Learning Phonological Features from Bottom-up Data using an Autoencoder

Abstract

This study explores the effectiveness of an autoencoder model in learning phonemes and distinctive features from unsegmented, non-transcribed wave data, resembling infants' language acquisition stages. The experiment on Mandarin and English reveals that features can be learned through repeated projection and reconstruction without prior segmentation knowledge. The model clusters segments of the same phoneme and projects different phonemes to separate regions in the hidden space. However, the model struggles to cluster allophones closely, indicating the boundary between bottom-up and top-down information in phonological learning. The study suggests that sound knowledge can be learned to some extent through unsupervised learning without labeled data or prior segmentation knowledge, providing insights into early human language acquisition stages.

Keywords: phonological acquisition, autoencoder, bottom-up learning, features, phonemes

1 Introduction

The acquisition of phonological knowledge is largely based on the continuous and gradient acoustic cues of sounds. One important aspect of this acquisition is the learning of phonological features, which are abstract and categorical, such as [±voice] or [±sonorant] (Kolachina & Magyar, 2019). While the acquisition of phonological features can occur even in the absence of specific phonetic cues (Warren, 1970), it poses a significant learning challenge to infants: During the early stages of phonological acquisition, infants do not possess knowledge of the distributional properties of sounds, such as their co-occurrence or complementarity, and are unable to incorporate lexical information into their learning process (Hayes, 2004). Consequently, the phonetic cues used to distinguish phonemes are primarily, if not exclusively, relied upon by infants, and they are acquired through "bottom-up" learning processes (Shain & Elsner, 2019), which rely on the most fundamental properties of sounds, such as formant patterns and pitch. As phonological acquisition

progresses, the distributional properties of sounds and lexical information can also be incorporated into phonological knowledge in a "top-down" manner. This approach involves integrating cues from higher levels, such as the syntactic and semantic levels, into the phonological learning process.

In the field of language modeling, top-down models have been proposed as an approach that relies on the distributional properties of phonemes. These models are typically trained on textual data: orthographies or International Phonetic Alphabet (IPA) transcriptions , as shown in studies conducted by Kolachina & Magyar (2019), Silfverberg et al. (2018), and Sofroniev & Çöltekin (2018). A core underlying assumption of these top-down models is that phonemes with similar distributional properties are more likely to have similar features. One such widely used model is the word2vec model, which projects word tokens into a semantic space based on their distributional similarity (Mikolov et al., 2013). Silfverberg et al. (2018) conducted an experiment that applied the word2vec model to phone embedding and found that, after training on text corpora, it could infer relations among phonemes based on their featural relations. For instance, given that *x* is /p/, *y* is /b/, and *a* is /t/, the model could predict that *b* is /d/, because the difference between the given pair of phonemes and the predicted pair is equally voicing. Although the word2vec model could successfully capture distributional properties of sounds and the learned clustering was found to be correlated with distinctive features, it does not incorporate phonetic properties of sounds, as this model's phonological learning relies solely on top-down information. Nonetheless, the model was found to capture cross-linguistic phonological differences, such as the clearer vowel clustering observed in languages with vowel harmony, compared to those without (Kolachina & Magyar, 2019).

However, for the acquisition of phonological features, top-down models may not accurately reflect infants' learning processes. A primary reason for this is that the relationship between phonemes and their distinctive features is not arbitrary, since the features correspond to the underlying physiological and phonetic properties of the phonemes, i.e., the bottom-layers of phonemes. Therefore, a bottom-up approach that incorporates the sounds' underlying physiological properties may be more appropriate for modeling the early stages of phonological feature learning. While mixed models that incorporate both top-down and bottom-up information may be more effective, they are often unable to clearly distinguish the relative contributions of each type of information (Shain and Elsner, 2019). Kamper et al. (2015) developed a model that

combined weak top-down information with a bottom-up approach by using phone segmentation data, but this approach made it difficult to determine the specific contributions of each type of learning. Furthermore, providing top-down information early in the learning process assumes prior knowledge of segment boundaries, which may not be present in infant learners who have limited explicit cues about these boundaries (Räsänen, 2014).

Recent research has used bottom-up models for learning distinctive features based on phonetic information. Although "pure" bottom-up models may have limitations as they completely ignore distributional and lexical cues, they may provide direct insights into the exact roles of phonetic cues in phonetic and phonological learning. An early version of a bottom-up model was proposed by Guenther & Gjaja (1996), where the model clusters vowels based on selected formants (F1, F2, F3). The trained model confirmed the perceptual magnetic phenomenon, as it successfully categorized scattered sounds more similar to the closest phoneme centers than they really are based on pure acoustic space. However, the input features, F1, F2, and F3, were pre-selected in a biased manner in favor of clear vowel clustering, and the model may not be applicable to other sounds, such as some consonants, and may not reflect a realistic setting of human learning, where no "pure" formant information is given at any explicit level.

A more recent experiment conducted by Shain and Elsner (2019) utilized powerful bottom-up learning models to extract phonological features automatically. The experiment focused on phonetic information rather than distributional information, and the model was trained on short, discrete frames of audio instead of longer recording sequences; thus, the boundary information (phoneme boundaries, word boundaries etc.) was minimized to learning. The autoencoder structure, which is a variant of encoder-decoder structure, was used in the experiment, where the input and output data were the same (see section 2.3 for the details). The model's encoder converts the input into a compressed, latent representation that encapsulates the input information and typically has lower dimensions than the original input. This process bears resemblance to how human learners abstract phonological information from phonetic cues. The decoder then utilizes this latent representation to generate the output, relying on the established phonological grammar rules. In essence, the model's output production process is akin to that of humans, who use phonological cues to construct language. By incorporating the autoencoder structure, the model was trained to reconstruct the audio input with minimal distortion and learn an informative hidden representation. In addition to the basic autoencoder, the model included a binary stochastic neuron layer as the

last layer of the encoder to project the representation into an 8-bit binary space, which is argued to be ideal since most distinctive features are binary. They conducted a comparative analysis of English and Xitsonga, and observed a high degree of similarities between the two languages. Based on this observation, they concluded that feature learning may exhibit a certain degree of cross-linguistic consistency.

The experiment in Shain and Elsner (2019) indicated that the model's latent representation can successfully display the learning of distinctive features. For example, in the English model's hidden representations, nasals' representations were found to concentrate on the same neurons, while the alveolar and post-alveolar fricatives' and affricates representations concentrated on the same neurons. When the latent representation was fed into a random forest classifier for a prediction task of theory-driven features, the different features resulted in varying accuracy in classification across the two languages, English and Xitsonga. For instance, [±voice] achieved high classification accuracy, while [±spread glottis] was extremely low. This finding was argued to be parallel to different feature availability to infants: The voicing feature (VOT) was found to be more salient to infants and even to other infant animals, compared to that of [±spread glottis]. Infant learners' difficulty in discriminating some nasal places was also well reflected in the modeling results (Shain & Elsner, 2019).

Despite the highly promising results of Shain and Elsner's (2019) recent bottom-up learning experiment, the validity of their conclusions remains uncertain due to a crucial ambiguity in the experimental design. Specifically, several underlying factors remain concealed beneath the observed linguistic features. For example, the voicing feature was shared by both vowels and consonants without clear distinction, and segments that utilize this feature were mixed with those that do not. Consequently, it becomes difficult to disentangle whether the model's performance on a particular feature, such as the [±voice] feature, is attributable to the feature itself or to some unforeseen factor. In addition, the precise cause of the considerable discrepancy in feature learning across distinct features remains uncertain. It is unclear whether the imbalanced distribution of features in the training dataset primarily contributes to this difference, or whether there are intrinsic variations in the learnability of distinct features. For instance, is the better acquisition of the voice feature by the bottom-up model primarily due to its more frequent occurrence in specific datasets, or are there inherent characteristics of the voice feature that make it easier to learn compared to, say, the nasality feature? Notably, certain features, such as voice, sonorant, and continuant, were

present in large numbers in the training datasets, while others were underrepresented, potentially influencing the modeling outcomes. Lastly, the learning of phonological features using an autoencoder model that relies on raw sound data has not been investigated in other studies. This current study is to further investigate the ability of an autoencoder model to learn phonological features.

The present study aims to examine the learnability of phonological features as well as the biases that emerge during autoencoder training and investigate the underlying reasons for the biased learning of distinctive features. These biases may simply be a result of feature availability or biased distribution of features: Due to the significant differences in the total count of distinct features among languages, it is possible that the biases arise from imbalanced input data. This indicates that the acquisition of features reflects the language-specific feature availability. In other words, a particular feature or set of features that is overrepresented in a specific language is learned more effectively. Alternatively, the differences found in the learnability of different features may be an intrinsic property of sounds, where certain types of features are inherently more challenging to learn and distinguish, i.e., despite the equal or similar representation of two features, one can be learned better than others. If this is the case, it is essential to examine the nature and characteristics of these biases evident in the modeling outcomes. To test our hypothesis, a autoencoder learning model is trained on two languages with significantly different feature distribution, namely English and Mandarin (see section 2.1).

2 Methods

2.1 Dataset

The experiment utilized two linguistically unrelated languages, English and Mandarin. The English dataset was based on the Buckeye Speech Corpus (Pitt et al., 2007), while the Mandarin dataset was obtained from AISHELL-3 (Shi et al., 2020). Notably, English and Mandarin differ in several ways regarding their distinctive feature distributions, such as the contrastiveness of voicing and aspiration (Deterding & Nolan, 2007) among many. For instance, in English, voiced sounds are prevalent across consonants, including nasals, voiced stops, fricatives, and affricates (Giegerich, 1992, pp. 121-124). Conversely, Mandarin features relatively few voiced consonants, with only nasals, lateral and the voiced retroflex fricative /ʐ/ being voiced (Dow, 1972, p. 40).

Thus, comparing the learning of the two languages may aid in exploring the extent to which learning models can learn the universality and language-specificity of the feature system.

2.1.1 *The English Dataset*

We employed Buckeye Speech Corpus, which is a time-aligned, phonetically labeled American English speech corpus (Pitt et al., 2007). The corpus comprises approximately 300,000 words spoken by 40 speakers from Central Ohio in conversational settings with an interviewer. High-quality recorders were used to make recordings, resulting in a total of approximately 38 hours of speech data sampled at 16000 Hz with 16-bit depth. All 40 speakers' recordings were included in the English dataset and were used in training, validation, and evaluation of the model. Notably, the phonetic labels specified in the corpus are more granular than the English phoneme inventory, differentiating between nasalized vowels, syllabic nasals and laterals, glottal stop, as well as the intervocalic tap /ɾ/.

2.1.2 *The Mandarin Dataset*

We employed AISHELL-3, a large-scale Mandarin speech corpus that comprises over 88,000 read utterances and roughly 85 hours of speech data (Shi et al., 2020). The recordings were made using 44100 Hz, 16-bit HI-FI microphones by 218 native Mandarin Chinese speakers from different provinces. Of the 218 speakers, 64 were randomly selected, ensuring that none of the chosen speakers had recordings that contained only a few tokens. The number of speakers selected was determined to make the total dataset file size compatible with that of the English dataset. Since each Mandarin speaker had fewer sentences than English speakers' data in Buckeye Speech Corpus, 5 English speakers' recordings were roughly equivalent in length to those of 8 Mandarin speakers.

As the Mandarin corpus did not provide phonetic time-alignment, the recordings were automatically aligned using a neural aligner *Charsiu*. This aligner was able to provide highly accurate alignment by utilizing sentence-level transcription in Chinese characters (Zhu et al., 2022). In addition to the Mandarin phoneme inventory, the pinyin transcription generated by *Charsiu* marked the different realizations of /i/ following alveolar fricatives and affricates ([ɻ]) as well as

retroflex fricatives and affricates ([ɻ̩]). The transcription also accounted for the elsewhere condition of [i] (Zee & Lee, 2001). Furthermore, the glide medials and nasal codas were combined with the vowel nucleus into a final form (Lin, 2007) and were not aligned separately.

2.2 Data Pre-processing

Prior to utilizing the raw sound data and time-aligned phonetic transcriptions for the purpose of training and evaluating the computational model, a series of pre-processing procedures were performed. These procedures encompassed primary feature extraction and chunking. Features were executed automatically through Python scripts.

2.2.1 *Prior Extraction of Sound Data*

In order to streamline the training process, we opted to extract the mel-frequency cepstral coefficients (MFCC) from the raw sound data prior to inputting it into the autoencoder. This decision was informed by prior research, which has shown that MFCC - a representation of sound frequency on a logarithmic scale - is a reliable tool for both speech recognition and linguistic analyses (Xu et al., 2004).

We extracted the MFCCs from the raw sound data using a window length of 25ms and a window step of 10ms, which follows the default setting as described by Lyons (2013). This process yielded 13 coefficients for each frame, which were generated by extracting wave information in a single window. In addition to the MFCCs, we also extracted differential coefficients (delta) by subtracting the MFCCs of adjacent frames and acceleration (second-order delta) by subtracting adjacent deltas. This allowed us to capture the dynamic trajectories of the MFCCs (Lyons, 2013). To create a frame of 39 coefficients, the deltas and second-order deltas were concatenated with the MFCCs, as both have the same dimension as the MFCCs.

2.2.2 *Sound Data Segmentation*

To ensure that our experiment focused solely on the bottom-up learning of phonetic systems, we did not segment the training data according to the time-alignment of phonetic transcriptions. Providing the model with such information would introduce existing boundaries and exclude sounds outside of the segments, which could potentially bias the model. Moreover, there is limited

evidence suggesting the existence of primitive sub-word boundaries prior to language acquisition (Kuhl, 2004), and speech input to infants does not appear to provide cues for natural segmental boundaries (Räsänen, 2014). To address this issue, we used random sampling to extract sound data, resulting in segments with varying lengths that were similar in distribution to the evaluation data with variable segment length. Segments of random lengths were sampled from a truncated normal sampler centered at the mean segment length of the total dataset and lower bounded by 0. The data for evaluation were segmented based on time-aligned phonetic transcriptions, and only segments containing phonetic tokens were included, excluding silence and pure noise. A CSV file listing the segments was created, along with the corresponding phonetic transcriptions.

To simplify modeling and improve accuracy of representation learning (Shain & Elsner, 2019), we resampled the segments, which consisted of varying numbers of extracted frames, to a fixed length of 25 frames per segment for input into the autoencoder. This resulted in segment dimensions of $F \times C$, where $F = 25$ and $C = 39$ for input. Although the input and output of an autoencoder model are typically identical, our model did not assume this because the deltas and second-order deltas were directly derived from the MFCCs.

The pre-processed sound data was chunked into groups of 5 (English) or 8 (Mandarin) speakers, each of which was of similar length. Each language material consisted of 5chunks that were used for training the model, while the other three were unseen, and thus used for validation and evaluation. The training and validation chunks underwent random sampling segmentation (random-sampled chunks), and the evaluation chunks were segmented based on the time-alignment of phonetic transcriptions (aligned chunks).

2.3 Model Settings

As previously stated, the model used in this experiment was an autoencoder (Bank et al., 2020), whose schematic structure is illustrated in **Figure 1.** As an autoencoder aims to encode input information (acoustic cues) into a latent space (mental representation of phonological knowledge) and reconstruct the input with minimal distortion (production), the resulting latent representation can be considered as the learned knowledge of an individual segment. These segments can be seen as analogous to the internal organization of a single phonological category in humans (Kuhl, 1991). As in **Figure 1,** the model's encoder and decoder consist of two linear layers and a residual block

as the main structure, which was simplified from Shain and Elsner (2019)'s model and was preliminarily tested to best perform in the segment classification task. The pre-processed input data in the shape of $F \times C$ were first flattened to a vector of length $FC = 975$ before being fed into the encoder. This step did not distort the information per se but simply allowed the linear layer to process the short segment as a holistic component, instead of a sequence of frames. Alternatives might include using Convolutional Neural Network (CNN; Palanisamy et al., 2020) or Recurrent Neural Network (RNN; Sherstinsky, 2020), which can treat input as sequences. However, to be more comparable to the previous work (Shain and Elsner, 2019), we used linear layers in the current modeling. After flattening the input data, the encoder transformed the input from FC to an intermediate number of dimension $I = 256$ through a linear layer with ReLU as the activation function following the process in **Figure 1**.

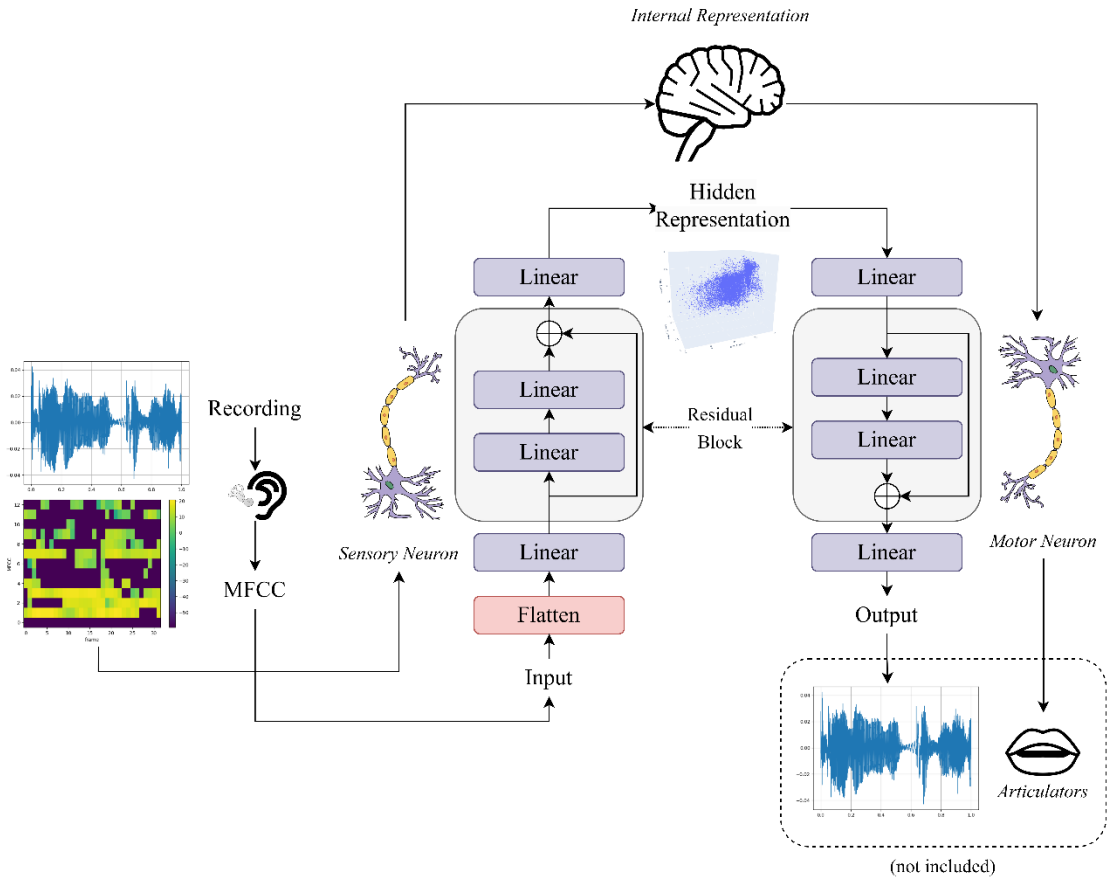


Figure 1. Structure of the autoencoder model used for this experiment. Dimensions of data marked in parenthesis.

The model structure in **Figure 1** was capable of approximating non-linear functions (Liu & Liang, 2019) and was thus able to simulate the neural activities involved in the speech perception process. The 256-dimensional latent representation was then passed through a residual block consisting of two stacked linear layers with ReLU as the activation function, as in **Figure 2.** The residual network structure outperforms the plain linear feedforward network structure in deeper networks due to the shortcuts it provides to propagate information through multiple "paths" with varying network depth (He et al., 2016). The current model allowed both (a) information that underwent the two linear layers in the residual block ($\boldsymbol{x_2} = Res(\boldsymbol{x_1})$) and (b) information that simply underwent the first linear layer ($\boldsymbol{x_1}$) to be present before they were fed into the last linear layer. A final linear layer transformed the 256-dimensional latent representation into the hidden representation of dimension $H = 3$; this is a three-dimensional hidden space, which can be understood as mental representation of phonological space. The deep neural network model of two linear layers and a residual block (equivalent to a total of four linear layers) simulated the complex, layered neural transformations underlying speech perception that convert the signal captured by sensory receptors into the underlying neural code of segments (Eggermont, 2001; Yost, 2007).

The decoder (production process) followed the same structure. Unlike the stacked autoencoder used in Kamper (2015), which stacked multiple encoders and decoders to yield multiple hidden representations (Liu et al., 2018), our current learner model only had a single set of encoder and decoder, with the output of the encoder being the hidden representation.

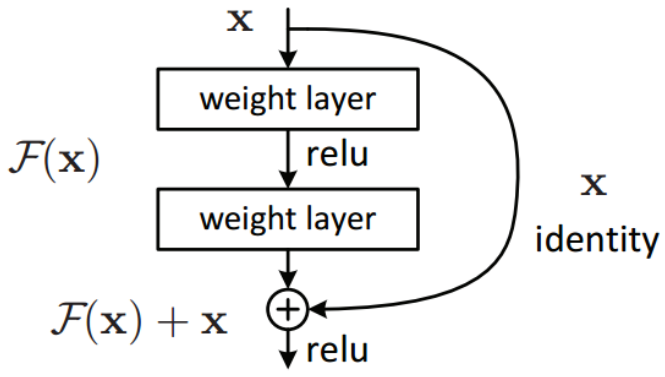


Figure 2. Structure of a residual block

The hidden representation was not composed of 8 binary bits, but rather several continuous real number values (set to 3), in contrast to Shain and Elsner (2019). Although they claimed that the binary representations were similar to the binary distinctive features proposed by Hayes (2009), the 8 binary hidden dimensions did not seem to have explicit correlations with the distinctive features, as the number of distinctive features far exceed the 8 dimensions. This lack of explicit correlation implies that an additional projection is required to transform representations in the binary hidden space to the distinctive feature space, which can be considered as a much smaller hidden space of the same structure as continuous spaces. Moreover, Kuhl's (1991) "perceptual magnets" hypothesis suggests that there are continuous internal categories that are perceived as one of the category centers, analogous to the extra projection from the hidden representation to the phonemes and implicitly, further to their features, rather than directly enforcing restrictions on each hidden dimension.

It should be noted that this entire model is closely related to the natural processing model of speech signals in humans, where the MFCC extraction and encoder simulate the neural coders that transform speech waves into internal codes of sounds (Eggermont, 2001; Yost, 2007), and the decoder simulates the reverse processing of generating sounds from internal representations (excluding articulation, as there is no simulation of articulators here). The whole process could be similar to infants' early acquisition of plain, non-segmented speech input to form distributional prototypes and eventually the phonetic system (Hayes, 2004).

The model was implemented using Pytorch (Paszke et al., 2017) and was optimized using Adam (Kingma & Ba, 2014) with a learning rate of 0.001. The autoencoder models were trained on five seen chunks that underwent random sampling (English 25 speakers, Mandarin 40 speakers) and were concurrently validated on one of the three unseen chunks that underwent the same random sampling as the training set. The model was trained for 30 epochs until it reached near-convergence.

3 Experimentation and Results

The experiment consisted of three main phases. Phase I involved training the autoencoder learner model. Phase II evaluated unsupervised clustering using measures of homogeneity, completeness, and V-measure (Rosenberg & Hirschberg, 2007). Phase III involved statistical tests

on the hidden representations of phonemes or natural classes. Each phase involved different models and data as follows. In Phase I, the autoencoder model was trained using random-sampled chunks. In Phases II and III, only the encoder part of the autoencoder was used to project the MFCC data onto the hidden representation space. Aligned chunks were used in both Phases II and III. Table 1 provides an overview of the tasks and data used in each phase of the experiment.

Phase	Language Task	English	Mandarin
1	Autoencoder train	_06_10[1], _16_20, 21_25, _26_30, _31_35	_01_08, 09_16, _41_48, _49_56, _57_64
1	Autoencoder development	_01_05	_17_24
2	Clustering	_01_05	_17_24
3	Natural Class Evaluations	_01_05, _11_15, _36_40	_17_24, _25_32, _33_40

Table 1. Data chunks and their usage

3.1 Clustering Evaluation and Results

Following the training of the autoencoder learner model on each dataset, we first performed an unsupervised clustering evaluation using measures of homogeneity, completeness, and V-measure (HCV measures) to assess the model's efficacy. Specifically, we evaluated the learner model's ability to accurately project different phonemes to distinct regions of the hidden representation space. This process bears resemblance to infants' acquisition of phonetic category centers from

[1] Name of data chunk.

sound distributions (Hayes, 2004). We employed this method as the learner model in the experiment had minimal access to contextual segment properties, if any at all.

The time-aligned tokens of the HCV test chunk were initially projected to the hidden space using the encoder of the trained model. Subsequently, Kmeans, a rudimentary but efficient unsupervised clustering algorithm (Lloyd, 1982), was utilized to cluster the data with the number of cluster (n_clusters) set to the number of phonemes in the dataset. The resulting cluster labels[2] were then compared with the ground truth phonetic transcription labels to compute the HCV scores as presented in **Table 2**.

	English			*Mandarin*		
	H	C	V	H	C	V
Baseline	0.006	0.004	0.005			
Our model	0.308	0.273	0.289	0.345	0.342	0.344

Table 2. Phone clustering scores. Baseline scores indicate the results in Shain and Elsner (2019). HCV scores on the respective clustering task data chunk.

As shown in **Table 2,** the HCV scores of both the English and Mandarin datasets were found to be comparable, and crucially, they are substantially higher than the baseline scores reported in Shain and Elsner (2019). This finding provides initial support for the use of continuous hidden space instead of relying on a binary feature system. It further suggests that phonetic category centers emerged from pure acoustic input that lacked phonological context and segmental boundary information during the autoencoder's learning of reconstructing or reproducing the raw input sounds. The emergence of phonetic category centers challenges the notion of innate phonetic boundaries (Eimas et al., 1971; Gilkerson, 2005; Nittrouer, 2001) by suggesting that the ability to perceive distinctions between phones may originate from readily-available acoustic distinctions between sounds. This is because the model did not have any assumptions about the shapes of phones, but was merely equipped with basic auditory system functions, such as keeping the

[2] $L \in C^{\mathrm{N}}, C = \{0, 1, \ldots, n-1\}$, N is the number of segments in chunk.

temporal structure of sound signals (Geisler, 1998, p. 4) and decomposing waveform signals into frequency components.

3.2 Natural Class Evaluations

To assess the model's ability to learn different phonemes and natural classes, we conducted direct plotting of hidden representations of sample tokens. To reflect the gradual binary fission process of human learning (Jakobson & Halle, 1956, p. 47) and the binary nature of distinctive features, we performed pairwise comparisons rather than comparing multiple phonemes together. This approach is also supported by evidence that phoneme knowledge in humans emerges from distribution-based clustering involving a series of binary decisions (Guenther & Gjaja, 1996; Hayes, 2004; Maye & Gerken, 2000). As a result, pairwise inspections offer a clearer and more straightforward evaluation method. Among all of the logically possible phoneme pairs, which are around 500, we selected those with a small number of distinctive featural differences, i.e., 1 to 2 differences, and present the data here. As it is argued that the more contrasting features are included, the more distinct the two phones are (Bailey & Hahn, 2005), success in distinguishing closer pairs would possibly imply success in farther pairs. For a larger set of pairs, please see the link[3].

Due to the frequency asymmetry of different phonemes in a language, around 3,000 tokens of each selected phoneme were randomly sampled from the time-aligned data chunks. As with the clustering task, the sampled tokens were processed by the model's encoder to yield the corresponding hidden representation. Each token was accompanied by its ground truth label for evaluation. The tokens' hidden representations were plotted in three-dimensional scatter plots, with the position of a point in the three-dimensional space representing the three hidden representation values and color representing the ground truth label that indicates to which phoneme or natural class this token belongs according to the transcription. Each hidden dimension value underwent min-max normalization (İleri et al., 2018), respectively. When printed on two-dimensional media, we chose a relatively best angle through which the boundaries were clear. The existence of such an angle and the clarity of a boundary would further indicate the success of the

[3] https://drive.google.com/drive/folders/1ROawP6Vh5tGUe4GlU2ggfJbsw8HAD_at?usp=share_link

learner model. This evaluation will explore hidden representation distributions of phonemes and natural classes in both English and Mandarin in parallel.

To preview, despite the typological unrelatedness and divergent distributions of phonological features between the two languages, our results demonstrate that similar performance patterns were achieved to a certain degree, indicating that the acquisition of phonological features through training on bottom-up data is not language-specific. Instead, it appears to rely more on universal properties of sounds to learn distinctive features.

3.2.1 *Vowels*

The results of the phoneme evaluation showed that vowel contrasts were relatively salient, with the [±back] ($d = 3.18$) and [±high] ($d = 3.48$) contrasts in Mandarin showing a very clear boundary and relatively large distances. In English, while less distinct, the division was still apparent, with the [±back] contrast having a centroid distance of $d = 1.33$ and the [±high] contrast having a centroid distance of $d = 1.05$.

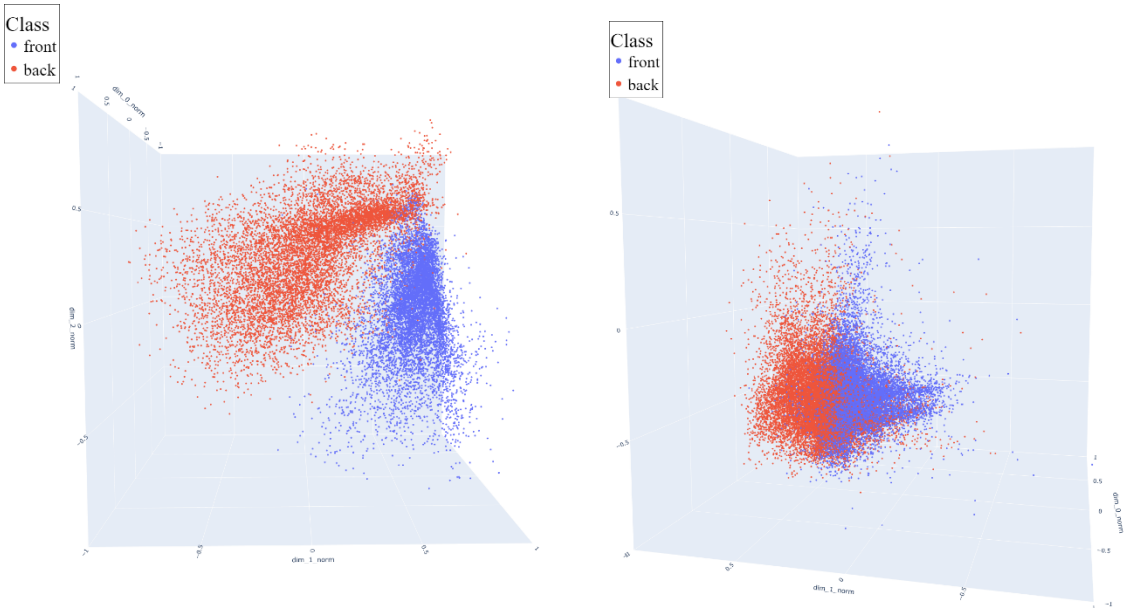


Figure 3. Plot of [±back] (/i/, /y/, /eɪ/~/o/, /u/, /ɤ/, /ou/) in Mandarin

Figure 4. Plot of [±back] (/æ/, /ɛ/, /eɪ/, /ɪ/, /i/~/ɔ/, /ɔɪ/, /oʊ/, /ʌ/, /u/, /ʊ/) in English

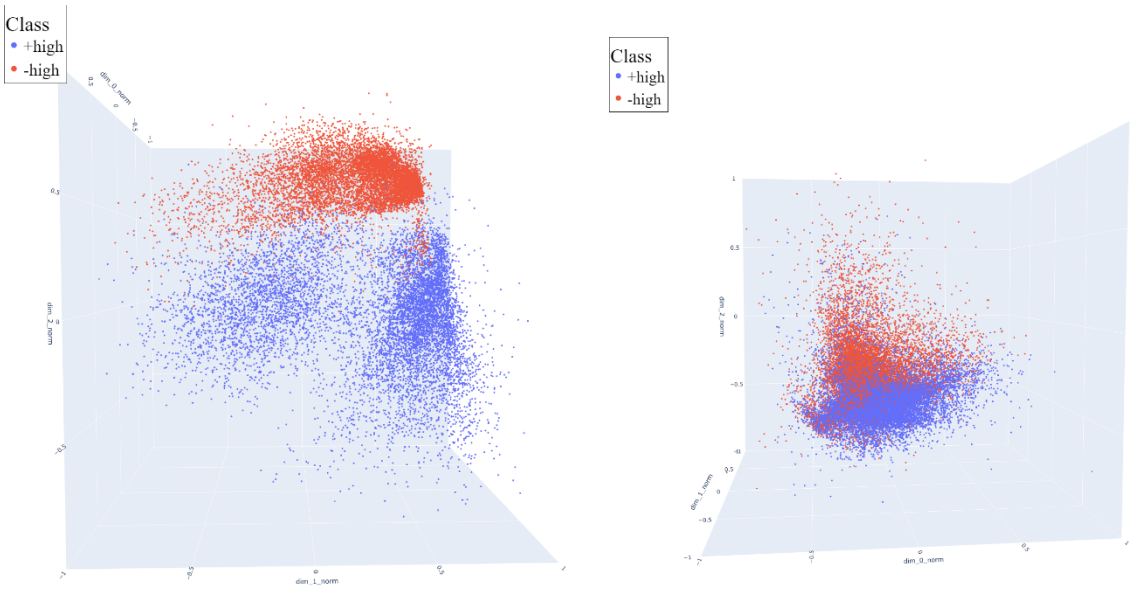


Figure 5. Plot of [±high] (/i/, /y/, /u/~/ai/, /au/, /a/) in Mandarin

Figure 6. Plot of [±high] (/eɪ/, /ɪ/, /i/, /ɔɪ/, /oʊ/, /u/, /ʊ/~/ɔ/, /ʌ/, /æ/, /ɛ/) in English

In contrast to the relatively clear distinctions observed for vowel features, the acquisition of the distinction between monophthongs and diphthongs was found to be not very successful in both Mandarin ($d_{man} = 0.75$) and English ($d_{eng} = 0.55$). As demonstrated in **Figure 7** and **Figure 8**, the boundaries between monophthongs and diphthongs are indistinct, resulting in a significant overlap between the two regions.

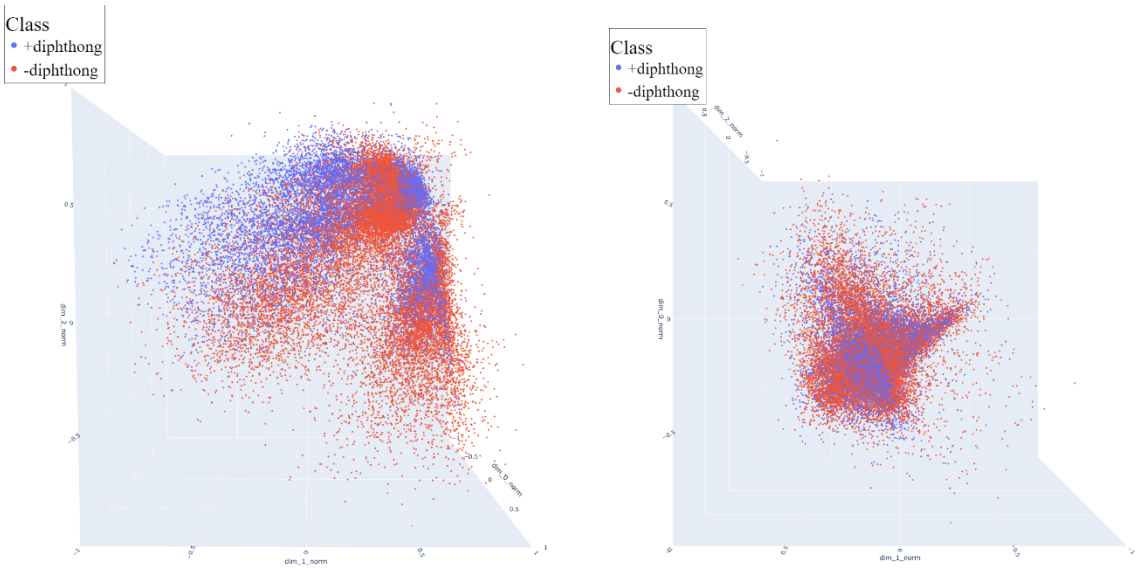


Figure 7. Plot of [±diphthong] (/ou/, /ai/, /eɪ/, /au/~/i/, /ɚ/, /ɻ/, /y/, /u/, /o/, /a/, /ɤ/) in Mandarin

Figure 8. Plot of [±diphthong] (/aʊ/, /ɔɪ/, /oʊ/, /eɪ/~/ɔ/, /ɛ/, /ɚ/, /ʌ/, /u/, /ʊ/, /ɪ/, /i/) in English

In addition to the scatter plots, we conducted statistical analyses to evaluate the model's capability in discriminating between various phoneme pairs and natural classes. Hotelling's T-square test, a multivariate T-test, was utilized, and the Mahalanobis distance between the centroids of each phoneme or each natural class was computed. Approximately 3,000 tokens of each phoneme were randomly sampled from the time-aligned data chunks and processed through the model's encoder to obtain their corresponding hidden representations. When the maximal number of tokens for one class was substantially lower than that of the other, the sample size was adjusted to achieve relative balance between the two classes. Results are provided in **Tables 3-4**.

d	y	ɚ	a	ɤ	u
i	1.313	4.698	5.263	4.131	3.745
y		4.354	5.123	3.717	3.776
ɚ			1.113	1.113	4.040
a				1.751	4.419
ɤ					2.829

Table 3. Pairwise distance of Mandarin monophthongs (/i, y, ɚ, a, ɤ, u/)

d	ɛ	ʊ	ɪ
æ	0.841	1.643	1.518
ɛ		0.988	0.896
ʊ			1.038

Table 4. Pairwise distance of English lax monophthongs (/æ, ɛ, ʊ, ɪ/)

Our findings in **Tables 3-4** indicate that, consistent with the descriptive analysis presented earlier, the model successfully mapped different vowels to distinct areas in the hidden representation space, with all evaluated pairs showing a significant difference from one another. In general, vowels in Mandarin display unambiguous boundaries and great centroid distances, much better than consonants (See section 3.2.2 for the results of consonants). The Mahalanobis centroid distance between vowels (see **Table 3**) ($\mu_v = 3.05$)[4] was high ($p < 0.001$), much greater than that between consonants ($\mu_c = 2.03$). The distances between vowels in English (**Table 4**) not as distinguishable ($p = 0.107$), although their distributional patterns remain identifiable and are more distinguishable than those of consonants (refer to section 3.2.2 for consonant results). This is consistent with the acquisition patterns of human infants, where vowels are typically acquired earlier than consonants (Pepperkamp et al., 2003). It is noteworthy that the differences among vowels seem to go beyond individual pairs, as the distribution of vowel tokens appears to align

[4] Including both monophthongs and diphthongs.

significantly with backness and height, even though they are considerably different from the corresponding backness and height features (Lee & Zee, 2003) or by F1 and F2 (Zee & Lee, 2001). Additionally, it is observed from the distance measures that acoustically similar vowels are closer to each other. For instance, /i/ and /y/, the two which are closer with each other, are only 1.31 units apart, while /i/ and /a/ are separated by a distance of 5.26 units.

3.2.2 *Consonants*

The patterns of consonants are generally more complex and varied than those of vowels. However, the model has successfully learned many consonant pairs, as demonstrated by the three-dimensional scatter plots of consonant pairs or natural classes, see **Figures 9 – 19**. These plots illustrate the model's ability to encode consonantal data into its hidden space, facilitating decoding after training on continuous raw speech data.

Whereas clustering evaluation results provide a general indication that the model has learned to map similar sounds to corresponding regions in the hidden space, the phoneme evaluation results further suggest the successful learning of phonetic category centers in the hidden representations for consonants. While there are variations in the central tendency and dispersion of the distributions and the boundaries between phonemes that extend along different directions, many phoneme pairs exhibit clear boundaries and large centroid distances. Details are as follows. The [±strident] distinction between plosive and fricative/affricate phonemes, which mainly contrasts fricatives/affricates and other consonants (except /x/ in Mandarin), was relatively clear for both languages ($d_{man} = 1.67; d_{eng} = 1.39$), as shown in **Figures 9 – 10**. Similarly, the distinction between plosives and affricates ([± delayed release]) was relatively clear ($d_{man} = 1.59; d_{eng} = 1.20$), as seen in Figures **11 – 12**. Despite a higher frequency of affricates in Mandarin compared to English, the English learner model did not exhibit a significant disadvantage in learning affricates compared to Mandarin, suggesting a potential cross-linguistic universality to some extent.

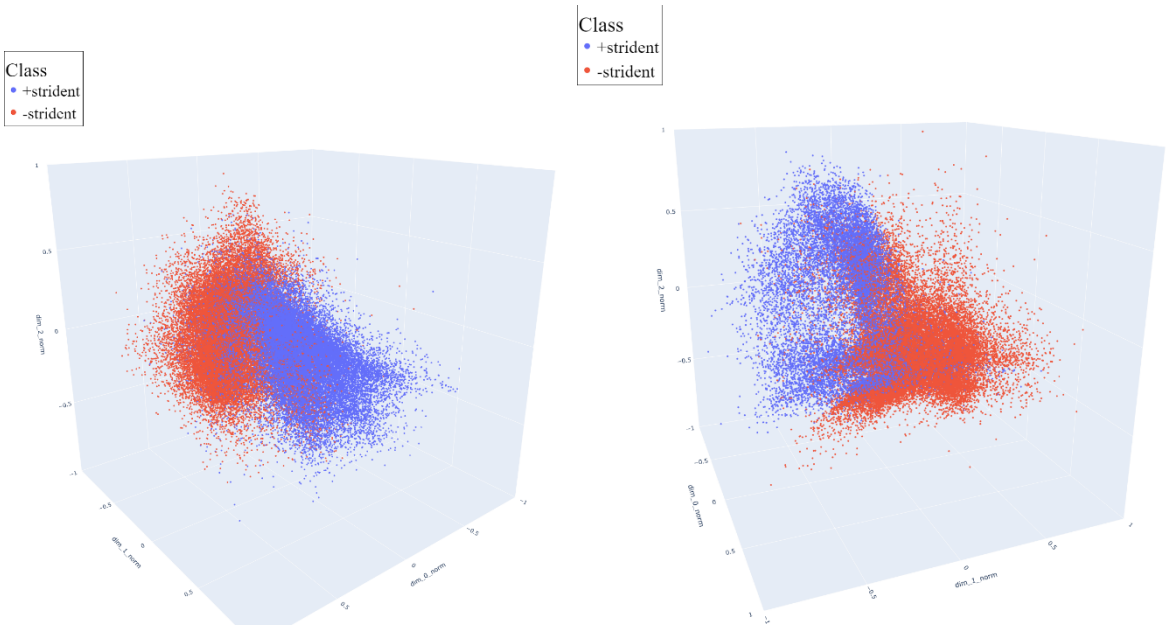


Figure 9. Plot of [±strident] (/ts/, /tsʰ/, /s/, /tɕ/, /tɕʰ/, /ɕ/, /ʈʂ/, /ʈʂʰ/, /ʂ/, /ʐ/, /f/~/tʰ/, /k/, /kʰ/, /t/, /p/, /pʰ/, /x/, /m/, /n/, /l/) in Mandarin

Figure 10. Plot of [±strident] (/s/, /z/, /ʃ/, /ʒ/, /tʃ/, /dʒ/, /f/, /v/~/t/, /g/, /k/, /d/, /b/, /p/, /h/, /m/, /n/, /l/, /ɹ/, /j/, /w/) in English

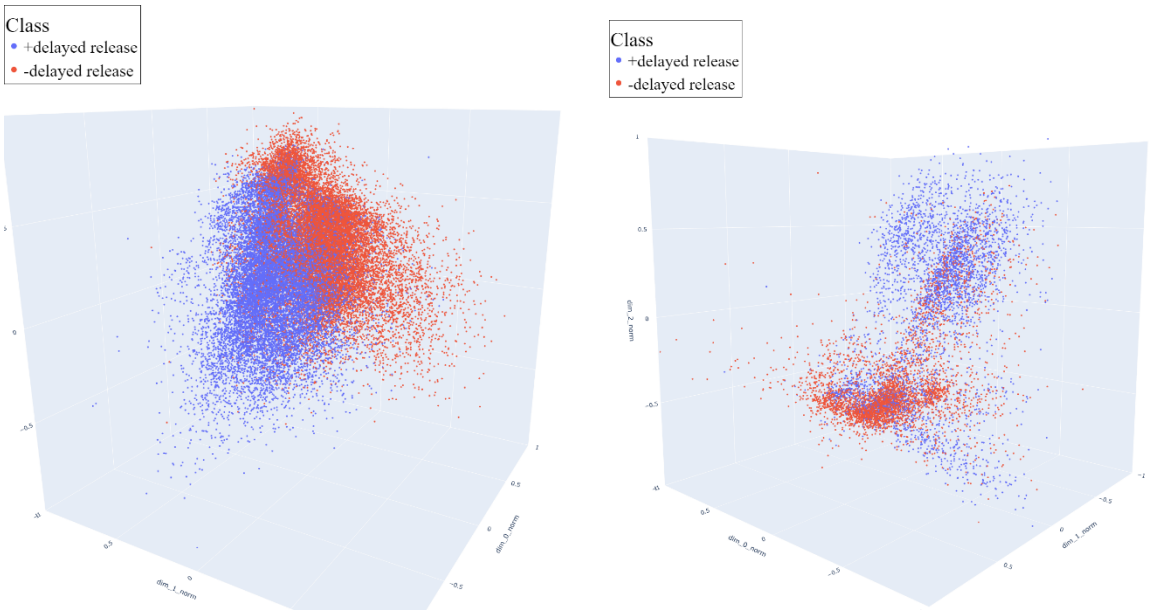


Figure 11. Plot of [±delayed release] (/ts/, /tsʰ/, /tɕ/, /tɕʰ/, /ʈʂ/, /ʈʂʰ/~/p/, /pʰ/, /t/, /tʰ/, /k/, /kʰ/) in Mandarin

Figure 12. Plot of [±delayed release] (/tʃ/, /dʒ/~/t/, /d/) in English

Recall that there are significant differences in the use of voicing contrast between the two languages (see section 2.1). Despite this, our modeling results demonstrate that the voicing contrast was well-learned in both Mandarin and English, with $d_{man} = 3.30$ and $d_{eng} = 1.22$. Although Mandarin has only one voicing contrast pair (/ʂ/~/ʐ/), in contrast to English's extensive inventory, the successful learning of this contrast by the Mandarin model suggests the possibility of universality in feature learning.

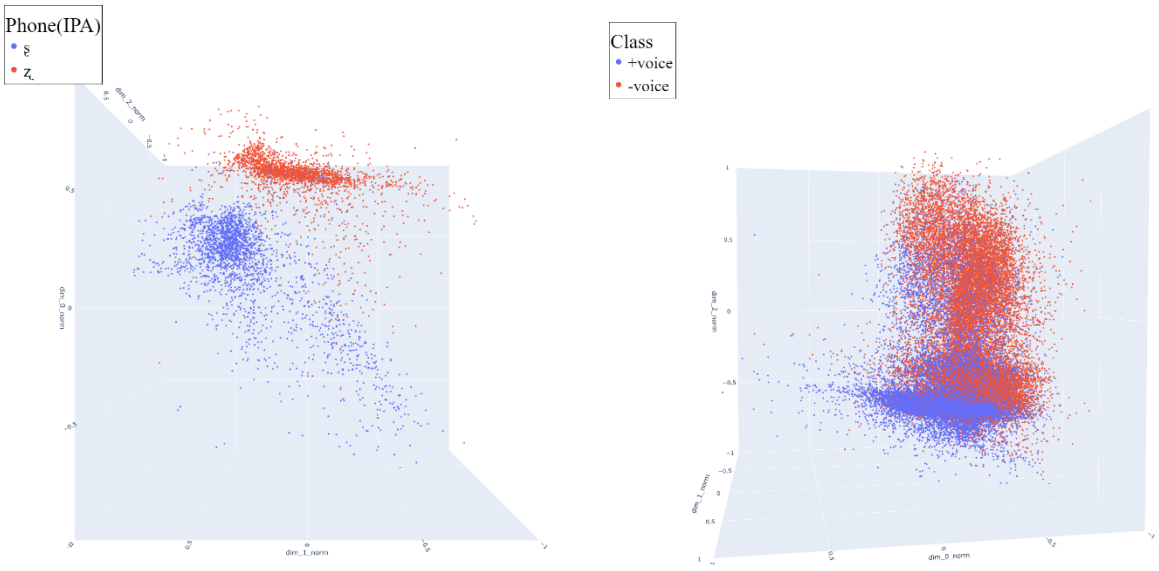


Figure 13. Plot of [±voiced] (/ʂ/~/ʐ/) in Mandarin

Figure 14. Plot of [±voiced] (/d/, /dʒ/, /ð/, /ʒ/, /z/, /ɡ/, /b/, /v/~/t/, /tʃ/, /θ/, /ʃ/, /s/, /k/, /p/, /f/) in English

Our evaluation of the nasal feature [±nasal] indicates that the distinction was not salient in both Mandarin and English, with considerable overlap in the class distributions ($d_{man} = 0.79$, $d_{eng} = 0.87$). Similarly, the [±labial] feature was poorly learned by the model in both languages, with even greater overlap ($d_{man} = 0.48$, $d_{eng} = 0.46$).

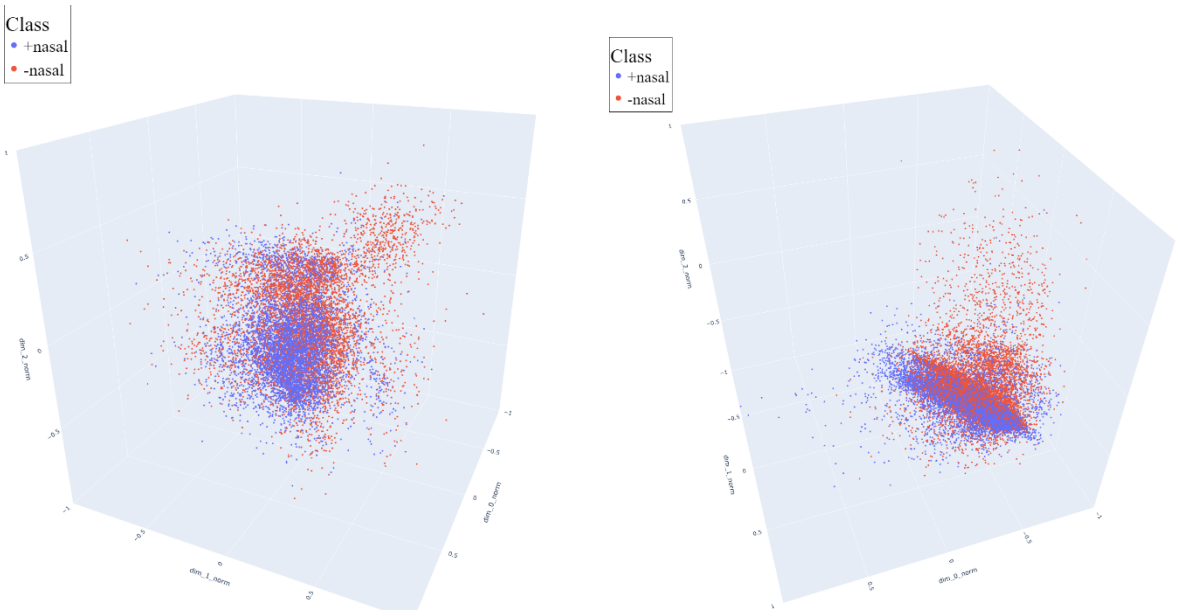


Figure 15. Plot of [±nasal] (/n/, /m/~/t/, /p/) in Mandarin

Figure 16. Plot of [±nasal] (/n/, /m/, /ŋ/~/d/, /b/, /g/) in English

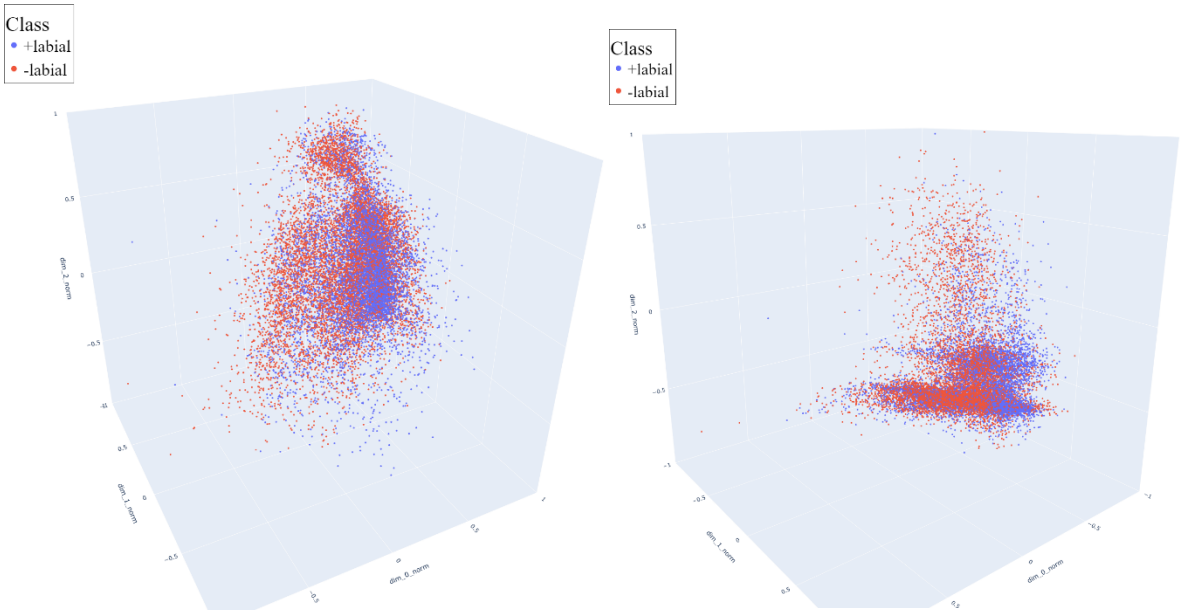


Figure 17. Plot of [±labial] (/p/, /pʰ/, /m/~/t/, /tʰ/, /n/) in Mandarin

Figure 18. Plot of [±labial] (/b/, /p/, /m/~/d/, /t/, /n/) in English

To evaluate the performance of our model in distinguishing between different consonant pairs, we conducted statistical analysis using the same methods as for vowels. Overall, the distribution

centroid distances in the two languages were compatible ($\mu_{man} = 2.03; \mu_{eng} = 1.96$). For a full data set, see the link in footnote 3. Specifically, phoneme pairs /ʂ/~/ʐ/ ($d = 3.30$), /t/~/s/ ($d = 3.08$), /t/~/t͡ɕ/ ($d = 2.34$), /t/~/ʂ/ ($d = 2.38$), and /f/~/ɕ/ ($d = 2.15$) in Mandarin, and /t/~/ʃ/ ($d = 2.21$), /ʒ/~/ʃ/ ($d = 2.18$), and /n/~/l/ ($d = 2.07$) in English were especially successfully learned. The phoneme boundaries were very clear, without much overlap in distribution, and the centroid distances were relatively large. Following are mid-level distant pairs, whose Mahalanobis centroid distances are roughly around 1~2. These pairs might have relatively large overlap in distribution, but the central tendencies are still rather distinct. These pairs include various contrastive features, not showing any coherent tendencies. When the centroid distance is lower than 1, the phoneme boundary generally would not be very obvious, and the pair would intrude further into the counterpart's region until near-complete overlap.

Among the low-distance pairs, two groups stand out. One is nasality, as shown from the descriptive data above (**Figures 15 – 16**): both the nasality contrast and place distinctions within nasal sounds seem hard to the model. In both Mandarin and English, nasal sounds at different places of articulation were projected very close to each other with great overlap (/m/~/n/: $d_{man} = 0.60, d_{eng} = 0.411$; /m/~/ŋ/: $d_{eng} = 0.39$, /n/~/ŋ/: $d_{eng} = 0.12$). Especially, the distinction between /n/ and /ŋ/ was extremely small in both languages. This bad performance seems to conform with human infants' difficulty in distinguishing these two nasal consonants (Narayan et al., 2010). In addition, the distinction between nasal stops and oral stops was small as well (/p, b/~/m/: $d_{man} = 0.76, d_{eng} = 1.03$; /t, d/~/n/: $d_{man} = 0.87; d_{eng} = 1.04$; /g/~/ŋ/: $d_{eng} = 0.70$). The results of the analysis further revealed that plosives are also difficult to distinguish by place. In Mandarin, all plosive place contrast pairs have a distance lower than 1 (**Table 5**), and such pattern is mostly observed in English as well (**Table 6**). Among voiceless plosives in English, they appear to be slightly more distinct than their voiced counterparts. Furthermore, aspirated plosives in Mandarin exhibit greater distinctiveness than unaspirated plosives, with the exception of /pʰ/~/tʰ/. Similarly, aspirated affricates display greater distinctiveness than their unaspirated counterparts (**Table 7**).

C1	C2	Mah. Dist.
p	t	0.322
pʰ	tʰ	0.530
p	k	0.571
t	k	0.851
pʰ	kʰ	1.088
tʰ	kʰ	1.292

Table 5. Mahalanobis distance ranking of plosive place contrast pairs in Mandarin

C1	C2	Mah. Dist.
b	d	0.464
d	g	0.371
b	g	0.621
t	k	0.732
p	t	0.699
p	k	0.840

Table 6. Mahalanobis distance ranking of plosive place contrast pairs in English

C1	C2	Mah. Dist.
ts	ʈʂ	0.743
ts	tɕ	0.934
tsʰ	ʈʂʰ	0.990
tɕ	ʈʂ	1.271
tsʰ	tɕʰ	1.469
tɕʰ	ʈʂʰ	1.984

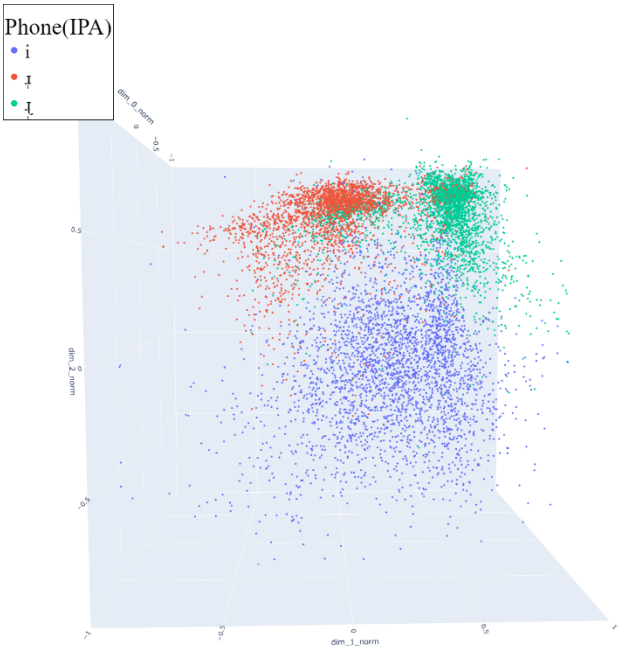
Table 7. Mahalanobis distance ranking of affricate place contrast pairs in Mandarin

The diverse range of learning outcome in **Tables 5 – 7** suggests so called "imperfect" learning from clustering evaluation: some phoneme pairs appear to have quite clear boundary, but some are

inaccurately merging with each other, as shown from the poor learning of nasality and place distinctions among plosives.

3.2.3 *Allophonic Distribution*

Now we move on to the learning of allophonic distinctions. As an illustration, we present the allophonic distributions of /i, ɨ, ʯ/ in Mandarin (**Figure 19**). This group was chosen for presentation purpose here, due to its comparable number of tokens to other phonemes in both the Mandarin and English datasets. Our objective was to investigate potential differences in the learning of allophonic distributions compared to phonemic distinctions.



$d(i,\ ii) = 3.660,\ d(i,\ iii) = 3.223,\ d(ii,\ iii) = 2.159$

Figure 19. Plot of variants of *i* (i~/i/, ii~/ɨ/, iii~/ʯ/)

The three allophones were well distinguished as in **Figure 19,** but importantly the current bottom-up model did not show a sensitivity to the difference between phoneme and allophone projections ($\mu\left(i, ɨ, ʯ\ dists\right) = 3.01;\ \mu(y,\ ɚ,\ a,\ ɤ,\ u\ dists) = 3.42;\ p = 0.653$); neither one is more distinct (as in Kolanchina & Magyar, 2019) or less distinct. In a sense, allophony was not

captured by the model. Unlike top-down models (Kolanchina & Magyar, 2019; Silfverberg et al., 2018), and despite human's different perceptual sensitivity and learnability of phonemes and allophones (Martin et al., 2013, Mitterer et al., 2018, Pepperkamp et al., 2003), the current model was not able to tell whether a contrast is phonemic or allophonic. This failure in learning allophonic distributions, however, makes sense, as the model was exposed to no, or least, contextual information, which would not further induce clustering of allophones into phonemes for humans (Maye & Gerken, 2001; Pepperkamp et al., 2003).

3.2.4 *Feature Distance*

The hidden space distance appears to be partially consistent with the theory-driven feature-based distance metric, where the similarity between phonemes *a* and *b* is greater than that between *a* and *c* if *a* and *b* have fewer contrastive features than *a* and *c* (Bailey & Hahn, 2005). For the purpose of this presentation, we selected four phonemes, namely /tʰ, t, p, m/, which form a near-minimal pair continuum where two adjacent phonemes differ in roughly one feature. Specifically, the difference between /tʰ/ and /t/ is only in aspiration ([±spread glottis]), /t/ and /p/ differ only in place of articulation ([±labial]), and /p/ and /m/ differ in voicing and nasality.

As shown in **Table 8**, the distance of different featural contrasts was accumulative, although not linear. This might suggest that feature-based distance metric is aligned with empirical acoustic-based quantitative prediction, as in the current study, to the evaluation of phonemic distances. For example, the two phonemes /tʰ/ and /t/, which differed only by aspiration in Mandarin, were closer than /tʰ/ and /p/, which differed by aspiration and place of articulation. Both pairs were closer than /tʰ/ and /m/, which differed by nasality and voicing, in addition to aspiration and place of articulation ($d(\mathrm{t^h}, \mathrm{t}) < d(\mathrm{t^h}, \mathrm{p}) < d(\mathrm{t^h}, \mathrm{m})$). The Mahalanobis distance between these phonemes seemed to be accumulative in relation to the number of contrastive features. Yet, by comparing /tʰ/ vs /t/, /t/ vs /p/, and /p/ vs /m/, it revealed that the distance created by various contrastive features may not be uniform. Due to a possibly high dimensionality of hidden representation and the non-uniform featural distances, the distance created by aspiration and that by place of articulation would not be directly comparable, i.e., it is not guaranteed that $d(\mathrm{t^h}, \mathrm{t}) > d(\mathrm{t^h}, \mathrm{p^h})$, for example.

d	*t*	*p*	*m*
t^h	1.57	1.84	2.88
t		0.34	1.05
p			0.71

Table 8. Pairwise distance of /t^h, t, p, m/

4 General discussion & Conclusion

This study has investigated the efficacy of an autoencoder model in learning phonemes and distinctive features from unsegmented, non-transcribed bottom-up wave data. The utilization of unsegmented, non-transcribed "raw" sound data aligns with the initial challenges encountered by human infants during the early stages of language acquisition, where they are exposed to continuous, unsegmented sound streams as input (Liberman et al., 1974). To ensure compatibility between the training and test data distributions, random sampling segmentation was employed as a model training technique during data preprocessing. This approach is more justifiable than fixed segmentation since learners lack prior knowledge regarding segmentation (Liberman et al., 1974) and are likely to attend to random pieces of sounds.

4.1 Empirical evidence supporting bottom-up feature learning

The experimental results presented here support the hypothesis that the knowledge about phonological features can be learned through a process of "repeating". More specifically, this process involved iteratively projecting sounds into a hidden space, and subsequently reconstructing this space into a production without any prior knowledge of segmentation. The acquired knowledge is represented by different distributions in the hidden space. The model was capable of clustering segments of the same phoneme and projecting different phonemes to separate regions in the hidden space. The projection appeared to group similar sounds sharing features together in closer regions and separate dissimilar sounds in farther regions. While the model was not flawless, its considerably higher HCV scores than those obtained by random chance in clustering evaluation, as well as its good performance shown from the Mahalanobis distances and phoneme evaluation plots, validated the current bottom-up model. The knowledge acquired by the

current model is consistent with Kuhl's (1991) phonetic category centers, in which a specific space is assumed and the "category members" (analogous to hidden representations of segments) are located in that space. These members are perceived as one of the phonetic category centers, as if a perceptual magnet was attracting them. Similarly, although lacking a precise, specific point in the hidden space, the perceptual magnets could be viewed as a distribution to which the clustered points of segment hidden representations belong.

The cross-linguistic feature learning similarities between Mandarin and English observed in the present study, as well as the similar biases found across the two languages, suggest that the model trained on bottom-up information is capable of acquiring universal properties of phonological features, rather than relying on phonological and language-specific properties. This ability of the model to perceive and learn sounds well across languages is comparable to the capability of infants in their first six months of life, during which they can distinguish phonetic contrasts of all languages, though they do have access to both top-down and bottom-up information during their learning (Kuhl, 2004). Moreover, the learned projection to map sounds to hidden representation space, as determined by the feature distance evaluation, supports the phoneme distance theory based on distinctive features (Bailey & Hahn, 2005). The transformations to the knowledge base through learning top-down information resembles the later stages of language learning, during which language-specific knowledge is formed (Hayes, 2004), which was not included in the current modeling.

4.2 Discrepancies between the availability of features and their learned outcomes

Notably, both the Mandarin and English models showed a remarkable ability to acquire voicing contrast, despite clear differences in the number of consonantal pairs exhibiting this feature in the two languages. The Mandarin model, for instance, demonstrated high accuracy in acquiring the voicing contrast, even though there was only one pair (/ʂ/ and /ʐ/) displaying voicing contrast among consonants. This exceptional performance is noteworthy, as the number of instances or variety of categories in the training data alone may not be sufficient to account for good generalization. Thus, other factors may contribute to the successful learning of voicing contrast.

One potential factor to consider is the acoustic salience of the voicing feature. In terms of acoustic cues, voicing is clearly reflected by a periodic soundwave and continuous fundamental frequency (F0), whereas voiceless sounds lack this characteristic feature (Johnson, 2011, p.13). This prominent contrast between voiced and voiceless sounds could have potentially facilitated the learning of voicing feature, even in cases where the amount of data signalling the contrast was limited. Another plausible factor to consider is the distinctive phonetic properties of the voiced retroflex fricative in Mandarin, which, despite being classified as a fricative, has characteristics that are more akin to an approximant (Chen & Mok, 2021). This gives rise to multiple distinguishing features from the voiceless retroflex fricative /ʂ/, thus facilitating the distinction between the two phonemes. Another relevant factor that may contribute to the exceptional learning of voicing contrast in Mandarin is the structure of the learner model. The model here employed a vector-based representation that flattened time slices into vectors, rather than considering time series sequentially. As a result, the model may have effectively learned to recognize instantaneous and non-sequential phonetic indicators, rather than sequential ones. Since the voicing feature is present continuously and consistently throughout an entire segment, the model could detect it upon encountering even a small slice with voicing.

In contrast, although aspiration is widely contrastive in Mandarin, with all plosives and affricates exhibiting minimal pairs that are contrastive in aspiration, the model's learning outcome was unsatisfactory ($d = 0.78$). If the previously mentioned hypothesis regarding sequential indicators holds true, this outcome could be attributed to the fact that detecting aspiration necessitates the model's processing of sequential elements, such as identifying a voiceless burst immediately followed by silence, which proved to be a challenging task for the model to learn.

Another feature that was poorly acquired in both language models was nasality, which exhibits salient, continuous, and consistent acoustic cues (Johnson, 2011, pp.185-191). One potential explanation for this observation is that nasal consonants spread their nasal features along the temporal sequence (Thompson, 1978). As a result, sounds preceding and following a nasal consonant tend to exhibit nasal features as well. Therefore, the model may have incorrectly learned to attribute this feature to a broad range of sounds, thus across multiple data frames, leading to a false overgeneralization of the nasal feature.

4.3 The role of top-down phonological learning

It is important to note that the differentiation between bottom-up and top-down learning does not necessarily correspond to distinct developmental stages in speech acquisition. In reality, infants have access to both phonetic and distributional cues of sounds from the beginning. However, the current experiment exclusively analyzed the performance of a model that underwent bottom-up learning only, without using distributional information. The purpose of the current exercise was to accurately assess the scope of learning achieved through a pure bottom-up approach, rather than to simulate the precise learning processes that infants undergo. The outcomes demonstrate the emergence of perceptual magnets, although they may differ from the distributional protocategories suggested by Hayes (2004).

It appears that the clustering in the model was based on phones rather than phonemes, despite being referred to as phonemes. For instance, the three allophones of Mandarin "i" (/i, ɨ, i̞/) (Zee & Lee, 2001) were projected to separate regions in the hidden space, and the distance between them was relatively large. The model's inability to cluster allophones more closely together than phonemes may indicate the boundary between the contributions of bottom-up and top-down information. Learners acquire a fundamental, natural, cross-linguistic hidden representation space from bottom-up information, but it is essential to use top-down information to transform this space to reflect paradigmatic relationships (such as contrastiveness and allophony) and co-occurrence restrictions (Kolachina & Magyar, 2019).

4.4 Contributions to the existing modeling results

Previous research has utilized machine learning models to investigate the acquisition of speech sounds. Notably, two streams of research have emerged in the literature. The first stream of research comprises a series of studies conducted by Beguš and colleagues. In these studies, a waveGAN model, a Deep Convolutional GAN architecture for audio data, and its modified versions were used in an unsupervised manner, and the models were applied to various linguistic tasks using different data and varying model components. In the initial work of this series, Beguš (2020b) trained the waveGAN model on a collection of continuous speech clips dedicated to the allophonic variations of plosives before /s/ in English. After training, the model underwent a "probing" test that involved identifying highly correlated latent variables through regression

analysis and manipulating the top variables to values far beyond the normal range to observe the resulting output features. This approach demonstrated that even a few of the 100 latent variables had a significant impact on the presence of /s/ and the resulting allophonic distribution of the succeeding plosive. This discovery indicates that when trained on raw sound input, as in our study, the model can effectively capture various sound features using distinct latent variables, and successfully encode phonological knowledge within them.

Following work in this series demonstrated the effective acquisition and representation of reduplication using categorical InfoWaveGAN through two specialized categorical latent variables that learn to control the presence of this linguistic phenomenon in an unsupervised manner (Beguš, 2021). Another study in the series compared the ability of GAN and human participants to acquire both local and non-local alternations through parallel experiments (Beguš, 2022). The results indicated that the GAN model accurately captured the relative difficulty of learning non-local dependencies compared to local alternations, which was consistent with human performance. Furthermore, an experiment that examined both GAN latent variables and brain stem signals found substantial similarities between the GAN model and human participants (Beguš et al., 2022b). The series of studies also explored iterative learning (Beguš, 2020a) and articulation acquisition (Beguš et al., 2022a), yielding results consistent with those of human experiments. These findings, overall, suggest that machine learning models based on raw sound data, such as waveGAN and its modified versions, can successfully simulate the acquisition of sounds, including their phonological features.

The approach taken in the current paper differs from the series of studies by Beguš in several aspects. First, the primary focus of this paper was at the segment level, with contextual information being excluded as much as possible. In contrast, Beguš's work primarily focused on the acquisition of higher-level phonological knowledge including phonological contextual information. Therefore, in that sense, the two lines of sound acquisition simulation based on raw data address different phonological acquisition stages and, thus, complement each other. Additionally, we assumed that sounds are encoded into a low-dimensional continuous space and use the trained encoder to encode sounds into this hidden space. We then observed the distributional properties of the sounds' hidden representations. In essence, our experiment aimed to replicate the process by which human learners perceive sounds from natural language input and subsequently develop a mental grammar that incorporates phonetic categories. In contrast, Beguš' approach involves "probing", which includes changing latent variables and observing corresponding outputs to diagnose the functions of

different latent variables. The training of the waveGAN model is thus similar to the way human learners use feedback and correction to adjust their own productions until they reach a level of accuracy and resemblance to the target forms, the approach which is different from the current study.

Another stream of work primarily focused on symbolized segment tokens (represented using text), with the learning target including phonemes (Kolachina & Magyar, 2019; Silfverberg et al., 2018; Sofroniev & Çöltekin, 2018), alternation rules (Chandlee & Jardine, 2021; Hua et al., 2021; Mamadou & Jardine, 2020), phonotactics (Magri, 2011), and articulation (Smith et al., 2021). However, these top-down approaches tended to focus more on the distributional properties of sounds, often neglecting their physical properties.

To conclude, the present study has demonstrated that an autoencoder model can to a certain extent acquire knowledge of phonological features from unsegmented and non-transcribed sound data. The acquired knowledge reflects the fundamental, universal properties of sounds across languages, while language-specific knowledge about paradigmatic relationships and co-occurrence restrictions was not learned. The current study has also confirmed that differences in feature storage alone cannot account for the acquisition of language-specific phonological knowledge. Furthermore, the model learned various distinctive features with varying degrees of success, indicating that the model's hidden space distribution can provide varying degrees of contrast for different features. However, it remains to be explored the extent to which poorly learned features reflect the natural properties of sounds that make them difficult to distinguish and how exactly the model's availability to distinguish different features is compatible with that of human learners at the early stages of phonological acquisition.

Overall, this study contributes to our understanding of the mechanisms underlying the acquisition of phonological features and provides a foundation for further research into the use of autoencoder models for phonological analysis. Further studies using more diverse language datasets could provide additional insights into the effectiveness of autoencoder models for feature learning and the extent to which the features learned by such models reflect natural properties of sound acquisition.

References

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362. https://doi.org/10.1016/j.jml.2004.12.003

Bank, D., Koenigstein, N., & Giryes, R. (2021). *Autoencoders* (arXiv:2003.05991). arXiv. http://arxiv.org/abs/2003.05991

Beguš, G. (2020a). *Deep Sound Change: Deep and Iterative Learning, Convolutional Neural Networks, and Language Change*. https://doi.org/10.48550/ARXIV.2011.05463

Beguš, G. (2020b). Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning With Neural Networks. *Frontiers in Artificial Intelligence*, *3*, 44. https://doi.org/10.3389/frai.2020.00044

Beguš, G. (2021). Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication. *Transactions of the Association for Computational Linguistics*, *9*, 1180–1196. https://doi.org/10.1162/tacl_a_00421

Beguš, G. (2022). Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Computer Speech & Language*, *71*, 101244. https://doi.org/10.1016/j.csl.2021.101244

Beguš, G., Zhou, A., Wu, P., & Anumanchipalli, G. K. (2022a). *Articulation GAN: Unsupervised modeling of articulatory learning*. https://doi.org/10.48550/ARXIV.2210.15173

Beguš, G., Zhou, A., & Zhao, T. C. (2022b). *Encoding of speech in convolutional layers and the brain stem based on language experience* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.01.03.474864

Chandlee, J., & Jardine, A. (2021). Computational universals in linguistic theory: Using recursive programs for phonological analysis. *Language*, *93*(3), 485–519.

Chen, S., & Mok, P. P. K. (2021). Articulatory and Acoustic Features of Mandarin /ɹ/: A Preliminary Study. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362070

Deterding, D., & Nolan, F. (2007). Aspiration and voicing of Chinese and English plosives. *Proceedings of the 16th International Congress of Phonetic Sciences*, 385–388.

Dow, F. D. M. (1972). *An outline of Mandarin phonetics* (2. ed). Australian National Univ. Press.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*(1–2), 1–42. https://doi.org/10.1016/S0378-5955(01)00259-3

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech Perception in Infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Geisler, C. D. (1998). *From sound to synapse: Physiology of the mammalian ear*. Oxford University Press.

Giegerich, H. J. (1992). *English phonology: An introduction*. Cambridge University Press.

Gilkerson, J. (2005). *Categorical perception of natural and unnatural categories: Evidence for innate category boundaries*.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, *100*(2), 1111–1121. https://doi.org/10.1121/1.416296

Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in Phonological Acquisition* (pp. 158–203). Cambridge University Press.

Hayes, B. (2009). *Introductory phonology*. Wiley-Blackwell.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hirschberg, J. B., & Rosenberg, A. (2007). *V-Measure: A conditional entropy-based external cluster evaluation*. https://doi.org/10.7916/D80V8N84

Hua, W., Dai, H., & Jardine, A. (2021). Learning Underlying Representations and Input-Strictly-Local Functions. In D. K. E. Reisinger & M. Huijsmans (Eds.), *Proceedings of the 37th West Coast Conference on Formal Linguistics* (pp. 143–151). Cascadilla Proceedings Project.

İleri, S., KarabiNa, A., & Kiliç, E. (2018). Comparison of Different Normalization Techniques on Speakers' Gender Detection. *Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi*, *2*(2), 1–12. https://doi.org/10.31200/makuubd.410625

Johnson, K. (2011). *Acoustic and Auditory Phonetics* (3rd Edition). Wiley.

Kamper, H., Elsner, M., Jansen, A., & Goldwater, S. (2015). Unsupervised neural network based feature extraction using weak top-down constraints. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5818–5822. https://doi.org/10.1109/ICASSP.2015.7179087

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. https://doi.org/10.48550/ARXIV.1412.6980

Kolachina, S., & Magyar, L. (2019). What do phone embeddings learn about Phonology? *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 160–169. https://doi.org/10.18653/v1/W19-4219

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Lee, W.-S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, *33*(1), 109–112. https://doi.org/10.1017/S0025100303001208

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, *18*(2), 201–212. https://doi.org/10.1016/0022-0965(74)90101-5

Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge University Press.

Liu, G., Bao, H., & Han, B. (2018). A Stacked Autoencoder-Based Deep Neural Network for Achieving Gearbox Fault Diagnosis. *Mathematical Problems in Engineering*, *2018*, 1–10. https://doi.org/10.1155/2018/5105709

Liu, B., & Liang, Y. (2019). *Optimal Function Approximation with Relu Neural Networks*. https://doi.org/10.48550/ARXIV.1909.03731

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lyons, J. (2013). *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. Practical Cryptography. http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Magri, G. (2011). An online model of the acquisition of phonotactics within Optimality Theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning Phonemes With a Proto-Lexicon. *Cognitive Science*, *37*(1), 103–124. https://doi.org/10.1111/j.1551-6709.2012.01267.x

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*, *2*, 522–533.

Maye, J., & Gerken, L. (2001). Learning Phonemes: How Far Can the Input Take Us? In *Proceedings of the 25th Annual Boston University Conference on Language Development, Vols. 1-2* (pp. 480–490). Cascadilla Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, *98*, 77–92. https://doi.org/10.1016/j.jml.2017.09.005

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *The Journal of the Acoustical Society of America*, *110*(3), 1598–1605. https://doi.org/10.1121/1.1379078

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). *Automatic differentiation in PyTorch*.

Peperkamp, S., Pettinato, M. & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In: B. Beachley, A. Brown & F. Conlin (eds.) *Proceedings of the 27th Annual Boston University Conference on Language Development*. Volume 2. Sommerville, MA: Cascadilla Press, 650-661.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)* [Data set]. Columbus, OH: Department of Psychology, Ohio State University (Distributor). www.buckeyecorpus.osu.edu

Räsänen, O. (2014). Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. *36th Annual Conference of the Cognitive Science Society, Quebec, Canada, July*.

Shain, C., & Elsner, M. (2019). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 69–85. https://doi.org/10.18653/v1/N19-1007

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Shi, Y., Bu, H., Xu, X., Zhang, S., & Li, M. (2021). *AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines* (arXiv:2010.11567). arXiv. http://arxiv.org/abs/2010.11567

Silfverberg, M., Mao, L. J., & Hulden, M. (2018). Sound analogies with phoneme embeddings. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, 136–144.

Smith, C., O'Hara, C., Rosen, E., & Smolensky, P. (2021). *Emergent Gestural Scores in a Recurrent Neural Network Model of Vowel Harmony*. https://doi.org/10.7275/QYEY-4J04

Sofroniev, P., & Çöltekin, Ç. (2018). Phonetic Vector Representations for Sound Sequence Alignment. *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 111–116. https://doi.org/10.18653/v1/W18-5812

Thompson, A. E. (1978). *Nasal air flow during normal speech production*.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392–393. https://doi.org/10.1126/science.167.3917.392

Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C., & Tian, Q. (2004). HMM-Based Audio Keyword Generation. In K. Aizawa, Y. Nakamura, & S. Satoh (Eds.), *Advances in Multimedia Information Processing—PCM 2004* (Vol. 3333, pp. 566–574). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30543-9_71

Yost, W., A. (2007). Perceiving sounds in the real world: An introduction to human complex sound perception. *Frontiers in Bioscience*, *12*(8–12), 3461. https://doi.org/10.2741/2326

Zee, E., & Lee, W.-S. (2001). An acoustical analysis of the vowels in beijing Mandarin. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 643–646. https://doi.org/10.21437/Eurospeech.2001-169

Zhu, J., Zhang, C., & Jurgens, D. (2022). Phone-to-Audio Alignment without Text: A Semi-Supervised Approach. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8167–8171. https://doi.org/10.1109/ICASSP43922.2022.9746112