1

1.1

1. (summary_of_doctoral_thesis.qmd) - 8
2. (writing_plan.qmd) - 1
3. (list_of_references.qmd) -
4. (information_materials.qmd) - 1

1.2 1. (Summary of Proposed Doctoral Thesis)

1.2.1 1.1

1.2.1.1

-
-

1.2.1.2

- Chomsky & Halle (1968)SPE
- (Prince & Smolensky 2004)
- (Goldsmith 1976)
- (Clements 1985)
-

- wav2vec 2.0 (Baevski et al. 2020)
- HuBERT (Hsu et al. 2021)
- WavLM (Chen et al. 2022)
-

1.2.1.3

-
-
-

1.2.2 1.2

1.2.2.1

-

1.2.2.2 (RQ) RQ1:

- VQ
-
- vs
-

RQ2:

-
- MaxEnt
-
-

-

RQ3:

- CHILDES
- ABX
-
- U

1.2.3 1.3

1.2.3.1

- SPE (Chomsky & Halle 1968)
- (Goldsmith 1976)
- (Clements 1985)
- (Prince & Smolensky 2004)
- MaxEnt (Hayes & Wilson 2008)

- word2vec (Mikolov et al. 2013)
- (Silfverberg et al. 2018)
- wav2vec 2.0HuBERTWavLM
-

- VQ-VAE (van den Oord et al. 2017)
- SpeechTokenizer (Zhang et al. 2024)
-

AI

- (Garcez et al. 2022)
- (Begu 2020)GAN
-

1.2.3.2 4

1.
   -
   -
   -
   -
2.
   -
   -
   -
   -
3.
   -
   -
   -
   -
4.
   -
   - FLOPs
   -
   -

1.2.4 1.4

1.2.4.1 1: (RQ1)

- LibriSpeech1000
- Common Voice50+
- TIMIT
-
-
-
-
- wug-test

- F1
-
-

2: (RQ2)

- -
  - SSL (WavLM-Large)
  -
  - MaxEnt HG
-
  -
  -
  -
-
  - Gumbel-softmaxOT
  -
  -

-
-
-

-
-
-
-

3: (RQ3)

- CHILDES
-
-

-
- ABX
-
-
-
-
-

1.2.4.2

- Docker
- Poetry
- Git + DVC
- Weights & Biases

- Montreal Forced Aligner
-
-
-

- wav2vec 2.0HuBERTWavLM (Hugging Face)
- VQ-VAEMaxEnt HG
-

1.2.4.3

- F1WERPER
-
-

- t-SNE
-
-

-
-
-
-
-

1.2.5 1.5

1.2.5.1 SSL

- wav2vec 2.0
  - (1-4)VOT

  - (5-8)
  - (9-12)
- 6-7
-

1.2.5.2

- 128
  - 85%
  - 75%
  -
-
  -
  -
  -

1.2.5.3

- MaxEnt HG
  -
  -
-
  - *COMPLEX-ONSET
  - *CODA-VOICE
  - AGREE

1.2.5.4

- CHILDES
  -
  -
  - U
-
  - “stop” [tp]
  -
  -

1.2.6 1.6

1.2.6.1

-
  -
  -
  -
-
  -
  -
  -
-
  -
  -
  -

1.2.6.2

-
  -
  -
  -
- **AI**
  -
  -
  -
-
  - PHOIBLE
  -
  -

1.2.6.3

-
-
-

1.2.6.4

-
  -
  -
  -
-
  -
  -
  -

1.2.7 1.7

1.2.7.1 11-12

- **1-3**
  - 150+
  -
  - 40
- **4-6**
  - Docker
  - SSL
  - 30
- **7-9**RQ1
  -
  -
  - 1
- **10-12**RQ2
  -
  - MaxEnt HG
  - ACL/INTERSPEECH

1.2.7.2 213-24

- **13-15**
- **16-18**
- **19-21**RQ3
- **22-24**

1.2.7.3 325-36

- **25-27**
- **28-30**1
- **31-33**2
- **34-36**

1.2.8 1.8

1.2.8.1

- 8NVIDIA A100 GPU40GB
- 200TB
- AWS/Google Cloud
- 50,000 GPU

1.2.8.2

- LibriSpeech1000
- Common Voice
- CHILDES
- IRB

1.2.8.3

-
-
-
-

1.2.8.4

-
-
-
-

1.2.9 1.9

1.2.9.1

-
-
-
-

1.2.9.2

-
-
-

1.2.10 1.10

-
-
-
-
-

1.3 2. (Writing Plan)

1.3.1

1.3.1.1

- 363
-

1.3.2 11-12

1.3.2.1 11-3

-
  - 3
    - (Tesar 1995; Jarosz 2019)
    - (Baevski 2020; Mohamed 2022)
    - (Panchendrarajan 2024)
  - 40
-
  - 150
  -

1.3.2.2 24-6

-
  -
  - Docker
  - SSLwav2vec 2.0HuBERTWavLM
  - (Venkateswaran 2025)
  - 30
-
  -

1.3.2.3 37-9

-
  - 1
  - SSLVQ (Hsu 2021; Chen 2022; Higy 2021)
  - LibriSpeechCommon Voice
  - 135
-
  -

1.3.2.4 410-12

-
  - 2

  - MaxEnt (Hayes 2008)
  -
-
  - ACLINTERSPEECH

1.3.3 213-24

1.3.3.1 13-18

-
  -
  - (Begu 2020; Chen 2023)
  - PHOIBLE
  - 240

1.3.3.2 19-24

-
  - 3CHILDES (Cruz Blandón 2023)
  - ABX
  - 335
-
  - Computational LinguisticsTACL

1.3.4 325-36

1.3.4.1 25-30

-
  -
  - 25
  - 30
  - 1

1.3.4.2 31-36

-
  -
  - 20
  -
  -
  -

1.3.5

-
  -
  -
  -
  - Git
-
  - 2-3ACLINTERSPEECHNeurIPS
  - 1

1.3.6

1.3.6.1

- GPU
- Common Voice
- Montreal Forced Aligner
- Kaldi
- PyTorch
- Hugging Face
-

1.3.7

-
  - 250-300
  - 83
-
  -
  -
  -

1.4 3. (Annotated Bibliography)

1.4.1

Chomsky & Halle (1968) - SPE

Prince & Smolensky (2004) -

Goldsmith (1976) -

Clements (1985) -

1.4.2

Baevski et al. (2020) - wav2vec 2.010ASR

Hsu et al. (2021) - HuBERT

S. Chen et al. (2022) - WavLM94kSUPERB

Mohamed et al. (2022) -

1.4.3

van den Oord et al. (2017) - VQ-VAE

Zhang et al. (2024) - SpeechTokenizerVQHuBERT

Chang et al. (2024) - 40+ASR/TTS/SSL

Higy et al. (2021) - VQ

1.4.4

Hayes & Wilson (2008) - MaxEnt

B. Tesar & Smolensky (1998) - OT

Daland (2015) -

Jarosz (2019) -

1.4.5

Begu (2020) - GAN

J. Chen & Elsner (2023) - GAN

Garcez et al. (2022) -

Panchendrarajan & Zubiaga (2024) - NLP200+

1.4.6

Silfverberg et al. (2018) -

Kolachina & Magyar (2019) -

Venkateswaran et al. (2025) - SSL

Astrach & Pinter (2025) -

1.4.7

MacWhinney (2000) - CHILDES500030+CHATCLAN

Dupoux (2018) - AI

Schatz et al. (2021) - ABX

Cruz Blandón et al. (2023) -

Benders & Blom (2023) -

McMurray (2023) -

1.4.8

Conneau et al. (2020) - XLSRwav2vec 53

McAuliffe et al. (2017) - Montreal Forced AlignerKaldi50+20ms

Belinkov & Glass (2019) - NLP

Panayotov et al. (2015) - LibriSpeech1000train/dev/test

1.4.9

Yang et al. (2024) - k2SSL34.8% WER3.5ZipformerU-Net

Liu et al. (2022) - SSL

Ebrahimi et al. (2023) - 200+ NLP

Cho et al. (2025) - Sylber

1.4.10

Mortensen et al. (2016) - Panphon5000+ IPA21

Moran & McCloy (2019) - 21551672IPA

Dunbar et al. (2019) - TTS

Parcollet et al. (2024) - LeBenchmark

1.4.11

B. B. Tesar (1995) - OTOT

Silverman (2012) -

Staples & Graves (2020) -

Nguyen et al. (2016) -

Reubold et al. (2010) -

Kazanina et al. (2018) -

Tsvilodub et al. (2025) -

Pandian (2025) - AI

Medin et al. (2024) - SSL

Pouw et al. (2024) -

Guriel et al. (2023) -

Gosztolya et al. (2024) - SSL

Pasad et al. (2024) - LibriSpeech

1.5 4. (Information Materials)

1.5.1

- : (Sora Nagano)
- : []
- : 20234
- :
- :
- :

1.5.2

- : 153-8902 3-8-1
- : +81-3-5454-6839
- : s-oswld-n@g.ecc.u-tokyo.ac.jp
- **GitHub**: https://github.com/m02uku

1.5.3

- : Phonological Features in the Deep Learning Era: A Multi-dimensional Investigation of Optimal Representational Units for Language Modeling
- :

1.5.4 200

wav2vec 2.03(1)(2)(3)

85%

1.5.5

- : []
- : []

1.5.6

- : 20263
- : 20265

1.5.7

1.5.8

- :
- : _____

-
- : _____

Astrach, G., & Pinter, Y. (2025). *Probing subphonemes in morphology models* (arXiv:2505.11297). arXiv. https://doi.org/10.48550/arXiv.2505.11297

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

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72.

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

Chang, X., Yan, B., Yoshimoto, Y., Lu, J., Mohamed, A., Du, S., & Watanabe, S. (2024). The interspeech 2024 challenge on speech processing using discrete units. *Proceedings of Interspeech 2024*, 4475–4479.

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* (p. 448). Harper & Row.

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

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). *Unsupervised cross-lingual representation learning for speech recognition* (arXiv:2006.13979). arXiv. https://doi.org/10.48550/arXiv.2006.13979

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

Daland, R. (2015). Long-distance statistical dependencies in natural language: Theory, computation, and neuroscience. *Phonology*, *32*(1), 1–36.

Dunbar, E., Karadayi, J., Bernard, M., Cao, X.-N., Algayres, R., Ondel, L., Besacier, L., Sakriani, S., & Dupoux, E. (2019). The zero resource speech challenge 2019: TTS without t. *Proceedings of Interspeech 2019*, 1088–1092.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *171*, 69–75.

Ebrahimi, M., Hitzler, P., & Sarker, M. K. (2023). Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, *14*(2), 111–141.

Garcez, A. S. d'Avila., Lamb, L. C., & Gabbay, D. M. (2022). *Neural-symbolic cognitive reasoning*. Springer.

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

Gosztolya, G., Kiss-Vetráb, M., Svindt, V., Bóna, J., & Hoffmann, I. (2024). *Wav2vec 2.0 embeddings are no swiss army knife-a case study for multiple sclerosis*.

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

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, *25*(2), 560–585. https://doi.org/10.3758/s13423-017-1362-0

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

Liu, A. T., Hsu, W.-N., Auli, M., & Baevski, A. (2022). Towards automated speech audiometry using self-supervised speech representations. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3169–3173.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association (Interspeech)*, 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

Medin, L. B., Pellegrini, T., & Gelin, L. (2024). Self-supervised models for phoneme recognition: Applications in children's speech for reading learning. *Interspeech 2024*, 5168–5172. https://doi.org/10.21437/Interspeech.2024-1095

Mohamed, A., Lee, H., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T. N., & Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179–1210. https://doi.org/10.1109/JSTSP.2022.3207050

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History. https://phoible.org/

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory features. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3475–3484.

Nguyen, D., Doruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

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

Pandian, S. M. (2025). *Hybrid symbolic-neural architectures for explainable artificial intelligence in decision-critical domains*.

Parcollet, T., Nguyen, H., Evain, S., Boito, M. Z., Pupier, A., Mdhaffar, S., Le, H., Alisamir, S., Tomashenko, N., Dinarelli, M., Zhang, S., Allauzen, A., Coavoux, M., Esteve, Y., Rouvier, M., Goulian, J., Lecouteux, B., Portet, F., Rossato, S., … Besacier, L. (2024). *LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech* (arXiv:2309.05472). arXiv. https://doi.org/10.48550/arXiv.2309.05472

Pasad, A., Chien, C.-M., Settle, S., & Livescu, K. (2024). What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, *12*, 372–391. https://doi.org/10.1162/tacl_a_00656

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*. Blackwell.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on *F*0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, *52*(7), 638–651. https://doi.org/10.1016/j.specom.2010.02.012

Schatz, T., Algayres, R., Dunbar, E., Nguyen, T. A., Lakhotia, K., Chen, M., Mohamed, A., & Dupoux, E. (2021). *The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling*. https://arxiv.org/abs/2011.11588

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

Silverman, D. (2012). *Neutralization*. Cambridge University Press.

Staples, R., & Graves, W. W. (2020). Neural components of reading revealed by distributed and symbolic computational models. *Neurobiology of Language (Cambridge, Mass.)*, *1*(4), 381–401. https://doi.org/10.1162/nol_a_00018

Tesar, B. B. (1995). *Computational optimality theory* [PhD thesis]. University of Colorado at Boulder.

Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, *29*(2), 229–268. https://doi.org/10.1162/002438998553734

Tsvilodub, P., Hawkins, R. D., & Franke, M. (2025). *Integrating neural and symbolic components in a model of pragmatic question-answering* (arXiv:2506.01474). arXiv. https://doi.org/10.48550/arXiv.2506.01474

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

Yang, S., Povey, D., Popov, S., Wang, P., & Khudanpur, S. (2024). *k2SSL: A faster and better framework for self-supervised speech representation learning*. https://arxiv.org/abs/2411.17100

Zhang, X., Dong, D., Meng, S., Li, S., Chen, X., Zhang, Z., Zhou, L., Liu, S., & Wei, F. (2024). SpeechTokenizer: Unified speech tokenizer for speech large language models. *The Twelfth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=AF9Q8Vip84