Architectural Control of Phonetic Invariance via Additive U-Net Multi-Task Learning

Vikram R. Lakkavalli*
vikram.ramesh@iiitb.ac.in, vikram.ckm@gmail.com
Neelam Sinha†
neelam@cbr-iisc.ac.in
*IIITB, Bangalore, India
†Center for Brain Research, IISc, Bangalore, India

***Abstract*—This work investigates whether phonetic invariance in speech representations can be characterized as a structured transformation within the input acoustic feature space under architectural constraints. To this end, we employ an additive U-Net based multi-task learning (MTL) framework that jointly optimizes phoneme recognition and spectral reconstruction, imposing complementary objectives: reconstruction preserves acoustic detail, while recognition promotes invariance to speaker and channel variability. Experiments on the TIMIT dataset using log-magnitude spectral features evaluate recognition performance, reconstruction fidelity, and speaker-related variability across representations. Results show that the encoder bottleneck forms a structured representation in which speaker-discriminative information is reduced while phonetic structure is retained, consistent with the emergence of phonetic invariance. Further analysis reveals that the proposed architecture not only shapes the representation but also alters error behavior, reducing substitutions while increasing deletions, reflecting suppression of spurious patterns. These findings suggest that architectural design, combined with multi-task learning, can systematically steer acoustic representations toward more invariant phonetic organization.**

I. Introduction

Automatic speech recognition (ASR) requires representations that preserve phonetic structure while remaining robust to variability introduced by speakers, recording conditions, and channels. Speech signals exhibit variability across multiple time scales: short-term spectral–temporal patterns encode phonetic content, while longer-term factors such as speaker identity, channel characteristics, and noise introduce systematic shifts in the acoustic feature space. As a result, identical phonetic units may occupy different regions of the feature space across utterances, making it necessary to align these variations to recover a consistent phonetic representation.

Modern deep learning approaches learn such representations directly from data and achieve strong performance [3], [4]. However, the mechanisms through which nuisance variability is suppressed remain largely implicit. Existing analyses are typically post-hoc and provide limited insight into how architectural design influences the formation of invariant representations. This raises the question of whether phonetic invariance can be systematically shaped through architectural constraints rather than solely emerging from data-driven optimization.

In this work, we study phonetic invariance from the perspective of representation alignment under competing objectives. We employ a multi-task learning (MTL) framework in which spectral reconstruction and phoneme recognition impose complementary constraints. Reconstruction encourages preservation of fine-grained acoustic detail, retaining variability present in the input, while recognition promotes invariance by suppressing long-term variability. The interaction between these objectives creates a structured tension that influences the organization of intermediate representations.

To further constrain this process, we adopt an additive U-Net architecture in which skip connections are fused through addition rather than concatenation. This design limits direct information bypass and enforces a more structured transformation of representations across layers, encouraging the encoder to progressively refine and reorganize information rather than simply propagate it.

The main contributions of this work are:

- Phonetic invariance as a problem of representation alignment under speaker- and channel-induced variability.
- An additive U-Net based multi-task framework that constrains information flow and regulates the formation of intermediate representations.
- Joint reconstruction and recognition objectives reshape the encoder bottleneck, reducing speaker-discriminative information while preserving phonetic structure.
- We show that joint reconstruction and recognition reshape phoneme error patterns, reducing substitutions while increasing conservative deletions.

II. Related Work

Learning robust speech representations has been a central focus in recent ASR research. Self-supervised approaches such as wav2vec 2.0 [3] and HuBERT [4] demonstrate that neural models can learn structured representations directly from raw speech by leveraging large-scale unlabeled data. These methods implicitly capture phonetic structure while achieving invariance to speaker and channel variability. However, the mechanisms through which such invariance emerges remain largely implicit, as these models rely on data scale and training objectives rather than explicit architectural constraints.

A complementary line of work investigates neural representations through probing methods [7], [8]. These studies reveal that phonetic and speaker-related information coexist

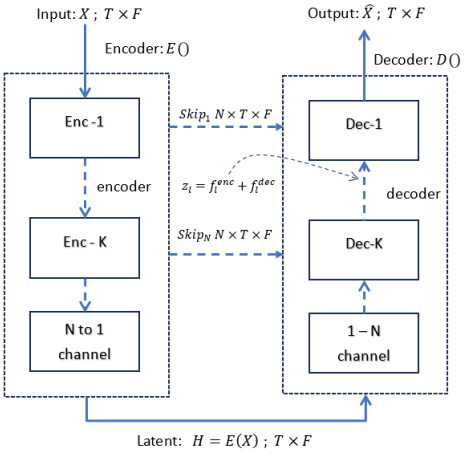

Fig. 1. *Additive U-Net architecture with encoder–decoder structure. The input $X \in \mathbb{R}^{T \times F}$ is processed by the encoder, which progressively transforms features across N channels while preserving temporal and frequency dimensions $(T \times F)$, and is projected to a single-channel latent representation. The encoder employs residual (subtractive) transformations to refine representations across layers. The decoder reconstructs $\hat{X} \in \mathbb{R}^{T \times F}$ by expanding bottleneck $H()$ $(1 \to N)$. At each level l, encoder and decoder features of size $N \times T \times F$ are fused through weighted additive skip connections $z_l = \alpha_l \times f_l^{enc} + f_l^{dec}$.*

and evolve across network depth, suggesting that invariance is not explicitly enforced but gradually emerges during training. While such analyses provide valuable post-hoc insights, they offer limited guidance on how architectural design and training objectives can be used to actively shape representation behavior.

Reconstruction-based approaches have also been explored from an analysis-by-synthesis (AbS) perspective [1], as well as in joint ASR–TTS frameworks [11], [12]. In these settings, reconstruction serves as an auxiliary objective that encourages preservation of acoustic detail. However, these methods typically operate at the signal or sequence level and do not explicitly examine how reconstruction interacts with recognition objectives to influence the structure of intermediate representations. In contrast, this work focuses on reconstruction directly in the acoustic feature space to study its role in shaping representation geometry.

Encoder–decoder architectures such as U-Net [9] employ skip connections to facilitate information flow between layers. Conventional designs rely on concatenative skip connections, which allow high-resolution features to bypass intermediate transformations. While effective for reconstruction tasks, such shortcuts can limit the extent to which intermediate representations are forced to encode structured information. In this work, we adopt additive skip fusion to restrict shortcut capacity, thereby encouraging representations that balance preservation of acoustic detail with invariance to long-term variability. The role of such architectural constraints in shaping phonetic invariance remains relatively unexplored and is the focus of this study.

III. Proposed Framework

A. Multi-Task Learning

Let $X \in \mathbb{R}^{T \times F}$ denote input acoustic features. The encoder produces a latent representation:

$$H = E(X). \tag{1}$$

A CTC head predicts phoneme sequences:

$$\hat{Y} = C(H), \tag{2}$$

while a decoder reconstructs the input:

$$\hat{X} = D(H). \tag{3}$$

The training objective combines recognition and reconstruction:

$$\mathcal{L} = \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_{rec}\mathcal{L}_{rec}. \tag{4}$$

The two losses impose competing constraints — recognition suppresses speaker- and channel-dependent variability, while reconstruction preserves acoustic detail — yielding a structured transformation that retains phonetic information while attenuating nuisance variability.

B. Additive U-Net Architecture

The encoder–decoder follows a U-Net structure with skip connections linking corresponding levels. Let h_l denote the encoder state at level l, and let $\phi_l(\cdot)$ denote the convolutional transformation applied at that level. The decoder state at the corresponding level is denoted by d_l. Instead of concatenative skips, we use additive fusion in the decoder:

$$z_l^{dec} = d_l + \alpha_l s_l, \tag{5}$$

where s_l is the skip representation from encoder level l, and α_l is a learnable scalar controlling the skip contribution. Additive fusion restricts shortcut capacity compared to concatenation, forcing intermediate representations to encode information in a more structured manner.

We consider two variants that differ in how the encoder propagates information after forming the skip representation.

Pseudo AddUNet. The transformed encoder output is used both as the skip representation and as the input to the next encoder level:

$$s_l = \phi_l(h_l), \qquad h_{l+1} = s_l. \tag{6}$$

Thus, the encoder propagates the same transformed representation that is also passed through the skip pathway.

Real AddUNet. The transformed component is used as the skip representation, but is subtracted from the current encoder state before propagation:

$$s_l = \phi_l(h_l), \qquad h_{l+1} = h_l - s_l. \tag{7}$$

This residual subtraction removes the component routed through the skip pathway from the deeper encoder stream,

encouraging progressive reorganization rather than direct propagation.

The pseudo variant serves as an additive-skip reference, while the real variant introduces the additional subtractive constraint used to study architectural control of invariance.

C. Experimental Setup

Data. All experiments use the TIMIT corpus with the standard train/dev/test split and 61-to-39 phoneme folding for CTC targets. Input features are log-magnitude STFT representations.

Architecture. The encoder has base channel count $N = 12$, five levels with no spectral or temporal downsampling (`down_factors=(1,1,1,1,1)`), context kernels $(3\times3, 3\times3, 3\times3, 5\times5, 7\times7, 7\times7)$, and update kernels $(5\times3)\times5$ followed by (3×3). The decoder mirrors the encoder over six layers. The CTC head is a CNN–BLSTM stack on a trainable 80-bin log-mel front-end: three 5×5 convolutional layers with 24 channels, followed by three BLSTM layers with hidden size 128 and dropout 0.3. Per-level skip-fusion scalars α_l are learnable.

Training. A two-stage protocol is used. In the warmup stage (8 epochs, $w_{ctc} = 1.0$, $w_{rec} = 0$, learning rate 1.5×10^{-4}), only the CTC head is supervised. In the joint stage (60 epochs, $w_{ctc} = 1.0$ with $\lambda_{ctc} = 0.6$, $w_{rec} = 1.0$ on the dense reconstruction target, learning rate 10^{-4}), the model is initialized from the warmup checkpoint and both losses are active. Both stages use Adam with weight decay 10^{-5}, batch size 6, and a plateau scheduler (factor 0.5, patience 4). Results are reported across five training seeds (111–555) at the best-validation epoch.

Speaker probe. A separate MLP probe (LayerNorm input $\to 256 \to 256 \to 462$ speakers, ReLU, dropout 0.5) is trained on masked-mean-pooled utterance representations using Adam optimizer(learning rate 10^{-3}, batch 128) for 250 epochs with best-validation selection. The data follows a fixed 70/15/15 speaker-stratified partition of the TIMIT training set (462 speakers, 3,696 utterances). To estimate probe-side variance, results are averaged over five probe seeds (111–555) on a single fixed AddUNet checkpoint (model seed 222) per variant.

IV. Results

The results aim to distinguish the effect of architectural constraints from that of multi-task learning by analyzing representation behavior across reconstruction, recognition, and invariance.

A. Recognition Performance

Table I reports PER for the CTC-only reference and the two AddUNet variants. Both AddUNet variants match the CTC-only reference within seed variance, indicating that MTL does not degrade recognition. The remainder of this section analyzes how MTL reshapes the representation and the structure of errors.

TABLE I
TIMIT PHONEME RECOGNITION (MEAN ± STD OVER FIVE SEEDS; CTC-ONLY FROM A SINGLE REFERENCE RUN).

Model	Val PER (%)	Test PER (%)
CTC-only	21.01	23.78
Pseudo AddUNet	21.26 ± 0.93	23.32 ± 1.18
Real AddUNet	22.05 ± 1.09	23.97 ± 1.18

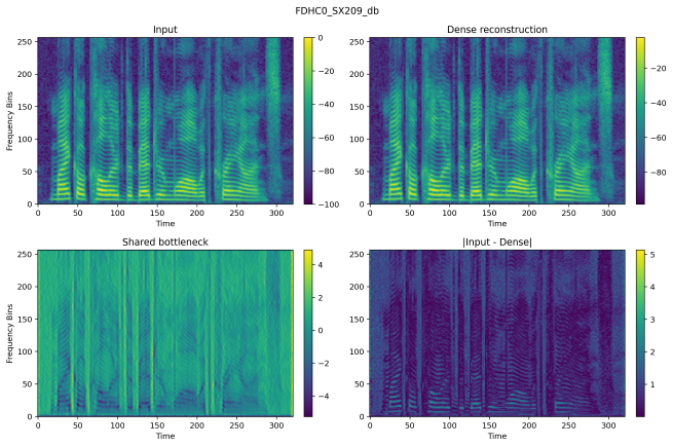

Fig. 2. *Input log-spectrum and dense reconstruction (top row), and the corresponding encoder bottleneck and absolute reconstruction error (bottom row) for the Real AddUNet on a TIMIT utterance (FDHC0/SX209). The bottleneck exhibits lower variance than the input while maintaining similar structure, indicating suppression of fine-scale variability.*

B. Reconstruction Behavior

Figure 2 shows the input, its reconstruction, and the encoder bottleneck. The reconstruction preserves dominant spectral structure while the bottleneck exhibits smoother spectral transitions, indicating suppression of fine-scale variability.

Figure 3 reports frequency-wise reconstruction error averaged over the test set. Error is lowest in the mid-frequency band, where phonetic information is most prominent, and increases toward both the low-frequency region (pitch and voicing) and the high-frequency region (fine spectral detail). This frequency-selective allocation of reconstruction fidelity suggests that the learned representation preferentially preserves acoustically relevant phonetic structure while relaxing reconstruction of variability less critical for recognition.

Figure 4 compares bottleneck representations across the three models. The CTC-only bottleneck shows sparse, low-energy activations, reflecting compression toward discriminative features. The Pseudo AddUNet retains fine spectral variations and time-local activations, preserving acoustic detail. The Real AddUNet produces a smoother, more structured representation in which harmonic structure and transitions are retained while fine-scale variability is attenuated — direct evidence that subtractive encoding reorganizes the bottleneck more aggressively than the additive-only variant.

C. Error Redistribution under Multi-Task Learning

Figure 5 shows the row-normalized confusion difference between MTL and the CTC-only reference on the TIMIT

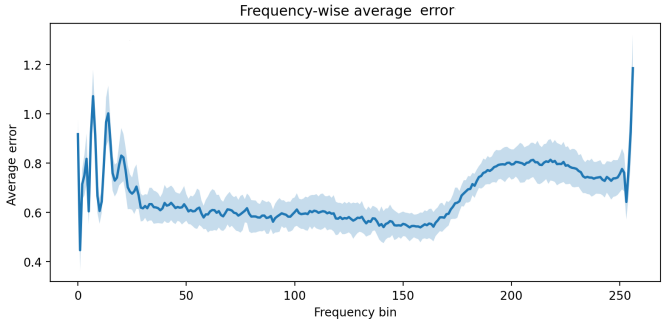

Fig. 3. *Frequency-wise reconstruction error of the Real AddUNet averaged over the TIMIT test set. Lower-frequency components are reconstructed with higher fidelity, while higher-frequency regions exhibit increased error, indicating selective preservation of dominant spectral structure with relaxed modeling of fine-grained detail.*

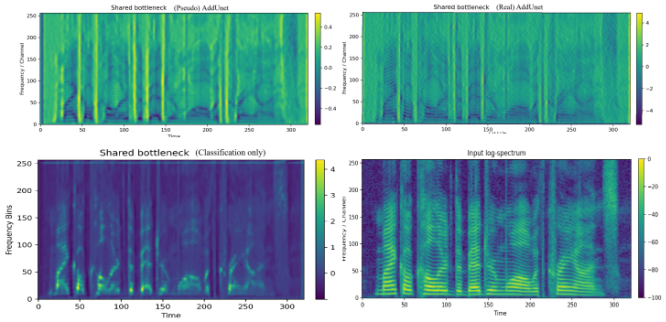

Fig. 4. *Bottleneck representations on TIMIT utterance FDHC0/SX209: Pseudo AddUNet (top-left), Real AddUNet (top-right), CTC-only (bottom-left), input log-spectrum (bottom-right).*

test set. Under MTL, off-diagonal mass decreases for most phonemes while the deletion column gains mass for several — a redistribution from misclassification toward rejection of uncertain evidence.

Table II lists representative cases. Phonemes with localized, transient acoustic cues (*uw*, *iy*, *k*, *aa*) gain correct classifications, with substitutions partly absorbed into correct decisions and partly into deletions. The *v* phoneme is the extreme case: nearly all instances are deleted under MTL (ΔDel $= +0.94$). This is consistent with /v/'s short duration and low acoustic energy, which under the smoothing induced by additive skip fusion — combined with the encoder's limited temporal context (3–7 frame kernels) — appears to fall below the threshold for confident emission. A targeted analysis of this failure mode, including its dependence on architectural variant and temporal context, is left to future work.

D. Speaker Invariance Analysis

Speaker identification probing is used to assess how much speaker-related information remains in the learned bottleneck representations. For each AddUNet variant, we train five independently initialized models using the same architecture and training configuration, differing only in random seed. After training, the encoder is frozen and a separate MLP speaker probe is trained on the input features and on the corresponding

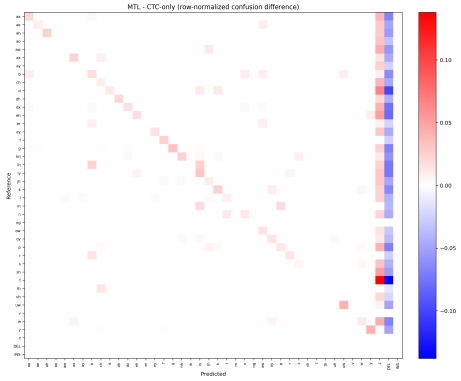

Fig. 5. *Row-normalized phoneme confusion difference (MTL minus CTC-only) on the TIMIT test set, using the Real AddUNet for MTL. Rows are true phonemes; columns are predicted phonemes plus a separate deletion column. Red entries indicate transitions more frequent under MTL, blue less frequent. Diagonal entries reflect correct classifications, off-diagonal columns capture substitutions, and the deletion column captures rejection of acoustic evidence. MTL reduces off-diagonal mass (fewer substitutions) and adds mass to the deletion column for several phonemes — most strikingly /v/, whose row is almost entirely shifted into the deletion column.*

TABLE II
REPRESENTATIVE PHONEME-LEVEL CHANGES UNDER MTL RELATIVE TO CLASSIFIER-ONLY TRAINING. SEVERAL PHONEMES SHOW INCREASED CORRECT CLASSIFICATION AND REDUCED SUBSTITUTION ERRORS, WHILE DELETION RATES INCREASE. THIS INDICATES A REDISTRIBUTION OF ERRORS FROM SUBSTITUTIONS TO DELETIONS, CONSISTENT WITH MORE CONSERVATIVE DECODING BEHAVIOR UNDER RECONSTRUCTION CONSTRAINTS.

Phone	ΔCorrect	ΔDel	ΔSub
uw	+0.0417	-0.0069	-0.0347
iy	+0.0303	+0.0364	-0.0667
k	+0.0245	+0.0490	-0.0735
aa	+0.0219	+0.0395	-0.0614
v	-0.9400	+0.9400	+0.0000

bottleneck representations. Table III reports the resulting probe accuracies averaged across the five model seeds.

Both variants substantially reduce test speaker probe accuracy relative to the input, with Real AddUNet producing a stronger reduction than Pseudo variant.

E. Effect of Architectural Constraints

To isolate the role of architectural design, we compare the two proposed AddUNet variants under the same training configuration. The Pseudo variant uses additive skip fusion without encoder-side subtraction; the Real variant adds the subtractive coupling. This comparison separates the effect of multi-task learning from the additional constraint imposed by subtractive information flow.

Table IV compares the similarity between input and bottleneck representations. The CTC-only model shows higher correlation and stronger PCA overlap with the input, indicating that its bottleneck remains close to the dominant input directions. The MTL representation has lower correlation while retaining substantial PCA overlap, suggesting that it remains in the acoustic feature coordinate system but undergoes stronger

TABLE III

SPEAKER IDENTIFICATION PROBE ACCURACY (%) FOR INPUT FEATURES AND FROZEN BOTTLENECK REPRESENTATIONS FROM INDEPENDENTLY TRAINED ADDUNET MODELS. RESULTS ARE AVERAGED OVER FIVE MODEL SEEDS USING THE SAME ARCHITECTURE AND TRAINING CONFIGURATION FOR PSEUDO AND REAL ADDUNET VARIANTS. LOWER BOTTLENECK ACCURACY INDICATES STRONGER SUPPRESSION OF SPEAKER-DISCRIMINATIVE INFORMATION.

Representation	Validation	Test
Input	80.51 ± 0.72	78.97 ± 0.60
Pseudo AddUNet	64.41 ± 0.54	62.11 ± 0.48
Real AddUNet	61.78 ± 0.49	60.60 ± 0.65

TABLE IV

SIMILARITY BETWEEN INPUT AND BOTTLENECK REPRESENTATIONS. CORRELATION MEASURES DIRECT LINEAR SIMILARITY, WHILE PCA OVERLAP ($k = 32$) MEASURES HOW MUCH BOTTLENECK ENERGY LIES IN THE PRINCIPAL SUBSPACE OF THE INPUT FEATURES.

Model	**Correlation**	**PCA Overlap**
CTC-only	0.894 ± 0.023	0.957
MTL (AddUNet)	0.654 ± 0.044	0.927

reorganization of fine-scale variability. Combined with the speaker-probe results (Table III), this indicates that the subtractive encoding constraint imposes a measurable additional reorganization on top of the MTL effect.

V. DISCUSSION

The bottleneck representations show a structured transformation of the acoustic feature space rather than direct preservation or compression. Under MTL, the bottleneck exhibits smoother spectral transitions and reduced fine-scale variability while retaining dominant phonetic patterns, consistent with the reconstruction behavior in which lower-frequency components are preserved with higher fidelity. This transformation is also reflected in recognition behavior: MTL reduces substitutions while increasing deletions, indicating a shift toward conservative decoding where uncertain acoustic evidence is suppressed rather than misclassified.

Similarity analysis supports this reading. The CTC-only bottleneck stays highly correlated with the input and aligned with its principal directions, while the MTL bottleneck shows lower correlation but retains substantial PCA overlap — reorganization within the input coordinate system, not projection into a different embedding space.

The Pseudo–Real comparison localizes the source of this reorganization. Both variants achieve recognition within seed variance on test PER (Table I), but the Real variant produces a smoother bottleneck (Fig. 4) and a slightly stronger reduction in speaker-probe accuracy (60.60% vs 62.11% test, Table III). The subtractive constraint, isolated by the controlled pair is associated with smoother bottleneck representations and a modest additional reduction in speaker-discriminative information, with recognition remaining within the observed seed-level variation.

Taken together, these observations indicate that architectural constraints influence not only recognition performance but also the structure of the learned representation and the type of invariance that emerges.

VI. CONCLUSION

This work examined phonetic invariance in ASR through multi-task learning and architectural constraints in an additive U-Net framework. The learned bottleneck reorganizes the acoustic feature space while preserving phonetic structure: long-term variability is attenuated, errors redistribute from substitutions toward deletions, and speaker information is suppressed relative to the input.

The Pseudo and Real AddUNet variants form a controlled architectural pair: both use additive skip fusion at the decoder, while only the Real variant introduces subtractive coupling in the encoder. This isolates the effect of subtractive encoding while holding the remaining architecture fixed. Both achieve comparable recognition performance, but the Real variant yields smoother bottleneck representations and slightly stronger suppression of speaker-discriminative information.

These results establish that architectural constraints imposed on the latent space measurably shape the learned representation in ASR. Characterizing how this effect generalizes — across acoustic conditions beyond TIMIT, across encoder kernel geometries, and across alternative encoder architectures — requires targeted experimental work beyond the scope of the present paper.

Future work will focus on explicit characterization of the learned representation geometry, evaluation under broader acoustic conditions, and extension of the framework with longer temporal context and speaker-aware conditioning.

REFERENCES

[1] H. Hermansky, "Analysis by synthesis for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 49–56, 1995.

[2] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," *Speech Communication*, vol. 12, no. 2, pp. 117–134, 1993.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[4] 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, 2021.

[5] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[6] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[7] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Computational Linguistics*, vol. 46, no. 1, pp. 49–72, 2020.

[8] A. Alishahi, D. Barkhuijsen, and G. Chrupała, "Encoding of phonology in a recurrent neural model of grounded speech," in *Proc. ACL*, 2017.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[11] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," in *Proc. ASRU*, 2017.

[12] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning," in *Proc. IEEE ASRU*, 2017.