Phonetic Analysis of Self-supervised Representations of English Speech

Citation for published version:

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.
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 rate 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 transcriptions. 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 semi-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-
### Table 1: Purity measures across different k-means sizes.

<table>
<thead>
<tr>
<th>Units</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone purity</td>
<td>0.58</td>
<td>0.63</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>Unit purity</td>
<td>0.31</td>
<td>0.22</td>
<td>0.17</td>
<td>0.10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Phone</th>
<th>Unit</th>
<th>Purity</th>
<th>Unit</th>
<th>Phone</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>30</td>
<td>0.71</td>
<td>32</td>
<td>h</td>
<td>0.39</td>
</tr>
<tr>
<td>11</td>
<td>0.06</td>
<td>k</td>
<td>47</td>
<td>p</td>
<td>0.17</td>
</tr>
<tr>
<td>17</td>
<td>0.04</td>
<td>t</td>
<td>11</td>
<td>0.06</td>
<td>k</td>
</tr>
</tbody>
</table>

#### 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.

<table>
<thead>
<tr>
<th>p</th>
<th>b</th>
<th>t</th>
<th>d</th>
<th>k</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>37/47</td>
<td>37/47</td>
<td>49/31</td>
<td>7/31</td>
<td>44/89</td>
<td>49/89</td>
</tr>
<tr>
<td>44/27</td>
<td>7/66</td>
<td>44/2</td>
<td>49/1</td>
<td>49/74</td>
<td>7/27</td>
</tr>
<tr>
<td>32/74</td>
<td>44/27</td>
<td>47/74</td>
<td>9/85</td>
<td>32/27</td>
<td>37/78</td>
</tr>
<tr>
<td>7/33</td>
<td>20/16</td>
<td>5/27</td>
<td>26/27</td>
<td>47/78</td>
<td>44/66</td>
</tr>
</tbody>
</table>

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.

#### 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, plosive 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”.

---

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 89) 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, s, z/. This unit is generally aligned with the onset of these phones, and appears to cover the transition from a previous voiced sound to the voiceless region of friction 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 friction 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.

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 friction noise. There are two affricate phones in English /ʃ, ʒ/, 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 [ɪnˈpʊt], 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 /n, ŋ/ 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 purple, and those shared with diphthongs in cyan.

<table>
<thead>
<tr>
<th>i</th>
<th>e</th>
<th>o</th>
<th>a</th>
<th>u</th>
<th>e</th>
<th>o</th>
<th>a</th>
<th>u</th>
<th>e</th>
<th>o</th>
<th>a</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>43</td>
<td>23</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>29</td>
<td>45</td>
<td>35</td>
<td>45</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>43</td>
<td>47</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>47</td>
<td>47</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>43</td>
<td>34</td>
<td>31</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>16</td>
<td>7</td>
<td>16</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>23</td>
<td>31</td>
<td>23</td>
<td>29</td>
<td>35</td>
<td>42</td>
<td>35</td>
<td>0</td>
<td>29</td>
<td>41</td>
<td>35</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>I</th>
<th>E</th>
<th>Ì</th>
<th>Ë</th>
<th>Ò</th>
<th>Î</th>
<th>Ô</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>40</td>
<td>29</td>
<td>29</td>
<td>23</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>*</td>
<td>43</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>*8</td>
<td>38</td>
<td>46</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>*</td>
<td>4</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 /ι, u/ and /I/,
where in English the former are short vowels susceptible to re-
duction to /I/ in unstressed positions. Similarly, unit 0 recurs
across multiple non-high, mostly back vowels 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 bot-
tom 50% of phones ranked by unit purity, suggesting a degree
of context-dependence in unit alignments. One readily identifi-
able source of variation here is pre-nasal vowels: unit 38 largely
covers /e, a, i, ò/ 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 par-
ticular 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 gener-
ally aligned to release portions leading into following vowels.

4.6. Diphthongs

Diphthongs are vowels which pass through two distinct acoustic
qualities throughout theirarticulation, presenting somewhat as
a trajectory between two monophthong steady states. We rep-
resent diphthongs ending in a high front vowel using /ι/ in line
with many descriptions of English, but comparing unit align-
ments 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 com-
pounded 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 con-
stituent monophthongs there as well.

4.7. Approximants

The two approximants /j, w/ are sometimes called semivow-
els, and may be seen as transient articulations of the vowels /i, 
/`. 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 alve-
olar approximant /l/ is characterized by sharply rising second
and third formants, while lateral /l/ shows a rising second for-
mant. For lateral approximants, Unisyn marks three allophonic
variants: /l/ occurs in syllable onsets, /l/ in coda position, and
/l/ constitutes a syllabic nucleus itself, typically matching the
articulation of /l/. We see some evidence for the onset/coda
distinction in unit alignments, with unit 41 being more indica-
tive of onset realizations and 42 of coda or syllabic variants.

5. Conclusion

In this work, we presented a fine-grained analysis of the align-
ment between phonetic labels and discrete units discovered in
a self-supervised manner from English speech. All represen-
tations are based on pre-trained models (HuBERT BASE and
associated k-means clusters) and freely-available speech cor-
pora (LibriSpeech and VCTK), so that our findings should be
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 vow-
els, 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 cap-
ture phones unseen in its monolingual training data.

Acknowledgements: This work was supported in part by the
UKRI Centre for Doctoral Training in Natural Language Pro-
cessing, 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


