# What Do Neural Speech Models Know About Phonology? Evidence from Structured Phoneme Confusions

Anonymous ACL submission

## Abstract

ASR errors are typically analyzed at the phoneme level, treating phonemes as atomic symbols. In this work, we instead adopt a featural representation of phonemes, grounded in phonological theory, which models speech sounds as structured bundles of distinctive articulatory and acoustic properties. This perspective allows us to analyze recognition errors at a finer granularity and to investigate whether certain phonological features are more vulnerable than others. Across multiple languages, we show that phoneme confusions are strongly structured in phonological feature space: errors are predominantly local and exhibit systematic asymmetries that reveal a small set of weakly modeled features. These findings have direct implications both for the design and diagnosis of ASR systems and for cognitive models of human speech perception, where similar feature-level asymmetries have long been observed.

## 1 Introduction

Despite their strong empirical performance, modern neural speech recognition models remain largely opaque. Trained end-to-end to optimise transcription accuracy, they incorporate no explicit phonological supervision or inductive bias of the kind assumed in feature-based or class-based models of speech perception. Any phonological structure present in their internal representations must therefore emerge implicitly from data, motivating diagnostic analyses that probe whether such models encode linguistically meaningful structure beyond surface input–output correlations (Belinkov and Glass, 2019; Belinkov, 2022).

Most existing work in this area adopts a probing perspective, analysing internal representations using auxiliary classifiers or geometric methods (Belinkov, 2022; de Seyssel et al., 2022). While these approaches have yielded valuable insights, they are inherently indirect and sensitive to architectural and methodological choices. In contrast, we adopt an output-centred perspective and analyse transcription errors directly. Because model outputs reflect the cumulative effect of all internal computations, systematic structure in errors provides a direct and architecture-agnostic window into the linguistic regularities captured by the model.

The central contribution of our paper is to show that phoneme recognition errors produced by neural speech models exhibit systematic asymmetries at the level of a phonologically grounded featural representation, in which speech sounds are modelled as structured bundles of distinctive features such as voicing, place and manner of articulation, or vowel quality. Rather than behaving as symmetric noise, errors preferentially preserve some phonological properties while systematically losing others. These asymmetries make it possible to identify which components of the phonological representation are more robustly encoded by the model, and which are intrinsically more vulnerable once recognition fails. Beyond their engineering relevance, such patterns are also informative from a cognitive modelling perspective, as asymmetric feature confusions have long been observed in human speech perception and are commonly interpreted as reflecting differential robustness of phonological features under noise (Miller and Nicely, 1955; Chomsky and Halle, 1968; Mielke, 2008).

As a prerequisite for this analysis, we first establish that phoneme substitution errors are phonologically local: substituted phonemes tend to be close to their targets in phonological feature space. This result serves as a validation step, showing that model errors preserve fine-grained phonological structure rather than collapsing phonemes into arbitrary symbols. Establishing phonological locality provides a meaningful basis for interpreting feature-level asymmetries and already indicates that the model's internal representations are compatible with a phonologically structured organisation of speech sounds; however, locality alone does not

reveal which specific features are preferentially preserved or lost.

To address these two questions, we introduce an error-based probing framework that represents phonemes as vectors of distinctive features and analyses transcription errors in phonological feature space. Our framework combines (i) feature-based distance measures to assess phonological locality, (ii) analyses of asymmetric feature insertions and deletions conditional on the occurrence of an error, and (iii) uncertainty-aware effect-size estimation using bootstrap confidence intervals and regions of practical equivalence. We apply this framework to phoneme recognition systems evaluated on 12 languages, including a single multilingual model trained with automatic phonemic supervision (Xu et al., 2022), as well as a language-specific system trained on expert phonemic annotations for Thulung, a newly documented Sino-Tibetan language.

Across languages, we show that phoneme confusions are strongly structured in phonological feature space and that errors exhibit consistent, directional asymmetries at the level of distinctive features. Broad class-level and sonority-related features tend to be preserved or over-generated, while fine-grained place, manner, and secondary articulatory features are systematically lost. Taken together, these findings demonstrate that phoneme recognition errors are not arbitrary, but reflect the internal organisation of phonological representations learned by neural speech models, offering a principled bridge between engineering-oriented evaluation and insights from phonological theory and human speech perception.

The remainder of the paper is organised as follows. Section 2 describes the speech recognition models, the multilingual evaluation data, and the phonological feature representations used in our analyses. Section 3 establishes that phoneme substitution errors are phonologically local, by showing that confused phonemes are significantly closer in feature space than would be expected by chance. Section 4 then analyses directional asymmetries in feature-level confusions, identifying which phonological features are more robustly preserved or systematically lost once recognition errors occur. Finally, Section 5 discusses the implications of these findings for the analysis of neural speech models and for connections between machine and human speech perception.

## 2 Experimental Setup and Phonological Feature Framework

### 2.1 Models and Data

Our methodology builds on recent self-supervised speech models that can predict *phonemic* transcriptions directly, rather than graphemic or character-based outputs.[1] In particular, we rely on the `wav2vec2/XLSR-53` architecture (Baevski et al., 2020; Conneau et al., 2021), whose Transformer-based encoder learns general-purpose acoustic representations that can be fine-tuned for phoneme-level automatic speech recognition.

We use the `Wav2Vec2Phoneme` model of Xu et al. (2022), which fine-tunes an XLSR-53 encoder to predict phoneme sequences from speech. Training data are drawn from Common Voice and Babel, with phonemic supervision obtained automatically via the `eSpeak` phonemiser. The model is trained with a standard CTC objective. We access the pre-trained model through the Hugging Face 🤗 API.

We evaluate this model on 12 languages spanning seven language families (Indo-European, Uralic, Austronesian, Afro-Asiatic, Dravidian, Sino-Tibetan, and Turkic). For 11 languages, evaluation is performed on a random sample of 250 utterances from the test split of Common Voice version 24.0. Six languages (English, French, Indonesian, Swedish, Tamil, and Turkish) are seen during fine-tuning, while the remaining languages (Dutch, Finnish, Italian, Maltese and Polish) are held out entirely. This design allows us to assess whether phonological error patterns generalise beyond the languages used for supervision.

Phonemic reference transcriptions are generated automatically using the same `eSpeak`-based phonemisation pipeline as in Xu et al. (2022), providing a consistent approximation of phoneme-level supervision across languages.

Table 1 summarises key characteristics of the languages considered, including phoneme inven-

[1] We deliberately restrict our analysis to models that predict phonemic transcriptions directly, rather than graphemic or word-level outputs. This choice is methodological rather than practical: our goal is not to evaluate end-to-end ASR systems as deployed in real-world applications, but to analyse the phonological structure of recognition errors. In systems that operate on graphemes or words, observed errors conflate multiple sources of variation, including phonology, orthography, lexical constraints, and language-model effects, making it difficult to attribute confusions to phonological representations alone. Direct phoneme prediction provides a controlled setting in which substitution errors can be interpreted as confusions between phonological units, allowing feature-level analyses that would not be possible in fully integrated ASR pipelines.

tory size and phoneme entropy, computed over the empirical distribution of phonemes in the test data as $H = -\sum_{p \in \mathcal{P}} P(p) \log P(p)$, where $\mathcal{P}$ denotes the set of phonemes in the inventory and $P(p)$ the relative frequency of phoneme $p$. Across languages, inventory sizes range from 29 to 61 phonemes, while phoneme entropy varies within a relatively narrow interval.

In addition to the multilingual model, we consider a phoneme recognition system trained specifically for Thulung, an endangered Tibeto-Burman language spoken in Nepal, using expert-produced phonemic annotations. This provides a complementary evaluation setting in which reference transcriptions reflect human phonological analysis rather than automatic phonemisation, allowing us to verify that observed error patterns are not artefacts of grapheme-to-phoneme conversion. Full details of the Thulung data and training procedure are provided in Appendix A.

All models are used to automatically transcribe the evaluation utterances into sequences of phonemic symbols. Utterances are segmented and preprocessed in a consistent manner across languages, ensuring that differences in error patterns can be attributed to model behaviour and linguistic factors rather than to preprocessing artefacts.

We report the phoneme error rate (PER), computed as the Levenshtein distance between predicted and reference phoneme sequences, as a basic sanity check on recognition performance. PERs vary substantially across languages (Table 1), but all systems achieve non-trivial phoneme recognition accuracy, providing a sufficient empirical basis for analysing the *structure* of recognition errors rather than absolute performance.

Crucially, the edit-distance computation underlying PER allows us to decompose errors into substitutions, insertions, and deletions. We focus on phoneme-to-phoneme substitution errors, which correspond to genuine phonological confusions and form the basis of our analyses. Substitutions involving non-phonemic control symbols (e.g., padding or CTC blanks), as well as insertion and deletion errors, are excluded from the present study and left for future work.

### 2.2 Feature-Based Phoneme Representations

To analyse transcription errors produced by speech recognition models, we adopt a featural representation of phonemes. Phonological features are a core concept in phonology and one of its most enduring analytical tools. Their central assumption is that speech sounds are not atomic symbols, but structured objects defined by a set of underlying articulatory and acoustic properties.

In traditional phonological models, as originally introduced by Jakobson et al. (1963) and later formalised in Chomsky and Halle (1968), these properties are represented as binary features indicating whether a given characteristic is present or absent.[2] By encoding such shared properties, phonological features provide a compact and linguistically meaningful way to express both similarity and contrast between phonemes.

This notion of structured similarity is particularly well suited to the analysis of phonemic transcription systems. In standard evaluation settings, phonemes are treated as unrelated symbols: for example, the voiceless stop /p/ (as in ***s**pin*) and its aspirated counterpart /p$^h$/ (as in ***p**in*) are considered no more similar to each other than either is to a vowel such as /æ/ (as in ***a**pple*), even though /p/ and /p$^h$/ differ by only a small number of phonological features, primarily related to laryngeal properties such as aspiration, whereas /p/ and /æ/ differ along many feature dimensions simultaneously, including major class, manner of articulation, sonority, and vowel space properties. Feature-based representations make such graded differences explicit, allowing phonological similarity to be quantified in a way that aligns with linguistic intuition and distinguishes minor phonetic variation from genuinely different sound categories.

Importantly, phonological features are not unstructured. They are commonly organised into broader groupings or hierarchies, such as laryngeal features related to voicing and aspiration, place features describing where a sound is produced, and manner features describing how airflow is shaped, which are commonly argued to reflect shared physical mechanisms of speech production. These groupings play a central role in phonological patterning,

[2] For example, the phonemes /m/, /n/, and /ŋ/ (as in ***m**an*, ***n**o*, and *si**ng***) all involve nasal airflow and therefore share the feature `[nasal]`, forming a natural class. The phoneme /m/ also shares a labial place of articulation with /p/ and /f/, while patterning with /l/ and /r/ as a sonorant sound produced without turbulent airflow. More generally, phonological features encode such articulatory and acoustic properties (covering major class distinctions, manner and place of articulation, and vowel quality) which allow phonemes to be represented as structured feature bundles. These are precisely the properties we manipulate in the analyses below to quantify phonological similarity and to characterise how specific features are preserved or lost in recognition errors.

| | glottolog | Language | Language family | PER (%) | inventory size | entropy |
|---|---|---|---|---|---|---|
| nld | mode1257 | Dutch | Indo-European (Germanic) | 18.6 | 57 | 4.96 |
| eng | stan1293 | English | Indo-European (Germanic) | 6.3 | 61 | 5.16 |
| fin | nucl1717 | Finnish | Uralic | 2.1 | 58 | 4.75 |
| fra | stan1290 | French | Indo-European (Romance) | 17.5 | 55 | 4.90 |
| ind | indo1316 | Indonesian | Austronesian | 18.8 | 61 | 4.79 |
| ita | ital1282 | Italian | Indo-European (Romance) | 18.5 | 60 | 4.92 |
| mlt | malt1254 | Maltese | Afro-Asiatic (Semitic) | 12.9 | 46 | 4.59 |
| pol | poli1260 | Polish | Indo-European (Balto-Slavic) | 21.9 | 60 | 4.92 |
| swe | swed1254 | Swedish | Indo-European (Germanic) | 17.7 | 49 | 4.72 |
| tam | tami1289 | Tamil | Dravidian | 11.5 | 41 | 4.60 |
| tdh | thul1246 | Thulung | Sino-Tibetan | 3.1 | 29 | 4.33 |
| tur | nucl1301 | Turkish | Turkic | 21.7 | 54 | 4.87 |

Table 1: Phoneme error rate, phoneme inventory size (computed on the test set), and phoneme entropy by language.

including assimilation, neutralisation, and systematic asymmetries in sound change and error distributions (Clements and Hume, 1995).

In this work, we adopt the phoneme representations of Rubehn et al. (2024), implemented in the `soundvectors` Python package.[3] This resource defines a set of 39 speech-relevant phonological features designed to capture a broad range of segmental distinctions across languages. We choose this feature set because it provides a linguistically grounded yet computationally tractable representation that has been shown to support cross-linguistic phoneme modelling.

Not all features are instantiated in every dataset: in our experiments, a feature is considered *observed* if it takes at least one non-zero value for any phoneme attested in the test set of a given language. As a consequence, only 33 of the 39 defined features are effectively observed in our experiments. We therefore restrict our analyses to features that are attested in the test data, ensuring that all reported effects are grounded in empirically observed phonological contrasts rather than in abstract dimensions that are not realised in the evaluation material. The complete list of features defined in `soundvectors`, together with the subset observed in our test data, is reported in Table 4 (Appendix B).

This featural representation allows us to compute distances between phonemes and to analyse transcription errors in a graded rather than purely categorical manner. Crucially, it also enables us to investigate which feature groupings are most systematically involved in substitution errors, thereby providing a phonologically grounded interpretation of model behaviour.

[3] https://pypi.org/project/soundvectors/

## 3 Are phoneme confusions phonologically local?

### 3.1 Establishing Phonological Locality

This section investigates whether phoneme recognition errors are phonologically local, in the sense that substituted phonemes tend to be closer to their targets in phonological feature space than would be expected by chance. In this work, we focus exclusively on cases where a target phoneme is replaced by a different phoneme in the model output.[4] These errors are identified through the standard computation of PER, by aligning predicted phoneme sequences with gold reference transcriptions. Phonemes are represented as vectors of binary distinctive features, and phonological similarity is quantified using Hamming distance, defined as the number of features on which two phonemes differ. Smaller distances therefore correspond to greater phonological proximity. For each substitution error, we compute the distance between the target and the predicted phoneme, yielding a distribution of observed distances $d_{\text{obs}}$.

To establish a reference point, we construct a null distribution $d_{\text{random}}$ by sampling random pairs of phonemes within the same language. This baseline captures the degree of phonological similarity that would be expected in the absence of any systematic relationship between model confusions and phonological features.

To summarise the difference between observed and random distances, we compute $\Delta = d_{\text{obs}} - d_{\text{random}}$ and estimate uncertainty using 95 % intervals of compatibility (ICs), computed via non-parametric bootstrapping (Efron and Tibshirani,

[4] Phoneme-to-phoneme substitutions account for over 70 % of the observed errors in our data.

1993).[5] A negative value of $\Delta$ indicates that confusions are, on average, phonologically closer than expected by chance. Beyond mean differences, we also report the full distribution of observed distances to characterise the variability of phonological confusions.

To assess whether any observed effect is not only statistically detectable but also practically meaningful, we compute the probability of superiority $P_S = \mathbb{P}(d_{\text{obs}} < d_{\text{random}})$, which measures the probability that a randomly drawn observed substitution is phonologically closer than a randomly drawn baseline pair (McGraw and Wong, 1992; Vargha and Delaney, 2000). Values of $P_S$ close to $0.5$ indicate no systematic effect, while larger values indicate increasing degrees of phonological locality.

In addition, we conduct an equivalence analysis by defining a region of practical equivalence (ROPE) around zero in terms of feature differences. Following general recommendations for equivalence testing (Lakens, 2017), the size of the ROPE should correspond to differences that are negligible from a domain-specific theoretical perspective. In the context of distinctive feature representations, differences of one or two features correspond to minimal phonological contrasts and are commonly treated as phonologically minor (Miller and Nicely, 1955; Mielke, 2008). Accordingly, we consider ROPEs of $\pm 1$ feature differences.

All analyses are carried out both at the level of individual languages and on data aggregated across all languages, allowing us to assess the cross-linguistic consistency of phonological locality effects while identifying potential language-specific variation. Cross-linguistic aggregation is performed over effect sizes rather than raw distances, ensuring that languages with different phoneme inventories contribute comparably to the overall analysis.

### 3.2 Phonological Locality of Substitution Errors

Table 2 and Figure 1 report the results of the phonological locality analysis for phoneme substitution errors, computed separately for each language and aggregated across all languages. Across languages, substituted phonemes are consistently closer to their targets in phonological feature space than would be expected by chance. The average difference between observed and chance distances ranges from approximately two to four distinctive features, with all confidence intervals lying well outside a conservative region of practical equivalence of $\pm 1$ feature.

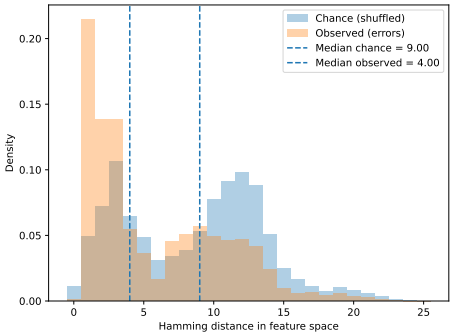


Figure 1: Distribution of phonological distances for phoneme substitution errors, pooled across all languages. Observed substitutions (blue) are compared to a chance baseline obtained by randomly shuffling predicted phonemes within the same phoneme inventory. Vertical dashed lines indicate the median distances for each distribution.

The aggregated analysis reveals a highly stable pattern. On average, observed substitutions differ from their targets by about four feature differences, whereas randomly paired phonemes differ by roughly nine features. This shift is reflected both in the mean difference ($\Delta \approx 2.8$) and in the full distribution of distances, whose median is reduced by more than half relative to the chance baseline. The corresponding probability of superiority ($P_S \approx 0.64$) indicates a moderate but robust effect: in nearly two thirds of cases, an observed substitution is phonologically closer to its target than a randomly paired phoneme. Given the size of the feature space and the absence of explicit phonological supervision, this represents a clear departure from chance rather than a marginal effect.

This pattern is consistent across languages with diverse phoneme inventories and typological properties. While the magnitude of the effect varies somewhat by language, no language shows evidence of substitutions being phonologically more distant than expected by chance.

Taken together, these results show that phoneme substitution errors produced by neural speech recognition models are strongly constrained in phonological feature space: Models predominantly confuse sounds that differ by only a small number of distinctive features, closely mirroring classic observa-

[5] As all Hamming distances are computed within each language, using a fixed set of attested features for that language, Observed and random distances are therefore directly comparable within languages.

| language | $d_{\text{obs}}$ | | $d_{\text{random}}$ | | $\Delta$ | $P_S$ | Practically meaningful |
|---|---|---|---|---|---|---|---|
| | mean | median | mean | median | | | |
| eng | 5.39 | 3 | 8.88 | 10 | $3.5 \pm 0.17$ | 0.68 | ✓ |
| fin | 4.64 | 3 | 7.42 | 6 | $2.8 \pm 0.3$ | 0.65 | ✓ |
| fra | 6.54 | 6 | 8.51 | 9 | $2.0 \pm 0.31$ | 0.58 | ✓ |
| ind | 4.72 | 3 | 7.46 | 7 | $2.7 \pm 0.094$ | 0.64 | ✓ |
| ita | 5.23 | 3 | 8.22 | 9 | $3.0 \pm 0.28$ | 0.61 | ✓ |
| mlt | 4.13 | 2 | 7.61 | 9 | $3.5 \pm 0.37$ | 0.66 | ✓ |
| nld | 5.97 | 4 | 8.94 | 10 | $3.0 \pm 0.33$ | 0.63 | ✓ |
| pol | 6.73 | 6 | 9.14 | 10 | $2.4 \pm 0.081$ | 0.61 | ✓ |
| swe | 5.12 | 3 | 7.95 | 9 | $2.8 \pm 0.32$ | 0.66 | ✓ |
| tam | 4.32 | 3 | 8.3 | 9 | $4.0 \pm 0.34$ | 0.71 | ✓ |
| tdh | 4.16 | 4 | 7.1 | 7 | $2.9 \pm 1.5$ | 0.49 | ✓ |
| tur | 4.83 | 3 | 7.54 | 8 | $2.7 \pm 0.3$ | 0.64 | ✓ |
| ALL | 5.75 | 4 | 8.56 | 9 | $2.8 \pm 0.05$ | 0.64 | ✓ |

Table 2: Phonological locality statistics for phoneme substitution errors by language and aggregated across all languages. We report the difference between observed and chance phonological distances ($\Delta$), 95% bootstrap intervals of compatibility, and the probability of superiority ($P_S$). *"Practically meaningful"* indicates whether the confidence interval of $\Delta$ lies entirely outside a region of practical equivalence of $\pm 1$ distinctive feature.

tions from human speech perception (Miller and Nicely, 1955). Importantly, this structure emerges despite the absence of any explicit phonological supervision or inductive bias during training: the models are not trained with access to phonological features, phoneme classes, or similarity relations. That phonological locality nevertheless arises suggests that self-supervised speech models implicitly capture aspects of linguistic structure that closely align with core notions of phonological similarity. Establishing this result is a crucial prerequisite for the feature-level analyses that follow, as it provides a principled basis for interpreting directional asymmetries in phonological feature transmission.

## 4 Identifying Weakly Modeled Phonological Features through Asymmetric ASR Errors

**Feature-level confusions** We now turn to a second question and examine whether phoneme recognition errors exhibit systematic asymmetries at the level of distinctive features. Such asymmetries make it possible to identify which phonological dimensions are more or less robustly encoded by the system. Beyond their engineering relevance, feature-level error patterns are also informative from a cognitive modelling perspective, as asymmetric feature confusions have long been documented in human speech perception and interpreted as reflecting differential robustness of phonological features under noise (Miller and Nicely, 1955; Cutler et al., 2004).

Let $p_i^{(t)}$ and $p_i^{(p)}$ denote the target and predicted phoneme a phoneme substitution event $i$, and let $f_j(p) \in \{0, 1\}$ denote the value of feature $j$ for phoneme $p$. As in our previous analysis, we restrict attention to phoneme substitution errors.

Within a substitution error, two types of feature-level events may occur for a given feature $j$: a *feature deletion*, when the feature is present in the target but absent in the prediction, and a *feature insertion*, when the feature is absent in the target but present in the prediction. We estimate the corresponding conditional probabilities

$$P_j^{\text{del}} = P\big(f_j(p^{(p)}) = 0 \mid f_j(p^{(t)}) = 1,\ \text{error}\big), \tag{1}$$

$$P_j^{\text{ins}} = P\big(f_j(p^{(p)}) = 1 \mid f_j(p^{(t)}) = 0,\ \text{error}\big), \tag{2}$$

which quantify the tendency of a feature to be lost or added once a recognition error has occurred.

All probabilities are estimated separately for each language, using substitution errors observed in that language. We then summarise the direction and magnitude of feature-level asymmetry using the asymmetry measure

$$A_j = P_j^{\text{ins}} - P_j^{\text{del}}. \tag{3}$$

Positive values of $A_j$ indicate insertion-dominant asymmetries (over-generation), negative values indicate deletion-dominant asymmetries, and values close to zero correspond to approximately symmetric transmission. Cross-linguistic aggregation is performed at the level of the asymmetry measure $A_j$, using a random-effects meta-analytic model.

Uncertainty is quantified using non-parametric bootstrapping over substitution errors. For each feature, we report a 95% interval of compatibility (IC) for $A_j$. To distinguish meaningful asymmetries from negligible ones, we adopt a ROPE around zero, set to $[-0.05, 0.05]$. Asymmetries whose 95% IC lies entirely outside the ROPE are interpreted as practically meaningful, whereas those whose IC lies entirely within the ROPE are considered negligible. Partial overlap leads to an inconclusive interpretation.[6]

In practice, probabilities are estimated using empirical proportions over substitution errors, conditioning only on informative cases: deletion probabilities are computed from instances in which the feature is present in the target phoneme, and insertion probabilities from instances in which it is absent. This conditioning ensures that features are evaluated only when they are logically eligible to be deleted or inserted.

Because some phonological features are rare or highly specific, naïvely estimating $A_j$ can yield unstable values driven by sparsity rather than by systematic error patterns. We therefore restrict the analysis to features that are sufficiently supported in the data.[7] Concretely, we exclude feature–language combinations with too few eligible observations and retain only features that meet these support requirements in a sufficient number of languages.

We refer to as *core features* those phonological features that satisfy two empirical criteria: (i) they are attested in at least 11 of the 12 languages considered, and (ii) for each retained language, both deletion and insertion probabilities are estimated from at least $N_{\min} = 50$ eligible substitution events. These criteria yield a set of broadly shared and well-supported features, which form the basis of the cross-linguistic meta-analysis reported below.[8]

Before turning to the results, we quantify between-language heterogeneity in feature-level asymmetries using the $I^2$ statistic. Following standard definitions (Cochran, 1954; Higgins and Thompson, 2002), $I^2$ measures the proportion of total variance in observed effect sizes that is attributable to genuine cross-linguistic differences rather than to sampling error. In the present setting, high $I^2$ values do not reflect inconsistent directions of asymmetry across languages, but rather substantial variation in their magnitude.

**Results.** Figure 2 reports cross-linguistically aggregated feature-level asymmetries for core features, with uncertainty summarised by 95% IC and practical relevance assessed against the ROPE. Many features exhibit clear directional biases (intervals outside the ROPE), indicating that substitution errors are not symmetric noise but are systematically structured in feature space. For a number of features, $I^2$ is substantial, suggesting that the *magnitude* of these asymmetries varies across languages even when their direction is broadly consistent.

Two broad tendencies emerge. First, several relatively coarse-grained properties show insertion-dominant asymmetries: [CONTINUANT] and [SONORANT] (and, to a lesser extent, [SYLLABIC] and [FRONT]) are more likely to be added to the predicted phoneme than deleted from the target once an error occurs. This pattern points to a bias towards acoustically continuous and vowel-like outcomes under recognition failure. Second, more fine-grained specifications tend to be deleted rather than inserted. In particular, place-related features (e.g., [CORONAL], [ANTERIOR]) and vowel-quality contrasts (e.g., [HIGH], [LOW], [BACK]) display deletion-dominant asymmetries, consistent with a systematic loss of articulatory and phonological specificity in erroneous predictions. Taken together, these results suggest a structured pattern of feature degradation: when the system fails to identify a phoneme correctly, broad class-level properties are relatively preserved (or even over-generated), while finer place and vowel-quality distinctions are disproportionately lost.

Finally, a small set of highly specific dimensions (notably segment length and diphthong-trajectory features) exhibits extremely strong deletion-dominant asymmetries. Because these effects are supported by limited and uneven empirical support for insertion vs. deletion events and restricted to a small subset of segments, we analyse them separately and report full results in Appendix C.

At a higher level, these results suggest that phoneme recognition errors reflect a structured

[6] Because inference is based on effect sizes, intervals of compatibility, and a theoretically defined ROPE rather than on null-hypothesis significance testing, no correction for multiple comparisons is required (Lakens, 2017).

[7] Empirical support refers to the number of substitution events in which a feature is logically eligible to be inserted or deleted.

[8] Extending the analysis to a broader set of phonological features would require either increasing the number of languages considered or substantially enlarging the amount of annotated speech data per language, in order to provide sufficient support for rare or highly specific features.

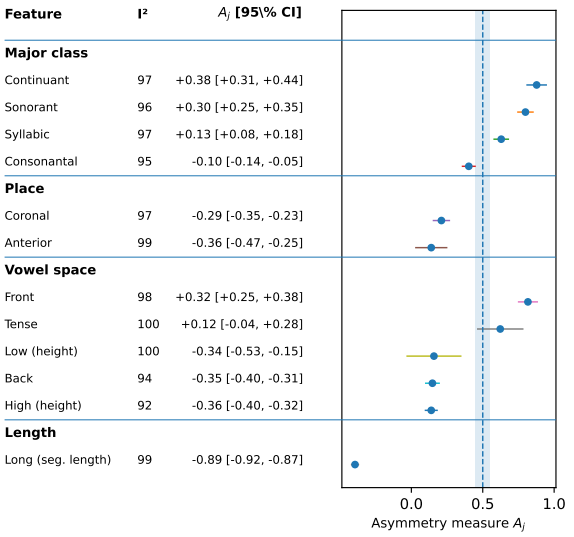


Figure 2: Feature-level asymmetries in phoneme substitution errors for core phonological features. Points and horizontal bars indicate random-effects meta-analytic estimates of the asymmetry measure $A_j$ and their 95% intervals of compatibility across languages. The shaded region denotes the region of practical equivalence (ROPE, $[-0.05, 0.05]$), and the dashed vertical line marks zero asymmetry. $I^2$ reports between-language heterogeneity.

form of phonological degradation rather than unstructured noise. When recognition fails, the system tends to preserve broad, acoustically salient properties, such as sonority or vocalic structure, while disproportionately losing finer-grained articulatory and phonological specifications, including detailed place and vowel-quality distinctions. This pattern points to an implicit hierarchy of phonological information in the learned representations, in which coarse-grained dimensions are more robustly encoded than fine-grained ones. Crucially, this hierarchy emerges despite the absence of any explicit phonological supervision during training, indicating that self-supervised speech models naturally capture non-uniform phonological structure from the statistical regularities of the acoustic signal.

## 5 Discussion and Conclusion

This paper investigated the phonological structure of errors produced by neural phoneme recognition systems through a feature-based analysis of phoneme confusions. Rather than treating phonemes as unrelated symbols, we adopted a representation in terms of distinctive features and showed that recognition errors exhibit systematic and interpretable phonological regularities.

Our first analysis demonstrated that phoneme substitution errors are phonologically local: substituted phonemes are significantly closer to their targets in feature space than would be expected by chance, both within individual languages and cross-linguistically. Crucially, this structure emerges even though the models are trained without any explicit access to phonemic representations or information about the internal structure of phonemes. This result indicates that self-supervised speech models can acquire representations that preserve substantial aspects of phonological organisation solely from distributional and acoustic learning signals.

We then examined feature-level asymmetries conditional on the occurrence of an error. This analysis revealed robust directional biases in the transmission of distinctive features: broad class-level and sonority-related features tend to be preserved or over-generated, while fine-grained place, manner, and secondary articulatory features are systematically lost. These asymmetries closely mirror long-standing findings from human speech perception, where perceptual confusions are strongly structured by phonological similarity and feature robustness under noise is highly uneven (Cutler et al., 2004).

Taken together, these results show that phoneme recognition errors are not arbitrary, but structured in ways that reflect the internal organisation of phonological representations. More importantly, they suggest that self-supervised learning of speech can give rise to representational biases that are compatible with those posited in linguistic theory and observed in human listeners, despite the absence of any explicit linguistic supervision during training.

An important direction for future work is to move beyond analyses that are conditional on the occurrence of an error and to directly model the vulnerability of phonological features, that is, the extent to which the presence of specific features in the target phoneme increases the probability that a recognition error occurs in the first place. Addressing this question will require token-level modelling of error likelihood while controlling for speaker-, utterance-, and acoustic-level factors.

More broadly, integrating phonological feature theory with modern self-supervised speech models offers a principled framework for analysing error patterns, comparing systems across languages, and drawing meaningful connections between machine and human speech processing.

## Ethical Considerations

This work analyses phoneme recognition errors produced by neural speech models through a phonologically grounded, feature-based framework. It is a diagnostic and analytical study: we do not introduce new speech recognition systems for deployment, nor do we propose changes to model architectures or training objectives that would directly affect end-user applications.

All experiments rely on existing datasets and pretrained models. The multilingual speech data used in this study come from publicly available resources (`Common Voice` and `Babel`) that were collected under established ethical guidelines. The Thulung data were obtained from the Pangloss Collection, an open archive for newly documented and underdocumented languages, and are used in accordance with the archive's licensing and data-sharing policies. No new data collection involving human participants was conducted as part of this work.

From an ethical perspective, a key motivation of this study is to improve transparency and interpretability in speech recognition systems. By characterising systematic phonological structure and asymmetries in recognition errors, our analysis aims to support more informed model diagnosis and evaluation, particularly in multilingual and low-resource settings. Understanding which phonological features are more vulnerable under recognition failure can help identify biases and limitations in current systems, rather than obscuring them behind aggregate performance metrics.

At the same time, we acknowledge that phoneme-based analyses do not capture all dimensions of speech variation, including sociophonetic variation, speaker identity, or language-specific phonological norms. Care should therefore be taken not to overgeneralise feature-level findings to individual speakers or communities. We emphasise that the results describe model behaviour under controlled evaluation conditions and should not be interpreted as normative claims about human speech or language use.

Overall, we believe that feature-based analyses of ASR errors contribute positively to responsible speech technology research by promoting interpretability, cross-linguistic comparability, and critical examination of model behaviour, rather than by enabling new forms of surveillance or automated decision-making.

## Limitations

This study has several limitations that define the scope of its conclusions. First, our analyses characterise the structure of phoneme recognition errors conditional on the occurrence of an error, rather than modelling the probability that an error occurs as a function of phonological features. As a result, the reported asymmetries describe how phonological features are preserved or lost once recognition fails, not their absolute vulnerability during normal recognition. Modelling feature-level vulnerability directly would require token-level analyses that control for acoustic, speaker, and contextual factors.

Second, the analysis is restricted to speech recognition systems that predict phonemic transcriptions directly. While this choice provides a controlled and interpretable setting for studying phonological structure, it limits the immediate applicability of the findings to end-to-end ASR systems whose outputs conflate phonological, orthographic, and lexical constraints.

Third, all experiments are conducted using a single self-supervised, Transformer-based phoneme recognition model and training paradigm. Although this model is representative of current approaches to phoneme-level ASR, we do not assess the extent to which the observed error patterns and feature-level asymmetries generalise across different architectures, training objectives, or model sizes. Evaluating the stability of these phonological patterns across a broader range of models remains an important direction for future work.

Fourth, feature-level results depend on the adopted phonological feature representation. We rely on the `soundvectors` feature inventory, which is linguistically grounded and designed for cross-linguistic modelling, but alternative feature systems or gradient representations could yield different quantitative patterns. Our conclusions should therefore be interpreted as conditional on this representational choice.

In addition, although the study spans multiple language families, the number of languages considered remains limited and does not cover the full diversity of phonological systems attested cross-linguistically. Highly specific or low-frequency phonological features, such as fine-grained temporal or diphthongal properties, could not be included in the primary cross-linguistic analysis due to limited empirical support and are therefore analysed separately.

Finally, while the observed error asymmetries parallel patterns reported in human speech perception, the present work does not aim to model human perceptual mechanisms. Similarities should be interpreted as indicating representational compatibility rather than cognitive equivalence.

## References


Evangelia Adamou, Séverine Guillaume, and Alexis Michaud. 2025. The Pangloss Collection: Opening up research data on endangered and underdocumented languages. *Language*, 101(1):e38–e59.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row.

G. N. Clements and Elizabeth Hume. 1995. The internal organization of speech sounds. In John Goldsmith, editor, *The Handbook of Phonological Theory*. Blackwell.

William G. Cochran. 1954. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. 2004. Patterns of english phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, 116(6):3668–3678.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*, pages 1402–1406.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of japhug (trans-himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Julian P. T. Higgins and Simon G. Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558.

Roman Jakobson, Gunnar Fant, and Morris Halle. 1963. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. MIT Press.

Daniël Lakens. 2017. Equivalence tests: A practical primer for $t$ tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Kenneth O. McGraw and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

Jeff Mielke. 2008. *The Emergence of Distinctive Features*. Oxford University Press.

George A. Miller and Patricia E. Nicely. 1955. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Arne Rubehn, Jessica Nieder, Robert Forkel, and Johann-Mattis List. 2024. Generating feature vectors from phonetic transcriptions in cross-linguistic data formats. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 205–216, Irvine, CA. Association for Computational Linguistics.

Andras Vargha and Harold D. Delaney. 2000. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022*, pages 2113–2117.


## A Development of a Phoneme-Based ASR System for Thulung

We develop an automatic speech recognition system for Thulung, a Tibeto-Burman language spoken in eastern Nepal. From a phonetic perspective, Thulung exhibits a relatively rich consonant inventory, including aspirated and unaspirated stops, a

contrastive voicing system, and a set of vowel qualities that are largely stable across contexts, making it a suitable test case for phoneme-level modeling. Thulung is a language currently undergoing documentation, and annotated speech data are available through the Pangloss Collection, an open archive dedicated to newly documented and underdocumented languages (Adamou et al., 2025). The corpus is transcribed using a transparent phonemic orthography, and a complete phoneme inventory is provided, which allows us to train an ASR system that directly predicts phoneme sequences rather than graphemic characters, an important distinction given that some phonemes are represented by multicharacter sequences in practical orthographies.

The dataset contains approximately 7 hours of transcribed speech[9]. We use 80 % of the available data to fine-tune a pretrained XLSR-53 model following the methodology of Guillaume et al. (2022), while the remaining data are split between a validation set and a held-out test set whose size is comparable to those used for other languages in our experiments. All hyperparameters used for fine-tuning are reported in Table 3.

| Hyperparameter | Value |
|---|---|
| Training batch size | 16 |
| Gradient accumulation steps | 8 |
| Number of epochs | 60 |
| Learning rate | $3 \times 10^{-4}$ |
| Warm-up steps | 500 |
| Mixed precision (FP16) | Enabled |
| Evaluation strategy | Steps |
| Evaluation frequency (eval_steps) | 50 |
| Checkpoint saving frequency (save_steps) | 100 |
| Logging frequency (logging_steps) | 50 |
| Maximum number of saved checkpoints | 2 |
| Optimiser | AdamW |
| Audio sampling rate | 16 kHz |

Table 3: Main hyperparameters used for fine-tuning Wav2Vec2

## B Phonological Feature Inventory

Table 4 reports the phonological features used in our analyses, as defined by the `soundvectors` representation, and indicates which of these features are effectively instantiated in the test-set phoneme inventory. A feature is considered *observed* if it takes a non-zero value for at least one phoneme occurring in the test data. The table is organised hierarchically into broad phonological groupings (such as major class, manner, laryngeal, place, and vowel space) to make explicit the structural relationships between features. This organisation helps clarify which phonological dimensions are available for analysing transcription errors and which distinctions are actually present in the evaluation material.

**Feature Classes** For clarity and interpretability, phonological features are grouped into a small number of broad classes reflecting well-established dimensions of phonological description. *Major class* features distinguish between consonantal, syllabic, and sonorant sounds, capturing coarse-grained differences in segment type. *Manner* features describe how airflow is shaped during sound production, differentiating, for instance, stops, nasals, laterals, and strident consonants. *Laryngeal* features encode properties related to the state of the glottis, such as voicing and aspiration, which are central to many phonological contrasts.

*Place* features specify where constrictions are formed along the vocal tract, including labial, coronal, and dorsal articulations, as well as more posterior regions. *Vowel space* features characterise vowel quality in terms of height, backness, rounding, and tenseness. In addition, we distinguish *length* features, which capture segmental duration contrasts, and *tone/contour* features, which encode pitch-related distinctions such as register and contour. Finally, *diphthong trajectory* features describe dynamic changes within a segment, capturing directional movements in vowel quality over time. These groupings are intended as an organisational device to facilitate analysis and do not presuppose a strict hierarchical structure among features.

## C Feature-Level Asymmetries for Non-Core Features

In addition to the core feature set analysed in Section 4, we examined a number of more specialised phonological features whose empirical support was insufficient for inclusion in the cross-linguistic meta-analysis. These *non-core* features primarily concern segmental length and diphthong trajectory properties, which are instantiated in a limited subset of phonemes and languages and are therefore associated with limited and uneven empirical support

[9] Additional annotated recordings have recently been collected as part of a new fieldwork campaign; however, these data were not included in the present experiments.

| Feature | Short name | Observed | |
|---|---|---|---|
| *Major class* | | | |
| | Consonantal (consonant-like constriction) | cons | ✓ |
| | Syllabic | syl | ✓ |
| | Sonorant | son | ✓ |
| | Continuant | cont | ✓ |
| *Manner* | | | |
| | Delayed release (affrication) | delrel | ✓ |
| | Lateral | lat | ✓ |
| | Nasal | nas | ✓ |
| | Strident | strid | ✓ |
| *Laryngeal* | | | |
| | Voiced | voi | ✓ |
| | Spread glottis (aspiration-related) | sg | ✓ |
| | Constricted glottis (glottalisation-related) | cg | ✓ |
| | Laryngeal class marker | laryngeal | ✓ |
| *Place* | | | |
| | Labial | lab | ✓ |
| | Coronal | cor | ✓ |
| | Dorsal | dorsal | ✓ |
| | Anterior | ant | ✓ |
| | Distributed | distr | ✓ |
| | Pharyngeal | pharyngeal | ✓ |
| | Velaric (click-related) | velaric | ✓ |
| *Vowel space* | | | |
| | High (vowel height) | hi | ✓ |
| | Low (vowel height) | lo | ✓ |
| | Back | back | ✓ |
| | Front | front | ✓ |
| | Rounded | round | ✓ |
| | Tense | tense | ✓ |
| *Length* | | | |
| | Long (segment length) | long | ✓ |
| *Tone / contour* | | | |
| | High tone | hitone | ✗ |
| | High register | hireg | ✗ |
| | Low register | loreg | ✗ |
| | Rising contour | rising | ✗ |
| | Falling contour | falling | ✗ |
| | Contour tone marker | contour | ✗ |
| *Diphthong trajectory* | | | |
| | Back shift (diphthong trajectory) | backshift | ✓ |
| | Front shift (diphthong trajectory) | frontshift | ✓ |
| | Opening diphthong | opening | ✓ |
| | Closing diphthong | closing | ✓ |
| | Centering diphthong | centering | ✓ |
| | Long-distance trajectory | longdistance | ✓ |
| | Second element rounded | secondrounded | ✓ |

Table 4: Phonological features `soundvectors` and whether they are observed in our test-set phoneme inventory.

for insertion vs. deletion events.

Figure 3 reports asymmetry measures for these features, estimated following the same procedure as for the core analysis but without cross-linguistic aggregation constraints. Across features, a consistent qualitative pattern emerges: non-core features exhibit extremely strong deletion-dominant asymmetries, indicating that once a recognition error occurs, such properties are almost systematically lost rather than spuriously inserted.

These effects are substantially larger in magnitude than those observed for core features. However, they are also supported by a small number of eligible observations and by distributions that are highly skewed towards deletion events. As a result, while the direction of the effect is stable, the corresponding estimates are not comparable in robustness to those reported for the core feature set.

Importantly, the behaviour of non-core features is fully consistent with the general interpretation advanced in the main text. Highly specific and phonetically complex properties—such as fine-grained temporal structure or dynamic vowel trajectories—appear particularly fragile under recognition failure. When the system errs, these dimensions are preferentially eliminated, reinforcing the picture of a structured degradation process in which increasingly detailed phonological specifications are progressively lost.

For these reasons, non-core features are reported separately and are not included in the primary cross-linguistic meta-analysis.

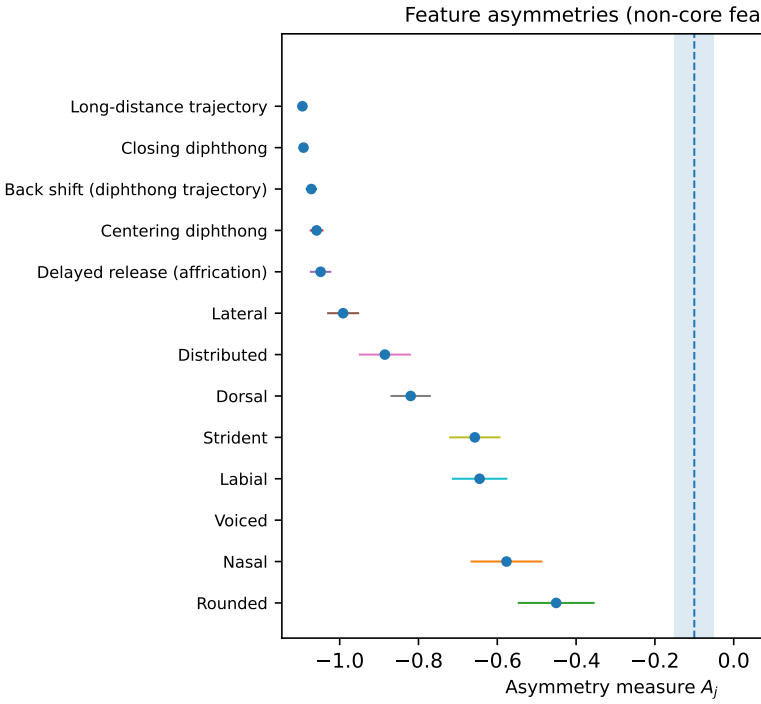


Figure 3: Feature-Level Asymmetries for Non-Core Features