INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Phonemic competition in end-to-end ASR models

Louis ten Bosch, Martijn Bentum, Lou Boves

Center for Language Studies/CLST, Radboud University Nijmegen
louis.tenbosch@ru.nl, martijn.bentum@ru.nl, lou.boves@ru.nl

Abstract

Advanced end-to-end ASR systems encode speech signals by means of a multi-layer network architecture. In Wav2vec2.0, for example, a CNN is used as feature encoder on top of which transformer layers are used to map the high-dimensional CNN representations to the elements of some lexicon. Compared to the previous generation of 'modular' ASR systems it is much more difficult to interpret the processing and representations in an end-to-end system from a phonetic point of view. We built a Wav2vec2.0-based end-to-end system for producing broad phonetic transcriptions of Dutch. In this paper we investigate to what extent the CNN features and the representations on several transformer layers of a pre-trained and fine-tuned model reflect widely-shared phonetic knowledge. For that purpose we analyze distances between phones and the phonetic features of the most-activated phones in the output of an MLP classifier operating on the representations in several layers.
Keywords: computational models, end-to-end audio decoding, broad phonetic classes, scientific understanding

1. Introduction

The recent emergence of so-called end-to-end systems (E2E), such as Wav2Vec 2.0 [1] (henceforth Wav2vec2), has revolutionized automatic speech recognition (ASR) in many ways. At the same time many researchers whose interest is not primarily to obtain the lowest possible transcription error rate are asking whether the representations on some or all layers of E2E models contain information that can be harnessed for other downstream tasks [2, 3, 4]. These approaches are collectively known as *probing*. In this paper we use a probing approach to investigate to what extent the latent representations on the layers of a Wav2vec2 model capture generally accepted phonetic knowledge, such as the fact that phones that are members of the same broad phonetic class (BPC) [5] are closer to each other than to phones that belong to a different BPC. This question differs fundamentally from asking if those representations can be used for phone classification.

ASR systems have long taken inspiration from models of human speech recognition and informed these in turn (e.g., [6, 7]). The relation between E2E-based ASR-systems and models of human speech recognition is less straightforward. An important concept within models of human speech recognition is the idea of competition (cf. [8] and references therein), which has both a temporal and a categorical (i.e. phonemes) dimension. Interestingly, competition in both dimensions is almost completely obscured in the discrete symbolic output of a Wav2vec2 model. In the current study, we focus on the role of competition on the categorical dimension in a Wav2vec2 model. We do so by using the information in the latent representations on the hidden layers of a Wav2vec2 model.

Many probing approaches take the representations on some layer of a deep neural network (DNN) as the data with which some classifier is trained for some specific task, such as phonetic feature extraction of phone classification. Previous research such as [2, 3, 4, 9] mainly focused on identifying the layer whose representations yielded the best classification performance. These studies show that phone classification is possible with convincing performance. However, their approach does not provide insight in the *structure* of the latent representations in terms of phonetic knowledge. Specifically, they do not provide insight in the competition between phones, a concept that plays an important role in models of human auditory word comprehension ([6, 10]) and in the classical cohort models ([11]). In this paper we develop probing methods that should do just that: highlight exactly how of phonetic information is structured in the latent representations, and how this structure may change when advancing from the lowest to the highest hidden layer.

In this study, we use an E2E model that is finetuned for broad phonetic transcription for Dutch. It has been shown that Wav2vec2 models can produce high-quality transcriptions on the phone level; [12] reports a phone error rate of 8.3% for the TIMIT test set. However, the final output of a Wav2vec2 model does not show how confusable phones compete for some of the probability mass. Typically, due to the CTC algorithm [13], the score distance between the winner and all competitors is very large, and the rank order of the competing phones may be seemingly random, which makes a Wav2vec2 model fine-tuned with CTC a good phone decision machine, but a poor machine when it comes to describing phonetic structure.

The probing techniques investigated in this paper aim to investigate the competition between phones in individual time frames. Phonetic theory predicts at least global aspects of the statistical distributions of the distances between different phones. The degree to which representations on specific layers of a Wav2vec2 system reflect that theory should shed light on the acoustic-phonetic structure captured in those representations. In the current study, we focus on the role of phone competition in an E2E model by investigating three related aspects: the Kullback-Leibler divergence between phone pairs, the within-frame phone ambiguity, and the interpretation of the competition between winning and runner-up phone in E2E in terms of 'classical' phonetic structure.

2. Searching for acoustic-phonetic structure

In order to obtain a first impression of how acoustic phonetic structure is covered in the representations on the hidden layers in a Wav2vec2 system, we decoded all utterances in the

10.21437/Interspeech.2023-1846

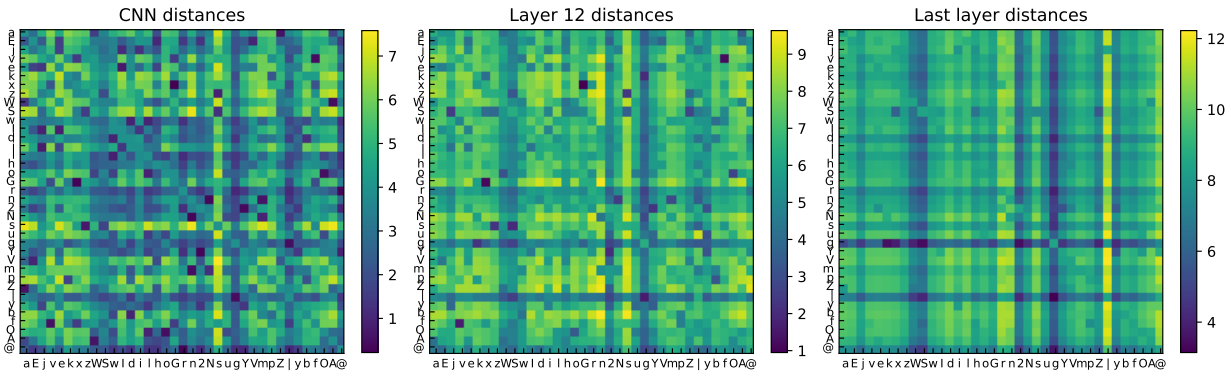

Figure 1: *KL distance between phone pairs. Left: CNN output layer; center: transformer layer 12; right: transformer layer 24.*

news broadcast component (component 'k') from the Spoken Dutch Corpus [14] ($\sim$ 27 hours of materials) with a properly fine-tuned system. This system was finetuned towards the classification of 37 phones in Dutch. In doing so, we stored the 1024-D representations on all hidden layers for all utterances. In addition, we stored the 512-D CNN representations, which are as close as we can come to the speech signals. We then used KMeans clustering to create codebooks with 512 entries, and replaced the 512-D or 1024-D frame vectors with the number of the closest code word (using Euclidean distance). For each of the 37 phone labels we constructed a histogram of the resulting code word counts from which we obtained 512-D vectors that describe $p(q_N|phone)$. Finally, we computed the Kullback-Leibler (KL) divergence (Eq. 1) between all pairs of phone probability vectors.

$$KL(P||Q) = -\sum_{x} P(x) \log(\frac{Q(x)}{P(x)}) \quad (1)$$

Figure 1 shows the (non-symmetric) KL divergence for all phone pairs for the representations on the CNN output layer and layers nr 12 and 24 in the transformer. It can be seen that on the CNN layer there is some evidence of acoustic-phonetic structure, but on the (higher) transformer layers that structure is much less apparent. Importantly, this does not mean that those representations contain weaker phonetic information. After all, [2, 3, 4] have shown that classifiers that use these representations for a range of downstream tasks outperform the results of similar classifiers that operate on conventional spectral representations. The figure shows that the acoustic organisation in the higher transformer layers differs from the lowest transformer layer, in line with the findings in the next sections.

3. Multi-layer perceptron phone classifiers

In our second method we trained MLP-based phone classifiers similar to the approach reported in other probing papers (e.g. [3]). First, we used both a pre-trained Wav2vec2 model without any fine-tuning, and a fine-tuned model trained on the core part of the read aloud books component (component o) in the Spoken Dutch Corpus [14]. For fine-tuning we used the manually corrected phonetic transcriptions and corresponding audio recordings ($\sim$7 hours of materials). Subsequently, we applied both the pre-trained and fine-tuned model on the complete set of read aloud book recordings ($\sim$ 64 hours of materials).

Based on the latent representations of both models, we trained for each layer a specific MLP phone classifier with the latent representation as input and the phone label from the fine-tuned Wav2vec2 model output as ground truth. We trained these classifiers only on those frames with a phone label in the output of the fine-tuned Wav2vec2 model. Table 1 presents the phone classification accuracy for a subset of the resulting MLP phone classifiers on an independent held out test set from component 'o' in the Spoken Dutch Corpus ([14]) (the other classifiers obtained similar results – all details are available on github). Table 1 serves as a sanity check for the MLPs used in follow-up experiments reported in the next sections. It shows that, globally, the accuracy increases with layer, in line with previous research. Interestingly, this does not only hold for the phone-based finetuned model but also for the original pretrained model. One observes relatively large differences between CTC-finetuned and pretrained model and the interaction with layer.

Table 1: *Classification accuracy of the MLP phone classifiers on a held-out test set in the Spoken Dutch Corpus [14].*

Layer	Pretrained	CTC
cnn	69 %	
1	84 %	83 %
12	92 %	94 %
24	94 %	98 %

4. Competition probing: KL-divergence with broad phonetic class distributions

Phonetic theory holds that phones within a Broad Phonetic Class (BPC) [5] are more alike compared to phones from a different BPC (e.g., /p/ is similar to /k/ but not to /a/). To investigate whether the representational structure of the hidden layers of the Wav2vec2 model respects the phonetic structure along BPCs we applied the Kullback-Leibler (KL) divergence as a dissimilarity measure between an observed PDF (generated by the MLP phone classifier) and a 'synthetic' phone PDF directly based on BPCs. Lower KL-divergence entails that the MLP-based PDF well matches the BPC-based PDF and that the underlying Wav2vec2 hidden state is in line with a categorical competition as in [5]. Conversely, higher KL divergence either means that most probability mass was assigned to phones outside the BPC or, alternatively, a flat distribution (i.e. assigning equal probability to each competitor phone).

The observed PDFs were obtained in the following manner. For a given model type (i.e. pretrained or ctc), frame and layer

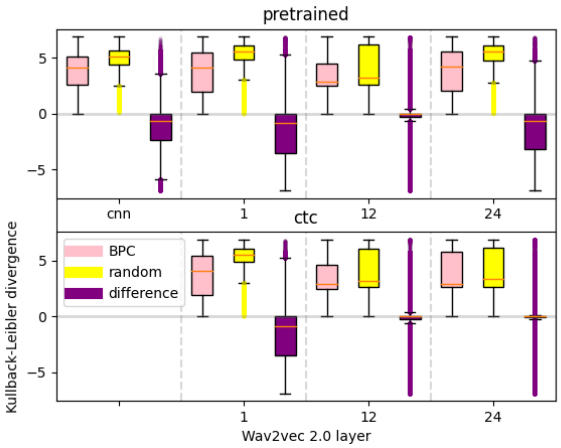

Figure 2: *The distribution of KL-divergences scores for BPC PDF, random PDF and the difference between the KL-divergence scores for BPC and random. The negative values are only present for the difference between the BPC and random based KL-divergence scores.*

we applied the corresponding MLP phone classifier to obtain a phone PDF. Subsequently, we removed the probability associated with the winning phone (since we are interested in the competition) and normalized the remaining probability vector (i.e. the set of competitor phones). In parallel, we constructed the synthetic phone PDF based on the winning phone's BPC by a step function, whereby all phones outside the BPC are assigned a small probability value ($p = .001$), and the remaining probability is divided equally among the set of BPC-phones (excluding the winning phone). Next, the KL-divergences was computed according to Eq. 1, with P for the MLP-based PDF and Q for the synthetic BPC-based PDF. For a sanity check, next to these BPC-based phone PDFs we also constructed phone PDFs based on *random* phone selections for the step function. These random phone sets provide a baseline comparison: if latent representations are phonetically faithful to BPCs, the KL divergence should be significantly lower for synthetic BPC-based phone PDFs compared to the synthetic PDFs based on random phone sets.

Based on [5], 7 BPCs were defined: **plosive** /ɡ b d p k t/, **nasal** /n m ŋ ɲ/, **approximant** /w j l ɾ/, **fricative** /x s f ɣ z v/, **high vowel** /i u ɪ y/, **mid vowel** /e ə ʏ ø ɛ ɔ o/, and finally **low vowel** /ɑ a/.

Figure 2 shows all KL divergence measures for a subset of the read aloud recordings from the Spoken Dutch Corpus ($\sim$ 20 hours of material, analysis on $\sim$ 3 million frames). It presents the distribution of the KL-divergence scores for BPC, random phone sets and the difference between them for the feature detector (CNN), transformer layer 1, 12, and 24 for both the pretrained and CTC fine-tuned Wav2vec2 model.

The figure shows that, across the board, there is a substantial difference between BPCs and the random phone sets in terms of the KL divergence with the observed MLP-based phone PDFs. Clearly, the comparison differs for the different layers in the Wav2vec2 model, and is different for the pretrained and finetuned variant. Interestingly, the KL-divergence does not differ for layer 24 from the CTC fine-tuned model, while the MLP phone classifier performed best in the phone classification task.

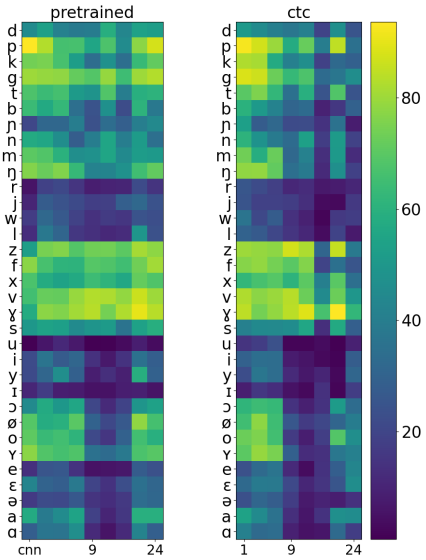

Figure 3: *The percentage of second best phones sharing the same broad phonetic class as the winning phone (y-axis), as classified by the multi-layer perceptron trained on a given layer (x-axis) of the Wav2vec2 model. Left: pre-trained right: ctc fine-tuned.*

5. Local and global structure

The results in the previous sections show that MLP-based classifiers perform adequately on the latent representations for all layers (Table 1, in line with [2, 3]), while the internal phone-phone structure shows a varying pattern across layers (Figs. 1, 2). This suggests that while phone classifiers are able to produce a likely correct winner, given a latent representation, a phonetically motivated phone-phone structure in the latent space is not necessary. In order to investigate this further, we study the competition between phones according to the MLP classifications. More explicitly, we zoom in into the statistics of the recognized phone and its runner-up, by counting the number of times a given phone appears as the runner-up of a winning phone. Figure 3 gives an pictorial overview of the percentage of runner-up phones sharing the same BPC as the winning phone, across all MLP-output vectors on a particular layer. For the plosives, nasals and fricatives this occurs fairly often even in the middle layers of the model, while this is not the case for the high, some middle and low vowels. The effect of fine-tuning can mainly been seen in the higher layers with a reduction of the percentage that the second best phone label is in the same BPC as the winning phone label.

Fig. 3 provides a glimpse on the intrinsic phone-phone competition per MLP input vector, that is, at a given location in the latent space. It raises the question to what extent these runners-up reveal information about the intrinsic phonetic structure. Evidently it can be expected that the overall statistics of the runner-up is informed by the phonetic neighborhood of the winning phone. To that end, we did an analysis on 3 million {winning phone, runner-up} combinations, layer by layer. After normalization for the absolute number of phones, the resulting histogram can be translated into a distance [15] between winning phone and runner-up, e.g. via

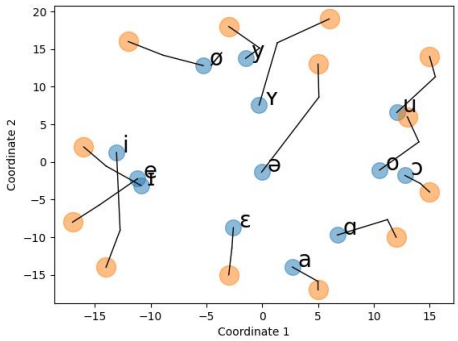

Figure 4: *MDS locations of Dutch vowels (monophthongs), after averaging over 50 different MDS runs; PCA-based corrective rotations are used before creating the overlay. The smaller light-blue markers denote the positions based on transformer layer 1; the larger light-orange markers denote the positions at layer 18; the connecting lines indicate the (simplified) route of the MDS solutions going from transformer layer 1 to transformer layer 18. Applicability of the Euclidean distance assumed.*

$$d \sim 1/(N+R) \tag{2}$$

in which d denotes the distance between winning phone and runner-up, N denotes the number of counts of the runner-up given the winner phone, and R is a regularisation term that corrects for very infrequent combinations. These infrequent combinations will certainly appear since, for example, /p/ will hardly occur as second candidate after /a/. In those cases, the formula predicts a large distance d between /a/ and /p/, and its numerical upper bound is determined by R.

Next, we use multi-dimensional scaling (MDS) [16, 17] to build a global 2D map, based on the phone-phone distances d resulting from Eq. 2. We used MDS rather than other visualization methods such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [18] since we focus on the local and global relations between phones instead of on clustering.

We carried out two MDS analyses, one for the vowels and one for the consonants. Both analyses provide a view of the structural change of the vowel and consonantal map, when moving upward through the layers. Figure 4 shows the MDS positions of vowels based on the distances obtained via Eq. 2 from the runner-up distribution in layers 1 and 18 of the transformer block, respectively. For layer 1 (represented by smaller light-blue markers) the vowel triangle (cf. [5]) is clearly respected, and the topology of the resulting configuration is very close to the Dutch vowel system. For higher layers, the MDS solution moves away from the configuration found for layer 1 – vowels are pushed apart towards a much more convex vowel configuration. This 'pushing apart' appeared a robust phenomenon for various initializations of the MDS algorithm. In terms of DNN-mappings (e.g. Theorems 10 and 11 in [19]), an increasing convexity of the configuration can be understood as preparing the latent representation for enabling the ultimate layer to make decisions using linear hyperplanes. A different but similar shift between plosives and fricatives is shown in [4], Fig 3.

Also in the case of consonants, we witness a substantial reorganisation of the internal structure of the group of consonants: also this group undergoes a global transformation as a function of layer. For consonants, the tendency towards a convex pattern is less obvious than for vowels.

6. Discussion and conclusion

This paper shows a potential paradox. One the one hand, the performance of MLP-based phone classifiers trained on transformer layers is adequate, as shown by recent papers and by our results. On the other hand, the latent representations appear to lose their connection with phonetic structure for the higher transformer layers. Figs. 1, 2 and 3 each provide different views of the (mostly weak) relation between the phonetic structure of phones and the latent representations in Wav2vec2.

The CNN provides an interpretable reflection of the phonetic structure in line with classical phonetic knowledge, but this structure changes for higher layers. Interestingly, the phonetic information itself in terms of phone-phone contrasts does not disappear: the MLP classification results are excellent also for higher layers. The relation between winning phone and runner-up, however, changes substantially across layers. While for the CNN output and lower transformer layers MDS provides interpretable vowel configurations, these configurations morph into a more circular convex pattern for higher layers. This reorganization is clearest in the case of vowels; the structure of the group of consonants also changes layer-by-layer but we could not detect clear patterns of convexity. Evidently, Fig. 4 only shows a glimpse of the underlying structural change, since it is a 2D representation of a morphing that actually takes place in a much higher dimension.

The results presented suggest that the Wav2vec2 model after finetuning using CTC is re-organising the acoustic-phonetic space in such a way that phones can still be classified using a spatially-oriented classifier, but the neighborhood structure in terms of *ambiguity* at a specific location is substantially changed. Apparently, strict obedience to classical phonetic structure in the latent space is not essential for a very good phone classification in that space. In the near future we aim to investigate this in more detail for other languages.

Although similarities between the neuronal structure of Wav2vec2 and the human brain might only exist at a high conceptual level (for a context see, e.g., [20]), it is worthwhile to note that [21], in their research on the computational aspects of the human superior temporal gyrus (STG) during speech perception, show that the neuronal processing in human brains does *not* directly show any relation with the structure of phones that is presumed in classical phonetics. Various recent neurophysiological studies suggest that speech processing in the human brain is based on a transformation from acoustic speech signals into internal representations of words mediated by nonlinear operations in the STG. The STG contains the nonprimary auditory cortex where phonological processing is known to take place. The STG represents the input speech via nonlinear processes that include spectro-temporal processing, normalization and restoration based on context, and involves complex auditory encoding for acoustic-phonetic and prosodic features. The high-level analogy between recent E2E ASR architectures on the one hand and STG-distributed patterns of neural activity on the other hand suggest that there is a parallel with respect to the extraction of abstract and higher-order phonemic and word-like representations from the speech signal [22, 21]. In future work, we will further explore this parallel with non-linear transformations in the STG. In this light, classical phonetics provides descriptive framework, rather than a firm basis for neurophysiologically plausible computational models.

The first author participates in the Dutch research NWA InDeep project, which aims at interpreting deep networks in speech, music and text processing.

7. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," *arXiv:2006.11477v3*, pp. 1 – 19, 2020.

[2] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 910–925.

[3] P. Cormac English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 83–91. [Online]. Available: https://aclanthology.org/2022.sigmorphon-1.9

[4] T. tom Dieck, P. A. Pérez-Toro, T. Arias, E. Noeth, and P. Klumpp, "Wav2vec behind the Scenes: How end2end Models learn Phonetics," in *Proc. Interspeech 2022*, 2022, pp. 5130–5134.

[5] P. Ladefoged and K. Johnson, *A Course in Phonetics (Seventh Edition)*. CENGAGE Learning, 2014.

[6] D. Norris and J. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review*, vol. 115, pp. 357 – 395, 2008.

[7] L. ten Bosch, L. Boves, and M. Ernestus, "DIANA, a process-oriented model of human auditory word recognition," *Brain Science*, vol. 23;12(5), 2022.

[8] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Words*. MIT Press, 2012.

[9] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," 2021. [Online]. Available: arXiv:2010.13007v4

[10] L. ten Bosch, L. Boves, and M. Ernestus, "DIANA, a process-oriented model of human auditory word recognition," *Brain Sciences*, vol. 12, no. 681, pp. 1–29, 2022.

[11] M. Taft and G. Hambly, "Exploring the cohort model of spoken word recognition," *Cognition*, vol. 22, no. 3, pp. 259–282, 1986.

[12] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," pp. 1–8, 2021. [Online]. Available: 10.48550/arxiv.2109.11680

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376.

[14] N. Oostdijk, "The design of the Spoken Dutch Corpus," in *Language and Computers, New Frontiers of Corpus Research*, P. Peters, P. Collins, and A. Smith, Eds. Rodopi, 2000, pp. 105–112.

[15] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms," *Computer Vision, Graphics, and Image Processing*, vol. 32, no. 3, pp. 328–336, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0734189X85900556

[16] A. Mead, "Review of the development of multidimensional scaling methods," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 41, no. 1, pp. 27–39, 1992. [Online]. Available: http://www.jstor.org/stable/2348634

[17] [Online]. Available: https://scikit-learn.org/stable/modules/manifold.html

[18] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, no. 9, pp. 2579 – 2605, 2008.

[19] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" *Proceedings of the International Conference on Learning Representations (ICLR)-2017*, 2017.

[20] B. C. Love, "The algorithmic level is the bridge between computation and brain," *Topics in Cognitive Science*, vol. 7, no. 2, pp. 230–242, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12131

[21] I. Bhaya-Grossman and E. F. Chang, "Speech computations of the human superior temporal gyrus," *Annual Review of Psychology*, vol. 73, no. 1, pp. 1–24, 2021.

[22] H. Yi, M. Leonard, and E. Chang, "The encoding of speech sounds in the superior temporal gyrus," *Neuron.*, vol. 102(6), pp. 1096–1110, 2019.