An Empirical Analysis of Speech Self-Supervised Learning at Multiple Resolutions

Theo Clark, Benedetta Cevoli, Eloy de Jong,
Timofey Abramski, Jamie Dougherty
Speechmatics
`{theoc,benedettac}@speechmatics.com`

Abstract

Self-supervised learning (SSL) models have become crucial in speech processing, with recent advancements concentrating on developing architectures that capture representations across multiple timescales. The primary goal of these multi-scale architectures is to exploit the hierarchical nature of speech, where lower-resolution components aim to capture representations that align with increasingly abstract concepts (e.g., from phones to words to sentences). Although multi-scale approaches have demonstrated some improvements over single-scale models, the precise reasons for these enhancements have poor empirical support. In this study, we present an initial analysis of layer-wise representations in multi-scale architectures, with a focus on Canonical Correlation Analysis (CCA) and Mutual Information (MI). We apply this analysis to Multi-Resolution HuBERT (`MR-HuBERT`) and find that (1) the improved performance on SUPERB tasks is primarily due to the auxiliary low-resolution loss rather than the downsampling itself, and (2) downsampling to lower resolutions neither improves downstream performance nor correlates with higher-level information (e.g., words), though it does improve computational efficiency. These findings challenge assumptions about the multi-scale nature of `MR-HuBERT` and motivate the importance of disentangling computational efficiency from learning better representations.

1 Introduction

Self-supervised learning (SSL) has become a cornerstone in state-of-the-art speech processing models [1, 2, 3]. These models serve as feature extractors or pre-trained encoders for various tasks, including Automatic Speech Recognition (ASR), Speaker Diarisation, Speech Enhancement, and as inputs to Large Language Models. The versatility of a single pre-trained model across multiple downstream tasks has led to concentrated efforts on improving this foundational component.

At the same time, there is growing interest in developing SSL models that more closely emulate human learning processes, as doing so could unlock more efficient and flexible learning mechanisms [4, 5]. While significant differences exist between the human brain and deep learning models, SSL aligns with some aspects of human cognition [6]. One key feature of human learning is the multi-timescale evolution of our world model [7, 8], resulting in a hierarchical learning structure that is more efficient than models operating on a single timescale.

Speech presents a particularly compelling domain for investigating these ideas as it is a mature field with well-established datasets [9, 10, 11, 12] and benchmarks [13] consisting of different downstream tasks that operate most naturally on varying timescales: longer audio sequences are required for tasks like language identification and speaker diarisation, in contrast to phoneme recognition. Speech also exhibits a strong and implicit natural hierarchy [14]: sentences comprise words, which in turn consist of phones and prosodic features.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

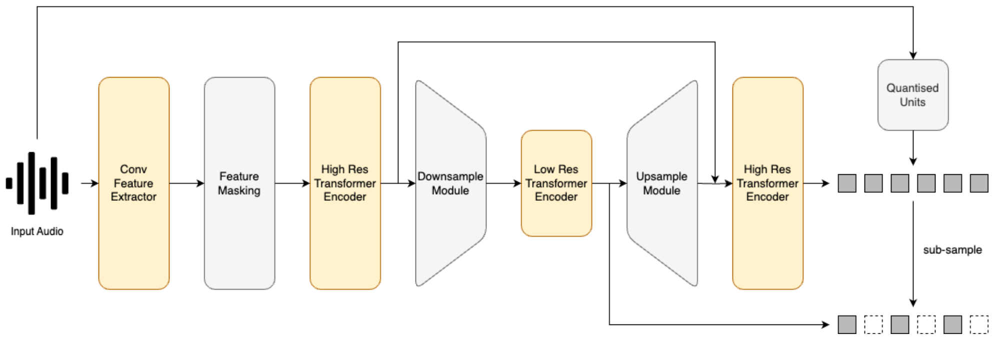


Figure 1: MR-HuBERT framework which incorporates masked unit prediction at multiple resolutions.

Designing architectures that optimally exploit this inherent hierarchy could enhance representation learning efficiency. Multi-scale architectures have been proposed across various domains [8, 15, 16], including speech processing [17, 18, 19, 20]. These approaches typically employ modular designs, with successive modules operating at progressively lower resolutions. Recent works [17, 18, 20, 21, 22, 23] that propose multi-scale architectures for speech processing tasks indicate that increasingly low-resolution representations align with increasingly abstract speech and language components, but these claims currently have limited empirical support.

Multi-Resolution HuBERT (MR-HuBERT) [21] is a multi-scale architecture that augments HuBERT [2] with a low-resolution block and an associated auxiliary loss. MR-HuBERT shows promise across various benchmarks and its success is attributed to a multi-scale structure. By using standard representation analysis techniques to examine these claims, we evaluate whether lower-resolution representations are more correlated with higher level speech and language units in multi-scale models. Our key contributions are:

- Lower-resolution components in MR-HuBERT models do not, as initially hypothesised, capture representations that align with increasingly abstract speech units.
- Downsampling to lower resolutions within MR-HuBERT does not improve downstream performance but improves computational efficiency.
- Improved downstream performance of MR-HuBERT over HuBERT is primarily due to the auxiliary loss located earlier in the network.

2 Multi-Resolution HuBERT

Hidden-Unit BERT (HuBERT) [2] has established itself as the leading architecture for audio SSL models. For this reason, we focus here on Multi-Resolution HuBERT (MR-HuBERT) [21, 24], a model that aims to improve HuBERT through a multi-resolution architecture by introducing:

- **Downsample and upsample modules** between encoder blocks to process features at different resolutions (skip connections are applied to link encoders of the same resolution);[1]
- **An auxiliary loss at low-resolutions**, applied at the end of each decoder and computed through a projection layer. Targets are formed by sub-sampling the base target stream.

We illustrate MR-HuBERT's architecture in Fig. 1. In this paper, we analyse the layer-wise acoustic and linguistic information content of a series of ablations of the MR-HuBERT-base model[2], listed in Table 1 , and HuBERT-base[3]. To do so, we employ methods used in previous representation analysis studies [25, 26, 27], such as Canonical Correlation Analysis (CCA) [28], Mutual Information (MI) [29], and spoken Semantic Textual Similarity (STS) [30][4]. We also run Speech processing Universal PERformance Benchmark (SUPERB) [13] downstream tasks and analyse learnt layer weightings. We provide further details on our methodology in Appendix A.

[1]We note a potential error in the official implementation of MR-HuBERT (see Appendix C).

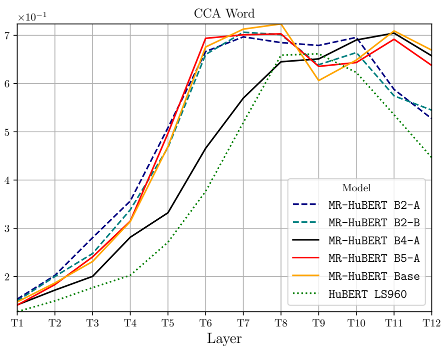
[2]MR-HuBERT models are downloaded from the Fairseq MR-HuBERT page.

[3]HuBERT models are downloaded from the Fairseq HuBERT page.

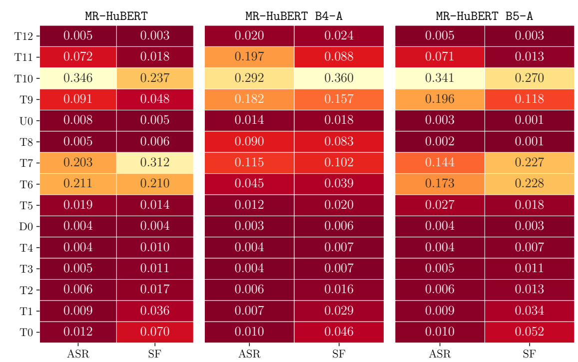
[4]We use the official implementation, available on the Layerwise Analysis repository.

Model	Resolutions (ms)	Layers	Downsampling	Auxiliary loss
HuBERT-base	20	12	✗	✗
MR-HuBERT-base[5]	20, 40	4, 4, 4	✓	✓
MR-HuBERT B2-a	20, 40, 80	3, 2, 2, 2, 3	✓	✓
MR-HuBERT B2-b	20, 40, 80	2, 2, 4, 2, 2	✓	✓
MR-HuBERT B4-a	20, 40	4, 4, 4	✓	✗
MR-HuBERT B5-a	20	4, 4, 4	✗	✓

Table 1: HuBERT and MR-HuBERT models used for analysis in this work. The total number of layers is the same for all models. In B2-a and B2-b a third resolution is introduced. B4-a is only trained on a single loss. B5-a has a single resolution but retains an auxiliary loss.



(a) CCA word-level similarity.



(b) SUPERB layer importance-weightings.

Figure 2: Impact of auxiliary loss, downsampling and added resolutions on information content and importance in downstream performance. Fig. 2a shows CCA scores for HuBERT and multiple MR-HuBERT variants. Comparing these models, we see that the auxiliary loss is the primary factor in increasing the word level information in earlier layers. Fig. 2b shows SUPERB weights for the ASR and SF tasks, and again shows that the auxiliary loss is responsible for middle layers being useful for downstream tasks. [6]

3 Findings

3.1 Lower-resolution layers fail to capture abstract speech units

In Fig. 2a, we show the layerwise word-level CCA values of HuBERT and MR-HuBERT models. We observe most MR-HuBERT models feature two peaks: one near the end of the network (a feature of HuBERT models generally), and another near the middle (a feature of other SSL models [26]). Notably, downsampling alone does not change this pattern. We see no difference in word-level CCA between MR-HuBERT-base (two-resolutions, downsampling) and B5-a (single resolution, no downsampling) nor do we see differences between MR-HuBERT and three-resolution ablations (B2-a and B2-b).

This pattern is consistent across other word-level measures (see Fig. 3 and 5 for further plots on other metrics and model sizes) as well as different speech units. We see no increase in learned word (Fig. 2), phone or semantic (Fig. 3) information in MR-HuBERT when downsampling to various degrees (MR-HuBERT-base, B2-a, B2-b, B4-a) compared to not downsampling at all (B5-a).

These results suggest that downsampling at these rates does not affect the information content of the representations learned. Most importantly, downsampling does not align with more abstract speech units. Whilst we find that middle layers of MR-HuBERT are more heavily associated with more

[5] MR-HuBERT-base refers to the mono-base model on the Fairseq MR-HuBERT page.

[6] To explain the difference in number of layers between Figs. 2a and 2b: as discussed in appendix D.4 of [21], MR-HuBERT encompasses transformer layers as well as outputs of the sampling modules, so a two-resolution MR-HuBERT adds two layers, denoted by D0 and U0 in Fig. 2b. Additionally, Fig. 2a does not include the layer before the first transformer layer, denoted by T0.

Model	ASR (WER ↓)	SF (F1/CER ↑ / ↓)	SE (STOI [34]/PESQ [35] ↑ / ↑)	IC (Acc ↑)	KS (Acc ↑)	SD (Acc ↑)
HuBERT-base^{+}	6.34	**89/23**	0.93/2.55	98.4	96.5	N/A
MR-HuBERT-base	5.85	89/24	0.94/2.53	**98.6**	95.7	94.8
MR-HuBERT B4-a	6.35	89/24	0.94/2.53	98.1	**96.7**	**95.1**
MR-HuBERT B5-a	**5.82**	88/26	**0.94/2.55**	98.3	96.3	94.9

Table 2: Performance on SUPERB downstream tasks with various upstream models based on MR-HuBERT. The results for HuBERT-base^{+} are taken from [21].

abstract information such as words, this appears to be independent of downsampling (B5-a) and unaffected by the resolution at which these layers operate, see e.g. Fig. 2b.

3.2 Down-sampling is only helpful from an efficiency perspective

Not only does downsampling in MR-HuBERT not enhance the information content of representations, it also does not improve downstream performance. Table 2 shows that performance in downstream tasks is hardly affected when downsampling is removed (B5-a). This suggests downsampling is not responsible for improvements seen in MR-HuBERT [21]. Nevertheless, it is important to note that downsampling is still useful for improving model inference speed and training time. The current downsampling methods, however, appear too limited in scope to effectively capture broader linguistic units which naturally vary across time scales up to 50 times larger [17]. This suggests more aggressive, context-aware downsampling techniques [31] could better capture higher-level speech information, leading to both improvements in downstream performance as well as further efficiency gains.

3.3 The auxiliary loss improves downstream performance

In contrast, the removal of the auxiliary loss impacts our analysis significantly. We see worse performance of the B4-a model compared to MR-HuBERT-base and B5-a on ASR tasks in Table 2. We also see clear differences in the content of the model representations in Fig. 2 which could explain the observed difference in performance. In Fig. 2a, we find the additional early peak typical of MR-HuBERT entirely disappears when the auxiliary loss is removed (B4-a; see also Fig 3). Moreover, middle layers of the network are more useful for phonetic-based tasks, such as ASR and Slot Filling (SF), when the auxiliary loss (B5-a) is present, independent of downsampling as shown in Fig. 2b. We also find that B4-a results are closer to those of HuBERT than any other MR-HuBERT model. This is the case for both ASR performance as shown in Table 2) as well as CCA scores as shown in 2a[7].

These results strongly suggest that the auxiliary loss is the key driver of downstream performance improvements [21]. By encouraging the model to learn more diverse and relevant features at earlier layers, the auxiliary loss enhances the model's ability to capture crucial phonetic and linguistic information. Notably, this loss mirrors the approach used in Deeply Supervised Nets [32], where early losses are thought to improve gradient flow and feature robustness. It may also act as a regulariser [33], helping the model learn more stable, generalised representations by adding constraints during training — a benefit that could be especially important in low-resource settings like LibriSpeech.

4 Conclusion

In this study, we find that the improved downstream performance of MR-HuBERT is primarily due to the auxiliary loss function, rather than downsampling in the multi-resolution architecture. Empirically, the auxiliary loss promotes better learning in intermediate layers, leading to superior downstream task performance. While downsampling enhances computational efficiency, it does not improve linguistic representations or downstream performance. Additionally, we find no evidence that lower-resolution layers capture more abstract speech information, highlighting the need for more effective unsupervised learning. This paper highlights the importance of analysing representation quality to gain deeper insights into how well multi-scale architectures capture different abstractions of speech information. We leave the exploration of improved architectures based on this analysis to future work.

[7] Remaining differences between B4-a and HuBERT may be due to MR-HuBERT's extra training iteration [21]. This may also explain why the peak scores for most CCA metrics are higher for MR-HuBERT than HuBERT.

5 Acknowledgements

The authors would like to thank Will Williams, John Hughes, Akis Kefalas and Ana Olssen for their valuable help in providing feedback and guidance on earlier drafts of this paper.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[4] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[5] Kjell Jørgen Hole and Subutai Ahmad. A thousand brains: toward biologically constrained ai. *SN Applied Sciences*, 3(8):743, 2021.

[6] Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9960–9971. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/7183145a2a3e0ce2b68cd3735186b1d5-Paper.pdf`.

[7] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

[8] Ramy Mounir, Sujal Vijayaraghavan, and Sudeep Sarkar. STREAMER: Streaming representation learning and event segmentation in a hierarchical manner. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=EfTMRQn00d`.

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[10] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020. doi: 10.1109/icassp40776.2020.9052942. URL `http://dx.doi.org/10.1109/ICASSP40776.2020.9052942`.

[11] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL `https://aclanthology.org/2021.acl-long.80`.

[12] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.520`.

[13] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

[14] Ray Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 01 2002. ISBN 9780198270126. doi: 10.1093/acprof:oso/9780198270126.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780198270126.001.0001`.

[15] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=S1di0sfgl`.

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[17] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:775–788, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2023.3235194. URL `http://dx.doi.org/10.1109/TASLP.2023.3235194`.

[18] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533, jun 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3288409. URL `https://doi.org/10.1109/TASLP.2023.3288409`.

[19] Tae Park, Nithin Koluguri, Jagadeesh Balam, and Boris Ginsburg. Multi-scale speaker diarization with dynamic scale weighting. pages 5080–5084, 09 2022. doi: 10.21437/Interspeech.2022-991.

[20] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, sheng zhao, Jiang Bian, Xixin Wu, Zhou Zhao, Shinji Watanabe, and Helen M. Meng. Uniaudio: An audio foundation model toward universal audio generation, 2024. URL `https://openreview.net/forum?id=nhgTmx1TZJ`.

[21] Jiatong Shi, Hirofumi Inaguma, Xutai Ma, Ilia Kulikov, and Anna Sun. Multi-resolution huBERT: Multi-resolution speech self-supervised learning with masked unit prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=kUuKFW7DIF`.

[22] Santiago Cuervo, Adrian Łańcucki, Ricard Marxer, Paweł Rychlikowski, and Jan Chorowski. Variable-rate hierarchical cpc leads to acoustic unit discovery in speech. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

[23] Saurabhchand Bhati, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Unsupervised speech segmentation and variable rate representation learning using segmental contrastive predictive coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2002–2014, 2022. doi: 10.1109/TASLP.2022.3180684.

[24] Jiatong Shi, Yun Tang, Hirofumi Inaguma, Hongyu Gong, Juan Pino, and Shinji Watanabe. Exploration on hubert with multiple resolution. pages 3287–3291, 08 2023. doi: 10.21437/Interspeech.2023-1337.

[25] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921, 2021. doi: 10.1109/ASRU51503.2021.9688093.

[26] Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096149.

[27] Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, 12:372–391, 04 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00656. URL `https://doi.org/10.1162/tacl_a_00656`.

[28] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992.

[29] Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1448. URL `https://aclanthology.org/D19-1448`.

[30] Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1269`.

[31] Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Lukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. Hierarchical transformers are more efficient language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1559–1571, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.117. URL `https://aclanthology.org/2022.findings-naacl.117`.

[32] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 562–570, San Diego, California, USA, 09–12 May 2015. PMLR. URL `https://proceedings.mlr.press/v38/lee15a.html`.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010. doi: 10.1109/ICASSP.2010.5495701.

[35] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023.

[36] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

[37] Shane Settle, Kartik Audhkhasi, Karen Livescu, and Michael Picheny. Acoustically grounded word embeddings for improved acoustics-to-word speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5641–5645, 2019. doi: 10.1109/ICASSP.2019.8682903.

[38] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9. URL `http://www.elementsofinformationtheory.com/`.

[39] Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Semantic sentence similarity: Size does not always matter. In *Interspeech 2021*, pages 4393–4397, 2021. doi: 10.21437/Interspeech.2021-1464.

[40] Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, et al. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 228–235. IEEE, 2021.

[41] Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, et al. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6152–6156. IEEE, 2022.

[42] Kuo-Hsuan Hung, Szu wei Fu, Huan-Hsin Tseng, Hsin-Tien Chiang, Yu Tsao, and Chii-Wann Lin. Boosting Self-Supervised Embeddings for Speech Enhancement. In *Proc. Interspeech 2022*, pages 186–190, 2022. doi: 10.21437/Interspeech.2022-10002.

[43] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Zhuo Chen, Peidong Wang, Gang Liu, Jinyu Li, Jian Wu, Xiangzhan Yu, and Furu Wei. Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition? In *Proc. Interspeech 2022*, pages 3699–3703, 2022. doi: 10.21437/Interspeech.2022-10019.

[44] Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Proc. Interspeech 2023*, pages 884–888, 2023. doi: 10.21437/Interspeech.2023-1316.

[45] Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE, 2023.

[46] Shinta Otake, Rei Kawakami, and Nakamasa Inoue. Parameter efficient transfer learning for various speech processing tasks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

A Analysis methods

In this section, we discuss the various metrics used to assess the acoustic and linguistic information present in the representations of different layers of a self-supervised model.

A.1 Canonical correlation analysis

Following previous layer-wise comparative studies [25, 26], we employ Projection Weighted Canonical Correlation Analysis (PWCCA), referred to throughout this paper as simply CCA, in order to correlate the model's internal representations with phonetic and word information and investigate how this varies across layers in the model. The internal representation for each word/phone is calculated by averaging the model's representations across the time steps corresponding to the span of that word/phone in the input sequence. Averaging across the time dimension effectively condenses the sequence information into a single vector representation for each word/phone, facilitating a more straightforward comparison of model behaviour across different layers. This process is then repeated to compare these internal representations against a range of external representations, capturing different linguistic and phonetic characteristics. Specifically, we perform comparisons using the following sets of representations: CCA mel (representations based on MFCCs to capture phonetic features), CCA phone (one-hot encoded phoneme embeddings), CCA word (one-hot encoded word embeddings), CCA glove (GloVe word embeddings [36] to capture semantic similarity), CCA agwe (acoustically grounded word embeddings [37] reflecting spoken word characteristics). For each of these, we follow [26] by using 7000 samples of words/phones from Librispeech.

A.2 Mutual information

Mutual information (MI) measures the information one random variable contains about another random variable. High MI is equivalent to a large reduction in uncertainty of one random variable given knowledge of the other, which implies dependence [38].

$$\mathrm{I}(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

We assess the dependence of phone and word labels on hidden representations in `MR-HuBERT` as described in [25, 29]. We first obtain averaged model features (as described above for CCA) which are then clustered using K-means to obtain a discrete distribution for MI analysis. Similarly to [26], we cluster phone representations with $k = 500$ and word representations with $k = 5000$ centres.

A.3 Spoken sentence-level semantic textual similarity

The spoken Semantic Textual Similarity (STS) allows us to examine the extent to which SSL representations capture utterance-level semantic content [39]. Following [27] we calculate Spearman's ρ correlation between annotated human judgments and the predicted similarity scores of utterance pairs. Sentence-level similarity scores are extracted by taking the cosine similarity between the mean-pooled representation of each utterance in a pair [27].

A.4 SUPERB

Speech processing Universal PERformance Benchmark (SUPERB) [13] is a set of benchmarking resources to evaluate the performance of a shared model across a variety of speech processing tasks. We report on the following SUPERB downstream tasks to give results across a broad spectrum of speech-related tasks: Automatic Speech Recognition (ASR), Slot Filling (SF), Speech Enhancement (SE), Intent Classification (IC), Keyword Spotting (KS) and Speaker Diarisation (SD).

We examine downstream performance as well as the learned weightings inside the downstream adaptor for each layer in the pre-trained model to gain insights into where the most useful information is located for specific downstream tasks. When training downstream models, we use the default hyperparameters, including the learning rate. SUPERB learns a weighted average of the representations from the different layers from the self-supervised upstream model. Following previous work [40, 41, 42, 43, 44, 45, 24, 46, 3], we use these learned weightings to determine if certain layers

Model	Final Train Loss	Final Validation Loss
`post-residual (baseline)`	7.312	6.784
`pre-residual`	7.302	6.757

Table 3: Effect on pre-train losses of altering the residual connection when added before the decoder.

contain significant information important to a specific downstream task and if so, which layers those are.

B Layer-wise analysis of single and multi-resolution models

In this section, we present the metric-specific results of the layer-wise analyses conducted on `MR-HuBERT` ablations (see Table 1) and `HuBERT` baselines. In addition to the findings reported in the paper, we observe in Fig. 4 that consistent with [25], the correlation between frame-level representations and fbanks increases with depth in the convolution layers of the feature extractor, but then decreases towards the mid-transformer layers for both `HuBERT` and `MR-HuBERT` models. While there are no significant differences between two- and three-resolution models in frame, word, and sentence-level metrics, we do see a notable decrease in phone-level scores in the mid-layers of the three-resolution models (specifically `B2-a` and `B2-b` in panels B and C of Fig. 3).

C Modifying the residual connection

The diagram and equations from the `MR-HuBERT` paper [21] show that the residual for a given resolution is added before the decoder. However, the official implementation[8] contradicts this and adds the residual *after* the decoder. We do not modify this in our experiments to retain consistency with the original results. However, we ran separate experiments which show an improvement in pre-train validation losses when the residual is added before the decoder (see Table 3).

These exploratory experiments used smaller `MR-HuBERT` model sizes due to resource constraints. Models were trained for only 10% of the usual 400k steps and the changes in architecture compared to `MR-HuBERT`-base are as follows: layers per encoder: 2, encoder embedding dim: 192, encoder feed-forward dim: 768.

D SUPERB layer weight analysis

Here, we provide further details on our layer weightings analysis of various SUPERB downstream tasks for a subset of the models listed in table 1. We show a layer weight analysis for `MR-HuBERT`, `B4-a` and `B5-a` in Figs. 6a, 6b and 6c respectively to ablate the effects of auxiliary loss and the down- and upsampling modules further. The layer weightings in Fig. 6a support the same conclusions as in [21], e.g., `MR-HuBERT` allocates over 40% of its attention to low-resolution layers 8 and 9 for ASR. As discussed in the main text, this number decreases when downsampling is removed. We see a similar effect for the SF task, where focus is shifted away from the low-resolution encoder towards the second high-resolution encoder. The low-resolution `MR-HuBERT` layers are associated to semantic context in the data and these results suggest that these semantics are pushed into the middle layers by training on the low-resolution loss and to a lesser extent by the downsampling.

The SE task generally focuses on the early layers - at least 66% of the weightings are assigned to the first three layers in all models. All the layers are used relatively evenly for SD and KS across all models. Weightings are slightly less concentrated towards the end of the network for `B5-a` compared to the other models.

[8] https://github.com/facebookresearch/fairseq/blob/main/fairseq/models/multires_hubert/multires_hubert.py#L783

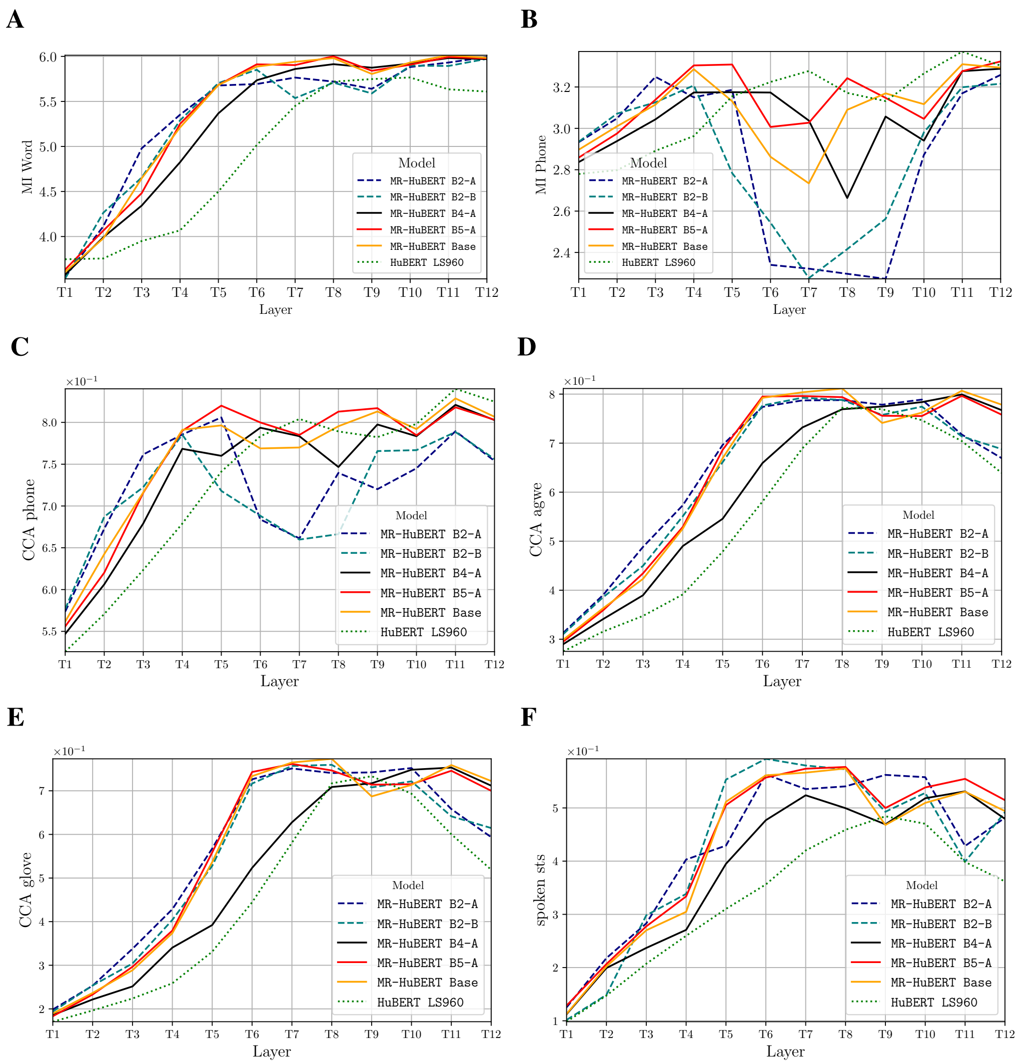


Figure 3: Layer-wise analyses of base models of `MR-HuBERT` and `HuBERT` models. (A) MI scores between mean-pooled word-level representations and word identities. (B) MI scores between mean-pooled phone-level representations and phone identities. (C) CCA similarity between mean-pooled phone-level representations and phone identities (one-hot encoded). (D) CCA similarity between mean-pooled word-level representations and AGWE embeddings. (E) CCA similarity between mean-pooled word-level representations and GloVE embeddings. (F) Spearman's ρ correlation between annotated human judgments and cosine similarity of spoken utterance pairs.

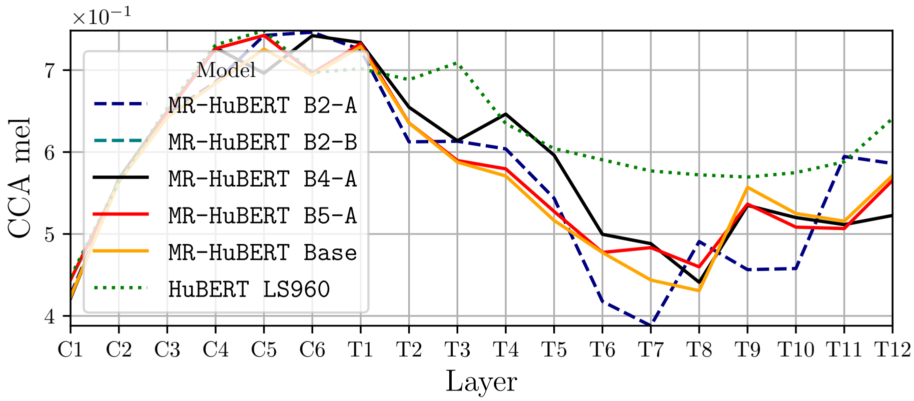


Figure 4: CCA similarity between frame-level representations and fbanks.

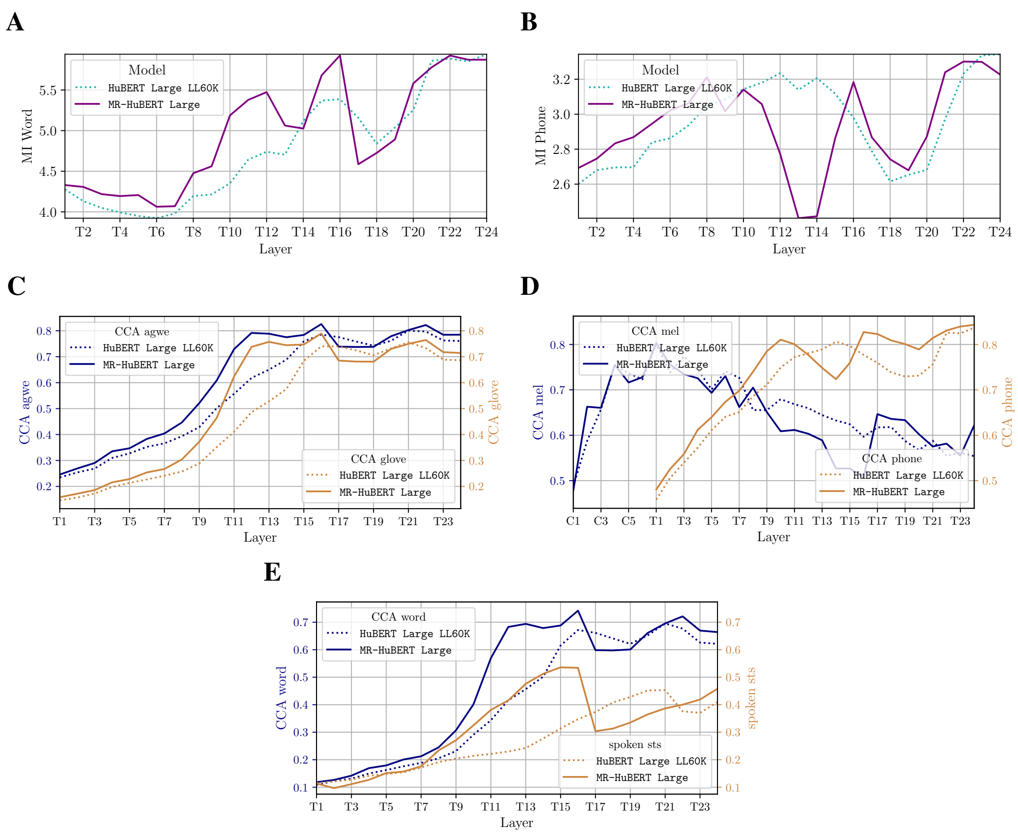
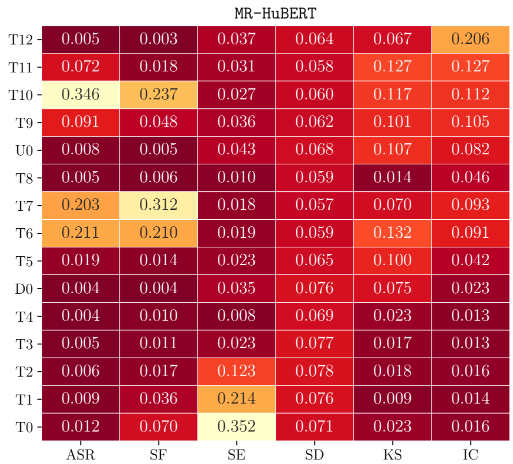
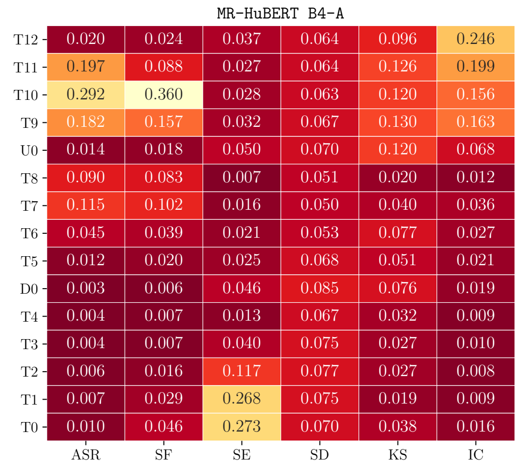


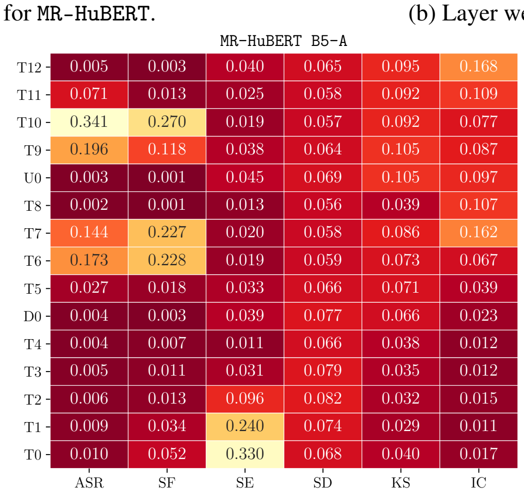
Figure 5: Layer-wise analyses of large models of MR-HuBERT and HuBERT models. (A) MI scores between mean-pooled word-level representations and word identities (one-hot encoded). (B) MI scores between mean-pooled phone-level representations and phone identities (one-hot encoded). (C) CCA similarity between mean-pooled word-level representations and AGWE embeddings and GloVE embeddings. (D) CCA similarity between mean-pooled frame-level representations and fbanks as well as phone-level representations and phone identities (one-hot encoded). (E) CCA similarity between mean-pooled word-level representations and word identities (one-hot encoded) as well as Spearman's ρ correlation between annotated human judgments and cosine similarity of spoken utterance pairs.



(a) Layer weights for `MR-HuBERT`.



(b) Layer weights for `B4-a`.



(c) Layer weights for `B5-a`.

Figure 6: Layer importance-weightings for all SUPERB downstream tasks we study in this work, for `MR-HuBERT` and two of its ablations.