The mutual exclusivity bias of bilingual visually grounded speech models

Dan Oneață[1], *Leanne Nortje*[2], *Yevgen Matusevych*[3], *Herman Kamper*[2]

[1]SpeeD Lab, Politehnica Bucharest, Romania
[2]Electrical and Electronic Engineering, Stellenbosch University, South Africa
[3]CLCG, University of Groningen, the Netherlands

{dan.oneata,nortjeleanne,kamperh}@gmail.com yevgen.matusevych@rug.nl

Abstract

Mutual exclusivity (ME) is a strategy where a novel word is associated with a novel object rather than a familiar one, facilitating language learning in children. Recent work has found an ME bias in a visually grounded speech (VGS) model trained on English speech with paired images. But ME has also been studied in bilingual children, who may employ it less due to cross-lingual ambiguity. We explore this pattern computationally using bilingual VGS models trained on combinations of English, French, and Dutch. We find that bilingual models generally exhibit a weaker ME bias than monolingual models, though exceptions exist. Analyses show that the combined visual embeddings of bilingual models have a smaller variance for familiar data, partly explaining the increase in confusion between novel and familiar concepts. We also provide new insights into why the ME bias exists in VGS models in the first place. Code and data: `https://github.com/danoneata/me-vgs`.

Index Terms: visually grounded speech models, language acquisition, mutual exclusivity, multilingual, cognitive science

1. Introduction

The mutual exclusivity (ME) bias is a constraint that young children use in language learning, where they prefer to associate novel words with unfamiliar referents rather than familiar ones. For instance, if a child hears a novel word *aardvark* during book reading, they will naturally map it to the unusual animal in the picture rather than the ordinary cat beside it. This strategy enables efficient learning by narrowing down the space of possible referents [1] and has been well-documented in children [2–4].

At the same time, the ME strategy might not apply to the same extent in bilingual situations, where each object can have more than one name. Bilingual children therefore generally show a weaker ME bias compared to monolingual children [5–7]. But this is not always the case, with results affected by a child's age [8, 9], their vocabulary [10], the amount of time since their exposure to the familiar objects [11], and the type of test used [9, 12]. In short, the big picture of the differences in ME bias between monolingual and bilingual children is still not clear.

In this study we investigate the bilingual ME bias from a computational perspective using multimodal machine learning models. While the ME bias in bilingual learners has not been studied computationally, there are studies on the monolingual bias. Most of these use text–image models associating written words with visual objects; findings are mixed in reproducing the ME bias [13–15]. Very recently, ME has been studied in visually grounded speech (VGS) models [16], which better approximate children's reliance on spoken (rather than written) language by operating on images and spoken words. Moreover, [16] reliably reproduces monolingual children's ME bias.

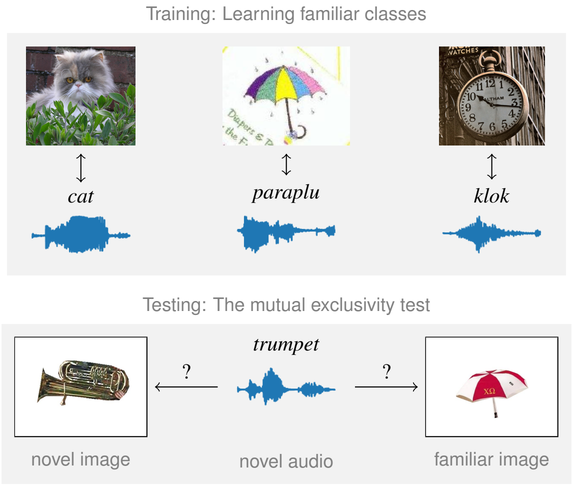
Training: Learning familiar classes
cat
paraplu
klok
Testing: The mutual exclusivity test
trumpet
?
?
novel image
novel audio
familiar image

Figure 1: *Looking for ME in bilingual VGS models. We first train a model on images and spoken words from two languages (e.g., English and Dutch). We then test for ME by pairing a novel word with a novel image and a familiar image. If the model gives a higher score to the novel–novel pair, then it has an ME bias.*

Taking Nortje *et al.* [16] as a starting point, we compare the ME bias in monolingual and bilingual VGS models. We first improve the model, leading to better and more consistent results in the monolingual case. We then train various combinations of monolingual and bilingual VGS models using English, French and Dutch speech–image data, as illustrated in Fig. 1. We find that in most cases the bilingual models exhibit a smaller ME bias, but the results are not consistent (as in the human studies).

We then present analyses to try and understand the organisation of the resulting VGS embedding spaces in mono- and bilingual models. In both cases, we find a modality gap in the joint audio–image space. Qualitatively, we also show for the first time that novel concepts are placed in-between familiar concepts in the embedding space. The main difference between monolingual and bilingual models is that the familiar images are packed tighter in the bilingual case. Our analyses lay the foundation for future work at larger scales on more language pairs.

2. Data and method

To investigate the ME bias computationally, we use a VGS model that takes audio and images as input. We train the model to associate spoken utterances of object names to their visual correspondences (Fig. 1-top). The concepts seen during training are then familiar to the learner. To test the model's ME bias after learning, we prompt it with a spoken query from an unseen novel class and ask it to select one of the two images, one showing a familiar and the other a novel object (Fig. 1-bottom). If the model tends to associate the novel spoken word with the novel object, it has an ME bias.

The setup is identical in the mono- and bilingual cases, except that in the bilingual setting, the training and test words come from two languages. For the bilingual English–Dutch example in Fig. 1, the novel *trumpet* (bottom) is not seen in either language during training, but the familiar *umbrella / paraplu* is seen in both English and Dutch during training.

Below we first describe how we construct the bilingual speech–image datasets used for training the VGS models. We then describe how the model is structured, trained, and evaluated.

2.1. Multilingual datasets

We need a dataset containing images paired with spoken words in multiple languages. We therefore extend the English speech–image data from [16] with Dutch and French speech.

The English data from [16] contains 13 familiar word classes and 20 novel word classes. These are concrete nouns like the ones in Fig. 1: *cat*, *umbrella*, *clock*. Speech segments for these words are sourced from the Flickr Audio Captions Corpus [17], Buckeye [18], and LibriSpeech [19]. The corresponding images come from MS COCO [20], Caltech-101 [21], and ImageNet [22]. For training, full images are used, while the test images contain the objects isolated using a white background mask based on the object segmentations provided with the datasets. Training therefore happens in a cluttered natural environment, while testing is done through unambiguous evaluations; this is similar to children learning a language in a naturalistic setting and then being tested in a laboratory.

We obtain Dutch and French spoken words for the 33 classes present in the English data. Dutch words are extracted from the Corpus Gesproken Nederlands [23] and the Dutch subsets of multilingual LibriSpeech [24] and Common Voice [25]. For French, we use the subsets from multilingual LibriSpeech and Common Voice. We isolate the target spoken words using forced alignment [26, 27]. Since for one novel word, *nautilus*, we have few samples in Dutch and French, we discard it, leaving us with 19 novel words. The class distribution follows the source data, with roughly 67k images and 4k audio samples for the most common concept (*dog*) and 47 images and 107 audio samples for the least common concept (*scissors*) across the three languages.

2.2. Visually grounded speech (VGS) model architecture

Our VGS model simulates word learning in a cognitively motivated manner, similar to a child learning to map object names to their visual referents. Given an audio $\mathbf{a}$ and image $\mathbf{i}$, the model produces a score $\phi(\mathbf{a}, \mathbf{i})$ of how well the two match.

As illustrated in Fig. 2, the architecture of the VGS model is a two-tower encoder network coupled with a contrastive loss. We use WavLM [28] to extract features from the audio $\mathbf{a}$, and DINO [29] to extract features from the image $\mathbf{i}$. Both have been pretrained on other unlabelled datasets using self-supervised learning, and are kept frozen throughout VGS model training. This can be seen as a proxy for how children during word learning can use visual and auditory perceptual abilities previously acquired from exposure in the respective modalities [30, 31]. The feature extractors return sequences of representations; to obtain a single embedding vector for each modality, we use a pooling layer. The resulting embeddings are then L2 normalised. A dot product gives the similarity score $\phi(\mathbf{a}, \mathbf{i})$ between the two inputs.

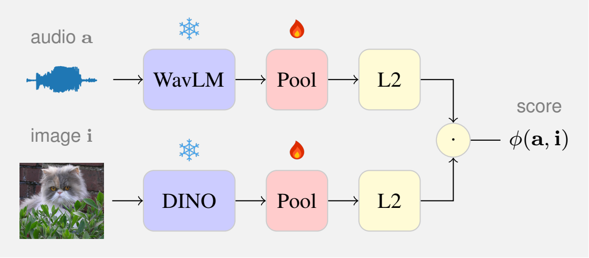

Figure 2: *The architecture of our VGS model. The only parameters that are updated are the transformer pooling layers.*

The pooling layer consists of a single transformer block with a learnable CLS token as the query vector. This layer also incorporates two down-projection layers: at the input, from the feature dimension D to the transformer width W, and at the output, from the transformer width W to the embedding dimension E. Based on validation experiments, we use transformer blocks with $W = E = 256$, four heads $(W/64)$, and an inner MLP dimension of 1024 $(W \times 4)$. The two transformer blocks (one for audio and one for image) are the only learnable layers in the model, amounting to approximately 2.5M parameters.

Compared to the VGS model from [16], our architecture is both more modern (updates the AlexNet image encoder to ResNet and the CPC audio encoder to WavLM) and simpler (pools both modalities in the same way instead of pooling only the audio modality and then max-pooling the scores). These changes substantially improve the model's ability to discriminate between familiar classes (as we show below in Sec. 3), allowing us to look for the ME bias in strong learners. Finally, since we rely on frozen encoders, training is also more efficient.

2.3. Model training

At each training step, the model receives an audio–image pair $(\mathbf{a}^+, \mathbf{i}^+)$ corresponding to the same word class and a set of negative audio samples $\{\mathbf{a}^-\}$ and image samples $\{\mathbf{i}^-\}$ from other classes. Both the positive and negative samples come from familiar classes. The negatives are sampled independently for the two modalities, so they are not necessarily matched. We define the probability that the model matches the spoken word $\mathbf{a}^+$ to the correct image $\mathbf{i}^+$ as follows:

$$p(\mathbf{i}^+|\mathbf{a}^+) = \frac{\exp\{\phi(\mathbf{a}^+, \mathbf{i}^+)/\tau\}}{\sum_{\mathbf{i}} \exp\{\phi(\mathbf{a}^+, \mathbf{i})/\tau\}}, \tag{1}$$

where $\mathbf{i}$ in the denominator ranges across the positive $\mathbf{i}^+$ and negative $\mathbf{i}^-$ samples, and τ is a learnable temperature parameter [32]. We define the reverse conditional probability $p(\mathbf{a}^+|\mathbf{i}^+)$ analogously and optimise the parameters (the pooling layer and the temperature τ) to maximise the log of these two probabilities averaged across the samples in a batch.

The model is trained for 24 epochs using a learning rate with a linear warm-up for the first four epochs up to a learning rate of 2×10^{-4}, followed by cosine annealing to 10^{-6}. In an epoch we go through all audio samples in the dataset, and for each audio sample we randomly pick a positive image and 11 negative images. We monitor the performance on a validation split and select the model with the lowest loss on this split. For the encoders, we use the `base-plus` WavLM variant [28] (pretrained on English data; $D = 768$) and the ResNet-50 DINO variant [29] $(D = 2048)$. The temperature τ is capped at 100.

We train monolingual (English, Dutch, and French) and bilingual (English–Dutch, English–French, and Dutch–French) VGS models on the 13 familiar classes from our speech–image datasets. Since the number of epochs is fixed and in each epoch we go through all the audio samples, the bilingual models will go through more updates than the monolingual models. Results were similar when an equal number of training steps was used.

2.4. Model evaluation

To test a model, we present it with an audio query and two images: a positive one, which matches the class of the audio query, and a negative one, belonging to a different class from the audio query. Using this protocol, we consider two types of tests.

Familiar test. First, we measure discrimination ability across familiar classes to ensure the models are properly trained. In this test, the audio query belongs to a familiar class, and both images are also from familiar classes (one matching and one not).

ME test. Second, we quantify the ME bias. In this test, the audio query and the positive image come from a novel class, and the negative image from a familiar class (as in Fig. 1-bottom).

In both tests, we sample 50 episodes for each class from the test set, with no overlap between train, validation, and test samples. To prevent dataset biases, the two images in each pair are drawn from the same source dataset (Sec. 2.1). Each experiment is repeated five times with a different seed affecting weight initialisation and data sampling. For the monolingual models, the test audio query matches the training language. For bilingual models, the audio query can come from either training language.

3. Results and analyses

We want to see whether both monolingual and bilingual models show an ME bias; whether the strength of the bias differs between the two; and what the similarities and differences are in how the embedding spaces are organised.

3.1. Experimental results

The results in Table 1 are ordered according to the language of the test query. Familiar performance is close to perfect in all settings, for all languages, and even in the bilingual case. The models have therefore properly learned the familiar words and can robustly recognise the corresponding objects despite the masked-out background at test time. This validates our computational setup and is crucial for a meaningful ME test. Note that overall, our results on the familiar test substantially improve on those reported in prior work [16], likely due to the use of self-supervised features in our model (see Sec. 2.2).

Having established that the models are well-trained, we now consider the ME test. All the monolingual models yield an accuracy of 67–68%, so they all show an ME bias (50% would indicate no bias). These results show for the first time consistent ME biases for monolingual VGS models trained on languages other than English. Our scores on English are higher than the ME results of roughly 60% in [16], suggesting that stronger learners (like the model in this study) show a stronger ME bias.

We now turn to our main question: a comparison of the ME results between monolingual and bilingual models. When tested on English, we see that the ME score of 66.2% for an English-only model drops to 65.7% when French is added, or to 63.5% when Dutch is added (Table 1 top). ME scores similarly drop when English is added on the French test (middle), and when English or French is added on the Dutch test (bottom). This trend matches the findings from experiments on children, indicating that bilingual children make less use of the ME bias than monolingual children [5–7]. But our result does not hold in every single case: the Dutch–French model tested on French gives an ME score of 69.4%, which is higher than the 67.6% from the French-only model.

Table 1: *Performance on the familiar and ME tests for monolingual and bilingual VGS models. We report mean accuracy (%) and standard error computed across five training seeds.*

Training languages	Test language	Familiar	ME
Monolingual: EN	English (EN)	99.4±0.1	66.2±1.1
Bilingual: EN, FR		99.6±0.1	65.7±1.3
Bilingual: EN, NL		99.6±0.1	63.5±1.5
Monolingual: FR	French (FR)	98.5±0.4	67.6±1.4
Bilingual: FR, EN		98.9±0.1	66.8±1.4
Bilingual: FR, NL		99.0±0.1	69.4±0.9
Monolingual: NL	Dutch (NL)	98.5±0.3	67.3±1.3
Bilingual: NL, EN		98.7±0.3	63.5±2.1
Bilingual: NL, FR		98.6±0.3	65.7±1.2

The results in Table 1 are for models with roughly 2.5M parameters, but we also repeated the experiments with smaller ($W = 128$, 0.8M parameters) and larger models ($W = 512$, 8.2M parameters). We observe similar trends: in most but not all cases the bilingual models exhibit a smaller ME bias compared to the monolingual ones. We carried out several statistical comparisons between bilingual and their corresponding monolingual models. Out of 18 tests (3 languages × 2 language pairs × 3 model sizes), 9 show that the bias is significantly stronger in the bilingual models, and 6 show a difference in the expected direction but not significantly so. Human studies have similar discrepancies: it isn't always the case that bilingual children show less ME, with age [8, 9], vocabulary [10], and the type of test [9, 12] all seeming to play a role.

We have shown that visually grounded bilingual models tend to have a lower ME bias than monolingual ones. But why does this happen? Below we first look at why the ME bias is seen in general, before considering the mono- vs bilingual question.

3.2. Understanding the embedding space

In an analysis of English-only models, Nortje *et al.* [16] showed quantitatively that novel audio samples are generally closer to novel images (regardless of class) than they are to any familiar class. This is why we see the ME bias. But how is the representation space organised qualitatively? The structure of the model of [16] did not allow for a sensible visualisation of the embedding spaces,[1] but our changes (Sec. 2.2) enable visualisation. We therefore contribute the following new analyses.

Fig. 3-left uses a PCA projection to visualise the 256-dimensional embeddings for both the audio and image samples. The linear PCA projection allows us to see the global structure of the data. Since the embeddings are L2 normalised, they live on the unit sphere. The audio and image embeddings are positioned on different sides. This is known as the modality gap and has been observed in other bimodal models trained with a contrastive loss [33]. Despite the gap, the familiar audio and image classes are well-aligned, e.g., the green points are at the top and the blue

[1] The model in [16] did not extract an image embedding, but rather patch embeddings that were aggregated non-linearly into the final score.

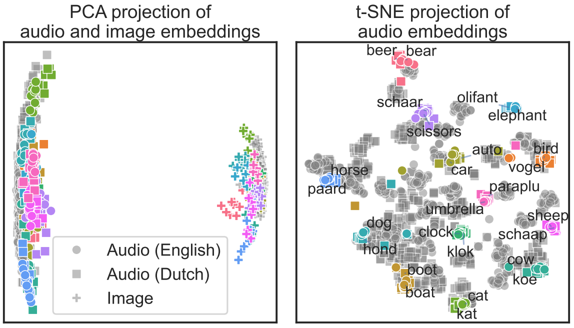

Figure 3: *Projections of embeddings from the bilingual English–Dutch model. Familiar classes are coloured; novel classes are grey. Left: Audio and image embeddings live in different cones of the unit sphere. Right: Bilingual audio embeddings are aligned for the familiar classes, while novel classes live in-between.*

points at the bottom for both modalities. This ought to be the case for the strong familiar results in Table 1.

The PCA picture shows global structure but is not complete: from quantitative measurement, we know that on average novel audio (grey circles and squares) are closer to novel images (grey crosses) than they are to familiar images (coloured crosses), but this is not captured in Fig. 3-left. We therefore analyse individual modalities using a non-linear visualisation. Fig. 3-right uses t-SNE to visualise the audio embeddings. The model separates out familiar classes throughout the representation space and places novel classes in an in-between region. We verify this interpretation quantitatively by computing the variance of familiar and novel samples, both when all the samples are combined or when considering samples per class.[2] Fig. 4 shows results for an English model and two bilingual models. We see that, indeed, the overall variance of the novel data is much smaller than that of the familiar data. At the same time, the per-class variance, shown in the second plot in Fig. 4, is somewhat smaller for the familiar than for the novel data. Taken together, these variance values support the hypothesis stated in [16] that during training the model spreads familiar classes around the space, with each class in a tight bundle of its own, while novel classes are placed between familiar ones, in a region where novel samples from different classes overlap. Fig. 4 shows that this is the case for both mono- and bilingual models.

3.3. Monolingual vs bilingual models

The analyses above give new insight into why we observe the ME bias in both mono- and bilingual models. But why is the bias slightly weaker in bilingual models? Given that ME is the result of comparisons between two modalities in a high-dimensional space (with a modality gap), it is inherently difficult to explain the small differences in results. But we can use the variance analysis of Fig. 4 to get some insights. A consistent change when adding either French or Dutch to the English model is that the spread across samples in the visual modality becomes tighter, in particular for familiar samples (compare the blue image bars in the first plot vs the third and fourth plots). While the variance of all the novel images is also slightly smaller (brown bars), we speculate that the familiar space shrinks more (relatively speaking), resulting in more novel items being confused with familiar ones (i.e., a lower ME bias) in the bilingual case.

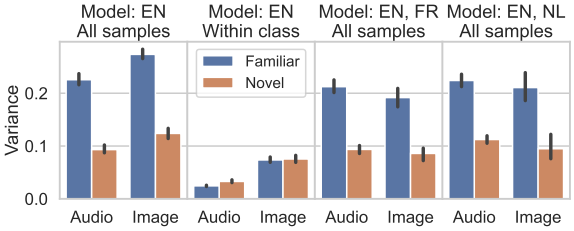

Figure 4: *Variance of familiar and novel samples for three models computed across all samples or samples within each class. Familiar samples occupy a larger space than novel ones (first plot), but they are tighter within class than novel samples (second plot). In the bilingual models (third and fourth plots), the image space gets tighter than in the monolingual case. Error bars are 95% confidence intervals computed by bootstrapping.*

This is not a comprehensive analysis: the ME bias exists because of comparisons across modalities, and not within a modality (which is what we do in the analysis here). Differences are also small, which complicates analyses. Repeating our analysis at a larger scale will therefore be necessary in future work.

3.4. Can a bilingual model implicitly translate?

An interesting secondary research question is how the bilingual models structure their acoustic spaces given supervision through the visual modality. In Fig. 3-right, we see that the audio embeddings of the familiar words in the two languages overlap. This happens regardless of whether the two words sound similar (*clock*–*klok*) or not (*horse*–*paard*). We quantify this translation performance by measuring the accuracy of a simple nearest mean centroid classifier: for each audio sample in one language, we find its closest audio centroid in the other language. With this approach, we achieve translation accuracies over 97% for all language pairs. The accuracies for the novel words are, as expected, poor: less than 30% in all cases. This is still better than random ($5.2\% = 1/19$ novel words), because some of the words are the same in all three languages (e.g., *bus*, *piano*).

4. Conclusion

In this study we investigated whether the ME bias—a heuristic employed by children in language learning—is also seen in visually grounded speech models trained on bilingual speech–image data. We found that bilingual models consistently exhibit an ME bias and that the strength of the bias tends to be weaker than for monolingual models, with some exceptions. These findings are consistent with those observed in children, where the ME bias has been reported to be generally lower in bilingual children, but this pattern is somewhat inconsistent [5–12]. While the results of our computational study cannot explain why certain patterns are observed in children, our model nonetheless can be used to generate predictions that can then be tested in experiments with children. Furthermore, in our computational model, we can carefully control the training data and analyse the model's internal representations. We relied on these advantages in this paper, but plan to use it even more in future work, increasing the number of language pairs and the vocabulary size to gain even more insights into the nature of the ME bias.

[2] The variance is the trace of the covariance matrix, or, equivalently, the mean of the distances between samples and their centroid.

5. Acknowledgements

This work was in part supported by the EU Horizon project AI4TRUST (No. 101070190) and by CNCS-UEFISCDI (PN-IV-P7-7.1-PTE-2024-0600).

6. References

[1] E. Markman and G. Wachtel, “Children’s use of mutual exclusivity to constrain the meanings of words,” *Cognitive Computation*, 1988.

[2] W. Merriman, L. Bowman, and B. MacWhinney, “The mutual exclusivity bias in children’s word learning,” *Monographs of the Society for Research in Child Development*, 1989.

[3] E. Markman, J. Wasow, and M. Hansen, “Use of the mutual exclusivity assumption by young word learners,” *Cognitive Psychology*, 2003.

[4] M. Lewis, V. Cristiano, B. M. Lake, T. Kwan, and M. C. Frank, “The role of developmental change and linguistic experience in the mutual exclusivity effect,” *Cognition*, 2020.

[5] K. Byers-Heinlein and J. F. Werker, “Monolingual, bilingual, trilingual: Infants’ language experience influences the development of a word-learning heuristic,” *Developmental Science*, 2009.

[6] K. Byers-Heinlein, “Bilingualism affects 9-month-old infants’ expectations about how words refer to kinds,” *Developmental Science*, 2017.

[7] C. Houston-Price, Z. Caloghiris, and E. Raviglione, “Language experience shapes the development of the mutual exclusivity bias,” *Infancy*, 2010.

[8] D. Davidson and D. Tell, “Monolingual and bilingual children’s use of mutual exclusivity in the naming of whole objects,” *Journal of Experimental Child Psychology*, 2005.

[9] M. Kalashnikova, K. Mattock, and P. Monaghan, “The effects of linguistic experience on the flexible use of mutual exclusivity in word learning,” *Bilingualism: Language and Cognition*, 2015.

[10] K. Byers-Heinlein and J. F. Werker, “Lexicon structure and the disambiguation of novel words: Evidence from bilingual infants,” *Cognition*, 2013.

[11] M. Kalashnikova, P. Escudero, and E. Kidd, “The development of fast-mapping and novel word retention strategies in monolingual and bilingual infants,” *Developmental Science*, 2018.

[12] D. Davidson, D. Jergovic, Z. Imami, and V. Theodos, “Monolingual and bilingual children’s use of the mutual exclusivity constraint,” *Journal of Child Language*, 1997.

[13] K. Gandhi and B. Lake, “Mutual exclusivity as a challenge for deep neural networks,” in *Proc. NeurIPS*, 2020.

[14] K. Gulordava, T. Brochhagen, and G. Boleda, “Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks,” in *Proc. CogSci*, 2020.

[15] W. K. Vong and B. Lake, “Cross-situational word learning with multimodal neural networks,” *Cognitive Science*, 2022.

[16] L. Nortje, D. Oneață, Y. Matusevych, and H. Kamper, “Visually grounded speech models have a mutual exclusivity bias,” *Transactions of the Association for Computational Linguistics*, 2024.

[17] D. Harwath and J. Glass, “Deep multimodal semantic embeddings for speech and images,” in *Proc. ASRU*, 2015.

[18] M. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, “The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability,” *Speech Communication*, 2005.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015.

[20] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common objects in context,” in *Proc. ECCV*, 2014.

[21] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, 2017.

[23] N. Oostdijk, “The spoken Dutch corpus: Overview and first evaluation,” in *Proc. LREC*, 2000.

[24] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “MLS: A large-scale multilingual dataset for speech research,” in *Proc. Interspeech*, 2020.

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proc. LREC*, 2020.

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using Kaldi,” in *Proc. Interspeech*, 2017.

[27] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli, “Scaling speech technology to 1000+ languages,” *Journal of Machine Learning Research*, 2024.

[28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[29] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proc. ICCV*, 2021.

[30] E. V. Clark, “How language acquisition builds on cognitive development,” *Trends in Cognitive Sciences*, 2004.

[31] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins, “Unsupervised neural network models of the ventral visual stream,” *Proceedings of the National Academy of Sciences of the USA*, vol. 118, no. 3, 2021.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proc. ICML*, 2021.

[33] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” in *Proc. NeurIPS*, 2022.