Emphasis Sensitivity in Speech Representations

Shaun Rafael Cassini, Thomas Hain, Anton Ragni
University of Sheffield, Sheffield, UK
{srcassini1, t.hain, a.ragni}@sheffield.ac.uk

Abstract—**This work investigates whether modern speech models are sensitive to prosodic emphasis—whether they encode emphasized and neutral words in systematically different ways. Prior work typically relies on isolated acoustic correlates (e.g., pitch, duration) or label prediction, both of whichwhich miss the relational structure of emphasis. This paper proposes a residual-based framework, defining emphasis as the difference between paired neutral and emphasized word representations. Analysis on self-supervised speech models shows that these residuals correlate strongly with duration changes and perform poorly at word identity prediction, indicating a structured, relational encoding of prosodic emphasis. In ASR fine-tuned models, residuals occupy a subspace up to 50% more compact than in pre-trained models, further suggesting that emphasis is encoded as a consistent, low-dimensional transformation that becomes more structured with task-specific learning.**

Index Terms—**emphasis, prosody, speech representations, self-supervised speech, speech understanding, representation analysis**

I. Introduction

Speech conveys much more than words, as it carries information about the speaker, their mood, and communicative intent. In particular, speakers use emphasis to highlight specific words or phrases, conveyed through a combination of prosodic cues such as pitch, duration, and loudness. Prior studies show that prosody in speech enables a listener to recover cues that signal the communicative function of an utterance [1]. Emphasis serves a range of communicative functions, including marking contrast, highlighting information structure, and resolving syntactic ambiguity which words alone may not express [2]–[4]. Automatic speech processing systems that are sensitive to emphasis cues are known to perform better on tasks ranging from intent prediction [5], speech translation [6], to text-to-speech synthesis (TTS) [7]. Yet it remains unclear to what extent emphasis is implicitly learned by such systems.

Emphasis is expressed in several prosodic cues (acoustic correlates) and is suprasegmental, spanning multiple speech segments in an utterance [8], [9]. Its realization is known to vary by speaker, utterance, language, and dialect [10], [11]. Cue-specific models which extract acoustic correlates such as the fundamental frequency F_0 can perform well when emphasis aligns with that cue [12]. However, they may miss instances where emphasis is conveyed through under-modeled cues [13]. For instance, post-focal compression refers to a reduction in prosodic cues on segments following an emphasized word. This representation of emphasis requires modeling context that extends beyond the emphasized word itself [14]. Approaches based on local acoustic correlates would be blind to such cues.

Given the limitations of cue-specific approaches and the distributed nature of emphasis, recent work focuses on supervised models with trainable parameters that map the speech signal to emphasis labels. Some approaches make direct use of the waveform [15], [16], while others jointly learn acoustic correlates like F_0 or spectral energy [17], [18]. Such supervised approaches require emphasis labels, yet the occurrence of emphasis in natural speech is scarce and labeling data is highly subjective—shaped by the same perceptual ambiguities they aim to model [19]. This motivates the question: *To what extent do modern speech models, trained without supervision for emphasis, implicitly encode it?*

Prior work has examined acoustic correlates of emphasis or trained classifier probes to predict emphasis labels on individual words [15], [20]. Both strategies ignore the fact that emphasis is inherently relational: a word sounds prominent only relative to how it would sound without emphasis (neutral) and to the prosodic context around it [21]. This work therefore probes for emphasis sensitivity in the residual space between representations of paired neutral-and-emphasized words. An analysis is conducted to assess whether the residual space encodes emphasis as a consistent, learnable relationship rather than a property of isolated words.

In this work, a residual space is derived from representations extracted from multiple self-supervised speech learning models (S3L models) [22] and their fine-tuned variants, specifically those fine-tuned for automatic speech recognition (ASR) and emphasis classification. Both model types capture prosodic cues in their representations [23], [24], making them strong candidates for investigating whether emphasis sensitivity arises implicitly across distinct objectives. This also enables a direct comparison between S3L models and their fine-tuned counterparts to assess how training objectives shape their emphasis sensitivity. Experiments are conducted on 3,732 word pairs derived from a synthetic dataset designed for emphasis control [15], comprising contrastive pairs of utterances with emphasized and neutral words. Representations are extracted from multiple state-of-the-art S3L models and their fine-tuned variants. Through analysis of residual vectors between representations derived from neutral–emphasized word pairs, this work finds that emphasis is encoded as a low-dimensional, consistent transformation that becomes more pronounced in fine-tuned models. The contributions of this work are as follows:

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications, funded by UK Research and Innovation [grant number EP/S023062/1], with additional support from Huawei Research & Development (UK). We thank Nicola Mendini and Mattias Cross for their proofreading and helpful discussions.

- A novel framework for quantifying emphasis sensitivity in speech models as a structured relationship between emphasized and neutral word pairs.
- A residual-based probing analysis to isolate and measure prosodic variation in representation space.
- Experiments showing that S3L models exhibit structured, layer-dependent emphasis sensitivity, which becomes stronger and more consistent after ASR finetuning.
- Evidence that residuals capture the shift from neutral to emphasized words, occupy a significantly lower-dimensional subspace, and correlate well with duration change.
- A geometric interpretation of how emphasis is encoded across architectures, offering insights for emphasis-aware model development.

II. Related Work

In practice, sensitivity to emphasis has been shown to benefit a variety of downstream tasks, including intent prediction [5], emotion recognition [25], [26], ASR [27], naturalistic transcription [28], [29], voice conversion [30], speech segmentation [12], [31], speech translation [6], [32]–[34], human-machine dialogue [35]–[37], TTS [7], [38], [39], and assisting with language learning [40]–[42]. For instance, modeling pitch-based emphasis cues reduced the word error rate of an HMM-based ASR system on the Boston News Corpus by 11% relative to prosody-independent systems [27].

Explicit approaches to emphasis detection rely on supervised learning or acoustic cue extraction. These include classifiers trained on prosodic correlates such as pitch, duration, and energy [17], [18] or on direct waveform inputs [15], [16]. Some also incorporate explicit theoretically grounded feature engineering, e.g., F_0 contours or post-focal compression effects [9], [12].

1) Representation Space Analysis: Recent studies have shown that S3L models, such as wav2vec 2.0 [43], HuBERT [44], and WavLM [45] encode information about prosodic structure, which includes prominence and intonation [23], [24], [46]. However, such findings often rely on supervised classifier probes trained to map intermediate representations to prosodic labels [20], [47]. Such probes introduce learnable parameters that risk overstating or misattributing what is encoded versus what is merely decodable [48].

2) Analysis of Residual Spaces: Since this work probes for emphasis sensitivity in residual representation space, it is useful to consider how residual analysis has been applied in related domains. In anomaly detection, PCA-based residual analysis is used under the assumption that structured data lie in a low-dimensional subspace, while residuals represent deviations or noise [49]. PCA separates signal components that are compressible from those that are distributed or unstructured. In contrast, this work does not treat residuals as anomalies, but instead asks whether the residuals themselves reflect a consistent transformation by exhibiting low-rank structure.

A related framework is residual component analysis (RCA) [50], which models structure in the residual covariance after accounting for known variation. RCA has been used to recover latent dynamics such as skeletal motion from residuals in motion capture data. While this work does not adopt RCA's probabilistic formulation, it shares the underlying view that residuals can encode meaningful, interpretable structure; the perspective applied to prosodic emphasis in speech in this proposed analysis.

III. Experimental Setup & Methods

A. Data

1) Synthetic Emphasis Dataset: The dataset used in this work is derived from the EmphAssess evaluation dataset, a benchmark for evaluating emphasis preservation in speech-to-speech models [15]. EmphAssess is comprised of variants of 299 sentences, with each variant changing which word is emphasized, as shown by the example in Figure 1. This yields 913 unique sentence variations. The sentences are synthesized with four American TTS voices (2 male, 2 female), yielding 3,652 short utterances (2.42 hours).

1) *"**The** dishonest politician who admits it?"*
2) *"The dishonest **politician** who admits it?"*
3) *"The dishonest politician who **admits** it?"*

Fig. 1: Example sentence from the EmphAssess dataset, with neutral words underlined and emphasized words in bold.

There are 546 unique neutral–emphasized word pairs. Of the 13,108 total words instances across all speakers and transcripts, 3,796 (0.52 hours) are emphasized and 9,312 (0.80 hours) are neutral. Analysis is conducted on representations of these words, as explained in the following section.

2) Deriving Word-Level Representations: As this work examines word–word comparisons, the following describes the method used to obtain word-level representations. First, model outputs are aligned to time-stamped word boundaries, following the procedure described in [51]. Word-level time boundaries are obtained using the Montreal Forced Aligner (MFA) [52]. All frames associated with a specific word are averaged to obtain a representation at each encoder layer, denoted by $\mathbf{z}_{i,j}^{(l)} \in \mathbb{R}^d$, where l indexes the encoder layer, i the utterance, j the word, and d the dimensionality of the layer's output (which is the same across all layers). The duration values for each word, denoted $d_{i,j} = t_{i,j}^{\text{end}} - t_{i,j}^{\text{start}}$, are also retained for further analysis.

3) Neutral–Emphasized Contrastive Pairing: To assess the emphasis sensitivity of representations while controlling for contextual and speaker-dependent factors, a contrastive pairing set is constructed in which each pair comprises one emphasized and one neutral word representation. Pairs are sampled from the dataset such that speaker, word, and transcript identity are matched, differing only in emphasis label. This yields 3,732 aligned neutral–emphasized pairings.

B. Representation Analysis

1) Sample-Wise Cosine Similarity: To evaluate how emphasis affects word representations, the distribution of cosine similarities between samples is analyzed. For each pair of aligned emphasized and neutral words, the cosine similarity between their representations is computed. The neutral–emphasized pairwise similarities are compared with neutral–neutral similarity baselines. If the emphasized and neutral variants of the same word are nearly identical ($\cos(\theta) \approx 1$), this suggests that emphasis has little effect. If they are consistently less similar, this may indicate a systematic prosodic shift. Analyzing the distribution of these similarities across the dataset provides an interpretable measure of the sensitivity of the model's representations to emphasis. The means of these distributions are reported as summary metrics, denoted by θ_{AA} (neutral–neutral), θ_{BB} (emphasized–emphasized), and θ_{AB} (neutral–emphasized).

In addition, the distribution of cosine similarities between all unique residual pairs, $\mathbf{R} = \mathbf{B} - \mathbf{A}$, is analyzed. Its mean, denoted θ_{RR}, is equivalent to the metric defined in [53]:

$$\theta_{RR} = \frac{1}{2N(N-1)} \sum_{i<j} \cos(\mathbf{r}_i, \mathbf{r}_j) \tag{1}$$

where $\mathbf{r}_i = \mathbf{b}_i - \mathbf{a}_i$ is the residual vector for the i-th pair. This metric captures the second-order structure of the residuals, quantifying whether emphasis transformations encoded by the model are directionally consistent across different word instances. However, because θ_{RR} involves comparisons over all residuals, it averages over potentially diverse lexical and speaker identities and may include variation uncorrelated with emphasis.

To reduce the impact of such variance, a first-order directional consistency metric, denoted $\theta_{\hat{R}}$, is used, defined as the cosine similarity between each residual $\mathbf{r}_i$ and the mean residual vector $\bar{\mathbf{r}}$:

$$\theta_{\hat{R}}^{(i)} = \cos(\mathbf{r}_i, \bar{\mathbf{r}}), \quad \bar{\mathbf{r}} = \frac{1}{N} \sum_i \mathbf{r}_i \tag{2}$$

This reflects how well each individual transformation aligns with the average emphasis direction.

2) Dimension-Wise Variance via PCA: To complement the sample-wise analysis, the variance across representation dimensions is examined. For this, Principal Component Analysis (PCA) [54] is applied to the following representation spaces:

- Neutral word representations $\mathbf{A} \in \mathbb{R}^{N \times d}$
- emphasized word representations $\mathbf{B} \in \mathbb{R}^{N \times d}$
- Concatenated representations $\mathbf{C} = [\mathbf{A} \mid \mathbf{B}] \in \mathbb{R}^{N \times 2d}$
- Residual vectors $\mathbf{R} = \mathbf{B} - \mathbf{A} \in \mathbb{R}^{N \times d}$

With λ_i denoting the eigenvalue corresponding to the i-th principal component (PC), the explained variance ratio is defined as:

$$v_i = \frac{\lambda_i}{\sum_j^d \lambda_j} \tag{3}$$

The effective dimensionality, $D_{95\%}$, is defined as the number of PCs needed to explain at least 95% of the total variance:

$$D_{95\%} = \min \left\{ k : \sum_{i=1}^{k} v_i \geq 0.95 \right\} \tag{4}$$

A higher $D_{95\%}$ in $\mathbf{C}$ than in either $\mathbf{A}$ or $\mathbf{B}$ suggests that emphasis introduces additional structured variation in representation space, potentially aligned with a prosodic axis. Additionally, a low $D_{95\%}$ in $\mathbf{R}$ implies that the transformation from neutral to emphasized representations lies in a low-dimensional subspace, indicating that emphasis is encoded consistently across samples (generalized) rather than as a unique variant of each sample (memorized).

3) Midpoint Centering: PCA typically involves mean-centering the data, which removes any global offset in the covariance estimate. However, when applied to residual vectors $\mathbf{r}_i = \mathbf{b}_i - \mathbf{a}_i$, mean-centering alters the interpretation of the resulting PCs. Let $\bar{\mathbf{r}} = \frac{1}{N} \sum_i \mathbf{r}_i$ denote the mean residual vector. The centered residual is then:

$$\tilde{\mathbf{r}}_i = \mathbf{r}_i - \bar{\mathbf{r}} = (\mathbf{b}_i - \bar{\mathbf{b}}) - (\mathbf{a}_i - \bar{\mathbf{a}}) \tag{5}$$

This effectively centers each group ($\mathbf{A}$ and $\mathbf{B}$) independently, eliminating the global offset between the emphasized and neutral representations. As a result, centering would remove the very structure under investigation. Hence, the sets $\mathbf{A}$ and $\mathbf{B}$ are *midpoint*-centered prior to analysis. For each sample,

$$\hat{\mathbf{a}}_i = \mathbf{a}_i - \mathbf{m}, \quad \hat{\mathbf{b}}_i = \mathbf{b}_i - \mathbf{m}, \tag{6}$$

where $\mathbf{m} = \frac{1}{2}(\bar{\mathbf{a}} + \bar{\mathbf{b}})$.

4) Reconstructing Duration Change from Residual Geometry: To test whether the residuals $\mathbf{r}_i = \mathbf{b}_i - \mathbf{a}_i$ encode interpretable prosodic transformations, a regression task is used to reconstruct relative word-level duration change, as it is a known acoustic correlate and proxy of emphasis [11]. The relative duration change between emphasized and neutral instances of the same word is defined as:

$$\delta_i = \frac{d_i^{\text{emph}} - d_i^{\text{neut}}}{d_i^{\text{neut}}} \tag{7}$$

where d_i^{neut} and d_i^{emph} are word durations obtained from forced alignment (see Section III-A2). This ratio reflects how much longer the emphasized word is relative to the neutral baseline. A ridge regression model is then fit to predict δ_i from the top-k PCs of the residuals $\mathbf{r}_i$, and R^2 scores are reported.

The same regression task is repeated on the remaining representation spaces ($\mathbf{A}, \mathbf{B}, \mathbf{C}$). Fitting on concatenated representations is expected to perform at least as well as $\mathbf{A}$ and $\mathbf{B}$, since the regressor has access to full information about both domains. Higher predictive performance from the residual space would support the hypothesis that emphasis is encoded as a structured, low-dimensional transformation, potentially making non-linear perceptual effects in speech linearly accessible.

5) Word Identity Prediction: To assess whether lexical information is accessible in representations, a simple word identity prediction task is performed. Similar to above, a logistic regression probing model is trained to predict word identity from representations.

Applied to residual representations, this provides an approximate measure of disentanglement: if residuals capture only emphasis transformations, they should contain little to no information about the underlying word. Recent work has shown that lexical or paralinguistic features can be explicitly removed from speech representations via linear projection, yielding disentangled representations [55]. In contrast, the current experiment evaluates inherent disentanglement without additional fine-tuning of the residuals.

The logistic regression model is trained using standard cross-entropy loss and a fixed learning rate of 1×10^{-4}. The dataset contains 546 unique word classes. Training is performed on 80% of the pairs (2985), with accuracy evaluated on a 20% held-out test set (747). This simple probe setup ensures that results reflect the information content of the representations rather than the capacity of the classifier.

The effective dimensionality required to achieve 95% of the task-specific performance is also computed. This is done by incrementally including the top-k PCs and identifying the smallest k for which cumulative performance reaches 95% of the maximum, providing insight into how concentrated the investigated information is within each representation space. This is then summarized using the area under the curve (AUC) over increasing k.

6) Layer- and Model-Wise Comparison: To investigate how emphasis sensitivity develops across layer-depth and training objectives, the duration change reconstruction and word identity prediction analyses are repeated across the following:

1) All encoder layers of each model,
2) A selection of pre-trained S3L models (e.g., wav2vec 2.0, HuBERT),
3) Fine-tuned variants, including: (a) ASR models, and (b) a model fine-tuned for emphasis classification.

This setup enables comparison of how task-relevant information is distributed across layers and whether fine-tuning shifts the encoding of emphasis from an implicit, low-dimensional transformation toward a more categorical or disentangled structure.

IV. Experiments

The analysis is demonstrated on layer 7 of wav2vec 2.0 as a worked example. It is then extended to all layers. Finally, different models and fine-tuning objectives are compared.

A. Cosine Similarity Distributions

The first experiment quantifies the extent to which emphasis alters word-level representations. Figure 2 suggests that emphasis induces a subtle but structured representational shift, with residual representations exhibiting both directional alignment and spread.

TABLE I: Summary of encoding properties for each group at a single layer. Effective dimensionality $D_{95\%}$ is computed from the explained variance. Top-k ($k = 20$) correlation is the average absolute correlation with duration change. R^2_{AUC} and $R^2_{95\%}$ is computed from regression onto duration change. WID_{AUC} and $\text{WID}_{95\%}$ show the performance and effective dimension of the word reconstruction task

Space	$D_{95\%}$	Corr	R^2_{AUC}	$R^2_{95\%}$	WID_{AUC}	$\text{WID}_{95\%}$
A	308	0.31	0.60	382	0.66	398
B	273	0.34	0.56	375	0.65	417
C	473	0.33	0.66	370	0.66	406
R	402	0.36	$\underline{0.71}$	341	$\underline{0.26}$	476

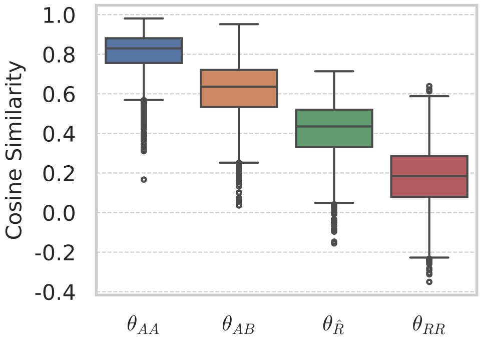


Fig. 2: Cosine similarity distributions across neutral and emphasized word representations on wav2vec 2.0, layer 7. θ_{AA}: neutral–neutral word pairs; θ_{AB}: neutral–emphasized word pairs; $\theta_{\hat{R}}$: residuals aligned to the mean residual vector; θ_{RR}: pairwise cosine similarity between residuals.

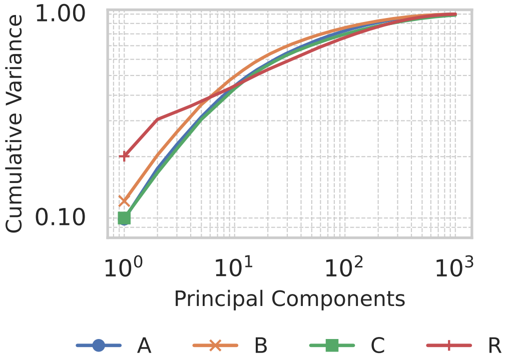


Fig. 3: Cumulative variance explained over PCs.

B. Dimensionality Analysis

Figure 3 shows the cumulative variance explained by PCs in each representation space. The residual space $\mathbf{R}$ appears more structured than the others in the early PCs, suggesting consistent variance between $\mathbf{B}$ and $\mathbf{A}$ along a low-dimensional subspace.

1) Correlations Between PCs and Duration: Figure 4 suggests a stronger correlation with duration change, providing evidence that these PCs best explain the emphasis transformation. Figure 5 illustrates the cumulative performance over

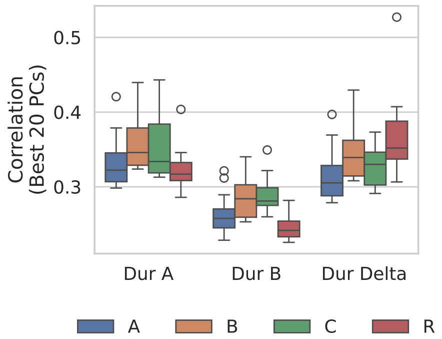


Fig. 4: Top-20 ranked PCs correlated with neutral duration (Dur A), emphasized duration (Dur B), and percentage duration change (Dur δ).

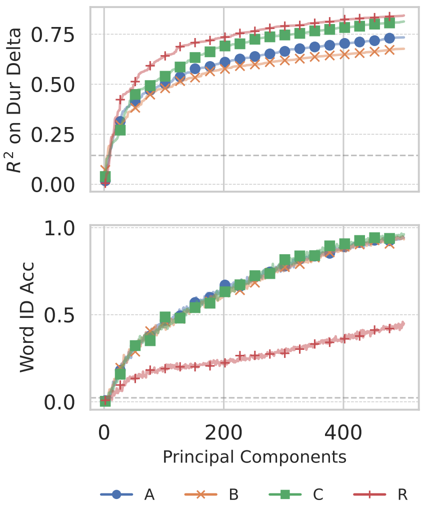
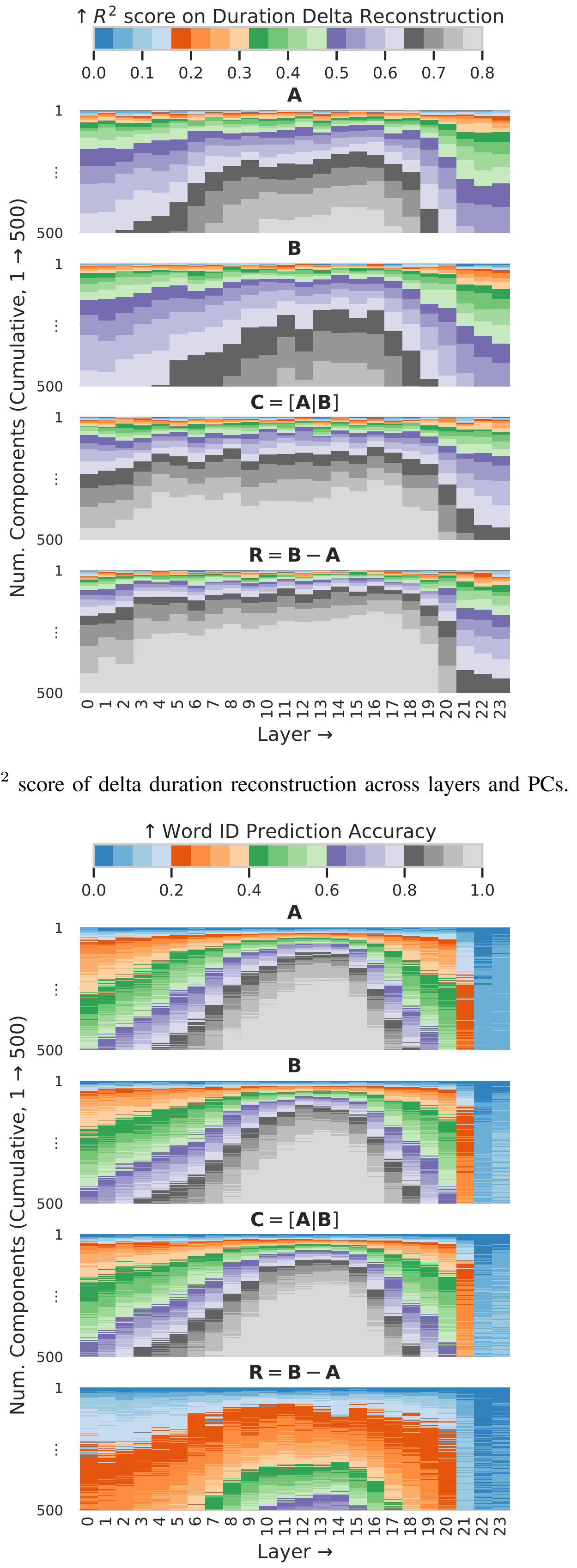


Fig. 5: The residual vectors contain enough information to reconstruct duration change (δ) but fail to predict word identity. Conversely, duration change is less well recovered by the emphasized, neutral, or concatenated representations.

PCs for both duration change reconstruction and word identity prediction. These results, summarized in Table I by taking the AUC over PCs, show that the residual vectors retain enough information to recover duration change but not word identity, which remains accessible to the original representation spaces.

C. Layer-Wise Comparison

Figure 6 shows the cumulative R^2 and accuracy scores for duration change reconstruction (top) and word identity prediction (bottom) across layers and PCs. The residual representations $\mathbf{R}$ achieve the highest reconstruction of duration change using fewer PCs, indicating that emphasis manifests as a structured, low-dimensional shift. In contrast, the residual yields near-zero performance on word identity prediction, indicating that lexical content is effectively removed. Meanwhile, the concatenated representations $\mathbf{C}$ matches $\mathbf{R}$'s reconstruction



(a) R^2 score of delta duration reconstruction across layers and PCs.

(b) Word identity prediction accuracy across layers and PCs.

Fig. 6: Layer-wise reconstruction and identity analysis. Top: regression-based reconstruction of emphasis variation. Bottom: word identity prediction performance.

TABLE II: Performance summary across models, showing area under the curve (AUC), effective dimensionality (Dim), and the corresponding layer of best performance for duration change (δ) reconstruction and word identity prediction

Model	**Residuals (R)**						**Concatenated (C)**					
	Duration δ			Word ID			Duration δ			Word ID		
	AUC	Dim	Layer	AUC	Dim	Layer	AUC	Dim	Layer	AUC	Dim	Layer
wav2vec 2.0 [43]	0.75	298	16	0.36	454	13	0.69	325	16	0.86	178	13
wav2vec 2.0 (ASR)	0.80	248	20	0.44	465	14	0.76	301	20	0.91	92	14
XLS-R [56]	0.76	309	32	0.26	476	32	0.70	341	35	0.84	206	28
XLS-R (ASR)	0.82	277	36	0.74	244	30	0.78	248	47	0.95	48	31
XLS-R (EC) [15]	0.76	293	23	0.74	207	23	0.69	358	23	0.93	63	21
HuBERT [44]	0.72	304	22	0.69	271	23	0.67	351	14	0.91	91	23
HuBERT (ASR)	0.81	263	23	0.69	295	22	0.73	337	22	0.94	59	22
data2Vec [57]	0.82	253	21	0.40	473	21	0.73	357	21	0.90	119	20
data2Vec (ASR)	0.83	240	20	0.39	457	21	0.75	296	20	0.90	117	21
WavLM-Base [45]	0.76	294	10	0.36	452	11	0.70	341	10	0.84	194	8
WavLM-Base (ASR)	0.77	293	10	0.34	499	7	0.72	330	11	0.87	132	8

performance, suggesting that the regressor can infer duration change from the full representations; yet only the residuals encode it directly and compactly.

Table II summarizes performance across models, comparing residual representations and concatenated representations for both duration change reconstruction and word identity prediction. Each entry shows the layer of best observed performance (AUC) and the number of PCs required to reach 95% of that performance (Dim). Across all models, residual representations consistently yield higher duration change reconstruction performance with lower effective dimensionality, indicating that emphasis is encoded as a structured, low-dimensional transformation. This effect is especially pronounced in fine-tuned ASR models, where residuals outperform raw representations while requiring fewer components. Conversely, word identity prediction accuracy is substantially lower for residuals than for concatenated representations, suggesting that lexical content is largely absent, or at least obfuscated, in the residual space. This supports the hypothesis that residual representations primarily isolate prosodic variation rather than word-specific features.

V. Discussion

The results provide strong evidence that emphasis is encoded as a structured, low-dimensional transformation within the internal representation spaces of the investigated speech models. Residual vectors between aligned emphasized and neutral word representations show strong directional consistency and occupy significantly fewer dimensions than the full embedding space. This supports the hypothesis that emphasis is not memorized in a word-specific manner but instead emerges as a reusable prosodic shift in representation space. Fine-tuning for ASR amplifies this effect: residuals become more predictive of duration change and less entangled with word identity, suggesting that ASR objectives may reinforce the accessibility of prosodic information. Word identity prediction from residuals remains low across models, further indicating that emphasis-related transformations are largely orthogonal to lexical encoding.

Interestingly, in the model fine-tuned for emphasis classification (XLS-R EC), duration change reconstruction is no longer dominated by the residual space. Emphasized words outperform neutral words in word identity prediction, suggesting the model may allocate more capacity on encoding emphasized content, potentially due to their relative rarity.

These findings echo results in style transfer and NLP analogy tasks, where residuals encode structured, interpretable variation. Unlike prior work that fine-tunes representations for disentanglement [55], this study shows that emphasis sensitivity can emerge inherently, particularly in middle-to-deep layers after fine-tuning.

A. Limitations

This work focuses on carefully controlled, aligned word pairs—matched by speaker, word, and sentence—to isolate the effect of emphasis. As a result, it does not explore how emphasis sensitivity behaves under relaxed conditions, such as varying speaker identity or contextual usage. In addition, all experiments are conducted on a benchmark synthetic dataset, which offers control over emphasis placement but may not fully capture the variability of natural speech. Finally, the analysis is limited to word-level emphasis, even though prosodic emphasis can span larger discourse units [58]. Investigating these broader and more variable conditions is left for future work. Nonetheless, the present findings demonstrate clear structure and interpretability under idealized settings, providing a strong foundation for further study.

VI. Conclusion

This work investigates whether modern speech models encode prosodic emphasis as a structured transformation in representation space. Using a novel residual analysis framework that combines parameter-free geometric metrics with lightweight probing tasks, this study shows that S3L models and ASR-tuned models exhibit clear emphasis sensitivity. Residual vectors are directionally aligned, low-dimensional, and predictive of changes in duration.

Fine-tuning for ASR enhances this effect, making emphasis encoding more consistent and less entangled with lexical identity. These findings suggest that emphasis is not only accessible but also implicitly structured in speech representations, offering implications for prosody-aware speech modeling, analysis, and control.

References

[1] S. Herment and L. Leonarduzzi, "The Pragmatic Functions of Prosody in English Cleft Sentences," in *Proceedings of Speech Prosody 2006*, May 2012, pp. 713–716. [Online]. Available: https://doi.org/10.21437/SpeechProsody.2012-178

[2] J. Pierrehumbert and J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds. The MIT Press, Jun. 1990, pp. 271–311. [Online]. Available: https://doi.org/10.7551/mitpress/3839.003.0016

[3] G. E. Lauerbach, "Emphasis," in *Pragmatics in Practice*, ser. Handbook of Pragmatics Highlights, J. O. Östman and J. Verschueren, Eds. John Benjamins Publishing Company, Dec. 2011, vol. 9, pp. 130–148. [Online]. Available: https://doi.org/10.1075/hoph.9.08eva

[4] M. Wagner, "Prosodic Focus," in *The Wiley Blackwell Companion to Semantics*, D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, and T. E. Zimmermann, Eds. John Wiley & Sons, Ltd, Nov. 2020, pp. 1–75. [Online]. Available: https://doi.org/10.1002/9781118788516.sem133

[5] S. Rajaa, "Improving End-to-End SLU Performance with Prosodic Attention and Distillation," in *Proceedings of Interspeech 2023*, Aug. 2023, pp. 1114–1118. [Online]. Available: https://doi.org/10.21437/Interspeech.2023-1760

[6] I. Tsiamas, M. Sperber, A. Finch, and S. Garg, "Speech is More than Words: Do Speech-to-Text Translation Systems Leverage Prosody?" in *Proceedings of the Ninth Conference on Machine Translation*, Nov. 2024, pp. 1235–1257. [Online]. Available: https://doi.org/10.18653/v1/2024.wmt-1.119

[7] A. Joly, M. Nicolis, E. Peterova, A. Lombardi, A. Abbas, A. v. Korlaar, A. Hussain, P. Sharma, A. Moinet, M. Lajszczak, P. Karanasou, A. Bonafonte, T. Drugman, and E. Sokolova, "Controllable Emphasis with zero data for text-to-speech," in *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW2023)*, Jun. 2023, pp. 113–119. [Online]. Available: https://doi.org/10.21437/SSW.2023-18

[8] V. J. van Heuven and A. Turk, "Phonetic Correlates of Word and Sentence Stress," in *The Oxford Handbook of Language Prosody*, C. Gussenhhoven and A. Chen, Eds. Oxford Handbooks, Feb. 2021, pp. 150–165. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780198832232.013.8

[9] K. J. Kohler, "What is Emphasis and How is it Coded?" in *Proceedings of Speech Prosody 2006*, May 2006, pp. 748–751. [Online]. Available: https://doi.org/10.21437/SpeechProsody.2006-225

[10] J. Cole and S. Shattuck-Hufnagel, "New Methods for Prosodic Transcription: Capturing Variability as a Source of Information," *Laboratory Phonology*, vol. 7, no. 1, Jun. 2016, article ID. 8. [Online]. Available: https://doi.org/10.5334/labphon.29

[11] D. R. Ladd and A. Arvaniti, "Prosodic Prominence Across Languages," *Annual Review of Linguistics*, vol. 9, no. 1, pp. 171–193, Jan. 2023. [Online]. Available: https://doi.org/10.1146/annurev-linguistics-031120-101954

[12] B. Arons, "Pitch-Based Emphasis Detection for Segmenting Speech Recordings," in *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, Sep. 1994, pp. 1931–1934. [Online]. Available: https://doi.org/10.21437/ICSLP.1994-485

[13] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness Predicts Prominence: Fundamental Frequency Lends Little," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, Aug. 2005. [Online]. Available: https://doi.org/10.1121/1.1923349

[14] Y. Xu, S.-w. Chen, and B. Wang, "Prosodic Focus with and without Post-Focus Compression: A Typological Divide within the Same Language Family?" *The Linguistic Review*, vol. 29, no. 1, pp. 131–147, Mar. 2012. [Online]. Available: https://doi.org/10.1515/tlr-2012-0006

[15] M. de Seyssel, A. D'Avirro, A. Williams, and E. Dupoux, "EmphAssess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Nov. 2024, pp. 495–507. [Online]. Available: https://doi.org/10.18653/v1/2024.emnlp-main.30

[16] M. Vaidya, K. Sabu, and P. Rao, "Deep Learning for Prominence Detection in Children's Read Speech," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8157–8161. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747780

[17] L. Zhang, J. Jia, F. Meng, S. Zhou, W. Chen, C. Zhang, and R. Li, "Emphasis Detection for Voice Dialogue Applications using Multi-channel Convolutional Bidirectional Long Short-Term Memory Network," in *Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Nov. 2018, pp. 210–214. [Online]. Available: http://doi.org/10.1109/ISCSLP.2018.8706625

[18] S. Shechtman, R. Fernandez, and D. Haws, "Supervised and Unsupervised Approaches for Controlling Narrow Lexical Focus in Sequence-to-Sequence Speech Synthesis," in *Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 431–437. [Online]. Available: https://doi.org/10.1109/SLT48900.2021.9383591

[19] T. B. Roettger, M. Tim, and J. Cole, "Mapping Prosody onto Meaning – the Case of Information Structure in American English," *Language, Cognition and Neuroscience*, vol. 34, no. 7, pp. 841–860, Aug. 2019. [Online]. Available: https://doi.org/10.1080/23273798.2019.1587482

[20] M. Yang, R. C. M. C. Shekar, O. Kang, and J. H. L. Hansen, "What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model," in *Proceedings of Interspeech 2023*, Jun. 2023, pp. 1923–1927. [Online]. Available: https://doi.org/10.21437/Interspeech.2023-2254

[21] R. Turnbull, A. J. Royer, K. Ito, and S. R. Speer, "Prominence Perception is Dependent on Phonology, Semantics, and Awareness of Discourse," *Language, Cognition and Neuroscience*, vol. 32, no. 8, pp. 1017–1033, Sep. 2017. [Online]. Available: https://doi.org/10.1080/23273798.2017.1279341

[22] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022. [Online]. Available: http://doi.org/10.1109/JSTSP.2022.3207050

[23] V. N. Vitale, F. Cutugno, A. Origlia, and G. Coro, "Exploring Emergent Syllables in End-to-End Automatic Speech Recognizers through Model Explainability Technique," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6875–6901, Apr. 2024. [Online]. Available: https://doi.org/10.1007/s00521-024-09435-1

[24] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the Utility of Self-Supervised Models for Prosody-Related Tasks," in *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 1104–1111. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10023234

[25] M. Kammoun and N. Ellouze, "Pitch and Energy Contribution in Emotion and Speaking Styles Recognition Enhancement," in *Proceedings of the Multiconference on "Computational Engineering in Systems Applications (CESA)"*, vol. 1, Oct. 2006, pp. 97–100. [Online]. Available: https://doi.org/10.1109/CESA.2006.4281631

[26] H. Cao, S. Beňuš, R. C. Gur, R. Verma, and A. Nenkova, "Prosodic Cues for Emotion: Analysis with Discrete Characterization of Intonation," in *Proceedings of Speech Prosody 2014*, May 2014, pp. 130–134. [Online]. Available: https://doi.org/10.21437/SpeechProsody.2014-14

[27] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody Dependent Speech Recognition on Radio News Corpus of American English," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 232–245, Jan. 2006. [Online]. Available: http://doi.org/10.1109/TSA.2005.853208

[28] Y. Cho, S. Ng, T. Tran, and M. Ostendorf, "Leveraging Prosody for Punctuation Prediction of Spontaneous Speech," in *Proceedings of Interspeech 2022*, Sep. 2022, pp. 555–559. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-11061

[29] O. Tilk and T. Alumäe, "LSTM for Punctuation Restoration in Speech Transcripts," in *Proceedings of Interspeech 2015*, Sep. 2015, pp. 683–687. [Online]. Available: https://doi.org/10.21437/Interspeech.2015-240

[30] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High Quality Voice Conversion using Prosodic and High-Resolution Spectral Features," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5265–5285, May 2016. [Online]. Available: https://doi.org/10.1007/s11042-015-3039-x

[31] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Communication*, vol. 32, no. 1, pp. 127–154, Sep. 2000. [Online]. Available: https://doi.org/10.1016/S0167-6393(00)00028-5

[32] Q. T. Do, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Improving Translation of Emphasis with Pause Prediction in Speech-to-Speech Translation Systems," in *Proceedings of the 12th International Workshop*

on Spoken Language Translation: Papers, Dec. 2015, pp. 204–208. [Online]. Available: https://aclanthology.org/2015.iwslt-papers.12/

[33] T. Do, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Preserving Word-Level Emphasis in Speech-to-Speech Translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 544–556, Mar. 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2643280

[34] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A Study on the Effect of Prosodic Emphasis Transfer on Overall Speech Translation Quality," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 8396–8400. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6639303

[35] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé, D. Dey, M. Harper, T. Howard, C. Kennington, I. Kruijff-Korbayová, D. Manocha, C. Matuszek, R. Mead, R. Mooney, R. K. Moore, M. Ostendorf, H. Pon-Barry, A. I. Rudnicky, M. Scheutz, R. S. Amant, T. Sun, S. Tellex, D. Traum, and Z. Yu, "Spoken Language Interaction with Robots: Recommendations for Future Research," *Computer Speech & Language*, vol. 71, Jan. 2022, article ID. 101255. [Online]. Available: https://doi.org/10.1016/j.csl.2021.101255

[36] E. Velner, P. P. Boersma, and M. M. de Graaf, "Intonation in Robot Speech: Does it Work the Same as with People?" in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2020, pp. 569–578. [Online]. Available: https://doi.org/10.1145/3319502.3374801

[37] E. Beier, M. Cohn, T. Trammel, F. Ferreira, and G. Zellou, "Marking Prosodic Prominence for Voice Assistant and Human Addressees," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 56, no. 6, pp. 986–1003, Jun. 2024. [Online]. Available: https://doi.org/10.1037/xlm0001396

[38] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling Prominence and Emphasis Improves Unit-Selection Synthesis," in *Proceedings of Interspeech 2007*, Aug. 2007, pp. 1282–1285. [Online]. Available: https://doi.org/10.21437/Interspeech.2007-230

[39] S. Latif, I. Kim, I. Calapodescu, and L. Besacier, "Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, Nov. 2021, pp. 544–551. [Online]. Available: https://doi.org/10.18653/v1/2021.conll-1.42

[40] J. Levis and L. Pickering, "Teaching Intonation in Discourse using Speech Visualization Technology," *System*, vol. 32, no. 4, pp. 505–524, Dec. 2004. [Online]. Available: https://doi.org/10.1016/j.system.2004.09.009

[41] T. Matzinger, N. Ritt, and W. T. Fitch, "The Influence of Different Prosodic Cues on Word Segmentation," *Frontiers in Psychology*, vol. 12, p. 622042, Mar. 2021. [Online]. Available: http://doi.org/10.3389/fpsyg.2021.622042

[42] S. Bannò and M. Matassoni, "Proficiency Assessment of L2 Spoken English using Wav2Vec 2.0," in *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 1088–1095. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10023019

[43] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[44] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[45] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022. [Online]. Available: https://doi.org/10.1109/JSTSP.2022.3188113

[46] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proceedings of Interspeech 2021*, Oct. 2021, pp. 1194–1198. [Online]. Available: http://doi.org/10.21437/Interspeech.2021-1775

[47] A. de la Fuente and D. Jurafsky, "A Layer-Wise Analysis of Mandarin and English Suprasegmentals in SSL Speech Models," in *Proceedings of Interspeech 2024*, Sep. 2024, pp. 1290–1294. [Online]. Available: https://doi.org/10.21437/Interspeech.2024-2341

[48] Y. Belinkov, "Probing Classifiers: Promises, Shortcomings, and Advances," *Computational Linguistics*, vol. 48, no. 1, pp. 207–219, Apr. 2022. [Online]. Available: https://doi.org/10.1162/coli_a_00422

[49] Y. Kanda, R. Fontugne, K. Fukuda, and T. Sugawara, "ADMIRE: Anomaly Detection Method using Entropy-Based PCA with Three-Step Sketches," *Computer Communications*, vol. 36, no. 5, pp. 575–588, Mar. 2013. [Online]. Available: https://doi.org/10.1016/j.comcom.2012.12.002

[50] A. A. Kalaitzis and N. D. Lawrence, "Residual Component Analysis: Generalising PCA for More Flexible Inference in Linear-Gaussian Models," in *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, Jun. 2012, pp. 539–546, URL references accepted version hosted on University of Cambridge Repository. [Online]. Available: https://doi.org/10.17863/CAM.47960

[51] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What Do Self-Supervised Speech Models Know About Words?" *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 372–391, Apr. 2024. [Online]. Available: https://doi.org/10.1162/tacl_a_00656

[52] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proceedings of Interspeech 2017*, Aug. 2017, pp. 498–502. [Online]. Available: https://doi.org/10.21437/Interspeech.2017-1386

[53] L. Fournier, E. Dupoux, and E. Dunbar, "Analogies Minus Analogy Test: Measuring Regularities in Word Embeddings," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Nov. 2020, pp. 365–375. [Online]. Available: https://doi.org/10.18653/v1/2020.conll-1.29

[54] I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Apr. 2016, article ID. 20150202. [Online]. Available: https://doi.org/10.1098/rsta.2015.0202

[55] C. Zhang, Y. Zhou, R. Zhao, Y. Chen, and X. Shi, "Representation Purification for End-to-End Speech Translation," in *Proceedings of the 31st International Conference on Computational Linguistics*, Dec. 2024, pp. 6255–6269. [Online]. Available: https://aclanthology.org/2025.coling-main.418/

[56] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale," in *Proceedings of Interspeech 2022*, Dec. 2022, pp. 2282–2287. [Online]. Available: http://doi.org/10.21437/Interspeech.2022-143

[57] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language," in *Proceedings of the 39th International Conference on Machine Learning*, Jul. 2022, pp. 1298–1312. [Online]. Available: https://proceedings.mlr.press/v162/baevski22a.html

[58] G. Klewitz and E. Couper-Kuhlen, "Quote – Unquote? The Role of Prosody in the Contextualization of Reported Speech Sequences," *Pragmatics*, vol. 9, no. 4, pp. 459–485, Jan. 1999. [Online]. Available: https://doi.org/10.1075/prag.9.4.03kle