Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features

Patrick Cormac English[1], **John D. Kelleher**[2], **Julie Carson-Berndsen**[1]
SFI Centre for Research Training in Digitally-Enhanced Reality (d-real),
[1]ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland
[2]ADAPT Research Centre, Technological University Dublin, Ireland

Abstract

In recent years large transformer model architectures have become available which provide a novel means of generating high-quality vector representations of speech audio. These transformers make use of an attention mechanism to generate representations enhanced with contextual and positional information from the input sequence. Previous works have explored the capabilities of these models with regard to performance in tasks such as speech recognition and speaker verification, but there has not been a significant inquiry as to the manner in which the contextual information provided by the transformer architecture impacts the representation of phonetic information within these models. In this paper, we report the results of a number of probing experiments on the representations generated by the wav2vec 2.0 model's transformer component, with regard to the encoding of phonetic categorization information within the generated embeddings. We find that the contextual information generated by the transformer's operation results in enhanced capture of phonetic detail by the model, and allows for distinctions to emerge in acoustic data that are otherwise difficult to separate.

1 Introduction

In recent years large transformer models have become available which provide a novel means of generating high-quality vector representations of input speech audio sequences. These transformers aim to exploit feature learning on large unlabelled datasets to perform sequence-to-sequence transformations on audio that capture and preserve salient features from the input sequence in a quantised and contextual output representation. While most work on transformer models in automatic speech recognition focus on performance improvements and applications in down-stream tasks, this paper focuses on whether the internal layers of a transformer model provide any information as to the emergence of phonetic and phonological properties of speech. Specifically we interrogate the wav2vec 2.0 model (Baevski et al., 2020) by probing the internal layers of the transformer using domain-informed features. The structure of this paper is as follows. Firstly, in section 2 we present some existing research related to our approach followed by a discussion of transformer-based models in section 3. Section 4 presents the resources used, and the experimental methodology is described in section 5. In section 6 we present our results, followed by conclusions and future work in section 7.

2 Related Work

There has been considerable work in recent years as to the extent and nature of phonetic information captured in the embeddings used by deep learning models. The word2vec model (Mikolov et al., 2013) has been applied below the level of the word to investigate phonological analogies and similarities. Silfverberg et al. (2018) have explored the sound analogies generated by phoneme embeddings. Kolachina and Magyar (2019) detailed the ability of embeddings to capture phonemic and allophonic relationships within an artificial language, noting that contrastive elements within the embedding space correlated with articulatory features. O'Neill and Carson-Berndsen (2019) demonstrate that embeddings derived purely from text using a grapheme-to-phoneme mapping and applying a word2vec approach exhibit similarity between phoneme classes. These phoneme embeddings were subsequently integrated with the data-driven acoustic similarities of Kane and Carson-Berndsen (2016) to generate a similarity matrix for use in phonemically driven spell checking (O'Neill et al., 2021).

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 83 - 91
July 14, 2022 ©2022 Association for Computational Linguistics

Specifically with respect to the capture of phonetic information in the embeddings of automatic speech recognition, Belinkov and Glass (2017) have investigated the internal layers of end-to-end recognition systems using a connectionist temporal classification (CTC) approach with DeepSpeech2 (Amodei et al., 2016). They found significant differences across layers in their architecture with respect to predictive performance of phoneme categories. Their work also demonstrated that certain categories became represented in the embedding space of their chosen model such that intra-category separation was significantly more difficult than for other categories. They noted that these categories saw better performance in later layers, at the expense of degraded performance in more easily separable categories. Scharenborg et al. (2019) have investigated the representation of speech in deep neural networks using a 3-layer model trained to distinguish consonants and vowels. They performed a wide-ranging comparison of PCA-transformed embedding spaces, and their work saw strong clustering on the basis of the vowel/consonant categorisation and manner of articulation. Most recently, Ma et al. (2021) investigated the extent to which phonetic properties emerge from the acoustic representations of transformer-based speech recognition architectures. Using four pre-trained acoustic representations from transformer-based speech recognition architectures, they designed probing tasks using linear regression, a support vector machine and a feedforward neural network consisting of two fully-connected layers. Their embeddings are associated with high-level categorisations derived from the TIMIT dataset (Garofolo et al., 1992), perform at a high level and see significant improvements across layers when considering less-separable classes such as fricatives.

Conneau et al. (2018) proposed a methodology known as probing as a way to examine what information is present in an embedding. In Conneau et al.'s framing a probing task involves training a classification model to predict properties (e.g., length, tense, parse tree depth, and so on) of a sentence based on the embedding of the sentence. Probing assumes that the accuracy of the classification model (i.e., a probe) on the task indicates whether the embeddings encode information relevant to task target. There is a growing body of work using probing to examine what types of information are encoded in the embeddings created by Transformer models (Hewitt and Manning, 2019; Liu et al., 2019; Tenney et al., 2019; Nedumpozhimana and Kelleher, 2021), and also exploring what layer in the Transformer architecture different types of information are encoded in (Jawahar et al., 2019). In this work, we adapt the probing methodology to speech embeddings, and use it to understand and compare the phonetic information encoded in different layers of a Transformer model. Through this comparision of probing performance across layers on phonetic tasks we hope to better understand whether the information encoded in these speech embeddings, and the sequencing of this encoding across layers, accords with domain-knowledge expectations regarding phonetics.

The work presented in this paper focuses specifically on the transformer module of the wav2vec2.0 model (Baevski et al., 2020) and the representations generated at each layer of the transformer. It will not probe the attention mechanism itself, which is outside the scope of this paper. The primary goal of this investigation is not to deliver an explanation of the operations undertaken by the transformer architecture in generating these representations, but instead to probe the representations generated at different layers across the architecture in order to examine the development of the architecture's ability to delineate between phonetic categories.

3 Transformer-Based Models

In recent years transformer-based models have reported state-of-the-art results on a range of speech processing tasks, and today pre-trained models are available for a variety of high-demand tasks such as automatic speech recognition (ASR). These models leverage the availability of large unlabelled acoustic datasets, in parallel with enhanced architectural features such as attention mechanisms, to produce information-dense distributed vector representations (embeddings) of input audio signals. In the architecture examined herein, embeddings are of a N*T dimensionality, with width N dependent upon input length, and each instance of T representing the dimensionality of the encoded information within a specific time-frame, and specific variance within that dimensionality relating to differences in the acoustic feature space for that frame.

The excellent performance of transformer based models on speech processing tasks suggests that these models have the ability to encode within the embeddings they generate aspects of the input signal relating to speech phenomena, while discarding low-information aspects of the input signal such as background noise and variation deemed to be unimportant during the training cycle. Furthermore, some architectures such as wav2vec 2.0 have been designed to exploit the high-quality of embeddings generated from unlabelled data by allowing for very small quantities of labelled data to be provided as fine-tuning information during a separate training stage while still achieving high levels of transcription performance.

However, while there has been significant inquiry as to the final-level performance of these models, relatively little is known as to the specific information captured within the embedding space, and whether that encoded information accords with domain-knowledge expectations. Previous works have explored the use of these embeddings as the basis for higher-order operations, such as accent-resilient ASR (Li et al., 2021), identification of speaker emotional state (Pepino et al., 2021), and modelling of prosody in speaker input (Gan et al., 2022).

For the probing task detailed in section 5, the phoneme embeddings (calculated by averaging the embeddings for the frames within the phoneme interval) for each layer in the multi-layer wav2vec2.0 transformer stack are used as inputs for the training of a multi-layer perceptron (MLP) on the task of identifying an associated TIMIT phonetic label. The performance of this model is taken as indicative of the relative richness of specific phonetic data within the output embeddings from wav2vec 2.0.

4 Resources

4.1 TIMIT

The TIMIT read-speech corpus (Garofolo et al., 1992) was used due to the high-quality metadata present in the dataset. The dataset is comprised of 5.4 hours of spoken English audio sampled at 16kHz in *wav* format. The audio is American-accented, with 8 major US English dialects represented, with each speaker recorded uttering ten high acoustic-information sentences. Each utterance is a single sentence of spoken audio, with manual character, phonetic, and orthographic transcriptions, in time-aligned format, provided for each recording.

4.2 wav2vec 2.0

This work uses wav2vec 2.0 (Baevski et al., 2020). This section outlines the pre-training task, training task, and architecture of the pre-trained wav2vec 2.0 model “base_960” [1] used at the pre-experimental stage. It then proceeds to the application of the model to the production of the ASR data used in the primary task.

4.2.1 Architecture

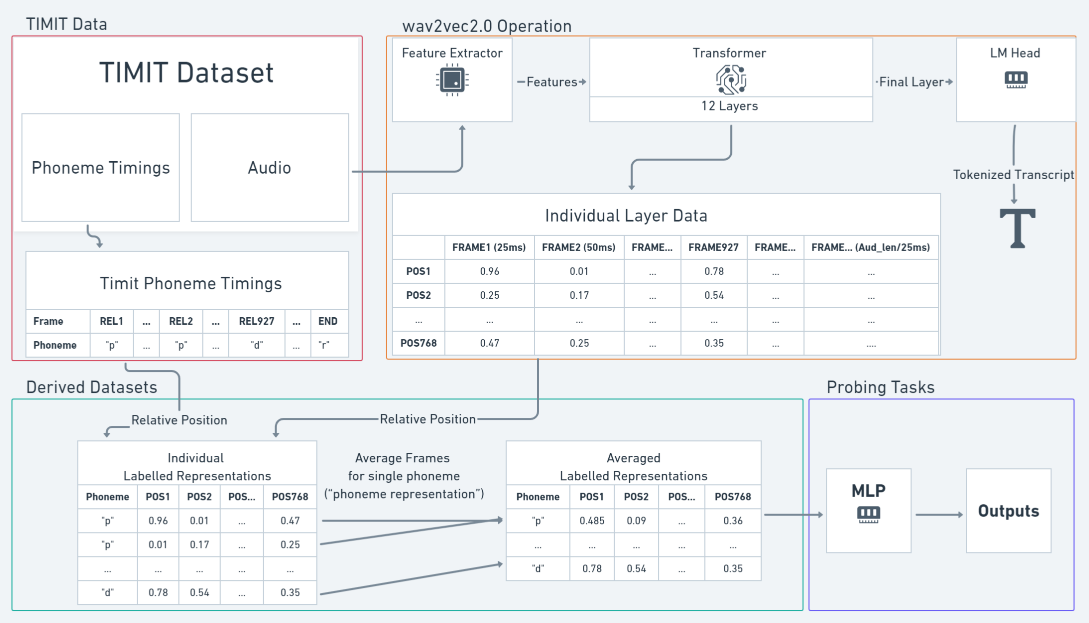
wav2vec 2.0 makes use of a transformer architecture for the purposes of transforming raw audio input W into a vector context representation C. A 1D ConvNet feature encoder first parses the waveform into a latent speech representation which is passed to the transformer. The transformer component is composed of a stack of 12 transformer layers each with an internal dimension of 768, a feed-forward dimension of 3072, and 8 attention heads. The component takes the output of the feature encoder, applies relative positional encoding and a GELU activation to the inputs, before a layer normalisation. This outputs context representation C.

The "base_960" model used can be loaded in a headless or LM-head configuration, the latter of which includes a language modelling head applied on top of the transformer architecture which divides output into a vocabulary of 32 characters including alphabetical characters and separators. This outputs character representations of C, which is the ASR transcription of W.

4.2.2 Training Task and Dataset

The wav2vec 2.0 model was pre-trained on the unlabelled Librispeech corpus containing 960 hours of audio. The wav2vec 2.0 model features both a pre-training and fine-tuning objective. The fine-tuning task is not relevant for this work, as it pertains to the language-modelling head which was not used in our experiments. The pre-training task requires the transformer module to correctly identify the “true” latent quantised speech representation, provided by the pre-transformer quantisation CNN module, for a masked time-step. A certain proportion of the inputs (representing quantisations of a particular time-step) to the transformer module are masked, and the transformer must identify them from a set of distractors sampled from the overall set of masked time-steps.

[1] https://huggingface.co/facebook/wav2vec2-base-960h



Timit Phoneme Timings

Frame	REL1	...	REL2	...	REL927	...	END
Phoneme	"p"	...	"p"	...	"d"	...	"r"

Individual Layer Data

	FRAME1 (25ms)	FRAME2 (50ms)	FRAME...	FRAME927	FRAME...	FRAME... (Aud_len/25ms)
POS1	0.96	0.01	...	0.78	...	...
POS2	0.25	0.17	...	0.54	...	...
...	...	...	...	...	...	...
POS768	0.47	0.25	...	0.35	...	....

Individual Labelled Representations

Phoneme	POS1	POS2	POS...	POS768
"p"	0.96	0.01	...	0.47
"p"	0.01	0.17	...	0.25
...	...	...	...	...
"d"	0.78	0.54	...	0.35

Averaged Labelled Representations

Phoneme	POS1	POS2	POS...	POS768
"p"	0.485	0.09	...	0.36
...	...	...	...	...
"d"	0.78	0.54	...	0.35

Figure 1: Overview of Experimental Methodology

5 Methodology

This section sets out the experimental methodology employed in this paper, outlining specifically how the relevant data was generated and the description of the probing task. Figure 1 provides an overview of the steps involved.

5.1 Data Generation

Firstly, utterance embeddings are generated using wav2vec 2.0. For each utterance in the full TIMIT training dataset (4620 separate 16kHz wav-formatted files), 12 sets of embeddings were generated, one per transformer layer. This was performed by operating the model without its language modelling head, and specifying the return of hidden-layer representations, where each transformer block is a single hidden-layer. Each audio file input generates an output of format [N*768] (N being the number of 25ms frames, proportional to the duration of the input audio); this results in the *Individual Layer Data* in figure 1. In contrast to the representations explored by Belinkov and Glass (2017), the representations here retain a constant dimensionality throughout each layer of the transformer, in distinction to the variety of layer architectures employed in DeepSpeech2.

The next step is to generate a frame-based dataset for the probing tasks. Since the TIMIT dataset provides frame-aligned annotations, marking the beginning and end of a given phoneme in the associated audio file, this data can be used to calculate phoneme-averaged durations. Taking the proportion between the maximum number of TIMIT frames in a given audio sequence and the number of wav2vec 2.0 frames N generated for that sequence, a relative positional mapping is generated for each [N*768] embedding, whereby a given frame of shape [1*768] is labelled with the phoneme[2] occurring at that position in the audio sequence, as according to the TIMIT labels. In this way a vector of shape [1*767] is generated, containing the vector representation of a given wav2vec 2.0 frame and the TIMIT-derived phoneme annotation. This process is depicted in figure 1 under *Derived Datasets*. 12 of these frame datasets were generated from the TIMIT dataset, to be used in the next section as the basis for deriving the phoneme-averaged representations used in the probing task.

Employing a variant of the method used in (Shah et al., 2021), the vector values of individual frames occurring during a specific phoneme interval are averaged, to create a representation in the embedding space of a given instantiation of a phoneme. This generated a dataset of 175,232 individual phoneme representations in the format [1*767], where the first field contains the phoneme label and the remaining 768 fields contain the column-wise average of all frames generated during a given phoneme

[2] We use the term phoneme here for labels that align with the English phoneme set. TIMIT also separates out the stop closures e.g. with the label "bcl". We retain these labels.

occurrence in the input audio. Figure 1 depicts this process for a simplified two-frame phoneme example. Twelve such datasets were derived, one per chosen layer. These datasets are then used as inputs to the probing task.

5.2 Probing Task

For the probing task, 12 multi-layer perceptron models were trained to predict TIMIT phoneme labels from the phoneme-averaged wav2vec 2.0 embeddings. A scikit-learn (Pedregosa et al., 2011) implementation of the multi-layer perceptron (MLP) was used, comprised of a single hidden layer of 200 neurons with ReLu activation, and an output layer of a single neuron with a logistic activation function. The models used the default hyper-parameters implemented in scikit-learn, with the exception of the hidden layer size which was expanded to 200 neurons.

To train the model, each multi-layer perceptron was provided with the phoneme-averaged dataset for a given layer as training material, with 43,808 samples reserved for testing. During training, the averaged vector representations of shape [1*768] were the input data with the [1*1] TIMIT phoneme label as the target category. The division of each layer's embeddings was static, with each model provided with its respective layer's wav2vec 2.0 outputs for the same audio files.

To generate the outputs described in section 6, the model was provided with the reserved rows, containing only the [1*768] vector information. The [1*1] phoneme label was removed and stored separately as the ground truth for each vector representation. The model then generated a predicted phoneme label per vector representation, which was stored with the ground truth in a collection of [1*2] ground-truth/predicted-label pairs.

Following best practice (Belinkov, 2021), we created a separate sub-experiment to assess the potential effects of chance correlation on our results. The primary probing task was re-conducted with an artificial dataset of the same dimensions as the phoneme-averaged dataset. This new dataset was comprised of values randomly sampled from the range of each feature column in the phoneme-averaged dataset, with the labels left unchanged. The performance of the probe on this task was very low (<2% accuracy per phone across layers). This result is substantially lower than the performance observed with the real embedding data, and we took this difference to indicate that the performance of our primary probing results reflect actual information relevant to the task, rather than chance correlation. Future work will seek to investigate the dataset in more detail, and incorporate any findings into a more robust probing task.

From the primary probing task, the following outputs were generated for each layer of the wav2vec 2.0 base model:

- Ground-truth/predicted-label pairs
- Average accuracy scores for each phoneme label, manner and place of articulation for each layer
- Phone label confusion matrices for each layer
- Dendrograms depicting sections of the confusion matrices for domain-informed categories

6 Results

This section presents a discussion of the results of the probing task. Firstly, categorisation accuracies for each predicted category per layer were considered. Then, heatmap representations of all phoneme confusions for layers of interest were considered in order to focus on the emergence of specific domain-informed categories, in this case a grouping of the consonants categorised with respect to manner of articulation based on hierarchical clustering.

6.1 Categorisation Accuracies

The accuracy scores for phoneme labels, manner of articulation (MOA) and place of articulation (POA) are presented in figures 2, 3 and 4 respectively. The accuracy scores here were derived by first obtaining a list of phoneme label predictions from the model, and then evaluating the number of correct labels with regard to the total number of predictions. Average accuracies for MOA and POA were derived by applying a category mapping to the original phoneme label predictions.

Of interest in figures 3 and 4 is that robust results are achieved around layer 7 which provides an indicator as to where to focus further investigation. This tallies with results from other work which have demonstrated a similar drop-off in performance in later layers (Belinkov and Glass, 2017).

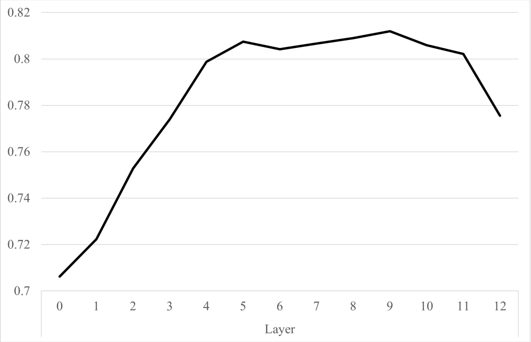


Figure 2: Average phoneme label accuracies per layer

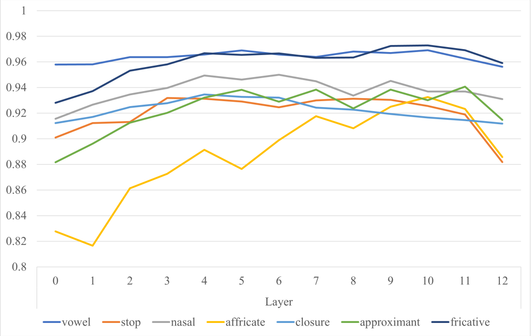


Figure 3: Accuracy per layer for MOA categorisation

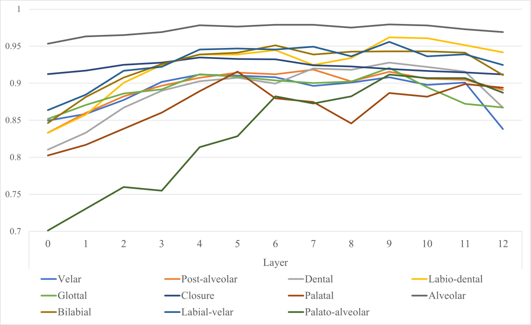


Figure 4: Accuracy per layer for POA categorisation

6.2 Confusion Heatmaps

To better understand the specific intra-categorical relationships captured in the MLP predictions, a confusion matrix was generated for each layer that detailed the confusions between ground-truth phoneme labels and the predicted label. This was done for each layer, with the labels arranged such that phonemes in the same manner-of-articulation category were adjacent. From this, a heatmap visualisation was generated for each matrix such that intra-MOA confusions occupy a contiguous subsection of the overall figure. Figure 5 depicts the overall confusions across all phoneme labels at layers 0, 7, and 12, whereby the bottom right represents vowels and the top left stops, closures, fricatives and affricates. Although the resolution in this figure is low, changes in patterns can be seen in the top left of the heatmap for each layer. For this reason, we have focused on those classes occupying that area in the next section.

6.3 Hierarchical Clustering

To allow assessment of changes in the MLP model's predictive certainty, dendrogram visualisations were created using hierarchical clustering with Ward linkage (Ward, 1963) for sounds with the manner of articulation stop, closure, fricative and affricate. This was done by first applying a transformation to the confusion matrix for all phonemes detailed above such that each cell now represented the probability of confusion at a given ground-truth/prediction intersection in the matrix. As this was a probability distribution, each row, representing the confusions for a given ground-truth label, sums to 1. The relevant rows and columns were then extracted as input to the clustering in no particular order. Figures 6, 7 and 8 show the dendrograms for these classes at layers 0, 7 and 12 respectively.

The hierarchical view in this context represents the clusters found by Ward's method in the probabilistic confusion matrices, and proximity in the hierarchy can be understood as representing "similarity", as the clustering method used seeks to minimise the loss of information incurred by merging nodes. Nodes adjacent to each other are minimally variant, with each sub-tree representing a grouping of less-variant nodes. As the data being clustered is the probability outputs from the model's confusion matrix, we can interpret proximity in the dendrogram images as indicating items that the model frequently confuses and hence with proximity within the model's representation of a given phoneme.

There are several patterns of interest captured in the hierarchical view, particularly with respect to the model's apparent enhanced understanding of phonetic structures and positional context. Viewing figures 7 and 8, it can be seen that the model has developed a representation of the various phoneme relationships within the category that better aligns with domain-informed expectations, with e.g. the closure/stop pairs for various stops having con-

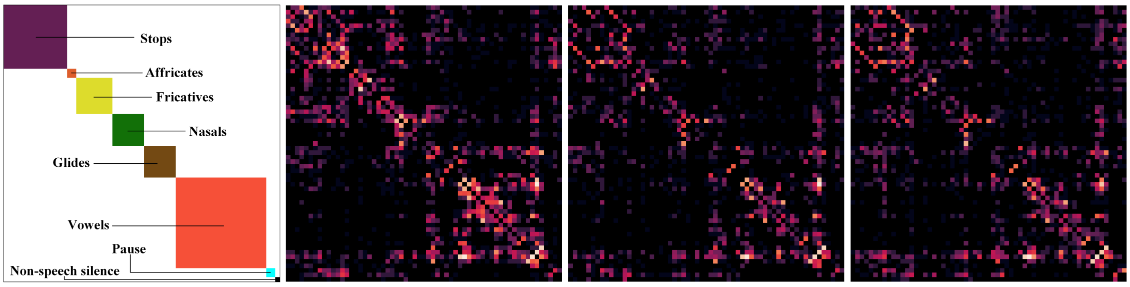


Figure 5: Heatmaps of confusions across all phoneme labels at layers 0, 7, and 12 with vowels in the bottom left quadrant and consonants in the top right quadrant. The leftmost grid describes layout of features within the matrix.

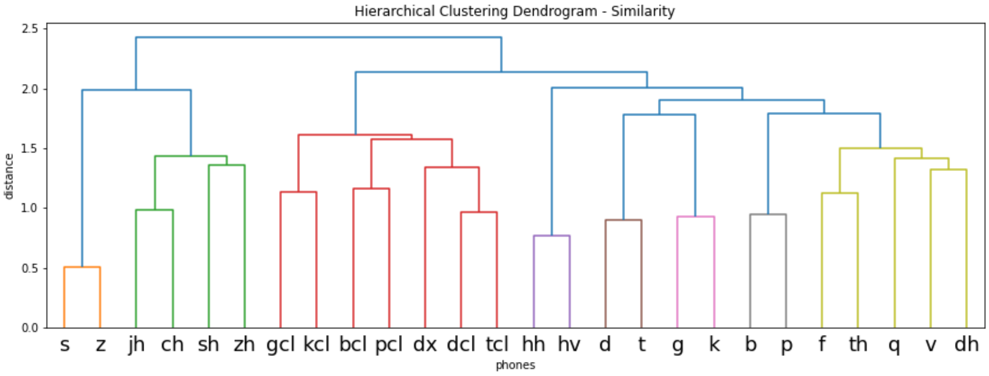


Figure 6: Obstruents at layer 0

verged. The labels /k/ and "kcl"[3], which were significantly detached in layer 0 have repositioned to be adjacent. Similarly, within the fricative region on the right hand side of the figure, the labio-dental fricatives (/f/, /v/) have become separated from the dental fricatives (/th/, /dh/).

Similarly certain acoustically-similar adjacent phonemes in layer 0, such as /d/ and /t/, see significant transformation within the clustering tree. The /d/ and /t/ labels occupy a separated sub-tree within the dendrogram produced for layer 0, but by layer 12 they have transitioned to become proximate to both their closures ("dcl" and "tcl") and their variants, such as /d-/dx/ and /t/-/q/. We can observe further development in this transition in the layer 7 representation (see figure 7) where certain proximate relationships have been established (as between the variants of /t/, /q/, and the closure "tcl") while other positionings remain (as with the inclusion of /t/ in the /d/-/dx/-"dcl" sub-tree).

The positioning of closures ("gcl", "kcl" etc.) is also of interest with regard to the apparent transition from acoustic to positional relations. Initially, given their strong acoustic similarity (representing a lack of sound production) it is intuitive that they should form a distinctive sub-group within the dendrogram, as they do in figure 6. At layer 7 this cluster has already separated significantly into several sub-trees of closure/stop pairs, such as /k/-"kcl". By layer 12, all closures have become proximate to their respective stop label.

7 Conclusions and Future Work

While the specific nature of the phonetic information captured by modern large transformer models will require significant further work to adduce, this paper has demonstrated that there is significant evidence to suggest that transformer architectures are capable of capturing significant levels of phonetic detail that accords with domain-informed understandings of phoneme relationships, and that permit distinction between less separable phonemes. Future work will look to establish more concretely the nature and effective mechanism of the layer-wise changes to these characteristics and the emergence of phonological generalisations, as well as looking to explore other aspects of the mechanisms asso-

[3] We do not describe these labels as phonemes.

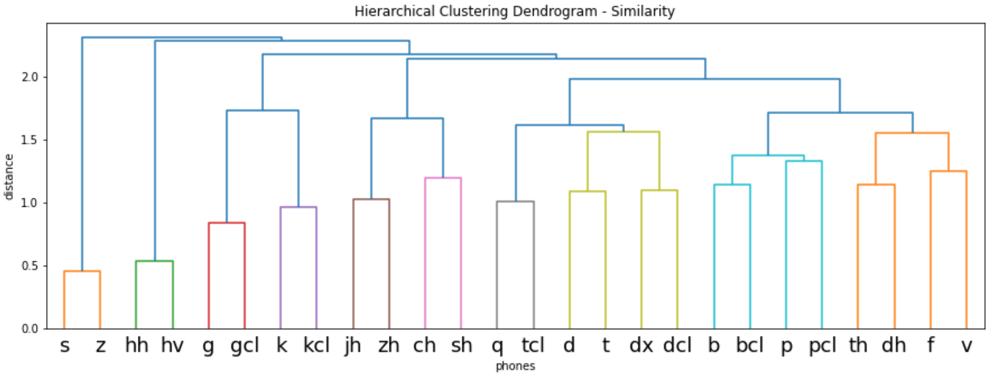


Figure 7: Obstruents at layer 7

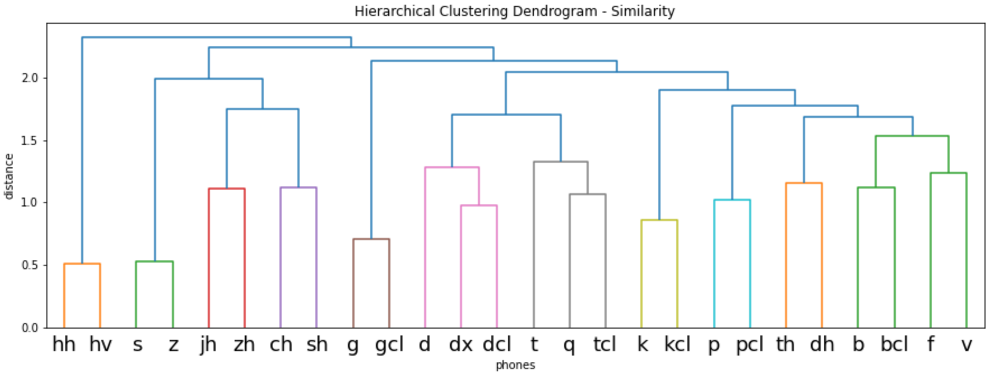


Figure 8: Obstruents at layer 12

ciated with these networks, such as the operation of their feature extractor modules and the attention matrices associated with each layer. While a chance-correlation experiment was conducted for this work, label imbalance in the TIMIT dataset was not specifically accounted for in the probing task; this will be assessed as a next step. Another focus of future work will be the investigation of the relationship of the emerging phonetic categories to infant language acquisition.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jin Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Erich Elsen, Jesse Engel, Linxi (Jim) Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Han, Patrick LeGresley, Xiangang Li, Libby Lin, Sharan Narang, A. Ng, Sherjil Ozair, Ryan J. Prenger, Sheng Qian, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Anuroop Sriram, Chong-Jun Wang, Yi Wang, Zhiqian Wang, Bo Xiao, Yan Xie, Dani Yogatama, Junni Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. *ArXiv*, abs/1512.02595.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Aul. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Neural Information Processing Systems (NeurIPS)*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Association for Computational Linguistics*, 48:207–219.

Yonatan Belinkov and James R. Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In *NIPS*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Wendong Gan, Bolong Wen, Yin Yan, Haitao Chen, Zhichao Wang, Hongqiang Du, Lei Xie, Kaixuan Guo, and Hai Li. 2022. Iqdubbing: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion.

J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. 1992. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium.*

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3651–3657. Association for Computational Linguistics.

Mark Kane and Julie Carson-Berndsen. 2016. Enhancing data-driven phone confusions using restricted recognition. In *INTERSPEECH*, pages 3693–3697.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology.*

Jialu Li, Vimal Manohar, Pooja Chitkara, Andros Tjandra, Michael Picheny, Frank Zhang, Xiaohui Zhang, and Yatharth Saraf. 2021. Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. *CoRR*, abs/1903.08855.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing acoustic representations for phonetic properties. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. corr abs/1301.3781 (2013). *arXiv preprint arXiv:1301.3781.*

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding bert's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62.

Emma O'Neill and Julie Carson-Berndsen. 2019. The effect of phoneme distribution on perceptual similarity in English. *Proc. Interspeech 2019*, pages 1941–1945.

Emma O'Neill, Joe Kenny, Anthony Ventresque, and Julie Carson-Berndsen. 2021. The influence of regional pronunciation variation on children's spelling and the potential benefits of accent adapted spellcheckers. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Leonardo Pepino, Pablo Ernesto Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Interspeech*.

O. E. Scharenborg, Nikki van der Gouw, M. A. Larson, Elena Marchiori, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis. 2019. The representation of speech in deep neural networks. *Lecture notes in computer science*, (Part II).

Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah. 2021. What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *ArXiv*, abs/2101.00387.

Miikka P Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *CoRR*, abs/1905.06316.

Joe H Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.