FOLLOWING THE EMBEDDING: IDENTIFYING TRANSITION PHENOMENA IN WAV2VEC 2.0 REPRESENTATIONS OF SPEECH AUDIO

Patrick Cormac English[⋆†] *Erfan A. Shams*[†] *John D. Kelleher*[‡] *Julie Carson-Berndsen*[†]

[⋆] SFI Centre for Research Training in Digitally-Enhanced Reality (d-real), Ireland
[†] ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland
[‡] ADAPT Research Centre, Hamilton Institute, Maynooth University, Ireland

ABSTRACT

Although transformer-based models have improved the state-of-the-art in speech recognition, it is still not well understood what information from the speech signal these models encode in their latent representations. This study investigates the potential of using labelled data (TIMIT) to probe wav2vec 2.0 embeddings for insights into the encoding and visualisation of speech signal information at phone boundaries. Our experiment involves training probing models to detect phone-specific articulatory features in the hidden layers based on IPA classifications. Furthermore, we propose an analysis framework for visualising the probabilities of the detected articulatory features in every layer and frame vector. Our primary focus is to probe and better understand the structure of speech signal information in the embeddings learned by unsupervised transformers, with a view to contributing to more explainable speech processing systems.

***Index Terms*—** Speech Recognition, Phonetic Representations, Probing, Explainable AI

1. INTRODUCTION

The transformer neural network architecture has significantly improved performance in various tasks, including natural language processing, computer vision, and automatic speech recognition (ASR) [1]. However, performance evaluation often relies on standard metrics, lacking insight into the specific information contained within encoded representations used by the model [2]. Probing hidden representations in pre-trained models has been extensively explored to understand domain-specific knowledge encoding [3], and recent studies have investigated linguistic and acoustic knowledge representations in ASR models [4]. Despite these advances, explainability in the context of speech embeddings in transformer models remains limited.

In this study, we address the challenge of analysing neural embeddings of speech for phonetic information given sparse human-annotated data for the labelling of speech embeddings. Typically, annotations are available only at the phone level, which may encompass several of the frames used by a model to represent the speech signal. Previous work [5] has associated labels with embeddings by averaging embeddings that occur over the duration of a phone, treating the averaged embedding as a representation of that phone. However, this approach limits resolution to some extent.

With this context in mind, we investigate the potential of using labelled data from the TIMIT dataset [6] to probe wav2vec 2.0 embeddings which represent 25 ms audio per embedding vector [7] for insights below the level of the phone annotation present in the dataset. Our research question is as follows: Can we find traces of learned articulatory features such as manner of articulation (MOA) and place of articulation (POA) in the averaged embeddings of the self-supervised model through probing techniques? If so, is it possible to use the probes to effectively identify these features in the individual 25 ms embeddings as well, leading to a more granular explanation of the hidden representations in the model?

Our methodology involves training small multilayer perceptron (MLP) models on speech embeddings averaged across the duration of each phone utterance. The timings and classes are derived from TIMIT and the latter is mapped to our chosen features. Then, we use our probes to perform feature detection on unaveraged time-step embeddings from our test set. The rest of the paper is organised as follows: Section 2 reviews related work. Sections 3 and 4 detail the data generation process, probing methodology, and the visualisation tool developed for examining the embeddings. Our study's results are presented in Section 5, and we conclude with a discussion and outline of future work in Section 6.

2. RELATED WORK

Explainability in machine learning remains a high profile topic due to the opacity of high-performing models [8, 9]. Probing trained models or embeddings, such as BERT [10], has been widely used for explaining system behaviour. Syntax tree structures and idiomatic information have been deduced from word representations [11]. In speech processing models, phonetic information has been analysed in deep end-to-end models based on convolutional and recurrent neural

networks [12]. Frame-level probing experiments on wav2vec and DeCoAR acoustic models revealed better probe performance with neural embeddings [13]. Transformer-based models, such as wav2vec 2.0 and Mockingjay, have been probed for encoding audio, fluency, and suprasegmental pronunciation [14, 15]. Phone embeddings extracted from the Allosaurus model were analysed for discourse function information [16]. Phonetic categorisation capabilities and phonetic feature event patterns were explored in the wav2vec 2.0 model [5, 17]. Accent identification fine-tuning of a wav2vec 2.0 model revealed richer representations of phoneme and prosody features [18].

Work by Abdullah et al. [19] explored the relationship between phonetic categories and discrete units learned by self-supervised speech models (including wav2vec 2.0 embeddings) to quantise speech, using an information-theoretic framework, and discovered that discrete units correspond to sub-phonetic events rather than high-level phonetic categories. This was significant for our work, as it confirms that the embeddings we probe do contain information at a more granular level than our phone labels. These findings contribute to the understanding of how neural representations capture phonetic information and their implications for explainability in large transformer models. More recently, work by ten Bosch et al. [3] investigated the latent representations of the wav2vec 2.0 model, and found that phone-identification classifier probes performed well across all layers, indicating the presence of phonetic information sufficient for phone-identification. However, as evaluated via classification runner-up evaluation, the underlying phonetic structure within the embeddings changes, apparently encoding less phonetic-acoustic structure at higher layers. This change in structure alters ambiguity at phone decision boundaries, often unintuitively. As an example, the authors note that vowels grouped in layer 1 of a transformer had moved apart to a "convex pattern" by layer 18, in a manner that did not occur as significantly for consonants. Furthermore, the study reveals that static embeddings do encode additional information in a manner that allows for higher-order categorisation of phones, such as an increased likelihood of runner-up phones being in the same "Broad Phonetic Class". However, this pattern fluctuates significantly across layers and differs for various phones.

We aim to build on our previous work, contemporaneous with the above [17], which probed for phenomena presence in time-step embeddings. In the current paper, we extend this by expanding the method to probe phone boundaries for a domain-informed set of features, providing further insights into the consistency and divergence in phonetic structure where phone identification proves unintuitive, and by developing a visualisation method for examining embedding representations of speech signals, offering additional explainability through structured exploration of wav2vec 2.0 representations of speech signals. Our emphasis on probing for sub-phonetic features aims to contribute to a deeper understanding of the wav2vec 2.0 model's organisation of acoustic-phonetic space.

3. DATA GENERATION

The following sections describe the generation of the various datasets used as part of this work.

3.1. wav2vec 2.0 Representation Generation

We generated wav2vec 2.0 embeddings for TIMIT training and test sets, using the default configuration, obtaining a $13 \times N \times 768$ tensor per wav file (12 layers + CNN output). N is the number of 25 ms frames with 20 ms stride. Utilising TIMIT annotations, we mapped time-step representations to phone labels and reserved $258,040 \times 768$ samples for probe evaluation. We created phone-averaged representations (PARs) using the process described by English et.al [17] from TIMIT training set samples, yielding 13 datasets (175,232 PARs per layer). A similar dataset of 63,555 PARs was generated for the TIMIT test set.

3.2. Choice of Phonetic Features

In this study, the International Phonetic Alphabet (IPA) phone classification framework is employed to select phonetic features for the probing task. This framework provided a systematic and widely-accepted basis for selecting phonetic features that are crucial for understanding speech sounds. Chosen features encompass category (consonant, vowel, silence), vowel features (height, front, rounding), and consonant features (MOA, POA, voicing), chosen for their comprehensive representation of speech sounds and capacity to capture critical distinctions. Utilising IPA-based categories enables effective evaluation of embeddings' phonetic information encoding and facilitates meaningful interpretation within speech science research. The full phone-feature classification mapping can be found in [20].

4. PROBING TASK DESCRIPTION AND METHODOLOGY

A probing task involves training a domain-specific classifier using hidden representations from a pre-trained model, such as the wav2vec 2.0 base model trained on the Librispeech corpus. The following sections describe the training of the probes used in this work, and detail the analysis methodology based on the outputs of these probes.

4.1. Probe Training

For the probing task, we trained 91 MLP models (7 features, 13 layers) to predict articulatory feature presence using

masked phone-averaged wav2vec 2.0 embeddings. Scikit-learn library [21] was employed for MLP implementation, featuring a single hidden layer with 200 ReLU activation neurons and k output neurons using logistic or softmax activation function based on k (binary or multi-class), where k represents the number of classes for each articulatory feature or category. Hyperparameters adhered to scikit-learn's default settings, except for increased hidden layer size. A dataset of 175,232 PARs was used for training, with each sample being a wav2vec 2.0 representation and phone label. Models were trained on $175,232 \times 768$ representations and received $175,232 \times 1$ feature presence labels as target categories, based on the phone label mapping [20]. Probing models predicted articulatory feature presence in time-steps from PARs. To ensure probe performance was not influenced by chance correlation, randomised datasets were devised for comparison, maintaining the same number of features, samples, and target labels as described above. The values of these datasets are in the range of the corresponding features in the embeddings. Probes trained on randomised datasets failed to predict features effectively, exhibiting performance levels aligned with random chance baseline.

4.2. Contextual Metadata Generation

In order to facilitate bulk evaluation of patterns discerned during the following analysis, it is necessary to generate metadata about frames for the test dataset, enabling the identification of specific phone boundaries. To achieve this, 13 additional probes are trained to predict TIMIT phones based on averaged phone representations. Subsequently, for the entire test dataset, pertinent information is generated including the target phone (derived from TIMIT metadata), the predicted phone (as determined by the phone-label probe), the preceding and following phones (as per the TIMIT metadata), word context, and frame indices corresponding to individual frames that constituted the averaged representation. This process allowed for subsequent extraction of frames comprising target phones (e.g., TH preceded by a K label[1] or correctly/incorrectly predicted phones as per the probe) for use in the second probing task.

4.3. Probe Outputs

Multiple probes are used for predicting phonetic features of a single time-step representation. The category probe (see Section 3.2) serves as an identifier to select the suitable probe. Then, feature probabilities are generated by the specific articulatory probes. For example, when a frame is predicted to be a consonant (on a per layer basis) the MOA, POA and voicing probe corresponding to a particular layer are used. It is important to note that probes yield a probability distribution for the feature group. As an example, the MOA probe for a layer produces eight probabilities corresponding to each of the eight consonantal MOAs (*plosive*, *nasal*, *trill*, *tap/flap*, *fricative*, *lateral fricative*, *approximant*, *lateral approximant*).

[1]We use ARPAbet notation throughout, e.g. F is [f]

4.4. Visualisation Tool

To facilitate analysis and enable manual inspection of probe outputs, we developed a visualisation tool which we refer to as *w2v2viz*. This takes a wav file and generates probe outputs for each time-step representation across output layers. These outputs are visualised as a 3D terrain, with X and Y axes representing MOA and POA categories, and the Z axis corresponding to the product of probability distributions for each feature intersection in consonant frames. Users can vary layers and frames using sliders, enabling evaluation of probe feature probabilities at each time-step in wav2vec 2.0 representations. Peaks observed at specific intersections signify high probe confidence in the presence of a particular articulatory feature pair within the time-step (e.g., in Figure 2, MOA/POA for consonants). Additional details are displayed in textual format, including target phone (provided metadata is available) and third-feature detection such as voicing and rounding.

5. RESULTS

This section presents the probing task results and discusses their significance within the analysis framework as a contribution towards modern ASR model explanations. Figure 1 presents the layer-wise accuracy scores for the feature and phone probes. These values are with respect to the PARs in the test-set. Layer 9 demonstrates suitable generalised accuracy across various features and is selected for the frame-by-frame exploration. A specific coarticulation example, the transition from NG to TH in the word *stre**ngth***, is chosen, with epenthesis of a K sound for smoother transition.

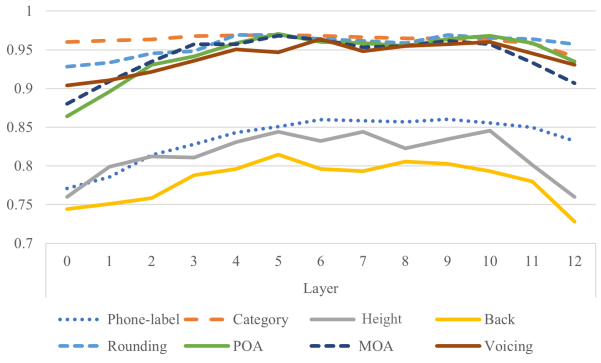

Fig. 1. Feature probe accuracies per layer.

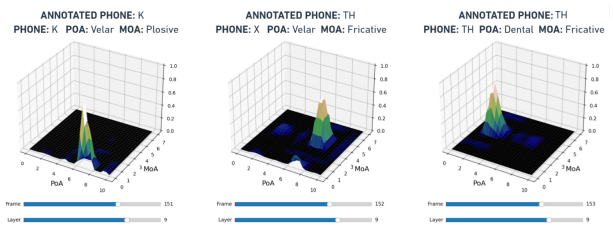

Fig. 2. Visualisation of K-TH transition across 3 frames.

5.1. Visualisation Tool Results

Figure 2 displays the *w2v2viz* outputs for the K-TH transition. Frame 152, part of the TH phone, accurately identifies the MOA as *fricative* and the POA as *velar*, with a smaller probability for *dental* POA, which becomes the primary peak in frame 153. This finding highlights the successful identification of an articulatory transition between K and TH phones, wherein the speaker adjusts the MOA to match the subsequent phone while maintaining the POA of the preceding phone. These results imply that individually-trained feature probes can identify phenomena at a granularity below phone-level training, holding promising implications for future research.

We also examine the transition frames between the same two phones in different utterances, particularly focusing on the N-F transition, as we expected to encounter instances of nasal assimilation. Figure 3 displays the transition between an "N" and "F" phone in two utterances for the word "uN-Fused." Subfigure A presents a bilabial assimilation, wherein the speaker's lips close as a result of the influence of the upcoming F on articulation. The middle frame reveals a detection for the articulatory features of a P phone, exemplifying epenthesis, an insertion that facilitates transition in articulation. In this instance, the airflow changes its path during the production of the M sound, initially passing through the nasal cavity, then transitioning to the vocal cavity, leading to the P sound before the lips open for the frication necessary to produce the F sound. This is detected by the probes as a confidence spike at the *bilabial-plosive* intersection. In contrast, Subfigure B demonstrates the transition without bilabial assimilation, presenting a direct transition where the probes identify the persistence of the POA feature as the speaker transitions to the production of the fricative F.

6. CONCLUSIONS AND FUTURE WORK

We have presented an approach for making use of limited labelled data to probe the phonetic information encoded within the transformer-based wav2vec 2.0 model, by training small MLP models to detect phone-specific features in the hidden layers based on IPA classifications. Our results demonstrate that our probing methodology is capable of identifying nuanced articulatory phenomena occurring below the level of annotation in our labelled data. The probes also successfully

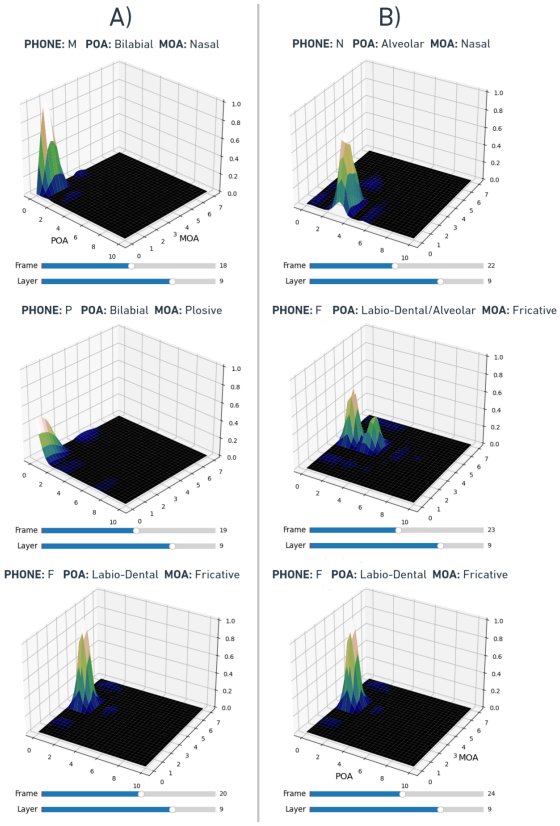

Fig. 3. Visualisation of N-F transition with: A) *Bilabial* assimilation and epenthesis of P. B) No **bilabial** occurrence.

identify articulatory transitions that were not recorded by annotators in the TIMIT dataset, notably constraints on POA transition. Furthermore, our visualisation tool facilitates investigation of a single frame across the transformer layers, tracking the emergence of features through the model. This is something we aim to explore in follow-up work which will track the development and identification of boundary phenomena across layers to uncover how different layers contribute to the encoding of phonetic features. We also aim to expand this work to assess a wider range of phonetic events to allow for the evaluation of phonetic theory and explainability within the embedding space.

7. ACKNOWLEDGEMENTS

We thank our colleague Margot Masson for contributing the phone coordinate scheme used in the visualisations. This work was conducted with the financial support of Science Foundation Ireland at the SFI Centre for Research Training in Digitally-Enhanced Reality (d-real) [18/CRT/6224], and at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at University College Dublin [13/RC/2106_P2]. 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.

8. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[3] Louis ten Bosch, Martijn Bentum, and Lou Boves, "Phonemic competition in end-to-end ASR models," in *INTERSPEECH*, 2023, pp. 586–590.

[4] Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller, "Probing Speech Emotion Recognition Transformers for Linguistic Knowledge," in *INTERSPEECH*, 2022, pp. 146–150.

[5] Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen, "Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features," in *SIGMORPHON*, 2022, pp. 83–91.

[6] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020, vol. 33, pp. 12449–12460.

[8] Amruta Kale, Tin Nguyen, Frederick C. Harris, Chenhao Li, Jiyin Zhang, and Xiaogang Ma, "Provenance documentation to enable explainable and trustworthy AI: A literature review," *Data Intelligence*, vol. 5, no. 1, pp. 139–162, 2023.

[9] Waddah Saeed and Christian Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, pp. 110273, 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805v2*, 2019.

[11] Filip Klubička, Vasudevan Nedumpozhimana, and John D. Kelleher, "Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space," *arXiv:2304.14333v1*, 2023.

[12] Yonatan Belinkov and James Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *NeurIPS*, 2017, vol. 30.

[13] Danni Ma, Neville Ryant, and Mark Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP*, 2021, pp. 311–315.

[14] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure," *arXiv:2101.00387v2*, 2021.

[15] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP*, 2020, pp. 6419–6423.

[16] Pin-Er Chen, Yu-Hsiang Tseng, Chi-Wei Wang, Fang-Chi Yeh, and Shu-Kai Hsieh, "Analyzing discourse functions with acoustic features and phone embeddings: non-lexical items in Taiwan Mandarin," in *ROCLING*, 2022, pp. 136–146.

[17] Patrick Cormac English, John D Kelleher, and Julie Carson-Berndsen, "Discovering Phonetic Feature Event Patterns in Transformer Embeddings," in *INTERSPEECH*, 2023.

[18] Mu Yang, Ram C. M. C. Shekar, Okim Kang, and John H. L. Hansen, "What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model," *arXiv:2306.06524v1*, 2023.

[19] Badr M. Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, and Dietrich Klakow, "An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech," *arXiv:2306.02405v1*, 2023.

[20] Patrick Cormac English, Erfan A. Shams, John D. Kelleher, and Julie Carson-Berndsen, "Probe Visualizer," https://github.com/erfanashams/w2v2viz, 2023.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.