Interspeech 2024
1-5 September 2024, Kos, Greece

Are Articulatory Feature Overlaps Shrouded in Speech Embeddings?

Erfan A. Shams[1], *Iona Gessinger*[1], *Patrick Cormac English*[2], *Julie Carson-Berndsen*[1]

[1]ADAPT Research Centre
[2] SFI Centre for Research Training in Digitally-Enhanced Reality (d-real),
School of Computer Science, University College Dublin, Ireland

erfan.shams@ucd.ie, iona.gessinger@ucd.ie,
patrick.english@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

Domain-informed probing can offer important insights into the types of phonetic information encoded in transformer-based speech recognition models. This paper focuses on phonetic feature probes and investigates whether feature spreading and assimilation are evident in the speech embeddings of the transformer model. Probes are trained for place and manner of articulation and voicing features according to the IPA feature classification, and exemplar fricative consonant clusters where local assimilation would be expected are selected. By then following the articulation trajectory of all of the features during inference, we explore how the transformer model encodes coarticulation and transitions between sounds in the latent representations, by tracking not only the features with the highest activation value but also alternative activations. The patterns identified appear to be in line with expectations from the literature and demonstrate the explanatory power of such an approach.

Index Terms: phonetic features, feature spreading, speech transformer models

1. Introduction

Probing pre-trained large language models, as a methodology for investigating the extent to which such models capture linguistic knowledge, has become an active area of research in natural language processing (NLP) and explainable artificial intelligence (XAI). In NLP, domain-informed probing tasks involve training simple classifiers which probe the internal layers of a transformer-based model to gain insights into the nature of the linguistic information or explain the features encoded in the representations, also known as embeddings (e.g., [1]). This approach is similarly becoming popular as a means of interrogating transformer-based speech recognition models to discover more about the encoding of underlying linguistic, articulatory or acoustic information that the model has learned during training. Large pre-trained transformer-based speech recognition models coupled with domain-informed probing tasks offer huge opportunities for speech science. Phonological feature theory has long sought to explain the nature of such features and their ability to describe real speech phenomena. There have been proposals for both acoustic and articulatory feature classifications and approaches which argued for non-segmental (multilinear or autosegmental) representations which capture overlap of independent features. Some of these approaches were designed to describe specific phonological phenomena in languages where segmental approaches were limited (such as vowel harmony) and others were designed for the purposes of modelling speech production (e.g., [2]) or recognition (e.g., [3]). In the latter approaches, emphasis was placed on independent temporal gestures or events which in combination (through overlap or precedence) resulted in multi-tiered patterns of phonetic features capturing a speech utterance. These approaches provided a framework for description and analysis of phenomena, such as feature spreading, which give rise to assimilation, elision and epenthesis. The question we now ask in this paper is whether such phenomena are evident in transformer-based speech recognition models, i.e., shrouded in the speech embeddings. Although the transformer-based model has been trained to recognise character or phone sequences, we illustrate how domain-informed probes can glean information about articulatory trajectories where varying degrees of assimilation would be expected, thus demonstrating the potential of probing for speech science more broadly by providing insights for phonetic and phonological theories which to date have not had access to such large data sets. For this illustration, we chose to investigate the potential assimilation in fricative consonant clusters (CCs) where [s] and [ʃ] may accommodate each other for ease of articulation during normal speech. Using agglomerative clustering, we identify local assimilation patterns, taking into account not only the top predictions but also the secondary predictions of the model. While we use exemplars from the TIMIT data set [4] with the transformer-based speech recogniser wav2vec 2.0 [5], the probes can be applied more broadly on any data set to investigate how the model captures phonetic relationships. The paper is structured as follows. In Section 2, we set out related work and point to some phonological theory background which has informed the approach. In Section 3, we describe the feature set, the data and the probing task. Section 4 outlines the exemplar selection and clustering methodologies, followed by the analysis of the emerging patterns in Section 5. We conclude the paper addressing limitations and future work in Section 6.

2. Related Work

One of the primary motivations for the work presented in this paper has been to explore the extent to which modern transformer-based speech recognition models encode phonetic and phonological information in their latent representations. From the perspective of speech science, deep learning models have typically been assumed to be black boxes which perform their main task of speech recognition extremely well but do not have a contribution to make towards confirming or refuting existing phonetic or phonological theories. While we do not claim in this paper to provide answers as to whether particular theories are correct or not, we do demonstrate how probing the models based on expectations which have already been well documented in the literature can provide a framework and suite of tools for inspecting the models in more depth making an explicit contribution towards explainability.

There have been a number of researchers who have pro-

10.21437/Interspeech.2024-1039

posed deeper investigation of speech and speaker embeddings, looking to disentangle certain speaker-specific and stylistic acoustic-phonetic features that explain observed variations [6, 7, 8]. Others have addressed explainability of deep learned models through analysis of the latent embedding spaces using PCA clustering [9] and probing tasks which focus on phone-level representations [10, 11, 12]. Most recently, focus has been on probing for more targeted broad phonetic classes and phonetic features [13, 14]. This paper goes beyond those approaches in that it explicitly considers whether expected phonetic patterns, such as assimilation, are encoded in the latent representations. If the feature probes are independent, do they encode how the articulators relate to each other?

As outlined above, other related work takes us back some decades to Articulatory Phonology [2] which proposed that gestural overlap or gestural blending underlies phonetic realisations as a result of local assimilation. Nolan et al. [15] examined this further, specifically with respect to assimilation where [s] and [ʃ] occur at word boundaries, to determine whether this phenomenon can be described by varying gestural configurations or whether this could be the result of cognitive planning. We do not get into the details of the pros and cons here, but rather look to see whether transformer-based models also capture assimilation patterns. Similar to the gestures of Articulatory Phonology, we assume that manner and place of articulation and voicing features are independent yet work together to form a speech pattern where various accommodations are made across the [sʃ] cluster.

3. Probing

A probing task is a post-hoc XAI method for investigating the latent representations or embeddings of a model in the scope of a downstream task to evaluate the capacity of the embedded information in identifying the specified features [11]. The probes are simple machine learning models such as multi-layer perceptrons (MLPs) with a small number of parameters trained on the latent representations of the model. In our case, the downstream task is phonetic articulatory feature identification using the TIMIT data set. The chosen articulatory features for consonants are place of articulation (POA), manner of articulation (MOA) and voicing, where the feature mapping is derived from the IPA classification for pulmonic consonants [16]. For the affricates [tʃ] and [dʒ], POA and MOA mappings correspond to the fricative segment.

The TIMIT data set with the standard train/test set split is used as input to the wav2vec 2.0 model to extract latent representations. Based on the audio inputs, the feature encoder (layer 0) and each of the 12 transformer layers of the model output an $N \times 768$ vector where N is the number of encoded frames roughly equal to $25\,\mathrm{ms}$ of the original audio per frame. Each encoded frame is then labelled according to the time-aligned phone-level transcription of TIMIT converting the TIMIT timestamps (t) to the latent representation timestamps (τ) using (1).

$$\tau = nint(t/\rho) \tag{1}$$

where $nint$ denotes the nearest integer and ρ is the number of audio samples per frame of the latent representations calculated by dividing the total number of audio samples by N. Consecutive frames with the same label are averaged into a single frame and mapped to their corresponding articulatory features.

Following the labelling process for the extracted latent representations, a total number of 39 probes are trained, i.e., one probe per articulatory feature (POA, MOA, voicing) for each layer of the model, including the output of the feature encoder (layer 0) and the transformer layers (1-12). In line with our previous research [14, 17], we used the MLP module of the scikit-learn library [18] for the probes where we set the number of input and hidden units to 768 and 200 for the single hidden layer, respectively. The number of units for the output layer varies based on the number of classes of the articulatory features. The output vector uses softmax activation where each element represents the probability of an articulatory feature. These activation values play an essential role in the clustering process explained below. We set the maximum number of training epochs to 200 and adhered to the default hyperparameters of the module. All probes process the same inputs but the outputs are independent of each other. We also trained three control probes (POA, MOA, voicing) with the same number of samples, input features, and identical labels where the input feature values are randomly generated based on the value range of the original embeddings. The performance of the randomised set closely resembles the random chance distribution whereas the original embeddings, despite having no knowledge of the acoustic realisations, show a high capacity in identifying the feature similarities based on distributional patterns. Overall, layer 9 demonstrates the highest capacity for identifying the chosen features while layers 0 and 12 illustrate the lowest performance. Hence, layer 9 is used for the analysis below.

4. Exemplar Selection and Clustering

To analyse the assimilation patterns in TIMIT, we identified utterances exhibiting the [sʃ] CC where the adjacent [s] and [ʃ] may affect the POA and/or voicing of one another, i.e., where local assimilation is expected. Examples such as "ga**s** **sh**ortage" or "thi**s** **sh**ellfish" were selected resulting in a set of 48 exemplars from 12 unique sentences. All utterances were produced by different speakers, yet some sentences occurred several times in the data set. Due to the automatic extraction process which is delimited by vowels, the CCs may also include consonants other than [s] and [ʃ], e.g., [t] in "mark**e**t**s** **sh**ould".

The latent representations from layer 9 of the transformer-based model for the selected exemplars are extracted whereby we identify the corresponding frames for the said CCs and probe them for the articulatory features discussed above. Each frame outputs three sets of probability vectors for POA, MOA and voicing which are then concatenated and arranged as a multivariate time series for each exemplar. However, since the utterance length varies across the data set, we use dynamic time warping (DTW) [19] to align the exemplars into the same number of frames. For this, a template must be selected as a basis for warping.

To ensure that the chosen template is a suitable representative of the exemplar set, we define a set of criteria where an exemplar is selected as the template if 1) it includes only the expected consonants related to the chosen CC realisations (i.e., preferably both [s] and [ʃ] rather than just one of the two, and no unexpected consonants or silent frames), 2) it has the highest frame count, 3) the number of frames for each unique phone is closest to the average value of the entire set, 4) the mean aggregated representation of the exemplar is closest to the mean aggregation of the entire set.

Following the template selection, DTW with asymmetric step pattern is used to find the best alignment from each exemplar to the template based on their latent representations. One exemplar from the set could not be aligned and was removed.

The final result is a set of $47 \times M \times 21$ multivariate time series where M is the number of frames in the template exemplar ($M = 11$ in the present case), and 21 is the result of concatenating MOA, POA and voicing with 11, 8, and 2 features, respectively.

Agglomerative clustering with complete linkage [20] and Euclidean distance is used to determine where the similarities exist among the exemplars. The Euclidean distances are normalised and used as follows for determining the number of clusters. First, we assume a merging threshold of below 0.6, then we verify whether the distance difference to the next cluster split is above 0.04. Implementing this, the exemplars were grouped into five clusters with 18, 19, 8, 1, and 1 exemplars for each cluster. The last two clusters seem to be outliers. However, the present data set is relatively small and these clusters may become more solid when examining larger data sets, In the next section, we further analyse the articulatory feature patterns in the similarity clusters.

5. Articulatory Feature Pattern Analysis

The similarity clusters are formed based on the concatenated probabilities of the feature vectors, as explained above. Hence, we expect that the utterances which are grouped together also show similar articulation trajectories associated with the probe output activations. That is to say, each cluster is associated with a specific articulatory feature pattern. In this section, we analyse the articulatory feature patterns based on main and alternative probe activations to further explain the information embedded in the latent representations of the transformer-based speech recognition model.

5.1. Alternative Activations

Since the multivariate time series described in Section 4 consists of the probabilities resulting from the probe classifications, their activation values influence the similarity cluster formation. Hence, it is intuitive to analyse the emerging articulatory feature patterns based on the combined probabilities of the articulatory features, as either main or alternative activation.

We derive the main and alternative activations for each frame from the combined probability of POA, MOA, and voicing denoted by $p(a_j)$ where $a_j \in A$ and A is a set of all possible pulmonic consonants in the IPA classification. The set of all activation probabilities $p(A)$ is an $8 \times 11 \times 2$ matrix calculated by (2).

$$p(A) = [\hat{y}_m^T \; \hat{y}_p] \; \hat{y}_v \tag{2}$$

where $\hat{y}_m$, $\hat{y}_p$, and $\hat{y}_v$ are 1×8, 1×11, and 1×2 probability vectors of MOA, POA, and voicing respectively.

The main activation probability $p(a_0)$ is $max(p(A))$ followed by alternative activations $p(a_j)$ which exceed the 5% threshold in a sorted manner, i.e., any activation with a probability of below 5% is not considered an alternative activation.

Figure 1 illustrates the alternative activations of all frames in articulatory feature pattern 2 with scaled circles related to the probabilities of each articulatory feature. POA and MOA are on the left and right axes and the main activations are connected using solid green and dashed red lines respectively. Voicing is indicated by filled (voiced) or hollow (voiceless) triangles. The combined main activation in frame 0 is voiced postalveolar fricative, i.e., [z]. However, there are other activations for each articulatory feature which are comparatively strong. Among the alternative activations, the probability of voicelessness has the highest value, followed by dental and palatal activations, respectively. These alternative activations suggest that the probe also detected traces of [s] (change to voiceless), [ð] (change to dental POA) and [ɹ] (change to palatal POA) encoded in the frame embeddings. The final four frames, however, do not show any trace of a significant alternative activation.

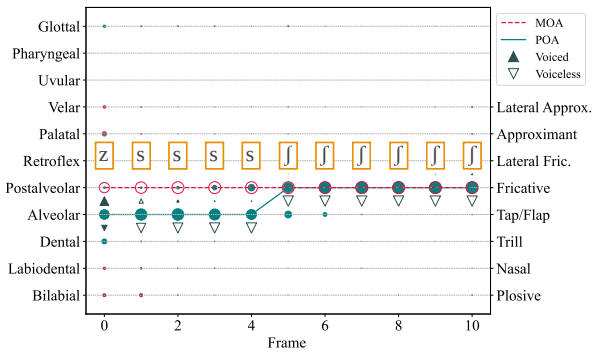

Figure 1: *Activations per frame in pattern 2 for POA (left axis; ●), MOA (right axis; ○), and voicing (▲▽). Symbol size corresponds to predicted probability. Main prediction indicated by connecting line (for POA and MOA) or symbol filling (for voicing). Given sound corresponds to highest combined probability.*

5.2. Emerging Articulatory Feature Patterns

Table 1 provides an overview of the mean predictions in each of the five articulatory patterns for the transitions found in the [sʃ] CCs by indicating the main activation per frame. Colour coding is applied to highlight the POA of the two expected fricative classes in the CC: alveolar (green) and postalveolar (light green). The main activations are followed by those alternative activations that exceed the 5% threshold; note that the relevant IPA symbols are used in the table to represent the combination of activations for POA, MOA and voicing respectively, e.g., [z] is a shorthand for the activations *alveolar, fricative, voiced*. The smaller font indicates alternative activations. This illustrates areas that show a lot of variation (e.g., pattern 1, frames 0-1), some variation that is straightforwardly related to the transition of sounds in the CC under examination (e.g., pattern 1, frames 2-4), and areas that are very uniform with respect to the predicted sound (e.g., pattern 1, frames 5-10). We can see that the alternative activations for frame 0 of pattern 2, are the same as described in Section 5.1: [s], [ð], [ɹ].

In some cases, the activations predict sounds as voiceless (0) where only voiced variants of the respective POA/MOA combination exist according to the IPA classification (e.g., [ɹ^0]) or sounds are predicted as postalveolar (P) where the respective POA/MOA combination spans from dental to postalveolar in the IPA classification and is usually denoted as alveolar (e.g., [t^p]). This further illustrates that the three feature sets POA, MOA, and voicing are independent and predicted separately in the present work, which is crucial to the interpretation of the results.

Based on the main activations, we can describe the overall patterns which have emerged from the clustering in terms of POA, MOA and voicing. The majority clusters, pattern 1 (n=18) and pattern 2 (n=19), differ in the spread of the POA with pattern 1 showing less alveolar and more postalveolar, whereas alveolar and postalveolar are more balanced in pattern 2. Both

Table 1: *Articulatory feature patterns with first sound indicating main activation, followed by alternative activations above 5%.*

	Frame										
Pattern	0	1	2	3	4	5	6	7	8	9	10
1 (n=18)	[z]-[s]-[ʒ]-[ɹ]-[ʃ]-[ɹ^{0}]	[s]-[ʃ]-[z]-[ʒ]	[ʃ]-[s]	[ʃ]-[s]	[ʃ]-[s]	[ʃ]	[ʃ]	[ʃ]	[ʃ]	[ʃ]	[ʃ]
2 (n=19)	[z]-[s]-[ð]-[ɹ]	[s]-[z]	[s]	[s]-[ʃ]	[s]-[ʃ]	[ʃ]-[s]	[ʃ]-[s]	[ʃ]	[ʃ]	[ʃ]	[ʃ]
3 (n=8)	[k]-[ŋ^{0}]-[g]-[ŋ]-[ɰ^{0}]	[k]-[g]	[k]-[x]-[t$^{\text{P}}$]	[x]-[ʃ]-[k]-[t$^{\text{P}}$]	[x]-[ʃ]-[k]-[t$^{\text{P}}$]	[ʃ]	[ʃ]	[ʃ]	[ʃ]	[ʃ]-[ʒ]	[ʃ]-[s]-[ɹ$^{0\text{P}}$]-[ʒ]
4 (n=1)	[ɹ^{0}]-[ɹ]-[ɹ$^{0\text{P}}$]	[ɹ^{0}]-[ɹ]-[ɹ$^{0\text{P}}$]	[s]-[ɹ^{0}]-[ʃ]-[ɹ$^{0\text{P}}$]	[s]-[ɹ^{0}]-[ʃ]-[ɹ$^{0\text{P}}$]	[ʃ]-[s]	[ʃ]	[ʃ]	[ʃ]	[ʃ]	[ç]-[ʃ]	[j]
5 (n=1)	[s]	[s]	[s]	[s]	[s]	[s]	[s]	[s]-[ʃ]	[s]-[ʃ]	[s]-[ʃ]	[s]-[ʃ]

alveolar fricative | postalveolar fricative | other fricatives | other | ◌^{0} voiceless | ◌$^{\text{P}}$ postalveolar

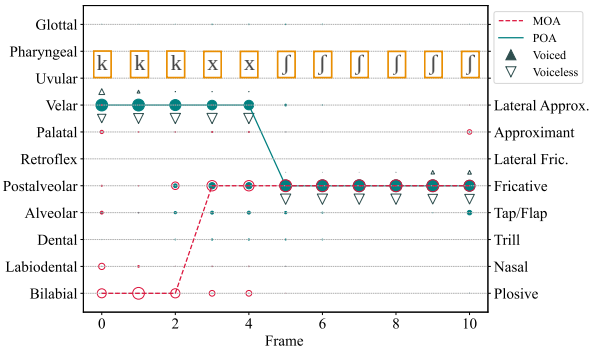

Figure 2: *Activations per frame in pattern 3.*

patterns have very limited voiced portions at the beginning and are mostly unvoiced. Taking a closer look at the transition from [s] to [ʃ] in frames 3 to 5 of pattern 2, we can see main alveolar activation in the first two, and main postalveolar activation in the last, each with the respective other as an alternative activation. Relating this back to Figure 1, we find that the probability of an alveolar POA actually decreases continuously, while that of a postveolar POA increases. The alveolar influence is much stronger in frame 3, only slightly stronger in frame 4, and finally overtaken by the postalveolar influence in frame 5. This aligns with the notion of *gestural blending* where there is a gradual change of the POA articulatory trajectory from alveolar to postalveolar (i.e., the degree of contribution of the specific POA feature to the realisation of the sound changes gradually).

Pattern 3 is fully unvoiced and shows a lot of seemingly unrelated velar activation in the beginning. The latter is due to the automatic extraction process of the CCs discussed in Section 4 where there may be other consonants immediately before the [s]. Here there is evidence of local assimilation where the [ʃ] is the main realisation of the CC, i.e., there is seemingly no realisation of the [s]. However, the pattern illustrates that the transition from the velar plosive [k] into the expected postalveolar fricative [ʃ] (e.g., in "ta**kes Shi**ela") takes place via the velar fricative [x] as predicted for frames 3 and 4. In other words, the velar influence extends into the fricative and covers an alternative postalveolar activation in these frames. This aligns with the notion of *gestural deletion* where the alveolar POA underlying the [s] appears to be deleted through the stronger activation of the velar POA which has spread from the preceding [k]. While it could perhaps be expected from the perspective of the physical articulation of "ta**kes Shi**ela" that there is a continuous movement of the tongue from the velar position to the postalveolar position, it is interesting that the activations predict this with a change in MOA with the POA strongly influenced by the preceding [k]. This can be seen in Figure 2 where the change of the main activation of MOA from plosive to fricative occurs, yet the POA remains with a main velar activation for a further two frames. We can also see in the figure that there is a much smaller alternative alveolar activation but it does not exceed the threshold and is thus hidden.

Regarding patterns 4 and 5 in Table 1, both singleton clusters, the first is similar to pattern 1 but shows activation of features of a voiceless alveolar approximant [ɹ^{0}] (which does not exist in the IPA classification) at the beginning, moving to a palatal activation in the end, while the latter only predicts *unvoiced, alveolar, fricative*, although with increasing postalveolar influence towards the end.

6. Conclusions and Future Work

We have presented an approach to explainability which uses domain-informed probes to investigate the articulatory feature patterns encoded in the speech embeddings of pre-trained transformer-based speech recognition models. The probes have been trained independently, yet they recognise patterns in the transformer layers. In this paper, we have limited our investigation to the [sʃ] CCs in the TIMIT data. As a result, there are two singleton clusters for which we cannot draw any real conclusions. As a next step, we intend to expand the investigation to include the [zʃ] CCs where we expect that this will lead to larger clusters which explain more about the alveolar-postalveolar patterns. This will also allow a more detailed examination using other visualisation tools of the impact of alternative activations in CCs which have the same main activations but belong to different clusters. We further intend to generate new data which will allow us to look specifically at CCs where we expect differences, e.g., as in "clap **Sh**aun" vs. "clap**s Sh**aun" from [15].

An important limitation of our approach is that it relies on the capacity of the probes to capture the embedded information in the latent representations of the model. Even though probes show significant performance at the higher layers of the model, their accuracy is not 100 %, which indicates that our results might be marginally affected by probe misclassification. However, we believe that the probes and the visualisations together constitute a powerful framework and suite of tools for opening the black box of transformer-based models to speech science.

7. Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at University College Dublin [13/RC/2106_P2].

8. References

[1] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, D. Das *et al.*, "What do you learn from context? probing for sentence structure in contextualized word representations," in *ICLR 2019*, 2019, pp. 3179–3195.

[2] C. P. Browman and L. Goldstein, *Tiers in articulatory phonology, with some implications for casual speech*, ser. Papers in Laboratory Phonology. Cambridge University Press, 1990, p. 341–376.

[3] J. Carson-Berndsen, *"Time Map Phonology: Finite State Models and Event Logics in Speech Recognition"*. Dordrecht: Springer Netherlands, 1998.

[4] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[6] J. Williams and S. King, "Disentangling Style Factors from Speaker Representations," in *Proc. Interspeech 2019*, 2019, pp. 3945–3949.

[7] C. Luu, S. Renals, and P. Bell, "Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations," in *Proc. Interspeech 2022*, 2022, pp. 610–614.

[8] M. Kuhlmann, A. Meise, F. Seebauer, P. Wagner, and R. Haeb-Umbach, "Investigating speaker embedding disentanglement on natural read speech," in *Speech Communication; 15th ITG Conference*, 2023, pp. 121–125.

[9] O. Scharenborg, N. van der Gouw, M. Larson, and E. Marchiori, "The representation of speech in deep neural networks," in *MultiMedia Modeling*, 2019, pp. 194–205.

[10] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP*, 2021, pp. 311–315.

[11] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure," *arXiv:2101.00387v2*, 2021.

[12] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features," in *SIGMORPHON*, 2022, pp. 83–91.

[13] L. ten Bosch, M. Bentum, and L. Boves, "Phonemic competition in end-to-end ASR models," in *INTERSPEECH*, 2023, pp. 586–590.

[14] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Discovering Phonetic Feature Event Patterns in Transformer Embeddings," in *INTERSPEECH*, 2023.

[15] F. Nolan, T. Holst, and B. Kühnert, "Modelling [s] to [ʃ] accommodation in English," *Journal of Phonetics*, vol. 24, no. 1, pp. 113–137, 1996.

[16] "Full IPA Chart – International Phonetic Association." [Online]. Available: https://www.internationalphoneticassociation.org/content/full-ipa-chart

[17] P. C. English, E. A. Shams, J. D. Kelleher, and J. Carson-Berndsen, "Discovering Phonetic Feature Event Patterns in Transformer Embeddings," in *ICASSP*, 2024.

[18] 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.

[19] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: the dtw package," *Journal of statistical Software*, vol. 31, pp. 1–24, 2009.

[20] E. K. Tokuda, C. H. Comin, and L. da F. Costa, "Revisiting agglomerative clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 585, p. 126433, 2022.