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

DOCTORAL THESIS ABSTRACT

DOCTORAL THESIS WRITING QUALIFICATION REVIEW

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

2025-08-17

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 (Prince and Smolensky 2004), 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 (Tesar and Smolensky 1998; Hayes and Wilson 2008). 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 (Baevski et al. 2020), HuBERT (Hsu et al. 2021), and WavLM (Chen et al. 2022) 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 (Martin et al. 2023).

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 (Pater 2019). 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 (Hayes and Wilson 2008; Goldwater and Johnson 2003).

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 (Maye, Werker, and Gerken 2002; Wu et al. 2022)
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 (Reh, Hensch, and Werker 2021; Matusevych et al. 2020)
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 (Chen et al. 2022) as a common feature extractor for fair comparison
- Candidate units: (1) Continuous: WavLM hidden state vectors, (2) Discrete: VQ layer clustering results (Oord, Vinyals, and Kavukcuoglu 2017), (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 (Chen et al. 2022)
2. **Symbolic Backend**: MaxEnt HG constraint system (Hayes and Wilson 2008)
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 (Matusevych et al. 2020)

3 Technical Approach

3.1 Utilization of Vector Quantization (VQ)

The technical core of this research lies in Vector Quantization (VQ) technology (Oord, Vinyals, and Kavukcuoglu 2017; Baevski et al. 2021) 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)
- 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 (Begu 2020, 2021) consists of three stages:

1. **Feature Extraction Stage**: Conversion from speech to continuous representations by pre-trained WavLM model (Chen et al. 2022)
2. **Bridge Stage**: Constraint violation prediction by small neural network
3. **Symbolic Reasoning Stage**: Optimal candidate selection by MaxEnt HG grammar (Goldwater and Johnson 2003)

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

4 Experimental Plan

4.1 Datasets

- **LibriSpeech** (Panayotov et al. 2015): English speech recognition dataset (960 hours)
- **Common Voice** (Ardila et al. 2020): 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 (Silfverberg, Mao, and Hulden 2018; Kolachina and Magyar 2019)

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 (Moran and McCloy 2019)
- 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 (Jarosz 2013)

4.3.3 Phase 3: Cognitive Plausibility Validation (6 months)

- Developmental simulation experiments
- Comparison with human experimental data (Maye, Werker, and Gerken 2002; Wu et al. 2022)
- 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 (Prince and Smolensky 2004; Hayes and Wilson 2008)
2. **Extension of Learning Theory**: Development of unsupervised learning algorithms for constraint-based grammars (Jarosz 2019)
3. **Cognitive Modeling**: Establishing new paradigms in computational explanations of language acquisition (Reh, Hensch, and Werker 2021)

5.2 Practical Contributions

1. **Speech Technology Enhancement**: More accurate and efficient multilingual speech recognition systems (Pratap et al. 2024; Conneau et al. 2021)
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 (Pater 2019)
- 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 (Martin et al. 2023; Silfverberg, Mao, and Hulden 2018). 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 (Begu 2020)

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 (Karttunen 1998; Boersma and Hayes 2001)
- 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 (Zhang et al. 2024)
- Publication of major experimental results

Year 3 (Integration and Development Period):

- Cognitive plausibility validation experiments (Matusevych et al. 2020)
- 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 (Kolachina and Magyar 2019)

8 Social Significance and Future Prospects

8.1 Short-term Impact

- Improved accuracy of multilingual speech technologies (Pratap et al. 2024)
- 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 (Pater 2019), 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 (Prince and Smolensky 2004; Baevski et al. 2020), 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.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. "Common Voice: A Massively-Multilingual Speech Corpus." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–22. Marseille, France: European Language Resources Association.

Baevski, Alexei, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. "Unsupervised Speech Recognition." In *Advances in Neural Information Processing Systems*. Vol. 34. https://proceedings.neurips.cc/paper/2021/hash/ea159dc9788ffac311592613b7f71fbb-Abstract.html.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." In *Advances in Neural Information Processing Systems*, 33:12449–60. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Begu, Gaper. 2020. "Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning with Neural Networks." *Frontiers in Artificial Intelligence* 3: 44. https://doi.org/10.3389/frai.2020.00044.

———. 2021. "Identity-Based Patterns in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication." *Transactions of the Association for Computational Linguistics* 9: 1180–96. https://doi.org/10.1162/tacl_a_00421.

Boersma, Paul, and Bruce Hayes. 2001. "Empirical Tests of the Gradual Learning Algorithm." *Linguistic Inquiry* 32 (1): 45–86. https://doi.org/10.1162/002438901554586.

Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, et al. 2022. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing." *IEEE Journal of Selected Topics in Signal Processing* 16 (6): 1505–18. https://doi.org/10.1109/JSTSP.2022.3188113.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. "Unsupervised Cross-Lingual Representation Learning for Speech Recognition." In *Proceedings of INTERSPEECH 2021*, 2426–30. https://doi.org/10.21437/Interspeech.2021-329.

Goldwater, Sharon, and Mark Johnson. 2003. "Learning OT Constraint Rankings Using a Maximum Entropy Model." In *Proceedings of the Workshop on Variation Within Optimality Theory*, edited by Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–20. Stockholm, Sweden: Stockholm University.

Hayes, Bruce, and Colin Wilson. 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.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 3451–60. https://doi.org/10.1109/TASLP.2021.3122291.

Jarosz, Gaja. 2013. "Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing." *Phonology* 30 (1): 27–71.

———. 2019. "Computational Modeling of Phonological Learning." *Annual Review of Linguistics* 5: 67–90. https://doi.org/10.1146/annurev-linguistics-011718-011832.

Karttunen, Lauri. 1998. "The Proper Treatment of Optimality in Computational Phonology." In *Proceedings of Finite State Methods in Natural Language Processing*. https://aclanthology.org/W98-1301/.

Kolachina, Sudheer, and Lilla Magyar. 2019. "What Do Phone Embeddings Learn about Phonology?" In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 160–69. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4219.

Martin, Kinan, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. "Probing Self-Supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration." In *Proceedings of INTERSPEECH 2023*, 251–55. https://doi.org/10.21437/Interspeech.2023-2359.

Matusevych, Yevgen, Thomas Schatz, Herman Kamper, Naomi H. Feldman, and Sharon Goldwater. 2020. "Evaluating Computational Models of Infant Phonetic Learning Across Languages." https://arxiv.org/abs/2008.02888.

Maye, Jessica, Janet F. Werker, and LouAnn Gerken. 2002. "Infant Sensitivity to Distributional Information Can Affect Phonetic Discrimination." *Cognition* 82 (3): B101–11. https://doi.org/10.1016/S0010-0277(01)00157-3.

Moran, Steven, and Daniel McCloy. 2019. "PHOIBLE 2.0." Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.2626687.

Oord, Aaron van den, Oriol Vinyals, and Koray Kavukcuoglu. 2017. "Neural Discrete Representation Learning." In *Advances in Neural Information Processing Systems*, 30:6306–15.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–10. IEEE. https://doi.org/10.1109/ICASSP.2015.7178964.

Pater, Joe. 2019. "Generative Linguistics and Neural Networks at 60: Foundation, Friction, and Fusion." *Language* 95 (1): e41–74. https://doi.org/10.1353/lan.2019.0009.

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, et al. 2024. "Scaling Speech Technology to 1,000+ Languages." *Journal of Machine Learning Research* 25 (97): 1–52.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA: Blackwell Publishing. https://doi.org/10.1002/9780470759400.

Reh, Rebecca K., Takao K. Hensch, and Janet F. Werker. 2021. "Distributional Learning of Speech Sound Categories Is Gated by Sensitive Periods." *Cognition* 213: 104715. https://doi.org/10.1016/j.cognition.2021.104653.

Silfverberg, Miikka P., Lingshuang Mao, and Mans Hulden. 2018. "Sound Analogies with Phoneme Embeddings." *Proceedings of the Society for Computation in Linguistics* 1 (1): 136–44. https://doi.org/10.7275/R5NZ85VD.

Tesar, Bruce, and Paul Smolensky. 1998. "Learnability in Optimality Theory." *Linguistic Inquiry* 29 (2): 229–68. https://doi.org/10.1162/002438998553734.

Wu, Yan Jing, Xinlin Hou, Cheng Peng, Wenwen Yu, Gary M. Oppenheim, Guillaume Thierry, and Dandan Zhang. 2022. "Rapid Learning of a Phonemic Discrimination in the First Hours of Life." *Nature Human Behaviour* 6: 1169–79. https://doi.org/10.1038/s41562-022-01355-1.

Zhang, Xuankai, Shi-Xiong Zhang, Liangyou Li, Meng Yu, Yue-Hua Zhou, Bin Ma, and Haizhou Li. 2024. "Comparing Discrete and Continuous Space LLMs for Speech Recognition." https://arxiv.org/abs/2409.00800.