Improving access to untranscribed speech by leveraging spoken term detection and self-supervised learning of speech representations

Nay San
Stanford University
nay.san@stanford.edu

Martijn Bartelds
University of Groningen
m.bartelds@rug.nl

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

We summarise findings from our recent work showing that a large self-supervised model trained only on English speech provides a noise-robust and speaker-invariant feature extraction method that can be used for a speech information retrieval task with unrelated low resource target languages. A qualitative error analysis also revealed that the majority of the retrieval errors could be attributed to the differences in phonological inventories between English and the evaluation languages. With a longer-term aim of leveraging typological information to better adapt such models for the target languages, we also report on work in progress which examines the phonetic information encoded in these representations.

Introduction. Language documentation efforts often yield a sizeable amount of untranscribed speech, which is difficult to index and search. These difficulties have a direct impact on how easily such resources may be used for language maintenance and revitalisation activities by many interested parties. One way to alleviate such difficulties is through query-by-example spoken term detection (QbE-STD), which, as shown in Figure 1, is a long-standing speech information retrieval task of finding all regions within a corpus of audio documents where a spoken query term occurs (Myers et al., 1980; Rohlicek, 1995; Fiscus et al., 2007; Rodriguez-Fuentes et al., 2014; Ram et al., 2020).

As QbE-STD involves directly comparing speech samples, retrieval performance can be poor when the query and corpus are spoken by different speakers and produced in different recording conditions. Using data selected from a variety of speakers and recording conditions from 7 Australian Aboriginal languages and Gronings (a regional variety of Dutch), all of which are endangered or vulnerable, we evaluated whether QbE-STD performance on these languages could be improved by leveraging features extracted from the pre-trained English wav2vec 2.0 model (Baevski et al., 2020, henceforth w2v2). The w2v2 model consists of an encoder network that reads in raw audio and outputs latent speech representations that are then fed to a 24-layer Transformer network to build contextualised representations. The self-supervised training objective was to correctly predict randomly masked portions of the speech representations.

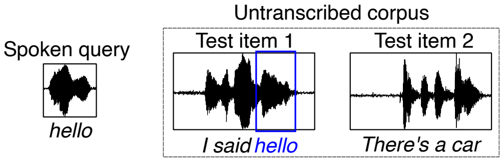

Figure 1: Query-by-example spoken term detection

QbE-STD with wav2vec 2.0 features. Compared to the use of Mel-frequency cepstral coefficients and bottleneck features, we find that representations from the middle layers of the w2v2 Transformer network offer large gains in task performance. In the worst of cases, retrieval performance improved 56–86% from only 27–28% of queries being retrievable to 42–52% when using w2v2 features (Transformer layer 11). This is a tolerable operating range, given the alternative is browsing untranscribed audio in near real-time.

Phonological similarity to the training language (English) resulted in higher QbE-STD performance, with performance on the regional variety of Dutch (Gronings) being higher than on the Australian languages with comparable dataset characteristics. A qualitative error analysis revealed that phonological differences between English and the Australian languages account for a substantial portion of the retrieval errors. For example, for the Kaytetye query [aɳanpə] 'medicinal sap', an erroneous retrieval was the word [anənkə] 'to sit', both of which share the VNVNTV template (where V represents a vowel, N a nasal, and T a plosive). In other words, the representations of the w2v2 English model do not appear to be sufficiently fine-grained to dif-

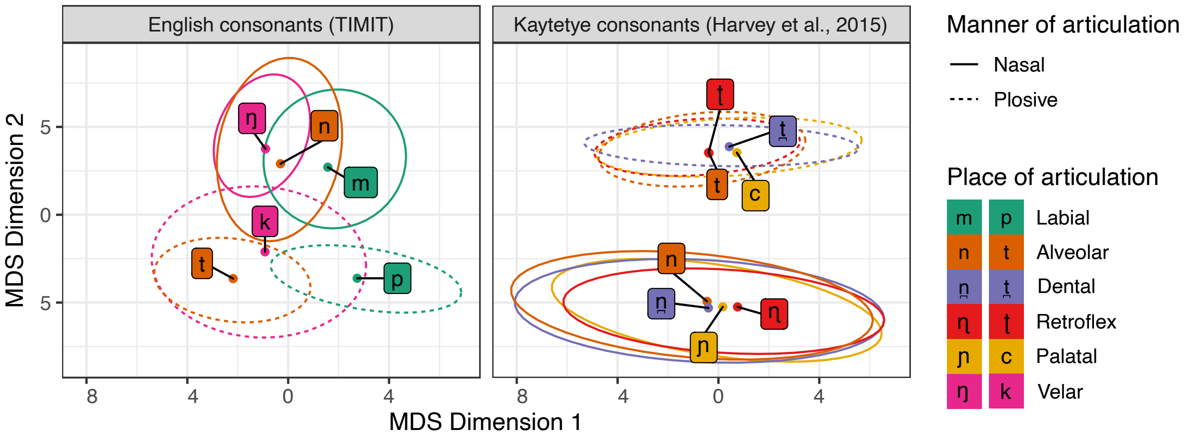

Figure 2: Multidimensional scaling (MDS) visualisations of features extracted using Transformer layer 11 of the wav2vec 2.0 model from English (TIMIT corpus) and Kaytetye consonants (Harvey et al., 2015), averaged over the duration of the consonant. Ellipses represent 95% confidence intervals and text labels represent means.

ferentiate between segments differing primarily in place of articulation, especially for those that do not occur contrastively in English (e.g. retroflex nasal [ɳ] vs. alveolar nasal [n]).

Exploring the wav2vec 2.0 features. To further examine the phonetic information encoded by the w2v2 English model, we used the Transformer layer 11 (the best performing layer for our QbE-STD evaluations) to extract features of English consonants (from the TIMIT corpus) and Kaytetye consonants (from audio stimuli collected by Harvey et al., 2015). Figure 2 shows the dissimilarities between each of the consonant types for English and Kaytetye in the 1024-dimensional space of the w2v2 English model (averaged across the duration of the consonant and reduced to 2 dimensions through multidimensional scaling for each language). For considerations of space and visual clarity, we focus on the nasals and plosives.

Despite being given no phone labels at training time, the w2v2 model learns to encode broad manner of articulation classes within its representations, as evidenced by ellipses forming two distinct clusters for nasals (solid lines) and plosives (dashed lines) in Figure 2. In line with findings from our qualitative error analysis, features from the w2v2 English model under-differentiate place of articulation contrasts, as evidenced by the overlapping ellipses within each of the nasal and plosive manners of articulation in Figure 2. Interestingly, this under-differentiation is present for both Kaytetye and English, the latter for which the distributions of the velar [ŋ, k] and alveolar consonants [n, t] are entirely overlapping within each manner class, and the labial consonants [m, p] also partially overlapping with the other two places.

As Australian languages typically have 4-6 place of articulation contrasts with relatively few contrasts based on manner of articulation, their phonological inventories have been described as ‘long and thin’ (Butcher, 2012).[1] Erroneous retrievals that occur with the Australian languages can be as explained as the result of an interaction between these languages’ phonological inventories (place-rich/manner-poor) and the representations of the evaluated w2v2 model (place-poor/manner-rich), learned from unlabelled English speech. In our future work we will investigate whether information relating to place of articulation is learned through additional training of the English w2v2 model on unlabelled speech in the target languages and how this affects QbE-STD performance.

For these low resource languages, features extracted using the w2v2 English model trained on 960 hours of speech offer noise-robust and speaker-invariant speech representations that are highly effective for speech information retrieval. A preliminary investigation revealed that phonetic information is selectively encoded in these representations: predominantly of manner of articulation, based on English speech. As such, leveraging typological information about how phonetic categories and phonological contrasts are distributed within various languages may be the key to training easily adaptable self-supervised models for use with various low resource target languages.

[1] In relation to the horizontal and vertical dimensions of the standard IPA consonantal chart.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Andy Butcher. 2012. On the phonetics of long, thin phonologies. *Quantitative approaches to problems in linguistics, edited by C. Donohue, S. Ishihara, and W. Steed (LINCOM, Munich, Germany)*, pages 133–154.

Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo, and George Doddingtion. 2007. Results of the 2006 spoken term detection evaluation. In *Proceedings of SIGIR 2007*, volume 7, pages 51–57.

Mark Harvey, Susan Lin, Myfany Turpin, Ben Davies, and Katherine Demuth. 2015. Contrastive and non-contrastive pre-stopping in Kaytetye. *Australian Journal of Linguistics*, 35(3):232–250.

Cory Myers, Lawrence Rabiner, and Andrew Rosenberg. 1980. An investigation of the use of dynamic time warping for word spotting and connected speech recognition. In *Proceedings of ICASSP 1980*, pages 173–177.

Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard. 2020. Neural Network based End-to-End Query by Example Spoken Term Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1416–1427.

Luis J. Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel, and Mireia Diez. 2014. High-performance query-by-example spoken term detection on the SWS 2013 evaluation. In *Proceedings of ICASSP 2014*, pages 7819–7823.

Jan Robin Rohlicek. 1995. Word spotting. In *Modern Methods of Speech Processing*, volume 327 of *The Springer International Series in Engineering and Computer Science*, pages 123–157. Springer.