# Visually Grounded Speech Models Have a Mutual Exclusivity Bias

**Leanne Nortje*** **Dan Oneaţă♦** **Yevgen Matusevych•** **Herman Kamper***
* Electrical and Electronic Engineering, Stellenbosch University, South Africa
♦ SpeeD Lab, University Politehnica of Bucharest, Romania
• CLCG, University of Groningen, The Netherlands
nortjeleanne@gmail.com dan.oneata@gmail.com
yevgen.matusevych@rug.nl kamperh@sun.ac.za

## Abstract

When children learn new words, they employ constraints such as the mutual exclusivity (ME) bias: A novel word is mapped to a novel object rather than a familiar one. This bias has been studied computationally, but only in models that use discrete word representations as input, ignoring the high variability of spoken words. We investigate the ME bias in the context of visually grounded speech models that learn from natural images and continuous speech audio. Concretely, we train a model on familiar words and test its ME bias by asking it to select between a novel and a familiar object when queried with a novel word. To simulate prior acoustic and visual knowledge, we experiment with several initialization strategies using pretrained speech and vision networks. Our findings reveal the ME bias across the different initialization approaches, with a stronger bias in models with more prior (in particular, visual) knowledge. Additional tests confirm the robustness of our results, even when different loss functions are considered. Based on detailed analyses to piece out the model's representation space, we attribute the ME bias to how familiar and novel classes are distinctly separated in the resulting space.

## 1 Introduction

When children learn new words, they employ a set of basic constraints to make the task easier. One such constraint is the *mutual exclusivity* (ME) bias: When a learner hears a novel word, they map it to an unfamiliar object (whose name they don't know yet), rather than a familiar one. This strategy was first described by Markman and Wachtel (1988) over 30 years ago and has since been studied extensively in the developmental sciences (Merriman et al., 1989; Markman et al., 2003; Mather and Plunkett, 2009; Lewis et al., 2020). With the rise of neural architectures, recent years saw renewed interest in the ME bias, this time from the computational modeling perspective: Several studies have examined whether and under which conditions the ME bias emerges in machine learning models (Gulordava et al., 2020; Gandhi and Lake, 2020; Vong and Lake, 2022; Ohmer et al., 2022).

The models in these studies normally receive input consisting of word and object representations, as the ME strategy is used to learn mappings between words and the objects they refer to. Object representations vary in their complexity, from symbolic representations of single objects (e.g., Gandhi and Lake, 2020) to continuous vectors encoding a natural image (e.g., Vong and Lake, 2022). Word representations, however, are based on their written forms in all these studies. For example, the textual form of the word *fish* has an invariable representation in the input. This is problematic because children learn words from continuous speech, and there is large variation in how the word *fish* can be realized depending on the word duration, prosody, the quality of the individual sounds and so on; see, e.g., Creel (2012) on how the ME bias affects atypical pronunciations such as [fesh] instead of [fɪsh]. As a result, children face an additional challenge compared to models trained on written words. This is why it is crucial to investigate the ME bias in a more naturalistic setting, with models trained on word representations that take into account variation between acoustic instances of the same word.

Recently, there has been considerable headway in the development of visually grounded speech models that learn from images paired with unlabeled speech (Harwath et al., 2016, 2018a; Kamper et al., 2019; Chrupała, 2022;



Transactions of the Association for Computational Linguistics, vol. 12, pp. 755–770, 2024. https://doi.org/10.1162/tacl_a_00672
Action Editor: Mauro Cettolo. Submission batch: 11/2023; Revision batch: 2/2024; Published 6/2024.

Peng and Harwath, 2022a; Peng et al., 2023; Berry et al., 2023; Shih et al., 2023). Several studies have shown, for instance, that these models learn word-like units when trained on large amounts of paired speech–vision data (Harwath and Glass, 2017; Harwath et al., 2018b; Olaleye et al., 2022; Peng and Harwath, 2022c; Nortje and Kamper, 2023; Pasad et al., 2023). Moreover, some of these models draw inspiration from the way infants acquire language from spoken words that co-occur with visual cues across different situations in their environments (Miller and Gildea, 1987; Yu and Smith, 2007; Cunillera et al., 2010; Thiessen, 2010). However, the ME bias has not been studied in these models.

In this work we test whether visually grounded speech models exhibit the ME bias. We focus on a recent model by Nortje et al. (2023), as it achieves state-of-the-art performance in a few-shot learning task that resembles the word learning setting considered here. The model's architecture is representative of many of the other recent visually grounded speech models: It takes a spoken word and an image as input, processes these independently, and then relies on a word-to-image attention mechanism to learn a mapping between a spoken word and its visual depiction. We first train the model to discriminate familiar words. We then test its ME bias by presenting it with a novel word and two objects, one familiar and one novel. To simulate prior acoustic and visual knowledge that a child might have already acquired before word learning, we additionally explore different initialization strategies for the audio and vision branches of the model.

To preview our results, we observe the ME bias across all the different initialization schemes of the visually grounded speech model, and the bias is stronger in models with more prior visual knowledge. We also carry out a series of additional tests to ensure that the observed ME bias is not merely an artefact, and present analyses to pinpoint the relationship between the model's representation space and the emergence of the ME bias. In experiments where we look at different modeling options (visual initialization and loss functions, in particular), the ME bias is observed in all cases. The code and the accompanying dataset are available from our project website.[1]

[1] https://sites.google.com/view/mutualexclusivityinvgs.

## 2 Related Work

Visually grounded speech models learn by bringing together representations of paired images and speech while pushing mismatched pairs apart. These models have been used in several downstream tasks, ranging from speech–image retrieval (Harwath et al., 2018b) and keyword spotting (Olaleye et al., 2022) to word (Peng and Harwath, 2022c) and syllable segmentation (Peng et al., 2023).

In terms of design choices, early models used a hinge loss (Harwath et al., 2016, 2018b), while several more advanced losses have been proposed since (Petridis et al., 2018; Peng and Harwath, 2022a; Peng et al., 2023). A common strategy to improve performance is to initialize the vision branch using a supervised vision model, e.g., Harwath et al. (2016) used VGG, Harwath et al. (2020) used ResNet, and recently Shih et al. (2023) and Berry et al. (2023) used CLIP. For the speech branch, self-supervised speech models like wav2vec2.0 and HuBERT have been used for initialization (Peng and Harwath, 2022c). Other extensions include using vector quantization in intermediate layers (Harwath et al., 2020) and more advanced multimodal attention mechanisms to connect the branches (Chrupała et al., 2017; Radford et al., 2021; Peng and Harwath, 2022a,b).

In this work we specifically use the few-shot model of Nortje et al. (2023) that incorporates many of these strategies (Section 5). We also look at how different design choices affect our analysis of the ME bias, e.g., using different losses (Section 7.4).

As noted already, previous computational studies of the ME bias have exclusively used the written form of words as input (Gulordava et al., 2020; Gandhi and Lake, 2020; Vong and Lake, 2022; Ohmer et al., 2022). Visually grounded speech models have the benefit that they can take real speech as input. This better resembles the actual experimental setup with human participants (Markman and Wachtel, 1988; Markman, 1989).

Concretely, since the models in Gulordava et al. (2020) and Vong and Lake (2022) are trained on written words, which are discrete by design, they need to learn a continuous embedding for each of the input classes. However, this makes dealing with novel inputs difficult: If a model never sees a particular item at training time, its embeddings are never updated and remain randomly initialized.

As a result, the ME test becomes a comparison of learned vs random embeddings instead of novel vs familiar. To address this issue, Gulordava et al. (2020) use novel examples in their contrastive loss during training, while Vong and Lake (2022) perform one gradient update on novel classes before testing. These strategies mean that, in both cases, the learner has actually seen the novel classes before testing. Such adaptations are necessary in models taking in written input. In contrast, a visually grounded speech model, even when presented with an arbitrary input sequence, can place it in the representation space learned from the familiar classes during training. We investigate whether such a representation space results in the ME bias.

## 3 Mutual Exclusivity in Visually Grounded Speech Models

Mutual exclusivity (ME) is a constraint used to learn words. It is grounded in the assumption that an object, once named, cannot have another name. The typical setup of a ME experiment (Markman and Wachtel, 1988) involves two steps and is illustrated in Figure 1. First, the experimenter will ensure that the learner (usually a child) is familiar with a set of specific objects by assessing their ability to correctly identify objects associated with familiar words. In this example, the familiar classes are 'clock', 'elephant', and 'horse', as illustrated in the top panel of the figure. Subsequently, at test time the learner is shown a familiar image (e.g., ELEPHANT) and a novel image (e.g., GUITAR) and is asked to determine which of the two corresponds to a novel spoken word, e.g., *guitar* (middle panel in the figure). If the learner exhibits a ME bias, they would select the corresponding novel object, GUITAR in this case (bottom panel).

Our primary objective is to investigate the ME bias in computational models that operate on the audio and visual modalities. These models, known as visually grounded speech models, draw inspiration from how children learn words (Miller and Gildea, 1987), by being trained on unlabeled spoken utterances paired with corresponding images. The models learn to associate spoken words and visual concepts, and often do so by predicting a similarity score for a given audio utterance and an input image. This score can then be used to select between competing visual objects given a spoken utterance, as required in the ME test.

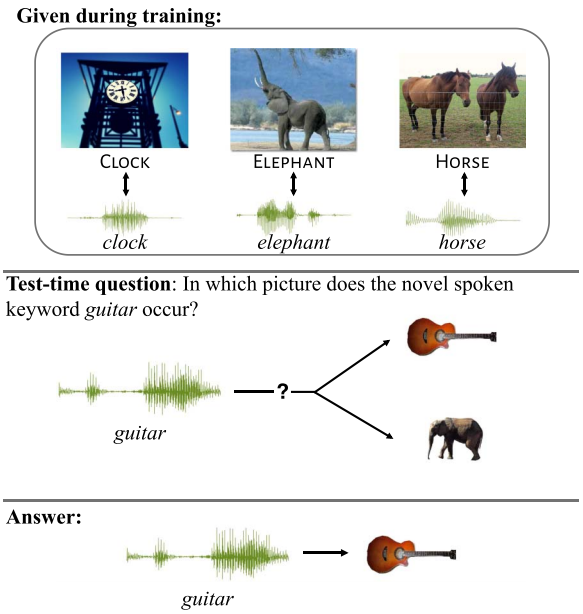



Figure 1: *Top:* A learner is familiarized with a set of objects during training. *Middle:* At test time, two images are given, one from a familiar class seen during training and the other from an unseen novel class. *Bottom:* If a learner has a ME bias, then when prompted with a novel spoken query, the novel object (GUITAR) would be selected.

In Section 4 below we describe how we set up our test of the ME bias. In Section 5 we then present the visually grounded speech model that we use in this study.

## 4 Constructing a Speech–Image Test for Mutual Exclusivity

To construct our ME test, we need isolated spoken words that are paired with natural images of objects. We also need to separate these paired word–image instances into two sets: familiar classes and novel classes. A large multimodal dataset of this type does not exist, so we create one by combining several image and speech datasets.

For the images, we combine MS COCO (Lin et al., 2014) and Caltech-101 (Fei-Fei et al., 2006). MS COCO contains 328k images of 91 objects in their natural environment. Caltech-101 contains 9k Google images spanning 101 classes. Ground truth object segmentations are available for both these datasets. During training, we use entire images, but during evaluation, we use segmented objects. This resembles a naturalistic learning scenario in which a learner is familiarized with objects

| | |
|---|---|
| Familiar | bear, bird, boat, car, cat, clock, cow, dog, elephant, horse, scissors, sheep, umbrella |
| Novel | ball, barrel, bench, buck, bus, butterfly, cake, camera, canon, chair, cup, fan, fork, guitar, lamp, nautilus, piano, revolver, toilet, trumpet |

Table 1: The familiar and novel classes in our ME test setup.

by seeing them in a natural context, but is presented with individual objects (or their pictures) in isolation at test time.

For the audio, we combine the FAAC (Harwath and Glass, 2015), Buckeye (Pitt et al., 2005), and LibriSpeech (Panayotov et al., 2015) datasets. These English corpora respectively span 183, 40, and 2.5k speakers.

To select familiar and novel classes, we do a manual inspection to make sure that object segmentations for particular classes are of a reasonably high quality and that there are enough spoken instances for each class in the segmented speech data (at least 100 spoken examples per class). As an example of an excluded class, we did not use CURTAIN, since it was often difficult to reliably see that curtains are depicted after these are segmented out. The final result is a setup with 13 familiar classes and 20 novel classes, as listed in Table 1.

During training (Figure 1, top panel), a model only sees familiar classes. We divide our data so that we have a training set with 18,279 unique spoken word segments and 94,316 unique unsegmented natural images spanning the 13 familiar classes. These are then paired up for training as explained in Section 5.1. During training we also use a development set for early stopping; this small set consists of 130 word segments and 130 images from familiar classes.

For ME testing (Figure 1, middle panel) we require a combination of familiar and novel classes. Our test set in total consists of 8,575 spoken word segments with 22,062 segmented object images. To implement the ME test, we sample 1k episodes: Each episode consists of a novel spoken word (query) with two sampled images, one matching the novel class from the query and the other containing a familiar object. We ensure that the two images always come from the same image dataset to avoid any intrinsic dataset biases. There

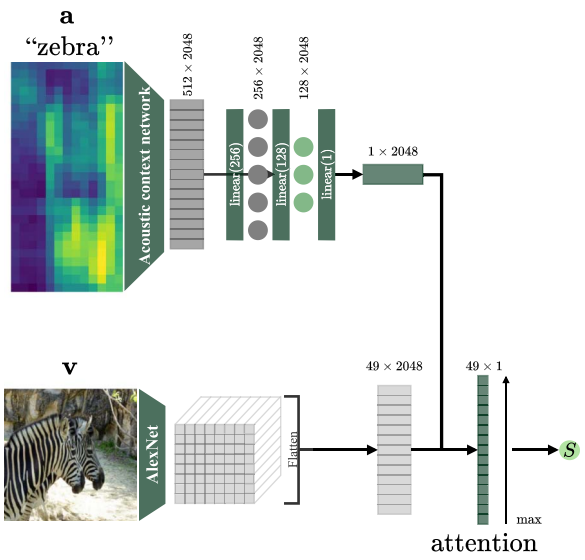



Figure 2: MATTNET (Nortje et al., 2023) consists of a vision network and an audio network. These are connected through a word-to-image attention mechanism. The model outputs a score $S$ indicating the similarity of the speech and image inputs.

is no overlap between training, development, and test samples.

## 5 A Visually Grounded Speech Model

We want to establish whether visually grounded speech models exhibit the ME bias. While there is a growing number of speech–image models (Section 2), many of them share the same general methodology. We therefore use a visually grounded speech model that is representative of the models in this research area: the Multimodal ATTention NETwork (MATTNET) of Nortje et al. (2023). This model achieves top performance in a few-shot word–object learning task that resembles the way infants learn words from limited exposure. Most useful for us is that the model is conceptually simple: It takes an image and a spoken word and outputs a score indicating how similar the inputs are, precisely what is required for ME testing.

### 5.1 Model

MATTNET consists of a vision and an audio branch that are connected with a word-to-image attention mechanism, as illustrated in Figure 2.

A spoken word **a** is first parameterized as a mel-spectrogram with a hop length of 10 ms, a window of 25 ms, and 40 bins. The audio branch

takes this input, passes it through an acoustic network consisting of LSTM and BiLSTM layers, and finally outputs a single word embedding by pooling the sequence of representations along the time dimension with a two-layer feedforward network. This method of encoding a variable-length speech segment into a single embedding is similar to the idea behind acoustic word embeddings (Chung et al., 2016; Holzenberger et al., 2018; Wang et al., 2018; Kamper, 2019).

The vision branch is an adaptation of AlexNet (Krizhevsky et al., 2017). An image $\mathbf{v}$ is first resized to 224×224 pixels and normalized with means and variances calculated on ImageNet (Deng et al., 2009). The vision branch then encodes the input image into a sequence of pixel embeddings.

The audio and vision branches are connected through a multimodal attention mechanism that takes the dot product between the acoustic word embedding and each pixel embedding. The maximum of these attention scores is taken as the final output of the model, the similarity score $S$. The idea behind this attention mechanism is to focus on the regions within the image that are most indicative of the spoken word.

The similarity score $S(\mathbf{a}, \mathbf{v})$ should be high if the spoken word $\mathbf{a}$ and the image $\mathbf{v}$ are instances of the same class, and low otherwise. This is accomplished by using a contrastive loss that pushes positive word–image pairs from the same class closer together than mismatched negative word–image pairs (Nortje et al., 2023):

$$\begin{aligned}\ell = {} & d\left(S(\mathbf{a}, \mathbf{v}), 100\right) \\ & + \sum_{i=1}^{N_{\text{neg}}} d\left(S(\mathbf{a}_i^-, \mathbf{v}), 0\right) + \sum_{i=1}^{N_{\text{neg}}} d\left(S(\mathbf{a}, \mathbf{v}_i^-), 0\right) \\ & + \sum_{i=1}^{N_{\text{pos}}} d\left(S(\mathbf{a}, \mathbf{v}_i^+), 100\right) + \sum_{i=1}^{N_{\text{pos}}} d\left(S(\mathbf{a}_i^+, \mathbf{v}), 100\right)\end{aligned} \tag{1}$$

where $d$ is the squared Euclidean distance, i.e., $S$ is pushed to 0 for negative pairs and to 100 for positive pairs. In more detail, for an anchor positive word–image pair $(\mathbf{a}, \mathbf{v})$, we sample positive examples $(\mathbf{a}^+_{1:N_{\text{pos}}}, \mathbf{v}^+_{1:N_{\text{pos}}})$ that match the class of the anchor and negative examples $(\mathbf{a}^-_{1:N_{\text{neg}}}, \mathbf{v}^-_{1:N_{\text{neg}}})$ that are not instances of the anchor class. We use $N_{\text{pos}} = 5$ and $N_{\text{neg}} = 11$ in our implementation.

As a reminder from Section 4, the model is trained exclusively on familiar classes and never sees any novel classes during training. Novel classes are also never used as negative examples. We train the model with Adam (Kingma and Ba, 2015) for 100 epochs and use early stopping with a validation task. The validation task involves presenting the model with a familiar word query and asking it to identify which of the two familiar object images it refers to. We use the spoken words and isolated object images from the development set for this task (see Section 4).

### 5.2 Different Initialization Strategies as a Proxy for Prior Knowledge

The ME bias has been observed in children at the age of around 17 months (e.g., Halberda, 2003). At this age, children have already gained valuable experience from both spoken language used in their surroundings and the visual environment that they navigate (Clark, 2004). For example, 4.5-month-olds can recognize objects (Needham, 2001), and 6.5-month-olds can recognize some spoken word forms (Jusczyk and Aslin, 1995). These abilities can be useful when learning new words. In light of this, we adopt an approach that initializes the vision and audio branches of our model to emulate prior knowledge.

For the vision branch, we use the convolutional encoder of the self-supervised AlexNet (Koohpayegani et al., 2020), which distills the SimCLR ResNet50×4 model (Chen et al., 2020) into AlexNet and trains it on ImageNet (Deng et al., 2009). For the audio branch, we use an acoustic network (van Niekerk et al., 2020) pretrained on the LibriSpeech (Panayotov et al., 2015) and Places (Harwath et al., 2018a) datasets using a self-supervised contrastive predictive coding (CPC) objective (Oord et al., 2019). Both these initialization networks are trained without supervision directly on unlabeled speech or vision data, again emulating the type of data an infant would be exposed to. When these initialization strategies are not in use, we initialize the respective branches randomly.

Considering these strategies, we end up with four possible MattNet variations: one where both the vision and audio branches are initialized from pretrained networks, one where only the audio branch is initialized from a CPC model, one where only the vision branch is initialized from AlexNet, and one where neither branch is initialized with pretrained models (i.e., a full random initialization).

| | | Model initialization | | Accuracy (%) | |
|---|---|---|---|---|---|
| | | Audio (CPC) | Vision (AlexNet) | Familiar–familiar | Familiar–novel |
| 1 | Random baseline | N/A | N/A | 50.19 | 49.92 |
| 2 | MattNet | ✗ | ✗ | 72.86 | 57.29 |
| 3 | | ✗ | ✓ | 85.89 | 59.32 |
| 4 | | ✓ | ✗ | 75.78 | 55.92 |
| 5 | | ✓ | ✓ | 83.20 | 60.27 |

Table 2: Performance for different initialization strategies of MattNet. The ME results are given in the familiar–novel column. As a reference, discrimination performance between familiar classes is given under familiar–familiar.

In the following sections, we present our results. We compare them to the performance of a naive baseline that chooses one of the two images at random for a given word query. To determine whether the differences between our model variations and a random baseline are statistically significant, we fit mixed-effects regression models to MattNet's scores using the lme4 package (Bates et al., 2015). Details are given in Appendix A.

## 6 Mutual Exclusivity Results

Our main question is whether visually grounded models like MattNet (Section 5) exhibit the ME bias. To test this, we present the trained model with two images: one showing familiar and one showing a novel object. The model is then prompted to identify which image a novel spoken word refers to (Section 3). We denote this ME test as the familiar–novel test. With this, we also introduce our notation for specific tests: `<image one type>-<image two type>`, with the type of the audio query underlined. The class of the audio query will match the one of the underlined image, unless explicitly stated. Table 6 in Appendix B contains a cheat sheet to understand the tests' notation.

Before we look at our target familiar–novel ME test, it is essential to ensure that our model has successfully learned to distinguish the familiar classes encountered during training; testing for the ME bias would be premature if the model does not know the familiar classes. We therefore perform a familiar–familiar test, where the task is to match a word query from a familiar class to one of two images containing familiar classes.

Table 2 presents the results of these two tests for the different MattNet variations described in Section 5.2. The results of the familiar–familiar test show that all the model variations can distinguish between familiar classes. The vision (AlexNet) initialization of the vision branch contributes more than the audio (CPC) initialization: The two best familiar–familiar models both use vision initialization. Our statistical tests confirm the reported patterns: All model variations are significantly better than the random baseline, and adding the visual (AlexNet) and/or audio (CPC) initialization to the basic model significantly improves MattNet's accuracy on the familiar–familiar test.

We now turn to the ME test. The results are given in the familiar–novel column of Table 2. All MattNet variations exhibit the ME bias, with above-chance accuracy in matching a novel audio segment to a novel image, as also confirmed by our statistical significance test (Appendix A). From the table, the strongest ME bias is found in the MattNet variation that initializes both the audio (CPC) and vision (AlexNet) branches (row 5), followed by the variation with the vision initialization alone (row 3). Surprisingly, using CPC initialization alone reduces the strength of the ME bias (row 2 vs row 4). Again, these results are confirmed by our statistical tests. To summarize: Even the basic MattNet has the ME bias, but the AlexNet initialization makes it noticeably stronger.

To investigate whether the reported accuracies are stable over the course of learning, we consider MattNet's performance over training epochs on the two tests: familiar–familiar and familiar–novel. We use the model variation with the strongest ME bias, i.e., with both the audio

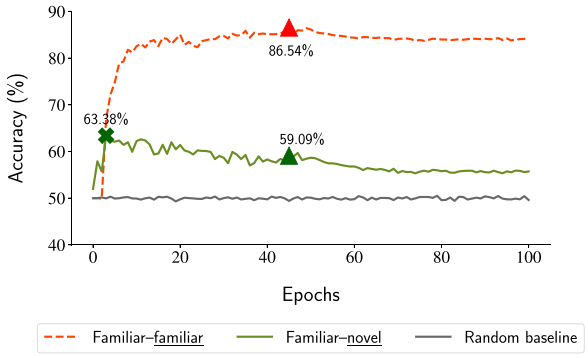



Figure 3: MATTNET's performance over training epochs. The cross indicates the highest overall ME familiar–novel score. The triangles show the scores at the point where the best familiar–familiar score occurs. Results are for the variant of MATTNET with both CPC and AlexNet initializations, and performance is averaged over five training runs.

and vision branches initialized. Figure 3 shows that the ME bias (familiar–novel, green solid line) is stronger early on in training and then decreases later on. The pattern is similar for the familiar–familiar score (red dashed line), but the highest score in this case is achieved later in training than the best familiar–novel score. The scores stabilize after approximately 60 epochs; at this epoch, the model's accuracy is 84.38% on the familiar–familiar task and 56.78% on the familiar–novel task (numbers not shown in the figure). This suggests that the results reported above for both tests are robust and do not only hold for a particular point in training.

In summary, we found that a visually grounded speech model, MATTNET, learns the familiar set of classes and thereafter exhibits a consistent and robust ME bias. This bias gets stronger when the model is initialized with prior visual knowledge, although the results for the audio initialization are inconclusive. Whereas the strength of the ME bias slightly changes as the model learns, it is consistently above chance, suggesting that this is a stable effect in our model.

## 7 Further Analyses

We have shown that our visually grounded speech model has the ME bias. However, we need to make sure that the observed effect is really due to the ME bias and is not a fluke. In particular, because our model is trained on natural images, additional objects might appear in the background, and there is a small chance that some of these objects are from the novel classes. As a result, the model may learn something about the novel classes due to information leaking from the training data. Here we present several sanity-check experiments to show that we observe a small leakage for one model variant, but it does not account for the strong and consistent ME bias reported in the previous section. Furthermore, we provide additional analyses that show how the model structures its audio and visual representation spaces for the ME bias to emerge.

### 7.1 Sanity Checks

The familiar–novel column in Table 3 repeats the ME results from Section 6. We now evaluate these ME results against three sanity-check experiments.

We start by testing the following: If indeed the model has a ME bias, it should not make a distinction between two novel classes. So we present MATTNET with two novel images and a novel audio query in a novel–novel test. Here, one novel image depicts the class referred to by the query, and the other image depicts a different novel class. If the model does not know the mappings between novel words and novel images, it should randomly choose between the two novel images. The results for this novel–novel test in Table 3 are close to 50% for all MATTNET's variations, as expected.

Surprisingly, our statistical test shows significant differences between the baseline and two out of the four variations: MATTNET with full random initialization scores higher than the baseline on this novel–novel task, and MATTNET with the vision initialization lower. Since the differences between each model and the baseline are small and in different directions (one model scores lower and the other higher), we believe these patterns are not meaningful. At the same time, one possible explanation of the above-chance performance of MATTNET with random initialization is that there may be some leakage of information about the novel classes that may appear in the background of the training images. To test whether our ME results can be explained away by this minor leakage, we observe that the model's scores in the familiar–novel task (the ME task) are noticeably higher than the scores in the novel–novel task. An additional statistical test (Appendix A) shows that the differences between MATTNET's scores across

| | | Model initialization | | Accuracy (%) | | | |
|---|---|---|---|---|---|---|---|
| | | Audio | Vision | Familiar–novel | Novel–novel | Familiar–novel* | Familiar–novel |
| 1 | Random baseline | N/A | N/A | 49.92 | 49.85 | 49.72 | 50.58 |
| 2 | MattNet | ✗ | ✗ | 57.29 | 51.05 | 55.52 | 69.68 |
| 3 | | ✗ | ✓ | 59.32 | 48.74 | 58.51 | 86.92 |
| 4 | | ✓ | ✗ | 55.92 | 50.52 | 53.41 | 70.93 |
| 5 | | ✓ | ✓ | 60.27 | 49.92 | 58.41 | 82.88 |

Table 3: To ensure that the ME bias is real and not because of a peculiarity of our setup, we compare the ME test (familiar–novel) to three sanity check experiments for the different variants of MattNet.

the two tasks are, indeed, statistically significant for three out of the four variations (except the one with the audio initialization alone). This suggests that the ME bias cannot be explained away by information leakage for most model variations.

To further stress test that the model does not reliably distinguish between novel classes, we perform an additional test: familiar–novel*. In the standard familiar–novel ME test, the model is presented with a familiar class (e.g., ELEPHANT) and a novel class (GUITAR) and correctly matches the novel query word *guitar* to the novel class. If the model truly uses a ME bias (and not a mapping between novel classes and novel words that it could potentially infer from the training data), then it should still select the novel image (GUITAR) even when prompted with a mismatched novel word, say *ball*. Therefore, we construct a test to see whether a novel audio query would still be matched to a novel image even if the novel word does not refer to the class in the novel image. Results for this familiar–novel* test in Table 3—where the asterisk indicates a mismatch in classes—show that the numbers are very close to those in the standard familiar–novel ME test. All the MattNet variations therefore exhibit a ME bias: A novel word query belongs to any novel object, even if the two are mismatched, since the familiar object already has a name. Our statistical tests support this result.

Finally, in all the results presented above, MattNet has a preference for a novel image. One simple explanation that would be consistent with all these results (but would render them trivial) is if the model always chose a novel object when encountering one (regardless of the input query). To test this, we again present the model with a familiar and a novel object, but now query it with a familiar word. The results for this familiar–novel test in Table 3 show that all MattNet variations achieve high scores in selecting the familiar object. Again, our significance test confirms that all the scores are significantly higher than random.

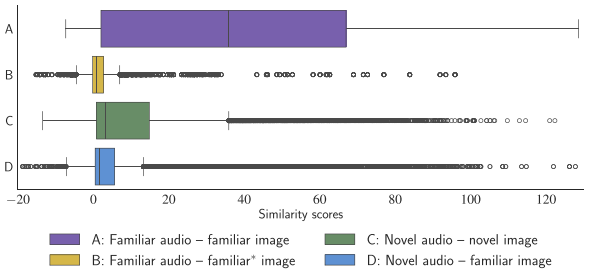



Figure 4: A box plot of similarities for four types of audio–image comparisons with MattNet. The audio–image examples of a familiar class have higher similarity (A) than mismatched familiar instances (B). Novel class instances are in-between (C), but they aren't placed as close as the learned familiar classes (A). Novel instances (C) are still closer to each other than to familiar ones (D).

### 7.2 Why Do We See a ME Bias?

We have now established that the MattNet visually grounded speech model exhibits the ME bias. But this raises the question: Why does the model select the novel object rather than the familiar one? How is the representation space organised for this to happen? We attempt to answer these questions by analyzing different cross-modal audio–image comparisons made in both the familiar–familiar and familiar–novel (ME) tests. Results are given in Figure 4, where we use MattNet with both visual and audio encoders initialized (row 5, Table 3).

First, in the familiar–familiar setting we compare two similarities: (A) the MattNet similarity scores between a familiar audio query and a familiar image from the same class against (B) the similarity between a familiar audio query and a

familiar image from a different class (indicated with familiar*). Perhaps unsurprisingly given the strong familiar–familiar performance in Table 2, we observe that the similarities of matched pairs (familiar audio – familiar image, A) are substantially higher than the similarities of mismatched pairs (familiar audio – familiar* image, B). This organization of the model's representation space can be explained by the contrastive objective in Equation 1, which ensures that the words and images from the same familiar class are grouped together, and different classes are pushed away from one another.

But where do the novel classes fit in? To answer this question, we consider two types of comparisons from the familiar–novel ME setting: (C) the MattNet similarity scores between a novel query and a novel image (from any novel class) against (D) the similarity between a novel query and a familiar image. We observe that the novel audio – novel image similarities (C) are typically higher than the novel audio – familiar image similarities (D). That is, novel words in the model's representation space are closer to novel images than to familiar images. As a result, a novel query on average is closer to *any* novel image than to familiar images, which sheds light on why we observe the ME bias.

The similarities involving novel words (C and D) are normally higher than those of the mismatched familiar classes (B). This suggests that novel samples are closer to familiar samples than familiar samples from different classes are to one another. In other words, during training, the model learns to separate out familiar classes (seen during training), but then places the novel classes (not seen during training) relatively close to at least some of the familiar ones. Crucially, samples in the novel regions are still closer to each other (C) than they are to any of the familiar classes (as indicated by D).

How does the contrastive loss in Equation 1 affect the representations of novel classes during training, given that the model never sees any of these novel instances? In Figure 5 we plot the same similarities as we did in Figure 4 but instead we use the model weights before training. It is clear how training raises the similarities of matched familiar inputs (A) while keeping the similarities of mismatched familiar inputs low (D), which is exactly what the loss is designed to

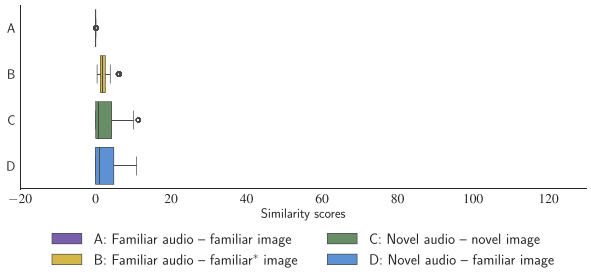


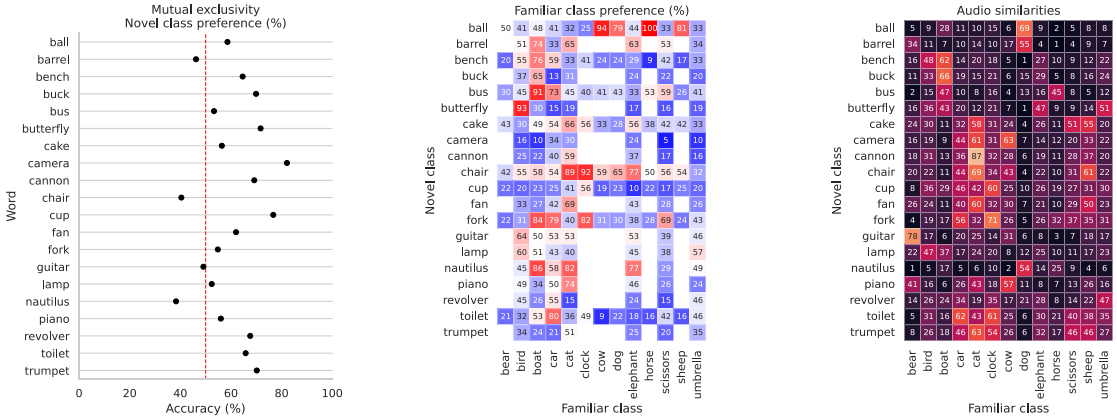

Figure 5: The same analysis as in Figure 4, but for MattNet before training. We can see how similarities are affected through training.

do. But how are novel instances affected? One a priori hypothesis might be that training has only a limited effect on the representations from novel classes. But, by comparing Figures 4 and 5, we see that this is not the case: Similarities involving novel classes change substantially during training (C and D). The model thus uses information from the familiar classes that it is exposed to, to update the representation space, affecting both seen and unseen classes.

### 7.3 Finer-grained Analysis

We have seen a robust ME bias in the aggregated results above. But what do results look like at a finer level? We now consider each of the 20 novel words individually and compute how often the model selects the corresponding novel image (Figure 6a) or any of the familiar images (Figure 6b). While most of the novel words are associated with the ME bias (Figure 6a, dots to the right of the vertical red line), a small number of words yields a strong anti-ME bias when paired with certain familiar words (Figure 6b, red cells). For example, for the novel word *bus*, in 91% of the test cases the model picks an image of a familiar class BOAT rather than an image of the novel class BUS. It is worth emphasizing that the ME bias isn't absolute: Even in human participants it isn't seen in 100% of test cases. Nevertheless, it is worth investigating why there is an anti-ME bias for some particular words (something that is easier to do in a computational study compared to human experiments).

One reason for an anti-ME result is the phonetic similarity of a novel word to familiar words. For example, *bus* and *boat* start with the same consonant followed by a vowel. If we look at Figure 6c, which shows the cosine similarities between the learned audio embeddings from MattNet, we




(a) Mutual exclusivity bias per word. Higher accuracy (dots to the right) indicates a stronger ME bias.

(b) Percentage of times a familiar image is selected for a novel audio. Some entries are empty because these were never compared in any of the sampled episodes.

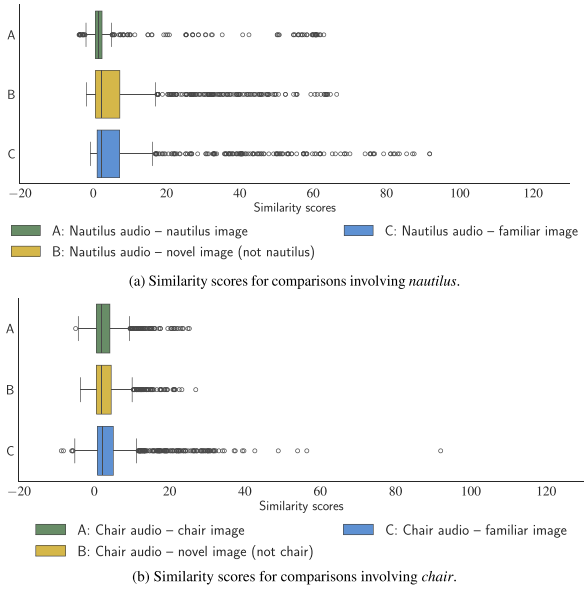
(c) Similarities of audio embeddings between novel and familiar words. The numbers are cosine similarity times 100. Lighter shades are associated with higher similarity.

Figure 6: A finer-grained analysis looking at the ME bias individually for each of the 20 novel words.

see that spoken instances of *bus* and *boat* indeed have high similarity. In fact, several word pairs starting with the same consonant (followed by a vowel) have high learned audio similarities, e.g., *buck–boat*, *bench–boat* and *cake–cat*, all translating to an anti-ME bias in Figure 6b.

However, the anti-ME bias cannot be explained by acoustic similarity alone: Some anti-ME pairs have low audio similarities, e.g., *nautilus–elephant*. For such cases, the representation space must be structured differently from the aggregated analysis in Section 7.2 (otherwise we would see a ME bias for these pairs). Either the spoken or the visual representation of a particular class can be responsible (or both). To illustrate this, we zoom in on the two novel words showing the strongest anti-ME results in Figure 6a: *nautilus* and *chair*.

Figure 7a presents a similar analysis to that of Figure 4 but specifically for *nautilus*. We see the anti-ME bias: *Nautilus* audio is more similar to familiar images (C) than to NAUTILUS images (A). This is the reverse of the trend in Figure 4 (C vs D). Is this due to the *nautilus* word queries or the NAUTILUS images? Here in Figure 7a, box B shows what happens when we substitute the NAUTILUS images from box A with any other novel image: The similarity goes up. This means that NAUTILUS images are not placed in the same area of the representation space as the other novel images. But this isn't all: Boxes B and C are also close




(a) Similarity scores for comparisons involving *nautilus*.

(b) Similarity scores for comparisons involving *chair*.

Figure 7: Box plots of similarities between combinations of novel and familiar class instances focused on two classes: (a) *nautilus* and (b) *chair*.

to each other. Concretely, if we compare B vs C here in Figure 7a to C vs D in Figure 4, then we still do not see the difference corresponding to the ME result, as in the latter case. This means that the NAUTILUS audio is also partially responsible for the anti-ME result here in that it is placed close to familiar images.

Let us do a similar analysis for *chair*: Figure 7b. We again see the anti-ME results by comparing A and C. But now swapping out CHAIR images

| | Accuracy (%) | |
|---|---|---|
| Loss | Familiar–familiar | Familiar–novel |
| MattNet (1) | 83.20 | 60.27 |
| Hinge (2) | 87.21 | 57.85 |
| InfoNCE (3) | 93.16 | 63.91 |

Table 4: The effect of different losses on the ME test (familiar–novel) and the sanity check (familiar–familiar).

| | Accuracy (%) | |
|---|---|---|
| Vision initialization | Familiar–familiar | Familiar–novel |
| Self-supervised | 83.20 | 60.27 |
| Supervised | 87.08 | 61.66 |

Table 5: The effect of using a self-supervised or supervised version of AlexNet for visual initialization. Scores for the ME test (familiar–novel) and the sanity check (familiar–familiar) are reported.

for other novel images (B) does not change the similarities. In this case, the culprit is therefore mainly the *chair* audio.

Further similar analyses can be done to look at other anomalous cases. But it is worth noting, again, that the aggregated ME scores from Section 6 are typically between 55% and 61% (not 100%). So we should expect some anti-ME trends in some cases, and the analysis in this section shows how we can shed light on those.

### 7.4 How Specific Are Our Findings to MattNet?

We have considered one visually grounded speech model, namely, MattNet. How specific are our findings to this particular model? While several parts of our model can be changed to see what impact they have, we limit our investigation to two potentially important components: the loss function and the visual network initialization.

**Loss Function.** Apart from the loss in Equation 1, we now look at two other contrastive losses. The hinge loss is popular in many visually grounded speech models (Harwath et al., 2016; Chrupała et al., 2017). It uses a piece-wise linear function to ensure a greater similarity for matched pairs:

$$\ell = \sum_{i=1}^{N_{\text{neg}}} \max\left(0, S(\mathbf{a}_i^-, \mathbf{v}) - S(\mathbf{a}, \mathbf{v}) + m\right) + \sum_{i=1}^{N_{\text{neg}}} \max\left(0, S(\mathbf{a}, \mathbf{v}_i^-) - S(\mathbf{a}, \mathbf{v}) + m\right) \quad (2)$$

where $m = 1$ is a margin parameter. We sample negatives within a batch, similar to Harwath et al. (2016).

InfoNCE is a loss typically employed by self-supervised models (Oord et al., 2019) and vision–text models (Jia et al., 2021; Li et al., 2021; Radford et al., 2021). It uses the logistic function to select a positive pair from among a set of negatives:

$$\ell = \log \frac{\exp S(\mathbf{a}, \mathbf{v})}{\exp S(\mathbf{a}, \mathbf{v}) + \sum_{i=1}^{N_{\text{neg}}} \exp S(\mathbf{a}, \mathbf{v}_i^-)} + \log \frac{\exp S(\mathbf{a}, \mathbf{v})}{\exp S(\mathbf{a}, \mathbf{v}) + \sum_{i=1}^{N_{\text{neg}}} \exp S(\mathbf{a}_i^-, \mathbf{v})} \quad (3)$$

Apart from changing the loss, the rest of the MattNet structure is retained. Results are shown in Table 4 for models that use self-supervised CPC and AlexNet initializations. The two new losses can learn the familiar classes and exhibit a ME bias. In fact, an even better familiar–familiar performance and a stronger ME bias (familiar–novel) are obtained with the InfoNCE loss. This loss should therefore be considered in future work studying the ME bias in visually grounded speech models.

**Visual Network Initialization.** In Section 6 we saw that vision initialization contributes most to the ME strength. Here we investigate whether we can get an even greater performance boost if we initialize MattNet using a supervised version of AlexNet instead of the self-supervised variant used thus far. Both the self-supervised (Koohpayegani et al., 2020) and supervised (Krizhevsky et al., 2017) versions of AlexNet are trained on ImageNet (Deng et al., 2009), so we can fairly compare MattNet when initialized with either option. Both MattNet variants shown in Table 5 make use of CPC initialization. We observe that the supervised AlexNet initialization performs better on the familiar–familiar task than the self-supervised initialization. However, the ME (familiar–novel) results with the supervised AlexNet initialization are only slightly higher than with the self-supervised initialization.

While there is a broad space of visually grounded models that could be used to consider

the ME task, it is encouraging that all the variants in this work show the bias.

## 8 Conclusion and Future Work

Mutual exclusivity (ME) is a constraint employed by children learning new words: A novel word is assumed to belong to an unfamiliar object rather than a familiar one. In this study, we have demonstrated that a representative visually grounded speech model exhibits a consistent and robust ME bias, similar to the one observed in children. We achieved this by training the model on a set of spoken words and images and then asking it to match a novel acoustic word query to an image depicting either a familiar or a novel object. We considered different initialization approaches simulating prior language and visual processing abilities of a learner. The ME bias was observed in all cases, with the strongest bias occurring when more prior knowledge was used in the model (initializing the vision branch had a particularly strong effect).

In further analyses we showed that the ME bias is strongest earlier on in model training and then stabilises over time. In a series of additional sanity-check tests we showed that the ME bias was not an artefact: It could not be explained away by possible information leakage from the training data or by trivial model behaviors. We found that the resulting embedding space is organized such that novel classes are mapped to a region distinct from the one containing familiar classes, and that different familiar classes are spread out over the space to maximize the distance between familiar classes. As a result, novel words are mapped on to novel images, leading to a ME bias. Lastly, we showed that the ME bias is robust to model design choices in experiments where we changed the loss function and used a supervised instead of self-supervised visual initialization approach.

Future work can consider whether using a larger number of novel and familiar classes affects the results. Another interesting avenue for future studies resolves around multilingualism. Following on from the original ME studies with young children, Byers-Heinlein and Werker (2009) and Kalashnikova et al. (2015), among others, have looked at how multilingualism affects the use of the ME constraint. This setting is interesting since in the multilingual case different words from the distinct languages are used to name the same object. These studies showed that in bi- and trilingual children from the same age group, the ME bias is not as strong as in monolingual children. We plan to investigate this computationally in future work.


## Acknowledgments

This work was supported through a Google DeepMind scholarship for LN and a research grant from Fab Inc. for HK. DO was partly supported by the European Union's HORIZON-CL4-2021-HUMAN-01 research and innovation programme under grant agreement no. 101070190 AI4Trust. We would like to thank Benjamin van Niekerk for useful discussions about the analysis. We would also like to thank the anonymous reviewers and action editor for their valuable feedback.


## References


Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01

Layne Berry, Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Hung-yi Lee, and David Harwath. 2023. M-SpeechCLIP: Leveraging large-scale, pre-trained models for multilingual speech to image retrieval. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP49357.2023.10096882

Krista Byers-Heinlein and Janet F. Werker. 2009. Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, 12(5):815–823. https://doi.org/10.1111/j.1467-7687.2009.00902.x, PubMed: 19702772

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

Grzegorz Chrupała. 2022. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.1.12967

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*. https://doi.org/10.18653/v1/P17-1057

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, and Hung-yi Lee. 2016. Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks. In *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2016-82

Eve Clark. 2004. How language acquisition builds on cognitive development. *Trends in Cognitive Sciences*, 8(10):472–478. https://doi.org/10.1016/j.tics.2004.08.012, PubMed: 15450512

Sarah Creel. 2012. Phonological similarity and mutual exclusivity: On-line recognition of atypical pronunciations in 3–5-year-olds. *Developmental Science*, 15(5):697–713. https://doi.org/10.1111/j.1467-7687.2012.01173.x, PubMed: 22925517

Toni Cunillera, Estela Càmara, Matti Laine, and Antoni Rodríguez-Fornells. 2010. Speech segmentation is facilitated by visual cues. *Quarterly Journal of Experimental Psychology*, 63(2):260–274. https://doi.org/10.1080/17470210902888809, PubMed: 19526435

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*. https://doi.org/10.1109/CVPR.2009.5206848

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611. https://doi.org/10.1109/TPAMI.2006.79, PubMed: 16566508

Kanishk Gandhi and Brenden Lake. 2020. Mutual exclusivity as a challenge for deep neural networks. In *Proceedings of NeurIPS*.

Kristina Gulordava, Thomas Brochhagen, and Gemma Boleda. 2020. Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks. In *Proceedings of Cognitive Science Society*.

Justin Halberda. 2003. The development of a word-learning strategy. *Cognition*, 87(1):B23–34. https://doi.org/10.1016/S0010-0277(02)00186-5, PubMed: 12499109

David Harwath, Galen Chuang, and James Glass. 2018a. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2018.8462396

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *Proceedings of Automatic Speech Recognition and Understanding*. https://doi.org/10.1109/ASRU.2015.7404800

David Harwath and James Glass. 2017. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/P17-1047

David Harwath, Wei-Ning Hsu, and James Glass. 2020. Learning hierarchical discrete linguistic units from visually-grounded speech. In *Proceedings of International Conference on Learning Representations*.

David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. 2018b. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of ECCV*. https://doi.org/10.1007/978-3-030-01231-1_40

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Proceedings of NeurIPS*.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. 2018. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. In *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2018-2364

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language

representation learning with noisy text supervision. In *Proceedings of ICML*. PMLR.

Peter Jusczyk and Richard Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1):1–23. https://doi.org/10.1006/cogp.1995.1010, PubMed: 7641524

Marina Kalashnikova, Karen Mattock, and Padraic Monaghan. 2015. The effects of linguistic experience on the flexible use of mutual exclusivity in word learning. *Bilingualism: Language and Cognition*. https://doi.org/10.1017/S1366728914000364

Herman Kamper. 2019. Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2019.8683639

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. 2019. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. https://doi.org/10.1109/TASLP.2018.2872106

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. 2020. CompRess: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *ACM*. https://doi.org/10.1145/3065386

Molly Lewis, Veronica Cristiano, Brenden M. Lake, Tammy Kwan, and Michael C. Frank. 2020. The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, 198:104191. https://doi.org/10.1016/j.cognition.2020.104191

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of NeurIPS*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV*.

Ellen Markman. 1989. *Categorization and Naming in Children: Problems of Induction*. MIT Press.

Ellen Markman and Gwyn Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Computing*, 20(2):121–157. https://doi.org/10.1016/0010-0285(88)90017-5, PubMed: 3365937

Ellen Markman, Judith Wasow, and Mikkel Hansen. 2003. Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, 47(3):241–275. https://doi.org/10.1016/S0010-0285(03)00034-3, PubMed: 14559217

Emily Mather and Kim Plunkett. 2009. Learning words over time: The role of stimulus repetition in mutual exclusivity. *Infancy*, 14(1):60–76. https://doi.org/10.1080/15250000802569702, PubMed: 32693468

William Merriman, Laura Bowman, and Brian MacWhinney. 1989. The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, 54(3–4):1–132. https://doi.org/10.2307/1166130

George Miller and Patricia Gildea. 1987. How children learn words. *Scientific American*, 257(3):94–99. https://doi.org/10.1038/scientificamerican0987-94, PubMed: 3659892

Amy Needham. 2001. Object recognition and object segregation in 4.5-month-old infants. *Journal of Experimental Child Psychology*, 78(1):3–22. https://doi.org/10.1006/jecp.2000.2598, PubMed: 11161419

Leanne Nortje and Herman Kamper. 2023. Towards visually prompted keyword localisation for zero-resource spoken languages. In *Proceedings of SLT*. https://doi.org/10.1109/SLT54892.2023.10023079

Leanne Nortje, Dan Oneţă, and Herman Kamper. 2023. Visually grounded few-shot word learning in low-resource settings. *arXiv preprint arXiv:2306.11371*.

Xenia Ohmer, Michael Franke, and Peter König. 2022. Mutual exclusivity in pragmatic agents. *Cognitive Science*, 46(1):e13069. https://doi.org/10.1111/cogs.13069, PubMed: 34973036

Kayode Olaleye, Dan Oneţă, and Herman Kamper. 2022. Keyword localisation in untranscribed speech using visually grounded speech models. *Journal of Selected Topics in Signal Processing*, 16(6):1454–1466. https://doi.org/10.1109/JSTSP.2022.3180220

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2015.7178964

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2023. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics (2024)*, 12:372–391. https://doi.org/10.1162/tacl_a_00656

Puyuan Peng and David Harwath. 2022a. Fast-slow transformer for visually grounding speech. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP43922.2022.9747103

Puyuan Peng and David Harwath. 2022b. Self-supervised representation learning for speech using visual grounding and masked language modeling. In *AAAI Conference on Artificial Intelligence SAS Workshop*.

Puyuan Peng and David Harwath. 2022c. Word discovery in visually grounded, self-supervised speech models. In *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2022-10652

Puyuan Peng, Shang-Wen Li, Okko Räsänen, Abdelrahman Mohamed, and David Harwath. 2023. Syllable discovery and cross-lingual generalization in a visually grounded, self-supervised speech mode. In *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2023-2044

Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2018.8461326

Mark Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95. https://doi.org/10.1016/j.specom.2004.09.001

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*.

Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath. 2023. SpeechCLIP: Integrating speech with pre-trained vision and language model. In *Proceedings of SLT*. https://doi.org/10.1109/SLT54892.2023.10022954

Erik Thiessen. 2010. Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, 34(6):1093–1106. https://doi.org/10.1111/j.1551-6709.2010.01118.x, PubMed: 21564244

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. In *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2020-1693

Wai Keen Vong and Brenden Lake. 2022. Cross-situational word learning with multimodal neural networks. *Cognitive Science*, 46(4):e13122. https://doi.org/10.31234/osf.io/udbh2, PubMed: 35377475

Yu-Hsuan Wang, Hung-yi Lee, and Lin-shan Lee. 2018. Segmental audio Word2Vec: Representing utterances as sequences of vectors with applications in spoken term detection. In *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2018.8462002

Chen Yu and Linda Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x, PubMed: 17576281

## A Testing for Statistical Significance

To determine whether the differences between our model variations and a random baseline are statistically significant, we fit two types of logistic mixed-effects regression models to the data, where each of them predicts the (binary) model's choice for each test episode. All models are fitted using the lme4 package (Bates et al., 2015). Unlike many other statistical tests, mixed-effects models take into account the structure of the data: For example, certain classes or even individual images/queries are used in multiple pairwise comparisons.

The first mixed-effects model tests whether each MattNet's variation is better than the random baseline: it uses the MattNet variation as a predictor variable and random intercepts over trials, test episodes, the specific acoustic realisation of the test query, individual image classes and their pairwise combinations, and specific images in the test episode.

The second mixed-effects model does not consider the random baseline, and instead tests whether adding the visual initialization, the audio initialization or a combination of both improves MattNet: It uses the presence (or lack of) visual initialization and audio initialization as two binary independent variables, as well as their interaction, and the same random intercepts as described above.

In Section 7.1 we additionally test whether MattNet's scores in the familiar–novel test are significantly higher than in the novel–novel test. For this, we fit a logistic mixed-effects model to MattNet's combined scores from both tests, with test type and model variation as predictor variables, together with their interaction, as well as random intercepts as described above.

## B Test Notation

| Setup | Query audio | Target image | Other image |
|---|---|---|---|
| Familiar–familiar | *familiar* | FAMILIAR | FAMILIAR* |
| Familiar–novel | *familiar* | FAMILIAR | NOVEL |
| Familiar–novel | *novel* | NOVEL | FAMILIAR |
| Novel–novel | *novel* | NOVEL | NOVEL* |
| Familiar–novel* | *novel* | NOVEL* | FAMILIAR |

Table 6: A summary of the evaluation setups in terms of the input types (familiar or novel) used for the audio query and the two images. The asterisk indicates different classes for the same input type. For example, FAMILIAR and FAMILIAR* are two different familiar classes.