# SUMMARY OF PROPOSED DOCTORAL THESIS

## PHONOLOGICAL FEATURES IN THE ERA OF DEEP LEARNING: A MULTI-DIMENSIONAL INVESTIGATION INTO OPTIMAL REPRESENTATIONAL UNITS FOR LANGUAGE MODELING

DOCTORAL THESIS WRITING QUALIFICATION REVIEW

**Sora Nagano**
Graduate School of Arts and Sciences
The University of Tokyo

s-oswld-n@g.ecc.u-tokyo.ac.jp

August 30, 2025

## ABSTRACT

This document presents a comprehensive summary of the proposed doctoral thesis for the writing qualification review. It outlines the research background, objectives, theoretical framework, methodology, preliminary results, and expected contributions of a multi-dimensional investigation into optimal representational units for language modeling in computational phonology.

***K*eywords** Phonology • Deep Learning • Self-Supervised Learning • Neuro-Symbolic Integration • Computational Linguistics

## 1 Introduction and Research Background

The fundamental question of how language should be represented computationally has been at the core of linguistic theory since the inception of generative grammar. In the era of deep learning, this question has gained renewed urgency as self-supervised learning models achieve unprecedented performance on speech and language tasks while operating on representations that diverge dramatically from traditional linguistic units. This doctoral thesis addresses the critical gap between symbolic phonological theories and neural representations by investigating the optimal representational units for language modeling through a multi-dimensional empirical and theoretical framework.

The motivation for this research emerges from a fundamental tension in contemporary computational linguistics. Traditional phonological theory, rooted in the work of Chomsky & Halle (1968) and further developed through Optimality Theory (Prince & Smolensky, 2004), provides elegant symbolic representations—distinctive features, phonemes, and constraints—that capture linguistic generalizations with remarkable parsimony. These representations offer interpretability and theoretical coherence, enabling linguists to formulate precise hypotheses about phonological patterns across languages. The symbolic tradition has produced sophisticated theoretical frameworks including Autosegmental Phonology (Goldsmith, 1976), which revolutionized our understanding of tonal phenomena through multi-tiered representations, and Feature Geometry (Clements, 1985), which captures the hierarchical organization of distinctive features and explains natural classes in phonological processes.

Conversely, modern neural networks, particularly self-supervised models like wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), learn continuous distributed representations directly from raw speech data, achieving superior empirical performance on downstream tasks while remaining largely opaque to linguistic interpretation. These models have demonstrated remarkable capabilities in learning phonetic and phonological patterns without explicit supervision, raising fundamental questions about the nature of linguistic knowledge. The recent development of WavLM (S. Chen et al., 2022) and other advanced SSL models has further pushed the boundaries of what can be learned from raw acoustic signals, achieving near-human performance on various speech tasks while using representations that bear little resemblance to traditional phonological units.

This dichotomy raises profound questions about the nature of phonological knowledge and its computational implementation. Are the representational units that emerge from neural learning fundamentally different from those posited by linguistic theory, or do they converge on similar abstractions through different computational pathways? Can we design hybrid architectures that combine the interpretability of symbolic approaches with the learning capabilities of neural networks? How do these different representational choices affect cognitive plausibility, computational efficiency, and cross-linguistic generalization?

The urgency of these questions is amplified by practical considerations in speech technology. Current speech recognition and synthesis systems, while achieving impressive performance, often fail in ways that suggest a lack of genuine phonological understanding. They struggle with out-of-distribution data, show poor cross-linguistic transfer, and provide little insight into their failure modes. A deeper understanding of optimal representational units could lead to more robust, interpretable, and efficient speech processing systems, with particular benefits for low-resource languages where data scarcity makes pure neural approaches less viable.

Furthermore, this research has implications for our understanding of human language acquisition and processing. The representations learned by neural models provide a computational hypothesis about what information is available in the speech signal and how it might be organized by learning mechanisms. By comparing these learned representations with human behavioral data and developmental trajectories, we can gain insights into the computational principles underlying human phonological knowledge.

## 2 Research Objectives and Questions

This thesis pursues three interconnected research objectives that collectively address the fundamental question of optimal representational units for phonological modeling. These objectives are designed to provide both theoretical insights and practical applications, bridging the gap between linguistic theory and computational implementation.

### 2.1 Primary Research Question

What are the optimal representational units for language modeling in the era of deep learning, considering multiple dimensions of evaluation including predictive accuracy, linguistic interpretability, cognitive plausibility, and computational efficiency? This overarching question recognizes that "optimality" is not a monolithic concept but rather a multi-faceted evaluation that must consider the diverse requirements of different applications and theoretical frameworks.

### 2.2 Specific Research Questions

**RQ1: Empirical Landscape Mapping**

How do different representational units—ranging from continuous neural embeddings to discrete symbolic features—perform across diverse phonological tasks including phonotactics, allophonic variation, and morphophonological alternations? This question involves systematic comparison of continuous representations from self-supervised models (Baevski et al., 2020; S. Chen et al., 2022; Hsu et al., 2021), discrete codes from vector quantization methods (Cho et al., 2025; Higy et al., 2021), traditional symbolic units (phonemes, distinctive features, syllables), and hybrid representations combining neural and symbolic elements.

The investigation will examine not only overall performance metrics but also the qualitative differences in how different representations handle specific phonological phenomena. For instance, do continuous representations better capture gradient phonetic variation while discrete units excel at categorical phonological processes? How do different representations handle long-distance dependencies, such as vowel harmony or consonant co-occurrence restrictions? These detailed analyses will provide insights into the strengths and limitations of each representational approach.

**RQ2: Neuro-Symbolic Integration**

Can we design architectures that effectively integrate symbolic phonological knowledge with neural representations, achieving both high predictive accuracy and linguistic interpretability? This investigation focuses on developing bridge networks that map between continuous and discrete representations, implementing Maximum Entropy Harmonic Grammar with neural parameterization, creating interpretable bottlenecks that enforce phonological structure, and evaluating trade-offs between performance and interpretability.

The neuro-symbolic integration explores several innovative architectural designs. One approach involves using neural networks to learn constraint weights in Optimality Theory or Harmonic Grammar frameworks, allowing the model to discover phonological generalizations while maintaining the interpretable structure of constraint-based theories.

Another approach uses vector quantization techniques to create discrete bottlenecks in neural architectures, forcing the model to compress information into phonologically meaningful units. These hybrid architectures aim to combine the learning power of neural networks with the theoretical insights of symbolic phonology.

**RQ3: Cognitive Plausibility Validation**

To what extent do different representational choices align with human language acquisition and processing patterns? This question examines developmental trajectories using CHILDES corpus simulations (Benders & Blom, 2023; Cruz Blandón et al., 2023), perceptual discrimination patterns through ABX tasks (McMurray, 2023), cross-linguistic transfer and generalization capabilities (Venkateswaran et al., 2025), and computational efficiency relative to human processing constraints.

The cognitive validation component goes beyond simple performance metrics to examine whether models exhibit human-like learning trajectories and processing patterns. This includes investigating whether models show perceptual narrowing effects similar to infants, whether they exhibit similar patterns of overgeneralization and regularization during learning, and whether their internal representations align with neural recordings from human speech processing. These investigations provide crucial constraints on the space of plausible models and offer insights into the computational principles underlying human phonological competence.

# 3 Theoretical Framework

## 3.1 Foundations in Computational Phonology

This research builds upon three theoretical pillars that span traditional and modern approaches to phonological computation, each contributing essential insights to our understanding of optimal representational units.

The symbolic phonological tradition provides the foundational understanding of phonological structure and organization. The Sound Pattern of English (Chomsky & Halle, 1968) established the framework of distinctive features and rule-based derivations that has influenced all subsequent phonological theorizing. This work demonstrated that complex phonological patterns could be captured through a finite set of binary features and ordered rules, providing a powerful formalism for cross-linguistic generalizations. Autosegmental Phonology (Goldsmith, 1976) revolutionized this framework by proposing that different types of phonological information exist on separate tiers, connected by association lines. This multi-tiered architecture naturally captures phenomena like tone spreading, vowel harmony, and the independence of various phonological processes. Feature Geometry (Clements, 1985) further refined our understanding by organizing distinctive features in hierarchical structures, explaining why certain features pattern together in assimilation and dissimilation processes.

Optimality Theory (Prince & Smolensky, 2004) represents a paradigm shift from rule-based to constraint-based phonology. Instead of serial derivations, OT proposes that phonological patterns emerge from the interaction of ranked, violable constraints. This framework has proven remarkably successful in capturing typological generalizations and explaining the range of variation observed across languages. The computational implementations of OT, particularly Maximum Entropy Harmonic Grammar (Hayes & Wilson, 2008), provide probabilistic extensions that can handle gradient phenomena and variation. These frameworks offer a natural bridge to neural implementation, as constraint weights can be learned through gradient descent while maintaining linguistic interpretability.

The revolution in neural representation learning has demonstrated that rich phonological representations can emerge from distributional learning without explicit linguistic supervision. The word2vec revolution (Mikolov et al., 2013) showed that semantic relationships could be captured through vector arithmetic in embedding spaces, and subsequent work demonstrated that similar principles apply to phonological representations (Silfverberg et al., 2018). The development of self-supervised learning for speech, particularly through models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (S. Chen et al., 2022), has shown that neural networks can learn hierarchical representations of speech that capture phonetic and phonological information at different levels of abstraction.

Recent work on vector quantization (Higy et al., 2021; van den Oord et al., 2017) and neural codecs [@; Cho et al. (2025)] provides methods for creating discrete representations that maintain neural learning capabilities while offering interpretability. These approaches address a fundamental challenge in neural phonology: how to maintain the categorical nature of phonological representations while leveraging the learning power of gradient-based optimization. The development of models like SpeechTokenizer (Zhang et al., 2024) demonstrates that hierarchical discrete representations can effectively separate linguistic content from paralinguistic information, suggesting principled ways to discover phonologically relevant units.

The emerging field of neuro-symbolic AI (Panchendrarajan & Zubiaga, 2024, ) provides methodologies for combining symbolic reasoning with neural learning. In phonology, this includes innovative work on Generative Adversarial

Phonology (Begu, 2020), which shows that phonological patterns can emerge from the adversarial training dynamics of GANs without explicit symbolic supervision. Studies examining what these models learn (J. Chen & Elsner, 2023) reveal that they discover representations with properties similar to distinctive features, suggesting deep connections between neural and symbolic approaches.

### 3.2 Multi-dimensional Evaluation Framework

The thesis employs a comprehensive evaluation framework that recognizes that "optimality" in representational units cannot be reduced to a single metric. This framework draws inspiration from multi-objective optimization in machine learning and the multiple evaluation criteria used in cognitive science.

**Predictive Accuracy** encompasses performance on downstream tasks that test different aspects of phonological knowledge. This includes phoneme recognition accuracy, measuring how well representations distinguish phonological categories; word segmentation performance, testing the ability to identify lexical boundaries; morphophonological prediction, evaluating knowledge of alternations and phonological processes; and phonotactic acceptability judgments, assessing knowledge of sound sequence constraints. Each task provides insight into different aspects of phonological competence, and the pattern of performance across tasks reveals the strengths and limitations of different representational approaches.

**Linguistic Interpretability** evaluates how well learned representations align with established phonological theories and whether they enable extraction of linguistic generalizations. This involves analyzing whether neural representations encode distinctive features, natural classes, and phonological processes in ways that correspond to linguistic theory. It also examines whether learned constraints or rules can be interpreted in terms of known phonological principles and whether the models' errors are phonologically plausible. Interpretability is crucial not only for scientific understanding but also for practical applications where model decisions must be explainable.

**Cognitive Plausibility** assesses consistency with human acquisition patterns, processing limitations, and behavioral data. This includes examining whether models show developmental trajectories similar to human learners, whether they exhibit similar patterns of generalization and error, and whether their processing demands align with human cognitive constraints. Cognitive plausibility provides important constraints on the space of viable models and offers insights into the computational principles underlying human phonological competence.

**Computational Efficiency** considers resource requirements, training time, and inference speed relative to task performance. This dimension is crucial for practical applications, particularly in resource-constrained settings or real-time processing scenarios. Efficiency metrics include memory footprint, FLOPs required for training and inference, latency in online processing, and scalability to larger datasets and models. The trade-offs between efficiency and other evaluation dimensions reveal fundamental constraints on phonological computation.

## 4 Methodology

### 4.1 Experimental Design

The research employs a systematic comparative methodology across three empirical studies corresponding to the research questions. Each study is designed to provide complementary insights into the nature of optimal phonological representations, using diverse datasets, tasks, and evaluation metrics.

**Study 1: Representational Landscape Analysis (RQ1)** provides a comprehensive empirical comparison of different representational units across a battery of phonological tasks. The study uses a controlled experimental design where all representations are evaluated on identical data and tasks, ensuring fair comparison.

The dataset collection includes LibriSpeech (Panayotov et al., 2015) for English phonological analysis, providing 1000 hours of read speech with aligned transcriptions; Common Voice (Ardila et al., 2020) for multilingual evaluation, covering 50+ languages with diverse phonological systems; specialized corpora like TIMIT for fine-grained phonetic analysis; and custom-collected data for specific phonological phenomena such as tone, stress, and vowel harmony. These diverse datasets ensure that findings generalize across languages and speaking styles.

The experimental tasks span multiple levels of phonological analysis. Phoneme classification tests basic segmental representation using frame-level and segment-level classification accuracy. Allophone prediction (Pouw et al., 2024) evaluates knowledge of context-dependent variation using rule-based and statistical patterns. Phonotactic acceptability judgments (Guriel et al., 2023) assess knowledge of sequential constraints through well-formedness ratings and discrimination tasks. Morphophonological alternation modeling tests understanding of systematic sound changes in paradigms through wug-test style generalization and naturalistic alternation prediction.

Evaluation employs multiple metrics to capture different aspects of performance. Accuracy metrics include frame-level and segment-level classification accuracy, F1-scores for imbalanced categories, and confusion matrices revealing systematic errors. Information-theoretic measures (Kolachina & Magyar, 2019) quantify mutual information between representations and linguistic categories, entropy of learned representations, and compression rates for different unit types. Linguistic analysis examines error patterns for phonological plausibility, generalization to unseen contexts, and cross-linguistic transfer capabilities.

**Study 2: Hybrid Architecture Development (RQ2)** focuses on designing and implementing novel neuro-symbolic integration strategies that combine the strengths of neural and symbolic approaches.

The architectural innovations include several novel designs. The Neural-Symbolic Bridge Network uses a frozen SSL encoder (WavLM-Large) for feature extraction, combined with a trainable bridge network that maps continuous representations to symbolic units, and a MaxEnt HG decoder implementing phonological constraints. The Vector Quantized Phonological Network employs hierarchical VQ-VAE with phonologically-informed codebooks, using separate quantization for different feature types (place, manner, voicing), with disentanglement objectives ensuring interpretable codes. The Constraint Discovery Network implements differentiable OT using Gumbel-softmax for discrete decisions, with automatic constraint induction from data and interpretable constraint weights learned through gradient descent.

Training procedures are carefully designed to balance different objectives. Multi-task learning simultaneously optimizes for reconstruction, classification, and constraint satisfaction. Curriculum learning progresses from simple to complex phonological patterns, starting with individual segments and building to sequences and alternations. Regularization techniques include sparsity constraints on symbolic representations, information bottlenecks enforcing compression, and phonological priors from typological databases (Moran & McCloy, 2019).

The evaluation focuses on the performance-interpretability trade-off. Pareto frontier analysis identifies models that optimally balance accuracy and interpretability. Ablation studies determine the contribution of different architectural components. Constraint weight analysis examines learned weights for linguistic plausibility and cross-linguistic validity. Human evaluation assesses the interpretability of extracted rules and constraints.

**Study 3: Cognitive Validation (RQ3)** examines whether different representational approaches align with human language acquisition and processing patterns.

The developmental simulation uses the CHILDES corpus (Macwhinney, 2000) to model language acquisition trajectories. Age-stratified training exposes models to data in developmentally appropriate sequences. Milestone tracking monitors the emergence of phonological contrasts, the development of phonotactic knowledge, and the acquisition of morphophonological alternations. Error analysis compares model errors with children's production patterns, examining overgeneralization, regularization, and U-shaped learning curves.

Perceptual experiments implement computational versions of classic psycholinguistic paradigms. ABX discrimination tasks (**schatz2021?**) test categorical perception and perceptual narrowing, comparing native vs. non-native contrasts. Gating paradigms examine incremental processing and predictive capabilities. Priming experiments investigate phonological representation and activation patterns.

Cross-linguistic evaluation tests transfer and generalization across typologically diverse languages. Zero-shot transfer evaluates performance on unseen languages without additional training. Few-shot learning examines rapid adaptation with minimal target language data. Universal tendency analysis tests whether models exhibit biases toward typologically common patterns, preferences for unmarked structures, and learnability differences for natural vs. unnatural patterns.

### 4.2 Technical Implementation

The computational infrastructure is designed for reproducibility, scalability, and efficient experimentation.

The software environment uses Docker containers ensuring reproducible environments across different systems, Poetry for dependency management with locked versions, and Git with DVC for version control of code and data. Experiment tracking uses Weights & Biases for comprehensive logging and visualization. The development follows test-driven practices with continuous integration.

The data processing pipeline employs sophisticated techniques for phonetic analysis. Montreal Forced Aligner (McAuliffe et al., 2017) provides automatic phonetic alignment with manual verification. Custom preprocessing handles diverse audio formats, sampling rates, and recording conditions. Data augmentation strategies include speed perturbation, noise addition, and synthetic generation for low-resource scenarios. Quality control involves automatic detection of alignment errors and systematic validation of annotations.

Model implementations leverage state-of-the-art frameworks and architectures. Baseline models include wav2vec 2.0, HuBERT, and WavLM implementations from Hugging Face Transformers, with careful hyperparameter tuning and fair comparison protocols. Custom architectures implement VQ-VAE variants with phonological inductive biases, hybrid networks combining neural encoders with symbolic decoders, and MaxEnt HG implementations with neural constraint discovery. All implementations prioritize efficiency through mixed precision training, gradient checkpointing, and distributed data parallel training.

### 4.3 Evaluation Protocols

The evaluation employs rigorous protocols ensuring reliable and interpretable results.

Quantitative metrics comprehensively assess model performance. Task-specific measures include accuracy, F1-score, and confusion matrices for classification tasks; WER and PER for sequence prediction; and perplexity and bits-per-character for language modeling. Information-theoretic measures (Maaten & Hinton, 2008) quantify mutual information with linguistic categories, encoding efficiency and compression rates, and disentanglement metrics for interpretable representations. Efficiency metrics track training time and convergence rates, inference latency and throughput, memory usage and model size, and energy consumption for environmental impact.

Qualitative analysis provides linguistic insights into model behavior. Representation analysis uses probing classifiers to test for specific linguistic information, visualization techniques including t-SNE and attention maps, and correlation analysis with linguistic features. Error analysis categorizes errors by phonological type, examines systematic biases and patterns, and compares with human error patterns. Case studies provide detailed analysis of specific phenomena such as vowel harmony, tone sandhi, and morphophonological alternations.

Statistical validation ensures robust and reliable conclusions. Bootstrap confidence intervals provide uncertainty estimates for all metrics. Permutation tests assess significance of differences between models. Multiple comparison correction controls for false discoveries in extensive evaluations. Effect size estimation quantifies practical significance beyond statistical significance. Cross-validation ensures generalization across data splits.

## 5 Preliminary Results and Pilot Studies

Initial experiments on micro-scale datasets have yielded promising results that validate the research approach and provide crucial insights for the full-scale investigation.

### 5.1 Probing Experiments on SSL Representations

Preliminary probing experiments (Venkateswaran et al., 2025) reveal hierarchical organization of phonological information in SSL models. Analysis of wav2vec 2.0 representations shows that lower layers (1-4) primarily encode acoustic-phonetic features such as formant frequencies, voice onset time, and spectral characteristics. Middle layers (5-8) capture phonemic categories and allophonic variation, showing highest accuracy for phoneme classification tasks. Upper layers (9-12) represent more abstract linguistic information including morphological and lexical patterns.

Layer-wise analysis using linear probing reveals interesting patterns in information encoding. Phonetic features like voicing and place of articulation are robustly encoded across multiple layers, with peak performance in layers 6-7. Manner features show more distributed encoding, suggesting they require integration of multiple acoustic cues. Suprasegmental features like stress and tone are better captured in upper layers, indicating they require longer temporal context. These findings suggest a natural progression from signal to symbol, supporting the hypothesis that neural models implicitly discover hierarchical phonological organization.

### 5.2 Vector Quantization Studies

Initial VQ experiments with 128 clusters demonstrate promising results for bridging continuous and discrete representations. The learned codebooks show interesting phonological structure, with codes clustering according to phonetic similarity. Discrete codes maintain 85% of the performance of continuous representations on phoneme classification while requiring 75% less memory and enabling symbolic manipulation (Higy et al., 2021).

Analysis of the learned codes reveals emergent phonological organization. Codes corresponding to vowels form a distinct cluster separate from consonants. Within consonants, natural classes emerge with stops, fricatives, and sonorants forming identifiable subclusters. The geometric structure of the codebook space shows correspondence with articulatory features, suggesting that VQ discovers phonetically meaningful discretization. These results indicate that carefully designed quantization can preserve phonological information while providing interpretable discrete units.

### 5.3 Hybrid Model Proof-of-Concept

Proof-of-concept implementations of neural-parameterized MaxEnt HG show that constraint weights can be learned end-to-end while maintaining interpretability. The model successfully learns phonotactic constraints from English data, discovering restrictions on onset clusters, coda sequences, and vowel combinations. Learned constraint weights show meaningful patterns, with markedness constraints generally weighted higher than faithfulness constraints, consistent with linguistic theory (Jarosz, 2019).

Early results indicate that hybrid models can discover phonologically meaningful constraints without explicit supervision. The model learns that *COMPLEX-ONSET is violated by word-initial consonant clusters,*CODA-VOICE prohibits voiced obstruents in coda position, and AGREE constraints enforce harmony patterns. These discovered constraints align well with known phonological generalizations, suggesting that the hybrid architecture successfully combines neural learning with symbolic structure (Begu, 2020).

### 5.4 Developmental Trajectory Modeling

Pilot studies using a subset of CHILDES data reveal intriguing parallels between model and human developmental trajectories. Models trained on age-stratified data show similar patterns of phonological acquisition, with early mastery of vowel contrasts, gradual acquisition of consonant clusters, and late development of morphophonological alternations. The models exhibit U-shaped learning curves for irregular patterns, initially memorizing specific forms, then overgeneralizing rules, before learning exceptions.

Error analysis shows that model errors resemble children's production patterns. Common error types include cluster reduction (e.g., “stop”  [tp]), final consonant deletion, and stopping of fricatives. The relative frequency of different error types matches developmental data, suggesting that models capture similar learning biases. These preliminary results support the cognitive plausibility of the proposed representational approaches.

## 6 Expected Contributions and Impact

### 6.1 Theoretical Contributions

The thesis will advance phonological theory by providing a unified framework for understanding the relationship between symbolic and neural representations. This framework will formally characterize the conditions under which neural and symbolic representations converge, demonstrating that under certain architectural and training constraints, neural models discover representations isomorphic to linguistic features. The framework will also identify fundamental divergences where neural representations capture patterns invisible to traditional symbolic approaches, such as gradient phonetic detail and probabilistic generalizations.

The research will establish optimal granularity principles for phonological representation, showing that different levels of granularity are optimal for different tasks and phenomena. Fine-grained continuous representations excel at capturing phonetic variation and speaker-specific patterns. Intermediate discrete units (phones, allophones) provide the best balance for speech recognition and synthesis. Abstract symbolic features are optimal for capturing phonological generalizations and cross-linguistic patterns. This multi-granular view reconciles apparently contradictory findings in the literature and provides a principled basis for representation selection.

Novel neuro-symbolic integration methodologies will demonstrate how symbolic phonological knowledge can be incorporated into neural architectures without sacrificing learning flexibility. These methods include differentiable implementations of constraint-based grammars, attention mechanisms that implement phonological processes, and modular architectures that separate phonological computation from phonetic realization. These contributions extend beyond phonology to other areas of linguistics where similar tensions between symbolic and statistical approaches exist.

### 6.2 Practical Contributions

The research will yield immediate practical benefits for speech technology applications. Improved speech recognition systems will result from hybrid models that combine neural acoustic modeling with phonological structure, achieving better performance on out-of-distribution data, improved handling of low-frequency words and morphological complexity, and more robust cross-linguistic transfer with less target language data.

Interpretable AI systems for speech processing will provide explicit phonological knowledge enabling better debugging and error analysis. Constraint weights and symbolic representations offer insights into model decisions, crucial for applications in education, clinical assessment, and forensic phonetics. The ability to examine and modify learned phonological constraints allows for targeted improvements and adaptation to specific domains or speakers.

Low-resource language applications will particularly benefit from architectures that leverage phonological universals. By incorporating typological knowledge from databases like PHOIBLE (Moran & McCloy, 2019), models can achieve reasonable performance with minimal training data. Transfer learning from high-resource languages becomes more effective when mediated by phonologically-informed representations. This has important implications for language preservation and documentation efforts.

### 6.3 Interdisciplinary Impact

For theoretical linguistics, the research provides large-scale empirical validation of phonological theories. Computational experiments test predictions of different theoretical frameworks at a scale impossible with traditional methods. The learned representations offer new insights into phonological organization, potentially revealing patterns invisible to human analysts. The success or failure of different representational approaches provides evidence for fundamental questions about the nature of phonological knowledge.

Machine learning contributions include novel architectures for discrete representation learning that balance expressiveness with interpretability. Methods for incorporating domain knowledge into neural networks without constraining learning flexibility have applications beyond speech processing. The multi-objective optimization framework for representation learning provides principled approaches to trading off competing desiderata. These contributions advance the broader agenda of interpretable and controllable AI systems.

The cognitive science community benefits from computational models that can be directly compared with human behavioral and neural data. The models provide testable predictions about acquisition trajectories, processing dynamics, and neural encoding. The relationship between optimal computational solutions and human cognitive patterns offers insights into the evolutionary and developmental pressures shaping phonological systems. These models serve as valuable tools for generating and testing hypotheses about human language processing.

### 6.4 Broader Societal Impact

The research has significant implications for education and accessibility. Improved models of phonological acquisition can inform language teaching methodologies and intervention strategies for speech disorders. Interpretable representations enable development of diagnostic tools that provide explicit feedback about pronunciation and phonological patterns. Applications in computer-assisted language learning can adapt to individual learner profiles and provide targeted practice.

For language preservation and revitalization efforts, the ability to build effective speech technology with minimal data is crucial. Many endangered languages lack the large corpora required by current neural approaches. Phonologically-informed models can leverage cross-linguistic regularities to bootstrap speech technology for these languages, supporting documentation, education, and community use.

## 7 Resources and Feasibility

### 7.1 Data Resources

Access to necessary datasets has been confirmed or is in process. Licensed datasets include LibriSpeech (1000 hours) for English experiments, Common Voice for multilingual studies, and CHILDES (approved access pending) for developmental modeling. Custom data collection is planned for specific phenomena not covered in existing corpora.

### 7.2 Risk Mitigation

Several risks have been identified with corresponding mitigation strategies. Technical risks include model training failures and computational bottlenecks, mitigated through incremental development and checkpoint saving. Data risks such as access restrictions or quality issues are addressed through multiple data sources and quality control procedures. Timeline risks from unexpected delays are managed through buffer time and parallel task execution.

# 8 Ethical Considerations and Broader Impact

## 8.1 Ethical Considerations

The research adheres to strict ethical guidelines for responsible AI development. Privacy protection ensures no use of personally identifiable information in speech data, with careful anonymization of any custom-collected data and compliance with relevant privacy regulations. Bias mitigation involves careful attention to demographic representation in training data, evaluation of model fairness across different speaker groups, and transparent reporting of limitations and potential biases.

Intellectual property considerations include proper attribution and citation of all data sources and prior work, with open-source release of code under permissive licenses and clear documentation of model provenance and training procedures.

## 8.2 Broader Impact Statement

This research has the potential for significant positive impact across multiple domains. Scientific advancement through improved understanding of phonological representation and learning, new methodologies for investigating linguistic questions, and tools and resources for the research community will accelerate progress in computational linguistics. Technological innovation in more effective and interpretable speech processing systems, improved support for low-resource languages, and enhanced human-computer interaction through better speech understanding will benefit society broadly.

Potential negative impacts are acknowledged and addressed. Concerns about AI system interpretability and trust are mitigated through our focus on interpretable representations. Risks of technology misuse are minimized through responsible disclosure and documentation. The potential for reinforcing linguistic biases is addressed through careful data curation and evaluation. Environmental costs of large-scale computation are offset through efficiency improvements and transparent reporting.

# 9 Conclusion

This doctoral thesis addresses fundamental questions at the intersection of linguistics, computer science, and cognitive science by investigating optimal representational units for phonological modeling in the deep learning era. Through systematic empirical investigation, novel architectural designs, and comprehensive evaluation across multiple dimensions, the research aims to bridge the gap between symbolic linguistic theory and neural representation learning.

The proposed research program is ambitious yet feasible, building on solid theoretical foundations and preliminary results that demonstrate the viability of the approach. The three-pronged investigation through empirical landscape mapping, neuro-symbolic integration, and cognitive validation provides complementary perspectives on the central question of optimal phonological representation.

The expected contributions span theoretical advances in our understanding of phonological representation, practical improvements in speech technology, and interdisciplinary insights relevant to cognitive science and machine learning. The multi-dimensional evaluation framework ensures that findings are robust and applicable across different contexts and requirements.

The preliminary results are encouraging, showing that neural models learn hierarchically organized phonological representations, that vector quantization can effectively bridge continuous and discrete representations, and that hybrid neuro-symbolic architectures can discover meaningful phonological constraints. These findings suggest that the integration of symbolic and neural approaches is not only possible but beneficial, offering a path toward more powerful, interpretable, and cognitively plausible models of phonological processing.

By pursuing this research program, the thesis will contribute to the development of a new generation of speech processing systems that combine the learning power of neural networks with the interpretability and theoretical grounding of linguistic knowledge. This work has implications not only for technology development but also for our fundamental understanding of how phonological knowledge is represented, learned, and processed.

The research is particularly timely given the rapid advances in self-supervised learning and the growing recognition of the need for interpretable AI systems. As speech technology becomes increasingly prevalent in daily life, understanding the optimal representations for phonological modeling becomes crucial for developing systems that are not only accurate but also trustworthy, adaptable, and aligned with human cognitive capabilities.

The ultimate goal of this thesis is to establish principled foundations for phonological representation in the age of deep learning, providing both theoretical insights and practical tools that will benefit researchers and practitioners across multiple disciplines. Through rigorous empirical investigation, innovative architectural design, and careful evaluation, this research aims to advance our understanding of one of the most fundamental aspects of human language while contributing to the development of more effective and interpretable speech technology.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common voice: A massively-multilingual speech corpus* (arXiv:1912.06670). arXiv. https://doi.org/10.48550/arXiv.1912.06670

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Begu, G. (2020). Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00044

Benders, T., & Blom, E. (2023). Computational modelling of language acquisition: An introduction. *Journal of Child Language*, *50*(6), 1287–1293. https://doi.org/10.1017/S0305000923000429

Chen, J., & Elsner, M. (2023). *Exploring how generative adversarial networks learn phonological representations* (arXiv:2305.12501). arXiv. https://doi.org/10.48550/arXiv.2305.12501

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Cho, C. J., Lee, N., Gupta, A., Agarwal, D., Chen, E., Black, A. W., & Anumanchipalli, G. K. (2025). *Sylber: Syllabic embedding representation of speech from raw audio* (arXiv:2410.07168). arXiv. https://doi.org/10.48550/arXiv.2410.07168

Chomsky, N., & Halle, M. (1968). *The sound pattern of english.*

Clements, G. N. (1985). The geometry of phonological features. *Phonology*, *2*, 225–252.

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (2023). Introducing meta-analysis in the evaluation of computational models of infant language development. *Cognitive Science*, *47*(7), e13307. https://doi.org/10.1111/cogs.13307

Goldsmith, J. A. (1976). *Autosegmental phonology* [PhD thesis]. Massachusetts Institute of Technology.

Guriel, D., Goldman, O., & Tsarfaty, R. (2023). *Morphological inflection with phonological features* (arXiv:2306.12581). arXiv. https://doi.org/10.48550/arXiv.2306.12581

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. https://doi.org/10.1162/ling.2008.39.3.379

Higy, B., Gelderloos, L., Alishahi, A., & Chrupaa, G. (2021). Discrete representations in neural models of spoken language. In J. Bastings, Y. Belinkov, E. Dupoux, M. Giulianelli, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the fourth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP* (pp. 163–176). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.blackboxnlp-1.11

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units* (arXiv:2106.07447). arXiv. https://doi.org/10.48550/arXiv.2106.07447

Jarosz, G. (2019). Computational modeling of phonological learning. *Annual Review of Linguistics*, *5*(1), 67–90. https://doi.org/10.1146/annurev-linguistics-011718-011832

Kolachina, S., & Magyar, L. (2019). What do phone embeddings learn about phonology? In G. Nicolai & R. Cotterell (Eds.), *Proceedings of the 16th workshop on computational research in phonetics, phonology, and morphology* (pp. 160–169). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4219

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Macwhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs.* Lawrence Erlbaum Associates Publishers.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech*, *2017*, 498–502.

McMurray, B. (2023). The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Language, Cognition and Neuroscience*, *38*(4), 419–445. https://doi.org/10.1080/23273798.2022.2105367

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Moran, S., & McCloy, D. (Eds.). (2019). *Phoible 2.0.* Max Planck Institute for the Science of Human History.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Panchendrarajan, R., & Zubiaga, A. (2024). *Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing* (arXiv:2401.11972). arXiv. https://doi.org/10.48550/arXiv.2401.11972

Pouw, C., Kloots, M. de H., Alishahi, A., & Zuidema, W. (2024). Perception of phonological assimilation by neural speech recognition models. *Computational Linguistics*, *50*(3), 1557–1585. https://doi.org/10.1162/coli_a_00526

Prince, A., & Smolensky, P. (2004). Optimality theory: Constraint interaction in generative grammar. In *Optimality theory in phonology* (pp. 1–71). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756171.ch1

Silfverberg, M. P., Mao, L., & Hulden, M. (2018). Sound Analogies with Phoneme Embeddings. *Society for Computation in Linguistics*, *1*(1). https://doi.org/10.7275/R5NZ85VD

van den Oord, A., Vinyals, O., & kavukcuoglu, koray. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, *30*.

Venkateswaran, N., Tang, K., & Wayland, R. (2025). *Probing for phonology in self-supervised speech representations: A case study on accent perception* (arXiv:2506.17542). arXiv. https://doi.org/10.48550/arXiv.2506.17542

Zhang, X., Zhang, D., Li, S., Zhou, Y., & Qiu, X. (2024). *SpeechTokenizer: Unified speech tokenizer for speech large language models* (arXiv:2308.16692). arXiv. https://doi.org/10.48550/arXiv.2308.16692