Transformer-based Speech Model Learns Well as Infants and Encodes Abstractions through Exemplars in the Poverty of the Stimulus Environments

Yi Yang*, **Yiming Wang***, **Jiahong Yuan**
University of Science and Technology of China
{yanggnay, wangyiming}@mail.ustc.edu.cn, jiahongyuan@ustc.edu.cn

Abstract

Infants are capable of learning language, predominantly through speech and associations, in impoverished environments—a phenomenon known as *the Poverty of the Stimulus* (POS). Is this ability uniquely human, as an innate linguistic predisposition, or can it be empirically learned through potential linguistic structures from sparse and noisy exemplars? As an early exploratory work, we systematically designed a series of tasks, scenarios, and metrics to simulate the POS. We found that the emerging speech model `wav2vec2.0` with pretrained weights from an English corpus can learn well in noisy and sparse Mandarin environments. We then tested various hypotheses and observed three pieces of evidence for abstraction: label correction, categorical patterns, and clustering effects. We concluded that models can encode hierarchical linguistic abstractions through exemplars in the POS environments. We hope this work offers new insights into language acquisition from a speech perspective and inspires further research.

1 Introduction

Humans are excellent language learners. Many complex phenomena that the best linguistic theories today cannot thoroughly describe yet- coarticulation, grammar (Hardcastle and Hewlett, 1999)—are easily perceived and proficiently used by a 3-year-old infant (Lust, 2006; Tardif, 1993), without any formal study in Phonetics and Syntax. This drives linguists represented by Chomsky (1986, 2011) to form a strong belief that there exists an innate predisposition in the human brain dedicated to language processing. One key argument supporting the theory is *the Poverty of the Stimulus* (POS) (Chomsky, 1980, 2014), which claims that the linguistic input available to children is inadequate to account for the sophisticated language abilities they acquire according to empiricist theories (Pullum and Scholz, 2002), including sparse environments (Longa and Lorenzo, 2008) and misleading noise (Roberts, 2016). Meanwhile, many studies (Clark and Lappin, 2010; Legate and Yang, 2002; Cook, 1991; Piantadosi, 2023) have challenged this argument by arguing the existing advantages in infants' learning are sufficient to overcome POS environments (Warstadt and Bowman, 2022).

Recently, we have witnessed the great breakthrough of Neural Networks (Achiam et al., 2023) and challenged the longstanding linguistic assumption that only humans can understand language well (Chomsky, 2002). However, unlike infants learning languages in POS environments, machines are mostly trained on sufficient data of carefully annotated labels with huge memories. This observation naturally raises an insightful question, which could test the hypotheses and assess the robustness and generalization capabilities of neural networks under real severe scenarios:

Can models also perform as well as infants within the Poverty of the Stimulus Environment?

As our ***first contribution***, we found that the speech model can perform well in *the Poverty of the Stimulus* environments, supporting the empiricist theories and showing superior learning capabilities. Considering that infants primarily acquire language through speech and associative learning (Mattock, 2012), we selected phoneme and tone recognition as two fundamental and representative tasks and designed sparsity and noise scenarios, to simulate *the Poverty of the Stimulus* environments using the `wav2vec2.0` (Baevski et al., 2020) speech model. We found that by fine-tuning pretrained weights from an English corpus, the model can recognize phonemes with 93.81% accuracy and tones with 92.32% accuracy in Mandarin using only one label per utterance. Even with 90% of training labels being incorrect, the model can recognize phonemes with 93.41% accuracy. Contextual tone acquisition

*These authors contributed equally to this work.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7881–7890
January 19–24, 2025.

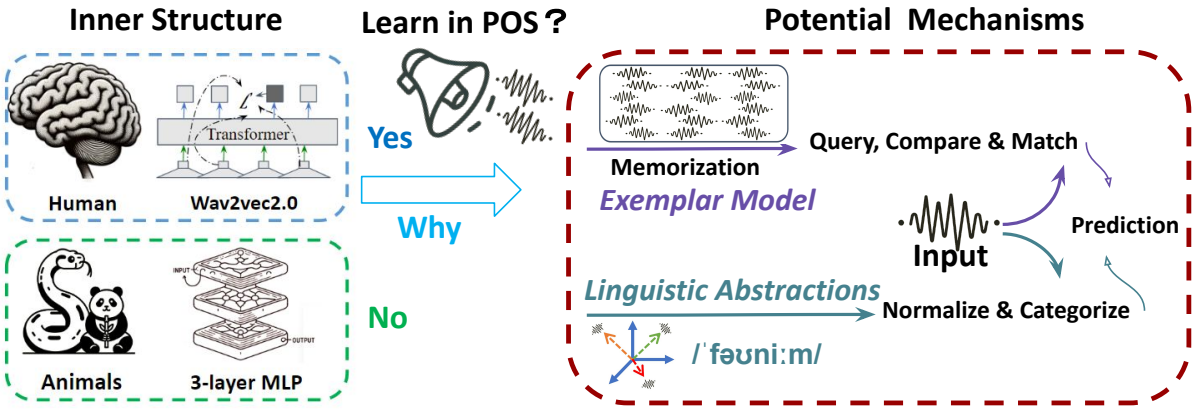

Figure 1: An Overview of inner structures and potential learning mechanisms in the Poverty of the Stimulus (POS) environments. **Left**: Human brains and `wav2vec2.0` models can learn in POS environments, while typical animals and basic 3-layer MLPs cannot. **Right**: Two potential perception mechanisms: Exemplar Model, which relies on stored instances and comparison, and Linguistic Abstractions.

experiments showed the impact of surroundings on model recognition is similar to human perception, known as carryover and anticipatory effects (Flege, 1988; Xu, 1993).

Based on these findings, we sought to offer new insights into the underlying mechanisms. Additionally, this can help determine whether observations are simply based on memorization (Bender and Koller, 2020) from limited datasets.

As our ***Second contribution***, we found the perception follows a mixture hypothesis instead of a dichotomy (Ambridge, 2020a): The model learns, corrects, and re-represents exemplars—speech and labels—into abstractions as parameters, moving beyond simple memorization. Firstly, with 90% of training labels being incorrect, the speech model not only performs well but correctly labels 93.82% of the training data compared to the original 10% correct training labels. Secondly, non-central frames showed high probabilities—92.14% for matching phoneme labels and 91.47% for tone labels, each with only one label per utterance. Additionally, the embeddings of non-central frames from different categories were separated and clustered around the same central frame clusters. The experiments also demonstrated that the speech model robustly encodes abstractions with noise levels up to a high threshold, beyond which performance rapidly deteriorates.

2 Background

Poverty of the Stimulus. *Poverty of the Stimulus* (POS) proposes that entities do not encounter sufficiently rich data within their environments to develop their capabilities, a phenomenon that has been found in many species. For example, bees could successfully navigate to food on their first exposure during heavily overcast mornings (Dyer and Dickinson, 1994; Gallistel, 2009), indicating an innate endowment (Berwick et al., 2011). Several studies have found that neural vision (Vong et al., 2024) or language models (Beguš et al., 2023a; Forster et al., 2018; Yedetore et al., 2023; Mahowald et al., 2024) can learn meaningful representations in human acquisition or POS environments. However, given that infants primarily acquire language through speech and associative learning, the study of simulating speech POS environments and their perception mechanisms is important yet less explored.

Exemplar Models. Exemplar Theory (Ambridge, 2020b) argues that unwitnessed forms are produced and understood through on-the-fly analogy across all exemplars. Under Exemplar Theory, hearing or saying "apple" involves no fixed phoneme retrieval. Instead, the brain constructs the word by referencing a multitude of stored exemplars—varied pronunciations of "apple" heard in the past and each exemplar carries phonemic components. Mahowald et al. (2020) noticed that neural networks cannot definitively preclude abstraction, raising an open question: *are modern neural networks truly examples of exemplar models?*

Categorical Perception: Liberman et al. (1957) proposed *Categorical Perception* by observing that humans simplify the variety and number of sounds into distinct phoneme categories, as a well-known pattern of speech abstractions. For example, when an infant hears the phrase 'apple', the continuous speech is not perceived uniformly but most frames

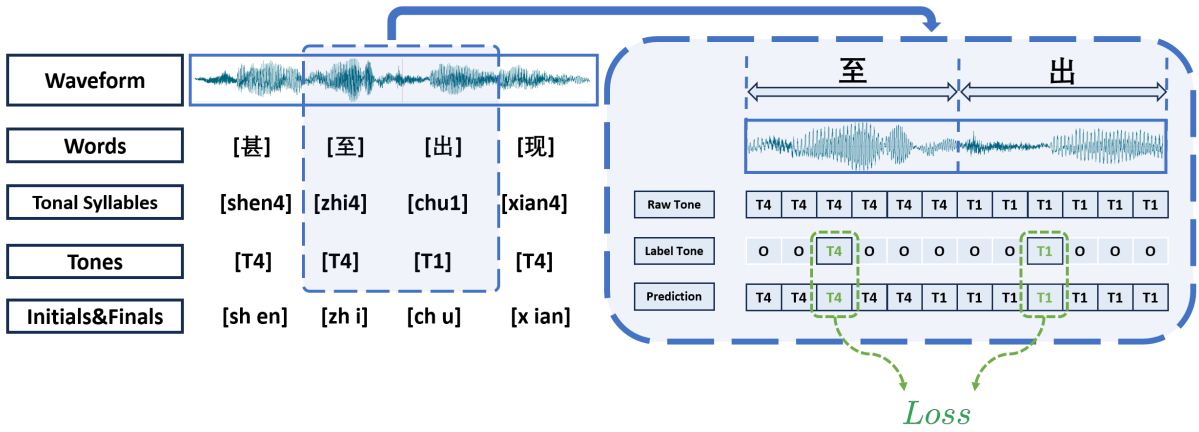

Figure 2: An overview of our task selection. Each frame is represented by a small square, and each syllable is indicated by a bold bracket, containing multiple frames. **Left**: A segment of Mandarin audio (waves) with corresponding words, tonal syllables, tones, and initials & finals. The tasks include tone classification (5 labels) and initials & finals classification (66 labels). **Right**: Using the tone classification task as an example, only the predictions for the middle frame of each syllable are used to compute loss and accuracy. The same approach is applied to the initials & finals classification task.

are segmented into distinct phonemes /æ/, /p/, and /l/ and the transitions are sharp (Hauser et al., 2002).

3 Methodology

3.1 Task Selection: Phoneme and Tone Recognition as Fundamental Language Acquisition Abilities

Humans can acquire language through speech and associative learning during infancy. During the perception, recognizing phonemes and pitch variations are basic and representative abilities. Phonemes, as the segmental elements, abstract continuous sounds into distinct units by filtering out non-essential details, a fundamental abstraction also observed in other mammals like sperm whales (Beguš et al., 2023b). Pitch variations, in contrast, enrich these units by conveying nuances such as emotions (Mozziconacci, 2002) and grammatical structures (Yip, 2002). In tonal languages like Mandarin, pitch variations are classified into specific categories known as tones, and different tones represent different meanings. Recognizing tones relies on context due to factors including tonal coarticulation (Hardcastle and Hewlett, 1999) and tone sandhi (Chen, 2000).

Therefore, we selected the two basic and representative tasks in experiments: tone classification (5 labels) and phoneme classification (66 labels). As illustrated on the left of Figure 2, each audio segment corresponds to a transcription of tones along with initials and finals (phonemes), forming a sequence of categories. During training, only the central frame of each tone or phoneme is labeled with a target, and all other frames are labeled as category '0', which is excluded from loss calculations. Assume we have N utterance-label pairs $(L_i, Y_i)_{i=1}^{N}$. We denote the label Y_i of the i-th utterance as $Y_i = (y_{i,1}, ..., y_{i,k_i})$ and the prediction as $\hat{Y}_i = (\hat{y}_{i,1}, ..., \hat{y}_{i,k_i})$, frame by frame. We adopt the loss function as Equation 1, following the study (Yuan et al.).

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{k_i} y_{i,c}\log(\hat{y}_{i,c}) \times I\{y_{i,c} \neq 0\} \quad (1)$$

3.2 Task Setting: Sparsity and Noisy Labels to Simulate POS environments

To simulate *the Poverty of the Stimulus* environments, we designed two scenarios: sparsity and noise, as illustrated in Figure 3. We also examined the contextual influence by adjusting the surroundings of the target, considering that tone recognition relies on context.

Sparsity Scenarios. In the sparsity scenarios, we quantified model performance by varying the degree of label sparsity, from complete labels to having only one label per utterance. This approach allowed us to systematically study how different levels of sparsity affect the model's recognition capabilities.

Noise Scenarios. In the noise scenarios, we examined the effects of label replacement by setting different replacement rates and adjusting the posi-

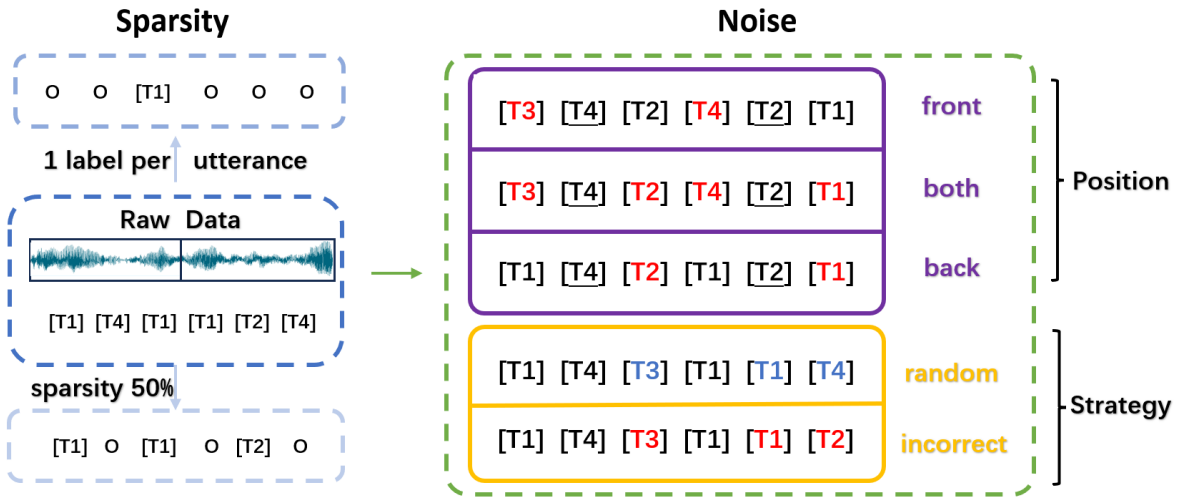

Figure 3: An overview of our task setting exemplified by tone classification. Each frame is represented by a small square, and each syllable is indicated by a bold bracket, containing multiple frames. **Left**: Sparsity experiment, showing each utterance with 50% of labels (bottom) and each utterance with only one label (top). **Right**: Noise experiment. For the position part, 'front' indicates randomly retaining a certain proportion of syllables and randomly selecting an incorrect label of the syllable to the front; 'back' indicates selecting an incorrect label of the syllable to the back; and 'both' indicates replacing the incorrect labels on both sides. For the strategy part, 'random' means randomly selecting from the overall labels (marked in blue), while 'incorrect' means randomly selecting from the remaining labels not belonging to the syllable (marked in red).

tion of randomly placed labels. We assessed how noise introduced at the front, back, and both positions of a syllable impacts recognition performance. Two types of noise conditions were tested: one involved randomly selecting labels from the overall labels, and the other involved selecting incorrect labels from the remaining labels not belonging to the syllable.

3.3 Task Evaluation: Probing Performance and Acquisition Theories

Due to the black-box nature of neural networks, it is challenging to understand their internal workings. To address this, we chose several metrics to evaluate the performance and acquisition theories.

Performance Evaluations. The effectiveness of the model and the impact of different POS environments can be directly evaluated by the accuracy variations on the test sets in various sparse and noisy simulation experiments.

Memorization Probes. In noise scenarios, we substitute a fraction of the true labels from the training set, denoted as $\mathcal{Y}_t$, with incorrect labels, referred to as the replaced labels $\mathcal{Y}_r$. Subsequently, we fine-tune the model using this modified training set. Then, we evaluate the degree of memorization by comparing the accuracy on the training set using the true labels $\mathcal{Y}_t$ and replace labels $\mathcal{Y}_r$ as the ground truth, respectively. If the model achieves high accuracy on the replace labels $\mathcal{Y}_r$ but low accuracy on true labels $\mathcal{Y}_t$, it suggests that the model tends to memorize samples. Conversely, it implies noisy environments contain learnable hierarchical linguistic rules, and models can correct replaced labels and generalize beyond mere memorization.

Categorical Patterns. We treat frames within a phoneme as different representations of the phoneme and compare their probabilities to test whether they were categorized through categorical perception or memorized as exemplars. First, we compute the average probabilities for the central and all frames of each tone or phoneme with the target label. If categorical patterns are present, both probabilities should be high. However, if the probability for all frames is significantly lower, it suggests continuous variation among phonemes, indicative of exemplar theory. Furthermore, we provide intuitive visualizations by plotting average probability changes from one syllable to the next. Categorical patterns will manifest as parallel lines with steep slopes corresponding to syllables, while exemplar theory would be represented by a nearly smooth, continuous line.

Clustering Effects. Considering the tone classification task has only five labels, we extracted the last layer features of frames from the test set

and conducted dimensionality reduction. If clear clustering patterns are observed and correspond to the target category, this further substantiates the model's abstraction capabilities.

4 Experiments Setup

Data. Our experiments utilized the Aishell-1 dataset (Bu et al., 2017), a commonly used resource for Mandarin automatic speech recognition (ASR). This dataset includes 165 hours of read speech along with word transcriptions in Mandarin Chinese, collected from 400 speakers. It is partitioned into training, validation, and testing sets, comprising 150 hours, 10 hours, and 5 hours of speech, respectively.

Forced Alignment. We performed forced alignment on the dataset, using an HMM-GMM based forced aligner trained with the HTK toolkit[*] and the pronouncing dictionary provided with the dataset. In the dictionary, words are transcribed into initials and tonal finals in Pinyin, the Romanized phonetic transcription system for Mandarin Chinese. From the results of forced alignment, a label was assigned to every frame in an utterance. The central frame within the boundary of a tone or phone received the corresponding label, while all other frames were labeled as 'O'. This 'O' label represents a special category that is excluded from cross-entropy loss computation during fine-tuning. The framerate was set at 20ms to align with the output framerate of `Wav2Vec2.0`.

Model. We finetuned the `Wav2Vec2.0` large model, which was pre-trained on 960 hours of Librispeech audio (`libri960_big.pt`), to perform experiments on tone and phoneme recognition. A fully connected layer was added on top of the `Wav2Vec2.0` model to convert the contextual representations into label tokens. Subsequently, the entire model is fine-tuned by minimizing a cross-entropy loss, with the 'O' category excluded from the loss calculation. The experiment was carried out using the Fairseq platform [†]. Initially, only the output classifier was trained for the first 10k updates, followed by fine-tuning of the entire network. The maximum token count was set to 1 million, equivalent to 62.5 seconds of audio sampled at 16 kHz, and the learning rate was 1e-5. The total number of updates was set to 100k. It took approximately 12 hours for a model to be trained on an A100 GPU. During inference, each audio utterance was fed forward through the fine-tuned model. The model's output at the central frame of every phone or tone was compared against its reference label to compute the classification accuracy.

[*]https://htk.eng.cam.ac.uk/

[†]https://github.com/facebookresearch/fairseq

5 The Speech Model Performs Well in POS Environments

In this section, we investigated whether the speech model could perform well in POS environments by examining the accuracy variations across various sparse and noisy simulation experiments. To simulate the sparse POS, we randomly selected a number of non-"O" labels (i.e., tone and phoneme labels at the center of the unit) from the label sequence and replaced them with "O" for each utterance, which is ignored in training. For noisy POS, we replace some labels in the training data with incorrect ones. For example, a label 'T1' was replaced with one randomly selected from other tones, or a label 'a' was replaced with one randomly selected from other phonemes. The same pre-trained `wav2vec2.0` model was fine-tuned using labels generated with these strategies and were then used to perform inference on the test set.

5.1 The Speech Model Performs Well in Sparse POS Environments

We set different proportions of labels in an utterance to be kept in the training set, without replacing them with "O": all labels, 75% of labels, 50% of labels, 25% of labels, and only 1 label was not replaced with "O". Specifically, tone and phoneme recognition utilized 7.1% and 3.5% of training labels with only one label per utterance.

Discussion (Left). The results in Fig 4-left show that the model performs well in sparse environments. As the proportion of labels per utterance used for training decreases, the model's performance only slightly lowers. Even with only one label per utterance for training, the models in sparse environments are quite robust, with performance decreasing by only 2.65% for phonemes (from 96.46% to 93.81%) and 3.63% for tones (from 95.95% to 92.32%).

5.2 The Speech Model Performs Well in Noisy POS Environments

To test the noisy environments, we choose different proportions in an utterance were replaced with another tone/phoneme label: 1 label per utterance,

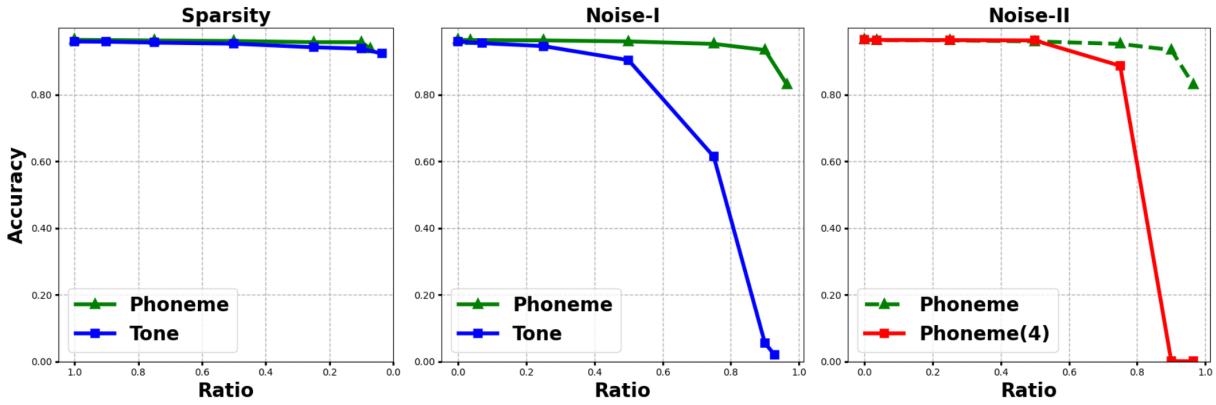

Figure 4: Phoneme and tone recognition accuracy on the test set in POS Environments. **Left**: Accuracy with All, 90%, 75%, 50% 25%, 10% labels and only 1 label per utterance. **Mid**: Accuracy with no, 1, 25%, 50%, 75%, 90% labels and all but 1 (1 Right) label per utterance replaced from incorrect labels. **Right**: Accuracy with no, 1, 25%, 50%, 75%, 90% labels and all but 1 (1 Right) label per utterance are incorrectly replaced and replaced uniformly from four fixed phoneme classes.

Task	`3 Correct`	`2 Correct`	`1 Correct`	`Front Random`	`Back Random`	`13 Random`	`13 Incorrect`
Acc (%)	94.09%	93.40%	92.32%	70.79%	72.61%	60.18%	44.86%

Table 1: Tone recognition accuracy on the test set under different surroundings.‘`k Correct`‘ indicates sequences with k consecutive correct tones per utterance. ‘`Front Random`‘ and ‘`Back Random`‘ describe scenarios where two consecutive tones have either the first or last label randomized and the other correct per utterance. ‘`13 Incorrect`‘ and ‘`13 Random`‘ signify 3 consecutive labels per utterance, where the first and third labels are incorrect and randomized respectively, with the middle label remaining correct.

25% of the labels, 50% of the labels, 75% of the labels, 90% of the labels, and finally only 1 label per utterance was not replaced with an incorrect one (1 Right).

Discussion (Mid). The experimental results on the test set in Fig 4-mid show that the model performs well in noisy environments. With 50% of the data incorrectly replaced, phoneme recognition achieves 95.98%, and tone recognition achieves 90.36%. Moreover, the model is more robust in recognizing phonemes than tones, with performance decreasing by only 0.74% for phonemes (from 95.98% to 95.24%) and 29% for tones (from 90.36% to 61.43%) when 75% of the data is incorrectly replaced. One key reason is that tones are contextual and influenced by factors like coarticulation (Hardcastle and Hewlett, 1999) and tone sandhi (Chen, 2000), making them harder to recognize than phonemes (McBride-Chang et al., 2008). It is worth noting that the tone recognition rate plummets to 5.54% under the 90% incorrect labels setting, signifying a collapse of the models due to incorrect attacks. This implies that the model might robustly handle biased noise up to a certain threshold, beyond which its performance rapidly deteriorates.

To further test this, we examined more biased phoneme labels, selecting incorrect labels uniformly from just four classes. The four classes were randomly chosen from all 66 classes and then fixed.

Discussion (Right). Results in Fig 4-right support the assumption: the model is robust but deteriorates beyond a certain attack threshold. The experimental results show that up to a certain limit, from 0% to 50%, the model performs well for both correct labels and incorrect labels replaced only from four specific phoneme classes. Even at the 50% level, the model with the Incorrect(4) setting outperforms. However, when the proportion increases further, the model's performance rapidly declines. In the 90% Incorrect(4) labels setting, the recognition rate drops sharply to 0.14%.

5.3 Contextual Tone Acquisition as Contextual Recognition

As Fig4-right shows, tone recognition is more affected by context compared to phonemes. Relying on this, we choose different noisy surroundings to test their influence of the speech model.

Discussion. As illustrated in Table 1, we found the surroundings matter. Both front and back labels

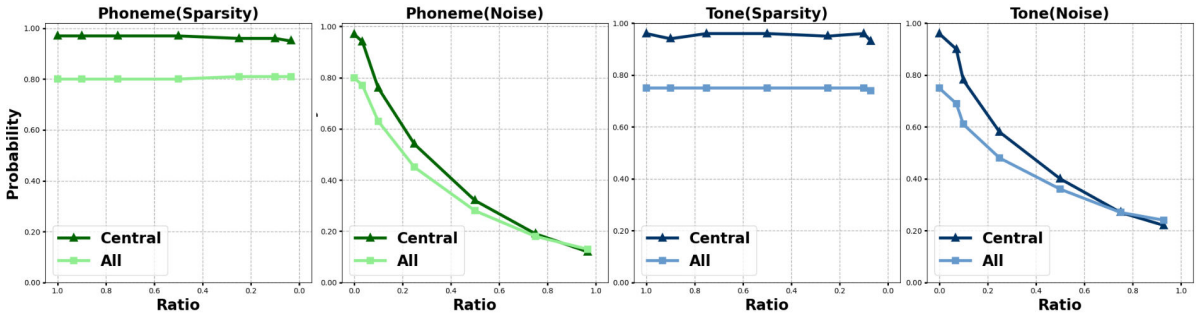

Figure 5: We compared the average probability values at the central frame with the corresponding average probability values within the boundaries produced by forced alignment for the given true label classes. This comparison was performed for two different tasks: phoneme (left) and tone (right).

significantly influence tone recognition, causing a performance decrease of more than 20%. This aligns with the carryover and anticipatory effects in tonal phonetics. The carryover effect appears to have a greater impact on tone recognition than the rear label, resulting in an additional 1.82% error rate, consistent with the findings described in the study by Xu (1993). An incorrect front tone label may lead to a shift in the preceding tone feature, affecting attention at the signal level and resulting in decreased accuracy. In addition, incorrect labeling, compared to random labeling, has a more severe impact on the language acquisition process, even though there is only a difference of one class. The accuracy drops by 33.51% (from 94.09% to 60.18%) with random labels, while incorrect labeling results in a decrease of 49.23% (from 94.09% to 44.86%).

6 Language Acquisition Theories in POS Learning Environments

From above, the models trained under different sparsity and noise conditions perform well on tone and phoneme recognition in POS environments. To explore the underlying acquisition theories, we conducted a series of experiments.

6.1 Beyond Memorization: Speech Model Corrects Noisy Training Labels

As illustrated in Fig 4-mid, the speech model performs consistently well across various noise levels. We further calculated the accuracy of the training set with replaced and true labels as the ground truth separately to test if the performance was simply based on memorization and analogy.

Discussion. The experimental results in Figure 6 show that the model does not solely depend on memorizing and comparing each specific exemplar, even though the model learns through examples. Instead, it efficiently encodes large amounts of noisy training data into meaningful abstractions represented by parameters. As the proportion of replaced labels increases, the accuracy with replaced labels significantly decreases. As noise levels increase, the accuracy associated with the true labels decreases slightly, whereas the accuracy of the replaced labels declines rapidly. This trend indicates that the model is highly robust in environments with substantial noise, indicating that the model is highly robust in such a noisy environment.

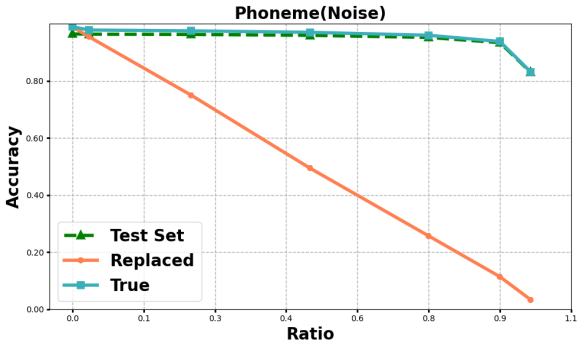

Figure 6: Phoneme recognition accuracy with replaced labels and true labels as ground truth on the training set, with no, 1, 1%, 25%, 50%, 75%, 90% labels and all but 1 (1 Right) label per utterance incorrectly replaced.

6.2 Speech Models Encode Abstractions in POS with Categorical Patterns

Then, we tested the categorical patterns of speech models in two aspects: (1) the average probabilities for central and all frames of each tone or phoneme with the target label, as shown in Figure 5, and (2) the transitions between categories of central frames, reflected by the decreasing probability of frames using the first central frame's label as the target and the increasing probability of frames using the

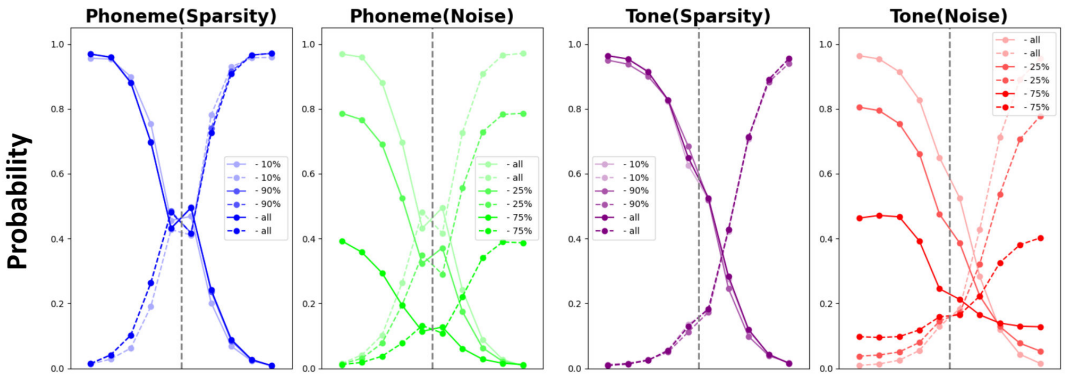

Figure 7: Phoneme and tone recognition probability changes between central frames. The gray line marks the boundary. The solid lines show the decreasing probability of frames using the first central frame's label as the target, while the dashed lines show the increasing probability of frames using the second central frame's label as the target.

second central frame's label as the target, as shown in Figure 7.

Discussion. As illustrated in Figure 5, the model maintains consistently high probabilities for all frames in both tone and phoneme recognition tasks under sparsity scenarios. However, in noise scenarios, the probability for central and all frames gradually declines, indicating a reduction in the model's abstraction capabilities. While the rate of probability decline slows down, the increasing rate of accuracy drop in noisy POS environments, as shown in Figure 4, further supports the notion that the model can robustly handle biased noise up to a certain threshold, beyond which its performance rapidly deteriorates. Similar patterns were observed in Figure 7, where noise affects abstractions more significantly compared to sparsity scenarios. Near the boundaries identified by force alignment, the probabilities drop rapidly. This suggests that there is a specific period where the model consistently groups tokens into a single category before abruptly switching to another.

6.3 Speech Models Encode Abstractions in POS with Clustering Effects

Additionally, we extract the features from the final layer of all frames under various conditions with Umap (McInnes et al., 2018) in Figure 8. The first row presents outputs with all labels, and the second row shows outputs with 25% and 75% noise.

Discussion. As illustrated in Figure 8, the model exhibits hierarchical abstractions in different POS environments. In both 25% sparsity and noise scenarios, the model shows clustering patterns similar to those using all labels, where frames are clearly divided into five clusters corresponding to the five

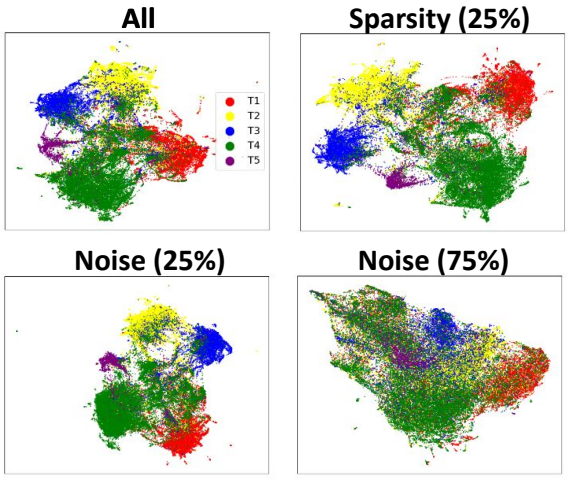

Figure 8: Tone recognition clusterings for both central and other frames with all labels, 25% labels(Sparsity), 25% incorrect, and 75% incorrect labels(Noise).

tone categories. However, in high-noise environments (75% noise), the model's performance deteriorates, with frames from different categories mixing, making recognition difficult.

7 Discussion and Conclusion

Through extensive speech experiments, we found that models can learn well in the Poverty of the Stimulus environments and encode hierarchical abstractions through exemplars, which implies that POS environments contain rich and learnable hierarchical linguistic rules. Further research is promising, especially when combined with findings of other modalities, such as syntactic abstractions (Leonard et al., 2024; Leong and Linzen, 2023), linguistic abilities (Beguš et al., 2023a), and so forth.

8 Limitation

Due to the workload and cost constrains, this paper focuses on phoneme and tone classification tasks using the model *wav2vec2.0* with pretrained weights from an English corpus exclusively. Further experiments involving additional tasks and models could be conducted. Additionally, we focus on the performance of speech models in the Poverty of the Stimulus environments and their perceptual mechanisms from a cognitive perspective. We do not delve into why machine learning develops such mechanisms, as this is another research direction requiring further exploration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ben Ambridge. 2020a. Abstractions made of exemplars or 'you're all right, and i've changed my mind': Response to commentators. *First Language*, 40(5-6):640–659.

Ben Ambridge. 2020b. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2023a. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*.

Gašper Beguš, Ronald L Sprouse, Andrej Leban, and Shane Gero. 2023b. Vowels and diphthongs in sperm whales.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Robert C Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive science*, 35(7):1207–1242.

Hui Bu, Jiayu Du, X. Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proceedings of O-COCOSDA 2017*.

Matthew Y Chen. 2000. *Tone sandhi: Patterns across Chinese dialects*, volume 92. Cambridge University Press.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. New York.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Noam Chomsky. 2011. *Current issues in linguistic theory*, volume 38. Walter de Gruyter.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Vivian J Cook. 1991. The poverty-of-the-stimulus argument and multicompetence. *Interlanguage studies bulletin (Utrecht)*, 7(2):103–117.

Fred C Dyer and Jeffrey A Dickinson. 1994. Development of sun compensation by honeybees: how partially experienced bees estimate the sun's course. *Proceedings of the National Academy of Sciences*, 91(10):4471–4474.

James Emil Flege. 1988. Anticipatory and carry-over nasal coarticulation in the speech of children and adults. *Journal of Speech, Language, and Hearing Research*, 31(4):525–536.

Dennis Forster, Abdul-Saboor Sheikh, and Jörg Lücke. 2018. Neural simpletrons: Learning in the limit of few labels with directed generative networks. *Neural computation*, 30(8):2113–2174.

Charles R Gallistel. 2009. The foundational abstractions. *Of minds and language*, pages 58–73.

William J Hardcastle and Nigel Hewlett. 1999. *Coarticulation: Theory, data and techniques*, volume 24. Cambridge University Press.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

Julie Anne Legate and Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):151–162.

Matthew K Leonard, Laura Gwilliams, Kristin K Sellers, Jason E Chung, Duo Xu, Gavin Mischler, Nima Mesgarani, Marleen Welkenhuysen, Barundeb Dutta, and Edward F Chang. 2024. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, 626(7999):593–602.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. *arXiv preprint arXiv:2306.05969*.

Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358.

Víctor M. Longa and Guillermo Lorenzo. 2008. What about a (really) minimalist theory of language acquisition? *Linguistics*, 46(3):541–570.

Barbara C Lust. 2006. *Child language: Acquisition and growth*. Cambridge University Press.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Kyle Mahowald, George Kachergis, and Michael C Frank. 2020. What counts as an exemplar model, anyway? a commentary on ambridge (2020). *First Language*, 40(5-6):608–611.

Karen Mattock. 2012. *Infant Language Learning*, pages 1542–1544. Springer US, Boston, MA.

Catherine McBride-Chang, Shelley Xiuli Tong, Hua Shu, Anita Wong, Ka-wai Leung, and Twila Tardif. 2008. Syllable, phoneme, and tone: Psycholinguistic units in early chinese and english word recognition. *Scientific Studies of Reading - SCI STUD READ*, 12:171–194.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Sylvie Mozziconacci. 2002. Prosody and emotions. In *Speech Prosody 2002, International Conference*.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50.

Ian Roberts. 2016. *The Oxford Handbook of Universal Grammar*. Oxford University Press.

Twila Zoé Tardif. 1993. *Adult-to-child speech and language acquisition in Mandarin Chinese*. Yale University.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Yi Xu. 1993. *Contextual tonal variation in Mandarin Chinese*. University of Connecticut.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*.

Moira Jean Winsland Yip. 2002. *Tone*. Cambridge University Press.

Jiahong Yuan, Xingyu Cai, and Kenneth Church. Improved contextualized speech representations for tonal analysis.