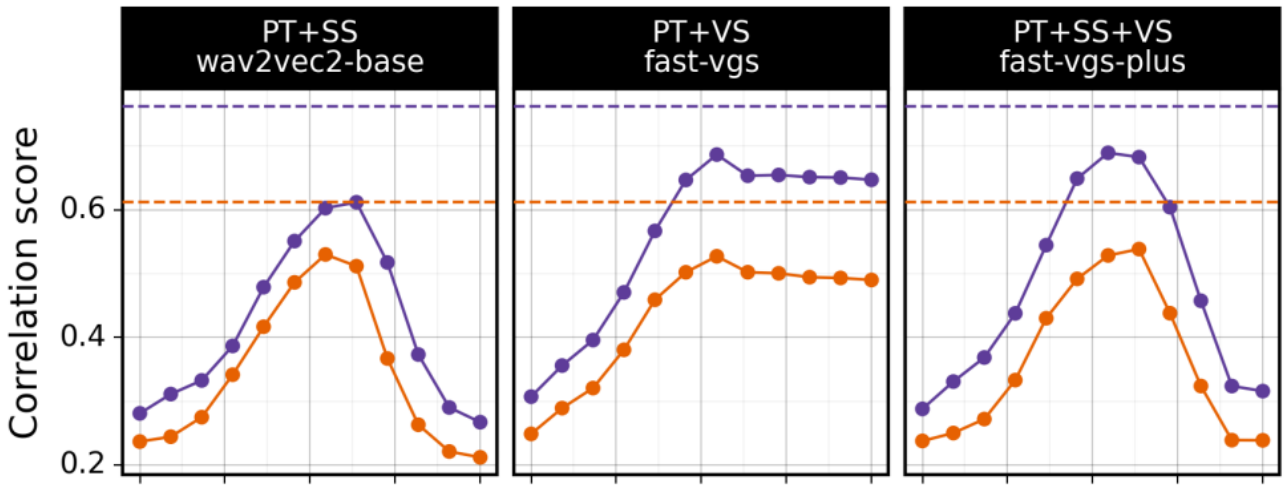


Figure from https://doi.org/10.48550/arXiv.2305.11435

"Visualization of feature self-similarity matrix (upper) and the attention (lower) in VG-HuBERT and HuBERT. The vertical white dotted lines are generated by minCutMerge, and vertical blue dotted lines are generated by taking the midpoint of boundaries of adjacent attention segments."

VG-HuBERT has been reported to show intriguing properties not present to the same extent in purely self-supervised models. The figure above illustrates the difference regarding the encoding of syllables and syllables boundaries. Peng and Harwath (2022) report analogous results with words.



The encoding of syntax has also been found to differ between self-supervised vs visually grounded models. This figure from https://doi.org/10.21437/Interspeech.2023-679 shows a laywewise comparison for wav2vec2 (self-supervision), fast-vgs (visual supervision) and fast-vgs-plus (self+visual supervision).

The models trained with no self-supervision (middle) does not display the drop in syntax encoding in the final layers, unlike the other two models.