Roles of Predictability and Acoustic Distance in Sound Discrimination via Contrastive Learning

Shuhao Zhang and Youngah Do
The University of Hong Kong
shuhaoz19@connect.hku.hk, youngah@hku.hk

Abstract

Research in sound discrimination demonstrates that listeners exhibit reduced sensitivity to acoustic differences between allophones, as opposed to phonemes. Previous studies indicates that highly predictable, complementary distribution of allophones contributes to this limited sensitivity by providing strong contextual cues. Building on these insights, this study investigates the role of predictability in sound discrimination within a supervised contrastive learning framework. Specifically, we examine how varying levels of predictability affect the ability to distinguish sounds and whether this influence is categorical or gradual. Additionally, we explore the interaction between acoustic distance and predictability, as well as how the presence of other contrasts within a language modulates this process. Our findings indicate that only full predictability leads to a significant decline in discrimination performance, demonstrating a categorical effect. This impairment can be alleviated as acoustic distance increases. Moreover, the presence of additional contrasts sharing the relevant acoustic dimension enhances discriminability, showing the importance of contextual contrasts in speech perception.

1 Introduction

When discriminating between different sounds, studies have shown that listeners perceive allophones within their native language as less distinct than phonemes, and this phenomenon emerges early in language development (Whalen et al., 1997; Pegg and Werker, 1997; Peperkamp et al., 2006; Boomershine et al., 2008; Seidl et al., 2009; Seidl and Cristia, 2012). For instance, in Boomershine et al. (2008), both Spanish-speaking and English-speaking listeners were asked to rate the similarity of sound pairs that differ in phonological status across the two languages (e.g., /d/ and /ð/) . In English, these are separate phonemes, whereas in Spanish, they are allophones in complementary environments. The results reveal that Spanish-speaking listeners rated these complementarily distributed sound pairs as more similar than English-speaking listeners did, indicating that listeners are less sensitive to allophonic differences than to phonemic distinctions, even when the acoustic distances between the sounds are comparable.

One possible explanation for this perceptual difference is that allophones generally occur in highly predictable, complementary environments, while phonemes can appear in overlapping contexts. Prior research suggests that contextual cues, especially complementary distributions, increase the predictability of sounds and thus attenuate individuals' sensitivity to acoustic differences (Peperkamp et al., 2003; Noguchi and Kam, 2018; Barrios et al., 2023). For example, Peperkamp et al. (2003) exposed participants to a continuum from [ʁ] to [χ], with stimuli presented in different distributions of occurrence (bimodal vs. unimodal distributions) and varying contextual conditions (overlapping vs. complementary conditions). The results of the post-exposure discrimination test suggest that, despite bimodal exposure, the high predictability of speech sounds from complementary contexts led to listeners' reduced sensitivity to the acoustic differences, compared to overlapping contexts. These studies indicate that increased predictability from complementary contexts reduces listeners' sensitivity to target sounds by enabling reliance on contextual cues (Peperkamp et al., 2003; Noguchi and Kam, 2018).

In natural languages, however, the predictability of sounds is not always defined in absolute terms. Instead, it is often gradient, with certain sounds more likely to occur in specific phonological environments than others. To systematically explore the role of predictability, beyond just the two extremes, the current study investigates how varying degrees of predictability influence discrimination performance along a continuum.

In addition to the influence of complementary contexts on sound discrimination, acoustic distance within such distributions plays a key role in perception. Allophones tend to share similar acoustic properties (Peperkamp et al., 2006; Skoruppa et al., 2011) and generally exhibit smaller acoustic distances than phonemes (Seidl and Cristia, 2012). When two sounds are acoustically very distant, they are thus less likely to be perceived as allophones of the same phoneme, even within complementary contexts (Peperkamp et al., 2006).

Building upon these backgrounds and specifically extending the work of Peperkamp et al. (2003), the current study aims to compare discrimination performance of sounds across different levels of predictability, while systematically varying the acoustic distances within bimodal distributions. Our hypothesis is that exposure to the complementary distribution of two sounds (100% predictability) will lead to reduced sensitivity to acoustic differences; however, large acoustic distances may attenuate this effect even within complementary contexts. To capture the gradient nature of acoustic differences, we examine multiple sound pairs with graded acoustic distances, allowing us to observe how discrimination performance varies along a continuum of acoustic similarity.

Our study design thus addresses two main questions:

(a) Does full predictability (i.e., complementary context) lead to reduced sensitivity? If so, is the effect of predictability categorical (i.e., reflecting phoneme-like versus allophone-like perception), or does it change gradually along with predictability?

(b) How does the influence of contextual cues on sound discrimination change across varying levels of acoustic distances?

To investigate these questions, we employ a supervised contrastive learning framework (Khosla et al., 2020). The model is trained to distinguish between different CV syllables for each condition over 200 epochs, simulating human perception of sound pairs. The trained encoder is then used to evaluate the discrimination performance on a test set. By analyzing the latent feature representations using the silhouette score and the distance distribution overlap (DDO) score, we quantify how well the two categories in the test set are separated in the feature space.

2 Experiment 1

Experiment 1 examines how predictability influences the discrimination of speech sounds of different acoustic distance. Within a contrastive learning framework, sound pairs are presented with varying acoustic similarity, represented by acoustic vectors, within different contextual conditions to modulate predictability. Discrimination performance is assessed by evaluating the silhouette scores and DDO scores of the sounds embedded within the same context, providing insights into how the two factors interact to influence perceptual separation.

2.1 Method

2.1.1 Dataset

The stimuli used are CV syllables, consisting of eleven fricatives differing primarily in their center of gravity (CoG). These form ten pairs of sounds, with a shared anchor sound (labeled as /s_0/) and a variable second sound (/s_1/ to /s_{10}/, referred to as /s_x/). The ten pairs are labeled as pair 1 through pair 10.

The mean CoG of /s_0/ is 9000 Hz. For each /s_x/, the mean CoG decreases in steps of 500 Hz, starting from 8000 Hz for /s_1/ down to 3500 Hz for /s_{10}/. The distribution of each sound along the CoG dimension follows a truncated Gaussian distribution spanning $[\text{mean} - 2 \times \text{sd}, \text{mean} + 2 \times \text{sd}]$, with a standard deviation $\text{sd} = 500$ Hz. For instance, /s_1/ ranges from 7000 Hz to 9000 Hz. Consequently, in the design, the two sounds in pair 1 and pair 2 have overlapping values along the CoG dimension, respectively. The truncated gaussian distribution is selected to maintain control over the scope while approximating realistic distributional variability.

In the training set, we manipulate the distribution of target sounds to create a gradient of predictability levels, including 0%, 25%, 50%, 75%, and 100%. This is achieved by varying the proportion of complementary contexts, using eight vowels (/i/, /ɪ/, /e/, /ɛ/, /u/, /ʊ/, /o/, /ɔ/) as contexts to form CV syllables. For each pair of sounds (/s_0/ and /s_x/), predictability is determined by the proportion of stimuli with complementary contexts. For example, in the 75% predictability condition, vowels /i, ɪ, e/ are only preceded by /s_0/, while /u, ʊ, o/ only follow /s_x/. Vowels /ɛ, ɔ/ are shared between the two fricatives. The number of tokens involving each vowel is equal, resulting in stimuli with complementary contexts constituting 75% of all stimuli

in this condition.

In the test set, both sounds in each pair are combined with the following vowel /ɑ/. All nine vowels (eight in the training set and one in the test set) only differ in their first three formant frequencies (F1 to F3). The values are based on measurements from 48 female American English speakers by Hillenbrand et al. (1994).

Thus, this creates 50 conditions in total, varying in predictability (five levels ranging from 0% to 100%) and acoustic distance (ten pairs). Each condition has 1,000 tokens per vowel for the training set, totaling 8,000 tokens per condition.

2.1.2 Input manipulation

Stimuli are represented as acoustic vectors: each sound as a 16-dimensional feature vector, with pairs of sounds combined into a 2×16 vector representing a CV syllable. The 16 dimensions of each sound include: center of gravity, frication duration, standard deviation, skewness, kurtosis, frication intensity, total duration, vocalic duration, vocalic intensity, fundamental frequency (F0), and frequencies and bandwidths of the first three formants (F1, F2, F3). The structure is shown in Table 1.

For fricatives, the last nine features have a zero value, while vowels are assigned zero values for the first six features. Most features follow a truncated Gaussian distribution ranging between $[\text{mean} - 2 \times \text{sd}, \text{mean} + 2 \times \text{sd}]$, with a standard deviation set to 5% of the mean value for each (except for CoG, whose standard deviation is 500 Hz). Total duration and vocalic duration are constant. The mean values of all components, other than the formant frequencies and CoG values, are assigned to be consistent across different syllables, with reference to relevant phonetic studies (e.g., de Cheveigné, 1999; Jongman et al., 2000; McMurray and Jongman, 2011; Li and Gu, 2015). Sample vector values for a token of /sɔi/ are shown in Table 1.

2.1.3 Model structure and loss function

The core encoder is a fully connected neural network comprising a flattening layer, followed by two linear layers with ReLU activations, which compresses input vectors into a four-dimensional feature space.

The training optimization employs a supervised contrastive loss based on normalized temperature-scaled similarity. A typical function of supervised contrastive learning is illustrated in Equation 1 (Khosla et al., 2020).

$$L = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z_i} \cdot \boldsymbol{z_p}/\tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z_i} \cdot \boldsymbol{z_a}/\tau\right)} \tag{1}$$

In this formulation, i denotes a specific token in the batch I, and $\boldsymbol{z_i}$ is its corresponding feature vector. The symbol $\cdot$ represents the inner (dot) product. The set $A(i)$ includes all samples in the batch I except for the token i, while $P(i)$ comprises all positive samples for i within $A(i)$, with $|P(i)|$ indicating its cardinality. τ is a scalar temperature parameter.

In our implementation, feature vectors are first normalized to unit length before similarity computation. The similarity matrix is then derived for all pairs within the batch. To improve numerical stability, the maximum similarity value in each row, corresponding to the similarity of the token i with itself, is subtracted from all entries in that row. A mask tensor is employed to distinguish positive pairs from negative pairs, while excluding self-similarities to prevent trivial solutions. To prevent division by zero, $|P(i)|$ is set to 1 when no positive samples are available for token i in this batch. The overall loss for the batch is calculated by averaging the individual losses of each sample. The loss function used in this study is shown in Equation 2.

$$\begin{aligned} Loss &= \frac{1}{N} \sum_{i \in I} \frac{-1}{D_i} \sum_{p \in P(i)} \log \frac{\exp S_{i,p}}{\sum_{a \in A(i)} \exp S_{i,a}} \\ S_{i,j} &= \text{sim}(\boldsymbol{z_i}, \boldsymbol{z_j})/\tau - \text{sim}(\boldsymbol{z_i}, \boldsymbol{z_i})/\tau \\ D_i &= \begin{cases} 1, & \text{if } |P(i)| = 0 \\ |P(i)|, & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

In Equation 2, N is the number of samples in the batch I. The function $\text{sim}(\boldsymbol{z_i}, \boldsymbol{z_j})$ calculates the cosine similarity between the vectors. The temperature parameter τ is set to 0.07 in this study. This loss function promotes embeddings of samples from the same class to cluster together, while pushing apart embeddings from different classes, aligning with recent contrastive learning methodologies (e.g., Chen et al., 2020; Khosla et al., 2020; Gao et al., 2022; Liao et al., 2023).

No.	Element	Explanation	Distribution	/s$_0$/	/i/
1	cog	Center of gravity	Gaussian	8847.59	0.00
2	fri_dur	Frication duration	Gaussian	178.82	0.00
3	sta_dev	Standard deviation	Gaussian	180.72	0.00
4	skewness	Skewness	Gaussian	0.51	0.00
5	kurtosis	Kurtosis	Gaussian	0.92	0.00
6	fri_int	Frication intensity	Gaussian	61.39	0.00
7	tot_dur	Total duration	Fixed	200.00	400.00
8	voc_dur	Vocalic duration	Fixed	0.00	400.00
9	voc_int	Vocalic intensity	Gaussian	0.00	79.72
10	f0	Fundamental frequency	Gaussian	0.00	190.57
11	b3	Bandwidth of F3	Gaussian	0.00	169.54
12	b2	Bandwidth of F2	Gaussian	0.00	114.70
13	b1	Bandwidth of F1	Gaussian	0.00	86.99
14	f3	Frequency of F3	Gaussian	0.00	3335.26
15	f2	Frequency of F2	Gaussian	0.00	2911.90
16	f1	Frequency of F1	Gaussian	0.00	407.04

Table 1: Vector components and sample vector values for a token of /s$_0$i/

2.1.4 Training and evaluation

The model is trained for 200 epochs using batches of size 32, with a fixed learning rate of 0.001. During training, the model of each condition learns distinguish between different CV syllables, under the contrastive loss described earlier. After training, the encoder generates the four-dimensional feature embeddings for test data. Discrimination performance of the model is quantified using two metrics: the silhouette score and the DDO score. The silhouette score measures how well the two classes are separated in the embedding space, with higher values indicating greater separation and, consequently, better discrimination. The DDO score quantifies the overlap between the probability density functions of all pairwise within-class and between-class distances, with lower scores signifying improved discrimination.

To mitigate the effects of randomness, each condition is repeated ten times with independently initialized models, providing more robust estimates of performance for subsequent statistical analysis.

2.1.5 Analysis

For each step, a linear regression model is utilized to examine the effect of predictability on the response variable, the silhouette score or the DDO scores. Following this, post-hoc pairwise comparisons using Tukey's Honestly Significant Difference (HSD) test are performed to assess differences in the scores across various levels of predictability.

2.2 Results

The silhouette scores for each condition in Experiment 1 are displayed in Figure 1. The y-axis depicts the silhouette score, while the x-axis indicates the step of the sound being compared (i.e., /s$_x$/).

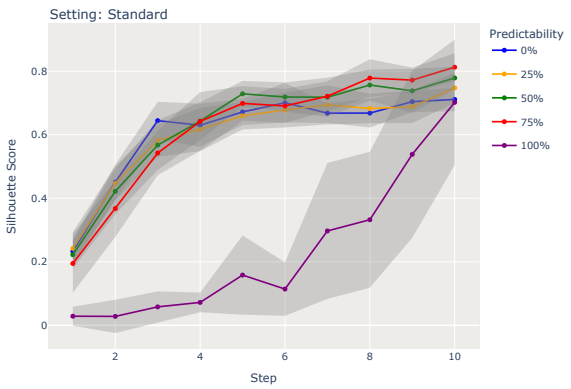

Figure 1: Silhouette scores for each condition in Experiment 1

Regarding the role of predictability, the results show that 100% predictability significantly leads to worse performance of sound discrimination compared to all other predictability levels at steps 1 to 8 (all p values < .0001). In contrast, the differences among the other predictability levels are not statistically significant (all p values > .05) across all acoustic distance conditions (i.e., steps).

There is a consistent trend that larger acoustic distances tend to correspond to higher silhouette score. When examining the interaction between

predictability and acoustic distance, it is notable that the influence of 100% predictability appears to diminish as the acoustic distance increases. Specifically, in conditions with large acoustic distances, there is limited or no significant difference in silhouette scores between the 100% predictability condition and other predictability levels. For example, at step 9, the difference in silhouette scores between the 100% and 25% predictability levels is not statistically significant (p = .077), and at step 10, no significant difference is observed between the 100% predictability condition and the other levels (all p values > .05). This suggests that salient acoustic distance can attenuate the influence of predictability on sound discrimination.

The DDO scores for each condition are shown in Figure 2. These scores provide similar information to the silhouette scores. However, even at step 10, 100% predictability does not achieve the same performance as other predictability levels (all p values < .01), although the differences are substantially reduced.

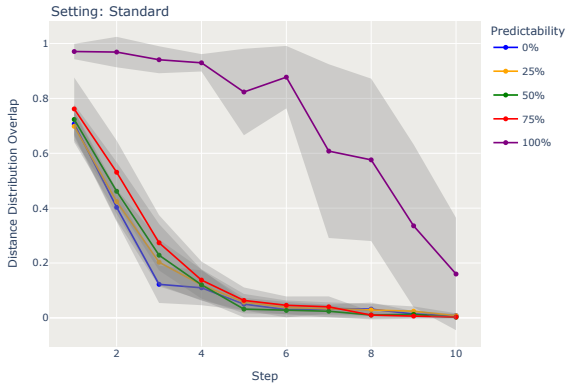

Figure 2: Distance distribution overlap (DDO) scores for each condition in Experiment 1

Overall, these findings indicate that predictability from contextual cues can influence sound perception, with full predictability reducing the sensitivity to the acoustic differences. The effect of predictability appears to be categorical, affecting performance only under complete predictability conditions. Meanwhile, acoustic distance gradually enhances discrimination performance, ultimately offsetting the influence from full predictability when the acoustic distance is sufficiently large.

3 Experiment 2

In Experiment 1, we have determined whether the effect of predictability from contextual cues is categorical or gradual and how this influence varies across different acoustic distances, focusing on a single contrast. However, in natural languages, contrasts are rarely learned in isolation. Instead, acoustic distinctions from other contrasts may also affect the perception of a new pair of sounds. To explore this, we conduct additional series of experiments to examine how acoustic distinctions from other sound contrasts impact the discrimination of the target sounds.

3.1 Method

Building on Experiment 1, we introduce an additional pair of sounds with overlapping contexts in training, where the predictability for this pair is set to 0%. For comparison, we train three sets of models, each incorporating a new sound pair with distinct levels in the CoG difference:

(1) A pair showing no difference in CoG, labeled as /z$_x$/ and /ts$_x$/, with their CoG values matching the second sound /s$_x$/ in the target sound pair for each condition.

(2) A pair exhibiting a fixed larger difference in CoG, labeled as /z$_0$/ and /z$_8$/, as their CoG values resemble those of /s$_0$/ and /s$_8$/.

(3) A pair with a fixed smaller difference in CoG, labeled as /z$_0$/ and /z$_5$/, as their CoG values are similar to those of /s$_0$/ and /s$_5$/.

Among them, /z$_0$/ and /z$_x$/ can be considered voicing counterparts to the target sounds /s$_0$/ and /s$_x$/, respectively, while /ts$_x$/ serves as an affricate counterpart of /s$_x$/. For simplicity, the three settings of models are designated as 'Unrelated', 'Long-Distance', and 'Short-Distance', respectively. The setting from Experiment 1 is referred to as 'Standard'. In Experiment 2, the training set is expanded by doubling the total number of tokens with an additional contrast, resulting in 16,000 tokens in each condition.

Regarding the input vectors, voicing differences are achieved by changing the vocalic duration of the consonant, whereas affricate differences are manipulated by altering the frication duration of the consonant. Specifically, the vocalic duration for voiceless fricatives is set to 0 ms, while for voiced counterparts it is fixed at 200 ms. Frication duration is 174 ms for fricatives, while it is 96 ms for affricates, both sampled from truncated gaussian distributions. The values for each component are

assigned based on established phonetic research (Jongman et al., 2000; Lee, 2011; Li and Gu, 2015). Other components are manipulated similarly to the target sounds in Experiment 1. The model structure, loss function, and training and evaluation procedures remain consistent throughout, except that only the silhouette scores are retained for more concise comparisons.

The same analysis pipeline used in Experiment 1 is applied to the results from each of the three model settings. Additionally, we compare the outcomes across the four settings based on their predictability, especially 100% predictability. For each step within each predictability level, we utilize a linear regression model to examine the effect of different experimental settings on the silhouette scores. Pairwise comparisons are conducted to assess the differences between various settings.

3.2 Results

The silhouette scores for each condition in the three settings of Experiment 2 are presented in Figure 3.

The results across all three settings suggest that 100% predictability tends to have a distinct impact compared to other predictability levels. In the 'Unrelated' setting, 100% predictability is significantly different from the other predictability levels from steps 1 to 9 with lower silhouette scores (except for the comparison with 75% predictability at step 1, where $p = 1$), indicating poorer discrimination. In the 'Long-Distance' setting, this significance is observed from steps 1 to 7 (except for the comparison with 75% predictability at step 6, where $p = .0824$), while it spans from steps 1 to 5 for the 'Short-Distance' setting. Aside from 100% predictability, other predictability levels show little or no difference at each step for all three settings.

In general, settings with another sound pair exhibit similar trends to the 'Standard' setting. However, a key difference lies in the influence of the 100% predictability. Figure 4 shows the silhouette scores for 100% predictability across the four settings at each step. Throughout all steps, the 'Standard' setting shows no significant difference from the 'Unrelated' setting. The 'Short-Distance' setting demonstrates significantly better discrimination than the 'Long-Distance' setting only at steps 3 ($p = .0002$) and 4 ($p = .0157$), with no other significant pairwise differences observed.

From steps 1 to 8, 'Standard' and 'Unrelated' settings generally show lower silhouette scores compared to the 'Long-Distance' and 'Short-Distance'

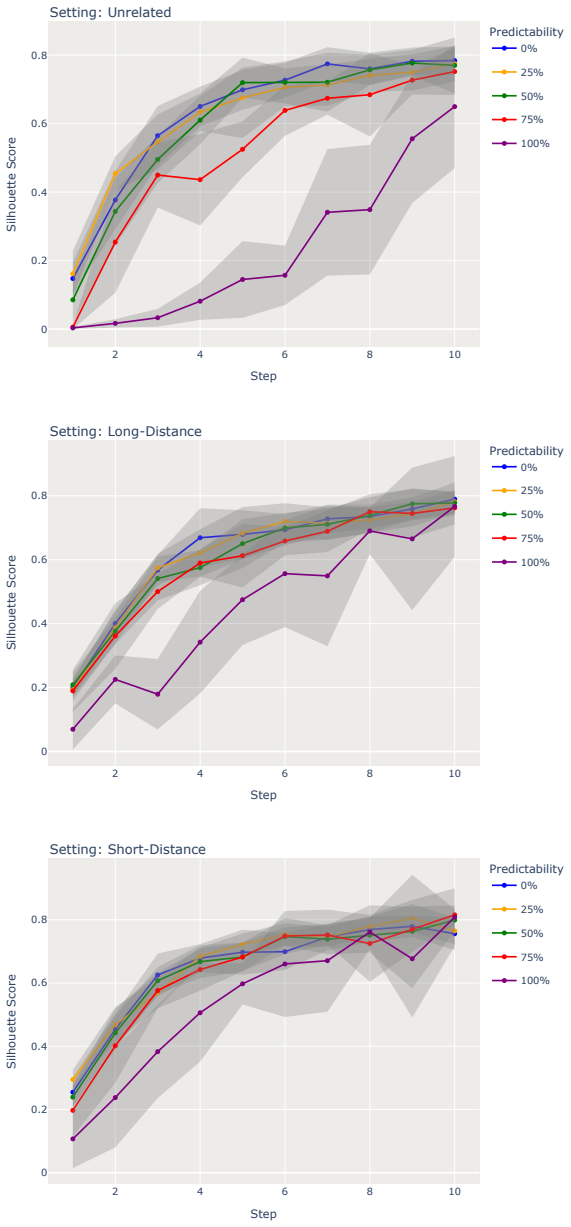

Figure 3: Silhouette scores for each condition in three model settings of Experiment 2

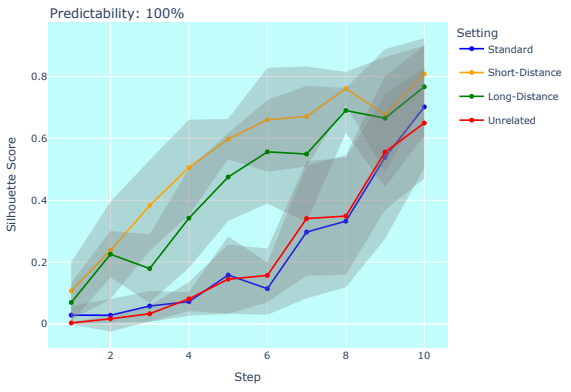

Figure 4: Silhouette scores for 100% predictability in different model settings

settings, with limited exceptions. However, this difference diminishes, when acoustic distance reaches steps 9 and 10. These findings suggest that another contrast sharing acoustic features in the target dimension facilitates the discrimination of a new sound pair, even when predictability of the target sounds is high and their acoustic distance is not very large. However, if this distinction cannot be learned from other sound contrasts, this facilitative effect is absent, resulting in discrimination performance comparable to that observed when learning in isolation. This underscores the essential role of shared acoustic features in supporting perceptual discrimination, particularly under conditions of high predictability and limited acoustic distance between target sounds.

4 Discussion

4.1 Role of predictability

The results reveal that discrimination performance under full predictability (100%) is significantly poorer compared to other levels of predictability, which do not differ significantly among themselves. This suggests that the influence of contextual cues on sound discrimination operates in a categorical manner: only when predictability reaches its maximum does perceptual sensitivity diminish noticeably. This phenomenon aligns with previous behavioral studies proposing that listeners treat sounds with fully predictable contexts (complementary distribution) as allophones, leading to reduced sensitivity to the acoustic differences, while phonemes remain encoded as contrastive (Peperkamp et al., 2003; Noguchi and Kam, 2018).

However, this categorical effect is not absolute. For example, in the 'Unrelated' setting, 75% predictability also shows significant differences from lower predictable levels at some steps (e.g., the pairwise comparison of 75% predictability and 50% predictability at step 4, with $p = .001$). This indicates that near-complete predictability might also influence discrimination, albeit with less consistency. A potential explanation for this general categorical manner is that the model's capacity or the task's simplicity may inherently limit the detection of subtle gradual changes in non-full predictability levels. Specifically, the model might tend to learn the contrast even when only a small amount of evidence is present showing that they are not fully predictable, thus the model separates the two sounds in its latent representations. As contrastive learning benefits more from longer training steps, later in training, the model is exposed to more negative pairs, thus more ready to learn the contrast even when overlapping contextual cues are limited (Chen et al., 2020).

Supporting this, analysis of early training phases for a closely paired contrast (first 10 epochs at step 3, under the 'Standard' setting) shows some gradual differences in the silhouette scores at lower predictability levels (see Figure 5). This suggests that the effect of predictability may have a gradual component that is hidden by the model's tendency to perform categorically at later stages. In other words, during early training, the model registers subtle gradations, but as training proceeds, it shifts to making a more binary, categorical distinction.

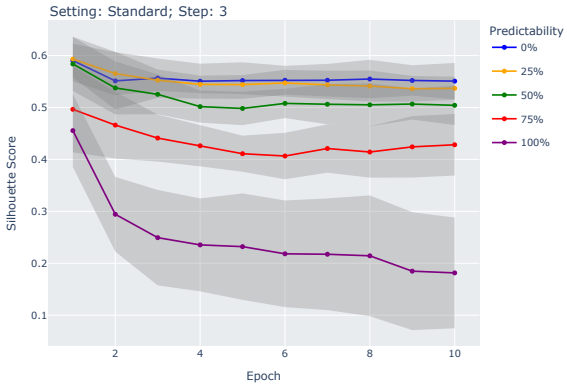

Figure 5: Silhouette scores across different epochs at step 3 in the 'Standard' setting

4.2 Role of acoustic distance and its interaction with predictability

Across all conditions, larger acoustic distances are associated with improved discrimination performance. Importantly, as acoustic distance increases, its influence begins to outweigh the effect of predictability, indicating that acoustic effects can mitigate context-driven perceptual attenuation.

Specifically, for the predictability condition that impacts discrimination the most (i.e., 100% predictability), the influence appears to change gradually along the acoustic distance. Rather than a sharp threshold, increasing acoustic differences lead to a progressive gain in the discrimination performance, constrained by the salience of acoustic cues. This implies that perceptual sensitivity is modulated by an incremental weighting of cue salience, where the effect of contextual predictability can be eventually overcome by salient acoustic cues, also sup-

porting that sound perception comprises utilization and interaction of multiple cues (Norris, 2003; Davis et al., 2005; Martin et al., 2013; Fourtassi and Dupoux, 2014; Frank et al., 2014).

4.3 Role of other sound contrasts

The presence of additional contrasts that share the same acoustic dimension enhances discrimination performance under full predictability. This suggests that phonemic or phonological information related to the relevant acoustic dimension does not operate in isolation but functions as a network of cues (Kuhl, 2000), reinforcing perceptual distinctions across contrasts. Moreover, when the acoustic distance of the additional sound pair is smaller, the discrimination accuracy will increase more.

This finding indicates that phonetic information from other contrasts can strengthen the perceptual salience of the target contrast. In other words, related phonetic cues from additional sounds can support and enhance the acoustic distinctions that are important for differentiating the target sounds (Maye et al., 2008; Finley and Badecker, 2009). This support helps to offset the decrease in perceptual sensitivity caused by high predictability, making subtle acoustic differences easier to detect. The effect is particularly prominent when the additional contrast involves a smaller acoustic difference. As a result, even slight acoustic distinctions become clearer because the influence of related contrasts facilitates the ability to discriminate, especially in environments where cues might otherwise be weakened by predictability.

Conversely, when the additional contrast exhibits no acoustic difference in the target dimension, indicating they are not related, it neither facilitates nor inhibits discrimination, highlighting that acoustic interactions are specific to contrasts sharing relevant phonetic features.

5 Limitations and conclusion

Our study utilizes a supervised contrastive learning framework to simulate the perception of sound pairs with different predictability levels and acoustic distances. The input consists of manipulated acoustic vectors with 16 features for each sound, derived from Gaussian distributions. Although this design is grounded in phonetic data and easy to control, it may not fully capture the variability and dynamic properties of natural speech. Additionally, the explicit supervision in the model limits its ability to reflect the complexity of natural language perception. Future work should consider more realistic inputs and improved model designs to enhance generalizability to real-world language perception. Apart from the influence of predictability and acoustic distance, the current study investigates the interactions with other contrasts, only considering one additional contrast for comparison. While the findings can be generalized to situations involving multiple sound pairs, incorporating a full phoneme inventory from a specific language would provide a more comprehensive understanding of how language-specific phonological systems influence perception.

Overall, the present study advances our understanding of how predictability and acoustic distance jointly influence sound discrimination via contrastive learning. The results demonstrate that predictability significantly impairs discrimination performance only when it reaches its maximum, suggesting a categorical influence of contextual cues on perceptual sensitivity. Furthermore, acoustic distance interacts incrementally with predictability, with larger acoustic distinctions counteracting the attenuating effect from full predictability on perception. The presence of related contrasts sharing phonetic features underscores the role of phonological networks in supporting perceptual robustness, especially under high predictability and limited acoustic distinction. These findings reinforce the notion that speech perception involves a complex interplay of multiple cues, where context, acoustic features, and phonological relationships among related sound pairs dynamically shape perceptual outcomes.

References

Shannon L Barrios, Joselyn M Rodriguez, and Taylor Anne Barriuso. 2023. The acquisition of L2 allophonic variants: The role of phonological distribution and lexical cues. *Second Language Research*, 39(3):899–924.

Amanda Boomershine, Kathleen Currie Hall, Elizabeth Hume, and Keith Johnson. 2008. The impact of allophony versus contrast on speech perception. In Peter Avery, B. Elan Dresher, and Keren Rice, editors, *Contrast in Phonology*, pages 145–172. Mouton de Gruyter.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint*. ArXiv:2002.05709 [cs].

Matthew H. Davis, Ingrid S. Johnsrude, Alexis Hervais-Adelman, Karen Taylor, and Carolyn McGettigan. 2005. Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, 134(2):222–241.

Alain de Cheveigné. 1999. Formant bandwidth affects the identification of competing vowels. *ICPhS*, pages 2093–2096.

Sara Finley and William Badecker. 2009. Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3):423–437.

Abdellah Fourtassi and Emmanuel Dupoux. 2014. A Rudimentary Lexicon and Semantics Help Bootstrap Phoneme Acquisition. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 191–200, Ann Arbor, Michigan. Association for Computational Linguistics.

Stella Frank, Naomi H. Feldman, and Sharon Goldwater. 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Baltimore, Maryland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint*. ArXiv:2104.08821 [cs].

James Hillenbrand, Laura A. Getty, Kimberlee Wheeler, and Michael J. Clark. 1994. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 95(5_Supplement):2875–2875.

Allard Jongman, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *arXiv preprint*. Version Number: 5.

Patricia K. Kuhl. 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857.

Sang-Im Lee. 2011. Spectral analysis of Mandarin Chinese sibilant fricatives. In *Proceedings of the 17th international congress of phonetic sciences*, pages 1178–1181, Hong Kong.

Shanpeng Li and Wentao Gu. 2015. Acoustic analysis of Mandarin affricates. In *Interspeech 2015*, pages 1680–1684. ISCA.

Qisheng Liao, Meiting Lai, and Preslav Nakov. 2023. MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 83–87, Toronto, Canada. Association for Computational Linguistics.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning Phonemes With a Proto-Lexicon. *Cognitive Science*, 37(1):103–124.

Jessica Maye, Daniel J. Weiss, and Richard N. Aslin. 2008. Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11(1):122–134.

Bob McMurray and Allard Jongman. 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246.

Masaki Noguchi and Carla L. Hudson Kam. 2018. The Emergence of the Allophonic Perception of Unfamiliar Speech Sounds: The Effects of Contextual Distribution and Phonetic Naturalness. *Language Learning*, 68(1):147–176.

D Norris. 2003. Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238.

Judith E. Pegg and Janet F. Werker. 1997. Adult and infant perception of two English phones. *The Journal of the Acoustical Society of America*, 102(6):3742–3753.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

Sharon Peperkamp, Michèle Pettinato, and Emmanuel Dupoux. 2003. Allophonic Variation and the Acquisition of Phoneme Categories. In *Proceedings of the 27th Annual Boston University Conference on Language Development*, volume 2, pages 650–661, Sommerville, MA. Cascadilla Press.

A. Seidl, A. Cristià, A. Bernard, and K. H. Onishi. 2009. Allophonic and Phonemic Contrasts in Infants' Learning of Sound Patterns. *Language Learning and Development*, 5(3):191–202.

Amanda Seidl and Alejandrina Cristia. 2012. Infants' Learning of Phonological Status. *Frontiers in Psychology*, 3.

Katrin Skoruppa, Anna Lambrechts, and Sharon Peperkamp. 2011. The role of phonetic distance in the acquisition of phonological alternations. In *Proceedings of the 39th Meeting of the North-East Linguistics Society (NELS)*.

D.H. Whalen, Catherine T. Best, and Julia R. Irwin. 1997. Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics*, 25(4):501–528.

A Supplementary materials

All scripts related to data manipulation, model architecture, loss function, and statistical analysis, along with additional resources, are accessible via the Open Science Framework: https://osf.io/krj2a/overview?view_only=d21b3d2a2ee64991a04f9ee456a93639.