The Development of Spectral and Temporal Encodings in Speech Sounds

Frank Lihui Tan and **Youngah Do**
The University of Hong Kong
frank.lhtan@connect.hku.hk, youngah@hku.hk

Abstract

This study uses a modeling approach to explore the development of spectral and positional encodings in speech sounds. Humans rely on their auditory system to differentiate between individual sounds in words by analyzing both spectral properties of phonemes and their relative positions. Previous neuroscientific research has identified specific neural populations in the auditory cortex that respond to spectral processing, while behavioral studies have confirmed humans' ability to perceive the relative positions of phonemes in speech sequences. To investigate these encodings, a Long Short-Term Memory (LSTM) autoencoder with a cross-attention mechanism trained on Mel-spectrogram transformed from raw speech data is employed in this research. By conducting ABX tests on the model's representations at various learning stages, we observe the emergence of spectral and positional encodings. The results show that the model excels in distinguishing spectral features similar to neuroscientific findings, and also reveals independent positional encoding through accurate temporal distinctions. Furthermore, we illustrate the developmental trajectory of spectral and positional encodings during the learning process, proposing the need for further investigating their neural correlates.

1 Introduction

When humans process words like "task" or "cast," the auditory system is responsible for identifying individual sounds, e.g., [t], [æ], [s], and [k], as part of the word comprehension. To do so, the auditory system needs to analyze spectral properties of sounds, including their acoustic features (e.g., formants) as well as acoustic-invariant phonemic representations (e.g., [k] in initial and final positions). Previous neuroscientific studies have shown that specific neural populations are selectively responsive to the spectral processing of sounds in auditory cortex, especially the superior temporal gyrus (Chan et al., 2014; Chang et al., 2010; Gwilliams et al., 2022; Keshishian et al., 2023; Mesgarani et al., 2008, 2014; Shestakova et al., 2004). For example, Mesgarani et al. (2014) demonstrated that electrodes placed at different sites within the superior temporal gyrus, each corresponding to distinct neural populations, show selective responses to various phoneme categories, such as plosives, sibilants, vowels, and nasals. Chang et al. (2010) found that when listening to speech stimuli along a continuum, speech representations are categorical and spatially organized in posterior superior temporal gyrus. This organization involves distinct regions which correspond to specific phonetic categories, regardless of their acoustic variations. Not only do populations of neurons exhibit selective responses, but individual neurons have also been found to respond to particular phonemes or phonemic categories, such as vowels versus consonants, and even to their phonological patterns or statistical regularities (Lakretz et al., 2021; Leonard et al., 2024). The development of such spectral encodings of sounds begins early in human development, within the first year of life (Di Liberto et al., 2023). In parallel, computational modeling studies have demonstrated that neural networks trained on raw speech can spontaneously learn phoneme-like representations in their latent spaces (e.g. Baevski et al., 2020; Beguš, 2020).

In addition to understanding spectral characteristics of sounds, it is also crucial to recognize the relative positions of sounds, such as [t-æ-s-k] or [k-æ-s-t], in order to successfully comprehend words. Humans' auditory system is capable of positional encoding of sounds, allowing the detection of sound sequences with intervals as short as 15–20 ms (Hirsh, 1959; Steinschneider, 2005). Behavioral studies showed that humans can detect the relative positions of phonemes in speech streams (Benavides-Varela and Mehler, 2015; Endress and Mehler, 2010; Fló, 2021; Zhang et al.,

2003). For example, Fló (2021) conducted a sequence perception study in which listeners were trained on novel words. Although all syllables in the test words were familiar, listeners preferred words that maintained some syllables in their familiarized serial positions over words where all syllables had changed position, indicating that humans can track phonemes' relative positions. Similarly,
to develop emergent positional awareness, enabling them to perform counting-like and order-sensitive tasks without explicit positional encoding (Gers and Schmidhuber, 2001; Weiss et al., 2018).

While there is a range of evidence supporting the positional encoding of phonemes in humans, it is still unclear whether distinct neural encodings are specifically dedicated to marking the timestamps of phoneme sequences. This idea is akin to the spectral encoding found in Gwilliams et al. (2022), where specific neural encoding of spectral information of phonemes was identified. Neuroscientific studies have shown that neural populations respond differently to the same sound depending on its relative position within a sequence Leonard et al. (2015): for example, an electrode with a strong preference for the /n/ sound at the syllable-initial position showed no specific preference when /n/ appeared at the syllable-final position, suggesting that temporal context influences neural responses to speech sounds. However, there is currently no conclusive evidence for the existence of neural mechanisms that explicitly encode the positional information of phonemes independently of their spectral features. In the current study, the first goal is to investigate whether a positional encoding of phonemes, similar to spectral encoding, can be observed during speech perception. The second goal is to trace the learning trajectories by which learners acquire spectral and positional encodings of phonemes. As shown, previous research has provided diverse evidence supporting the existence of spectral and positional encodings of phonemes in human cognition, but the process by which such encodings are established in learners remains unclear, including whether similar patterns are observed in the development of both spectral and positional encodings. To our knowledge, no longitudinal data exist which directly compare the emergence and evolution of spectral and positional encodings. In this context, neural networks offer a promising framework for investigating the developmental trajectory of spectral and positional encoding in the acquisition of speech sounds and their results may help formulate hypotheses about analogous neural processes in human perceptual development.

The current study uses a modeling approach to replicate spectral and positional encodings observed in humans and investigate how these two different types of encodings evolve and interact over time. To do so, we simulate their learning process using a Long Short-Term Memory (LSTM) autoencoder equipped with cross-attention mechanism. The autoencoder is trained on Mel-spectrogram transformed from raw speech data, in order to examine the emergence and evolution of these encodings during language acquisition. We conduct ABX tests on the model's representations from various layers to assess the information captured in different stages of learning. Through this experiment, we expect to observe spectral encodings, consistent with previous neuroscientific findings. If the model successfully replicates humans' positional encoding abilities of phonemes, we also anticipate the emergence of independent positional encoding, which should be observed from the ABX distinction performance on positional contrasts, e.g., the same phoneme in different positions. In terms of developmental progression, we expect the model's ability to distinguish spectral and positional features to improve steadily across training epochs.

2 Methods

The model was trained to understand and reconstruct sound sequences accurately. Its ability to reconstruct the input was evaluated through a reconstruction task, measuring how well it replicated the original data. The model was expected to capture the sequential patterns effectively and accurately reproduce the input during the reconstruction process.

2.1 Model Architecture

An LSTM-based deep autoencoder model (Saini and Singh, 2024) was employed to learn auditory input by processing it sequentially without explicit guidance. The overall architecture follows the same design as that adopted in Tan and Do (2025). As there is no explicit feedback or learning guidance, the learning process is similar to how infants learn through passive listening. As in Figure 1, the model consists of an encoder, decoder, and cross-attention mechanism, all of which were designed to capture

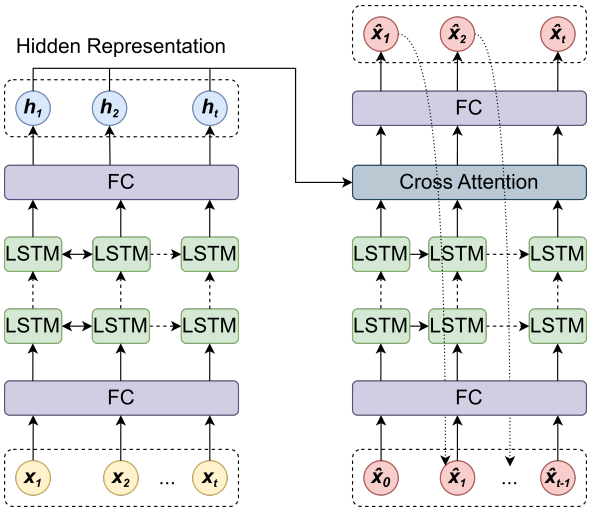

Figure 1: Illustration of model architecture.

spectral information as well as the temporal dependencies of the auditory sequences. The encoder and decoder were built with fully connected layers (FC) and LSTM layers. FC layers connect every neuron in one layer to every neuron in the next layer, enabling the learning of relationships between features in the data. LSTM layers capture sequential dependencies of the data. Cross-attention, implemented as scaled dot-product attention (Vaswani et al., 2017), was included to ensure effective information flow between the encoder and decoder. See Appendix A for details.

2.2 Dataset Construction and Preprocessing

The "train-clean-100" subset from the LibriSpeech corpus (Panayotov et al., 2015) was used, containing 100 hours of 16kHz English speech recorded by 251 speakers (125 female and 126 male). We relied on existing transcriptions (Lugosch et al., 2019) generated by the Montreal Forced Aligner (McAuliffe et al., 2017) to produce sub-word clusters from continuous recordings of sentences. We adopted a segment-based concatenation to construct the stimuli rather than extracting naturally occurring VCV sequences from continuous speech. Segments were extracted as individual phoneme realizations from the corpus and then artificially concatenated to construct VCV tokens following the templates VSV and VSHV, rather than being taken from naturally occurring continuous VCV sequences. The concatenation process ensured within-token speaker consistency such that each concatenated VCV token was constructed exclusively from segments produced by a single speaker. We selected from the English phoneme inventory six vowels (four monophthongs: AA, AE, IY, UW; and two diphthongs: AY, EY) and two consonants (S and SH). In total, this process yielded 66,792 V-S-V and around 25,238 V-SH-V sequences. We randomly selected a subset of 25,238 V-S-V sequences to match the size of the V-SH-V dataset. Mel-spectrograms were then extracted from the audio recordings and were standardized with mean and variance normalization.

2.3 Training

The model was trained using the Adam optimizer with a learning rate of 0.0005 (Kingma and Ba, 2014). To reduce the impact of random variation, each experiment was repeated five times independently, with the model trained for 100 epochs per run to ensure convergence. Performance was recorded at every epoch, including epoch 0, which represents the model's untrained state before any learning has taken place.

2.4 Layers and Evaluation Metrics

To evaluate the encoding capabilities of the model, we conducted ABX tests on outputs from all layers (Millet and Dunbar, 2020; Nguyen et al., 2020; Schatz et al., 2013, 2018). All representations for the tests were normalized using z-score normalization prior to distance computation. To reduce random variation, each ABX test was conducted six times, with 20 samples drawn from each phoneme class for every run.

All intermediate representations of the model were tested with the ABX test. The 96-dimensional Mel-spectrogram of the input signal was included as a baseline and labelled as *input-mel-spectrogram*. The final-layer outputs from the bidirectional encoder LSTM were sepatated into forward and backward representations and labelled as *encoder-LSTM-last-forward/backward-output*, respectively. The output of the final FC layer of the encoder was labelled as *encoder-output*. In the decoder, the representation produced after cross-attention was labelled as *cross-attention-output*. This layer-by-layer evaluation allows us to trace how phonological information is processed, refined, and abstracted as it propagates through the model during learning.

For each layer, we conducted two types of ABX tests: phoneme contrast test and phoneme position test. The phoneme contrast ABX test assessed the model's ability to distinguish between phoneme categories, e.g., /p/ vs. /k/ or /i/ vs. /u/. Good

performance in the phoneme contrast ABX test indicates that the corresponding layer is sensitive to phonemic or spectral contrasts.

The phoneme position ABX test, on the other hand, evaluated the model's ability to distinguish the same phoneme at different positions within the sequence (e.g., the first versus the third /i/ in the constructed VCV structure). Specifically, ABX triplets were constructed such that A and B corresponded to the same vowel category realized at two different positions in the VCV structure. Good performance in this test suggests that the layer is sensitive to positional information independent of spectral differences.

As a control test, the distinction between the same phoneme at the same position across different tokens was examined. For each training epoch, including epoch 0, we collected the model's outputs on the evaluation dataset and conducted the aforementioned ABX tests.

3 Results

We present our findings from two main aspects. First, we report model's ability to learn phonemic contrast (spectral encoding) and phoneme position contrast (positional encoding) and show how performance differs across layers. Next, we explore how the two encoding capabilities of different layers evolve over the course of training.

3.1 Division of Codes: Final Epoch Distinction Performances

To evaluate whether the model developed distinct spectral and positional encodings, the ABX test results from the final epochs were taken as the model's learning outcomes during the later stages of training. Specifically, ABX error rates were collected from epochs 95 to 100, as these represent the concluding phases of training, offering a reliable reflection of the model's final performance. The inclusion of five consecutive epochs helped reduce the impact of random fluctuations. While all hidden dimensions were assessed and showed similar trends (see OSF), a mid-range value of 16 was chosen for presentation purposes.

The results are depicted in five representations as shown in Figure 2: (a) Input Mel-Spectrogram illustrating the initial state, (b-c) two-directional last encoder outputs prior to forming the encoder output, (d) encoder output, which represents the final knowledge encapsulated by the encoder, and (e) cross-attention output representing a near-final representation in the decoder, which is used for reconstruction.

The input Mel-spectrogram (a) clearly distinguished between phoneme categories, with a low mean ABX error rate of 0.177, but struggled to encode phoneme position information, resulting in high error rates of 0.482. This implies that while the input data exhibited distinct phonemic contrasts (e.g., /i/ vs. /u/), positional distinctions between phonemes were less evident in the input (e.g., initial /i/ vs. final /i/). In the intermediate LSTM layers, the performance trends remained consistent with those of the input Mel-spectrogram, showing better accuracy in distinguishing phoneme category differences than phoneme position differences (see OSF for the details). Upon further data processing, the last LSTM layers (b-c) revealed a significant difference in how they each encoded spectral and positional elements within the forward and backward representations. Specifically, the forward output (b) excelled in capturing phoneme position details with a low error rate of 0.098 but struggled to differentiate phoneme categories, resulting in a high error rate of 0.476. Conversely, the backward output (c) demonstrated strong encoding of phoneme category distinctions with relatively low error rates (0.183) but encountered challenges in assessing phoneme position, showing high error rates (0.476). The two types of outputs, thus, demonstrated complementary distributions in encoding spectral and positional representations.

The encoder output (d), i.e., the learned outcome at the final stage of the encoding process, showed the lowest average error rates, indicating a successful learning outcome at the final stage of the encoder. This layer effectively integrated spectral and positional information, achieving moderate error rates for phoneme category differentiation (0.303) and low error rates for phoneme position distinction (0.105). Nevertheless, when cross-attention was applied to this encoder output in the decoding process, the representation in the decoder reverted to emphasizing spectral encoding, a different encoding type compared to the final encoder stage. The output from the cross-attention layer exhibited the lowest error rates in phoneme category classification among all layers (0.125) but faced challenges in preserving phoneme position information (0.485).

The ABX test results show a combination of spectral and positional encoding. Notably, layers within the model specialize in either spectral or

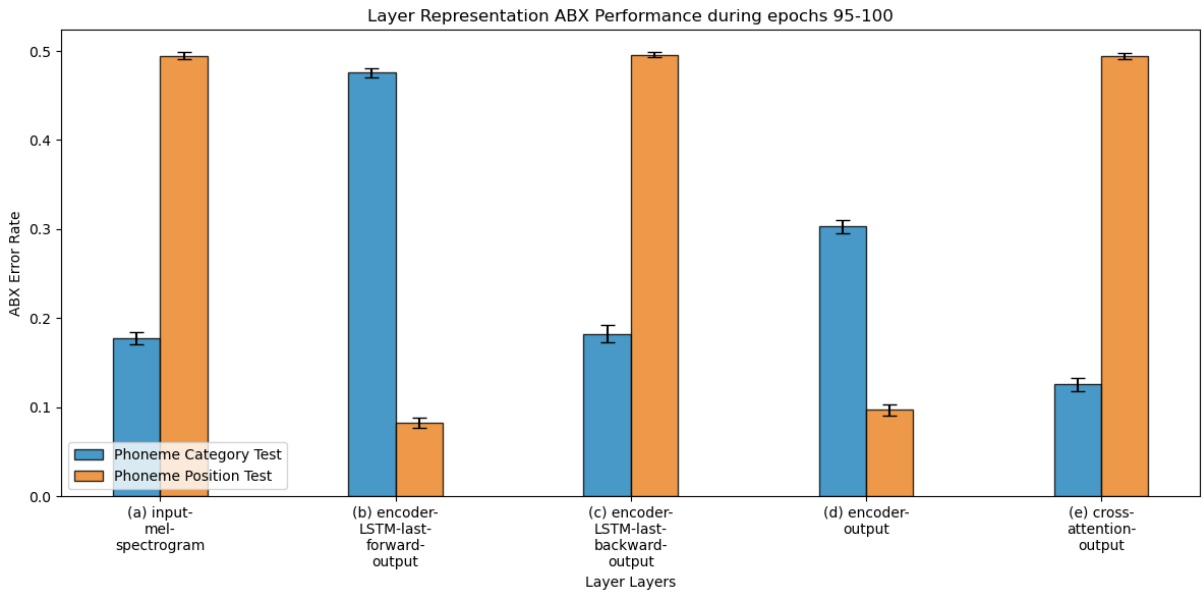

Figure 2: Layer representations' ABX test error rate on phoneme category test and phoneme position test for epochs 95-100.

positional information encoding to a certain extent. This relatively specialized division among layers could allow the model to optimize the reconstruction of input data by assigning specific encoding tasks to different layers. Specifically, successful spectral encoding in the input Mel-spectrogram is expected, as Mel-spectrogram features represent primarily spectral information; consequently, the same phoneme occurring in different positions yields highly similar representations, leading to weak positional distinctions in ABX comparisons. For the encoder output, it can also be inferred from its bridging role between the encoder and decoder that the encoder output encodes both spectral and positional information. The specialization of the cross-attention output in spectral encoding can be understood from its structural role—positioned before the final FC layer, it prioritizes spectral information to optimize reconstruction, mirroring the input. In contrast, the separate encoding of spectral and positional information in the final LSTM outputs, and their relation to forward and backward directions, is not easily explained solely by the model's structure. While the exact reasons for why and how each direction is linked to its specific encoding, i.e., forward output for spectral encoding and backward output for positional encoding, are not entirely predictable, it is evident that there is a clear division of labor based on their respective directions. We will revisit this topic in the Discussion section.

3.2 Model Development: Evaluation along Training Trajectory

The developmental trajectories of ABX error rates provide evidence of how spectral and positional information encoding in layers evolved during training. We examine how different layers progressed from an initial stage to their eventual specialization in encoding spectral or positional information.

As shown in Figure 3, the model initially struggled to distinguish between phoneme categories and phoneme positions, with middle to high ABX error rates across all layers. Following the earlier epochs, the L-shaped learning trajectories were found from some layers, both for phoneme contrast (solid lines) and phoneme position tests (dotted lines), with phoneme position tests showing more rapid learning. They are characterized by initial high ABX error rates, followed by a steady decline (comparison of ABX error rates between epoch 0 and the epoch with the lowest error rate, $p < 0.001$) as the model improves its ability to detect phoneme contrast or position, then stabilizes at a lower error rate with comparatively minor fluctuations. This pattern is primarily observed in layers that strongly encode either phoneme category or phoneme position in the final epochs, as described in the previous section, e.g., the encoder output. This pattern distribution suggests that each layer exhibits L-shaped learning trajectories for the specific type of information they specialize in encoding. For example, in the phoneme position test, the encoder LSTM last

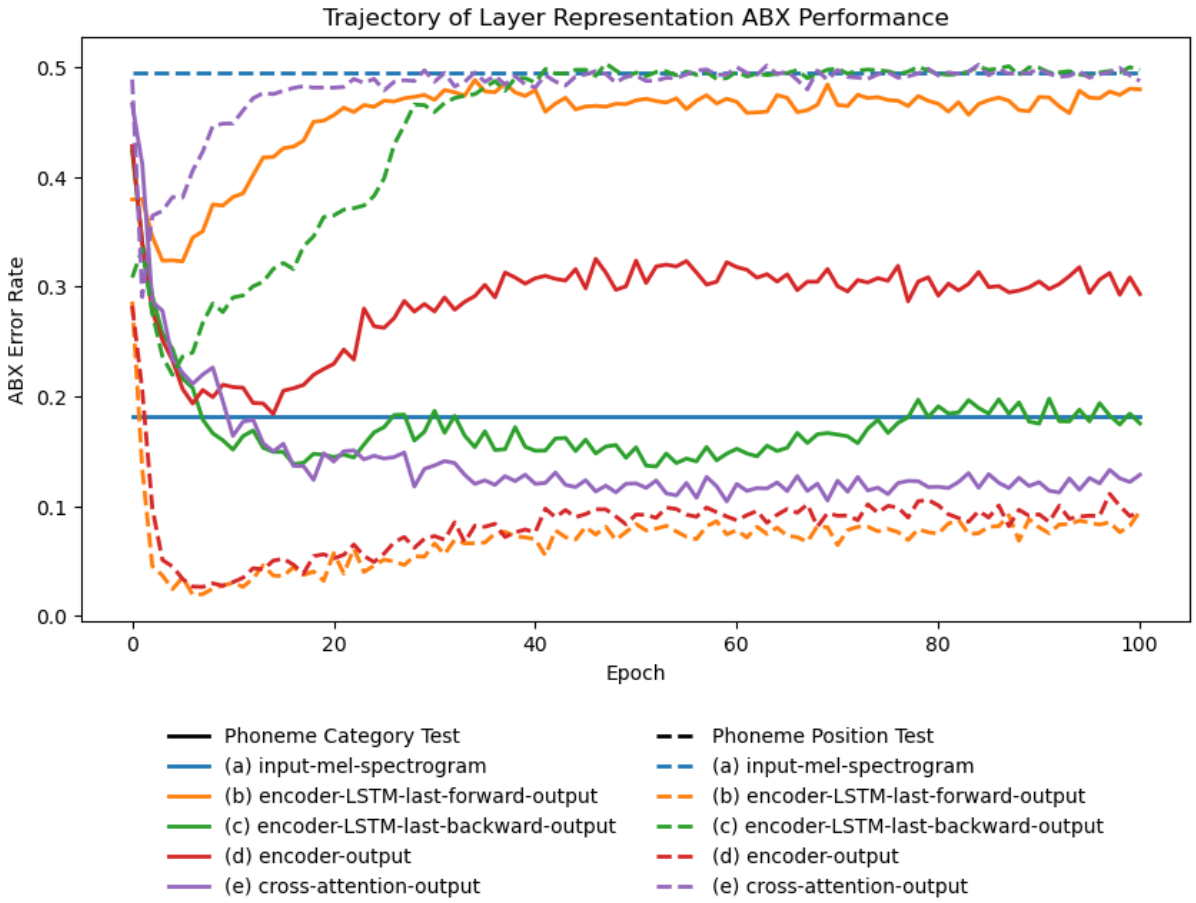

Figure 3: Trajectory of layer representations' ABX test error rate on phoneme category test (solid lines) and phoneme position test (dotted lines).

forward output, and the encoder output followed an L-shaped trajectory, demonstrating steady improvement over training. Similarly, in the phoneme category test, this pattern was observed consistently in the encoder LSTM last backward output and the cross-attention output.

However, not all learning trajectories followed a straightforward progression. As shown in Section 3.1, most layers ultimately specialize in encoding only one type of information, either spectral or positional. Accordingly, each layer typically exhibits only one L-shaped trajectory, corresponding to the information type it retains. The other trajectory does not remain flat at chance level; rather, it often shows an initial rise in performance before regressing. For instance, the encoder LSTM last backward output and the cross-attention output, which specialize in spectral encoding, initially displayed some positional encoding ability. Yet, unlike the trajectories that steadily improved in their specialized information, these non-specialized trajectories showed early progress followed by decline, resulting in the loss of the transiently encoded information. This phenomenon, evidenced by significant differences in ABX error rates between epoch 0, the final epoch, and the epoch with the lowest error rate ($p < 0.001$ for both comparisons), may suggest an "exploratory" learning strategy. Rather than having predefined roles from the outset, each layer underwent an adaptive process, eventually converging on the specific type of information it would specialize in encoding. These findings indicate that the specialization of layers observed in the final epoch did not emerge immediately but rather resulted from a gradual process of trial and refinement. Instead of encoding a single type of information from the start, many layers initially attempted to encode both spectral and positional details before settling into distinct roles.

4 Discussion

4.1 Summary of Main Findings

This study addressed two primary research questions: (1) whether independent positional encod-

ings exist to mark the timestamps of sound sequences separately from spectral representations and (2) how spectral and positional encodings develop over time in neural network learners newly introduced to language. As positional understanding is essential for perceiving phonemes in a sound stream (Chambers et al., 2003; Saffran et al., 1996), and behavioral studies have provided evidence supporting humans' ability for positional encoding of phonemes (Benavides-Varela and Mehler, 2015; Endress and Mehler, 2010; Fló, 2021; Zhang et al., 2003), we hypothesized that an independent positional encoding mechanism would be observed as a parallel to the spectral encoding observed by Gwilliams et al. (2022). We further anticipated a steady improvement in both spectral and positional encodings along with their separation across different layers of the model.

By evaluating the model's different layers on the phoneme contrast test (spectral knowledge) and the phoneme position test (positional knowledge), we found that the model successfully learned distinct encodings of phoneme categories (spectral knowledge) and phoneme positions (positional knowledge), with different layers specializing in each task. Moreover, the trajectory of different layers' outputs showed how such spectral and positional encodings evolved during the early stages of learning. We show that most layers exhibited expected learning trajectories showing a specialized type of information, either spectral or positional, and they steadily improved their specialized knowledge over time. However, as to the opposite information that a certain layer is not successfully encoding (e.g., spectral information for layers specialized in positional encoding and vice versa), instead of consistently displaying low distinction, these layers initially encoded both spectral and positional distinctions before gradually suppressing the encoding of the two types of information as training progressed. We elaborate each point below.

4.2 Phoneme Category and Position Encoding Specialization

The observed specialization of layers aligns with previous neuroscientific research demonstrating the existence of spectral encoding, i.e., position-invariant neural phoneme encoding (Gwilliams et al., 2022), as it supports a designated space for encoding the spectral information of phonemes. For example, the encoder LSTM last forward layer encoded the positional information of phonemes within a sequence while largely disregarding their spectral characteristics. This pattern suggests that positional encoding operates relatively independently from phonemic categorization. Literature on positional encoding suggests that its mechanisms are not restricted to language; Instead, positional encoding has also been identified in studies on non-linguistic sound sequences (Furl et al., 2011; Overath et al., 2007; Skerritt-Davis and Elhilali, 2018). For instance, Furl et al. (2011) observed that neural responses to temporal sequences of pure, non-linguistic tones were stronger when the sequence violated expectations based on prior training. Similarly, Overath et al. (2007) found that pure tone sequences with higher entropy led to increased neural activity in the planum temporale. Similar encoding of temporal patterns and sequence statistics extends beyond humans, as evidenced by animal auditory perception as well (Bouchard and Brainard, 2013; Schnupp et al., 2006), highlighting the broader significance of identifying temporal encoding of sounds. Additionally, behavioral and neural studies on object sequence recall in animals further support the existence of relative position sensitivity beyond human language and auditory perception (Ninokura et al., 2003, 2004; Orlov et al., 2000). Ninokura et al. (2004) identified neural populations in the dorsal lateral prefrontal cortex that selectively responded to relative position of objects, regardless of their physical properties. Such a diverse range of evidence beyond phonemes and human perception suggests that the positional encoding mechanisms identified in the current study have the potential to offer insights into the broader cognitive processes involved in auditory perception and temporal pattern recognition beyond language.

Moreover, Ninokura et al. (2004) discovered neural populations that encode objects' identity as well as relative position, with some neurons specialized in each property individually. This indicates that combined encoding arises through the interaction of separate yet complementary representations. Our findings on the encoder output, which merges spectral and positional information, align with those observed in object perception as in Ninokura et al. (2004). Instead of inherently encoding both spectral and positional properties from the start, the encoder output, representing the learned knowledge of the encoder, construct a joint representation by transforming independently learned spectral and positional encodings. This mechanism may represent a broader principle within the

human auditory system, where distinct neural populations process various aspects of speech before integrating their outputs at higher processing levels. This organization aligns with the structure proposed in the Phonological Loop Model (Burgess and Hitch, 1999), which posits that separate item property and order processing mechanisms contribute to sequence perception and recall.

4.3 Developmental Trajectories of Encoding Specialization

The analysis of the model's learning trajectory revealed patterns that align with previous research on phoneme acquisition. The L-shaped learning trajectory observed in spectral encoding—particularly in layers such as the encoder LSTM last backward layer and the cross-attention layer—indicates a progressive improvement in phonemic encoding during the early stages of training. This trend mirrors findings in human development, where phonemic representations become increasingly detailed and acoustically invariant over time. For instance, Di Liberto et al. (2023) demonstrated that infants in their first year of life gradually refine their phoneme encoding, developing more robust and stable phonemic categories. Similarly, behavioral and neural research supports the notion that infants' ability to distinguish phonemes in their native language steadily improves as they are exposed to linguistic input (Cheour et al., 1998; Kuhl et al., 2006; Rivera-Gaxiola et al., 2005; Sundara et al., 2006), similar to the L-shaped developmental trajectory in the present study. A similar developmental trajectory was observed in layers specializing in positional encoding as well, such as the encoder LSTM last forward layer and the encoder output layer.

Furthermore, our analysis revealed that positional encoding developed more rapidly than spectral encoding, reaching a stable level of proficiency earlier in training. This pattern also aligns with developmental findings in infants, where sensitivity to temporal structure emerges before robust phoneme discrimination. Studies have shown that newborns already exhibit positional awareness (Gervain et al., 2008), and by five months of age, the better learners among infants begin to show sensitivity to relative positions (Fló, 2021). This sensitivity continues to develop, with evidence of positional awareness at seven months (Benavides-Varela and Mehler, 2015) and nine months of age (Gerken, 2006). In contrast, while infants display an early universal sensitivity to phonemes, the ability to reliably distinguish phonemic contrasts in their native language develops more gradually and continues to refine over time (Cheour et al., 1998; Kuhl et al., 2006; Rivera-Gaxiola et al., 2005; Sundara et al., 2006). The developmental trajectory of the current model reflects a similar pattern, in which positional encoding emerges earlier, likely supporting later refinements in phoneme representation.

The learning trajectories of spectral and positional encoding indicate that different layers specialize in encoding distinct types of information, either spectral or positional. Despite the anticipation that layers would not encode information opposite to their specialization, our results unveil that many layers initially encode both types of information. For instance, the encoder LSTM last forward layer initially showed some capacity for encoding phonemic contrast, even though this layer was specialized for positional encoding. Similarly, the encoder LSTM last backward layer exhibited a certain degree of positional encoding at first, despite specializing in spectral encoding. This suggests that layers were not strictly selective for the type of information to encode innately but rather engaged in an exploratory phase where they encoded both spectral and positional information to some extent. However, as learning progressed, rather than continuously improving in encoding both types of information or simply maintaining the initial level of encoding for the non-specialized dimension, the layers actively suppressed the encoding of the non-specialized information. They restructured their representation space and de-learned certain contrasts. This trajectory suggests that specialization was not predefined but emerged dynamically from exploratory learning as the network optimized its representations for the reconstruction task. Instead of having rigidly assigned roles from the beginning, each layer initially explored how to encode both spectral and positional information before gradually refining its specialization.

4.4 Limitations

While our findings align with previous neural and behavioral studies, several limitations should be acknowledged. First, we presented modeling evidence supporting positional encoding and traced the model's developmental trajectories for both spectral and positional encoding. To our knowledge, no empirical neuroscientific evidence currently suggests the presence of distinct neurons or neural groups dedicated to positional encoding in

speech perception, nor are there longitudinal studies comparing the emergence and development of spectral and positional encoding in humans. Yet, the LSTM autoencoder used here is not intended as a mechanistic model of infant auditory cortex, and we do not assume that human learners use the same architecture, objective function, or optimization procedure. Rather, consistent the view that cognitive models are tools for exploring the implications of theoretical ideas under simplified and explicitly specified conditions (McClelland, 2009), the present model serves as a controlled artificial learner. In this sense, the contribution of the model does not depend on the claim that human infants learn in the same implementational manner as the network. Instead, its relevance lies in providing a computational demonstration: it tests whether a particular representational organization can emerge from the information structure of sequential speech under the pressure to preserve that information for reconstruction. The current results therefore show that a division of labor between phoneme-category information and relative-position information is a computationally plausible outcome for an artificial learner exposed to sequential speech input. Given that human learners are also sensitive to statistical regularities in auditory input (Saffran et al., 1996; Saffran and Kirkham, 2018), this modeling result offers a hypothesis for future empirical work: under similar information-structural constraints, human learners may also arrive at separable phoneme-category and relative-position representations through a comparable developmental trajectory. While neurophysiological evidence can inspire modeling studies, we advocate for reciprocal influence, where modeling data could also generate hypotheses and inspire neurophysiological research. We hope that our findings can stimulate neuroscientists and acquisitionists to investigate parallels between spectral encoding and positional encoding.

Second, we did not test alternative model architectures, such as Transformer-based models or convolutional networks. The emergence of positional encoding in our LSTM-based model may be inherently linked to the sequential nature of recurrent networks, which are explicitly designed for processing temporal dependencies. However, we argue that using an LSTM was the appropriate choice in the current study, given that sequential processing is a fundamental and early-acquired cognitive ability in humans. Research has shown that even newborns exhibit increased neural responses to reduplicated sequences (Gervain et al., 2008) and demonstrate sensitivity to ordinal positions by six months of age (Lewkowicz, 2013). Additionally, sequential learning plays a crucial role in linguistic processes, such as phonotactic pattern learning (c.f. Saffran and Kirkham, 2018), which infants begin acquiring as early as seven months (Thiessen and Saffran, 2003). Future work should explore whether similar spectral-positional encoding specializations emerge in non-recurrent architectures to determine if this effect is unique to LSTMs or a more general feature of neural network-based learning.

Finally, we want to note that the observed over-learning and subsequent de-learning trajectory of non-target information lacks empirical support from human neurophysiological studies. Nevertheless, our findings have shown that layers specializing in spectral or positional encoding followed learning trajectories similar to those observed in phoneme acquisition and positional learning research in the field. Therefore, we argue that the observed patterns in the current study warrant further neuroscientific investigation. A deeper understanding of how neural systems transition from broad, exploratory encoding to specialized representations may provide valuable insights into both human cognitive development and artificial learning architectures.

5 Conclusion

In conclusion, this study used an LSTM-based autoencoder to examine how spectral and temporal encodings can emerge and specialize during learning. By applying layer-wise ABX analyses, we showed that different components of the model come to preferentially encode phoneme category or phoneme position information, while higher-level representations integrate both. Yet, these encoding specializations were not present from the start but emerged gradually through learning.

Code Availability

Data, analysis scripts, and all results can be found at the following OSF link: https://osf.io/pja86/?view_only=3d182a56f7bb46e796149af313516bc2.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint*. Version Number: 3.

Gašper Beguš. 2020. Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning With Neural Networks. *Frontiers in Artificial Intelligence*, 3:44.

Silvia Benavides-Varela and Jacques Mehler. 2015. Verbal Positional Memory in 7-Month-Olds. *Child Development*, 86(1):209–223.

Kristofer E. Bouchard and Michael S. Brainard. 2013. Neural Encoding and Integration of Learned Probabilistic Sequences in Avian Sensory-Motor Circuitry. *The Journal of Neuroscience*, 33(45):17710–17723.

Neil Burgess and Graham J. Hitch. 1999. Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, 106(3):551–581.

Kyle E. Chambers, Kristine H. Onishi, and Cynthia Fisher. 2003. Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87(2):B69–B77.

Alexander M. Chan, Andrew R. Dykstra, Vinay Jayaram, Matthew K. Leonard, Katherine E. Travis, Brian Gygi, Janet M. Baker, Emad Eskandar, Leigh R. Hochberg, Eric Halgren, and Sydney S. Cash. 2014. Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus. *Cerebral Cortex*, 24(10):2679–2693.

Edward F. Chang, Jochem W. Rieger, Keith Johnson, Mitchel S. Berger, Nicholas M. Barbaro, and Robert T. Knight. 2010. Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13(11):1428–1432.

Marie Cheour, Rita Ceponiene, Anne Lehtokoski, Aavo Luuk, Jüri Allik, Kimmo Alho, and Risto Näätänen. 1998. Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience*, 1(5):351–353.

Giovanni M. Di Liberto, Adam Attaheri, Giorgia Cantisani, Richard B. Reilly, Áine Ní Choisdealbha, Sinead Rocha, Perrine Brusini, and Usha Goswami. 2023. Emergence of the cortical encoding of phonetic features in the first year of life. *Nature Communications*, 14(1):7789.

Ansgar D. Endress and Jacques Mehler. 2010. Perceptual constraints in phonotactic learning. *Journal of Experimental Psychology: Human Perception and Performance*, 36(1):235–250.

Ana Fló. 2021. Evidence of ordinal position encoding of sequences extracted from continuous speech. *Cognition*, 213:104646.

Nicholas Furl, Sukhbinder Kumar, Kai Alter, Simon Durrant, John Shawe-Taylor, and Timothy D. Griffiths. 2011. Neural prediction of higher-order auditory sequence statistics. *NeuroImage*, 54(3):2267–2277.

LouAnn Gerken. 2006. Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67–74.

F.A. Gers and E. Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340.

Judit Gervain, Francesco Macagno, Silvia Cogoi, Marcela Peña, and Jacques Mehler. 2008. The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):14222–14227.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. 2022. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, 13(1):6606.

Ira J. Hirsh. 1959. Auditory Perception of Temporal Order. *The Journal of the Acoustical Society of America*, 31(6):759–767.

Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, Anurag Kumar, Chin-Yun Yu, Chuang Zhu, Chunxi Liu, Jacob Kahn, Mirco Ravanelli, Peng Sun, Shinji Watanabe, Yangyang Shi, and 5 others. 2023. TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch. *arXiv preprint*. ArXiv:2310.17864.

Menoua Keshishian, Serdar Akkol, Jose Herrero, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. 2023. Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature Human Behaviour*, 7(5):740–753.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. Version Number: 9.

Patricia K. Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2).

Yair Lakretz, Ori Ossmy, Naama Friedmann, Roy Mukamel, and Itzhak Fried. 2021. Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation. *NeuroImage*, 226:117499.

Matthew K. Leonard, Kristofer E. Bouchard, Claire Tang, and Edward F. Chang. 2015. Dynamic Encoding of Speech Sequence Probability in Human

Temporal Cortex. *The Journal of Neuroscience*, 35(18):7203–7214.

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.

David J. Lewkowicz. 2013. Development of ordinal sequence perception in infancy. *Developmental Science*, 16(3):352–364.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-training for End-to-End Spoken Language Understanding.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

James L. McClelland. 2009. The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, 1(1):11–38.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. 2014. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174):1006–1010.

Nima Mesgarani, Stephen V. David, Jonathan B. Fritz, and Shihab A. Shamma. 2008. Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, 123(2):899–909.

Juliette Millet and Ewan Dunbar. 2020. The Perceptimatic English Benchmark for Speech Perception Models. *arXiv preprint*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint*.

Yoshihisa Ninokura, Hajime Mushiake, and Jun Tanji. 2003. Representation of the Temporal Order of Visual Objects in the Primate Lateral Prefrontal Cortex. *Journal of Neurophysiology*, 89(5):2868–2873.

Yoshihisa Ninokura, Hajime Mushiake, and Jun Tanji. 2004. Integration of Temporal Order and Object Information in the Monkey Lateral Prefrontal Cortex. *Journal of Neurophysiology*, 91(1):555–560.

Tanya Orlov, Volodya Yakovlev, Shaul Hochstein, and Ehud Zohary. 2000. Macaque monkeys categorize images by their ordinal number. *Nature*, 404(6773):77–80.

Tobias Overath, Rhodri Cusack, Sukhbinder Kumar, Katharina Von Kriegstein, Jason D Warren, Manon Grube, Robert P Carlyon, and Timothy D Griffiths. 2007. An Information Theoretic Characterisation of Auditory Encoding. *PLoS Biology*, 5(11):e288.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Queensland, Australia. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Maritza Rivera-Gaxiola, Juan Silva-Pereyra, and Patricia K. Kuhl. 2005. Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. *Developmental Science*, 8(2):162–172.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928.

Jenny R. Saffran and Natasha Z. Kirkham. 2018. Infant Statistical Learning. *Annual Review of Psychology*, 69(1):181–203.

Kapil Saini and Ajmer Singh. 2024. A content-based recommender system using stacked LSTM and an attention-based autoencoder. *Measurement: Sensors*, 31:100975.

Thomas Schatz, Francis Bach, and Emmanuel Dupoux. 2018. Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *The Journal of the Acoustical Society of America*, 143(5):EL372–EL378.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proceedings of INTERSPEECH 2013*, pages 1–5, Lyon, France.

Jan W. H. Schnupp, Thomas M. Hall, Rory F. Kokelaar, and Bashir Ahmed. 2006. Plasticity of Temporal Pattern Codes for Vocalization Stimuli in Primary Auditory Cortex. *The Journal of Neuroscience*, 26(18):4785–4795.

Anna Shestakova, Elvira Brattico, Alexei Soloviev, Vasily Klucharev, and Minna Huotilainen. 2004. Orderly cortical representation of vowel categories presented by multiple exemplars. *Brain Research. Cognitive Brain Research*, 21(3):342–350.

Benjamin Skerritt-Davis and Mounya Elhilali. 2018. Detecting change in stochastic sound sequences. *PLOS Computational Biology*, 14(5):e1006162.

M. Steinschneider. 2005. Intracortical Responses in Human and Monkey Primary Auditory Cortex Support a Temporal Processing Mechanism for Encoding of the Voice Onset Time Phonetic Parameter. *Cerebral Cortex*, 15(2):170–186.

Megha Sundara, Linda Polka, and Fred Genesee. 2006. Language-experience facilitates discrimination of /d-/ in monolingual and bilingual acquisition of English. *Cognition*, 100(2):369–388.

Frank Lihui Tan and Youngah Do. 2025. Attention-LSTM autoencoder simulation for phonotactic learning from raw audio input. *Linguistics Vanguard*.

Erik D. Thiessen and Jenny R. Saffran. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39(4):706–716.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the Practical Computational Power of Finite Precision RNNs for Language Recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745, Melbourne, Australia. Association for Computational Linguistics.

Da Ren Zhang, Zhi Hao Li, Xiang Chuan Chen, Zhao Xin Wang, Xiao Chu Zhang, Xiao Mei Meng, Sheng He, and Xiao Ping Hu. 2003. Functional comparison of primacy, middle and recency retrieval in human auditory short-term memory: an event-related fMRI study. *Cognitive Brain Research*, 16(1):91–98.

A Model Architecture Details

A.1 Encoder

The encoder was designed to extract latent information from the input, capturing essential features for further processing. To facilitate modeling, the original audio data was transformed into the Mel-Spectrogram features, which captures how different frequencies of sound change over time. The Mel-spectrogram features were structured in the shape of (B,T,F), where B is the batch size (set to 128), T is the sequence length, and F is 96 which corresponds to the number of Mel filter banks, defining the feature dimension for each frame. Zero-padding was applied to shorter sequences within the batch to maintain uniform sequence lengths. The input Mel-spectrogram was first passed through an FC layer, transforming the feature dimension from 96 to the target hidden dimension, reflecting differing memory capacities. The transformed features were then processed sequentially through five bidirectional LSTM layers, each integrating information from both past and future timesteps. This allowed the model to capture dependencies and patterns in the input data, enhancing its understanding of the sequential information. The output from the final LSTM layer was passed through another FC layer to integrate the concatenated bi-directional representation into a final output of the target dimension. This structure allowed the encoder to capture spectral as well as temporal patterns and generate a compact yet abstract representation of the input sequence, showing how the model learned the input data.

A.2 Hidden Representation

Each layer's output contained information at varying levels of abstraction, with deeper layers expected to capture more abstract and generalized features. The output of the encoder, i.e., the deepest hidden representation, encapsulated the knowledge learned by the encoder. This hidden representation formed the foundation for the decoder's reconstruction process, enabling the generation of output sequences. To further evaluate the influence of hidden dimensionality on the model's ability to learn phoneme contrast from contextual input, we experimented with six different hidden dimension settings: 4, 8, 16, 32, 48, and 64. Increasing the size of hidden dimensions aimed to mimic higher memory capacities.

A.3 Decoder and Cross-Attention

The aim of the decoder was to accurately reconstruct the input data using the learned information extracted by the encoder. It received the hidden representation the encoder generated as input and processed it autoregressively, meaning that each timestep's input was the output of the previous timestep. As shown in Figure 1, the decoder consists of an FC layer, five unidirectional LSTM layers, the cross-attention layer, and a final FC layer to generate the output. The decoding process began with a starting token, initialized to zeros, which served as the input for the first timestep. For each subsequent timestep, the decoder's LSTM layers incorporated historical information and made a prediction for reconstructing the next timestep. The scaled dot-product cross-attention mechanism

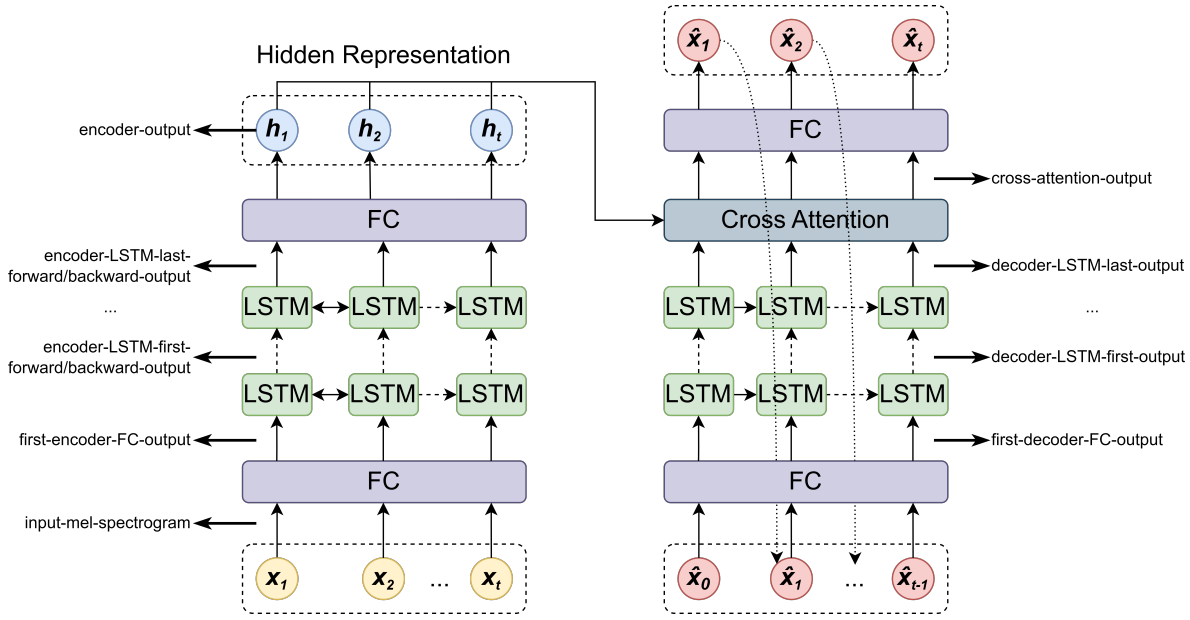

Figure 4: Illustration of layer output names.

(Vaswani et al., 2017) played a crucial role in bridging the encoder and decode. It allowed the decoder to query relevant information from the hidden representation. Specifically, the decoder's LSTM output served as the query, i.e., a representation of the current output that is used to retrieve relevant information, while the hidden representation acted as the keys and values, which helped the model to focus on important aspects of the input data. The cross-attention mechanism captured the similarity between the query and keys, guiding the decoder's reconstruction for the current timestep. After the attention layer, the data underwent a final FC transformation, shaping the output to match the input dimensions. This autoregressive output was then used as input for the next decoding timestep until the entire sequence was processed.

A.4 On Bi-directional LSTM

The use of bi-directional LSTMs was chosen to enhance the model's ability to integrate contextual information more effectively; this does not introduce additional non-humanlike characteristics, as the decoder's cross-attention mechanism will already incorporate global information.

A.5 Loss

This output was assessed against the input using the masked mean squared error (MSE) loss function (Paszke et al., 2019) to measure reconstruction accuracy.

B Dataset and Preprocessing Details

B.1 Concatenation

This approach allowed precise control over phoneme identity and phonotactic structure while minimizing systematic dependencies on particular lexical items or local contextual realizations. Segment selection was not constrained by positional or prosodic factors (e.g., word boundary, phrase position, or local coarticulatory context) but were instead sampled across a wide range of naturally occurring environments. This design was intended to avoid systematic coupling between specific positional contexts and acoustic realizations, thereby reducing the influence of strong but non-target position-acoustic correlations that are inherent in continuous speech.

B.2 Preprocessing

Mel-spectrograms were extracted from the audio recordings using the transforms.MelSpectrogram function from the torchaudio library (Hwang et al., 2023). This function transforms the input signal into a sequence of spectral samples through the application of filter banks for ease of processing. The n_fft and window length parameters were set to 512, and the hop length to 128. Using 96 Mel filter banks, the process produced windows of 32 ms with an 8 ms shift between consecutive windows. The resulting Mel-spectrogram had dimensions of (wave frame // 128, 96), where "wave frame" represents the total number of waveform frames. Finally,

mean and variance normalization was applied to the Mel-spectrogram to standardize input features and enhance consistency during processing.

C Evaluation Layer Outputs Visualization

Figure 4 displays the naming of layer output representations in more detail.

D Model Reconstruction Results Performances

We evaluated the model's performance on its primary training objective, namely reconstruction. Reconstruction quality visualization at the end of training can be found in the OSF link. Although the reconstruction lacked finer acoustic details, the model successfully captured the coarse energy distributions of sounds, which are crucial for their identification. The mean squared error (MSE) loss on the test set over 100 training epochs also demonstrated consistent improvement in reconstruction quality, reducing the MSE loss from 0.677 ($\sigma = 0.0459$) at epoch 0 to 0.179 ($\sigma = 0.0118$) by the end of training. This reconstruction performance confirms that the model was learning effectively, providing a reliable foundation for subsequent evaluations.