Proceedings Paper:
Ollerenshaw, A. orcid.org/0000-0001-5779-1905, Jalal, M.A. and Hain, T. orcid.org/0000-0003-0939-3464 (2021) Insights on neural representations for end-to-end speech recognition. In: Heřmanský, H., Černocký, H., Burget, L., Lamel, L., Scharenborg, O. and Motlicek, P., (eds.) Interspeech 2021. Interspeech 2021, 30 Aug - 03 Sep 2021, Brno, Czechia. ISCA - International Speech Communication Association , pp. 4079-4083.

https://doi.org/10.21437/interspeech.2021-1516

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Insights on Neural Representations for End-to-End Speech Recognition

Anna Ollerenshaw[1], *Md Asif Jalal*[1], *Thomas Hain*[1]

[1]Speech and Hearing Research Group, University of Sheffield, UK

alollerenshaw1@sheffield.ac.uk

Abstract

End-to-end automatic speech recognition (ASR) models aim to learn a generalised speech representation. However, there are limited tools available to understand the internal functions and the effect of hierarchical dependencies within the model architecture. It is crucial to understand the correlations between the layer-wise representations, to derive insights on the relationship between neural representations and performance. Previous investigations of network similarities using correlation analysis techniques have not been explored for End-to-End ASR models. This paper analyses and explores the internal dynamics between layers during training with CNN, LSTM and Transformer based approaches using Canonical correlation analysis (CCA) and centered kernel alignment (CKA) for the experiments. It was found that neural representations within CNN layers exhibit hierarchical correlation dependencies as layer depth increases but this is mostly limited to cases where neural representation correlates more closely. This behaviour is not observed in LSTM architecture, however there is a bottom-up pattern observed across the training process, while Transformer encoder layers exhibit irregular coefficiency correlation as neural depth increases. Altogether, these results provide new insights into the role that neural architectures have upon speech recognition performance. More specifically, these techniques can be used as indicators to build better performing speech recognition models.

Index Terms: End-to-End, speech recognition, analysis

1. Introduction

Traditionally, ASR frameworks have been developed using Hidden Markov Models (HMM) in combination with Gaussian Mixture Models (GMM) to identify and map acoustic features to phonemes. Recent work [1] has introduced deep neural networks to replace traditional approaches by factorising the system into specialised modules, such as acoustic and language models. The End-to-End approaches for ASR attempt to simplify the pipeline and directly model the input features to characters or phonemes [2]. This approach allows the development of a complete ASR system without the requirement of expert domain knowledge, while attempting to globally optimise the training process. As the development and integration of End-to-End approaches have become increasingly popular, many different architectures have been developed [3] [4] [5] [6]. Research from [7] has shown that neural layer depth can attribute to a richer neural representational capacity, but generalisation or memorisation behaviour of the models remains elusive [8]. This hypothesis does not always translate to performance improvements in all cases [9] and has little exploration in the End-to-End ASR domain. End-to-End architectures have inherently complex internal dynamics and whether the model learns to generalise from the training process, is imperative to yield recognition performance improvements [10] [11]. Furthermore, there is limited research with regard to the interactions between the training dynamics of End-to-End models and speech data and it is unclear how structural components or residual connections within the models contribute to more optimal representations. To explore this further, a window of observation into the neural representations of the network architectures would be required to provide information on the interaction between the training structures and the data.

Current correlation analysis techniques from [12] and [13] have been utilised to compare deep neural network representations. Comparing population representations has been explored in several methods, however this is a non-trivial task as it not clear which aspects of the representations the similarity index should attempt to focus on. Canonical Correlation Analysis (CCA) [14] and Centered-Kernel Alignment (CKA) [13] have been used as tools to compare network representations, as they enable the identification of shared structures across representations which are trivially dissimilar. The application of singular value decomposition (SVD) before CCA, referred to as SVCCA [12] has been used to compare representations across networks and it was found that network solutions for image classification, diverged predominantly in the intermediate neural layers. SVCCA has also been used as a tool to show the evolution of linguistic features as they were encoded in language models [15] and it has been observed, for an image classification task, that as neural layer depth increased, the network similarities decreased. Finally, the CKA approach demonstrated that task trained neural layers developed more similar representations than layers that were randomised.

However, these approaches have not yet been used to analyse neural representations of End-to-End architectures with speech data for speech recognition. In this work, a comparative study of neural representation analysis is provided with some of the most prevalent End-to-End ASR networks. The analysis focuses on understanding the similarity of the neural representations and answering questions such as: given the same input, how similar are the learned representations across training? Which model architectures have the highest impact upon similarity? How does neural similarity correspond to model performance? Using correlation analysis techniques from [13] and [12], the key contributions are summarised as follows:

- Development of a framework for correlation analysis across neural representations within state-of-the-art End-to-End architectures, Section 3.1.
- A comparative analysis of similarity indexes upon an ASR task for different architectures, Sections 3.2 and 3.3.
- Verification that internal representation analysis of End-to-End network structures can be used to magnify pathological components throughout model training, Section 3.3.
- Discussion of the distinctions regarding the representations of End-to-End architectures and possible further work these observations can contribute to, Section 4.

http://dx.doi.org/10.21437/Interspeech.2021-1516

2. Similarity Indexes for End-to-End ASR

The End-to-End ASR task is to identify the acoustic input sequence $X = \{x_1, ..., x_T\}$ of length T as a label sequence $Y = \{y_1, ..., y_N\}$ of length N and directly map to the posterior distribution $p(Y|X)$. Due to the undefined separation of modules within End-to-End ASR networks, it is relatively unclear which, what and where the traditionally separate ASR system tasks are occurring, such as acoustic or language modelling. The internal parameter dependencies upon the structures of the model, and their effect upon the resulting performance, are ambiguous and inherently complex.

Using statistical correlation analysis methods, it is possible to relate two sets of observations within a network to find their correlation relationship. For the dataset $X = \{x_1, ..., x_N\}$ and neuron i in layer l, the activation output vector $z_i^l = (z_i^l(x_1), ..., (z_i^l(x_N))$. By conducting correlation analysis techniques that are invariant to affine transforms, this enables comparisons between different neural networks and observations on the dynamic behaviour.

2.1. SVCCA

SVCCA [12] is used to find bases w, s for two matrices such that, when the original matrices are projected onto these bases, their correlation is maximised:

$$\frac{w^T \sum_{XY} s}{\sqrt{w^T \sum_{XX} w} \sqrt{s^T \sum_{YY} s}} \tag{1}$$

where $\sum_{XX}, \sum_{XY}, \sum_{YY}$ are the covariance and cross-covariance. In the case of ASR neural networks, this is between the neural layers for N data points where $l_1 = \{z_1^{l1}, ..., z_{N1}^{l1}\}$ and $l_2 = \{z_2^{l2}, ..., z_{N2}^{l2}\}$. The projected views of l_1 and l_2 are reduced to the top 99% representative dimensions, using SVD, in an attempt to reduce potential noise in the representations, to form subspaces $l_1' \subset l_1, l_2' \subset l_2$. CCA [16] is then used to maximise the correlation of the projections of the linear transform of l_1', l_2' by identifying vectors w, s to maximise:

$$\rho = \frac{\langle w^T l_1', s^T l_2' \rangle}{||w^T l_1'||\ ||s^T l_2'||} \tag{2}$$

The correlations of ρ are higher when the representations have encoded more similar information.

2.2. CKA

CKA, first introduced in [17], resembles CCA but is weighted by the eigenvalues of the corresponding eigenvectors. It is also similar in effect to SVCCA but incorporates the weighting symmetrically and doesn't require matrix decomposition. Instead of comparing multivariate features of the neural layers, the coefficiency between every pair of examples in each representation is measured, then the correlation computation is conducted. To measure the similarity index between the internal representations, the inner product is taken:

$$\langle vec(l_1 l_1^T), vec(l_2 l_2^T) \rangle = tr(l_1 l_1^T l_2 l_2^T) = ||l_2^T l_1||_F^2 \tag{3}$$

where the elements of $l_1 l_1^T$ and $l_2 l_2^T$ are dot products between neural representations $z(x_1), ..., z(x_T)$, using calculations from [13]. The first half of Equation 3 measures the similarity between examples while the second half has the same result by measuring between features by taking the sum of the squared dot products between every pair. For centered l_1, l_2:

$$\frac{1}{(n-1)^2} tr(l_1 l_1^T l_2 l_2^T) = ||cov(l_1^T, l_2^T)||_F^2 \tag{4}$$

The Hilbert-Schmidt Independence Criterion (HSIC) [18] generalises Equations 3 and 4 to the inner products from the kernel spaces, where the squared Frobenius norm of $\sum_{XY}$ becomes the squared Hilbert-Schmidt norm of the operator. This is equivalent to calculating a distance covariance. Where $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ where k and l are kernels, the estimator of HSIC is defined as:

$$HSIC(K, L) = \frac{1}{(n-1)^2} tr(KHLH) \tag{5}$$

where H is the centering matrix $H_n = I_n - \frac{1}{n} 11^T$. When $HSIC = 0$, this suggests independence when k and l are universal kernels. However HSIC is not invariant to scaling until it has been normalised:

$$CKA(K, L) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K) HSIC(L, L)}} \tag{6}$$

These properties limit the use of CKA to be conducted across features rather than examples, for large models due to the size and the dramatically large computational costs that would be required.

3. Experiments & Results

3.1. Experimental Framework

To analyse the internal representations of the models on an End-to-End ASR task, the experiments were done using the ESPRESSO framework [19]. Each network was trained using the Switchboard dataset [20] with 300 hours of transcribed speech. This enabled the development of consistent state-of-the-art architectures and environment variables for End-to-End ASR.

To investigate the time dependencies of the neural representations within End-to-End models, SVCCA and CKA, described in Section 2, were applied to the activation outputs of each model layer across time. In order to conduct a comparable correlation analysis for each network and analysis method, several further steps were necessary: firstly, the network models were preserved at each epoch of the ASR task; and then they were fed to a separately developed pipeline for the extraction of the activation embeddings for each neuron. To ensure consistency, this was done by feeding in a controlled input of 100 speech frames to all architectures and extracting the activation output at each neuron, enabling the representation analysis methods to be conducted concurrently. Linear interpolation of the narrower layer to the same dimensionality as the wider layer was conducted due to the different spatial dimensions of the neural layers and thereby data-points, as both SVCCA and CKA methods require representation vectors to be the same dimensions. To compare the coefficiency correlation across the number of layers in the network, the spatial dimensions of the activation outputs were flattened into the number of data-points, in order to provide a spatial representation of each data-point.

3.2. Encoder-Decoder Neural Representation Analysis

The convolution network used in this work is a multi-layer stacked 2-dimensional convolution, with kernel size (3,3) on both the feature and the time axis from [21]. The final layer of

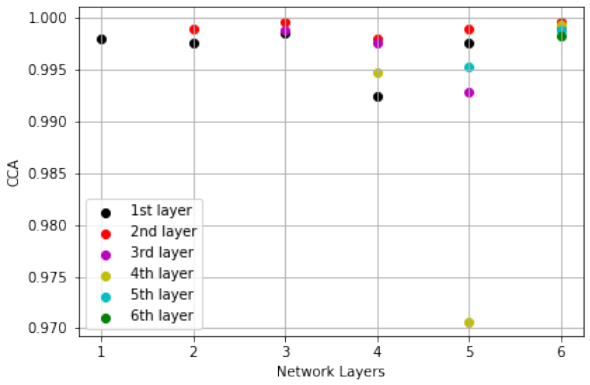

Figure 1: *The final converged correlation coefficiency across all models with ascending CNN layers*

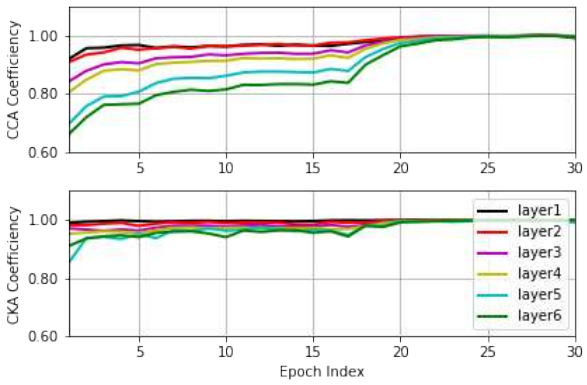

Figure 2: *CNN neural representations evaluated with SVCCA (**top**) and CKA (**bottom**) through time as performance converges*

the sequence-to-sequence model is then projected to an LSTM decoder from [22], and context at each time-step is generated with Bahdanau attention [23]. The encoder-decoder model function can be described by:

$$\hat{y}_u^{dec} = LSTM(c_u, y_{u-1}, \hat{y}_{u-1}^{dec}) \tag{7}$$

where $\hat{y}$ is the hypothesised output by the model, c_u is the context vector obtained by the encoder output, calculated by the attention mechanism.

This architecture allowed the observation of the layer-wise representation analysis methods across scaled convolutional layers within an ASR task. Comparison across layers allows the observation of the converged layer correlations, while comparison across epochs shows the hierarchical representations within the layers as the models train. Upon evaluation with the Hub5'00 set [24], the word error rate performance of this architecture is displayed in Table 1. Increasing the neural depth improved accuracy slightly up to 3 layers but results varying the spatial dimensions of each layer showed little improvement. The performance was observed to be limited when varying the amount of neurons in each layer across the variable sized CNN architectures, with the best WER performance achieved with a 3 layer CNN. Upon comparing these coefficiency correlations across the CNN architectures with varied layers, shown in Figure 1, the architectures with deeper spatial dimensions have more variation in the neural representations than the architectures that had the greater performance results, as suggested by results in [13], except for the 6 layer CNN model which had similar coefficiency but worse performance than the other models; this would require further investigation.

Table 1: *Variable sized CNN layers for End-to-End ASR evaluated on the Hub5'00 test set*

CNN Architecture	SWBD WER%	Clhm WER%
6 layers	11.4	22.4
5 layers	10.7	21.3
4 layers	10.9	21.2
3 layers	**10.5**	**20.8**
2 layers	10.6	20.9
1 layer	11.6	22.5

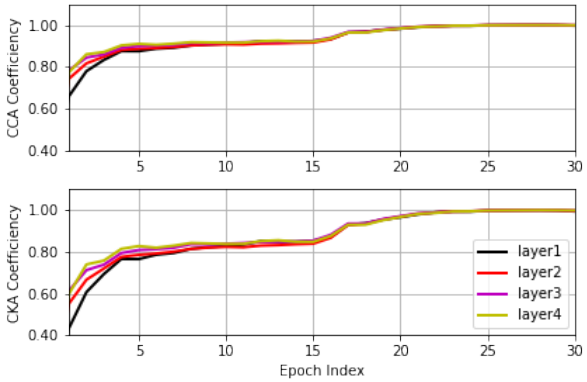

Figure 3: *LSTM correlation coefficients of neural representations through time as performance converges with SVCCA (**top**) and CKA (**bottom**)*

As the number of the layers increased, the coefficiency of each layer approached 1, as shown in Figure 2. Using SVCCA analysis, described in Section 2.1, to correlate the activations across the epochs, it was observed that layers 1, 2 and 3 converge together at epoch 17, whereas deeper layers (layers closer to the output) converged slightly later but at the same point in training.

Figure 2 also shows the CKA coefficiency, described in 2.2, of the CNN architecture, where it can be generally observed that the SVCCA analysis is more sensitive to the initialisation parameters than CKA. With both strategies, a hierarchical correlation within the layers across training can be observed, although the CKA results suggest that there is some pathological behaviour present in deeper layers; for example the small spikes in coefficiency across layer 6. The CKA results potentially differ from the SVCCA results, due to the pruning of the SVD component of SVCCA while also assuming that all the coefficiency vectors are equally important to the representation of the ASR task.

The LSTM neural representations, within the encoder-decoder framework, are displayed in Figure 3. Comparing the SVCCA correlation results with the CKA results, it can be observed that the variance is slightly under-estimated by the SVCCA implementation, although both techniques display similar attributes. By comparing the internal representations with

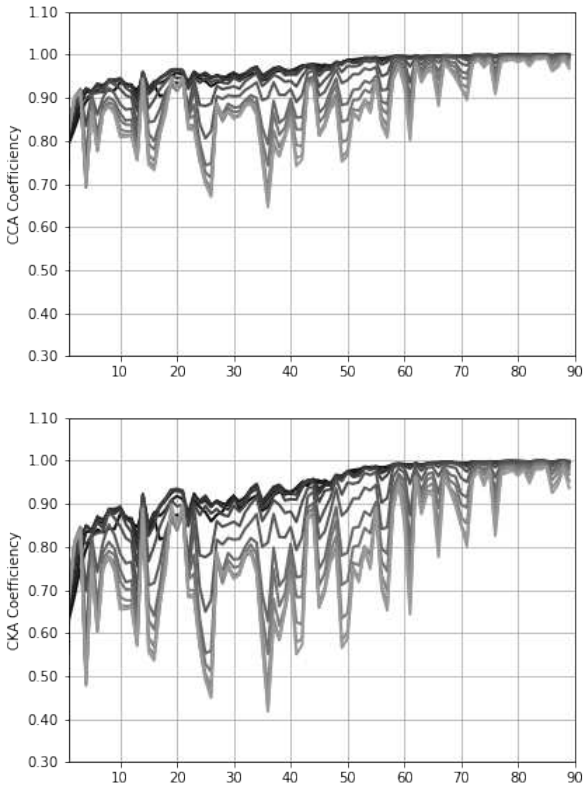

Figure 4: *Transformer correlation coefficients through time as performance converges produced with SVCCA (**top**) and CKA (**bottom**), the darker colour gradients are higher layer representations, while the lighter the gradient the deeper the layer*

SVCCA and CKA, the behaviour of the internal neural dynamics of the architecture can be observed to be invariant to transformations, in a robust method. The coefficiency across epochs suggests that there is a bottom-up behaviour within the LSTM representations, with convergence occurring around epoch 22.

3.3. Transformer Neural Representation Analysis

The Transformer architecture from [25] was trained using the same dataset, with all 12 encoder blocks containing identical spatial widths. Due to the size of this architecture, the training was conducted across 90 epochs, to ensure model convergence. The Transformer model uses stacked self-attention and point-wise, fully connected layers for the encoder and decoder. Each block has a multi-head self-attention layer and feed forward layer. To analyse the representations of the Transformer encoder layers, the representations were unrolled across time steps.

The Transformer layer encoder output representations, shown in Figure 4, emphasise the attending mechanism pathology present after the self-attention and linear operator. It can be observed that the higher layers of the Transformer encoder are less susceptible to the attention pathology than the deeper layers, which don't converge smoothly even after 80 epochs. There is a more noticeable distinction that can be ascertained from the CKA analysis, which retained more emphasised results, that there is similar overall hierarchical learning dynamics as observed in the CNN architecture.

4. Discussion

By using SVCCA as a method of analysing the internal representations for an End-to-End ASR framework, a window is observed on the dynamics of the training behaviour, although the top vectors pruned appear to under-represent the neural representations compared to using CKA. This is partly due to the assumption that all of the CCA vectors are equally important to the neural representation but also the SVD component of the SVCCA technique in Section 2.1 relies on the reflection of class information, which, for End-to-End speech recognition, is a potential limitation. By implementing the CKA analysis method in Section 2.2 it is possible to visualise the pathology of neural representations during training, particularly in the Transformer model, Figure 4, which is indicative of the attention mechanism augmenting context information, particularly for wider layers.

The techniques described in Section 2, allow the observation of hierarchical behaviour of CNN and Transformer neural representations across training, Figures 2 and 4, whilst also providing insight on the bottom-up invariant behaviour dynamics within the LSTM layers (without residual connections), shown in Figure 3. The learning dependencies between layers across time exhibit similar learning dynamics as language models [26]. The similarity indexes could also be used in future work to compare the correlation of the trained neural layers of various architectures across different speech datasets, such as noisy or augmented data, to observe how the neural layers respond dynamically during the training process. These experiments could then be directly correlated with the performance results.

Additionally, it has been noticed that scaling the depth of the convolutional layers had a limited effect upon network performance also in the case of End-to-End ASR, as shown in Table 1. Expanding the results from [27], Figure 1 provides some evidence that better performing networks converged to similar solutions across the layers, however the 6 layer CNN showed this is not always the case. The poorer performance of the 6 layer CNN could be attributed to an over-fitting issue and to investigate this further, the potential memorisation within the neural representations would be need to be undertaken. These results can be expanded to further develop and explore better architecture solutions for End-to-End ASR performance, whilst gaining some insight of the effect architecture changes have upon network dynamics.

5. Conclusion

A comparative analysis of SVCCA and CKA has been undertaken for an End-to-End task and pathological components have been identified in CNN and Transformer models. Further investigation of the attributes for the pathology would be required, for instance, do the unstable deeper layer neural representation correlations correspond to noisy components within ASR task? Furthermore, an extension to this work could be the analysis of neural representations on out of domain data, as the structural properties of the different layers could be beneficial to building models for few-shot-learning in ASR.

6. Acknowledgements

This work was partly supported by Voicebase Inc. at the Voicebase Research Center.

7. References

[1] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[3] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[4] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[5] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[6] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[7] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," *arXiv preprint arXiv:1402.1869*, 2014.

[8] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 233–242.

[9] S. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, 2020.

[10] S. Furui, "Generalization problem in asr acoustic model training and adaptation," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 1–10.

[11] A. Ogawa and T. Hori, "Asr error detection and recognition rate estimation using deep bidirectional recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4370–4374.

[12] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," *arXiv preprint arXiv:1706.05806*, 2017.

[13] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.

[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[15] N. Saphra and A. Lopez, "Understanding learning dynamics of language models with svcca," *arXiv preprint arXiv:1811.00225*, 2018.

[16] B. Thompson, "Canonical correlation analysis," *Encyclopedia of statistics in behavioral science*, 2005.

[17] N. Shawe-Taylor and A. Kandola, "On kernel target alignment," *Advances in neural information processing systems*, vol. 14, p. 367, 2002.

[18] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *International conference on algorithmic learning theory*. Springer, 2005, pp. 63–77.

[19] Y. Wang, T. Chen, H. Xu, S. Ding, H. Lv, Y. Shao, N. Peng, L. Xie, S. Watanabe, and S. Khudanpur, "Espresso: A fast end-to-end neural speech recognition toolkit," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 136–143.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[21] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4845–4849.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[24] R. Sundaram, A. Ganapathiraju, J. Hamaker, J. Picone *et al.*, "Isip 2000 conversational speech evaluation system," in *Speech Transcription Workshop, College Park, Maryland, USA*, 2000.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[26] S. Mittal, A. Lamb, A. Goyal, V. Voleti, M. Shanahan, G. Lajoie, M. Mozer, and Y. Bengio, "Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6972–6986.

[27] A. S. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," *arXiv preprint arXiv:1806.05759*, 2018.