PHONOLOGICAL FEATURES IN THE AGE OF DEEP LEARNING: A MULTI-DIMENSIONAL EXPLORATION OF OPTIMAL REPRESENTATIONAL UNITS FOR LANGUAGE MODELING

DOCTORAL THESIS ABSTRACT

Sora Nagano
Graduate School of Humanities and Sociology, Department of Language and Information Sciences
The University of Tokyo

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

2025-08-14

1 Research Background and Objectives

This research addresses a fundamental question in computational phonology: "What are the optimal representational units for language modeling?" using the lens of modern deep learning technologies. At the heart of this inquiry lies a fundamental tension within linguistics. On one hand, there are symbolic features such as distinctive features, phonemes, and syllables used in Optimality Theory (OT), which are interpretable and elegant for human understanding. On the other hand, there are the powerful yet often opaque continuous vector representations that modern neural networks learn from vast amounts of data.

1.1 Problem Statement

Current computational phonology faces a fundamental divide between two paradigms. The symbolic tradition has employed discrete symbols such as distinctive features, phonemes, and syllables as canonical units for phonological analysis. These provide a foundation for describing and explaining linguistic structure and possess characteristics that are easily understood by humans. However, they face difficulties in representing fine-grained sound differences and gradient phenomena.

Conversely, modern neural networks demonstrate powerful and flexible learning capabilities through the use of sub-symbolic representations. Self-supervised learning (SSL) models such as wav2vec 2.0, HuBERT, and WavLM have shown the ability to learn rich representations from large amounts of unlabeled speech data and model diverse phonological phenomena. However, the internal workings of these models are often opaque, and it is not self-evident whether the learned representations are organized in linguistically interpretable ways.

1.2 Research Novelty and Academic Significance

The originality of this research lies in proposing a **hybrid neuro-symbolic architecture** that integrates traditionally opposed symbolic and connectionist approaches. Specifically, we develop a novel modeling methodology that uses continuous representations from SSL models to learn the weights of symbolic constraint-based grammars such as Maximum Entropy Harmonic Grammar (MaxEnt HG).

This research is academically significant in the following respects:

1. **Theoretical Integration**: Bridging the gap between symbolism and connectionism, two major paradigms in linguistics that have long been in opposition

2. **Interpretable AI**: Bringing linguistic interpretability to black-boxed neural models
3. **Cognitive Plausibility**: Validating consistency with human language acquisition mechanisms
4. **Computational Efficiency**: Pursuing applicability to practical speech technologies

2 Research Methodology

2.1 Multi-dimensional Definition of "Optimality"

This research operationally defines "optimal representational units" not as a single criterion but as a multi-objective optimization problem across four dimensions:

1. **Predictive Accuracy**: Generalization capability to unseen data across diverse phonological tasks
2. **Linguistic Interpretability**: Consistency with existing linguistic theories and explainability
3. **Cognitive Plausibility**: Alignment with human language acquisition and processing patterns
4. **Computational Efficiency**: Efficiency in terms of data volume, time, and computational resources

2.2 Three Major Research Questions

2.2.1 RQ1: Empirical Landscape Elucidation

Research Question: How do different representational units (continuous vectors, VQ codes, phonemes, features) compare in their ability to model diverse phonological phenomena from the perspectives of predictive accuracy and computational efficiency?

Methodology:

- Use WavLM-Large as a common feature extractor for fair comparison
- Candidate units: (1) Continuous: WavLM hidden state vectors, (2) Discrete: VQ layer clustering results, (3) Symbolic: Supervised mapping to phoneme/feature labels
- Evaluation tasks: phoneme classification, phonotactic modeling, morphological inflection

2.2.2 RQ2: Neuro-symbolic Bridge

Research Question: Can hybrid architectures that parameterize symbolic constraint-based grammars using continuous representations from SSL models achieve better trade-offs than pure neural or symbolic models?

Architecture Design:

1. **Neural Frontend**: Continuous representation extraction by frozen WavLM
2. **Symbolic Backend**: MaxEnt HG constraint system
3. **Bridge Network**: Small-scale neural network predicting constraint violation profiles from continuous vectors

2.2.3 RQ3: Cognitive Plausibility Validation

Research Question: Which representational paradigm best captures developmental trajectories in human infant language acquisition and demonstrates the highest cognitive plausibility?

Methodology: Developmentally realistic learning simulation using CHILDES corpus, reproduction of perceptual narrowing through ABX discrimination tasks

3 Technical Approach

3.1 Utilization of Vector Quantization (VQ)

The technical core of this research lies in Vector Quantization (VQ) technology that bridges continuous and discrete representations. VQ recovers symbolicity while maintaining learning capability by mapping continuous neural network outputs to finite discrete "codebooks."

VQ Implementation:

- Representative vector learning through K-means clustering (K=128)

2

- Mapping continuous vectors to nearest discrete IDs
- Optimal cluster number determination using information-theoretic measures (NMI)

3.2 Detailed Design of Hybrid Models

The proposed neuro-symbolic model consists of three stages:

1. **Feature Extraction Stage**: Conversion from speech to continuous representations by pre-trained WavLM model
2. **Bridge Stage**: Constraint violation prediction by small neural network
3. **Symbolic Reasoning Stage**: Optimal candidate selection by MaxEnt HG grammar

This design enables integration of data-driven representation learning with explicit description of linguistic constraints.

4 Experimental Plan

4.1 Datasets

- **LibriSpeech**: English speech recognition dataset (960 hours)
- **Common Voice**: Multilingual speech dataset (100+ languages)
- **CHILDES**: Child-directed speech corpus (for developmental simulation)
- **Specialized Corpora**: Buckeye Corpus (American English flapping), Turkish corpus (vowel harmony)

4.2 Evaluation Metrics

- **Accuracy Metrics**: F1-score, classification accuracy, perplexity
- **Efficiency Metrics**: Training time, inference time, memory usage
- **Interpretability Metrics**: Linguistic validity analysis of constraint weights
- **Cognitive Metrics**: ABX error rates, developmental trajectory alignment

4.3 Experimental Procedure

4.3.1 Phase 1: Baseline Establishment (6 months)

- Individual performance evaluation for each representational unit
- Benchmarking on standard phonological tasks
- VQ model optimization

4.3.2 Phase 2: Hybrid Model Development (12 months)

- Architecture design and implementation
- Phonological process modeling experiments
- Constraint learning algorithm optimization

4.3.3 Phase 3: Cognitive Plausibility Validation (6 months)

- Developmental simulation experiments
- Comparison with human experimental data
- Psycholinguistic validation of model predictions

5 Expected Results and Contributions

5.1 Theoretical Contributions

1. **Computational Implementation of Phonological Theory**: Proposing new phonological theory through integration of symbolic constraints and distributed representations
2. **Extension of Learning Theory**: Development of unsupervised learning algorithms for constraint-based grammars
3. **Cognitive Modeling**: Establishing new paradigms in computational explanations of language acquisition

5.2 Practical Contributions

1. **Speech Technology Enhancement**: More accurate and efficient multilingual speech recognition systems
2. **Language Learning Support**: Applications to pronunciation correction and second language acquisition
3. **Low-resource Language Processing**: Efficient model development utilizing linguistic knowledge

5.3 Impact on Computational Linguistics

This research is expected to have the following impacts as the first systematic comparative study of phonological units in the SSL era:

- Proposing methods for improving interpretability of deep learning models
- Expanding applications of neuro-symbolic AI to phonology
- Establishing integration methodology between linguistic knowledge and data-driven learning

6 Research Originality and Challenge

6.1 Differences from Existing Research

Previous research has focused on probing the existence of individual phonological properties in specific models. This research is original in the following respects:

1. **Systematic Comparison**: Comparative evaluation of diverse representational units within a unified framework
2. **Multi-axis Evaluation**: Comprehensive assessment of not only accuracy but also interpretability, cognitive plausibility, and efficiency
3. **Integrative Approach**: Proposing new models through integration of symbolic and distributed approaches

6.2 Technical Challenges

1. **Scalability**: Efficient implementation of hybrid models for large-scale data
2. **Optimization**: Representational selection algorithms as multi-objective optimization problems
3. **Evaluation**: Development of quantitative evaluation methods for interpretability and cognitive plausibility

7 Research Schedule and Feasibility

7.1 Overall Schedule (36 months)

Year 1 (Foundation Research Period):

- Deepening literature review and refining theoretical framework
- Building basic experimental environment and baseline experiments
- Presenting preliminary results at international conferences

Year 2 (Core Research Period):

- Design, implementation, and evaluation of hybrid models
- Large-scale experiments and ablation studies
- Publication of major experimental results

Year 3 (Integration and Development Period):

- Cognitive plausibility validation experiments
- Exploration of theoretical implications and application possibilities
- Doctoral thesis writing and final result presentation

7.2 Risk Management

Technical Risks: Preparing multiple backup approaches and ensuring research progress through incremental goal setting

Data Risks: Securing multiple data sources and conducting parallel validation with synthetic data and low-resource languages

Evaluation Risks: Combining existing and novel metrics for validation from multiple perspectives

8 Social Significance and Future Prospects

8.1 Short-term Impact

- Improved accuracy of multilingual speech technologies
- Enhanced language learning and education tools
- Expansion of computational methods in phonological research

8.2 Long-term Vision

This research contributes to the larger academic and social goal of bridging human language capabilities and artificial intelligence. By integrating linguistic knowledge with data-driven learning, it aims to provide foundational research toward realizing more human-like and interpretable AI systems.

In particular, it is expected to contribute to the broader AI research community by providing concrete answers from a phonological perspective to questions about how rapidly developing large language models can integrate and utilize linguistic knowledge.

9 Conclusion

This research presents an ambitious plan that takes a multi-faceted approach to a fundamental question in computational phonology of the deep learning era: What are the optimal representational units? By integrating symbolic traditions with modern neural approaches, it aims to open new horizons in both phonological theory and practical technology.

Through this research, we are confident that we can make concrete and empirical contributions to the 21st-century academic challenge of linguistics-AI fusion while providing theoretical foundations for next-generation language technology development.