Under the Hood: Phonemic Restoration in Transformer-Based Automatic Speech Recognition

Iona Gessinger[a,b,c], Erfan A. Shams[a,d], Julie Carson-Berndsen[a,b]

[a]*ADAPT Research Centre, Ireland*
[b]*School of Computer Science, University College Dublin, Ireland*
[c]*Johannes Gutenberg University Mainz, Germany*
[d]*School of Computer Science and Statistics, Trinity College Dublin, Ireland*

Abstract

This study investigates how the automatic speech recognition (ASR) models wav2vec 2.0 *large-960h-lv60-self* and Whisper *large-v3* perform when segment-level signal perturbations (added noise, noisy gaps, and two types of silent gaps) are introduced in English words and pseudowords. We probed the speech embeddings throughout their encoder transformer layers to examine how they encode articulatory features (place and manner of articulation, and voicing). We found that wav2vec 2.0 was more successful than Whisper at restoring perturbed segments across conditions. For wav2vec 2.0 embeddings, classification accuracy was higher in words than in pseudowords. The articulatory features encoding of both ASR models was least disturbed by added noise, and most disturbed by noisy gaps, with silent gaps falling in between. Coarticulatory cues improved classification of articulatory features and classification accuracy increased from early to late layers for both models. Among the examined target sounds, [n] stood out from [m], [ɹ], and [l], as it was classified particularly well under all conditions. We compare ASR performance to the Phonemic Restoration Effect in human speech perception and discuss potential reasons for the performance differences between the two ASR models. This approach aims to foster a better understanding of otherwise opaque systems.

Keywords: explainability, automatic speech recognition, signal perturbation, phonemic restoration

1. Introduction

Noisy or otherwise perturbed speech signals can be challenging to process for automatic speech recognition (ASR). Modern ASR models, like those based on transformers, are designed to directly map the audio waveform to text. These models can

learn complex patterns of variation in speech and are often pre-trained on massive datasets to handle perturbed signals effectively. Techniques such as training models with noisy examples, where intentionally perturbed inputs simulate real-world noise or distortions, make them even more robust to such challenging conditions (Peso Parada et al., 2022).

It is straightforward to assess the model robustness to noise at the highest level, that is, with respect to its textual output, for example, by calculating the word error rate (WER). However, we lack a clear understanding of how the models internally process perturbed signals. With the present study, we aim to address this gap. We believe that a better understanding of the inner workings of modern ASR models will help to make them more explainable.

To take a look under the hood of ASR models, the speech representations (or embeddings) generated throughout their hidden layers have to be converted into information that can be interpreted by humans. This can be done via domain-informed probing tasks that indicate the degree to which relevant information — in the present case, information about articulatory features — is encoded in these embeddings (English et al., 2024).

Even though machines and humans process speech in fundamentally different ways — through metaphorical versus actual neural mechanisms — using experimental paradigms from research on human speech perception to investigate ASR models can provide valuable insights into the inner workings of this technology. For instance, Scharenborg et al. (2018) explored perceptual learning (i.e., a shift in sound category perception due to repeated exposure) in a deep learning (DL)-based ASR model. They found that when the model was trained with data containing the ambiguous speech sound [l/ɹ], it classified this ambiguous sound more frequently as either [l] or [ɹ] depending on the words it encountered during training (i.e., when lexical context favoured one interpretation over the other). This is consistent with the results of experiments on perceptual learning in human listeners (Norris et al., 2003; Drozdova et al., 2016). Other recent work by de Heer Kloots and Zuidema (2024) has demonstrated a phonotactic bias in the transformer-based ASR model *wav2vec 2.0* by using a classic phonetic categorisation experiment. In the same vein, the present work was inspired by the way human perception handles missing speech sounds, that is, by perceptually restoring them based on remaining contextual information (Warren, 1970; Warren and Obusek, 1971; Samuel, 1981, 1987).

While the robustness of ASR models against signal perturbations is usually assessed at the word level (i.e., by measuring the WER; Gong et al., 2023; Olivier and Raj, 2023; Patman and Chodroff, 2024; Shah et al., 2024), the present work examined the segmental level in order to gain a more detailed understanding of how

such models process missing or obscured sounds throughout their layers. We used the Phonemic Restoration Effect (PRE) as a theoretical framework for this study to compare the way in which machines process speech signals and the way in which humans perceive them. In particular, we explored versions of the transformer-based ASR models *wav2vec 2.0* (Baevski et al., 2020) by Meta AI and *Whisper* (Radford et al., 2023) by OpenAI. In order to make their inner workings interpretable for humans, we applied a domain-informed probing task that relates the speech embeddings generated by the models to articulatory features. The articulatory features we considered are place of articulation (POA), manner of articulation (MOA), and voicing (VOI), corresponding to the IPA classification for pulmonic consonants (International Phonetic Association, 2015). They provide a linguistically meaningful basis for probing the information encoded in the embeddings. We aimed to answer the following research questions (RQs):

- **RQ1:** How do different types of segment-level perturbations (added noise, noisy gaps, and two types of silent gaps) affect the articulatory feature encoding in the embeddings of the two selected ASR models?
- **RQ2:** How does the fact that the stimulus is a word or a pseudoword affect the articulatory feature encoding in the embeddings of the two selected ASR models?
- **RQ3:** How does the target sound itself affect the articulatory feature encoding in the embeddings of the two selected ASR models?
- **RQ4:** How does the layer region (early, middle, and late) affect the articulatory feature encoding in the embeddings of the two selected ASR models?
- **RQ5:** How does model performance regarding RQ1, RQ2, and RQ3 compare to human speech perception?

We propose hypotheses for these RQs in Section 3.4, as they can be defined more clearly once the stimuli (Section 3.1), the ASR models (Section 3.2), and the probing task (Section 3.3) have been introduced in detail. The hypotheses are followed with a description of our statistical analysis (Section 3.5), the results (Section 4), discussion (Section 5), and our conclusion (Section 6). In the following sections, we first provide background information on the Phonemic Restoration Effect (Section 2.1), the robustness of ASR against perturbed signals (Section 2.2), and the concept of domain-informed probing (Section 2.3).

2. Background

2.1. Phonemic Restoration Effect

The Phonemic Restoration Effect (PRE) is a phenomenon of human speech perception in which listeners perceive missing sounds in spoken language as if they were present. The effect was first described by Warren (1970), who showed that listeners perceive speech as complete even when a segment of it is replaced by noise. This demonstrates that listeners are able to compensate for degraded auditory input and perceptually restore the replaced segment. The effect is rooted in the ability to use top-down processing, integrating lexical context in order to maintain a coherent perceptual experience. However, it also depends on bottom-up acoustic-phonetic information, since it only occurs when a replacement noise is present and not when the target sound is merely removed (i.e., replaced by silence; Warren and Obusek, 1971).

To explore the relative impact of bottom-up and top-down information on phonemic restoration, Samuel (1981) proposed an experimental paradigm in which a sound is either replaced by noise (type 1) or has noise added to it (type 2). The second stimulus type thus represents the signal constellation that the PRE is expected to generate: a speech sound overlaid with noise. This paradigm can evaluate the perceptual discriminability of the two stimulus types. The PRE should make type 1 sound like type 2, which would result in low discriminability. At the same time, this paradigm can assess the cognitive bias of judging a speech signal as intact based on a postperceptual decision process. The latter could be compared to the decoder or language model components of ASR that select the most likely sequence of sounds or words given the acoustic input.

The relevance of bottom-up information in the restoration process is evident in findings showing a stronger PRE for certain phone classes. Specifically, classes that are acoustically more similar to the replacing noise lead to more effective restoration of missing sounds. In contrast, the relevance of top-down information is demonstrated by findings showing a stronger PRE for words compared to phonotactically compliant pseudowords, as well as when words are primed (Samuel, 1981). Further support for top-down effects comes from studies showing a stronger PRE for items with multiple possible restorations (_egion → "legion"/"region" vs. _esion → "lesion"), and for items that become lexically unique earlier (e.g., uncommon first syllable "boysenberry" vs. common first syllable "indelible") (Samuel, 1987).

Samuel (1997) suggested that speech sounds perceived in English words as a result of phonemic restoration can even lead to a perceptual learning effect, i.e., a shift in the identification of stimuli on a /bɪ/–/dɪ/ continuum. However, this result

was later challenged by McQueen et al. (1999) who could not replicate the effect for Dutch words.

As the PRE provides both a measure of basic acoustic–phonetic performance and lexical impact (Samuel, 1981), it can help to disentangle these aspects. For instance, Mattys et al. (2014) employed phonemic restoration to explore scenarios where attention is divided between a visual task and speech input. They demonstrated a stronger PRE under higher cognitive load, implying a decrease in perceptual sensitivity. At the same time, the impact of lexical information (as assessed by words and pseudowords) remained stable under varying cognitive load.

Other work investigated phonemic restoration in non-native speakers. For instance, Ishida and Arai (2016) found that phonemic restoration occurred to a similar extent in native speakers of English and non-native speakers with first language Japanese — even for sounds that are not included in the phoneme inventory of Japanese. In contrast to native speakers, however, the PRE in non-native speakers was not influenced by lexical context, i.e., by whether the stimulus was a word or a pseudoword.

2.2. ASR and Perturbed Signals

One way to improve ASR performance in the presence of noise is to use front-end pre-processing techniques aimed at denoising the input signal. Additionally, integrating noise robustness into an ASR model by introducing realistic noisy signals during training can improve performance further, as it increases the likelihood that the testing and training environments closely match (Delcroix et al., 2013).

In an early work focusing on the differences in perturbed signal processing between humans and machines, Vaidya et al. (2015) introduced specific perturbations into the signal, which are considered random noise by humans, but can be interpreted as speech by an ASR. This type of perturbation is often referred to as adversarial noise. Vaidya et al. (2015) showed that such adversarial noise can be abused to activate voice assistant applications while being imperceptible by humans. Both adversarial and random noise introduce unique challenges and affect various models differently. Below, we provide examples of studies focusing on explaining noise robustness in state-of-the-art ASR models, either through post-hoc explainability methods or standard benchmarking.

Regarding the innate robustness of ASR models against noise, Gong et al. (2023) note: "It is commonly believed that the representation of a robust ASR model should be noise-*invariant*, and researchers often set noise-invariance as an explicit inductive bias for robust ASR." That is to say, models are expected to learn speech embeddings in a noisy environment, regardless of the type of noise. However, Gong et al.

(2023) point out that OpenAI's *Whisper* is not noise-invariant, as it also encodes the noise type and consequently speech recognition is conditioned on the noise type. They found that the representations (or embeddings) of Whisper are actually noise-*variant*, even in the later layers where higher contextualisation is expected, that is, where the model extracts more relevant information from the input context. They trained sound classifiers on representations of Whisper and found a positive correlation between the classifier performance (i.e., higher F1-score) for different background sounds and the robustness of Whisper to those background sounds (i.e., smaller increase in WER). However, some background sounds overlap with speech in such a way that Whisper is not robust to them, even though the classifier recognises them well, for example, the sounds of sheep, clapping, or a hand saw.

Olivier and Raj (2023) also explored Whisper's robustness to noise and demonstrated that adding adversarial noise at a signal-to-noise ratio (SNR) that is nearly imperceptible for humans (around 35-45 dB SNR) can degrade the performance of ASR substantially. They further showed that such adversarial perturbations affect smaller versions of Whisper more, with the large version being least affected (i.e., lowest WER).

Expanding on previous robustness assessments, Shah et al. (2024) evaluated various model sizes of Whisper and wav2vec 2.0. This was done using a speech robustness benchmark that includes, among other scenarios, environmental noises and adversarial perturbations. The two models outperformed each other in different scenarios (as measured by normalised WER degradation). On the one hand, Whisper large was more resilient in non-adversarial scenarios, where wav2vec 2.0 models were severely affected. On the other hand, wav2vec 2.0 large stood out as being least affected by adversarial perturbations, with Whisper models performing worse in these scenarios.

Overall, although ASR systems are often trained on noisy data, robustness to signal perturbations is not a given. Even the performance of a model such as Whisper, which is described as being more robust to noise than others (Patman and Chodroff, 2024; Gong et al., 2023), degrades with certain types of perturbations.

2.3. Domain-Informed Probing

The idea of using simple classifiers to explain black-box deep learning (DL) models was first proposed by Montavon et al. (2011). This approach was later expanded and popularised by Alain and Bengio (2017), who inserted simple linear classifiers on top of each model layer without affecting the trained weights. These classifiers, which are called probes, have the same output targets as the trained classifier in the final layer of the DL model. This type of probe can be used as an indicator of whether layer-wise transformation is progressive, and whether the intermediate layer embeddings

are already suitable for the final classification.

A domain-informed probe is similar to the linear classifiers described above, but with different output targets than the original trained model. Domain-informed probes are trained to measure the generalisability of the model embeddings in a task other than the originally trained task within the same domain. For instance, Belinkov and Glass (2017) trained supervised feed-forward classifiers on the frame-level embeddings of the *DeepSpeech2* model (Amodei et al., 2016) in order to measure their layer-wise capacity in encoding phonetic information rather than only recognising speech. The current work uses domain-informed probes with a simpler architecture (see Section 3.3) to reduce the effect of probe complexity on the evaluation of the embedding quality. In other words, the simplicity of the probes used here ensures that their performance is mostly reliant on the quality of the embeddings rather than the probing model complexity.

Probing ASR models also gained popularity in recent years due to advances in state-of-the-art ASR systems, which are predominantly black-box models. More recent development of probing applications in this domain include analysing self-supervised models for segmental units such as phones (English et al., 2022; ten Bosch et al., 2023), as well as articulatory features (English et al., 2023, 2024; Shams et al., 2024b) and phonological assimilation (Pouw et al., 2024). Other works investigate how syllables (Vitale et al., 2024; Shams et al., 2024a) or prosody (Yang et al., 2023; de la Fuente and Jurafsky, 2024) are encoded in the embeddings, or even explore speech assessment models for embedded acoustic characteristics (Lee et al., 2025).

One of the challenges of analysing DL-based ASR models is that information concerning one coherent speech unit (e.g., a speech sound) can be spread across multiple embedding frames. The frame length is relatively small and depends on the initial feature extractor or input processor of the model.[1] Hence, it is common practice (although not always necessary; de la Fuente and Jurafsky, 2024; Pouw et al., 2024) in domain-informed probing to pool a multi-frame speech unit — e.g., to transform a 3×1024 embedding for a given sound into a 1×1024 embedding. Different pooling (or aggregation) methods might be used depending on their suitability for the respective task. Mean pooling is the most common method (Prasad and Jyothi, 2020; Singla et al., 2022; Yang et al., 2023; Mohebbi et al., 2023; Shen et al., 2023). Other possible pooling methods include max pooling, weighted sum pooling (Xing et al., 2025), or self-attention pooling (Dhawan et al., 2024). The latter work suggests that mean pooling is less erratic than learned self-attention-based pooling. Furthermore,

[1]For example, in Whisper, every frame of the transformer layer embeddings is consistently equivalent to 320 samples or 20 ms of the input audio with a 16,000 Hz sampling rate.

Table 1: Examples of word/pseudoword pairs containing the target sounds at the onset of their final syllable.

Target	**Word**	**Pseudoword**
[m]	taxono**m**y	nastono**m**y
[n]	abdomi**n**al	gobflomi**n**al
[ɹ]	herbivo**r**ous	waarpivo**r**ous
[l]	molecu**l**ar	jotteku**l**ar

our previous work suggests that articulatory feature probes trained over mean pooled frames can generalise well to non-pooled frames (English et al., 2024).

In this work, we used domain-informed probing to take a look under the hood of wav2vec 2.0 and Whisper.

3. Materials and Methods

3.1. Stimuli

We used the test stimuli from Mattys et al. (2014), which contain 30 words and 30 pseudowords, each four to five syllables long. Pseudowords conform to the phonotactic rules of a language, but carry no meaning in that language, that is, they are not part of its lexicon. For both item types, the segment that will be manipulated, also called the target sound,[2] is either a nasal (13 × [m], 17 × [n]) or a liquid (15 × [ɹ], 15 × [l]) located at the onset of the final syllable.[3] Words and pseudowords come in pairs sharing the same stress pattern and the same final two syllables. Table 1 shows examples of such word/pseudoword pairs for each target sound. For an overview of all word/pseudoword pairs used in this study, please refer to Table A.1 in Appendix A. The stimuli were recorded at a sampling rate of 44.1 kHz with 16-bit depth by a female speaker of Standard Southern British English.

Mattys et al. (2014) provide the stimuli in three conditions. The first condition (C1) contains the clean audio recordings of words and pseudowords as reference. In the second condition (C2), the target sound was overlaid with signal-correlated noise of the same intensity (i.e., at 0 dB SNR). In the third condition (C3), the target sound was replaced by this same noise. The onset and offset of each target sound were identified through auditory analysis in order to minimise the occurrence

[2]Mattys et al. (2014) use *critical phoneme* to refer to the segments in question. We choose to use *target sound* instead, since the phonemic status is not central to the present study.

[3]In a few cases, the target sound is ambisyllabic between penultimate and final syllable.

of coarticulatory cues outside of the manipulation window. For the present study, we created two more stimulus conditions to test the impact of a silent gap on the encoding of articulatory features. Condition four (C4) uses the same manipulation window as C2 and C3 to replace the target sound by silence, while condition five (C5) uses the boundaries of the manually annotated target sound as the window for manipulation. In other words, C4 largely excludes coarticulatory cues in the remaining signal, while C5 retains these cues. In the following, we refer to C4 as "replaced by silence (wide)" and C5 as "replaced by silence (narrow)". For both conditions that were created for the present study, we made sure that the respective boundaries were located at zero crossings in order to avoid click artefacts. The stimulus conditions are illustrated in Figure 1.

3.2. ASR Models

We selected two state-of-the-art, publicly available transformer-based ASR models: Meta AI's *wav2vec 2.0* (Baevski et al., 2020) and OpenAI's *Whisper* (Radford et al., 2023). For wav2vec 2.0, we probed the *large-960h-lv60-self* variant, which was pre-trained on 60k hours of unlabelled data (Libri-Light dataset) using self-supervised learning and fine-tuned on 960 hours of English speech data (LibriSpeech 960h dataset).[4] For Whisper, we probed the *large-v3* variant, which was trained for ASR and language translation on 5M hours of weakly or pseudo-labelled multilingual data.[5] Throughout the paper, we refer to these models as wav2vec 2.0 and Whisper, respectively.

Apart from the training method and data, the two models also differ in their architecture — although both are based on transformers. While wav2vec 2.0 is designed in an encoder-only manner, Whisper is an encoder-decoder model. An encoder-only architecture is sufficient for the ASR task. However, Whisper's encoder-decoder architecture is relevant for its ability to translate speech of other languages into English text. This difference in architecture also introduces variability in the speech embeddings of the two models. Mohebbi et al. (2023), for example, showed that the lexical information required to identify French homophones is mostly located in the intermediate encoder layers for wav2vec 2.0, while it is traced in the cross-attention and decoder layers for Whisper. This may be relevant to our study, since lexical information relating to words and pseudowords may affect the encoder embeddings of the two models in different ways.

[4] `https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md`

[5] `https://github.com/openai/whisper/discussions/1762`

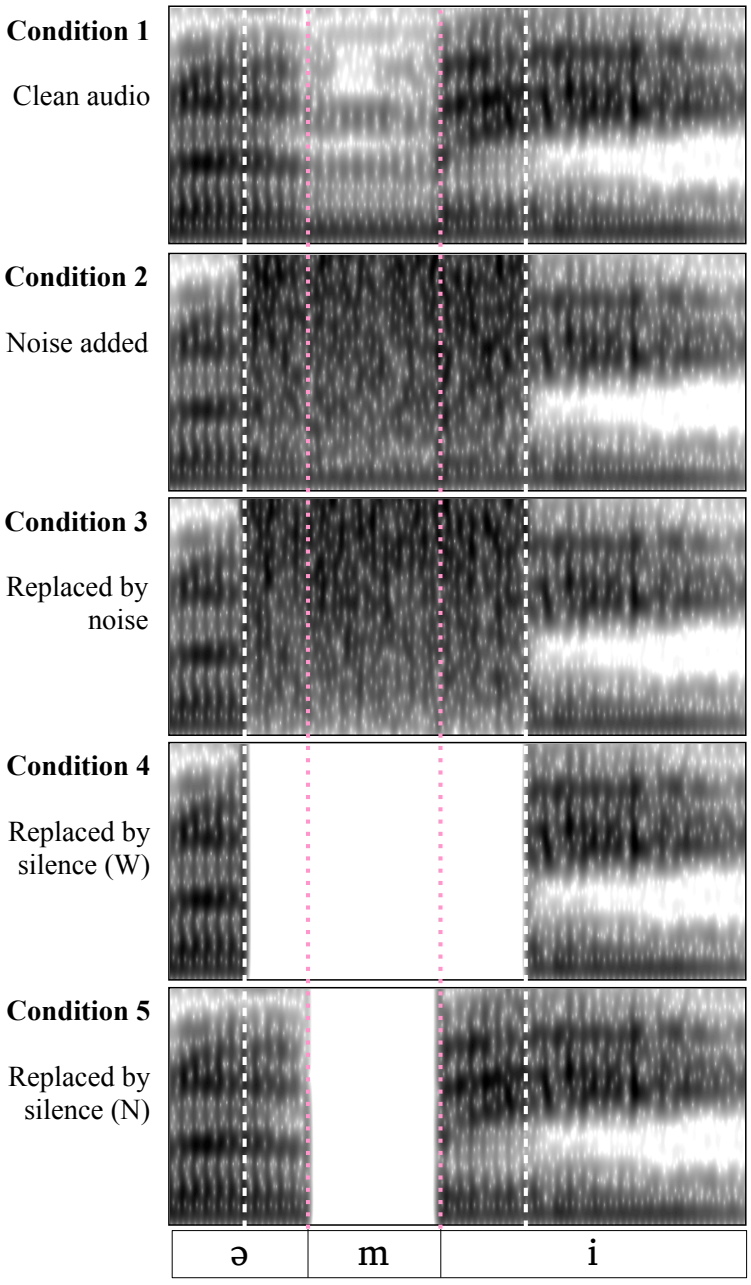


Figure 1: Stimulus conditions illustrated using the example word "taxon**omy**" [taksɒnə**mi**] with the target sound being [m]. In conditions 2, 3, and 4 (W = wide), the manipulation extends beyond the manually annotated [m] to exclude coarticulatory cues (white long-dashed lines), whereas in condition 5 (N = narrow), it remains within these boundaries (pink short-dashed lines).

However, since we focus on the embeddings representing speech segments rather than the word/subword segments, it is impractical to apply our method to the decoder layers of Whisper. The decoder utilises word tokens (i.e., text) instead of speech segments, which means that its embeddings do not have the same granularity as the audio-level segments we aim to investigate. Although Whisper's cross-attention heads have both audio and word token dimensions, these are not transformed into speech embeddings, which means they would require further processing before probing them in the same way we probe encoder embeddings. This would lead to an inconsistent comparison, which we sought to avoid.

In order to compare the contextual representation learning mechanisms of the two models, we only considered the transformer layers of the encoder block in both cases. We excluded the initial convolutional layers as well as all layers following the encoder.

To explore the effect of signal perturbation throughout the models, we grouped their transformer layers into early, middle, and late regions. For wav2vec 2.0, these regions correspond to layers 1–8 (early), 9–16 (middle), and 17–24 (late), while for Whisper they correspond to layers 1–11 (early), 12–21 (middle), and 22–32 (late).

3.3. Domain-Informed Probing Task

The stimuli were processed by the selected ASR models and the speech embeddings generated by the models were extracted at each transformer layer. The extracted embeddings of a target sound were mean pooled and then separately probed for encoded information about POA, MOA, and VOI features using the domain-informed probing method presented in English et al. (2024). The probes are simple multilayer perceptrons trained to detect these articulatory features. For every pooled embedding, the POA-probe outputs a probability distribution across the 11 primary POAs, as defined by the classification of the International Phonetic Alphabet (IPA). The MOA-probe does the same for the 8 primary MOAs, and the VOI-probe for *voiceless* and *voiced*. We identified the POA, MOA, and VOI features of the target sound embedding based on the highest probability of each probe. The predicted articulatory features were then classified as either *matching* or *not matching* the target, which will serve as binary dependent variable for the subsequent statistical analysis. Figure 2 illustrates this process.

3.4. Hypotheses

In RQ1, we test the effect of different segment-level perturbations. We expect the selected ASR models to encode articulatory features relatively well in stimulus condition C2, since the full audio signal is still present and the models encounter

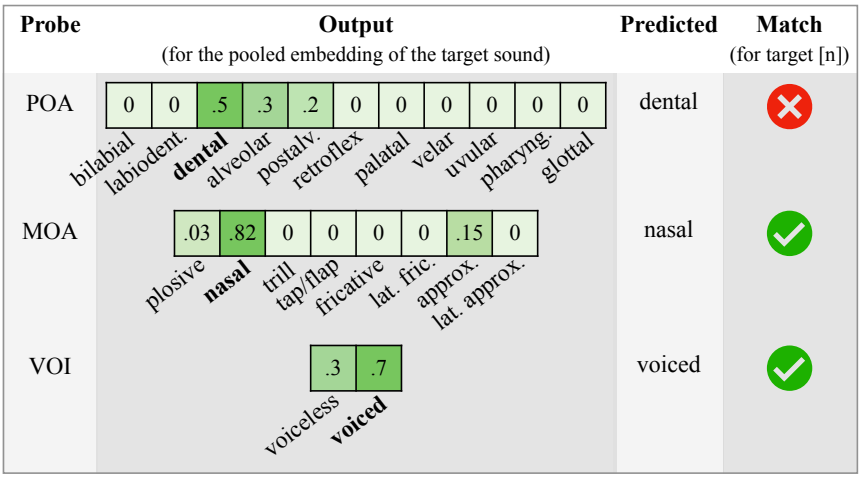


Figure 2: The probability distributions output by each of the three probes for the pooled embedding of the target sound. The predicted articulatory feature either matches or does not match the target, here [n].

background noise in their training data. While the noisy gap in C3 would allow for phonemic restoration in human perception, the silent gap in C4 would inhibit it (Warren and Obusek, 1971). However, it is not clear whether signal continuity plays a similar role for ASR processing. This aspect also relates to RQ5, which compares ASR model performance and human speech perception. Furthermore, condition C5 is expected to perform better than C4, since more coarticulatory cues are available and prior work suggests that ASR models pick up on coarticulatory cues (English et al., 2024; Shams et al., 2024b; Zellou et al., 2024).

In RQ2, we test the effect of lexical status, that is, whether the stimulus is a word or a pseudoword. We can assume that, similar to humans (Samuel, 1981), ASR models also make use of lexical context to restore perturbed segments. Therefore, we expect articulatory features to be classified more successfully in words than in pseudowords. This aspect also relates to RQ5.

In RQ3, we test the effect of different target sounds. The selected target sounds, two nasals and two liquids, fall in the middle range of restorability for human perception, with fricatives being most successfully and vowels least successfully restored when replaced by noise (Samuel, 1981). There are no differences between the four sounds that would lead us to expect ASR processing to be particularly affected. Since replacing speech sounds by noise or silence leads to an absence of a voice component in the signal, classification accuracy of VOI would be affected in different ways for voiced and voiceless sounds by these perturbations. However, all four target sounds

investigated in this study are voiced ([m], [n], [ɹ], [l]) and VOI classification should therefore be similarly affected for all of them.

Relating back to RQ1, we assume that replacing sounds by noise may lead to MOA confusions with fricatives, while replacing them by silence may lead to MOA confusions with plosives. However, for RQ3, with all target sounds being sonorants (nasals [m] and [n], approximant [ɹ], and lateral approximant [l]), this should affect classification accuracy of MOA similarly for all of them. In terms of POA classification accuracy, we do not expect the investigated perturbations to have systematically different effects for bilabial ([m]) and alveolar ([n], [ɹ], [l]) targets.

In RQ4, we test the effect of the three layer regions early, middle, and late. Prior work suggests that earlier transformer layers have higher similarity to spectrogram features (Pasad et al., 2023), while some self-attention heads in these layers already focus on specific articulatory information such as distinguishing alveolar and postalveolar fricatives (Shams and Carson-Berndsen, 2024). Middle to higher layers then encode information more relevant to identifying phonetic and sub-phonetic features (ten Bosch et al., 2023; English et al., 2024). We therefore expect that the perturbed signal will be partially recovered in the lower layers and then fully recovered, if possible, in the higher layers.

3.5. Statistical Analysis

To evaluate the significance and robustness of our findings, we conducted a statistical analysis appropriate for our data structure and research questions. The three binary dependent variables discussed above — *same POA* as target vs. *other POA* than target, *same MOA* as target vs. *other MOA* than target, and *voiced* (since all targets are voiced) vs. *voiceless* — were analysed using separate generalised linear mixed-effects models (GLMMs). The models were formulated with the *lme4* package (Bates et al., 2015) and evaluated with the *lmerTest* (Kuznetsova et al., 2017), *ggeffects* (Lüdecke, 2018), and *emmeans* (Lenth, 2025) packages in *RStudio* (Posit Team, 2024) with *R* (R Core Team, 2024).

Model selection was carried out bottom-up, starting with a model which only included the random factor intercepts for ITEM. Here, ITEMs refer to the individual words and pseudowords used as stimuli in the experiment. Then, the fixed factor MODEL (sum coded levels: *wav2vec 2.0, Whisper*) was added to the model — with random slopes for MODEL by ITEM. Subsequently, the following four fixed factors as well as their interaction with MODEL were added: STIMULUS CONDITION (treatment coded levels: *clean audio, noise added, replaced by noise, replaced by silence (wide), replaced by silence (narrow)*), ITEM TYPE (treatment coded levels: *word, pseudoword*), TARGET SOUND (sum coded levels: *[m], [n], [ɹ], [l]*), and LAYER RE-

GION (treatment coded levels: *early, middle, late*). Treatment coding was applied where one factor level represents a natural reference. This was the case for *clean audio*, *word*, and the *early* layer region. Sum coding, on the other hand, compares all levels to the grand mean where no natural reference exists.

A factor was kept in the model if the model fit improved significantly, as determined by a likelihood-ratio test, and the Akaike information criterion value (Akaike, 1998) decreased by at least two points as compared to the model without the factor in question. Factors kept in the model are considered significant predictors of the respective dependent variable at $\alpha = 0.05$.

The final GLMMs for all three dependent variables had the following structure:

$$\begin{cases} \text{same POA} \\ \text{same MOA} \\ \text{voiced} \end{cases} \sim \text{MODEL} * \text{STIMULUS CONDITION} + \text{MODEL} * \text{ITEM TYPE} + \text{MODEL} * \text{TARGET SOUND} + \text{MODEL} * \text{LAYER REGION} + (1 + \text{MODEL} \mid \text{ITEM})$$

To further explore differences between factor levels, we computed estimated marginal means and conducted pairwise comparisons. These comparisons accounted for other terms in the model, enabling a detailed examination of differences for each level of the interacting factors. Adjustments for multiple comparisons were applied using the Tukey method.

4. Results

The results are presented in two parts: first, we provide a visual inspection of the data to give an initial overview of existing patterns; second, we report the results of the statistical analysis which formally tests the observed effects.

For the visual inspection, we summarise the percentages with which the target sounds were correctly classified in early, middle, and late transformer layers, focusing on POA in Figure 3, on MOA in Figure 4, and on VOI in Figure 5. The results are grouped by item type and stimulus condition. The colours green, blue, and orange indicate a correctly identified articulatory feature, while grey denotes that the identified articulatory feature does not match the target sound. The illustrations suggest a strong impact of stimulus condition (RQ1) and layer region (RQ3) on the performance of both models across all three articulatory features. An impact of item type (RQ2) is not as clearly identifiable by visual inspection. The results of the statistical analysis for these three factors as well as for the impact of the target

sound itself (RQ4) are presented below. For a more detailed, layer-wise illustration, please refer to Figures B.1 and B.2 in Appendix B.

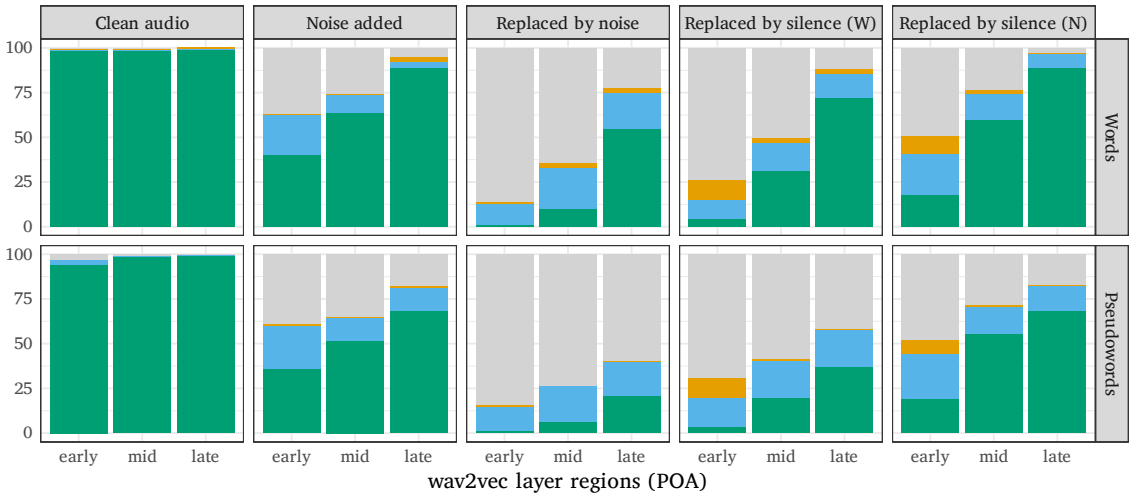
Although not directly related to the primary RQs of this study, the colour coding in Figures 3 to 5 offers additional insight into how successfully the three independent probes extracted all relevant articulatory features to characterise the target sound. Green highlights cases where the other two articulatory features were also correctly classified by the respective probing models. Blue and orange denote cases where either or both of these are incorrectly classified. This can either lead to sounds that are attested (blue) in the classification of the IPA (e.g., [d] for target [n] → POA and VOI match) or sounds that are unattested (orange) in this classification (e.g., [n0] with 0 indicating voiceless for target [n] → POA and MOA match). The concept of attested and unattested sounds is illustrated by Figure C.1 in Appendix C. This approach to visualising the results makes it possible to focus on one articulatory feature while still maintaining a holistic view of the three probing models.

To evaluate the significance of the observed patterns, we conducted statistical tests according to the method detailed in Section 3.5. The analysis results are presented in one big table (Table 6; at the end of this section on p. 25) and four smaller tables (Tables 2-5). Table 6 shows the output of the GLMMs provided above, while Tables 2-5 show the pairwise comparisons for each of the four factors under investigation (ITEM TYPE, STIMULUS CONDITION, LAYER REGION, TARGET SOUND) and their interaction with the two ASR systems (MODEL).

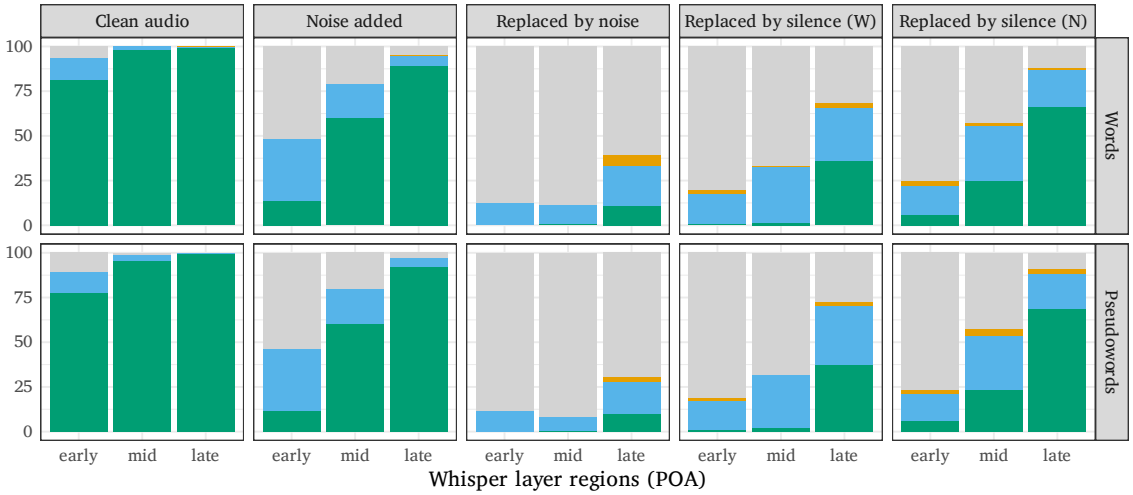
In Table 6, MODEL has a statistically significant influence on both POA and MOA classification, with target matches being more frequent for wav2vec 2.0 compared to Whisper,[6] Furthermore, STIMULUS CONDITION has a statistically significant effect on all three dependent variables, with target matches being least frequent when the target sound is replaced by noise, followed by the replaced by silence (wide) and (narrow) conditions, and classification being least affected by the addition of noise in all cases.[7] Addressing RQ1, which asked how the articulatory feature encoding in the embeddings of the two models is affected by different types of segment-level perturbations, the interaction between MODEL and STIMULUS CONDITION indicates that the effect of adding noise on both POA and MOA, as well as the effect of

[6]Since MODEL is sum coded, the estimate reflects the deviation of *wav2vec 2.0* from the grand mean, with the effect of *Whisper* being the same size but in the opposite direction — similarly for TARGET SOUND.

[7]Since STIMULUS CONDITION is treatment coded, the estimate reflects the effect of each factor level relative to the reference level *clear audio*, which is coded as the baseline — similarly for ITEM TYPE and LAYER REGION.

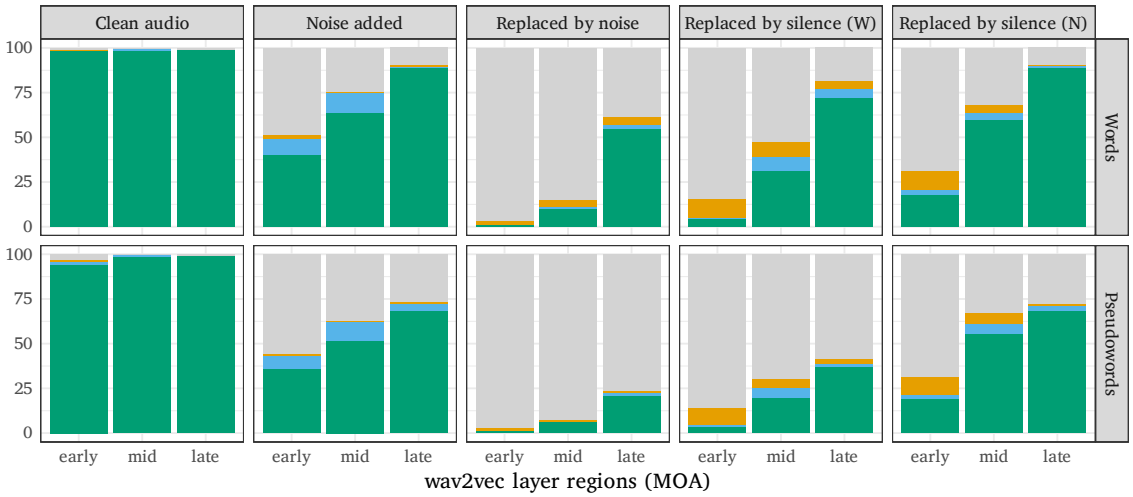


(a) wav2vec 2.0: early (1–8), mid (9–16), and late (17–24) transformer layers.

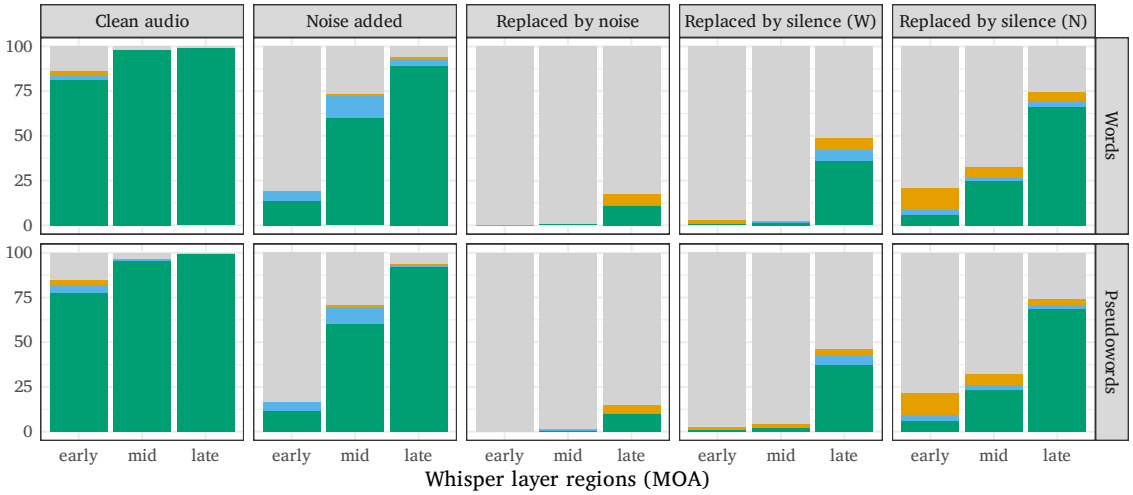


(b) Whisper: early (1–11), mid (12–21), and late (22–32) transformer layers.

Figure 3: Percentage target sounds correctly classified with focus on place of articulation (POA): target full match ■, same POA (attested) ■, same POA (unattested) ■, or different POA ■.

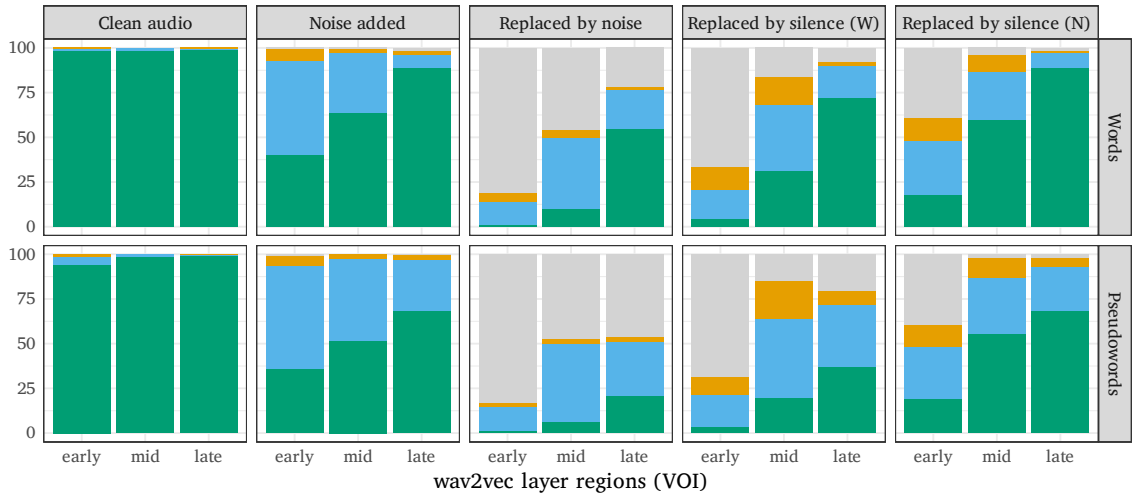


(a) wav2vec 2.0: early (1–8), mid (9–16), and late (17–24) transformer layers.

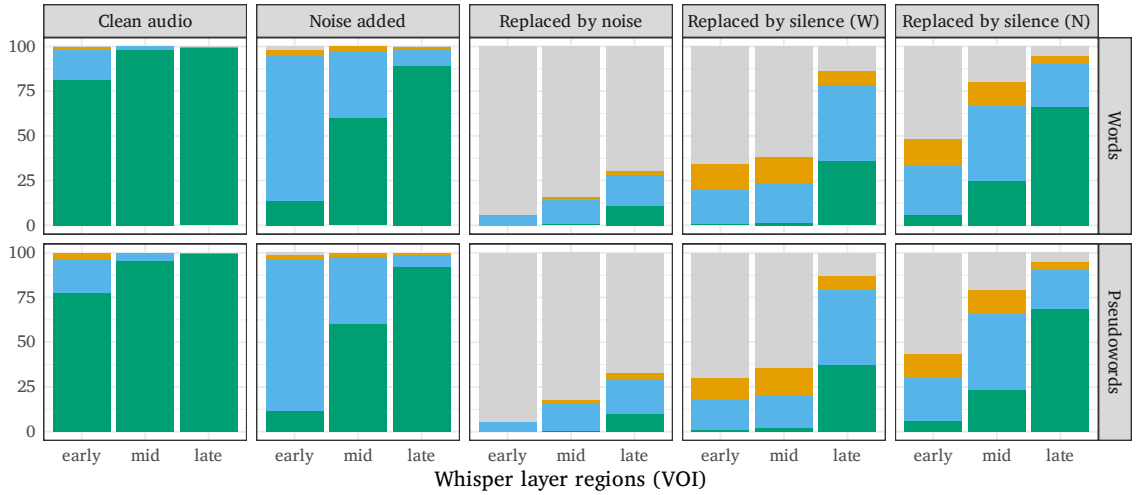


(b) Whisper: early (1–11), mid (12–21), and late (22–32) transformer layers.

Figure 4: Percentage target sounds correctly classified with focus on manner of articulation (MOA): target full match ■, same MOA (attested) ■, same MOA (unattested) ■, or different MOA ■.



(a) wav2vec 2.0: early (1–8), mid (9–16), and late (17–24) transformer layers.



(b) Whisper: early (1–11), mid (12–21), and late (22–32) transformer layers.

Figure 5: Percentage target sounds correctly classified with focus on voicing (VOI): target full match ■, voiced (attested) ■, voiced (unattested) ■, or voiceless ■.

replacing by silence (wide) on POA, are more pronounced for wav2vec 2.0. At the same time, the effects of replacing by noise and by silence (wide) on MOA are more pronounced for Whisper.

Table 2 summarises the related pairwise comparisons for the interaction of MODEL and STIMULUS CONDITION. The results show that wav2vec 2.0 and Whisper do not differ in terms of VOI classification in the clean audio condition. Further, they perform the same at POA and MOA classification when noise is added to the speech signal. In all other conditions, wav2vec 2.0 outperforms Whisper. When comparing conditions within a system, only adding noise and replacing the target sound with silence (narrow) does not result in a performance difference for wav2vec 2.0. All other conditions lead to different performances for both systems. For both systems, adding noise has the smallest negative impact over clean audio, followed by replacing the speech segment by silence (narrow), that is, leaving coarticulatory information in the signal. Replacing the speech segment by silence (wide), that is, removing coarticulatory cues, impedes classification relative to clean audio further, and replacing the speech segment with noise has the biggest impact on classification. Crucially, replacing the speech segment by noise leads to a statistically significantly worse classification than replacing it by silence (wide). An illustration of the interaction based on estimated marginal means is provided by Figure D.1 in Appendix D.

Returning to Table 6, ITEM TYPE has a statistically significant influence on both POA and MOA classification, with target matches being less frequent for pseudowords compared to words. In relation to RQ2, which asked how the articulatory feature encoding in the embeddings of the two models is affected by the stimulus being a word or a pseudoword, the interaction between MODEL and ITEM TYPE indicates that the effect of pseudowords is more pronounced for wav2vec 2.0. The classification of VOI is not statistically significantly influenced by MODEL or ITEM TYPE.

Table 3 summarises the related pairwise comparisons for the interaction between MODEL and ITEM TYPE. The results show that wav2vec 2.0 recognises POA and MOA significantly better in words than pseudowords, while ITEM TYPE does not influence the performance of Whisper. In the case of VOI, neither of the two systems is affected by ITEM TYPE. An illustration of the interaction based on estimated marginal means is provided by Figure D.2 in Appendix D.

Returning again to Table 6, TARGET SOUND has a statistically significant effect on all three dependent variables, with [n] leading to more target matches than the average across all sounds, [m] falling below this grand mean (not statistically significant for MOA), and [ɹ] even further so. Regarding RQ3, which asked how the articulatory feature encoding in the embeddings of the two models is affected by the

Table 2: Pairwise comparisons for the interaction of MODEL and STIMULUS CONDITION (CA = clean audio, NA = noise added, RN = replaced by noise, SW = replaced by silence (wide), SN = replaced by silence (narrow)), reported as log Odds (lgO) with standard error (SE), z-statistic, and p-value. Cases that are not statistically significant at $\alpha = 0.05$ are highlighted in blue.

	POA				MOA				VOI			
	lgO	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**
wav2vec 2.0–Whisper												
CA	1.29	.21	5.88	<.001	1.16	.22	5.18	<.001	1.90	1.07	1.76	.077
NA	−.06	.09	−.64	.517	.09	.11	.83	.404	.63	.31	2.01	.044
RN	1.04	.09	10.97	<.001	2.63	.14	18.00	<.001	1.67	.10	16.04	<.001
SW	.42	.08	4.76	<.001	2.11	.12	17.62	<.001	.91	.09	9.13	<.001
SN	.90	.09	9.91	<.001	1.15	.11	10.55	<.001	1.24	.11	10.99	<.001
wav2vec 2.0												
CA–NA	4.13	.18	21.84	<.001	4.82	.18	25.45	<.001	3.16	.97	3.23	.010
CA–RN	6.53	.19	33.70	<.001	8.08	.20	39.95	<.001	9.91	.95	10.36	<.001
CA–SW	5.65	.19	29.60	<.001	6.61	.19	33.92	<.001	8.47	.95	8.87	<.001
CA–SN	4.23	.18	22.37	<.001	5.24	.19	27.49	<.001	6.91	.95	7.23	<.001
NA–RN	2.39	.07	32.67	<.001	3.26	.08	39.16	<.001	6.75	.23	28.45	<.001
NA–SW	1.52	.06	21.99	<.001	1.79	.07	25.06	<.001	5.31	.23	22.85	<.001
NA–SN	.10	.07	1.44	.602	.41	.06	6.03	<.001	3.75	.23	16.19	<.001
RN–SW	−.87	.06	−13.15	<.001	−1.47	.07	−19.46	<.001	−1.44	.07	−20.04	<.001
RN–SN	−2.29	.07	−31.63	<.001	−2.84	.08	−35.39	<.001	−3.00	.08	−33.56	<.001
SW–SN	−1.42	.06	−20.74	<.001	−1.37	.06	−19.90	<.001	−1.56	.08	−18.71	<.001
Whisper												
CA–NA	2.78	.10	26.38	<.001	3.75	.10	37.15	<.001	1.89	.53	3.54	.003
CA–RN	6.28	.11	54.45	<.001	9.55	.15	60.31	<.001	9.67	.49	19.40	<.001
CA–SW	4.79	.10	44.26	<.001	7.56	.13	57.21	<.001	7.47	.49	15.06	<.001
CA–SN	3.85	.10	36.39	<.001	5.23	.11	47.42	<.001	6.25	.49	12.60	<.001
NA–RN	3.50	.07	49.38	<.001	5.80	.11	49.27	<.001	7.79	.20	37.09	<.001
NA–SW	2.00	.06	33.01	<.001	3.81	.08	43.27	<.001	5.59	.20	27.43	<.001
NA–SN	1.06	.05	18.31	<.001	1.47	.06	21.49	<.001	4.36	.20	21.49	<.001
RN–SW	−1.49	.06	−23.72	<.001	−1.99	.10	−19.77	<.001	−2.20	.06	−33.51	<.001
RN–SN	−2.43	.06	−37.24	<.001	−4.32	.10	−40.44	<.001	−3.42	.07	−47.25	<.001
SW–SN	−.94	.05	−16.69	<.001	−2.33	.07	−29.66	<.001	−1.23	.05	−21.36	<.001

Table 3: Pairwise comparisons for the interaction of MODEL and ITEM TYPE, reported as log Odds (lgO) with standard error (SE), z-statistic, and p-value. Cases that are not statistically significant at $\alpha = 0.05$ are highlighted in blue.

	POA				MOA				VOI			
	lgO	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**
wav2vec 2.0–Whisper												
Words	.99	.10	9.10	<.001	1.85	.13	13.50	<.001	1.41	.25	5.55	<.001
Pseudowords	.44	.10	4.15	<.001	1.02	.13	7.46	<.001	1.13	.25	4.48	<.001
Words–Pseudowords												
wav2vec 2.0	.61	.16	3.67	<.001	.92	.20	4.56	<.001	.32	.17	1.81	.069
Whisper	.07	.12	.58	.556	.09	.16	.58	.561	.04	.12	.38	.698

target sound, the interaction between MODEL and TARGET SOUND indicates that the effect is more pronounced for wav2vec 2.0 in the case of POA [m] and [n], MOA [m], and VOI [n], while it is stronger for Whisper in the case of MOA [n] and [ɹ].

Table 4 summarises the related pairwise comparisons for the interaction between MODEL and TARGET SOUND. For all sounds and features, wav2vec 2.0 has more target matches than Whisper, except for POA [m], where the performance of the two systems is equal. In both systems, [n] stands out, as POA and MOA are correctly classified more often than for the other three sounds tested, and VOI is correctly classified more often than for [l] and [ɹ]. Both systems also recognise voicing correctly more often in [l] than in [ɹ]. Other statistically significant differences between target sounds are less systematic, with wav2vec 2.0 embeddings leading to better classification of POA and VOI for [l] than for [m], and Whisper embeddings leading to better classification of POA for both [m] and [l] compared to [ɹ], and MOA for [m] compared to both [l] and [ɹ]. An illustration of the interaction based on estimated marginal means is provided by Figure D.3 in Appendix D.

Finally, returning to Table 6 once again, the statistically significant effect of LAYER REGION indicates that for all three dependent variables, the number of target matches increases from early to middle layers, and even more so from early to late layers. Referring to RQ4, which asked how the articulatory feature encoding in the embeddings of the two models is affected by the layer region, the interaction between MODEL and LAYER REGION suggests that, for POA and MOA, this effect is more pronounced in the case of Whisper, while for VOI, wav2vec 2.0 is more impacted by LAYER REGION.

Table 5 summarises the related pairwise comparisons for the interaction between

Table 4: Pairwise comparisons for the interaction of MODEL and TARGET SOUND, reported as log Odds (lgO) with standard error (SE), z-statistic, and p-value. Cases that are not statistically significant at $\alpha = 0.05$ are highlighted in blue.

	POA				MOA				VOI			
	lgO	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**
wav2vec 2.0–Whisper												
[m]	−.05	.15	−.32	.742	.47	.20	2.33	.019	.97	.28	3.39	<.001
[n]	1.33	.14	9.24	<.001	.58	.17	3.27	.001	1.66	.27	5.97	<.001
[ɹ]	.81	.14	5.55	<.001	2.10	.19	11.02	<.001	1.04	.27	3.73	<.001
[l]	.78	.14	5.31	<.001	2.57	.19	13.46	<.001	1.41	.28	5.03	<.001
wav2vec 2.0												
[m]–[n]	−2.09	.24	−8.69	<.001	−1.69	.29	−5.84	<.001	−1.25	.25	−4.89	<.001
[m]–[ɹ]	−.25	.24	−1.02	.735	−.27	.29	−.91	.796	.02	.26	.08	.999
[m]–[l]	−.72	.24	−2.94	.016	−.59	.29	−1.99	.189	−.85	.26	−3.25	.006
[n]–[ɹ]	1.84	.23	7.96	<.001	1.42	.27	5.10	<.001	1.27	.24	5.17	<.001
[n]–[l]	1.37	.23	5.93	<.001	1.10	.27	3.95	<.001	.39	.24	1.61	.370
[ɹ]–[l]	−.47	.23	−2.00	.187	−.32	.28	−1.12	.676	−.87	.25	−3.46	.003
Whisper												
[m]–[n]	−.70	.17	−4.12	<.001	−1.58	.22	−6.93	<.001	−.56	.17	−3.20	.007
[m]–[ɹ]	.61	.17	3.49	.002	1.35	.23	5.75	<.001	.08	.18	.49	.960
[m]–[l]	.11	.17	.64	.917	1.50	.23	6.37	<.001	−.40	.18	−2.24	.110
[n]–[ɹ]	1.32	.16	8.00	<.001	2.93	.22	13.19	<.001	.65	.17	3.85	<.001
[n]–[l]	.82	.16	4.97	<.001	3.08	.22	13.82	<.001	.15	.17	.92	.790
[ɹ]–[l]	−.50	.17	−2.96	.016	.14	.22	.65	.914	−.49	.17	−2.84	.023

Table 5: Pairwise comparisons for the interaction of MODEL and LAYER REGION, reported as log Odds (lgO) with standard error (SE), z-statistic, and p-value.

	POA				MOA				VOI			
	lgO	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**
wav2vec 2.0–Whisper												
early	.98	.09	10.95	<.001	2.22	.11	18.73	<.001	.63	.23	2.66	.007
mid	.68	.09	7.44	<.001	1.73	.11	15.36	<.001	2.23	.24	9.03	<.001
late	.49	.10	4.91	<.001	.34	.11	2.95	.003	.94	.25	3.76	<.001
wav2vec 2.0												
ea–mi	−.94	.05	−16.72	<.001	−1.54	.06	−24.49	<.001	−2.65	.07	−33.44	<.001
ea–la	−2.47	.06	−37.51	<.001	−2.87	.07	−40.63	<.001	−3.07	.08	−36.36	<.001
mi–la	−1.52	.06	−24.48	<.001	−1.33	.06	−21.54	<.001	−.42	.07	−5.36	<.001
Whisper												
ea–mi	−1.25	.05	−24.68	<.001	−2.03	.07	−28.50	<.001	−1.05	.05	−17.89	<.001
ea–la	−2.96	.05	−51.16	<.001	−4.75	.08	−52.95	<.001	−2.76	.06	−40.93	<.001
mi–la	−1.71	.05	−32.37	<.001	−2.72	.07	−38.27	<.001	−1.71	.06	−27.00	<.001

MODEL and LAYER REGION. As expected, the number of correct classifications increases statistically significantly for both systems on all three features from early to middle to late transformer layers. Overall, wav2vec 2.0 always outperforms Whisper in this comparison. While for POA and MOA the gap between systems becomes increasingly smaller from early to late layers, the difference is highest in the middle layers for VOI and remains higher in late than in early layers. An illustration of the interaction based on estimated marginal means is provided by Figure D.4 in Appendix D.

Probes usually perform better in some layers than in others. In the present case of three independent probes, the best-performing layers are defined as those with the highest average probe performance for POA, MOA, and VOI. In an analysis across the full set of English phonemes, we identified layer 16 as best-performing for wav2vec 2.0, and layer 26 for Whisper.[8] In order to evaluate whether certain stimulus conditions lead to systematic target confusions, Tables E.1 to E.4 in Appendix E provide an overview of the classified sounds for each target sound in the best-performing lay-

[8]Note that the differences to surrounding layers are only minimal and the pattern observed for the present phoneme subset /m/, /n/, /r/, and /l/ is largely consistent with the probe results for the full phoneme set.

ers for both models. Across words and pseudowords, most confusions in C3 (replaced by noise) occurred with fricatives (w2v: 74 %, Whisp: 56 %), followed by plosives (w2v: 20 %, Whisp: 31 %), while most confusions in C4 (replaced by silence (wide)) occurred with plosives (w2v: 31 %, Whisp: 41 %), closely followed by nasals (w2v: 22 %, Whisp: 37 %).

Table 6: GLMM results for the three dependent variables, reported as log Odds (lgO) with standard error (SE), z-statistic, and p-value. Upper: Main effects of MODEL (WV = wav2vec 2.0), STIMULUS CONDITION (NA = noise added, RN = replaced by noise, SW = replaced by silence (wide), SN = replaced by silence (narrow)), ITEM TYPE (pseu = pseudoword), TARGET SOUND, and LAYER REGION (mi = middle, la = late). Lower: Interactions of MODEL with the four other fixed factors. Cases that are not statistically significant at $\alpha = 0.05$ are highlighted in blue.

	POA				MOA				VOI			
Predictor	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**	**lgO**	**SE**	**z**	**p**
(Intercept)	3.79	.14	27.70	<.001	3.59	.15	24.09	<.001	7.05	.54	12.97	<.001
Model[WV]	.92	.11	7.97	<.001	1.19	.12	9.98	<.001	.70	.54	1.30	.192
Cond[NA]	−3.46	.11	−31.90	<.001	−4.29	.11	−39.92	<.001	−2.52	.56	−4.54	<.001
Cond[RN]	−6.41	.11	−56.79	<.001	−8.82	.13	−68.64	<.001	−9.79	.54	−18.17	<.001
Cond[SW]	−5.22	.11	−47.54	<.001	−7.09	.12	−60.17	<.001	−7.97	.54	−14.82	<.001
Cond[SN]	−4.04	.11	−37.26	<.001	−5.24	.11	−47.55	<.001	−6.58	.54	−12.24	<.001
Type[pseu]	−.34	.13	−2.70	.007	−.51	.16	−3.24	.001	−.19	.13	−1.45	.146
Sound[m]	−.38	.12	−3.30	.001	−.16	.14	−1.13	.259	−.37	.12	−3.19	.001
Sound[n]	1.02	.11	9.60	<.001	1.48	.13	11.23	<.001	.54	.11	5.02	<.001
Sound[ɹ]	−.56	.11	−5.12	<.001	−.70	.14	−5.14	<.001	−.43	.11	−3.84	<.001
Region[mi]	1.10	.04	28.93	<.001	1.79	.05	37.57	<.001	1.85	.05	37.53	<.001
Region[la]	2.72	.04	61.96	<.001	3.81	.06	66.74	<.001	2.91	.05	53.96	<.001
M:C[NA]	−.68	.11	−6.25	<.001	−.54	.11	−5.00	<.001	−.64	.56	−1.14	.253
M:C[RN]	−.12	.11	−1.07	.283	.73	.13	5.71	<.001	−.12	.54	−.22	.829
M:C[SW]	−.43	.11	−3.96	<.001	.47	.12	4.02	<.001	−.50	.54	−.92	.357
M:C[SN]	−.19	.11	−1.77	.076	.00	.11	−.04	.966	−.33	.54	−.61	.541
M:T[pseu]	−.27	.07	−3.80	<.001	−.42	.09	−4.49	<.001	−.14	.09	−1.63	.104
M:S[m]	−.39	.07	−5.93	<.001	−.48	.08	−5.67	<.001	−.15	.08	−1.95	.051
M:S[n]	.31	.06	5.06	<.001	−.42	.08	−5.38	<.001	.19	.07	2.71	.007
M:S[ɹ]	.05	.06	.76	.448	.33	.08	4.12	<.001	−.12	.07	−1.58	.115
M:R[mi]	−.15	.04	−4.02	<.001	−.25	.05	−5.16	<.001	.80	.05	16.22	<.001
M:R[la]	−.25	.04	−5.63	<.001	−.94	.06	−16.47	<.001	.15	.05	2.87	.004

5. Discussion

In this work, we assessed how two state-of-the-art automatic speech recognition (ASR) models, wav2vec 2.0 and Whisper, process segment-level perturbations throughout their encoder transformer layers. Through domain-informed probing, we investigated whether information about articulatory features — place of articulation (POA), manner of articulation (MOA), and voicing (VOI) — was detectable in the model embeddings. We directly compared the performance of the two selected models in early, middle, and late layers when different perturbations interfered with individual target sounds of words or pseudowords. The theoretical background and inspiration for this study were taken from human speech perception, where the Phonemic Restoration Effect (PRE) predicts how humans perceptually cope with perturbed speech segments (Warren, 1970; Warren and Obusek, 1971; Samuel, 1981, 1987).

It is worth noting that the performance of the three probes detecting POA, MOA, and VOI of the target sounds investigated in this study ([m], [n], [ɹ], and [l]) was very good throughout all individual layers in the clean audio reference condition. This is especially true for all three articulatory features in the case of wav2vec 2.0, while in the case of Whisper early layers still have difficulties to correctly identify POA and MOA. This demonstrates that the probes are able to successfully detect the articulatory features explored in this study under favourable conditions, which provides an excellent basis for investigating different perturbed conditions.

Overall, we found that wav2vec 2.0 encoded the articulatory features of the target sounds more clearly — i.e., more recognisably for the three probes — than Whisper across all conditions. This difference may stem from the distinct training methods and objectives. Wav2vec 2.0 *large-960h-lv60-self* was initially pre-trained to identify masked embedding time steps (self-supervised), then fine-tuned on orthographically transcribed English speech data, which associates acoustic patterns with text (supervised). This may lead the model to implicitly learn about phonetic patterns (Pasad et al., 2021). In contrast, Whisper *large-v3* was trained on a large, diverse transcribed dataset (supervised) across multiple languages and tasks, focusing on higher-level context, such as word sequences and semantic content. In the following, we will discuss the results pertaining to the five research questions (RQs) that guided this study.

5.1. RQ1: Effect of segment-level perturbations on articulatory feature encoding

First, we investigated how segment-level perturbations affect the encoding of articulatory features in the ASR model embeddings. Overall, the different perturbation types (added noise, noisy gaps, and two types of silent gaps) tended to affect both

ASR systems in similar ways. Compared to human speech perception, the effect of noisy versus silent gaps was reversed.

Adding noise (C2) to the reference audio had the smallest negative impact on classification among all segment-level perturbations investigated here. This was in keeping with our prediction, as the full audio signal was still present. However, considering that the ASR models are exposed to background noise during training, which is aimed at improving their performance when faced with noisy signals (Delcroix et al., 2013), it is noteworthy that the deterioration was statistically significant for all three articulatory features in both systems. As Delcroix et al. (2013) point out, the performance of DL-based ASR models depends on a good match between training and testing conditions — also in terms of noise characteristics — and still benefits from speech enhancement pre-processing. Even the state-of-the-art model Whisper, which performs very well under adverse conditions (Patman and Chodroff, 2024), is not noise-invariant and consequently affected by noise (Gong et al., 2023). Our result for C2 supports this notion for both wav2vec 2.0 and Whisper.

While phonemic restoration in human perception depends on the continuity of the signal and works where a sound is replaced by noise (C3), but not where a silent gap disrupts the signal (C4, C5; Warren, 1970; Warren and Obusek, 1971), this does not seem to be the case for the two ASR systems investigated here. Stimulus condition C3 showed the lowest classification accuracy for POA, MOA, and VOI — significantly worse in all cases than stimulus condition C4, where the same signal window was replaced by silence.

It is reasonable to assume that the presence of noise would lead ASR systems to recognise articulatory features of a sound class that closely resembles this perturbation. Therefore, we expected that C3 would most likely lead to confusions with fricatives, and this was indeed the case across words and pseudowords. At the same time, a gap in the signal, such as in C4, may also be confused with a sound class that is similar to this perturbation, in particular with plosives. In accordance with this expectation, we found that most confusions in C4 were indeed with plosives.

It remains unclear why the ASR systems restore the articulatory features of the missing sonorants better when there is silence (C4) than in the presence of noise (C3), since the original signal was missing and the coarticulatory cues in the adjacent sounds were largely removed in both cases.

In accordance with our expectations, condition C5, which also replaced the target sounds with silence but retained coarticulatory cues, resulted in significantly more correctly classified articulatory features than C4 in all cases. Regarding the POA classification accuracy for wav2vec 2.0 embeddings, C5 performance was even on par with that of condition C2, where noise was added to the target sounds. This

demonstrates the important role of coarticulatory information for left-right contextualised representation learning (i.e., simultaneously taking left and right context into account), and ultimately ASR performance. This is consistent with prior work using the domain-informed probing task applied in this study, which showed that the probes pick up traces of coarticulatory information in the form of alternative probe activations (English et al., 2024; Shams et al., 2024b). It is further in line with Zellou et al. (2024) who demonstrated that wav2vec 2.0 (large, pre-trained) is able to detect anticipatory nasal coarticulation in vowels before nasal codas.

5.2. RQ2: Effect of word vs. pseudoword on articulatory feature encoding

Second, we examined how the lexical status of the stimulus — whether it is a word or a pseudoword — affects the encoding of articulatory features in the ASR model embeddings. We found that this factor influenced wav2vec 2.0, whereas Whisper remained unaffected by it.

In line with our assumption that lexical context would benefit the correct classification of articulatory features, wav2vec 2.0 embeddings performed significantly better in POA and MOA probing for words than for pseudowords across all stimulus conditions. We believe that VOI may not have been impacted by word type, because VOI classification accuracy was close to the ceiling for both words and pseudowords in the case of wav2vec 2.0. The performance of Whisper, however, was not impacted by word type for any of the three articulatory features. This may again be rooted in the different training methods. Whisper uses subword tokenisation (using the Byte-Pair Encoding text tokeniser from GPT-2) to represent words as smaller units (Radford et al., 2023). This may reduce the contrast between words and pseudowords, as the latter follow the phonotactic rules of English and therefore consist of subword units that may be familiar to the model. In contrast, wav2vec 2.0 is fine-tuned to predict characters from a list of 29 tokens along with a word boundary marker. The model uses this marker to predict word boundaries based on known words rather than associating embedding frames with smaller subwords that can be used to construct pseudowords.

5.3. RQ3: Effect of target sound on articulatory feature encoding

Third, we explored how the four target sounds [m], [n], [ɹ], and [l] affect the encoding of articulatory features in the ASR model embeddings. Unexpectedly, [n] stood out clearly from the other sounds.

While we had no reason to believe that classification accuracy would be affected differently for the four target sounds examined in this study, relevant information about [n] has proven to be particularly well encoded by both systems. This may be

due to the specific acoustic properties of [n]. Although both nasals [m] and [n] have characteristic nasal formants, anti-resonances, and formant transitions, their specific locations in the case of [n] may favour classification. However, several stimulus conditions did not contain the actual target signal (C3-C5), and in some conditions even coarticulatory cues were largely removed (C3 and C4). Therefore, other aspects such as frequency of occurrence in the training data may have played a role. It has been demonstrated that [n] is the most frequently occurring consonant in conversational English (Mines et al., 1978). In fact, the first six most frequently occurring consonants are all alveolar ([n], [t], [s], [ɹ], [l], and [d])[9], which may introduce a bias for this POA in the absence of other cues. While it is known that read speech data was used to train wav2vec 2.0 (see Section 3.2), the training data of Whisper is not fully disclosed. However, it is likely that conversational English, which we used here to illustrate a potential frequency effect, constitutes a significant portion of the training data for the latter. Another possibility is that a bias towards [n] was introduced by the probes which were trained with the phonetically transcribed English TIMIT speech corpus (Garofolo et al., 1993). The seven most frequently occurring consonants in the TIMIT training set are, in the following order: [s], [n], [ɹ], [l], [k], [t], and [m] — indicating the same possible bias for the four target sounds examined in this study.

5.4. RQ4: Effect of layer region on articulatory feature encoding

Fourth, we evaluated how the layer region affects the encoding of articulatory features in the ASR model embeddings. Throughout the transformer layers of the two models, which we grouped into early, middle, and late regions, classification accuracy for all three articulatory feature probes steadily increased. This is further evidence for the finding of previous work that — based on probe performance — information for the identification of (sub)phonetic features is already present to a certain extent in early layers, but becomes more prominent in middle and late layers (ten Bosch et al., 2023; English et al., 2024; Shams and Carson-Berndsen, 2024). We found that the embeddings of wav2vec 2.0 encoded more relevant information about the articulatory features than the embeddings of Whisper in all layer regions, and observed the greatest performance difference for POA and MOA in the early layers, while the greatest difference for VOI occurred in the middle layers. This is due to a marked increase in the VOI classification accuracy based on wav2vec 2.0 embeddings from the early to middle layers. The bigger performance difference between the models observed in the early and middle layers may result from contrastive learning during

[9] [m] is the ninth most frequently occurring consonant in conversational English.

the pre-training of wav2vec 2.0. This leads the model to produce embeddings — particularly those before the transformer layers — that capture phonetic information. Contrastive learning leverages discrete speech embeddings, which are generated from the output of the feature encoder. These have been demonstrated to align strongly with English phonemes (see Appendix D[10] in Baevski et al., 2020). More recent work suggests that these discrete speech embeddings relate to sub-phonemic categories (Abdullah et al., 2023). The loss calculated during the contrastive learning process is backpropagated to the feature encoder layers.[11] Therefore, extracting phonetic information is encouraged from the beginning. This early advantage of wav2vec 2.0 over Whisper is reflected in our results.

5.5. RQ5: Model performance vs. human speech perception

Finally, we assessed how the ASR model performance regarding different perturbations, stimulus types, and target sounds compares to human speech perception. This revealed similarities and differences.

In human speech perception, the Phonemic Restoration Effect (PRE) would create the impression that the stimulus condition in which signal-correlated noise replaced the target sound (C3) is the same as the condition in which the full signal is present and overlaid with noise (C2; Samuel, 1981). The embeddings of the transformer-based ASR models wav2vec 2.0 and Whisper, however, encoded the articulatory features of the target sounds least successfully across all conditions in C3. The fact that most of the confusions in this condition occur with fricatives suggests that the models tend to interpret the noise as a target, while human perception regards it as an interference masking the expected target sound.

The presence of noise is crucial for phonemic restoration to occur in human speech perception, since the illusion of continuity is interrupted by a silent gap (Warren, 1970; Warren and Obusek, 1971). Our finding that the stimulus condition in which noise replaced the target sound (C3) impaired the correct encoding of the articulatory features in the ASR model embeddings significantly more than the conditions in which noise replaced the target sound (C4, C5) thus represents a crucial difference between ASR processing and human speech perception. The ASR models do not require simulated continuity to achieve better encoding of the articulatory features

[10] `https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Supplemental.pdf`

[11] Note that the feature encoder weights learned during pre-training are not updated during the fine-tuning process. Consequently, the influence of the discrete speech embeddings remains the same in the wav2vec 2.0 model we probed.

of missing sounds. In fact, the continuous signal could lead the model to misinterpretations (e.g., noise → fricative).

As for the effect of lexical context, consistent with our prediction, we found that wav2vec 2.0 encodes articulatory features of (perturbed) sounds in words more successfully than in pseudowords. This is also the case for human speakers where the existence of a lexical entry favours restoration (Samuel, 1981). Whisper, however, did not show an effect of lexical context, which is similar to previous findings on phonemic restoration in non-native speakers (Ishida and Arai, 2016).

The sound classes investigated in this study — nasals ([m], [n]) and liquids ([ɹ], [l]) — are particularly interesting to examine from the perspective of human perception, since they are in the middle range of restorability when replaced by noise, while other sound classes perform either at the ceiling (fricatives) or the floor (vowels) in the context of the PRE (Samuel, 1981). Ishida and Arai (2016), who used the same stimuli as Mattys et al. (2014) and the present study, found that listeners restored nasals more so than liquids. This partly matches the present results, where the nasal [n] stood out as being most successfully encoded by the ASR models across conditions.

5.6. Limitations and Future Work

The present study has some limitations that should be acknowledged. Some of these have the potential to be addressed in future research. As mentioned at the outset, the domain-informed probes used in this study have a relatively simple architecture, which enables their performance to mainly reflect the quality of the ASR model embeddings and not their own complexity (see Sections 2.3 and 3.3). However, as we use these probing algorithms to make the embeddings interpretable for humans, the results inevitably still rely on their performance. In other words, we can only provide as much insight into the information encoded by the ASR models as we can observe through the lens of the probes.

The three probes for POA, MOA, and VOI were trained on the phoneme set of American English based on the TIMIT corpus (Garofolo et al., 1993), whereby the probabilities for articulatory features that are not present in this set were assumed to be zero. This means that a *retroflex*, *uvular*, and *pharyngeal* POA, as well as the MOAs *trill* and *lateral fricative* did not have any training data representation and therefore cannot be learned or predicted by the probes — even if the embeddings should encode these POAs and MOAs, the probes would not detect them. However, to have a more intuitive probability vector which corresponds to the IPA classification for consonants and can be used in future work investigating other languages, the missing features were retained and inserted at their respective positions.

As explained in Section 3.2, the probing method applied in this work is limited to the encoder layers of the ASR models. In the case of Whisper, a considerable part of the internal processing is therefore not taken into account. This is warranted because our focus was on contextualised representations of speech segments within the two models as opposed to the lexical segments processed in the decoder of Whisper. However, excluding the decoder may lead to an underestimation of Whisper's full ASR capabilities.

We focused on a limited set of target sounds in this study, namely two nasals ([m] and [n]) and two liquids ([ɹ] and [l]), as these were investigated in Mattys et al. (2014). Future work could explore how ASR models process other sound classes in the presence of segment-level perturbations. Especially the encoding of fricatives in the presence of noisy gaps or plosives in the presence of silent gaps would be interesting cases to examine, since the similarity of the perturbation to the sound class may prevent disruptive effects.

In terms of the analysis, in this work, we focused on the comparison of the two transformer-based ASR models wav2vec 2.0 and Whisper. Hence we explored how the encoding of articulatory features in their embeddings is affected by segment-level perturbations, word and pseudoword items, different target sounds, and in distinct layer regions. A future analysis could take a closer look at the interaction of item types, target sounds, and layer regions with different perturbations to investigate whether specific perturbations affect certain factor combinations more than others.

Future work could also consider a post-hoc analysis of a wav2vec 2.0 model without fine-tuning such as *large-lv60*. This would help clarify the extent to which the encoding of articulatory features in the model embeddings is influenced by its pre-trained representations versus the adaptations introduced during fine-tuning. Since fine-tuning has been shown to improve the encoding of relevant phonetic information in the embeddings (Pasad et al., 2021), we expect that probing the embeddings of a model without fine-tuning would lead to worse articulatory feature classification.

Furthermore, complementing the current results with an analysis of the character error rate of the target sounds in the textual outputs of wav2vec 2.0 and Whisper would provide deeper insight into how the information encoded throughout their layers relates to their final predictions.

Finally, this study investigated segment-level perturbations which can typically be perceived by humans. Another possible avenue for future work would be to experiment with adversarial perturbations that target specific speech sounds and cause ASR systems to mistake a word for its minimal pair counterpart — for example, "ship" [ʃɪp] instead of "sip" [sɪp] — while the interference remains undetectable to human perception.

6. Conclusion

The two state-of-the-art ASR models examined in this study showed similarities and differences in the way they internally processed segment-level signal perturbations. In contrast to phonemic restoration in human speech perception, the articulatory feature encoding of both systems was most affected by noisy gaps in the speech signal, and less so by silent gaps. Added noise had the least impact in both cases. Coarticulatory cues surrounding a silent gap improved articulatory feature encoding of the missing sound for both models. The target sound [n] was more clearly encoded in the embeddings across conditions than [m], [ɹ], and [l]. Furthermore, some information about articulatory features was found to be already encoded in early layers of both models, but information became more prominent over middle to late layers. Similar to human perception, articulatory feature encoding across conditions was stronger for words than pseudowords in the case of wav2vec 2.0 embeddings. Whisper did not show such an effect of lexical status — similar to previous findings on phonemic restoration in non-native speakers. Differences between the two models could be related to the distinct training methods and objectives of the models. Although the insights into the internal processing of the two models gained in the present study do not directly predict the transcriptions produced by them beyond the encoder components, they provide valuable context for understanding modern ASR systems and making them more interpretable to humans. Identifying where model processing differs from human speech processing by evaluating ASR performance in experimental paradigms from human speech perception research can highlight the limitations and distinctive features of ASR systems. This, in turn, fosters a more realistic understanding of their capabilities.

Acknowledgments

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

References

Abdullah, B.M., Shaik, M.M., Möbius, B., Klakow, D., 2023. An information-theoretic analysis of self-supervised discrete representations of speech, in: Interspeech, pp. 2883–2887. doi:10.21437/Interspeech.2023-2131.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle, in: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer, pp. 199–213. doi:10.1007/978-1-4612-1694-0_15.

Alain, G., Bengio, Y., 2017. Understanding intermediate layers using linear classifier probes. URL: https://openreview.net/forum?id=ryF7rTqgl.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., et al., 2016. Deep Speech 2: End-to-end speech recognition in English and Mandarin, in: International Conference on Machine Learning, PMLR. pp. 173–182. URL: https://proceedings.mlr.press/v48/amodei16.html.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations, in: NeurIPS, pp. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48. doi:10.18637/jss.v067.i01. R package version 1.1-36.

Belinkov, Y., Glass, J., 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems, in: NeurIPS, p. 2438–2448. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/b069b3415151fa7217e870017374de7c-Paper.pdf.

de Heer Kloots, M., Zuidema, W., 2024. Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2.0, in: Interspeech, pp. 4593–4597. doi:10.21437/Interspeech.2024-2490.

de la Fuente, A., Jurafsky, D., 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models, in: Interspeech, pp. 1290–1294. doi:10.21437/Interspeech.2024-2341.

Delcroix, M., Kubo, Y., Nakatani, T., Nakamura, A., 2013. Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?, in: Interspeech, pp. 2992–2996. doi:10.21437/Interspeech.2013-276.

Dhawan, K., Koluguri, N.R., Jukić, A., Langman, R., Balam, J., Ginsburg, B., 2024. Codec-ASR: Training performant automatic speech recognition systems with

discrete speech representations, in: Interspeech, pp. 2574–2578. doi:10.21437/Interspeech.2024-330.

Drozdova, P., Van Hout, R., Scharenborg, O., 2016. Lexically-guided perceptual learning in non-native listening. Bilingualism: Language and Cognition 19, 914–920. doi:10.1017/S136672891600002X.

English, P.C., Kelleher, J.D., Carson-Berndsen, J., 2022. Domain-informed probing of wav2vec 2.0 embeddings for phonetic features, in: SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 83–91. doi:10.18653/v1/2022.sigmorphon-1.9.

English, P.C., Kelleher, J.D., Carson-Berndsen, J., 2023. Discovering phonetic feature event patterns in transformer embeddings, in: Interspeech, pp. 4733–4737. doi:10.21437/Interspeech.2023-1985.

English, P.C., Shams, E.A., Kelleher, J.D., Carson-Berndsen, J., 2024. Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio, in: ICASSP, pp. 6685–6689. doi:10.1109/ICASSP48485.2024.10446494.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V., 1993. TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium doi:10.35111/17gk-bn40.

Gong, Y., Khurana, S., Karlinsky, L., Glass, J., 2023. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers, in: Interspeech, pp. 2798–2802. doi:10.21437/Interspeech.2023-2193.

International Phonetic Association, 2015. IPA Chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2015 International Phonetic Association. URL: https://www.internationalphoneticassociation.org/content/full-ipa-chart.

Ishida, M., Arai, T., 2016. Missing phonemes are perceptually restored but differently by native and non-native listeners. SpringerPlus 5, 1–10. doi:10.1186/s40064-016-2479-8.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software 82, 1–26. doi:10.18637/jss.v082.i13. R package version 3.1-3.

Lee, S., Kim, S., Chung, M., 2025. Exploring acoustic foundations in speech production assessment models for children with cochlear implants, in: ICASSP, pp. 1–5. doi:10.1109/ICASSP49660.2025.10889076.

Lenth, R.V., 2025. emmeans: Estimated Marginal Means, aka Least-Squares Means. URL: https://rvlenth.github.io/emmeans/. R package version 1.10.1.

Lüdecke, D., 2018. ggeffects: Tidy data frames of marginal effects from regression models. Journal of Open Source Software 3, 772. doi:10.21105/joss.00772. R package version 2.2.0.

Mattys, S.L., Barden, K., Samuel, A.G., 2014. Extrinsic cognitive load impairs low-level speech perception. Psychonomic Bulletin & Review 21, 748–754. doi:10.3758/s13423-013-0544-7.

McQueen, J.M., Cutler, A., Norris, D., 1999. Lexical activation produces impotent phonemic percepts. The Journal of the Acoustical Society of America 106, 2296–2296. doi:10.1121/1.427858.

Mines, M.A., Hanson, B.F., Shoup, J.E., 1978. Frequency of occurrence of phonemes in conversational English. Language and Speech 21, 221–241. doi:10.1177/002383097802100302.

Mohebbi, H., Chrupała, G., Zuidema, W., Alishahi, A., 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers, in: EMNLP, pp. 8249–8260. doi:10.18653/v1/2023.emnlp-main.513.

Montavon, G., Braun, M.L., Müller, K.R., 2011. Kernel analysis of deep networks. Journal of Machine Learning Research 12, 2563–2581. URL: https://www.jmlr.org/papers/volume12/montavon11a/montavon11a.pdf.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychology 47, 204–238. doi:10.1016/S0010-0285(03)00006-9.

Olivier, R., Raj, B., 2023. There is more than one kind of robustness: Fooling Whisper with adversarial examples, in: Interspeech, pp. 4394–4398. doi:10.21437/Interspeech.2023-1105.

Pasad, A., Chou, J.C., Livescu, K., 2021. Layer-wise analysis of a self-supervised speech representation model, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 914–921. doi:10.1109/ASRU51503.2021.9688093.

Pasad, A., Shi, B., Livescu, K., 2023. Comparative layer-wise analysis of self-supervised speech models, in: ICASSP, pp. 1–5. doi:10.1109/ICASSP49357.2023.10096149.

Patman, C., Chodroff, E., 2024. Speech recognition in adverse conditions by humans and machines. JASA Express Letters 4. doi:10.1121/10.0032473.

Peso Parada, P., Dobrowolska, A., Saravanan, K., Ozay, M., 2022. pMCT: Patched multi-condition training for robust speech recognition, in: Interspeech, pp. 3779–3783. doi:10.21437/Interspeech.2022-117.

Posit Team, 2024. RStudio: Integrated Development Environment for R. Posit Software, PBC. Boston, MA. URL: http://www.posit.co/. Version 2024.12.0+467.

Pouw, C., de Heer Kloots, M., Alishahi, A., Zuidema, W., 2024. Perception of phonological assimilation by neural speech recognition models. Computational Linguistics 50, 1557–1585. doi:10.1162/coli_a_00526.

Prasad, A., Jyothi, P., 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems, in: Annual Meeting of the Association for Computational Linguistics, pp. 3739–3753. doi:10.18653/v1/2020.acl-main.345.

R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/. Version 4.4.2.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision, in: ICML, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a/radford23a.pdf.

Samuel, A.G., 1981. Phonemic restoration: insights from a new methodology. Journal of Experimental Psychology: General 110, 474–494. doi:10.1037//0096-3445.110.4.474.

Samuel, A.G., 1987. Lexical uniqueness effects on phonemic restoration. ournal of Memory and Language 26, 36–56. doi:10.1016/0749-596X(87)90061-1.

Samuel, A.G., 1997. Lexical activation produces potent phonemic percepts. Cognitive Psychology 32, 97–127. doi:10.1006/cogp.1997.0646.

Scharenborg, O., Tiesmeyer, S., Hasegawa-Johnson, M., Dehak, N., 2018. Visualizing phoneme category adaptation in deep neural networks, in: Interspeech, pp. 1482–1486. doi:10.21437/Interspeech.2018-1707.

Shah, M.A., Noguero, D.S., Heikkila, M.A., Raj, B., Kourtellis, N., 2024. Speech robust bench: A robustness benchmark for speech recognition. URL: https://arxiv.org/abs/2403.07937. Under review at the NeurIPS datasets and benchmark track 2025.

Shams, E.A., Carson-Berndsen, J., 2024. Attention to phonetics: A visually informed explanation of speech transformers, in: Text, Speech, and Dialogue, pp. 81–93. doi:10.1007/978-3-031-70566-3_8.

Shams, E.A., Gessinger, I., Carson-Berndsen, J., 2024a. Uncovering syllable constituents in the self-attention-based speech representations of Whisper, in: BlackboxNLP Workshop, Association for Computational Linguistics. pp. 238–247. doi:10.18653/v1/2024.blackboxnlp-1.16.

Shams, E.A., Gessinger, I., English, P.C., Carson-Berndsen, J., 2024b. Are articulatory feature overlaps shrouded in speech embeddings?, in: Interspeech, pp. 4608–4612. doi:10.21437/Interspeech.2024-1039.

Shen, G., Alishahi, A., Bisazza, A., Chrupała, G., 2023. Wave to syntax: Probing spoken language models for syntax, in: Interspeech, pp. 1259–1263. doi:10.21437/Interspeech.2023-679.

Singla, Y.K., Shah, J., Chen, C., Shah, R.R., 2022. What do audio transformers hear? Probing their representations for language delivery and structure, in: IEEE International Conference on Data Mining Workshops (ICDMW), pp. 910–925. doi:10.1109/ICDMW58026.2022.00120.

ten Bosch, L., Bentum, M., Boves, L., 2023. Phonemic competition in end-to-end ASR models, in: Interspeech, pp. 586–590. doi:10.21437/Interspeech.2023-1846.

Vaidya, T., Zhang, Y., Sherr, M., Shields, C., 2015. Cocaine noodles: exploiting the gap between human and machine speech recognition, in: USENIX Conference on Offensive Technologies, p. 16. doi:10.5555/2831211.2831227.

Vitale, V.N., Cutugno, F., Origlia, A., Coro, G., 2024. Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique. Neural Computing and Applications doi:10.1007/s00521-024-09435-1.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393. doi:10.1126/science.167.3917.392.

Warren, R.M., Obusek, C.J., 1971. Speech perception and phonemic restorations. Perception & Psychophysics 9, 358–362. doi:10.3758/BF03212667.

Xing, J., Luo, D., Xue, C., Xing, R., 2025. Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective. URL: https://arxiv.org/abs/2411.14654. Accepted at the International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence 2025, proceedings forthcoming.

Yang, M., Shekar, R.C.M.C., Kang, O., Hansen, J.H.L., 2023. What can an accent identifier learn? Probing phonetic and prosodic information in a wav2vec2-based accent identification model, in: Interspeech, pp. 1923–1927. doi:10.21437/Interspeech.2023-2254.

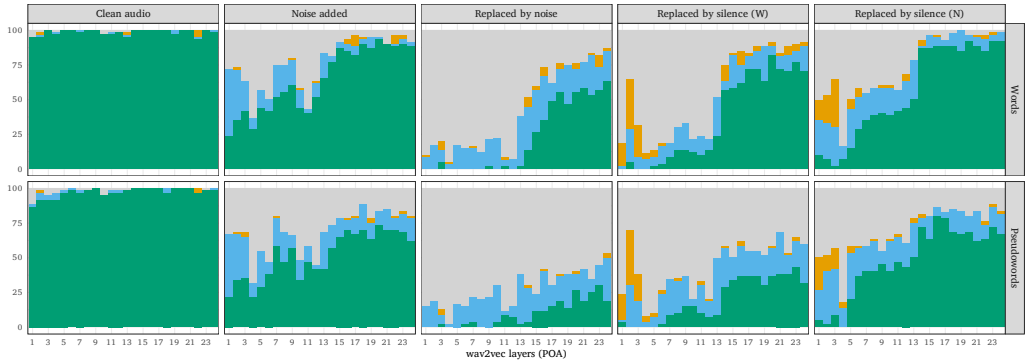
Zellou, G., Kim, L., Gendrot, C., 2024. Comparing human and machine's use of coarticulatory vowel nasalization for linguistic classification. The Journal of the Acoustical Society of America 156, 489–502. doi:10.1121/10.0027932.

Appendix A. Word/pseudoword pairs

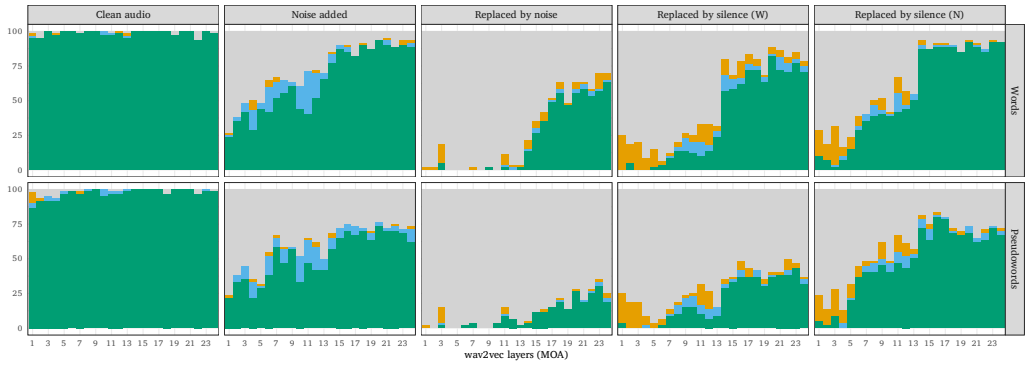
Table A.1: Pairs of words and pseudowords from Mattys et al. (2014) sharing the same stress pattern and the same final two syllables. The target sound is either a nasal (13 × [m], 17 × [n]) or a liquid (15 × [ɹ], 15 × [l]).

	Nasal		**Liquid**	
	Word	Pseudoword	Word	Pseudoword
1	abdomi**n**al	gobflomi**n**al	accele**r**ate	vabbelle**r**ate
2	acade**m**ic	wopplade**m**ic	accumu**l**ate	muzoomu**l**ate
3	accompani**m**ent	ishimpni**m**ent	alcoho**l**ic	balmeho**l**ic
4	aerodyna**m**ic	wissafyna**m**ic	apoca**l**ypse	twestoka**l**ips
5	anato**m**y	hustatto**m**y	approximate**l**y	guttropimate**l**y
6	astrono**m**er	yiplono**m**er	atmosphe**r**ic	nipzusphe**r**ic
7	cacopho**n**y	bitopho**n**y	binocu**l**ars	yaboku**l**ars
8	constitutio**n**al	gaspofootio**n**al	caterpi**ll**ar	laggerpi**ll**ar
9	conventio**n**al	hakzenshu**nn**el	congratu**l**ate	humstratu**l**ate
10	conversatio**n**al	zandustatio**n**al	conside**r**ate	bupwide**r**ate
11	coordi**n**ate	banstordi**n**ate	contradicto**r**y	vampradicto**r**y
12	develop**m**ent	sikwellup**m**ent	curricu**l**um	hotricu**l**um
13	dichoto**m**y	stuffloto**m**y	disciplina**r**y	huffakina**r**y
14	discrimi**n**ate	notromi**n**ate	enginee**r**ing	nonkineer**r**ing
15	disfigure**m**ent	pashigu**mm**ent	epheme**r**al	yokweme**r**al
16	disproportio**n**ate	hastigortio**n**ate	equiva**l**ent	mestiva**l**ent
17	econo**m**ize	marzono**m**ise	exagge**r**ate	tupstagge**r**ate
18	educatio**n**al	voplortatio**n**al	herbivo**r**ous	waarpivo**r**ous
19	embezzle**m**ent	tumkezzle**m**ent	illite**r**ate	ossite**r**ate
20	emotio**n**al	piskotio**n**al	imagina**r**y	fittagin**r**ee
21	enlighten**m**ent	sinfighten**m**ent	immacu**l**ate	hassacu**l**ate
22	imperti**n**ence	umferti**n**ence	immobi**l**ize	kwatobi**l**ise
23	matrimo**n**y	bleffrimo**n**y	incredu**l**ous	winfedu**l**ous
24	monoga**m**y	stantoga**m**y	irregu**l**ar	appegu**l**ar
25	perfectio**n**ist	hargoctio**n**ist	molecu**l**ar	jottoku**l**ar
26	receptio**n**ist	rudipshu**nn**ist	periphe**r**al	wegniffe**r**al
27	saxopho**n**ist	luddopho**n**ist	persevе**r**ance	glertevee**r**ance
28	taxono**m**y	nastono**m**y	recove**r**y	stroppuvve**r**y
29	tobacco**n**ist	messacko**n**ist	rhinoce**r**os	reekosu**rr**us
30	vindictive**n**ess	kwamsholtive**n**ess	spectacu**l**ar	nupshacku**l**ar

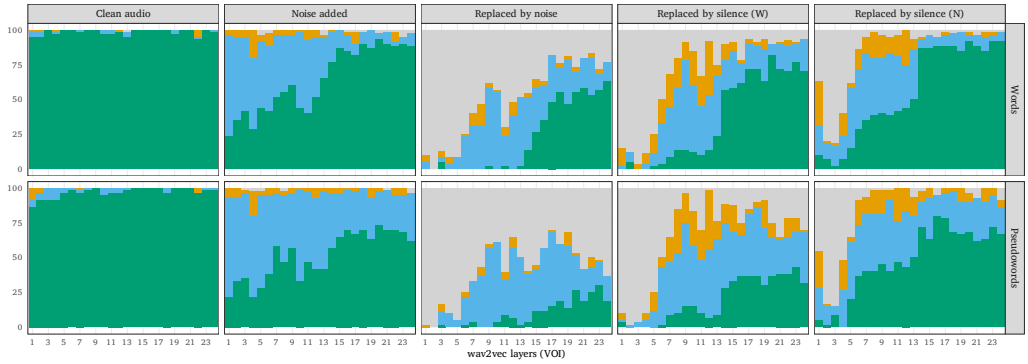
Appendix B. Target classification throughout model layers



(a) POA: target full match ■, same POA (attested) ■, same POA (unattested) ■, or different POA ■.

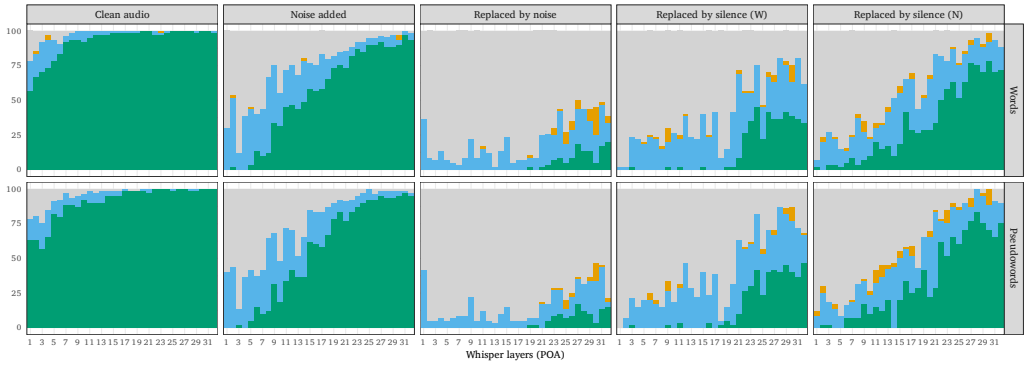


(b) MOA: target full match ■, same MOA (attested) ■, same MOA (unattested) ■, or different MOA ■.

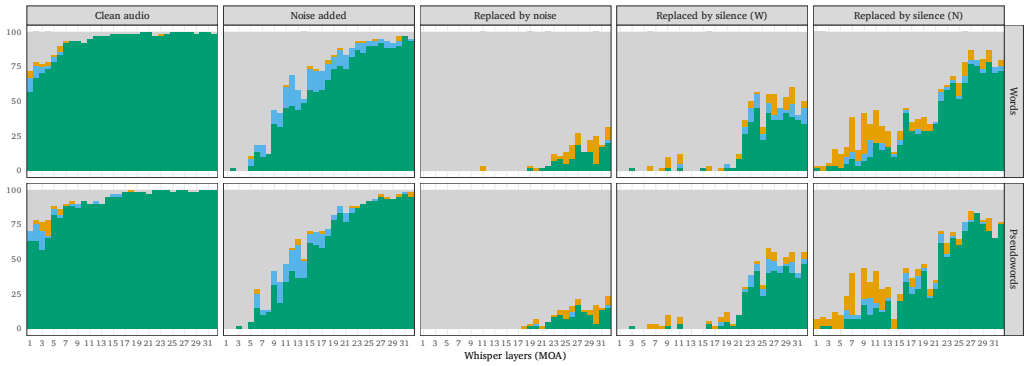


(c) VOI: target full match ■, voiced (attested) ■, voiced (unattested) ■, or voiceless ■.

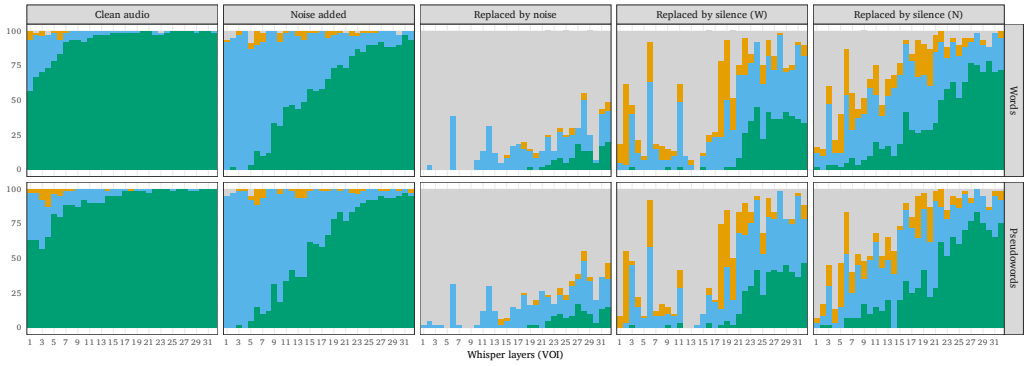
Figure B.1: **wav2vec 2.0** — percentage target correctly classified in layers 1 to 24 with focus on place of articulation (POA), manner of articulation (MOA), and voicing (VOI).



(a) POA: target full match ■, same POA (attested) ■, same POA (unattested) ■, or different POA ■.



(b) MOA: target full match ■, same MOA (attested) ■, same MOA (unattested) ■, or different MOA ■.



(c) VOI: target full match ■, voiced (attested) ■, voiced (unattested) ■, or voiceless ■.

Figure B.2: **Whisper** — percentage target correctly classified in layers 1 to 32 with focus on place of articulation (POA), manner of articulation (MOA), and voicing (VOI).

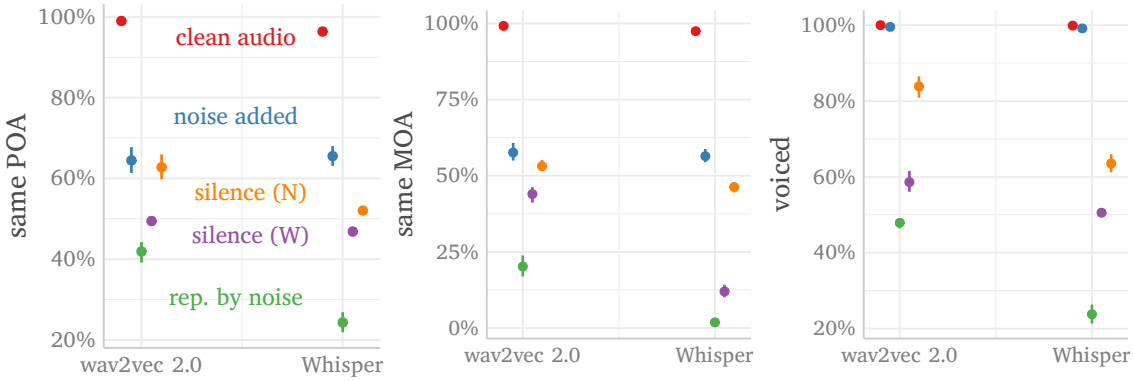
Appendix C. Attested vs. unattested sounds

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

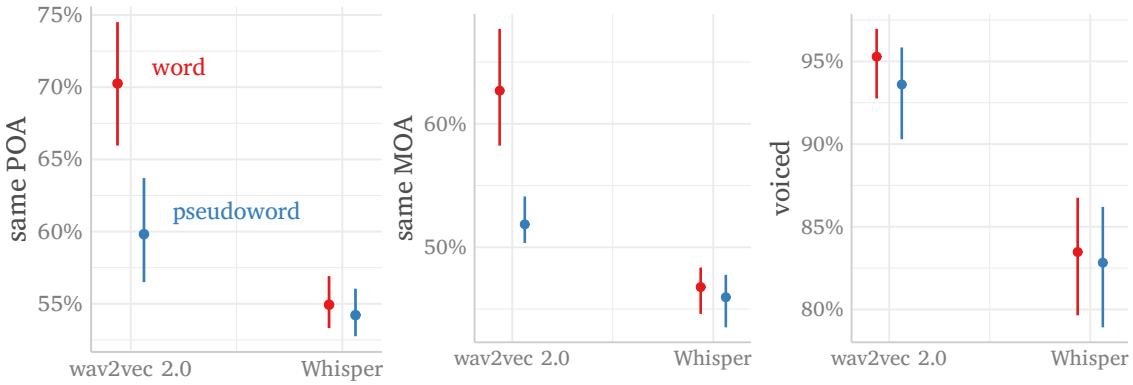
Figure C.1: Overview of sounds we consider to be *attested* (blue) and *unattested* (orange) in the IPA. Unattested sounds with a grey frame denote articulations judged impossible in the IPA. The sounds [m], [n], [ɹ], and [l] are also the target sound realisations (green) examined in this study.

Appendix D. Interaction effects



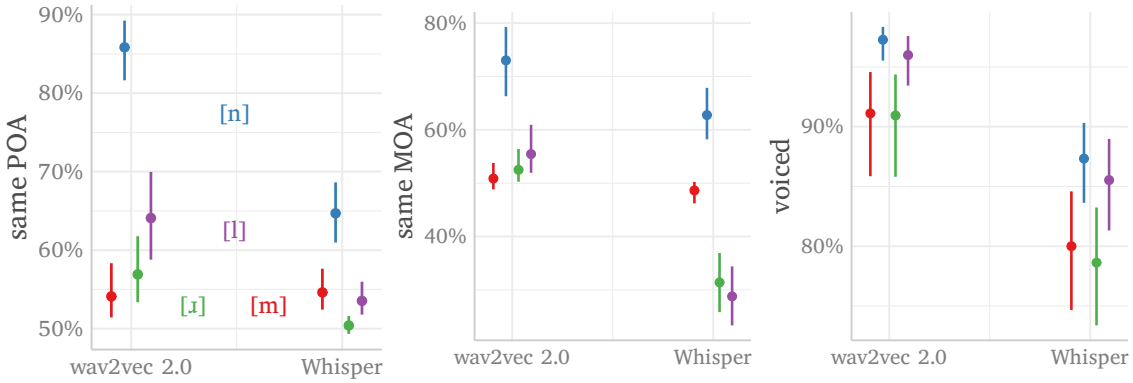
(a) Place of articulation (b) Manner of articulation (c) Voicing

Figure D.1: Estimated marginal means for stimulus condition: clean audio, noise added, replaced by noise, replaced by silence (wide), and replaced by silence (narrow) across the two ASR models. Error bars represent the 95 % confidence intervals.



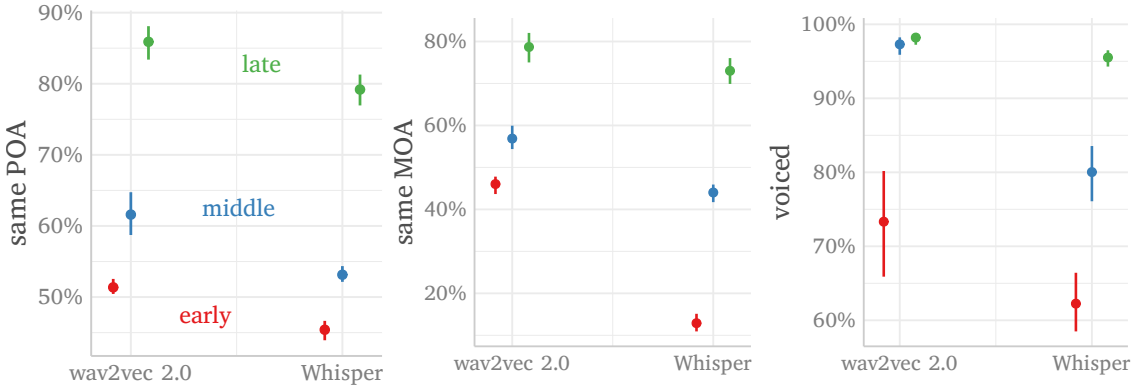
(a) Place of articulation (b) Manner of articulation (c) Voicing

Figure D.2: Estimated marginal means for stimulus being a word or a pseudoword across the two ASR models. Error bars represent the 95 % confidence intervals.



(a) Place of articulation (b) Manner of articulation (c) Voicing

Figure D.3: Estimated marginal means for target sound being [m], [n], and [ɹ], and [l] across the two ASR models. Error bars represent the 95 % confidence intervals.



(a) Place of articulation (b) Manner of articulation (c) Voicing

Figure D.4: Estimated marginal means for early layers, middle layers, and late layers across the two ASR models. Error bars represent the 95 % confidence intervals.

Appendix E. Target confusions in the best-performing layers

Table E.1: **wav2vec 2.0, words** — combined POA, MOA and VOI probe outputs for layer 16 embeddings. Outputs are grouped by MOA and ordered by POA (front to back) within each group. Note: Approx. → Approximant, Lat. A. → Lateral Approximant, bila → bilabial, lab → labiodental, pala → palatal, glot → glottal, 0 → voiceless, 1 → voiced.

	Words	Clean audio				Noise added				Replaced by noise				Replaced by silence (W)				Replaced by silence (N)			
		m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l
Plosives	p	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	2	0	1	0	1	0	0	0	1	0	0	0
	lab0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	t	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0	1
	d	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	0
	k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	g	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	ʔ	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1
	ʔ1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Nasals	m0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
	m	**13**	0	0	0	**8**	0	0	0	1	0	0	0	**5**	0	0	0	**11**	0	0	0
	n0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
	n	0	**17**	0	0	0	**16**	1	0	0	**12**	0	0	3	**16**	1	0	1	**17**	0	0
	ɲ	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Taps	ɾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	pala1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fricatives	ɸ	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	β	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	f	0	0	0	0	0	0	0	0	**4**	0	1	0	0	0	0	0	0	0	0	0
	v	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
	θ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	s	0	0	0	0	0	0	0	0	0	1	**2**	3	0	0	0	1	0	0	0	0
	z	0	0	0	0	0	1	0	0	1	1	1	2	0	0	0	0	0	0	0	0
	ʒ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɦ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Approx.	bila1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɹ0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	3	0	0	0	1	0
	ɹ	0	0	**15**	0	0	0	**14**	0	0	0	1	0	0	0	**6**	1	0	0	**12**	0
	j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɥ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Lat. A.	bila1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	l0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
	l	0	0	0	**15**	2	0	0	**13**	0	0	0	**7**	0	0	0	**10**	0	0	0	**13**
	ʎ	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	ʟ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table E.2: **wav2vec 2.0, pseudowords** — combined POA, MOA and VOI probe outputs for layer 16 embeddings. Outputs are grouped by MOA and ordered by POA (front to back) within each group. Note: Approx. → Approximant, Lat. A. → Lateral Approximant, bila → bilabial, lab → labiodental, pala → palatal, glot → glottal, 0 → voiceless, 1 → voiced.

Pseudowords		Clean audio				Noise added				Replaced by noise				Replaced by silence (W)				Replaced by silence (N)			
		m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l
Plosives	p	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
	lab0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	t	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**3**	0	0	0	1	0
	d	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	1
	k	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
	g	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	ʔ	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
	ʔ1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Nasals	m0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	m	**13**	0	0	0	**5**	0	1	2	0	0	0	0	**2**	0	0	0	**11**	0	0	0
	n0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	n	0	**17**	0	0	3	**14**	1	1	0	**7**	0	0	**2**	**15**	2	1	0	**17**	1	0
	ɲ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Taps	ɾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	pala1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Fricatives	ɸ	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	1	0
	β	0	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0
	f	0	0	0	0	0	0	0	0	**6**	1	**8**	0	0	0	0	0	0	0	0	0
	v	0	0	0	0	2	0	1	0	0	0	2	4	1	0	1	0	2	0	1	0
	θ	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
	ð	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	1	0	0	0	1
	s	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0
	z	0	0	0	0	0	0	0	0	1	2	0	**5**	0	0	0	1	0	0	0	0
	ʒ	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	j	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	h	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
	ɦ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Approx.	bila1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɹ0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	ɹ	0	0	**15**	0	1	1	**11**	0	0	0	0	0	0	0	2	0	0	0	**10**	0
	j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	ɥ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lat. A.	bila1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	l0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	l	0	0	0	**15**	1	1	0	**12**	0	0	0	0	1	1	0	**3**	0	0	1	**10**
	ʎ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
	ɺ	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table E.3: **Whisper, words** — combined POA, MOA and VOI probe outputs for layer 25 embeddings. Outputs are grouped by MOA and ordered by POA (front to back) within each group. Note: Approx. → Approximant, Lat. A. → Lateral Approximant, lab → labiodental, glot → glottal, 0 → voiceless, 1 → voiced, + → POA fronted, − → POA retracted.

	Words	Clean audio				Noise added				Replaced by noise				Replaced by silence (W)				Replaced by silence (N)			
		m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l
Plosives	p	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0
	lab0	0	0	0	0	0	0	0	0	2	0	1	1	0	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	2	0	0	0	0
	t+	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	t	0	0	0	0	0	0	0	0	0	3	1	1	1	0	0	0	0	0	0	0
	d	0	0	0	0	0	0	0	1	0	3	1	1	0	0	**6**	**3**	0	0	**4**	3
	k	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
	ʔ	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	ʔl	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1
Nasals	m	**13**	0	0	0	**11**	0	0	0	0	0	0	0	**5**	0	0	0	**10**	0	0	0
	ɱ0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	ɱ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
	n0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
	n	0	**17**	0	0	2	**17**	0	1	0	**5**	0	0	3	**17**	3	0	0	**17**	2	0
	n−0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	ŋ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Tap	ɾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Fricatives	f	0	0	0	0	0	0	0	0	**8**	0	**5**	**5**	0	0	0	0	0	0	0	0
	v	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0
	θ	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	ð	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0
	s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	z	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0
	ʃ	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Approx.	ʋ0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	ʋ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0
	ɹ0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
	ɹ	0	0	**15**	0	0	0	**14**	0	0	0	0	0	0	0	2	0	0	0	**4**	0
	ɹ−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0
Lat. A.	lab1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
	l0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	l	0	0	0	**15**	0	0	1	**12**	0	0	0	0	1	0	0	1	0	0	0	**7**
	ɭ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table E.4: **Whisper, pseudowords** — combined POA, MOA and VOI probe outputs for layer 25 embeddings. Outputs are grouped by MOA and ordered by POA (front to back) within each group. Note: Approx. → Approximant, Lat. A. → Lateral Approximant, lab → labiodental, glot → glottal, 0 → voiceless, 1 → voiced, + → POA fronted, − → POA retracted.

Pseudowords		Clean audio				Noise added				Replaced by noise				Replaced by silence (W)				Replaced by silence (N)			
		m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l	m	n	ɹ	l
Plosives	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	lab0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0
	lab1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1
	t+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	t	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
	d	0	0	0	0	0	0	1	0	0	3	0	0	1	0	**4**	2	0	1	**4**	1
	k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ʔ	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0
	ʔl	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
Nasals	m	**13**	0	0	0	**13**	0	0	1	0	0	0	0	**5**	0	0	0	**12**	0	0	0
	ŋ0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ŋ	0	0	0	0	0	0	0	0	1	1	0	0	2	0	0	0	0	0	0	0
	n0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
	n	0	**17**	0	0	0	**17**	0	2	0	**5**	0	0	3	**17**	**4**	**6**	0	**16**	3	0
	n−0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ŋ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Tap	ɾ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fricatives	f	0	0	0	0	0	0	0	0	**8**	1	**11**	**6**	0	0	0	0	0	0	0	0
	v	0	0	0	0	0	0	0	0	1	0	0	4	0	0	0	0	0	0	0	0
	θ	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	s	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	z	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1
	ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Approx.	ʋ0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ʋ	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**4**	0
	ɹ0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɹ	0	0	**15**	0	0	0	**13**	0	0	0	0	0	0	0	1	0	0	0	3	1
	ɹ−	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	glot1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Lat. A.	lab1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
	l0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	l	0	0	0	**15**	0	0	0	**12**	0	0	0	0	0	0	0	1	0	0	0	**11**
	ɭ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	glot0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0