Computational Insights to Acquisition of Phonemes, Words, and Word Meanings in Early Language: Sequential or Parallel Acquisition?

Khazar Khorrami (khazar.khorrami@tuni.fi)
María Andrea Cruz Blandón (maria.cruzblandon@tuni.fi)
Okko Räsänen (okko.rasanen@tuni.fi)
Unit of Computing Sciences, Tampere University
P.O. Box 553, FI-33101, Finland

Abstract

Previous computational models of early language acquisition have shown how linguistic structure of speech can be acquired using auditory or audiovisual learning mechanisms. However, real infants have sustained access to both uni- and multimodal sensory experiences. Therefore, it is of interest how the uni- and multimodal learning mechanisms could operate in concert, and how their interplay might affect the acquisition dynamics of different linguistic representations. This paper explores these questions with a computational model capable of simultaneous auditory and audiovisual learning from speech and images. We study how the model's latent representations reflect phonemic, lexical, and semantic knowledge as a function of language experience. We also test how the findings vary with differential emphasis on the two learning mechanisms. As a result, we find phonemic learning always starting to emerge before lexical learning, followed by semantics. However, there is also notable overlap in their development. The same pattern emerges irrespectively of the emphasis on auditory or audiovisual learning. The result illustrates how the acquisition dynamics of linguistic representations are decoupled from the primary learning objectives (mechanisms) of the learner, and how the emergence of phonemes and words can be facilitated by both auditory and audiovisual learning in a synergetic manner.

Keywords: computational modeling; language acquisition; visual grounding; statistical learning; unsupervised learning

Introduction

Computational models of early language acquisition (LA) focus on understanding how infants learn to bootstrap their language acquisition process. This includes questions such as how phonemic categories are acquired, how auditory word forms or lexemes are learned, or how words and phrases become related to their meanings. Typical modeling research tries to identify what kind of inputs, outputs, learning mechanisms, and innate biases or constraints are essential for human-like LA, and how variability in such factors affects learning outcomes.

Classical modeling studies often focus on learning of specific linguistic units using dedicated processing mechanisms for them, such as phonemic category acquisition with Bayesian inference (Feldman, Griffiths & Morgan, 2009) or word segmentation with statistical cues (Frank et al., 2010). However, recent advances in machine learning (ML) have enabled increasingly powerful autonomous learning algorithms that do not explicitly target any particular language structures but use more generic learning principles instead. For instance, speech representations can be learned from unlabeled acoustic speech using so-called self-supervised learning (SSL; e.g., van den Oord, Li & Vinyals, 2018, Baevski et al., 2020). In SSL, the task of the model is to predict or reconstruct unobserved parts of the speech stream, not unlike to models of predictive processing ascribed to the human brain. Another branch of algorithms, known as visually-grounded speech (VGS) models, learn to link patterns of spoken language with semantically related visual inputs through associative learning (see Chrupała, 2022, for a review), again not unlike to cross-situational learning of word meanings in developmental literature (Smith & Yu, 2008). Given that SSL and VGS algorithms do not use any prior knowledge of structures in spoken language (beyond some basic constraints on timescales of speech), they are also relevant as computational models of infant learning.

Previous research with SSL and VGS models has shown that they are indeed capable of acquiring a range of language representations, such as phonemes, syllables, words, and word semantics (see, e.g., Chrupała et al., 2017; Chrupała, 2022; Dunbar, Hamilakis & Dupoux, 2022; Khorrami & Räsänen, 2021; Merkx, 2022; Nikolaus & Fourtassi, 2021). Recently Lavechin et al. (2023) investigated SSL-based model of infant statistical learning using contrastive predictive coding (CPC; van den Oord et al., 2018) as the SSL mechanism. They showed that the model can acquire highly selective phonetic representations together with clearly above-chance lexical and syntactic category knowledge, and these all improve with increasing language experience. Peng and Harwath (2022b) recently combined auditory (SSL) and audiovisual (VGS) learning into a single neural network model architecture that can use both mechanisms in a sequence or in parallel. They showed that auditory learning before audiovisual learning can lead to notable learning gains in acquiring sematic relationships between speech and visual input. Peng and Harwath (2022c) also showed how visual grounding facilitates word segmentation in a similar model. Overall, the existing research demonstrates that auditory and audiovisual learning mechanisms can support acquisition of linguistic patterns, both independently and in conjunction.

However, what is unclear from the existing research is how the auditory and audiovisual learning mechanisms cooperate during learning, and how different levels of language emerge as a function of language experience in the presence of the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

two mechanisms (considering, e.g., synergies across linguistic representations; e.g., Johnson et al., 2010).

In this paper, we explore the temporal dynamics of early LA in the parallel presence of auditory and audiovisual learning mechanisms in a computational model of language learning. We apply the model to a large dataset of acoustic speech and related visual inputs and measure how phonemic, lexical, and semantic knowledge emerge in the model as a function of model training time ("language experience"). We also explore how the acquisition pattern depends on the relative emphasis on, and timing of, auditory and audiovisual learning.

Theoretical Considerations

Based on the developmental literature, one could formulate two archetypes of the acquisition process: A primarily sequential and compositional "bottom-up" process, and a holistic meaning -oriented "top-down" process. In the first, some amount of phonetic perceptual organization is expected to precede lexical learning (cf., e.g., NLM-e theory by Kuhl et al., 2007), which then enables word segmentation (cf., Saffran et al., 1996) and acquisition of word meanings through situated grounding (e.g., Smith & Yu, 2008). In the top-down case, early language would be bootstrapped by grounding of holistic phrase-like speech patterns with other multimodal and/or embodied experiences such as visual input. In this case, phonemic and lexical representations would not be proximal targets of learning, but gradually emerge through analysis and decomposition of the situated language patterns into constituents that enable more efficient encoding of the language (cf., e.g., PRIMIR theory by Werker & Curtin, 2005; see also Khorrami & Räsänen, 2021; Merkx, 2022; Tomasello, 2000; see also Räsänen & Rasilo, 2015, for a discussion on the bottom-up and top-down strategies).

In terms of computation, VGS models can be seen as an example of the top-down approach, as their learning criterion is focused on grounding of utterance-level auditory representations to their visual referents (Merkx, 2022). In contrast, SSL models could be seen as "bottom-up" in the sense that they focus on low-level auditory pattern discovery and without a mechanism to ground speech into referential meanings. However, the two archetypes do not have to be mutually exclusive, as auditory statistical learning and cross-modal associative learning can be present in parallel. Also, as we will show, a learner focusing on holistic learning of usage-based meanings may still *exhibit* acquisition order compatible with the bottom-up view.

Methods

Computational Model

As our model of early LA, we adopt the so-called FaST-VGS+ neural network architecture by Peng and Harwath (2022b). The model is a fusion of two earlier models with two distinct learning mechanisms: Wave2Vec 2.0 (Baevski et al., 2020) for SSL-based learning from acoustic speech only (from now on: W2V2) and FaST-VGS from Peng and Harwath (2022a) for visually grounded learning from utterances and images related to speech contents. The high-level architecture of FaST-VGS+ is visualized in Fig. 1.

In auditory learning, the task of the network is to predict its own representational activations (quantized outputs of the wave encoder in Fig. 1) for speech segments that are masked from the rest of the processing chain, and when the network can make use of the non-masked parts of the utterance. When optimized for the prediction task, the network learns contextualized signal representations to "fill in the blanks" (see Baevski et al., 2020, for details). In case of audiovisual learning, the task of the network is to learn vector embeddings derived from images and utterances in such a manner that the embeddings from the two modalities are similar when the speech and image arrive concurrently (as a proxy for semantic relatedness). In contrast, when an utterance and image do not match, the embeddings should be different. By optimizing the network on the embedding similarity/dissimilarity task, it learns a latent semantic space in which relatedness of speech and images can be measured by comparing their embeddings with a chosen distance metric (see, e.g., Chrupała, 2020, for an overview).

In the original FaST-VGS+ formulation, the model consists of four optimization loss terms that were combined as a weighted sum during the training: 1) a standard audiovisual contrastive loss $\mathcal{L}_{\mathrm{AV}}$ for audiovisual grounding based on similarity-scores of holistic utterance- and image-level embeddings, 2) an additional "fine-grained" audiovisual loss $\mathcal{L}_{\mathrm{AV,F}}$ that applies cross-modal Transformer attention to the acoustic and visual embeddings before calculating their similarity, 3) an auditory reconstruction loss $\mathcal{L}_{\mathrm{AUD,R}}$ of W2V2 that is primarily responsible for the auditory-only learning, and 4) an auditory diversity loss $\mathcal{L}_{\mathrm{AUD,D}}$ that encourages variability to self-learned representations to avoid trivial local optima.

In the present work, we do not use the $\mathcal{L}_{\mathrm{AV,F}}$ loss due to its substantial computational burden and minor performance gains. Following Baevski et al. (2020), we also tie the W2V2-related losses ($\mathcal{L}_{\mathrm{AUD,R}}$ and $\mathcal{L}_{\mathrm{AUD,D}}$) together with a ratio of 1:0.1. Given that the numeric scales of the auditory and audiovisual loss terms are similar during the training (approx. $\mathcal{L}$ ~ 0–10 overall), we can then control the relative weights of auditory and audiovisual learning with just one parameter α:

$$\mathcal{L} = (1-\alpha)\big(\mathcal{L}_{\mathrm{AUD,R}} + 0.1\,\mathcal{L}_{\mathrm{AUD,D}}\big) + \alpha\mathcal{L}_{\mathrm{AV}} \quad (1)$$

where $\mathcal{L}$ denotes total loss of the model. For instance, by setting $\alpha = 0$, the model only uses the auditory learning mechanism, whereas $\alpha = 1$ only uses audiovisual learning.

As for the details of different processing modules in Fig. 1, the waveform encoder is a 6-layer convolutional neural network (CNN) that maps acoustic waveforms (sampled at 16 kHz) into a sequence of latent 512-d embeddings (one embedding every 10 ms). Audio encoder is a stack of 8 Transformer layers that feeds into two paths: a 5-layer CNN ("ResDAVEnet") responsible for gradual temporal pooling (downsampling) of the utterance-level latent representations, and a Transformer-based 4-layer audio decoder ("*Trm3*" in Peng & Harwath, 2022b) whose task is to predict the masked speech segments. The downsampling CNN feeds to a single Transformer layer that summarizes the CNN output into a single 768-d auditory embedding **a** (the so-called CLS-token) using self-attention. The visual encoder consists of 6 Transformer layers and converts pixels from detected object regions of an input image into an overall 768-d visual embedding **v**, again captured by the CLS token of the final layer. Finally, the embeddings from the auditory and visual paths are used for audiovisual similarity scoring $S(\mathbf{a},\mathbf{v})$ (with a dot product) which are used in triplet loss ($\mathcal{L}_{\mathrm{AV}}$) calculation. Please see Peng and Harwath (2022b) for model details.

During training, the waveform encoder and audio encoder are updated by both auditory and audiovisual learning, depending on the relative weights of the two losses, whereas audio decoder is only updated by auditory learning. Like Peng and Harwath (2022b), we focus on the analysis of linguistic representations (here: phonemes and words) in the hidden layers of the audio encoder and decoder, as they are primarily responsible for creating contextualized speech representations for the two tasks. For semantic analyses, we focus on capability of auditory embeddings **a** to carry cross-modal semantic knowledge with respect to visual domain.

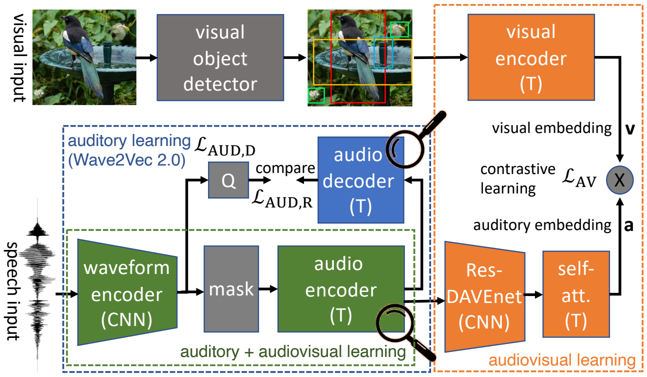

Figure 1: An overview of the model architecture, as adapted from Peng and Harwath (2022b). Green waveform and audio encoder neural blocks are shared and optimized by both auditory learning (W2V2) and audiovisual learning (VGS) based on their relative weights. Orange blocks are only updated by the audiovisual learning, and the blue block by auditory learning only. In the experiments, hidden layers of the audio encoder and decoder are analyzed for selectivity towards phonemes and words (marked with magnifiers). "*Q*" denotes vector quantization, "*mask*" training-time masking of representations from random speech segments, "*T*" Transformer layers, and "*CNN*" convolutional layers.

Data

The experiments used acoustic speech and image data of SpokenCOCO dataset (Hsu et al., 2020). The dataset consists of 123k images (photos of everyday scenes; originally from MSCOCO by Lin et al., 2014) that each come with five spoken captions describing the images in English, resulting in a total of 742 h of speech (2353 speakers). We use the data split of Karpathy and Fei-Fei (2017) with 118k images for training and 5k for testing. Given that the utterances and images consist of several words and visual objects, there is substantial referential ambiguity in the data.

Model Evaluation

Evaluation focused on analyzing how phonemic, lexical and semantic information emerge in the model as a function of training time (a proxy for model's language experience).

Phonemic Evaluation. For phonemic evaluation, we applied the widely used ABX-task (Schatz et al., 2013), also used by Lavechin et al. (2022) and Peng and Harwath (2022b). The task consists of feeding the model with a large number of minimal pair speech triplets (e.g., A: "*bat*" B: "*bit*" X: "*bit*") and testing whether the resulting hidden layer activations of the model are more similar for the tokens containing the same target phoneme than across phonemic categories. Representation similarity is measured with dynamic time warping -based alignment of the representations in time, followed by calculation of the alignment path cost (see Schatz et al., 2013, for more details). Output of the test is ABX error rate (%): the percentage of ABX-trials in which the model failed to discriminate the minimal pair token from the same-phoneme tokens. For the ABX test, we used the English Librispeech version of the test (www.zerospeech.com; Dunbar et al., 2022), for which we report the ABX scores for the within-speaker clean data condition.

Lexical Evaluation. For lexical evaluation, we created a new[1] benchmark called "*CDI-Lextest*" to evaluate how well a model's hidden layer activations discriminate between different lexical items. For this purpose, we synthesized isolated tokens of 89 lexical types from the receptive vocabulary of English CDI short forms (list taken from Wordbank; Frank et al., 2016). These words represent common early vocabulary items of North American English learners. Google cloud text-to-speech API was used for the synthesis, and each lexical type was synthesized using 10 different neural voices (5 male, 5 female) and in both question and statement prosodic forms. This resulted in a total of 1780 word tokens with 20 tokens per lexical type.

During the test, the model is fed with a word token at a time and should output a fixed-dimensional representation vector for the token irrespectively of the duration (e.g., mean hidden layer activation across time). Pairwise cosine distances between all the 1780 token vectors are then calculated. A quality score (0–1) is assigned for each token, where the score is defined as the proportion (0–1) of the nearest 19 other tokens being of the same word type as the given token, since there are a total of 19 other tokens of the same lexeme in the test set (e.g., given a "*ball*" token, what is the proportion of

[1] CDI-Lextest is available for download at https://github.com/SPEECHCOG/CDI_lextest.

other "*ball*" tokens within the nearest 19 tokens). The final lexical quality score $q_{\text{LEX}} \in [0, 1]$ (chance ~1/89) is calculated as the mean of token scores across all the 1780 tokens. As a result, the test measures how well tokens of each word type cluster together in the representation space while avoiding confusions across word types.

Semantic Evaluation. Semantic evaluation was performed in terms of standard audiovisual retrieval using recall@10 as the metric (see, e.g., Chrupała, 2020). For each utterance in the test set of SpokenCOCO, the corresponding embedding ("query") **a** was first extracted from the output of the audio model branch (see Fig. 1). Cosine distances between the query and image embeddings **v** of all test set images (from the visual encoder) were then calculated. If the embedding of the matching image was within the 10 nearest embeddings to the query embedding, the retrieval task was considered as successful. The proportion of successful retrievals across all 25000 unique speech queries in the test set was then defined as the audio-to-visual recall@10 ∈ [0, 1]. The process was repeated for visual-to-audio search using all the 5000 images as the queries and checking if at least one of the five correct captions were in the top 10. We report the mean of the two retrieval directions as the final recall@10.

Experimental Setup

As our baseline model, we train the model with $\alpha = 0.5$ for 50 epochs and starting from randomly initialized model parameters, which corresponds to equal emphasis on auditory and audiovisual learning throughout the process. In addition, we run the experiment for $\alpha = 0.1$ (strong emphasis on auditory learning) and $\alpha = 0.9$ (strong emphasis on audiovisual learning). We also report results for a scenario where the model is first trained for 20 epochs with $\alpha = 0$ (auditory learning only), followed by 50 epochs of equal weighing of the mechanisms ($\alpha = 0.5$), denoted as $\alpha = 0 \rightarrow 0.5$ -variant. This aims to simulate learning where auditory learning is predominant up to a certain stage, after which the learner starts to utilize visual information as well (e.g., due to improved head stability and motor skills). Finally, we test a model with equally weighted auditory and audiovisual pre-training (for 20 epochs), followed by 50 epochs of auditory learning ($\alpha = 0.5 \rightarrow 0$) to see if the model's representations can handle long-term absence of visual information. The epochs counts were determined based on saturation of recall@10 (or its pre-training benefits) in pilot experiments.

For all training, we use Adam optimizer with an initial learning rate of 10^{-4} with a warm-up period and then a linearly decaying learning rate schedule. The optimizer is always reset between pre-training and the final 50 epochs of training to ensure sufficient learning rate and Adam stability with a suddenly changing optimization space.

The models are probed for phonemic, lexical and semantic scores for each of the first 5 epochs, then every 10 epochs, and for the final 50th epoch. Phonemic evaluation is conducted with the temporal sequence of hidden layer activations at the 100 Hz sampling rate of the audio encoder and decoder (Fig. 1). Lexical evaluation uses the time-average of the hidden layer activations across the duration of the input word. Both phonemic and lexical evaluations are separately conducted for all layers of the audio encoder (N=8) and decoder (N=4). When reporting the phonemic and lexical learning curves, only the best scores are shown across the tested layers for each epoch. Semantic evaluation with recall@10 uses the high-level auditory and visual embeddings from their respective processing paths.

To facilitate comparison across the three linguistic levels of interest, plotted learning curves are normalized between 0 (chance level in the task) and 1 (the best performance across all the training variants and epochs), referred to as PHON, LEX, and SEM for phonemes, words, and semantics, respectively. The best non-normalized performance scores at the end of full training (epoch 50) are reported separately to indicate absolute quality of the three representation types.

Results

Table 1 shows the unnormalized performance of the model variants in the three original evaluation tasks at the end of full training. Note that the intermediate representation qualities of some variants can exceed the final performance (cf. Fig. 2).

All variants are successful in all the three tasks with substantially above-chance performance. ABX scores reach 5.06–6.90% phoneme discrimination error rate (chance 50%) and recall@10 reaches 0.422–0.798 depending on the variant, both measures being comparable to earlier studies focusing on the respective tasks (e.g., Chrupała, 2022; Peng & Harwath, 2022a, 2022b), especially considering that the previous ABX scores are typically reported for models trained on the same data as used by the ABX task (Librispeech corpus). The new lexical test results in a score of $q_{\text{LEX}} = 0.829$ for the best variant and 0.519 for the worst. Substantially above chance performance in all three categories provides a solid basis for closer analysis of the learning process as a function of language experience.

Table 1: Absolute performance levels obtained by each variant after full training. ABX = phoneme error rate, q_{LEX} = lexical quality score, recall@10 = audiovisual retrieval performance. α =X → Y denotes α during (X) and after (Y) pre-training. Numbers in parentheses denote the layer number responsible for the reported best performance.

Model variant	ABX (%) (layer)	q_{LEX} (layer)	recall@ 10
$\alpha = 0.5$	**5.06** (8)	0.519 (8)	0.750
$\alpha = 0.1$	5.15 (8)	0.713 (8)	0.759
$\alpha = 0.9$	**5.06 (8)**	0.647 (7)	0.744
$\alpha = 0.5 \rightarrow 0$	6.23 (8)	0.574 (8)	0.422
$\alpha = 0 \rightarrow 0.5$	6.90 (2)	**0.829** (6)	**0.798**

Concurrent Auditory and Audiovisual Learning

Moving on to the linguistic representation trajectories, Fig. 2. shows the results for the normalized PHON, LEX and SEM scores. The overall learning pattern is similar across all the different learning scenarios: phonemic representations are

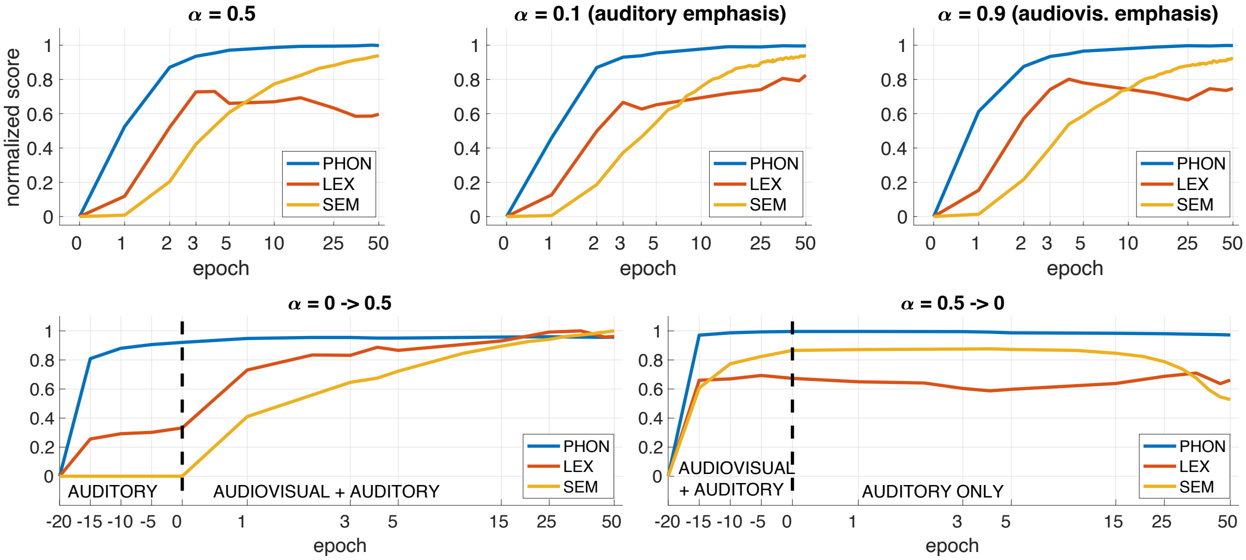

Figure 2: Normalized learning curves for phonemic (PHON), lexical (LEX) and semantic (SEM) representations for different learning scenarios. Normalized score of 0 corresponds to chance level and 1 corresponds to the best performance across all the epochs and training variants. Note that x-axis for epochs –20 to 0 are linear scale on the bottom row.

always emerging first, followed by lexical organization, whereas semantics always start improving after there is at least some amount of lexical organization. Yet, in all the cases where both auditory and audiovisual mechanisms have finite contributions to the process (Fig. 2, top row) there is always notable concurrency in the improvement of all three levels of representation. Also, whenever both mechanisms are present, the changes in α result only in minor numerical differences between the learning curves of PHON and SEM.

As for the LEX, it appears that a stronger emphasis on the auditory learning ($\alpha = 0.1$) leads to somewhat slower but more consistent over-time improvement of the lexical representations compared to other cases. In fact, $\alpha = 0.5$ and $\alpha = 0.9$ result in a peak of LEX at around epochs 3–4, after which LEX starts to degrade gradually even though PHON and SEM continue to improve. While the reason for this is unclear, it is possible that the audiovisual learning mechanism prioritizes encoding of vocabulary with visual referents at the cost of more general lexical encoding. For instance, some words of the CDI-Lextest, such as "*all gone*", "*I*", "*help*", or "*fast*", have little identifiable correspondence to the entities present in still visual scenes of SpokenCOCO. Hence, model overfitting to groundable lexemes may occur.

Results on Auditory to Audiovisual Learning

In the case of purely auditory learning (epochs –20 to 0 of $\alpha = 0 \rightarrow 0.5$ in bottom left panel of Fig. 2), there is a clear and relatively fast improvement in PHON, yet learning is still slower than in the scenario where concurrent audiovisual grounding is available (bottom right). LEX also gradually improves with purely auditory learning (see Lavechin et al., 2023, for a similar finding), but again at a slower rate than with audiovisual learning. Naturally, audiovisual semantics show no improvement with purely auditory learning. When the auditory learning becomes supported by audiovisual learning (epoch 0 onwards), the model reaches the best overall learning outcomes for LEX and SEM. PHON also becomes very good in absolute terms (cf. Table 1) but falls slightly short of the non-pretrained variants. In addition, learning of semantics is very rapid in this scenario compared to the models that do not have auditory pre-training.

The results suggest that general auditory statistical learning is beneficial in initializing speech representations that can then later support different levels of language structure. While phonemic discrimination is largely obtained already with purely auditory processing, the obtained auditory representations also support fast acquisition of high-quality lexical and semantic representations when access to visual information becomes available. The finding also aligns with the $\alpha = 0.1$ variant (Fig. 2, top center), where a stronger emphasis on auditory learning was found to improve the stability of lexical representations over time.

Results on Audiovisual to Auditory Learning

In the last learning scenario, audiovisual grounding is used to bootstrap the learning and is then followed by purely auditory learning ($\alpha = 0.5 \rightarrow 0$; Fig. 2, bottom right). The first 20 epochs of pre-training now correspond to the $\alpha = 0.5$ condition with the same findings as earlier (Fig. 2, top left). In this case, the good phonemic representations obtained with concurrent auditory and audiovisual learning stay as such when access to visual information is removed. In addition, the LEX of approx. 0.6 obtained by $\alpha = 0.5$ stage stays relatively constant during the following auditory learning stage but does not show any further improvements. Interestingly, quality of semantic representations obtained during audiovisual learning increases very slightly during the first epochs of purely auditory learning. After that, SEM stays stable until ~15 epochs, and finally start to degrade when visual input remains unavailable for too long.

The observed persistence of visually grounded semantic representations in the absence of further audiovisual input is atypical from the perspective of catastrophic forgetting in artificial neural networks. It appears that the auditory learning process does not aggressively alter the audiovisual representations when optimizing the network purely for the auditory learning task. Instead, the auditory and audiovisual representations are somewhat aligned across the tasks, but only when the network is initially trained for both tasks.

Discussion and Conclusions

This study set out to investigate how phonemes, words, and utterance-level semantics evolve as a function of language experience when the learning is driven by two mechanisms: 1) purely auditory learning of speech patterns and 2) associative learning between utterances and concurrent visual input. We studied the question with a computational model of speech representation learning using model training time as a proxy for language experience. The employed model was able to utilize regularities within the acoustic speech stream but also across perceptual modalities through visual grounding, both without strong priors on linguistic structure of speech or without any other strong constraints or biases (see Peng & Harwath, 2022b for the original description).

The results show that irrespectively of how auditory and audiovisual learning are weighted during the learning process, phonemic representations always precede lexical learning. In addition, improvements in lexical selectivity systematically precede learning of utterance-level semantics. At the same time, there is also substantial concurrency in learning the different levels of representation. Note that the general pattern *could* be interpreted as sequential acquisition, if a performance score threshold (e.g., above chance) would be used to determine a specific "age of acquisition" for the phonemic, lexical and semantic knowledge. However, such a sequential interpretation would be a simplification that hides many details of the acquisition dynamics, and hence would not reflect the interactions between levels of language. In fact, a related issue may exist in interpretation of behavioral research with infants, where the focus is often on finding above-chance behavioral differences between experimental conditions, which is then used to draw conclusions about the existence or lack of particular language capabilities.

With respect to bottom-up and top-down theories of language acquisition, the results from all testing conditions are *superficially* compatible with the acquisition patterns associated with "bottom-up" view in the sense that sensitivity towards phonemic structure appears before lexical and semantic knowledge. However, the results show how initially emerging phonetic discrimination capabilities can result from statistical learning operating on top of acoustic speech (see also, e.g., Lavechin et al., 2023), from audiovisual associative learning (see also Khorrami & Räsänen, 2021), or from both simultaneously. This disconnects the observable order of acquisition (phone(me)s → words → meanings) from the underlying learning mechanisms and learning targets driving the acquisition. Hence, *one should not equate the emergence of different language capabilities at different developmental milestones with a cascade of specific learning processes targeting at those capabilities.* Thereby, the present results are equally compatible with the usage-based accounts of language acquisition, where gradual emergence of lexical and sub-lexical structure results from holistic "top-down" meaning-centered learning (see Chrupała, 2022; Khorrami & Räsänen, 2021; or Merkx, 2022, for a recent discussion, models and references; cf. also Werker & Curtin, 2005).

The present results also suggest that learning of language patterns from auditory or audiovisual input is flexible in terms of the two types of learning applied, as linguistic structure emerges in all the tested learning scenarios. Visual input is highly useful whenever available, but the acquired visually grounded representations are tolerant against periods of absent visual information. There is some benefit from having a stronger early emphasis on auditory learning, latter followed by learning from sights. This may be due to a more generic nature of auditorily-learned representations compared to visual grounding, where only some words of the language have concrete visual referents. Hence, the groundwork laid by auditory learning can make learning from the more infrequent visual events more effective. This also fits to the language experiences of real infants, to whom communicative scenarios with closely paired visual referents are less frequent than access to speech signal in general.

Naturally, the present findings are subject to the simplifying assumptions and constraints of the present setup. For instance, the present speech data was not naturalistic speech and visual input to children (cf., e.g., Clerkin et al., 2017) but spoken descriptions of photographs from the MS-COCO database. When audiovisual learning was employed, the density of referentially relevant utterance-image pairs was also much higher than what would be expected in real world, even though there was still substantial referential ambiguity between the two modalities. On the other hand, the model lacks a mechanism for infant-caregiver joint attention that can shape the referential ambiguity substantially in the favor of the learner (e.g., Yu et al., 2021). The analogy between model training and "amount of language experience" was based on iterative processing of the same dataset instead of having a continuous stream of new language experiences. However, all the representations were tested on different data from the training set (with PHON and LEX on completely different corpora and SEM on a separate held-out fold), which means that the representations still generalize to novel data. Finally, the metrics used for the three levels of representation are only way to test learner's knowledge, and robustness of the results with respect to other ways to probe language skills should be verified.

Overall, the study provides new insights into the complementarity of auditory and audiovisual learning in early language acquisition, and how they might relate to the emergence of linguistic knowledge. To better understand the generality of the findings, the present experiments should be replicated with more naturalistic training data (in quality and quantity) and with alternative model architectures.

Acknowledgement

The research was supported by Academy of Finland grants no. 314602 and 345365, and by Kone Foundation grant L-SCALE. The authors would like to thank Marvin Lavechin for the useful discussions related to the topic and Puyuan Peng for his help with the model implementation. Code and instructions for replicating the experiments will be available at https://github.com/SPEECHCOG/VGS_order_CogSci.

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: a framework for self-supervised learning of speech representations. *In Proc. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*.

Chrupała, G. (2022). Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73, 673–707.

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 613–622.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B.,* 372: 20160055.

Dunbar, E., Hamilakis, N., & Dupoux, E. (2022). Self-supervised language learning from raw audio: lessons from the zero resource speech challenge series. *IEEE Journal of Special Topics in Signal processing*, 16(6), 1211–1226.

Feldman, N., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752–782.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2), 107–125.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3), 677–694.

Hsu, W-N., Harwath, D., Sog, C., & Glass, J. (2020). Text-free image-to-speech synthesis using learned segmental units. *Proc. Self-Supervised Learning for Speech and Audio Processing Workshop at NeurIPS-2020*, held as a virtual event.

Johnson, M., Demuth, K., Frank, M., & Jones, B. K. (2010). Synergies in learning words and their referents. *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pp. 1018–1026.

Karpathy, A., & Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 39(4), 664–676.

Khorrami, K., & Räsänen, O. (2021). Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? – A computational investigation. *Language Development Research*, 1(1), 123–191.

Kuhl, P. K., Conboy, B. W., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 979–1000.

Lavechin, M., de Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2023). Statistical learning bootstraps early language acquisition. *PsyArxiv preprint*: https://psyarxiv.com/rx94d/.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Proc. ECCV-2014*, Zurich, Switzerland, pp. 740–755.

Merkx, D. (2022). *Modelling multi-modal language learning: from sentence to words.* PhD thesis, Radboud University Nijmegen, MPI series in Psycholinguistics, Nijmegen: MPI.

Nikolaus, M., & Fourtassi, A. (2021). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. *Proc. Workshop on Cognitive Modeling and Computational Linguistics*, held as a virtual event, pp. 200–210.

Peng, P., & Harwath, D. (2022a). Fast-slow transformer for visually grounding speech. *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-2022)*, 23–27 May, Singapore.

Peng, P., & Harwath, D. (2022b). Self-supervised representation learning for speech using visual grounding and masked language modeling. *The 2nd Workshop on Self-Supervised Learning on Audio and Speech Processing (AAAI-SAS 2022)*, held as a virtual event.

Peng, P. & Harwath, D. (2022c). Word discovery in visually grounded, self-supervised speech models. *Proc. Interspeech-2022*, Incheon, South-Korea, pp. 2823–2827.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, 122(4), 792–829.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. *Proc. Interspeech-2013*, Lyon, France, pp. 1781–1785.

Smith, L. B. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1/2), 61–82.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint*: http://arxiv.org/abs/1807.03748

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1, 197–234.

Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the problem of referential uncertainty in early word learning. *PNAS*, 118(52), e2107019118.