# INTERPRETING INTERMEDIATE CONVOLUTIONAL LAYERS IN UNSUPERVISED ACOUSTIC WORD CLASSIFICATION

*Gašper Beguš and Alan Zhou*

University of California, Berkeley
{begus,azhou314}@berkeley.edu

## ABSTRACT

Understanding how deep convolutional neural networks classify data has been subject to extensive research. This paper proposes a technique to visualize and interpret intermediate layers of unsupervised deep convolutional neural networks by averaging over individual feature maps in each convolutional layer and inferring underlying distributions of words with non-linear regression techniques. A GAN-based architecture (ciwGAN [1]) that includes three convolutional networks (a Generator, a Discriminator, and a classifier) was trained on unlabeled sliced lexical items from TIMIT. The training results in a deep convolutional network that learns to classify words into discrete classes only from the requirement of the Generator to output informative data. The classifier network has no access to the training data – only to the generated data – which means lexical learning needs to emerge in a fully unsupervised manner. We propose a technique to visualize individual convolutional layers in the classifier that yields highly informative time-series data for each convolutional layer and apply it to unobserved test data. Using non-linear regression, we infer underlying distributions for each word which allows us to analyze both absolute values and shapes of individual words at different convolutional layers as well as perform hypothesis testing on their acoustic properties. The technique also allows us to tests individual phone contrasts and how they are represented at each layer.

***Index Terms*—** unsupervised acoustic word embedding, interpretable deep learning, ASR, generative models, generalized additive mixed models

## 1. INTRODUCTION

Several prominent speech processing models involve deep convolutional networks, both in supervised [2, 3, 4] and unsupervised settings [5, 6, 7]. Models using CNNs achieve high performance for phone and vocabulary recognition tasks, but substantially fewer studies focus on interpretability of the networks.

### 1.1. Prior work

Interpreting and visualization of intermediate convolutional layers in CNNs has primarily focused on the visual domain [8], but techniques have recently been proposed for speech data as well [9, 10, 11, 12]. Most proposals focus on visualizing and interpreting filters of supervised models [13, 14, 15, 10, 9, 16, 17]. [14, 15, 10] visualize filters of models trained on spectrograms; [11] focuses on relevance maps using gradient-based visualization. [18] focuses on activation maps and compare them to brain data. [19] cluster feature representations of each layer in a DeepSpeech model.

### 1.2. Goals

Unlike most previous proposals, our work uses unsupervised models trained on raw acoustic data. We are primarily interested in how representation of linguistically meaningful units self-emerges in unsupervised generative models. Additionally, instead of on filters, we focus on visualizing feature maps and argue that they yield highly informative time-series data that allows acoustic analysis on the intermediate layers. Finally, we introduce non-linear regression techniques to the study of learned intermediate representations in CNNs.

This work also falls in line of the larger scope of unsupervised acoustic word embedding [20, 21, 22, 5, 6, 7]. The majority of proposals in this framework operate with variational autoencoders (VAEs). This paper is thus part of a larger attempt to build interpretable unsupervised acoustic word embedding models within the GAN framework [23, 1].

## 2. MODELS & TECHNIQUES

### 2.1. Models

We use Categorical and Featural InfoWaveGAN (ciwGAN and fiwGAN) [1]. CiwGAN/fiwGAN is an InfoGAN [24] extension of the WaveGAN [25] architecture which, unlike the InfoGAN proposal, features a separate Q-network that can model lexical learning. Like WaveGAN (based on [26] and [27]), the model consists of two networks, a Generator $G$ which attempts to generate audio samples given a latent distribution $z$, and a Discriminator $D$ which takes both Generator outputs $G(z)$ and real data $x$ as input and attempts to estimate the Wasserstein distance between the input and the distribution of the real data. Just as in WaveGAN, the Generator uses 1D transpose convolutions to upsample from a low-dimensional latent space to audio, while the Discriminator uses regular 1D convolutions to stride along the audio. FiwGAN differs from ciwGAN in the structure of latent code: ciwGAN takes a one-hot vector as the latent code ($c$), while fiwGAN introduces a new latent space structure: binary codes that allow featural learning [1].

Following [24], we aim to maximize the mutual information between specific latent codes $c$ and the generated outputs $G(z, c)$ by forcing an auxiliary distribution $Q(c|G(z, c))$ to be as close to the true distribution of $P(c|G(z, c))$ as possible. In InfoGAN, this is done by borrowing convolutional layers from the Discriminator to estimate the distribution $Q$. In ciwGAN/fiwGAN, however, we make use of an entirely separate Q-network that shares the convolutional

This research was funded by faculty development grants at UC Berkeley and University of Washington. We would like to thank Sameer Arshad for slicing data from TIMIT.

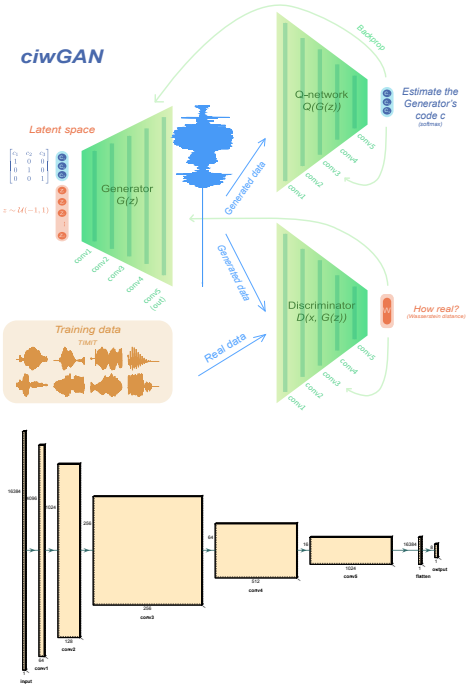


**Fig. 1**. (top) The architecture of the ciwGAN network [1] (based on [26, 25, 24]) schematized for three code variables $c$. A network with eight code variables $c$ was used in training. (bottom) The structure of the Q-network [1] with intermediate convolutional layers based on [25] (drawn with [28]).

structure as the Discriminator, the only difference being the dimensionality of the final layer. In addition to inducing the use of latent codes in the Generator, the Q-network can be seen as a classification network that aims to categorize audio into latent codes.

Separating the Q-network (on which we perform the visualization technique) from the Discriminator comes with several advantages. First, while the Discriminator comes into contact with both real and generated samples over the course of training, the Q-network only ever comes into contact with generated samples. This means that the Q-network never sees "real" data, making it unlikely to overfit so long as the Generator continues to produce diverse output. In addition, separating the Q-network follows from the differing objectives of the Q-network and Discriminator. While the Discriminator aims to estimate the input's distance from a distribution, the Q-network aims to categorize the input among several classes. This means that we can perform interpretation and visualization techniques on a classification network trained in an unsupervised manner.

WaveGAN also [25] introduces the phase shuffle technique. As the Generator makes exclusive use of transpose convolutions, the outputs that it generates contain periodic artifacts. As these artifacts often occur at the same phase, it is easy for the Discriminator to learn these artifacts, making the Discriminator's training objective trivial. To combat this, WaveGAN makes shifts the activations in each layer of the Discriminator by a random number of samples. We also make use of phase shuffle in the Q-network for the sake of consistency with the Discriminator.

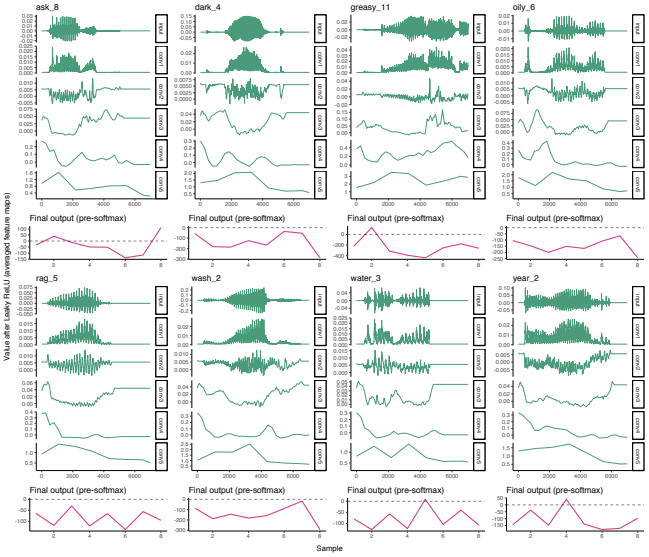


**Fig. 2**. Averaged and upsampled activations (0-6000) in each layer of the Q-network on an instance of each word from the test set. For each layer, going from top to bottom, we see activations in order from the earliest layers to the latest layers (green), followed by output of the final unnormalized logits in the last layer (purple). While earlier activations follow the stimulus quite closely, later activations become more abstract representations.

### 2.2. Visualization techniques

Following [29] which focuses exclusively on the Generator, we propose that averaging over individual feature maps in each convolutional layer after the ReLU activation in the Q-network yields highly interpretable time-series data that summarizes which linguistic properties are encoded at which layer. Specifically, for a convolutional layer $C$, we are able to obtain time-series data $t$ via

$$t = \frac{1}{\|C\|} \sum_{i=1}^{\|C\|} C_i \tag{1}$$

where $\|C\|$ is the total number of feature maps (or equivalently, the number of channels in the convolutional layer) and $C_i$ is the $i$th feature map.

We apply this technique on the Q-network by feeding the network new data withheld from training. We took samples of the word-slices withheld from the training process, and normalized them to fit with the scale of the Generator output. These normalized samples were then passed through the Q-network with phase shuffle disabled. The activations of each layer are averaged as per (1).

To infer the underlying distribution for each word in a given convolutional layer, we fit raw visualizations (averaged feature maps after ReLU) to non-linear regression models: generalized additive mixed models (GAMMs; [30]). This allows us to analyze both absolute values (parametric terms) and shapes (smooths) of acoustic words at various convolutional layers and to perform hypothesis testing on differences between words at different layers.

The proposed technique thus allows testing with inferential statistics of which linguistic contrast are encoded at what layer in an unsupervised deep convolutional network in which learning of linguistically meaningful representations needs to emerge in an unsupervised manner.

The technique also allows testing of how individual sounds or contrasts (such as difference between the two fricatives [s] and [ʃ] in *ask* and *wash*) are encoded. To test individual contrasts, we annotate phone boundaries in the input and visualize intermediate layers only for the interval that corresponds to the presence of that feature in the input (Section 3.2).

## 3. EXPERIMENTS

### 3.1. CiwGAN

The first network (ciwGAN) is trained on eight unlabeled words from TIMIT [31]: *ask*, *dark*, *greasy*, *oily*, *rag*, *wash*, *water*, and *year*. The smaller number of training words is chosen to increase interpretability.

Four-fifths (80%) of the sliced TIMIT words (altogether 4,052 tokens) were used in training; the remaining 20% (altogether 1,067 tokes) were withheld from the training process as test data. The words were sliced into individual files and padded with 25ms window of silence. During training, they are additionally padded with silence so that each item contains 16,384 data points (approximately 1s with 16kHz sampling rate) that constitutes the input to the Discriminator network and Q-network (Figure 1).

The model is trained with 8 categorical codes $c$ — one for each word — for 121,116 steps, after which collapse (common in GAN training) was observed.

### 3.2. Visualization & regression

To visualize how representations of words self-emerge in intermediate convolutional layers of the Q-network, we feed 1,067 test items (withheld entirely from the training) to the Q-network in the ciwGAN architecture and average over feature maps at each convolutional layer.

For the purposes of visualization against the original stimulus, we also upsample outputs at each layer to 16834 samples using linear interpolation. Examples of these averaged activations for individual words are shown in Figure 2.

Raw visualizations reveal that acoustic properties in the input are encoded relatively locally in the first few convolutional layers. For example, the visualization of *dark* in Figure 2 shows how bursts are encoded with positive values in Conv5, but with a depression in Conv4.

To infer underlying shapes of each lexical item, we fit raw visualizations into a generalized additive mixed model with Value at Conv5 as the dependent variable; word identity (treatment-coded with *ask* as reference) as the parametric term; and thin plate smooths for the Value of sample with by-word difference smooth, by-token random smooths, and correction for autocorrelation. We perform the analysis for the fifth convolutional layer because of its highly reduced dimensionality.

Estimates of the non-linear regression allow hypothesis testing on both the absolute values and shapes of words at different layers. The parametric coefficients in Table 1 estimate the absolute values of individual lexical items and their differences: value of *ask* is significantly different from 0 and differs significantly in absolute values from *greasy*, *oily*, and *year*, but not from other four words (see Table 1). In shape, however, which is estimated with thin plate smooths, *ask* differs significantly from all seven other words. Visualization of the smooths for each word suggest that the network represents each lexical item with a distinct shape in Conv5 which is then transformed into the final layer (Figure 2) that, after softmax, classifies each word into one of the eight classes.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0.6832 | 0.0215 | 31.7138 | $< 0.0001$ |
| Word=dark | -0.0409 | 0.0300 | -1.3626 | 0.1730 |
| Word=greasy | 0.6323 | 0.0314 | 20.1116 | $< 0.0001$ |
| Word=oily | 0.0811 | 0.0312 | 2.6021 | 0.0093 |
| Word=rag | -0.0240 | 0.0305 | -0.7870 | 0.4313 |
| Word=wash | 0.0603 | 0.0311 | 1.9425 | 0.0521 |
| Word=water | -0.0464 | 0.0307 | -1.5126 | 0.1304 |
| Word=year | 0.1298 | 0.0311 | 4.1712 | $< 0.0001$ |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(Sample) | 8.9700 | 8.9934 | 235.5282 | $< 0.0001$ |
| s(Sample):Word=dark | 8.9684 | 8.9936 | 159.7072 | $< 0.0001$ |
| s(Sample):Word=greasy | 8.9446 | 8.9886 | 231.4914 | $< 0.0001$ |
| s(Sample):Word=oily | 8.9562 | 8.9911 | 122.4444 | $< 0.0001$ |
| s(Sample):Word=rag | 8.9534 | 8.9904 | 110.0716 | $< 0.0001$ |
| s(Sample):Word=wash | 8.9625 | 8.9924 | 165.8256 | $< 0.0001$ |
| s(Sample):Word=water | 8.9309 | 8.9855 | 71.5510 | $< 0.0001$ |
| s(Sample):Word=year | 8.9613 | 8.9922 | 140.2448 | $< 0.0001$ |
| s(Sample,UniqueWord) | 5327.3250 | 9595.0000 | 4.8275 | $< 0.0001$ |

**Table 1**. Regression estimates of the model in Figure 3.

To test how individual sounds are encoded across the convolutional layers, we manually annotated the period of frication noise of [s] and [ʃ] in 120 instances of two lexical items, *ask* and *wash* (60 each). We extract those values of averaged feature maps (as per (1)) that correspond to the period of frication noise in the input and normalize time. We analyze layers up to the fourth layer (Conv1-4) as Conv5 features too few variables. We fit the data to a generalized additive mixed model.

The visualization of the [s]/[ʃ] contrast in Figure 4 suggest that the underlying distribution is encoded with a substantial difference in shapes of the two sounds in the first layer. In the second layer (Conv2), the difference in shapes ceases to be significant. Conv3 and Conv4 illustrate how a shape distinction translates into an absolute value distinction: at Conv4, the shapes of the two sounds are not significantly different ($F = 0.35, p = 0.56$), while their absolute values differ significantly ($\beta = -0.068, t = -4.83, p < 0.00001$).

### 3.3. FiwGAN on entire TIMIT

To test how the proposed techniques scales to larger corpora, we analyze intermediate layers in a fiwGAN network trained on all words from TIMIT (a pretrained network from [1]): 54,378 tokens of 6,229 unique words. The fiwGAN model allows highly reduced vector representation of lexical items [1]. The fiwGAN network is trained with 13 latent feature variables ($\phi$) which enables $2^{13} = 8,192$ unique classes. Lexical learning emerges despite the mismatch between the unique lexical items in TIMIT (6,229) and the number of unique classes allowed by the architecture (8,192) (for error rates of the Generator, see [1]). Unlike in the 8-word ciwGAN model, the test data in the fiwGAN model is not withheld from training. However, testing the Q-network on training data that only the Discriminator has access to is justified because the Q-network never accesses the training data in the first place (only the generated data).

The same test data as used for the 8-word ciwGAN model is used in the fiwGAN model for three words: *ask*, *greasy*, and *wash*. 391 tokens of the three words are fed to the model. Averaged feature maps after ReLU from Conv5 are fit to a generalized additive model. Absolute values of *ask* are significantly different from zero ($\beta = -0.095, t = -9.16, p < 0.00001$) and from absolute values of *greasy* ($\beta = 0.034, t = 2.25, p = 0.024$), but not from *wash* ($\beta = -0.013, t = -0.88, p = 0.38$). The shapes of both *wash* and *greasy*, however, differ significantly from *ask* as estimated with smooth terms: $F = 90.3, p < 0.00001$ for *greasy* and

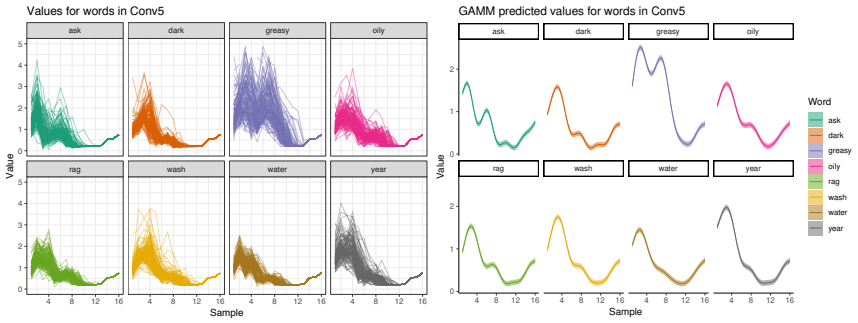


**Fig. 3**. (left) Raw values of the fifth convolutional layer (Conv5) for all tested words across the eight unique words (three datapoints are above the plot's upper limit). (right) Predicted values of a Generalized Additive Mixed Model (GAMM) [30, 32].

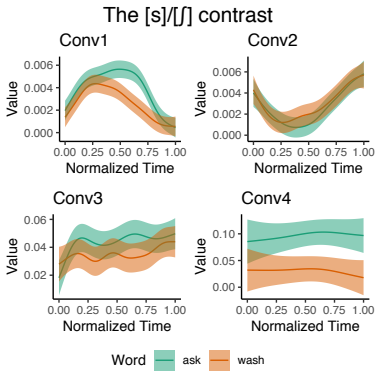


**Fig. 4**. Predicted values from a GAMM model at the first four convolutional layers for slices that correspond to the fricative part [s]/[ʃ] of *ask* and *wash* in the input.

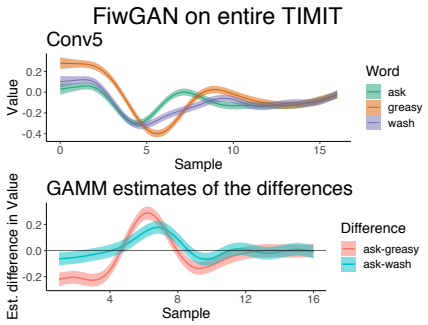


**Fig. 5**. (top) Predicted Conv5 values of the fiwGAN model trained on the entire TIMIT based on a GAMM model for each of the three words tested. (bottom) Difference plot estimating differences between pairs of words.

$F = 22.9, p < 0.00001$ for *wash*.

The difference plot in Figure 5 estimates at which sample points a pair of words differs significantly. We observe that the shapes for *ask* and *wash* differ substantially only in a relatively narrow band that likely corresponds to the identity of the fricative, because that is the most salient acoustic difference between the two words. Like in the ciwGAN model, the values for [ʃ] are significantly lower than values for [s].

## 4. DISCUSSION & CONCLUSION

Interpreting and visualizing how deep neural networks classify data has primarily focused on the visual domain. Here, we propose a technique to interpret and visualize intermediate convolutional layers when networks learn to classify words form unlabeled data in an unsupervised manner without ever having access to the actual training data in the ciwGAN/fiwGAN framework. This means learning representations of linguistically meaningful properties (such as words) needs to self-emerge in our models. We focus on applying inferential statistics — generalized additive mixed models — to infer underlying distributions of word representations. This allows inferential statistical tests of both absolute values and shapes of word representations at each convolutional layer.

The proposed technique opens up several ways to explore and interpret the relationship between the input, the latent space, and intermediate convolutional layers in unsupervised acoustic word embedding tasks. Any acoustic contrast can be tested and compared, both using smaller controlled setting that are more interpretable (e.g. models trained on a subset of words) as well as using models trained on entire speech corpora. The technique has the potential to serve as a diagnostic for detecting layers at which speech contrasts (such as phonemes) fail to get encoded. There are several further directions this work should take: from performing acoustic analysis on spectra of the averaged feature maps to exploring encoding of further phonemic contrasts in speech.

## 5. REFERENCES

[1] Gašper Beguš, "CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks," *Neural Networks*, vol. 139, pp. 305–325, 2021.

[2] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. Interspeech 2013*, 2013, pp. 1766–1770.

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Jul 2014.

[4] Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, and Aaron Courville, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 410–414.

[5] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020, pp. 1–12.

[6] Benjamin van Niekerk, Leanne Nortje, and Herman Kamper, "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," *Interspeech 2020*, Oct 2020.

[7] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[8] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 818–833, Springer International Publishing.

[9] J. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4989–4993.

[10] Andreas Krug and Sebastian Stober, "Visualizing deep neural networks for speech recognition with learned topographic filter maps," 2019.

[11] Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai-Doss, and Sébastien Marcel, "Understanding and Visualizing Raw Waveform-Based CNNs," in *Proc. Interspeech 2019*, 2019, pp. 2345–2349.

[12] Anurag Chowdhury and Arun Ross, "Deepvox: Discovering features from raw audio for speaker recognition in degraded audio signals," 2020.

[13] Mirco Ravanelli and Yoshua Bengio, "Interpretable convolutional filters with SincNet," 2019.

[14] Andreas Krug and Sebastian Stober, "Introspection for convolutional automatic speech recognition," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, Nov. 2018, pp. 187–199, Association for Computational Linguistics.

[15] A. Krug, René Knaebel, and S. Stober, "Neuron activation profiles for interpreting convolutional speech recognition models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language*, 2018.

[16] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings of Interspeech*. ISCA, 2015, pp. 11–15, ISCA.

[17] Pavel Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *Interspeech*, 2015, pp. 26–30,.

[18] Juliette Millet and Jean-Remi King, "Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech," 2021.

[19] Yonatan Belinkov and James Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 2438–2448, Curran Associates Inc.

[20] H. Kamper, A. Jansen, S. King, and S. Goldwater, "Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 100–105.

[21] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Interspeech 2016*, 2016, pp. 765–769.

[22] Cory Shain and Micha Elsner, "Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders," in *Proceedings of the 2019 NAACL-HLT, Volume 1*, Minneapolis, Minnesota, June 2019, pp. 69–85, ACL.

[23] Gašper Beguš, "Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks," *Frontiers in Artificial Intelligence*, vol. 3, pp. 44, 2020.

[24] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 2172–2180. Curran Associates, Inc., 2016.

[25] Chris Donahue, Julian J. McAuley, and Miller S. Puckette, "Adversarial audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019, OpenReview.net.

[26] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[28] Haris Iqbal, "Harisiqbal88/plotneuralnet v1.0.0," Dec. 2018.

[29] Gašper Beguš and Alan Zhou, "Interpreting intermediate convolutional layers of CNNs trained on raw speech," *CoRR*, vol. abs/2104.09489, 2021.

[30] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society (B)*, vol. 73, no. 1, pp. 3–36, 2011.

[31] J. S. Garofolo, Lori Lamel, W M Fisher, Jonathan Fiscus, D S. Pallett, N L. Dahlgren, and V Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1993.

[32] Márton Sóskuthy, "Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction," 2017.