INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech

Badr M. Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, Dietrich Klakow

Language Science and Technology (LST), Saarland University, Germany
Saarland Informatics Campus, Germany
{babdullah|mmshaik|moebius|dietrich}@lsv.uni-saarland.de

Abstract

Self-supervised representation learning for speech often involves a quantization step that transforms the acoustic input into discrete units. However, it remains unclear how to characterize the relationship between these discrete units and abstract phonetic categories such as phonemes. In this paper, we develop an information-theoretic framework whereby we represent each phonetic category as a distribution over discrete units. We then apply our framework to two different self-supervised models (namely, wav2vec 2.0 and XLSR) and use American English speech as a case study. Our study demonstrates that the entropy of phonetic distributions reflects the variability of the underlying speech sounds, with phonetically similar sounds exhibiting similar distributions. While our study confirms the lack of direct one-to-one correspondence, we find an intriguing indirect relationship between phonetic categories and discrete units.
Index Terms: discrete speech representations, self-supervised learning, information theory

1. Introduction

Self-supervised learning (SSL) for the speech modality is an active area of research that aims to develop models that build meaningful speech representations from raw audio without any explicit labels or transcriptions (see [1] for an overview). These models can be further adapted for downstream tasks such as automatic speech recognition and speaker identification, and have become the state-of-the-art approach even when limited labeled data are available [2, 3, 4, 5]. Recently, it has become a common practice to include a quantization module within the architecture of SSL speech models that transforms the acoustic input into a sequence of discrete entities. Besides representing the complex acoustic signal in a compact and computationally efficient manner, learning discrete representations of speech can also facilitate training large SSL speech models using a masked language modeling objective similar to those employed in natural language processing (e.g., BERT [6]).

Nevertheless, the nature of the discrete units learned via self-supervision remains an under-explored area of research. A key question is whether these discrete representations correspond to abstract phonetic categories such as phonemes. A few recent studies have investigated the discrete units from a neural network interpretability point of view [7, 8, 9, 10]. The analysis in [9] showed that the discrete units correspond to low-level "sub-phonetic" events—rather than high-level phonetic categories—since they are sensitive to context-dependent and non-phonemic variations in speech. In [10], the authors concluded that there exists a strong correspondence between discrete units and phonemes, and attributed the lack of consistent phoneme-to-unit mapping to variations in phonological contexts. These findings seem to be contradictory and rely on different definitions of the term "phoneme", and thus remain inconclusive.

Although information theory was initially proposed as a mathematical theory of communication [11], it also provides a quantitative framework for measuring the amount of information conveyed by linguistic units, such as words or sounds. Information theory has been adopted as a framework to study various aspects of linguistic structure, including phonology [12, 13], morphology [14, 15], and syntax [16, 17]. In this paper, we build on this line of research and develop information-theoretic metrics to analyze the correspondence between phonetic categories and discrete units. Concretely, we make the following contributions:

- We develop an empirical approach to represent each phonetic category as a probability distribution over discrete units using two self-supervised pre-trained models: English wav2vec 2.0 (henceforth W2V2) and multilingual wav2vec-XLSR (henceforth XLSR) (§2).
- We characterize each phonetic category using the notion of information entropy and demonstrate that entropy quantifies acoustic-phonetic variability (§4).
- We quantify the dissimilarity between phonetic distributions using Jensen-Shannon divergence and illustrate that this metric highly reflects feature-based phonetic similarity (§5).

2. Research methodology

2.1. Speech quantization via self-supervised learning

Consider a continuous acoustic signal represented as a sequence of T acoustic frames $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Here, $\mathbf{x}_t$ could either be an interval of the raw waveform or a spectral vector such as MFCCs. Given a pre-trained speech encoder, the signal $\mathbf{x}$ is first transformed via a local, temporal convolutional encoder $\mathcal{F} : \mathcal{X} \mapsto \mathcal{Z}$ into a sequence of latent speech representations in a continuous space as $\mathcal{F}(\mathbf{x}) = \mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$, where $\mathbf{z}_t \in \mathbb{R}^d$. As a part of the quantization step, the sequence of continuous representations gets discretized to produce a sequence of discrete units $\mathcal{D}(\mathbf{z}) = \boldsymbol{\omega} = (\omega_1, \ldots, \omega_T)$, where $\mathcal{D} : \mathcal{Z} \mapsto \Omega$ is a vector-to-centroid mapping and $\omega_t \in \Omega$ is the index of the centroid. Here, we use Ω to denote the finite set of discrete units within the model codebook. During pre-training using masked learning objectives, the corresponding quantized representations of these discrete units become the targets of the model prediction.

2.2. Phonetic categories as distributions over discrete units

Consider a speech corpus that is transcribed and aligned to phonetic segments given an inventory of phonetic categories Φ. In this scenario, a phonetic category can be considered as a set of K different acoustic exemplars obtained from the corpus,

10.21437/Interspeech.2023-2131

$\boldsymbol{\varphi} = \{\varphi^1, \ldots, \varphi^K\}$. These exemplars represent different acoustic realizations of the underlying phonetic category, and should optimally be produced by various speakers in diverse phonological contexts. Using the feature encoder and quantization module of a self-supervised speech model, we transform the associated acoustic segments of all exemplars $\{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$ into a discrete representation to obtain a collection of discrete sequences $\{(\omega_1^1, \ldots, \omega_{\tau_1}^1), \ldots, (\omega_1^k, \ldots, \omega_{\tau_k}^k)\}$ for each phonetic category. We then discard the exemplar identity as well as the sequential nature of each discrete sequence and view each phonetic category as a bag of discrete units. In this approach, each phonetic category can be described as a frequency distribution over the units in Ω. To facilitate our information-theoretic analysis, we turn the frequency distribution into a probability distribution where the probability of observing a discrete unit ω under a phonetic category $\boldsymbol{\varphi}$ is calculated using maximum likelihood estimation as follows

$$\boldsymbol{p}_{\varphi}(\omega_i) = \frac{N_{\varphi}(\omega_i)}{\sum_{\pi \in \Omega} N_{\varphi}(\pi)} \tag{1}$$

Here, $N_{\varphi} : \Omega \mapsto \mathbb{Z}^+$ is a function that returns the number of occurrences of a discrete unit under the phonetic category $\boldsymbol{\varphi}$, and therefore $\boldsymbol{p}_{\varphi} : \Omega \mapsto [0, 1]$ is a probability mass function defined over Ω such that $\sum_{\omega \in \Omega} \boldsymbol{p}_{\varphi}(\omega) = 1$. Note that each phonetic category in our analysis has its own $\boldsymbol{p}_{\varphi}$ and N_{φ} functions. For example, the vowels /æ/ and /ɔ/ are represented as two empirical distributions $\boldsymbol{p}_{/æ/}$ and $\boldsymbol{p}_{/ɔ/}$, respectively. Given our representation of a phonetic category as a distribution over discrete units $\boldsymbol{p}_{\varphi}$, we can employ information-theoretic metrics to characterize each phonetic distribution. For simplicity, we henceforth omit the subscript notation in $\boldsymbol{p}_{\varphi}$ and use $\boldsymbol{p}$ to denote a distribution associated with a single phonetic category.

3. Experimental data and models

Speech data. We use the TIMIT speech corpus which consists of recordings from 630 American English speakers each speaking 10 different sentences, for a total of 6,300 sentences covering a diverse range of ages, genders, and regional accents from across the United States [18]. Following [19], the original phonetic categories of TIMIT annotation are mapped to the reduced set of 40 categories. We exclude silences and closures from our analysis.

SSL speech models. We conduct our analysis using two publicly available (via the HuggingFace Model Hub) SSL speech models: (1) monolingual English wav2vec 2.0-BASE [4], which is a 12-layer transformer model, and (2) multilingual wav2vec XLSR-53-LARGE [20], which is a 24-layer transformer model trained on different languages. Both models employ two codebooks with 320 discrete units each, for a total of 640 units in each model. We consider the concatenation of the two codebooks as the set of discrete units in our analysis, thus $|\Omega| = 640$.

Code and reproducibility. Our analysis code is publicly available on GitHub[1].

4. Analysis I: Phonetic variability as information entropy

4.1. Information content and entropy

For any discrete unit within the codebook $\omega \in \Omega$, we measure its information content, or surprisal under a specific phonetic category as

$$\eta(\omega) = -\log_2 \boldsymbol{p}(\omega) \tag{2}$$

which quantifies the unexpectedness of the discrete unit to be observed under the phonetic category associated with the distribution $\boldsymbol{p}$. It is measured in bits. The uncertainty or "randomness" of the distribution $\boldsymbol{p}$ can be quantified as the average surprisal, or entropy

$$H(\boldsymbol{p}) = \sum_{\omega \in \Omega} \boldsymbol{p}(\omega)\, \eta(\omega) \tag{3}$$

where $0 \leq H(\boldsymbol{p}) \leq \log_2 |\Omega|$. If all acoustic realizations of a phonetic category are associated with a single discrete unit, then its entropy is minimal $H(\boldsymbol{p}) = 0$. On the other hand, a distribution of a phonetic category is maximally entropic (i.e., $H(\boldsymbol{p}) = \log_2 |\Omega|$) when all discrete units are equally likely to be aligned to this category. Therefore, entropy can be viewed as a measure of (within-category) acoustic-phonetic variability in our case. That is, the more entropic a phonetic category is, the higher the difficulty of predicting its alignment to discrete units. Note that our measure of variability is similar to the measure of diversity (i.e., the unit purity measure) introduced in [5], but we express the variability of phonetic distributions using information-theoretic metrics.

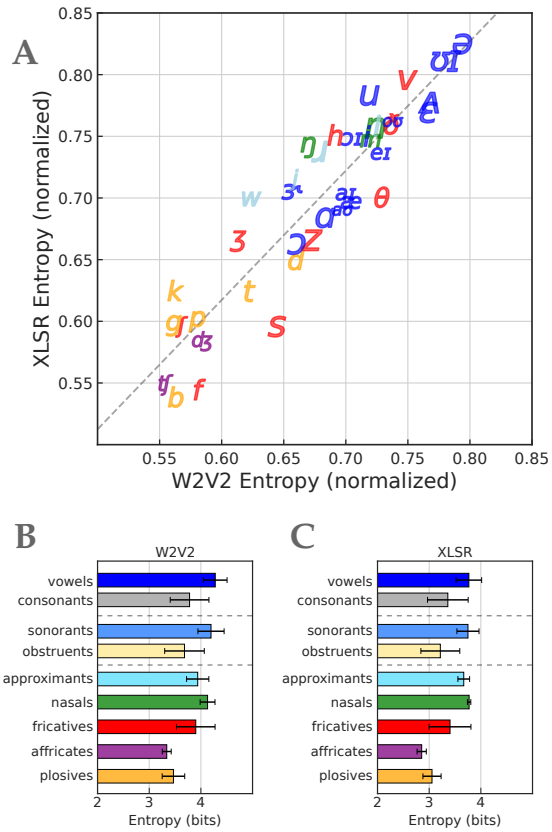

Figure 1: *(A) The (normalized) entropy of each phonetic category in* W2V2 *(x-axis) vs* XLSR *(y-axis). (B-C) The entropy of several selected articulatory classes in* W2V2 *(B) and* XLSR *(C).*

4.2. Entropy per phonetic category

We compute the entropy of each phonetic category using Eq. 3. First, we find that phonetic categories are more entropic on average under W2V2 (mean $H = 3.97$) compared to XLSR

[1] https://github.com/uds-lsv/phone2unit

(mean $H = 3.52$). After inspecting the phone-to-unit alignment of the TIMIT corpus, we attribute this behavior to different utilization of the codebooks across the two models. While there are 56.6% of the discrete units under W2V2 with non-zero counts across all phonetic categories, only 24.2% of the units have non-zero counts under XLSR. This difference gets reflected in lower entropy values in XLSR compared to W2V2.

Fig. 1 illustrates the results of our analysis with entropy as a measure of phonetic variability. Fig. 1A shows the entropy of each phonetic category in W2V2 (x-axis) and XLSR (y-axis). We report the normalized entropy in Fig. 1A to account for differences in entropy values between the two models. In addition, we group phonetic categories according to several articulatory classes, average the entropy over the categories within each class, and depict the result for W2V2 (Fig. 1B) and XLSR (Fig. 1C). From Fig. 1A, we observe a strong correlation between the two models (Pearson's $r = 0.92, p \ll 0.001$). When considering entropy values, we see that none of the phonetic categories is minimally entropic (i.e, $H(\boldsymbol{p}) = 0$), which confirms the findings in the literature about the lack of one-to-one correspondence between high-level abstract phonetic categories and discrete units in self-supervised speech models.

Regarding the variation of entropy across different phonetic categories, we observe that vowels tend to be more entropic than consonants in W2V2 ($H_V = 4.28 > H_C = 3.78$) and XLSR ($H_V = 3.77 > H_C = 3.36$). This reflects a higher variability in the acoustic realizations of vowels compared to consonants, since vowels are subject to a higher degree of variation due to vowel reduction in unstressed syllables and co-articulation, as well as other factors such as cross-speaker and dialect variability [21, 22, 23]. For consonants, the nasal sounds (i.e., /n, m, ŋ/) are the most entropic consonant group, followed by the approximant sounds (i.e., /l, j, w, ɹ/), and then by the fricative sounds (i.e., /ð, z, ʒ, v, θ, s, ʃ, f, h/). We also observe that resonating consonants (i.e., nasals and approximants) exhibit higher variability on average than obstruents (i.e., plosives, fricatives, and affricates). Furthermore, we find an effect of voicing on variability since the voiced fricatives (i.e., /ð, z, ʒ, v/) are more entropic than their voiceless counterparts (i.e., /θ, s, ʃ, f/). For example, consider the voiceless-voiced contrast /f-v/ where /v/ is substantially more entropic than /f/ under W2V2 ($H(/\text{v}/) = 4.40 > H(/\text{f}/) = 3.41$) and XLSR ($H(/\text{v}/) = 4.01 > H(/\text{f}/) = 2.75$). This effect of voicing can be explained by the presence of low-frequency voicing energy in voiced fricatives which is likely to vary due to cross-speaker variability. Finally, the affricates (i.e., /dʒ, tʃ/) are found to be the least entropic consonant category under both W2V2 ($H = 3.33$) and XLSR ($H = 2.86$).

5. Analysis II: Phonetic dissimilarity as Jensen-Shannon divergence

5.1. Relative entropy and divergence

Consider two phonetic distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ that are defined over the same set of discrete units Ω. To quantify how different $\boldsymbol{p}$ is from $\boldsymbol{q}$, we measure the expected surprisal from using $\boldsymbol{q}$ as a model distribution when the true distribution is $\boldsymbol{p}$. This quantity is known as the relative entropy or Kullback–Leibler divergence

$$D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{q}) = -\sum_{\omega\in\Omega} \boldsymbol{p}(\omega)\log_2 \frac{\boldsymbol{q}(\omega)}{\boldsymbol{p}(\omega)} \tag{4}$$

Here, $D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{q}) \geq 0$, with $D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{q}) = 0$ only if $\boldsymbol{p} = \boldsymbol{q}$. Note that relative entropy is not symmetric, that is, $D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{q}) \neq D_{KL}(\boldsymbol{q}\,||\,\boldsymbol{p})$. Since a symmetric metric is more suitable for our analysis, we therefore measure the distance between two probability distributions using Jensen-Shannon divergence (JSD)

$$D_{JS}(\boldsymbol{p}\,||\,\boldsymbol{q}) = \frac{1}{2}D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{m}) + \frac{1}{2}D_{KL}(\boldsymbol{q}\,||\,\boldsymbol{m}) \tag{5}$$

where $\boldsymbol{m} = \frac{1}{2}\boldsymbol{p} + \boldsymbol{q}$ and $0 \leq D_{JS}(\boldsymbol{p}\,||\,\boldsymbol{q}) \leq 1$. Here, our goal is to investigate the degree to which the distance between distributions reflects phonetic similarity. Therefore, we use JSD as a measure of phonetic (dis)similarity in our analysis.

5.2. Exploratory similarity analysis

Table 1 presents a qualitative similarity analysis for a few selected phonetic categories under both models we analyze in this study. Concretely, we retrieve five phonetic categories that exhibit the lowest JSD scores (and by implication the highest similarity) for each of the categories in the set /w, ɛ, ʃ, g/. We then provide a ranking in the table from the most similar to the least. In the case of the approximant or semivowel /w/, we observe that the approximant sound /l/ exhibits the highest similarity under both models, but four vowels appear in ranks $2-5$. This indicates a high similarity in phonetic distributions between the approximant /w/ and vowels, which we further study in the clustering analysis below. For the front vowel /ɛ/, the top-5 similar categories are all vowels under both models, although no strong preference for other front vowels can be observed since similar vowels are a mixture of front and central vowels. The two models exhibit the highest agreement in the case of the unvoiced post-alveolar fricative /ʃ/, since both models have identical ranks that include the voiced post-alveolar fricative /ʒ/ and the affricates /tʃ, dʒ/ among the most similar. For the voiced velar plosive /g/, the unvoiced velar plosive /k/ is the most similar, as expected.

5.3. Hierarchical clustering

To study the similarity patterns among the phonetic categories, we apply agglomerative hierarchical clustering with the Ward algorithm [24] over the distance matrix generated by category-wise JSD values. The result of this clustering is illustrated in Fig. 2, where each phonetic category is colored by the manner of articulation. We observe that the clustering analysis yields a similar high-level grouping between W2V2 and XLSR, except for the placement of nasals which differs across the two models. For W2V2 in Fig. 2A, the highest level of organization divides the phonetic categories into two groups: a group that represents obstruent sounds (i.e., plosives, fricatives, and affricates) as well as nasals, and another group that represents vowels and approximants. On the other hand, the highest level of organization in

Table 1: *Top-5 most similar phonetic categories to each of the categories /w, ɛ, ʃ, g/ in both* W2V2 *(*W*) and* XLSR *(*X*).*

	/w/		/ɛ/		/ʃ/		/g/	
	W	X	W	X	W	X	W	X
1	/l/	/l/	/æ/	/æ/	/tʃ/	/tʃ/	/k/	/k/
2	/ɯ/	/u/	/ʌ/	/ɪ/	/ʒ/	/ʒ/	/b/	/b/
3	/ʊ/	/ʊ/	/ɪ/	/ʌ/	/dʒ/	/dʒ/	/d/	/d/
4	/ɔ/	/ə/	/eɪ/	/eɪ/	/s/	/s/	/p/	/p/
5	/ɔɪ/	/oʊ/	/aʊ/	/aɪ/	/z/	/z/	/ð/	/h/

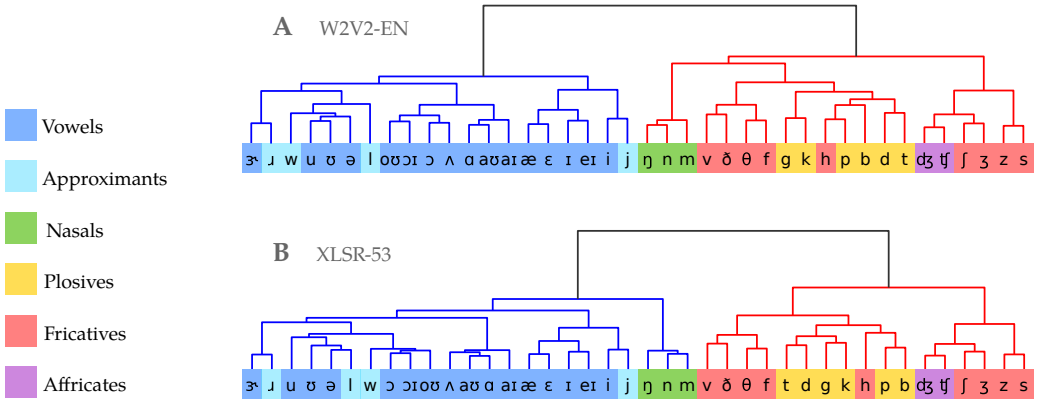

Figure 2: *The resulting clusters from applying agglomerative hierarchical clustering over the distance matrix, where our measure of the distance is the Jensen-Shannon divergence between phonetic distributions: (A)* W2V2 *and (B)* XLSR.

XLSR in Fig. 2B reveals a pure obstruent vs. sonorant division, since both approximants and nasals exhibit a higher similarity to vowels than to other consonants. The consistent grouping of approximant sounds with vowels is not surprising given their acoustic-phonetic properties. Even though approximant sounds are considered consonants from a phonological point of view, they are produced with a (relatively) unconstricted articulation and exhibit a formant structure similar to vowels [25].

Considering lower-level grouping for obstruent consonants, labio-dental and dental fricatives /f, v, θ, ð/ exhibit a higher similarity to plosive sounds /p, b, t, d, k, g/ than alveolar and postalveolar fricatives /s, z, ʃ, ʒ/ in both models. The affricates /ʤ, ʧ/ are grouped together with alveolar and postalveolar fricatives under both models, indicating the prominence of the fricative component of affricates in their underlying distributions over the discrete units. The only phonetic category that exhibits unexpected behavior in this analysis is the glottal fricative /h/, which is grouped within plosives under both models. However, the placement of the fricative /h/ among plosives should not be surprising given that the voiceless plosives /p, t, k/ are typically aspirated in syllable-initial position before a stressed vowel. Plosive aspiration is acoustically realized as a friction noise following the release of the plosive, similar to the friction of the sound /h/. Furthermore, the lowest level of grouping reflects the high similarity of phonetic minimal pairs (i.e., voicing contrasts) among all plosive contrasts (i.e., /t, d/, /p, b/, and /k, g/), but only two fricative contrasts (i.e., /s, z/ and /ʃ, ʒ/). As for the vowels, the lower-level grouping seems to reflect vowel backness more than vowel height in both models, although only a slight tendency to separate front vowels from back vowels can be observed.

5.4. Correlation with feature-based phonetic distance

To study the degree to which our measure of (dis)similarity (JSD) reflects phonetic distance, we correlate the distance among phonetic distributions over discrete units against a measure of feature-based phonetic distance. To this end, we map each phonetic category in the TIMIT inventory onto a discrete, multi-valued feature vector based on the PHOIBLE feature set [26]. We then compute the feature-based distance as the Hamming distance between their feature vectors. When we consider all phonetic categories, we find a strong positive correlation between the JSD and feature-based phonetic distance in W2V2 ($r = 0.63$) and XLSR ($r = 0.61$). Surprisingly, the correlation becomes stronger when we consider only the vowels in our analysis for both W2V2 ($r = 0.77$) and XLSR ($r = 0.80$), while it becomes weaker for consonants in W2V2 ($r = 0.47$) and XLSR ($r = 0.43$). The weaker correlation among the consonants could be attributed to the high similarity between the phonetic distributions of vowels and approximants in both W2V2 and XLSR, and vowels and nasals in XLSR. The correlation coefficients reported in this section are all Pearson's r and significant with $p \ll 0.001$.

6. Discussion and Conclusion

We presented an information-theoretic framework for characterizing the relationship between phonetic categories and discrete units in self-supervised speech models. By representing each phonetic category as a distribution over discrete units, we have shown that the distribution entropy reflects the acoustic-phonetic variability of the underlying speech sounds, with vowels being more entropic on average than consonants. Moreover, phonetically similar sounds have been found to exhibit similar distributions, with the highest level of division separating obstruents and sonorants. Our findings confirm the characterization of discrete units as sub-phonemic events, rather than high-level categories such as phonemes, which is consistent with the findings of Wells et al. [9]. Given that speech sounds are dynamic acoustic signals that vary considerably due to many factors such as context and speaker, we argue that the characterization of phonetic categories as distributions over sub-phonemic events allows for a more nuanced understanding of the relationships between phonetic categories and discrete units in self-supervised speech models. Our presented analysis has a few limitations. For example, since we do not control for the different sources of variability of speech, it is difficult to disentangle the effect of these sources on the entropy of the phonetic distributions. Future work can further tackle this limitation with a controlled analysis with respect to the speaker and context variations.

7. Acknowledgements

We thank the anonymous reviewers for their positive feedback. We extend our thanks to Marius Mosbach, Miaoran Zhang, and Vagrant Gautam for their comments on the paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

8. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[7] B. Higy, L. Gelderloos, A. Alishahi, and G. Chrupała, "Discrete representations in neural models of spoken language," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021, pp. 163–176.

[8] T. A. Nguyen, B. Sagot, and E. Dupoux, "Are discrete units necessary for spoken language modeling?" *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1415–1423, 2022.

[9] D. Wells, H. Tang, and K. Richmond, "Phonetic analysis of self-supervised representations of english speech," in *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*. ISCA, 2022, pp. 3583–3587.

[10] A. Sicherman and Y. Adi, "Analysing discrete self supervised speech representation for spoken language modeling," *arXiv preprint arXiv:2301.00591*, 2023.

[11] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[12] T. Pimentel, B. Roark, and R. Cotterell, "Phonotactic Complexity and Its Trade-offs," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 1–18, 01 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00296

[13] T. Pimentel, C. Meister, E. Salesky, S. Teufel, D. Blasi, and R. Cotterell, "A surprisal–duration trade-off across and within the world's languages," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 949–962. [Online]. Available: https://aclanthology.org/2021.emnlp-main.73

[14] N. Rathi, M. Hahn, and R. Futrell, "An information-theoretic characterization of morphological fusion," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10 115–10 120. [Online]. Available: https://aclanthology.org/2021.emnlp-main.793

[15] S. Wu, R. Cotterell, and T. O'Donnell, "Morphological irregularity correlates with frequency," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5117–5126. [Online]. Available: https://aclanthology.org/P19-1505

[16] M. Hahn, J. Degen, N. D. Goodman, D. Jurafsky, and R. Futrell, "An information-theoretic explanation of adjective ordering preferences," *Cognitive Science*, 2018.

[17] R. Futrell, K. Mahowald, and E. Gibson, "Quantifying word order freedom in dependency corpora," in *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, Aug. 2015, pp. 91–100. [Online]. Available: https://aclanthology.org/W15-2112

[18] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[19] O. Räsänen, T. Nagamine, and N. Mesgarani, "Analyzing distributional learning of phonemic categories in unsupervised deep neural networks," in *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*, vol. 2016. NIH Public Access, 2016, p. 1757.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[22] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[23] R. Hagiwara, "Dialect variation and formant frequency: The american english vowels revisited," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 655–658, 1997.

[24] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[25] L. J. Raphael, "Acoustic cues to the perception of segmental phonemes," *The handbook of speech perception*, pp. 603–631, 2021.

[26] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/