THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Phonetic Analysis of Self-supervised Representations of English Speech

Citation for published version:
Wells, D, Tang, H & Richmond, K 2022, Phonetic Analysis of Self-supervised Representations of English Speech. in H Ko & JHL Hansen (eds), *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* vol. 2022-September, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, ISCA, pp. 3583-3587, 23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022, Incheon, Korea, Republic of, 18/09/22. https://doi.org/10.21437/Interspeech.2022-10884

Digital Object Identifier (DOI):
10.21437/Interspeech.2022-10884

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 19. Feb. 2025

Interspeech 2022
18-22 September 2022, Incheon, Korea

Phonetic Analysis of Self-supervised Representations of English Speech

Dan Wells, Hao Tang, Korin Richmond

The Centre for Speech Technology Research, University of Edinburgh
{dan.wells, hao.tang, korin.richmond}@ed.ac.uk

Abstract

We present an analysis of discrete units discovered via self-supervised representation learning on English speech. We focus on units produced by a pre-trained HuBERT model due to its wide adoption in ASR, speech synthesis, and many other tasks. Whereas previous work has evaluated the quality of such quantization models in aggregate over all phones for a given language, we break our analysis down into broad phonetic classes, taking into account specific aspects of their articulation when considering their alignment to discrete units. We find that these units correspond to sub-phonetic events, and that fine dynamics such as the distinct closure and release portions of plosives tend to be represented by sequences of discrete units. Our work provides a reference for the phonetic properties of discrete units discovered by HuBERT, facilitating analyses of many speech applications based on this model.

Index Terms: speech units, self-supervised learning

1. Introduction

Self-supervised speech representation learning aims to discover representations of unlabeled speech audio which are useful for some downstream task, such as automatic speech recognition (ASR) or speech synthesis [1]. Recent approaches incorporate quantization of continuous representations to learn discrete speech units, either as part of the pre-training process before fine-tuning on a low-resource ASR task [2, 3], or with acoustic unit discovery as the end goal itself [4]. While these approaches may provide benefit where linguistic information is important [5], little work has been done to analyze how these discovered discrete units correspond to phonetic categories in speech.

The quality of discovered units is often evaluated using metrics based on frame-level alignment with phone transcripts. Purity measures indicate the degree to which discrete units are shared across multiple phone labels, which might reveal confusions between individual sounds, or the diversity of units aligned to a single phone label, possibly corresponding to context-dependent or sub-phone level representations [6]. ABX discrimination tasks move beyond individual frames, instead testing how well extracted unit sequences distinguish phonemic contrasts in the target language in triphone contexts [7]. These metrics are typically computed in aggregate across all frames in the test corpus, as in [3, 8, 1], hiding potentially significant differences between speech sounds. In [9], ABX evaluation was extended to phoneme-level measures as well as confusability between broader phonetic categories based on manner or place of articulation. While this work found differences in ABX accuracy across different phonemes, and linguistically plausible patterns of confusion between articulatory classes (e.g. affricates were most often confused with plosives and fricatives), there was no analysis of phone behavior *within* classes. Moreover, the analysis was based on continuous features rather than discrete discovered units. In [10], a spectral clustering approach struggled to discover units corresponding to phones with transitory articulations, for example plosives with their varying acoustics over closure and release phases.

In this paper, we provide more fine-grained analysis of the phonetic bases of discrete units extracted from English speech using a pre-trained HuBERT model [3, 6]. We focus primarily on a set of 50 units, approximately equal to the size of the language's phonemic inventory. By aligning phones to a discrete unit vocabulary with restricted capacity, we gain insight into the relative priority of different aspects of phone articulation in HuBERT representations. Though we limit our analysis to English, these insights are made more cross-linguistically relevant by framing them in terms of common physical characteristics of broad articulatory classes rather than individual phonemes.

2. Data preparation

We use the VCTK corpus of English speech [11], comprising 41.6 hours of read speech from 110 speakers with several distinct accents. This was recorded in a hemi-anechoic chamber at 96 kHz sample rate and 24-bit precision, giving high quality audio originally intended for speech synthesis systems. We convert all audio to 16 kHz and 16-bit precision to match the audio parameters expected by HuBERT for its input. Our choice of data is motivated by planned future work on driving speech synthesis from discrete unit inputs, as recently explored in [12, 13].

For forced alignment, we use Kaldi [14] to train speaker-adaptive HMM-GMM acoustic models for each of three English accents: Received Pronunciation (*RP*), General American (*GAm*) and Edinburgh (*Edi*). Each accent is represented by a particular reflex of the Unisyn lexicon [15], with accent-specific pronunciations derived from meta-phonemic representations designed to account for variation between accents of English. We assign each speaker in VCTK to the accent which is likely to be closest in pronunciation to their own, broadly based on rhoticity and expected vowel differences between North American and other varieties of English. The resulting accent groups each cover around 1/3 of the total number of speakers, with 43 in *RP*, 33 in *GAm* and 34 in *Edi*. While Unisyn can provide other accent-specific lexicons (including Australian and a variety from Northern England), we limited ourselves to three accents to prevent splitting the data too much and possibly negatively impacting the quality of our alignments (there are only 2 Australian speakers, for example). In total, our phone inventory comprises 70 phones used across the three accent groups. Of these, 44 are common to all accents, although their specific realizations may differ even when symbols are shared.

3. Learned discrete units

HuBERT learns speech representations by predicting sequences of discrete units for masked regions of input speech. The model is trained iteratively with initial target units derived through k-means clustering of MFCC features, then improved by clus-

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10884

Table 1: *Purity measures across different k-means sizes.*

Units	50	100	200	500
Phone purity	0.58	0.63	0.67	0.69
Unit purity	0.31	0.22	0.17	0.10

Table 2: *Examples of (left) high unit purity for the phone /f/ and (right) low phone purity for unit 32 in the 50-unit model.*

Phone	Unit	Purity
f	30	0.71
	11	0.06
	47	0.04
	17	0.04

Unit	Phone	Purity
32	h	0.39
	k	0.20
	p	0.17
	t	0.10

tering hidden representations learned during the first round of training. We extract hidden representations from the sixth layer of a HuBERT BASE model pre-trained on the full 960-hour LibriSpeech train partition [16]. While the HuBERT waveform encoder has a receptive field of 25 ms [2], output features are contextualized representations over the entire audio sequence after passing through a series of BERT transformer blocks [17]. Features are generated at a 20 ms framerate for 16 kHz audio. We then discretize the continuous features using k-means models with 50, 100, 200 & 500 clusters derived from similar HuBERT BASE representations extracted from the *train-clean-100h* partition of LibriSpeech, made available by [3, 8].[1]

We focus on purity metrics as defined in [6] to direct our analysis. For phone labels p, we calculate unit purity as the conditional probability of discrete units u aligned to that phone, $p(u \mid p)$. For discrete units, we likewise calculate phone purity $p(p \mid u)$. As seen in Table 1, phone purity tends to increase with the number of k-means clusters. Intuitively, increasing the size of the unit vocabulary should allow for more specific contextual variation to be accounted for, and confusability of phone labels given acoustics should be reduced. By the same principle, unit purity goes down as the number of clusters is increased. Table 2 shows examples of a phone with relatively high unit purity under the 50-unit model, where /f/ is most often represented by unit `30`, and a unit with relatively low phone purity, where unit `32` tends to represent both the glottal fricative /h/ and aspiration bursts across the three plosives /p, t, k/.

In [6], it is said that for a pair of target label sets of the same size, higher purity should always represent better label quality. However, the degree to which it makes sense always to align a given phone label to a single unit or vice versa depends on the nature of the phone labels used. Often, forced alignment is done with phonemic targets, and most pronunciation lexicons (including Unisyn) make use of a single phone label in all contexts. Nonetheless, HuBERT representations and k-means models may still be sensitive to non-phonemic acoustic variation, as between aspirated [k$^{\text{h}}$] and unaspirated [k] in the words *'key'* and *'ski'*. In that case, decreased unit purity could actually indicate increased sensitivity to such variation, for example if one unit represents the closure portion of a plosive such as /k/ and two others its aspirated and unaspirated releases.

4. Phonetic analysis of unit alignments

Given a set of audio frames aligned to both discrete units and phone labels, we begin by identifying the most frequent units aligned to each phone as the most likely to be distinctive for that phone. We can then check the most frequent phone labels aligned to these units, which allows us to determine whether any particular unit is in fact distinctive for a given phone, or if it is aligned with similar frequency across multiple phones.

To better structure our analysis, we divide the phones in our phone set into seven articulatory classes: plosives, fricatives, affricates, nasals, approximants, monophthongs and diphthongs. By comparing frequent unit alignments to individual frames of phones within broad articulatory classes, we expect to find cases where apparently low purity values can be attributed to similarities in the articulation of particular sets of phones. For example, plosives are produced with periods of closure regardless of place of articulation, so these silent regions should be similar across multiple phones within this class. If a particular unit is used to represent plosive closures across multiple phones, this would indicate that discovered units are at the sub-phone level. We expect these patterns to vary with the size of the unit inventory, with more context-specific units in models with higher capacities, and more shared units for smaller inventories.

Table 3: *Most frequent units aligned with plosives in 50/100-unit models. Units representing closure portions are marked in **blue**, release bursts in **red** and aspiration in **green**.*

p	b	t	d	k	g
37 / 47	37 / 47	49 / 31	7 / 31	44 / 89	49 / 89
44 / 27	7 / 66	44 / 2	49 / 1	49 / 74	7 / 27
32 / 74	44 / 27	47 / 74	9 / 85	32 / 27	37 / 78
7 / 33	20 / 16	5 / 27	26 / 27	47 / 78	44 / 66

4.1. Plosives

English distinguishes plosives at three places of articulation: labial /p, b/, alveolar /t, d/ and velar /k, g/. Each of these pairs are further distinguished by voice onset time after release of the closure portion of their articulation: /p, t, k/ are generally aspirated, with a burst of air and some frication noise after release (except when following /s/), while /b, d, g/ are unaspirated, with voicing beginning more quickly after release.

Table 3 shows the four most frequent unit IDs aligned to plosives for 50- and 100-cluster k-means models. With 50 units, many are shared across multiple plosives regardless of place of articulation. We find that units `7` and `44` are associated with closure portions, covering the transition into short silences produced by interruption of airflow through the vocal tract. On the other hand, units `37` and `49` are associated with release bursts, and unit `32` with aspiration noise following release of /p, t, k/ (it is the 7th most frequent unit for /t/). In terms of unit purity, plosives measure low compared to other phones, all ranking in the bottom 20% except for /b/. This reflects their acoustically dynamic nature, with each distinct part of their articulations best being modeled by different units. For the 100-unit model, there are fewer units shared across all plosives, and four which are used for only a single place of articulation: `47` for labial /p, b/, `31` for alveolar /t, d/, `78` and `89` for velar /k, g/. These units are all associated with closure release bursts, likely capturing characteristic formant transitions into following vowels which are key features for discriminating plosives. Note that even with these more specialized units, the usual voiced/unvoiced distinction is not evident in these alignments. Instead, it is captured by presence or absence of unit `33`, representing aspiration noise for /p, t, k/, in the sequence aligned to each phone. Unit `27` represents closure portions, and is shared across all plosives as in the 50-unit model. Figure 1 illustrates many of these features for the plosives /k, d, b, t/ in the phrase *'could use a boost'*.

[1] `https://github.com/pytorch/fairseq/tree/main/examples/textless_nlp/gslm/speech2unit`

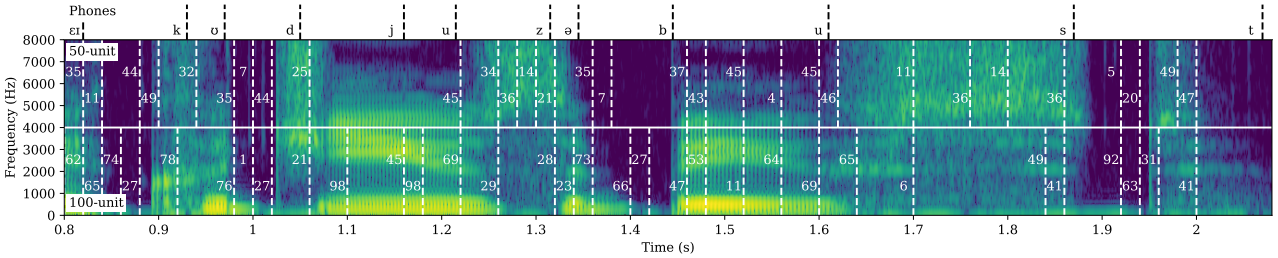

Figure 1: *Spectrogram for the phrase 'could use a boost' from VCTK utterance p236_132, labeled with phones and units from the 50- and 100-unit models. Sequences of frames with identical unit labels are shown as variable-length spans. Dynamic articulations are represented by sequences of discrete units, for example under the 50-unit model the initial plosive /k/ has distinct closure (unit 44), release burst (49) and aspiration (32) portions.*

Table 4: *Most frequent units aligned with fricatives and affricates in the 50-unit model. Units suggestive of place of articulation for fricatives are marked in **green**, as are units shared with affricates. Units shared with plosive closures and release bursts are marked in **blue** and **red**.*

f	v	θ	ð	s	z	ʃ	ʒ	h	tʃ	dʒ
30	**30**	30	**48**	**14**	**36**	**25**	**25**	**32**	**25**	**25**
11	3	**48**	18	11	34	47	34	18	**44**	**7**
47	35	11	33	**36**	**14**	11	47	15	**47**	**47**
17	7	44	20	21	9	18	11	33	**49**	**49**

We find that the units most frequently aligned to plosives are generally much less pure (all in the bottom 50%) than units aligned to less dynamic phone classes. In the 50-unit model, the majority of the top 10 most frequent phone labels for the units aligned to closure and release portions (7, 37, 44 and 49) are either plosives or affricates, which also begin with closure articulations. For unit 32, which we found to be associated with aspiration, the most frequent phone label is /h/, i.e. the fricative which is produced in an identical manner to aspiration noise. In the 100-unit model, on the other hand, the units which best distinguish plosives by place of articulation (31, 47, 78 and 89) all have their associated plosive pairs as the top two most frequently aligned phones, with the remainder of the top 10 typically representing an assortment of non-plosive phones in small proportions, likely reflecting simple boundary errors in our forced alignments. This effect increases for the 200-unit model, where /p/ and /k/ each have high-purity units to themselves. We see similar behavior across different phone classes as the number of units increases, so in general we limit the rest of the discussion below to observations from 50-unit models.

4.2. Fricatives

Most varieties of English distinguish 9 fricative consonants, with voiced/unvoiced pairs at most places of articulation: labiodental /f, v/, dental /θ, ð/, alveolar /s, z/, palato-alveolar /ʃ, ʒ/ and glottal /h/. Compared to plosives, fricatives maintain a consistent acoustic quality throughout their duration, and so these phones tend to measure much higher on unit purity; under the 50-unit model, /f, v, ʃ, ʒ/ are among the top 10 phones by unit purity, and all other fricatives are within the top 50%.

Table 4 lists the most frequent unit alignments for fricative phones in the 50-unit model. Unit 11 appears for all unvoiced oral fricatives /f, θ, s, ʃ/. This unit is generally aligned with the onset of these phones, and appears to cover the transition from a previous voiced sound into the voiceless region of frication noise which makes up the remainder of their duration. Unlike for plosives, here there appear to be distinctive units for each place of articulation. We note that discovered units tend to capture local aspects of phone articulation and that phone boundaries generally align closely with unit boundaries, i.e. units tend not to cover considerable portions of adjacent phones. Whereas formant transitions in following or preceding vowels are important acoustic cues for plosive identity, the center of gravity of frication noise is a distinguishing feature captured entirely within these phones' own durations. This explains the relative success of discovered units in the limited 50-unit vocabulary in distinguishing fricatives compared to plosives.

Table 5: *Most frequent units aligned with nasals in the 50-unit model. Units shared across places of articulation are marked in **light blue**.*

m	m̩	n	n̩	ŋ
27	**12**	**28**	**12**	**28**
46	27	26	**28**	**46**
28	**46**	**12**	**46**	**12**
3	16	31	16	4

4.3. Affricates

The articulation of affricate consonants begins similarly to plosives with complete closure of the vocal tract, but on release moves quickly to a period of frication noise. There are two affricate phones in English /tʃ, dʒ/, with their respective plosive and fricative components represented in their symbols. The relationship between these phone classes is reflected clearly in the most frequent units aligned to each affricate phone shown in Table 4: in the 50-unit model, every one is shared with their corresponding plosive or fricative phones.

4.4. Nasals

Nasals are another class of relatively steady articulations, maintaining a consistent acoustic quality throughout. Unlike fricatives, however, the discrete units in the 50-unit model do not appear to capture place distinctions between different nasal phones. This is likely due to the use of canonical pronunciations in our forced alignments. For example, a word such as *'input'* may be produced as [ɪmpʊt], with a labial nasal matching the following /p/, rather than the alveolar /n/ in its canonical form. Additionally, words ending in *-ing* may be pronounced with a final alveolar /n/ rather than a velar /ŋ/. These processes, combined with the non-phonemic distinction of syllabic /m̩, n̩/ which *is* marked in Unisyn, account for the shared unit alignments (12, 28, 46) seen across all nasals in Table 5.

Table 6: *Most frequent units aligned with monophthongs in the 50-unit model. Recurring units suggesting reduced articulations are marked in orange, and those shared with diphthongs in purple and cyan.*

i	ɪ	ɛ	ə	a	ɑ	ɒ	ʌ	ɔ	u	ʊ
45	35	0	35	23	8	8	0	29	45	35
4	6	43	47	0	3	3	3	8	47	47
43	34	31	16	6	0	0	8	0	16	7
9	16	23	31	3	23	29	35	42	35	0

Table 7: *Most frequent units aligned with diphthongs in the 50-unit model. Units shared with initial monophthongs are marked in purple and with final monophthongs in cyan; asterisks indicate units shared with similar but not identical monophthongs.*

ɛɪ	aɪ	ɔɪ	əʊ	aʊ
43	40	29	29	23
*4	*43	*4	0	0
16	*8	38	46	3
*9	*4	16	16	8

4.5. Monophthongs

Monophthongs are vowels which maintain a singular acoustic quality throughout, apart from some coarticulation effects at boundaries with surrounding phones. Similar to nasals, purity metrics are affected by the use of canonical pronunciations in our forced alignments. For example, unit `35` in the 50-unit model is the most frequently aligned for both /ɪ, ʊ/ and /ə/, where in English the former are short vowels susceptible to reduction to /ə/ in unstressed positions. Similarly, unit `0` recurs across multiple non-high, mostly back vowels /ɑ, ɒ, ʌ, ɔ, a, ɛ/, possibly indicating another reduced variant for phones in this part of the vowel space. Unit `45` is most frequently aligned to high vowels /i/ and /u/, which phonemically should differ in terms of both frontness and rounding. This apparent confusion could arise from frequent fronting of /u/ by English speakers across accents. The phone /u/ is however captured by its own unit `69` in the 100-unit model.

For the 100-unit model, most monophthongs lie in the bottom 50% of phones ranked by unit purity, suggesting a degree of context-dependence in unit alignments. One readily identifiable source of variation here is pre-nasal vowels: unit `38` largely covers /e, a, i, ə/ before /n/ (unit `31` has a similar role in the 50-unit model). However, Table 9 shows that these phones are among those with the fewest different units aligned to any particular instance, which would imply that contextual variation is covered elsewhere. This is supported by the place-distinctive units seen in the 100-unit model for plosives, which are generally aligned to release portions leading into following vowels.

4.6. Diphthongs

Diphthongs are vowels which pass through two distinct acoustic qualities throughout their articulation, presenting somewhat as a trajectory between two monophthong steady states. We represent diphthongs ending in a high front vowel using /ɪ/ in line with many descriptions of English, but comparing unit alignments across Tables 6 and 7 we see that these diphthongs in fact tend to share more units with /i/. Phonemic notation aside, we see that unit alignments for diphthongs indeed tend to be composed of units used for their constituent vowels in the 50-unit model. There is some indication of diphthong-specific units in the 100-unit model, perhaps explicitly modeling the transitions between steady states, but many units are still shared with constituent monophthongs there as well.

Table 8: *Most frequent units aligned with approximants in the 50-unit model. Units shared with similar monophthongs are marked in cyan.*

j	w	ɹ	l	ɬ	l̩
45	38	24	41	42	42
4	10	17	19	19	9
18	18	9	42	29	46
43	15	16	0	41	41

Table 9: *Average number of unique units aligned to instances of phones per phone class in 50- and 100-unit models.*

Units	Pl	Fr	Af	Na	Mo	Di	Ap
50	2.16	2.00	2.86	1.84	1.60	2.49	1.54
100	2.28	2.19	3.11	2.05	1.74	2.88	1.70

4.7. Approximants

The two approximants /j, w/ are sometimes called semivowels, and may be seen as transient articulations of the vowels /i, u/. This is reflected in the unit alignments for /j/ under the 50-unit model shown in Table 8, where most units are shared with /i/, although the same is not true for /w/, which has unit `38` as an especially distinctive unit by phone purity. The alveolar approximant /ɹ/ is characterized by sharply rising second and third formants, while lateral /l/ shows a rising second formant. For lateral approximants, Unisyn marks three allophonic variants: /l/ occurs in syllable onsets, /ɬ/ in coda position, and /l̩/ constitutes a syllabic nucleus itself, typically matching the articulation of /ɬ/. We see some evidence for the onset/coda distinction in unit alignments, with unit `41` being more indicative of onset realizations and `42` of coda or syllabic variants.

5. Conclusion

In this work, we presented a fine-grained analysis of the alignment between phonetic labels and discrete units discovered in a self-supervised manner from English speech. All representations are based on pre-trained models (HuBERT BASE and associated k-means clusters) and freely-available speech corpora (LibriSpeech and VCTK), so that our findings should be of direct relevance to other researchers using these resources.

By focusing on a relatively small unit vocabulary, close to the size of the phonemic inventory of English, we provided some insight into the relative priorities of articulatory-acoustic features of different phone classes when discretizing HuBERT representations. For example, manner of articulation appears to have greater priority than distinctive place features for plosives and nasals. The dynamic nature of plosive articulations give rise to discrete units aligned with distinct regions of closure and release, which are shared across multiple phones within this class. Low unit purity for nasals and monophthong vowels, on the other hand, may stem from particular choices in the phonetic transcription system used in our analysis, something which should be considered when using purity metrics for gross evaluation of discovered acoustic unit quality. In future work, we would like to extend this analysis to languages other than English, to see how well HuBERT representations might capture phones unseen in its monolingual training data.

Acknowledgements: This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

6. References

[1] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux, "The Zero Resource Speech Challenge 2019: TTS Without T," in *Interspeech 2019*, 2019, pp. 1088–1092.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[3] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HUBERT: How Much Can a Bad Teacher Benefit ASR Pre-Training?" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6533–6537.

[4] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge," in *Interspeech 2020*, 2020, pp. 4836–4840.

[5] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*. ISCA, 2021, pp. 1194–1198.

[6] 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.

[7] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Interspeech 2013*, 2013, pp. 1781–1785.

[8] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "Generative Spoken Language Modeling from Raw Audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[9] S. Feng and O. Scharenborg, "The Effectiveness of Unsupervised Subword Modeling With Autoregressive and Cross-Lingual Phone-Aware Networks," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 230–247, 2021.

[10] S. Feng and T. Lee, "On the Linguistic Relevance of Speech Units Learned by Unsupervised Acoustic Modeling," in *Interspeech 2017*, 2017, pp. 2068–2072.

[11] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," https://datashare.is.ed.ac.uk/handle/10283/3443, Nov. 2019.

[12] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech 2021*, 2021, pp. 3615–3619.

[13] C. Wang, W.-N. Hsu, Y. Adi, A. Polyak, A. Lee, P.-J. Chen, J. Gu, and J. Pino, "Fairseq S^2: A Scalable and Integrable Speech Synthesis Toolkit," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2021, pp. 143–152.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[15] S. Fitt, "Unisyn Lexicon," https://www.cstr.ed.ac.uk/projects/unisyn/.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 4171–4186.