

Learning English with Peppa Pig

Mitja Nikolaus

Aix-Marseille University

mitja.nikolaus@univ-amu.fr

Afra Alishahi

Tilburg University

a.alishahi@uvt.nl

Grzegorz Chrupała

Tilburg University

grzegorz@chrupala.me

Abstract

Attempts to computationally simulate the acquisition of spoken language via grounding in perception have a long tradition but have gained momentum in the past few years. Current neural approaches exploit associations between the spoken and visual modality and learn to represent speech and visual data in a joint vector space. A major unresolved issue from the point of ecological validity is the training data, typically consisting of images or videos paired with spoken descriptions of what is depicted. Such a setup guarantees an unrealistically strong correlation between speech and the visual world. In the real world the coupling between the linguistic and the visual is loose, and often contains confounds in the form of correlations with non-semantic aspects of the speech signal. The current study is a first step towards simulating a naturalistic grounding scenario by using a dataset based on the children’s cartoon *Peppa Pig*. We train a simple bi-modal architecture on the portion of the data consisting of naturalistic dialog between characters, and evaluate on segments containing descriptive narrations. Despite the weak and confounded signal in this training data our model succeeds at learning aspects of the visual semantics of spoken language.

1 Introduction

Attempts to model or simulate the acquisition of spoken language via grounding in the visual modality date to the beginning of this century (Roy and Pentland, 2002) but have gained momentum recently with the revival of neural networks (e.g. Synnaeve et al., 2014; Harwath and Glass, 2015;

Harwath et al., 2016; Chrupała et al., 2017; Alishahi et al., 2017; Harwath et al., 2018; Merckx et al., 2019; Havard et al., 2019a; Rouditchenko et al., 2020; Khorrami and Räsänen, 2021; Peng and Harwath, 2021). Current approaches work well enough from an applied point of view, but most are not generalizable to real-life situations that humans or adaptive artificial agents experience. Commonly used training data consist of images or videos paired with spoken descriptions of the scene depicted. The type of input that a language learner receives from its environment is much more challenging. Firstly, speech is only loosely coupled with the visual modality. Secondly in addition to correlations between the visual scenes and the *meaning* of spoken utterances, there are also correlations with non-semantic aspects of the speech signal, such as the voice of specific speakers, as well as with non-speech ambient sounds. Although it is plausible that such non-semantic correlations can sometimes be useful to the learner in the general endeavour of making sense of the world, for the specific task of learning the semantics of linguistic units they are likely more often an obstacle, as they make it harder to zoom in on the meaning-bearing aspects of the audio signal.

In the current study we make a first step towards simulating the acquisition of language via grounding in perception in a more naturalistic scenario. Our main focus is on learning the meaning of individual words as well as multi-word expressions from spoken utterances grounded in video. We use the well-known children’s cartoon *Peppa Pig* as a case study as a source of training and evaluation data. Compared to commonly used video datasets, this data has a number of interesting characteris-

tics. The visual modality is very schematic, and the language is also simple in terms of vocabulary size and syntactic complexity. Crucially, however, most of the speech in the videos consists of naturalistic dialogs between the characters. The utterances are only loosely and noisily correlated to the scenes and actions depicted in the videos.

This choice of data thus allows us to directly address the ecological limitations of the current approaches. In addition, the cartoon videos also contain comments interjected by the narrator. We use these for a evaluation of the acquisition of meaning as they are more descriptive and less noisy and allow us to measure performance while controlling for speaker characteristics.

One further aspect of learning spoken language via visual grounding is the fact that grounding is only part of the story. Human children arguably infer substantial amounts of information about language structure and meaning from purely linguistic co-occurrence statistics. A similar mechanism is what allows written language models such as BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020) to capture and exhibit relatively sophisticated linguistic knowledge. Loosely similar approaches are starting to also make an impact for the spoken modality (e.g. Baevski et al., 2020; Hsu et al., 2021). Here we take a simple pre-training based approach to integrating this type of self-supervision with learning-via-grounding.

Our contributions are the following:

- We implement a simple bi-modal architecture which learns spoken language embeddings from videos;
- We evaluate model performance in terms of video fragment retrieval and additionally design controlled evaluation protocols inspired by the intermodal preferential looking paradigm (Hirsh-Pasek and Golinkoff, 1996);
- We carry out ablations of model components in order to understand the effects of pre-training for the audio and video encoders, the role of temporal information, and of segmentation strategies while training.

We show that despite the challenges of our naturalistic training data our model succeeds at learning associations between the form of spoken utterances and their visual semantics.

Our findings include the facts that temporal information contributes substantially to video modeling, and that unsupervised pre-training followed by fine-tuning of the audio encoder is key to the best performance.

2 Related Work

Early attempts at simulating grounded language learning focus on interactions between adults and young children while playing with a set of objects from different categories (Roy, 1999, 2000a,b, 2002; Gorniak and Roy, 2003; Mukherjee and Roy, 2003). In a representative study from these series, Roy and Pentland (2002) use speech recorded from such interactions paired with different views of the visible objects to identify linguistic units (i.e. words) and visual categories, and to map these two modalities together. A hard-coded visual system extracts object representations from images, and spoken utterances are represented as phoneme probabilities generated by a RNN pre-trained on spectrograms. Their experiments on small-scale data (around 20 words and seven visual categories) show that the model can segment words and map them to visual categories.

2.1 Spoken Language Grounded in Images

The availability of datasets of images associated with spoken captions such as Flickr Audio Captions (Harwath and Glass, 2015), Places (Zhou et al., 2014) and Spoken COCO (Hsu et al., 2019) led to a rapid development of deep models of grounded language learning; see Chrupała (2021) for a comprehensive overview. Most of these models adapt the architecture of Karpathy et al. (2014) by implementing separate pathways for encoding the visual and speech modality, and mapping these encodings into a joint representation space. Visual features are extracted from a pre-trained image classification model that processes the whole or a specific region of an image (however see Harwath et al. (2018), who train the model end-to-end on images and their spoken captions on the Places dataset). The audio encoder component in most models is either an adaptation of Harwath et al. (2016) which feeds a spectrogram of the speech signal to a convolutional architecture, or a hybrid architecture of convolutional followed by recurrent layers using Mel-Frequency Cepstral Coefficient (MFCC) features from the audio signal as input as introduced by Chrupała et al. (2017).

GC:

Can we support this assertion in a quantitative way?
MN:
Difficult I think.
To measure semantic correlation between images/videos and speech we need to pass them through a model. So we could run our model on a different dataset (e.g. spoken captions) and show that it performs better there, but we wouldn't have directly comparable evaluation

MN:
maybe highlight here the difference of these models to the early approaches by deb roy (e.g. handling larger + more realistic data, ..)?

Models of speech grounded in images are optimized for and evaluated on image retrieval from spoken caption and vice versa. Additionally, a range of diagnostic analyses have been performed on the hidden representations of these models to study whether they encode knowledge about the identity and boundaries of subword units such as phonemes and syllables (Alishahi et al., 2017; Harwath and Glass, 2019; Khorrami and Räsänen, 2021) as well as individual words (Chrupała et al., 2017; Havard et al., 2019b). Moreover, in addition to examining form-meaning associations at the utterance level, Harwath and Glass (2017) explicitly learn a lexicon by extracting audio and image segments, clustering each modality separately, and mapping them together by calculating the pairwise similarities of their members in the joint semantic space.

2.2 Spoken Language Grounded in Video

There have been also recent attempts to learn spoken language grounded in video instead of static images. Boggust et al. (2019) sample audio-visual fragments from cooking videos, however their grounded model treats video frames as still images and discard their temporal order. Rouditchenko et al. (2020) integrate the temporal information when encoding videos from the Howto100m dataset (Miech et al., 2019), and perform better than previous work in language and video clip retrieval. Models trained on such instructional video datasets often do not generalize well to other domains. Monfort et al. (2021) highlight this limitation and show that training on their larger and more diverse Spoken Moments in Time dataset leads to better generalization. The point remains that these video datasets contain descriptive speech, thus ensuring that there is a strong correlation between the spoken language and their visual context, a characteristic that is not representative of the experience of learning language in the real world.

Child language learning from video. There are many studies on young children learning language by watching videos; see Vanderplank (2010) for a survey. The main takeaway of these studies is that language learning is much more effective in a social, conversational setting than by passively watching videos (Kuhl et al., 2003; Ander-

son and Pempek, 2005; Robb et al., 2009),¹ but learning does happen in such contexts. Importantly for our goal, techniques such as the inter-modal preferential looking paradigm have been developed to systematically test young language learners’ knowledge of words, syntactic structure and semantic roles (Hirsh-Pasek and Golinkoff, 1996; Bergelson and Swingley, 2012; Noble et al., 2011). Nikolaus and Fourtassi (2021) employ this evaluation strategy to test semantic knowledge at word and sentence level in their computational model of word learning from images. We adapt this approach to evaluate how our grounded model associates semantic information to spoken words and utterances from video.

Self-supervised learning from multimodal data.

3 Method

The main focus of this study is on the data and evaluation. We thus keep the components of our architecture simple, and follow established modeling practices whenever possible.

3.1 Dataset

We use the dataset provided by Papasrantopoulos and Cohen (2021) containing a metadata for the set of 209 episodes (seasons 1–5) of the English-language version of *Peppa Pig*. We purchased the corresponding Peppa Pig episodes in the form of DVDs.

These annotations created by Papasrantopoulos and Cohen (2021) feature written transcriptions aligned with the audio as well as segmentation into *dialog* and *narration*.² Dialogs are the parts spoken by the characters, while narrations are comments inserted by the narrator, which are more descriptive in nature. All the narration segments are uttered by the same voice actor. We use the dialogs for training the model, and set aside the narrations for evaluation purposes only. A small portion of the dialog data is also used for validation. Specifically, out of the total 209 episodes, we use dialog from episodes 1–196 for training, and

¹Adding social interaction while watching videos improves learning; Lytle et al. (2018).

²It should be noted that the quality of the alignment and segmentation in the original dataset is variable. In cases where exact alignment is needed, such as for word-level analyses, we re-align the transcriptions using github.com/lowerquality/gentle.

GC: I think we need a paragraph or so on recent work on self-supervised learning from multimodal data. Some possible works to mention: <https://arxiv.org/abs/2006.04839>, <https://arxiv.org/abs/2006.04839>.

197–209 for validation. We set aside narrations from episodes 1–104 for validation and 105–209 for testing. Table 1 shows the sizes of the training and validation splits.

Split	Type	Size (h)
train	dialog	10.01
val	dialog	0.66
val	narration	0.94
test	narration	0.64

Table 1: Duration in hours of the dataset splits.

3.2 Preprocessing

Our model is trained to discriminate positive video-audio pairs from negative ones. The positive pairs are those that are temporally coincident in the original video file. In order to generate these training items we need to split the videos into fragments. For segmenting data for training, we *do not* use word or sentence-level subtitle alignment in order to make the setting naturalistic. Processing long segments of video and audio is not tractable on commodity GPU hardware, and we thus segment the data into brief snippets roughly comparable in length to the duration a short sentence or a phrase. We use the following two segmentation strategies:

Fixed Using this approach we simply split sections into fixed-length non-overlapping fragments of 2.3 second duration. This length is close to the mean duration of audio aligned to a single line of subtitles.

Jitter In this approach the mean duration of the segments is the same (2.3 seconds) but we randomly vary the length of the video, and independently, of the corresponding audio around this average duration. This means that (i) the segments can be partially overlapping and (ii) the video and the audio it is paired with are normally of different length. Specifically we sample the fragment duration d (in seconds) from the following distribution:

$$d \sim \min(6, \max(0.05, \mathcal{N}(2.3, 0.5))) \quad (1)$$

The video is subsampled to 10 frames per second, and to 180×100 resolution.³ The audio is

³Performance is better with higher resolution (we tried 360×200), but it makes GPU memory requirements prohibitive.

converted to mono by averaging the two channels and the raw waveform is used as input. We use the original sample rate of 44.1 kHz (instead of downsampling to the 16 kHz sample rate used for pre-training WAV2VEC2) as we found out that this helps with generalization performance on the narration validation data.

For evaluation we have a number of different conditions and evaluation metrics described in detail in Section 3.4 and in some of these conditions we use the subtitles to guide segmentation.

3.3 Model Architecture

We adapt the high-level modeling approach from work on spoken image-caption data (Harwath et al., 2016; Chrupała et al., 2017): our objective function is based on a triplet-like contrastive loss with margin which encourages the matching audio and video clip to be projected nearby in the embedding space, and mis-matching audio and video clips to be far away:

$$\ell = \sum_{av} \left[\sum_{a'} \max(0, S_{a'v} - S_{av} + \alpha) + \sum_{v'} \max(0, S_{av'} - S_{av} + \alpha) \right] \quad (2)$$

where α is a margin, S_{av} is a similarity score between a matching audio-video clip pair, and $S_{a'v}$ and $S_{av'}$ denote similarity scores between mis-matched pairs, i.e. negative examples from the current batch. Our heuristic to generate positive and negative examples is very simple: we consider the example positive if the audio is aligned with a video clip in our data. Other pairs of audio-video clips are considered negative.

3.3.1 Audio Encoder

The audio encoder portion of the model consists of a small wav2vec2 model (Baevski et al., 2020) pre-trained in a self-supervised fashion, without any supervised fine tuning.⁴ The WAV2VEC 2.0 architecture learns audio embeddings by self-supervised learning driven by a contrastive loss applied to quantized latent representations of masked frames, loosely inspired by the BERT approach to language modeling (Devlin et al., 2019).

⁴Available from https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt.

The output of this module is a temporal sequence of 28-dimensional vectors. We pool this output across time using an attention mechanism with dimension-wise weights (Merx et al., 2019):

$$\begin{aligned} \mathbf{A} &= \text{softmax}_t(\text{MLP}(\mathbf{X})) \\ \mathbf{z} &= \sum_t (\mathbf{A}_t \odot \mathbf{X}_t), \end{aligned} \quad (3)$$

where \mathbf{X} is the tensor with the encoder output vectors for each time-step t : an MLP followed by a time-wise softmax is used to compute an attention weight for each time step and for each dimension. The pooling is followed by a linear projection and L_2 normalization. For our experiments we also use versions of the encoder where the wav2vec weights are frozen, as well a randomly initialized rather than pre-trained.

3.3.2 Video Encoder

As a video encoder we use the 18-layer ResNet (2+1)D architecture (Tran et al., 2018) pretrained on the action recognition dataset Kinetics-400 (Kay et al., 2017). The pre-trained model is available via Pytorch.⁵ This architecture implements 3D convolution by decomposing it into a 2D spatial convolution followed by 1D temporal convolution. The output of this module is aggregated using the attention mechanism with the same architecture as for the audio module, linearly projected to the same dimensionality as the audio (512) and L_2 normalized. For our experiments we also use a version of the video encoder without pre-training.

STATIC baseline As a baseline to investigate the contribution of temporal information to video modeling we swap the video ResNet (2+1)D with the D2 ResNet pre-trained on ImageNet, which embeds each video frame separately. These frame embeddings are then attention-pooled as with the standard video encoder.

3.4 Evaluation

The most common approach to evaluation for visually grounded models trained on spoken image captions is caption-to-image retrieval (often combined with image-to-caption retrieval): in fact this technique has been carried over from text-based image-caption modeling. With the standard spoken caption dataset this approach is unproblematic since the content of the captions is not corre-

lated with extra-linguistic clues in the speech signal, such as speaker identity (since speakers are randomly assigned to captions) or non-speech environmental sounds. In such an artificial setting, a retrieval metric measures the ability of the model to match spoken utterances to images based on their semantic content. This is not the case for the *Peppa Pig* dataset: here we can expect that when a video segment depicts a particular character (e.g. George) then the audio in this segment is more likely to contain utterances spoken by the voice actor playing George. Moreover, some characters might have a tendency to talk about certain topics more often than others, and the model might pick up on these associations instead of paying attention to the actual meaning of the uttered words. Due to these factors, in a naive retrieval setting, a model could obtain a high score by mostly capturing these non-linguistic correlations.

In order to control for these factors we leverage the narrator speech in the videos. These utterances are always spoken by the same actor, so speaker identity cannot be used as a clue for matching video and audio. Furthermore, the narration segments are akin to video captions in that they tend to describe what is happening in the video and thus their semantic content is more strongly correlated with the content of the video than in the case of the dialog, which is also a desirable feature for the purposes of system evaluation.

3.4.1 Video Retrieval

For the retrieval evaluation, as for training, we use the FIXED and JITTER segmentation strategies. We encode each audio clip in a candidate set sampled from the validation (or test) data using the speech encoder part of the model; we encode each video clip using the video encoder. We then measure cosine similarity between the audio clip and all the video clips. If the video clip corresponding to the audio is among the n most similar video clips, we count that as a success. The proportion of successes across all audio clips gives us the retrieval metric known as $\text{recall}@n$: specifically in this paper we focus on $n = 10$. We set the candidate set size to 100, and thus the random baseline for the $\text{recall}@10$ is 10%. In order to quantify uncertainty in this evaluation due to the test data we repeat this procedure 500 times with randomly sampled candidate sets and visualize the score distribution.

⁵See <https://pytorch.org/vision/0.8/models.html#resnet-3d>.

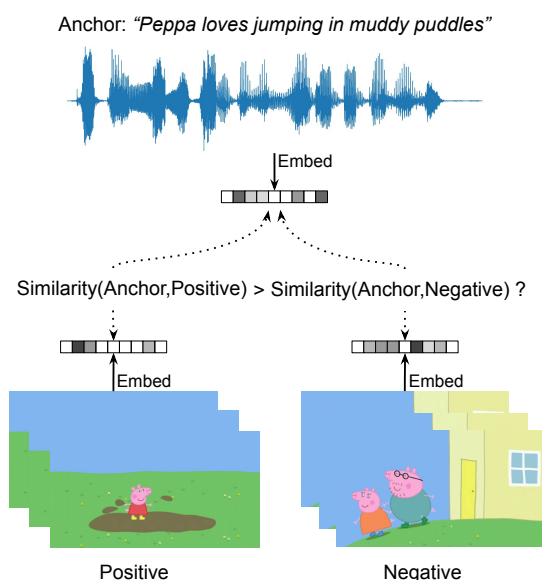


Figure 1: Triplets Evaluation: Given a reference audio sequence (anchor), we measure the model’s performance at choosing the matching video (positive) over a random distractor video (negative).

3.4.2 Triplets

Retrieval metrics such as recall@10 have some disadvantages. Firstly the absolute value of this metric may be hard to interpret as it depends on the size of the candidate set. Secondly, if we wanted to compare model performance with human performance, we could not feasibly ask human participants to provide the quadratic number of audio-video similarity judgments needed. For these reasons, we evaluate model performance using a more simple and controlled scenario, inspired by intermodal preferential looking paradigms in child language acquisition (Hirsh-Pasek and Golinkoff, 1996).

We extract clips aligned to a single subtitle line, group them by length, and for each pair of same-length video clips⁶, we extract the audio from one of them (selected at random) – this is our *anchor*. The video clip from which the anchor was taken is the *positive* one, which the other video clip is the *negative* one. This triplet of stimuli form a single test item. We use the model’s audio encoder to encode the anchor, and the video encoder to encode both video clips. We then check whether anchor is more similar to the positive or negative clips in terms of cosine similarity (see also Figure 1).

⁶To keep test items independent, the pairing of video clips is done such that each clip only occurs as a member of a single triplet.

More precisely, *triplet accuracy* is the mean over all triplets of the following quantity:

$$\frac{\text{signum}(\cosine(A, P) - \cosine(A, N)) + 1}{2} \quad (4)$$

with A being the anchor, P positive and N negative. The triplet accuracy metric is inspired by the ABX score of Schatz (2016). For triplet accuracy, regardless of the specific set of test items, we expect random-guessing performance to be at 0.5, and perfect performance to be 1.0. For this metric we also quantify uncertainty by resampling the triplets 500 times from the dataset, and display the score distribution.

3.4.3 Minimal Pairs

While the triplet evaluation gives us a general idea about whether the model has learned a mapping between videos and audio, the metric does not provide insight into whether the model has acquired the grounded semantics of specific words.

To address this question, we probe the model’s performance in a more targeted triplet setup, where the model is required to select the correct video from a pair of videos with *minimal differences* regarding a target word in the corresponding transcripts. We pair every triplet with a corresponding counter-example to control the evaluation for linguistic biases in the dataset (see also Nikolaus and Fourtassi, 2021).

To construct the evaluation set, we search the transcripts of the validation data for phrases with minimal differences with respect to the most commonly occurring (at least 10 times) nouns, verbs, and adjectives. We set the minimum phrase duration to 0.3 seconds.⁷

Based on the video and audio of each pair of phrases, we create two counter-balanced test trials (example and counter-example), as depicted in Figure 2. Here, the anchor A_{example} of the example triplet is the audio of *Peppa loves jumping*, the positive video P_{example} is the corresponding video, and the negative video N_{example} is the video corresponding to *George loves jumping*. In the counter-example triplet, the anchor $A_{\text{counterex}}$ is the audio of *George loves jumping*, and the positive and negative video are flipped.

We measure per-word accuracy by calculating the triplet accuracy (cf. Equation (4)) for all

⁷For shorter sequences, we do not expect that the video contains enough semantic information to distinguish target and distractor. A phrase can also be a single word.

MN: that might be confusing for the reader: we first say it’s inspired by preferential looking and then by ABX score. Maybe we can only say once that’s inspired by the two of them?

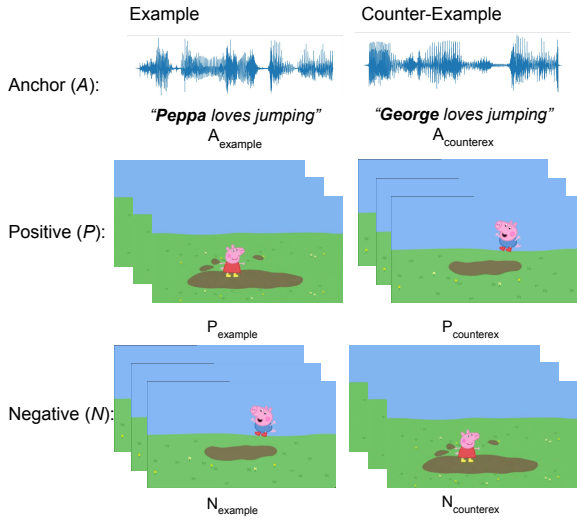


Figure 2: Minimal Pairs Evaluation

triplets that contain a given word (e.g. *Peppa*) either as target or distractor word, i.e. cases in which the model needs to succeed in either choosing a video containing the given word (the example triplet in Figure 2) or rejecting a video containing the given word (the counter-example triplet in Figure 2). We report accuracy for all words (nouns and verbs) for which we find at least 100 pairs of triplets.⁸

4 Experimental Settings

We implement the architecture in PyTorch (Paszke et al., 2019). We use the Adam optimizer (Kingma and Ba, 2015) with the scheduling described in (Devlin et al., 2019). We train every configuration on a single GPU and stop training after 48 hours, with batch-size 8 and accumulating gradients over 8 batches, in 16 bit precision mode. For each model configuration we save model weights after each epoch and report results for the checkpoint which gets the best triplet accuracy on the narration validation data. Our code is publicly available at github.com/gchrupala/peppa, and can be consulted for further details of the experimental setup.

5 Results

In Section 5.1 we first focus on understanding the impact of the main components of our training setup on performance as measured by recall@10 and triplet accuracy. Afterwards in Section 5.2 we

⁸We did not find enough examples for any adjectives in the data to perform an evaluation on them.

switch our attention to the targeted evaluation via minimal pairs.

5.1 Ablations

For completeness, we report results on both the dialog and narration data. However, it should be kept in mind that in the case of the narration data the scores are not confounded by speaker-based clues and thus the scores on narration are those that indicate to what extent to model learns aspects of utterance meaning.

5.1.1 Pre-training and Fine-tuning

Results on different pre-training configurations are shown in Figure 3.

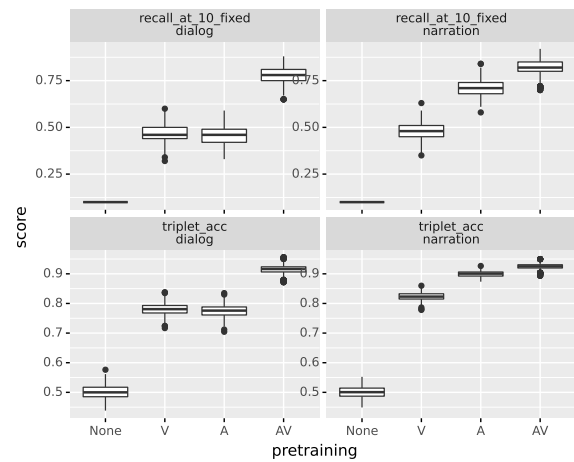


Figure 3: Effect of pre-training on performance on the dialog and narration validation data. The top row shows recall@10; the bottom row triplet accuracy.

The best overall performance on both the dialog and the narration data is achieved with a model where both the video and audio encoder are pre-trained before being fine-tuned on our data. The model with no pre-training in either modality failed to converge and thus performs at random.

On narration data, for both metrics, we see a clear ranking of configurations from best to worst: (AV) audio and video pre-training, (A) audio pre-training, (V) video pre-training and (None) no training. Meanwhile for dialog data, the performance between A and V is comparable. In this absence of any pre-training (None) the model failed to converge and therefore performs at chance level.⁹

To further understand and disentangle the effects of audio pre-training and fine-tuning, we

⁹Some of the runs of this configuration did perform above chance but training is in general unstable.

Anonymity violation (for the final submission).

GC: We really should have at least 3 runs of each configuration with a different random seed.

GC: Generate and insert results on test data

train a model with frozen parameters of the WAV2VEC module. The effect of this condition is shown in Figure 4. We find without fine-tuning

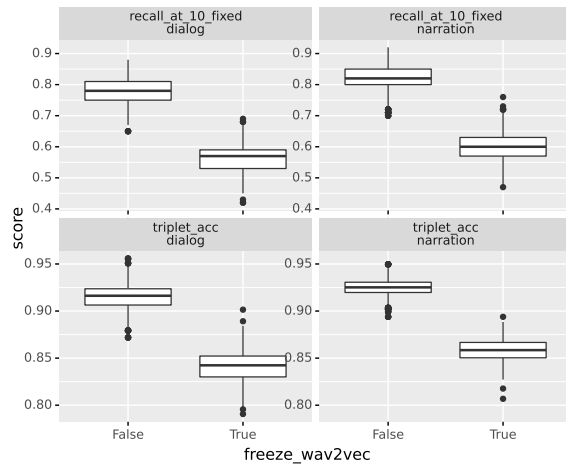


Figure 4: Effect of freezing the parameters of the WAV2VEC module on model performance, on the dialog and narration validation data. The top row shows recall@10; the bottom row triplet accuracy.

of the WAV2VEC module, performance decreases substantially on both metrics. In other words, best performance is only achieved with pre-trained and fine-tuned models.

5.1.2 Jitter

Next, we evaluate a model that has been trained with varying video and audio lengths (JITTER). For fair comparison, we report recall@10 for both FIXED and JITTER validation configurations. As seen in Figure 5, the effect of JITTER is only minor and that performance is comparable.

5.1.3 Temporal Information

Finally, we explore the role of the temporal nature of the visual modality. Figure 6 compares the model with the regular video encoder with one using the STATIC baseline encoder. Across all metrics, we observe substantial performance drops for the STATIC model, which has access to the same video frames, but does not have access to their temporal ordering.

5.2 Minimal Pairs

Table 2 presents results for the minimal pair evaluation. We find that a models that are pre-trained and fine-tuned with JITTER perform best. Generally, there is not much difference in the scores for verbs and nouns, with the exception of models trained on STATIC data. Here the performance for

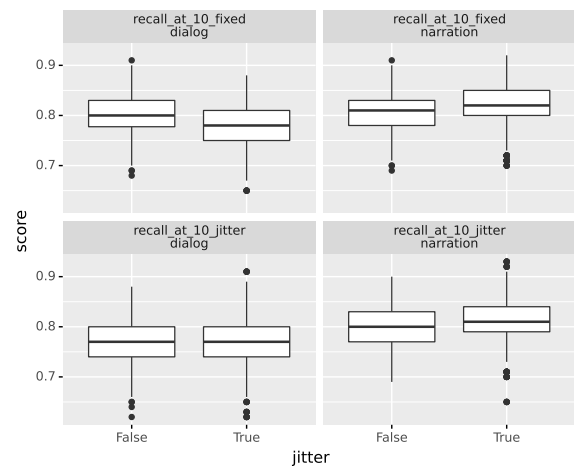


Figure 5: Effect of jitter on model performance, on the dialog and narration validation data. The top row shows recall@10; the bottom row triplet accuracy.

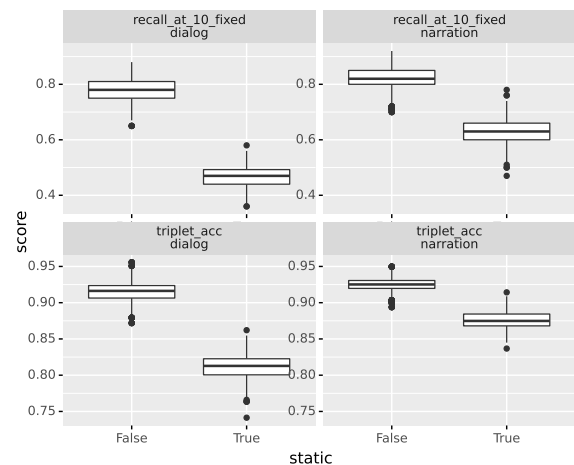


Figure 6: Effect of temporal information on model performance, on the dialog and narration validation data. The top row shows recall@10; the bottom row triplet accuracy.

verbs is worse, confirming that temporal information is more important for the semantics of verbs.

Figure 7 and 8 show per-word accuracy for nouns and verbs for the best performing model. We perform bootstrapping ($n_{\text{resampling}} = 100$) to estimate mean and standard deviation for each accuracy score.

We observe substantial variance in the accuracy scores, suggesting that the difficulty to learn certain words varies. For example, the scores for “house” and “cake” are very low. This could be because these concepts are not easy to ground, as they do not often refer to a similar visual entity. When looking at our evaluation examples, we find

GC: Insert reference to results with jitter for minimal pair scenario.

GC: Note that a confounding factor here is that the static encoder is pre-trained on ImageNet rather

Keep table? Add more rows?

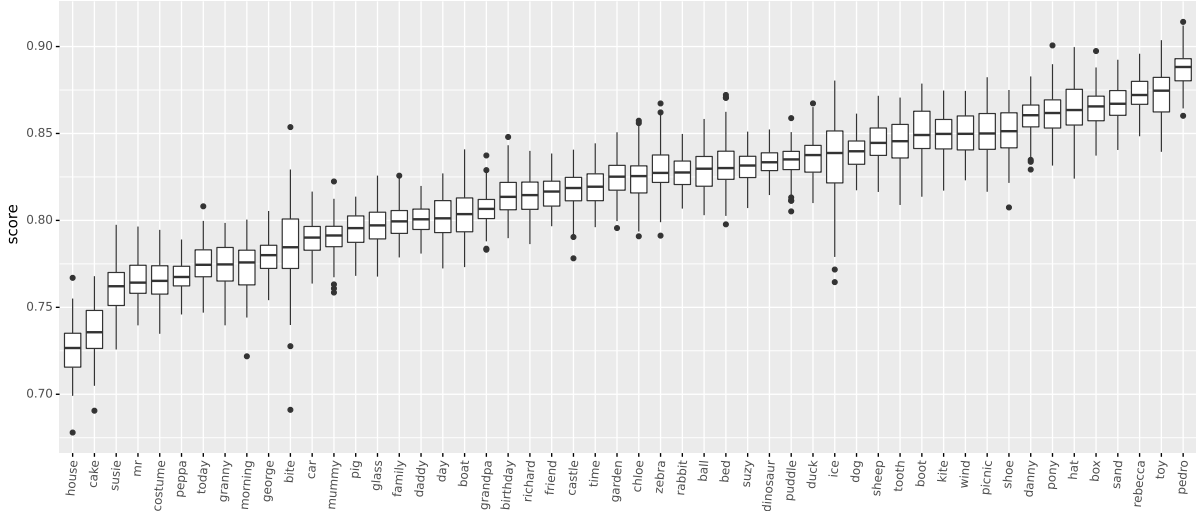


Figure 7: Per-word accuracies on the minimal pairs evaluation data for nouns.

Finetune	Jitter	Static	Accuracy	
			Nouns	Verbs
✓	✓		0.81	0.80
✓			0.73	0.73
	✓		0.71	0.70
✓	✓	✓	0.79	0.76

Table 2: Minimal pair accuracies for different model configurations.

that indeed the word “house” is used in varying visual contexts (house entrance, view of the whole house, inside the house, rabbit’s house) and in displaced speech (talking about going to somebody’s house). Regarding “cake”, it refers to either a whole cake, a slice, dough, or crumbs.

On the other end, the words “Pedro” and “Rebecca” are learned very well: They refer to “Pedro pony” and “Rebecca rabbit”, easily visually distinguishable from other characters which are mainly pigs. Additionally, it is usually present and central to the video when mentioned.

Further investigations with larger datasets are necessary to reveal the underlying reasons for difficulty, and relating them to predictors of age of acquisition in the child language acquisition literature (Roy et al., 2015; Frank et al., 2021).

6 Conclusion

Our results suggest that despite the challenges inherent to the naturalistic aspects of the *Peppa Pig* dataset, a simple bimodal architecture trained on

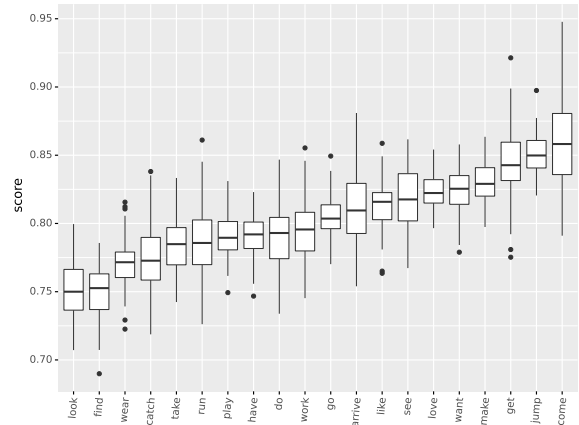


Figure 8: Per-word accuracies on the minimal pairs evaluation data for verbs.

it generalizes well on narrative utterances featuring an unseen speaker and a descriptive rather than conversational style. We saw that generalization is substantially boosted by fine-tuning audio representations pre-trained on unlabeled single-modality speech data. Fine-tuning a pre-trained video encoder also makes a contribution, but is less crucial to generalization from dialog to narration.

6.1 Limitations and future work

Our setting models the acquisition of linguistic knowledge both language-internal correlations as well as from grounding in vision in a simplistic way: we fine-tune an audio encoder pre-trained on read English speech. In future, it would be interesting to make the setting more realistic by using

pre-training data which reflect a young learner's experience more closely, and to interleave learning via self-supervision from speech and via grounding in vision as happens with human learners also.

Regarding the video encoder, ideally we would want to dispense with supervised pre-training and rather use a model pre-trained in a self-supervised way also for this modality.

In order to learn into what aspects of spoken language our model acquires, we would like to carry out in-depth probing-type analyses of learned representations on sub-word, lexical, and phrasal levels. It would also be worthwhile to figure out the details of how specifically temporal information in video contributes to acquiring linguistic knowledge.

Acknowledgements

We would like to thank Nikos Papasrantopoulos and Shay B. Cohen for creating the Peppa Pig annotations and for sharing them with us. Part of the NWO/E-Science Center grant number 027.018.G03 was used to purchase the video data on DVD. Thanks to Bertrand Higy for taking care of this purchase, as well as for sharing his ideas with us in the initial stages of this work. We also thank Abdellah Fourtassi for his useful feedback.

This work, carried out within the Labex BLRI (ANR-11-LABX-0036) and the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX)

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille - A*MIDEX (Archimedes Institute AMX-19-IET-009), a French "Investissements d'Avenir" Programme

References

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 368–378, Vancouver, Canada. Association for Computational Linguistics.

Daniel R Anderson and Tiffany A Pempek. 2005.

Television and very young children. *American behavioral scientist*, 48(5):505–522.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Angie W Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogério Schmidt Feris, Danny Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, et al. 2019. Grounding spoken words in unlabeled video. In *CVPR Workshops*, pages 29–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Grzegorz Chrupała. 2021. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. Preprint: <https://arxiv.org/abs/2104.13225>.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

GC: suggest some existing options here.

- Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2021. *Variability and consistency in early language learning: The Wordbank project*. MIT Press.
- Peter Gorniak and Deb Roy. 2003. A visually grounded natural language interface for reference to spatial scenes. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 219–226.
- David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.
- David Harwath and James R Glass. 2017. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*.
- David Harwath and James R. Glass. 2019. Towards visually grounded sub-word speech unit discovery. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3017–3021. IEEE.
- David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665.
- David F. Harwath, Antonio Torralba, and James R. Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1858–1866.
- William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. 2019a. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 8618–8622. IEEE.
- William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. 2019b. Word recognition, competition, and activation in a model of visually grounded speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 339–348, Hong Kong, China. Association for Computational Linguistics.
- Kathy Hirsh-Pasek and Roberta Michnick Golinkoff. 1996. The intermodal preferential looking paradigm: A window onto emerging language comprehension.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Wei-Ning Hsu, David Harwath, and James Glass. 2019. Transfer Learning from Audio-Visual Grounding to Speech Recognition. In *Proc. Interspeech 2019*, pages 3242–3246.
- Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1889–1897.
- Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *CoRR*, abs/1705.06950.
- Khazar Khorrami and Okko Räsänen. 2021. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation. Preprint psyarxiv.com/37zn.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Patricia K Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in in-

- fancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101.
- Sarah Roseberry Lytle, Adrian Garcia-Sierra, and Patricia K Kuhl. 2018. Two are better than one: Infant language learning from video improves in the presence of peers. *Proceedings of the National Academy of Sciences*, 115(40):9859–9866.
- Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. 2019. Language Learning Using Speech to Image Retrieval. In *Proc. Interspeech 2019*, pages 1841–1845.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.
- Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Niloy Mukherjee and Deb Roy. 2003. A visual context-aware multimodal system for spoken language processing. In *Eighth European Conference on Speech Communication and Technology*.
- Mitja Nikolaus and Abdellah Fourtassi. 2021. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210, Online. Association for Computational Linguistics.
- Claire H Noble, Caroline F Rowland, and Julian M Pine. 2011. Comprehension of argument structure and semantic roles: Evidence from english-learning children and the forced-choice pointing paradigm. *Cognitive science*, 35(5):963–982.
- Nikos Papasarantopoulos and Shay B. Cohen. 2021. Narration generation for cartoon videos. Preprint: <https://arxiv.org/abs/2101.06803>.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Puyuan Peng and David Harwath. 2021. Fast-slow transformer for visually grounding speech.
- Michael B Robb, Rebekah A Richert, and Ellen A Wartella. 2009. Just a talking book? word learning from watching baby videos. *British Journal of Developmental Psychology*, 27(1):27–45.
- Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos.
- Brandon C Roy, Michael C Frank, Philip DeCamp, Matthew Miller, and Deb Roy. 2015. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668.
- Deb Roy. 1999. *Learning from sights and sounds: A computational model*. Ph.D. thesis, MIT Media Laboratory.
- Deb Roy. 2000a. Grounded speech communication. In *Sixth International Conference on Spoken Language Processing*.
- Deb Roy. 2000b. Learning visually grounded words and syntax of natural spoken language. *Evolution of communication*, 4(1):33–56.

- Deb K Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3-4):353–385.
- Deb K Roy and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.
- Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).
- Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2014. Learning words from images and speech. In *NIPS Workshop on Learning Semantics*.
- Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.
- Robert Vanderplank. 2010. Déjà vu? a decade of research on language laboratories, television and video in language learning. *Language teaching*, 43(1):1–37.
- Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 487–495.