The representation of speech in deep neural networks

Scharenborg, Odette; van der Gouw, Nikki; Larson, Martha; Marchiori, Elena

DOI
10.1007/978-3-030-05716-9_16

Publication date
2019

Document Version
Accepted author manuscript

Published in
MultiMedia Modeling

Prefinal version

The Representation of Speech in Deep Neural Networks

Odette Scharenborg[1,2], Nikki van der Gouw[2], Martha Larson[1,2], Elena Marchiori[2]

[1] Multimedia Computing Group, Delft University of Technology, the Netherlands

[2] Radboud University, Nijmegen, the Netherlands
o.e.scharenborg@tudelft.nl

Abstract. In this paper, we investigate the connection between how people understand speech and how speech is understood by a deep neural network. A naïve, general feed-forward deep neural network was trained for the task of vowel/consonant classification. Subsequently, the representations of the speech signal in the different hidden layers of the DNN were visualized. The visualizations allow us to study the distance between the representations of different types of input frames and observe the clustering structures formed by these representations. In the different visualizations, the input frames were labeled with different linguistic categories: sounds in the same phoneme class, sounds with the same manner of articulation, and sounds with the same place of articulation. We investigate whether the DNN clusters speech representations in a way that corresponds to these linguistic categories and observe evidence that the DNN does indeed appear to learn structures that humans use to understand speech without being explicitly trained to do so.

Keywords: Deep neural networks, Speech representations, Visualizations.

1. Introduction

Recently, Deep Neural Networks (DNNs) have achieved striking performance gains on multimedia analysis tasks involving processing of images [1], music [2], and video [3]. DNNs are inspired by the human brain, which the literature often suggests to be the source of their impressive abilities, e.g. [4]. Although DNNs resemble the brain at the level of neural connections, little is known about whether they actually solve specific tasks in the same way the brain does. In this paper, we focus on speech recognition, which was one of the first multimedia processing areas to see remarkable gains due to the introduction of neural networks. We investigate whether a generic DNN trained to distinguish high-level speech sounds (vowels and consonants) naturally learns the underlying structures used by human listeners to understand speech. Speech is a uniquely useful area for such an investigation since decades of linguistic research in the area of phonetics provide us with a detailed and reliable inventory of the abstract categories of sounds with which human listeners conceptualize speech.

This paper is an exploratory study, and its contribution lies in the larger implications of its findings. Here, we mention two of these implications explicitly. First, insights into the extent to which neural networks learn human conceptual categories without being taught these categories may extend to other areas of multimedia, such as image or video, for which we lack the detailed structural characterization we have for speech. Second, insight into the ways in which neural networks fail to learn the same underlying categories used by humans could potentially point us to ways of improving speech recognition systems. Such insight is particularly valuable. Although today's automatic speech recognition systems have achieved excellent performance, they still perform much worse than human listeners when listening conditions are more difficult, e.g., when background noise is present or when speakers are speaking with an accent (cf. [5]).

The design of our investigation is straightforward. First, we train a naïve, generic feed-forward DNN on the task of vowel/consonant classification. We chose this task because it is a relatively simple and well understood task and will allow us to focus on what exactly a generic DNN learns when it is faced with the large variability of the speech sounds in the speech stream. Subsequently, we visualize the clusters of speech representations at the different hidden layers and observe the patterns that are formed. In analogy to visualization techniques used in the field of vision, e.g, [6],[7], we need to reduce the data to a lower dimension. We adopt the t-distributed neighbor embedding (t-SNE) algorithm in order to visualize the high-dimensional speech signal, typically two or three dimensions are used [8]. We choose t-SNE because of its previously shown usefulness for related tasks. For example, in [9], it has been successfully used to visualize the similarities between Mel feature cepstral coefficient (MFCC) feature and filterbank feature vectors created by deep belief neural networks (DBNs) to determine the most suitable feature vector as input representation for the DBN. The closest work to our own is [10], which visualizes how phoneme category representations in the hidden layers of a feed-forward network adapt to ambiguous speech. Our work is different in that we are not interested in individual phoneme categories, but rather train the network to distinguish vowels and consonants and observe the structures with which they emerge. They however visualize with principle component analysis (PCA) rather than t-SNE.

The paper is structured as follows. Section 2 describes the experimental framework including the data, the DNN architecture, as well as our t-SNE visualization. Section 3 presents the vowel/consonant classification results and the analysis of the speech representations learned by the DNN. Finally, Section 4 provides a brief discussion and our conclusions.

2. Experimental Framework

2.1. Speech data and labels

The DNN was trained using a selected subset of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN, [11]), which is a dataset containing nearly 9M words of

Dutch spoken in the Netherlands and in Flanders (Belgium) in 14 different speech styles. For the experiments reported here, we only used the read speech material from the Netherlands, and only the part from the so-called core corpus, which has a manual phonetic transcription of the speech signal. In total, our dataset contains 135,071 spoken words. The speech signal was transformed into 24 dimensional Mel Filterbank acoustic features calculated for every 10 ms..

The CGN uses 46 different phonemes for Dutch. Some phonemes only occur rarely in Dutch due to only being part of loan words. Since not enough training material is available for these phonemes, we mapped these rare phonemes onto similar Dutch phonemes. Such a mapping is common practice in automatic speech recognition. Table 1 lists all Dutch phonemes and their manner and place of articulation/tongue position. The phoneme label indicates the sound that is spoken.

Table 1. The phonemes of Dutch in the CGN with their manner and place of articulation label. Multiple phonemes in one cell are all mapped onto the first phoneme of that cell. Consonants are on the left-side and vowels are on the right-side of the table.

Phoneme	Manner of articulation	Place of articulation	Phoneme	Manner of articulation	Tongue position
p	Plosive	Bilabial	I	Short vowel	Front
b	Plosive	Bilabial	E, E~ , E:	Short vowel	Front
t	Plosive	Alveolar	A, A~	Short vowel	Central
d	Plosive	Alveolar	O, O~, O:	Short vowel	Back
k	Plosive	Velar	Y, Y~ , Y:	Short vowel	Front
g	Plosive	Velar	i	Long vowel	Front
f	Fricative	Labiodental	y	Long vowel	Central
v	Fricative	Labiodental	e	Long vowel	Front
s	Fricative	Alveolar	2	Long vowel	Central
z, Z	Fricative	Alveolar	a	Long vowel	Back
S	Fricative	Palatal	o	Long vowel	Back
x	Fricative	Glottal	u	Long vowel	Back
G	Fricative	Glottal	@	Long vowel	Central
h	Fricative	Glottal	E+	Diphthong	Front
N	Nasal	Glottal	Y+	Diphthong	Central
m	Nasal	Bilabial	A+	Diphthong	Back
n	Nasal	Alveolar			
l	Approximant	Alveolar			
r	Approximant	Labiodental			
w	Approximant	Labiodental			
j, J	Approximant	Palatal			

The manner and place of articulation are acoustic/phonological descriptions of the articulations of the phoneme. Vowels and consonants differ in the way they are produced primarily by the absence and presence, respectively, of a constriction in the vocal tract. In consonants, this constriction can occur in various ways, such as, a full closure

of the vocal tract followed by an audible release in the case of plosives (e.g., the /p/ in pot), or a narrowing of the vocal tract which results in an audible frication noise (e.g., the /s/ in stop). The amount of closure of the vocal tract in consonants determines the manner of articulation. Vowels are produced without a constriction in the vocal tract. Since Dutch has both long and short vowels, as well as diphthongs, rather than having one vowel class, we specified three vowel classes. The place of articulation indicates the location of the constriction in consonants. Since vowels do not have a place of constriction, we use position of the tongue to specify place of articulation, or rather tongue position on the front-back plane.

It is important to understand why we refer to manner and place of articulation as acoustic/phonological descriptions of speech sounds. They are acoustic in that manner and place of articulation do have a defining impact on the acoustic properties of the signal (i.e., its dominant frequencies). However, speech is highly variable, and acoustic properties are far from being unique to specific categories. Upon learning one's first language, human listener learn to associate certain acoustic variability with certain conceptual categories, in other words, phonological categories, e.g., an English native speaker will learn that the /ɛ/ (as in 'bed') and the /æ/ (as in 'bad') are different phonological categories, while a Dutch person will map the acoustic signal associated with both these categories onto a single category, /ɛ/, as Dutch does not have the /æ/ category. We expect a DNN to learn to leverage acoustics. The goal to our investigation is to start to understand whether a DNN also learns phonological categories similar to those used by human listeners.

2.2. Deep Neural Networks

The model used in the experiments is a feed-forward deep neural network, based on the architecture used in [12], but without the pre-training used in [12]. Another difference is that we used ReLU functions (following the rationale in [13]). The DNN consists of 3 fully connected hidden layers, each containing 1024 units with ReLU. The network is trained to optimize a cross entropy loss function for 20 epochs with batches of size 128 and an Adam optimizer with learning rate 0.0001. No dropout was applied.

The input to the DNN is a frame of 10 ms duration in a context of its five preceding and succeeding frames. The output layer consists of two units (consonant and vowel) with soft-max activation functions. Our CGN dataset was randomly split into 80% training set and 20% test set. Training was carried out 5 times with different random splits of the training and test set. Classification results are reported in terms of percentage of correctly classified frames, the frame accuracy rate (%FAR).

2.3. Visualizations

For the visualizations, we use the t-distributed neighbor embedding (t-SNE) algorithm to reduce the high-dimensional into two dimensions [8]. T-SNE places data points that are highly similar close to one another while placing data points that are less similar further apart. In the input layer, the activations are directly related to the input features. We therefore expect any clusters in the input layer to be directly related to the input

features. In later (hidden) layers, we expect this relationship between input features and clustering to become less strong and rather more abstract. We hypothesize that these clusters, instead, relate to the linguistic categories, specifically, phonemes and manner and/or place of articulation categories, and that clusters of these categories can be observed without the network having been specifically trained to recognize them.

To implement the visualizations, we randomly selected 1024 frames from our test set. To create the t-SNE visualizations, a learning rate of 10 and a perplexity of 25 were used. The algorithm iterated until no further changes were observed.

3. Results

3.1. Frame accuracy rates

The five trained networks had an average frame accuracy of 85.5% (SD = .003). The low SD shows that the randomly created training sets yield similar performance of the trained network. Looking at the two classes of the vowel/consonant classification task individually: 85.19% of the frames with the consonant label were classified correctly, and 86.69% of the frame with the vowel label. As automatic speech recognition systems typically recognize phonemes or words, we cannot directly compare our results to existing results. Nevertheless, we should note that these recognition rates are reasonable and as to be expected. The discussion that follows includes information on the reasons that certain errors occur.

Table 2. The frame accuracy rates per phoneme label.

Phoneme	Accuracy (%)	Phoneme	Accuracy (%)	Phoneme	Accuracy (%)
P	90.69	h	69.82	I	88.03
B	93.00	N	81.02	i	87.81
T	93.71	m	90.77	y	73.16
D	86.66	n	84.07	e	95.24
K	89.52	l	65.73	2	89.34
G	78.10	r	59.52	a	93.20
F	93.75	w	77.97	o	91.69
V	92.19	j, J	53.84	u	78.00
S	97.50	A, A~	90.14	@	75.34
Z, Z	92.37	O, O~, O:	88.59	E+	93.12
S	89.52	Y, Y~, Y:	83.03	Y+	90.39
X	93.91	E, E~, E:	86.31	A+	92.75
G	89.38				

We then calculated the frame accuracies per phoneme label and per manner and place of articulation label by labelling all correctly and incorrectly classified frames with their appropriate phoneme, manner of articulation and place of articulation label.

Table 2 shows the results per phoneme label. Generally, all phonemes were well classified, with the exception of /j/ and /r/ and to a lesser extent /l/, although for these phonemes classification performance was still well above chance. These sounds all belong to the manner of articulation category approximant. As Table 3 shows approximants are the manner of articulation category which have the lowest frame accuracy. This is not surprising as another word for approximants is half-vowels. The articulation, i.e., no clear constriction in the vocal tract, and consequently the spectral mark-up of approximants, or half-vowels, is very close to that of vowels. From a linguistic point of view, however, approximants cannot be the nucleus of syllables or stand-alone syllables where vowels can be, they are consonants. It is thus not surprising that approximants are relatively often misclassified as vowels.

Regarding the place of articulation/tongue position categories, the frames with the palatal label are most often misclassified in the vowel/consonant classification task. Most likely this is explained by the fact that this category only consists of two phonemes, one of which is the /j/ which, as we saw earlier, is not so well classified. Overall, the misclassifications in the vowel/consonant classification task seem to be fairly evenly distributed over the different phonemes and manner and place of articulation categories.

Table 3. The frame accuracy rates per manner and place of articulation/tongue position label.

Manner of articulation	Accuracy (%)	Place of articulation / Tongue position	Accuracy (%)
Plosive	91.00	Bilabial	91.11
Fricative	92.06	Alveolar	84.19
Nasal	85.67	Labiodental	87.96
Approximant	63.82	Velar	90.29
Short vowel	82.52	Glottal	69.82
Long vowel	90.89	Palatal	58.38
Diphthong	92.50	Front (vowel)	90.48
		Central (vowel)	80.28
		Back (vowel)	90.50

3.2. Visualizations

Vowel/consonant classification. We first investigated the clustering of the speech sounds in the input and the three hidden layers, and the relationship with the vowel and consonant labels, i.e., the task on which the DNN was trained. Fig. 1a. shows the representation of the original input of the network, with the blue dots corresponding to frames with the vowel label and the orange dots corresponding to the frames with the consonant label. We can observe a slight clustering structure in the representation of the vowels and consonants, with most of the consonants on the left side of the figure and most of the vowels on the right side. The clusters are however not well separated. Fig. 1b.-d. visualize the representations at the first, second, and third hidden layers of the DNN, respectively. As expected, in later layers the clusters become more compact

and more distinct, reflecting that the model is learning to create representations of the speech signal which abstract away the high variability of the speech signal. This is especially evident in the third layer, with one fairly well-defined vowel cluster and two consonant clusters, one of which is highly homogeneous and far removed from the rest.

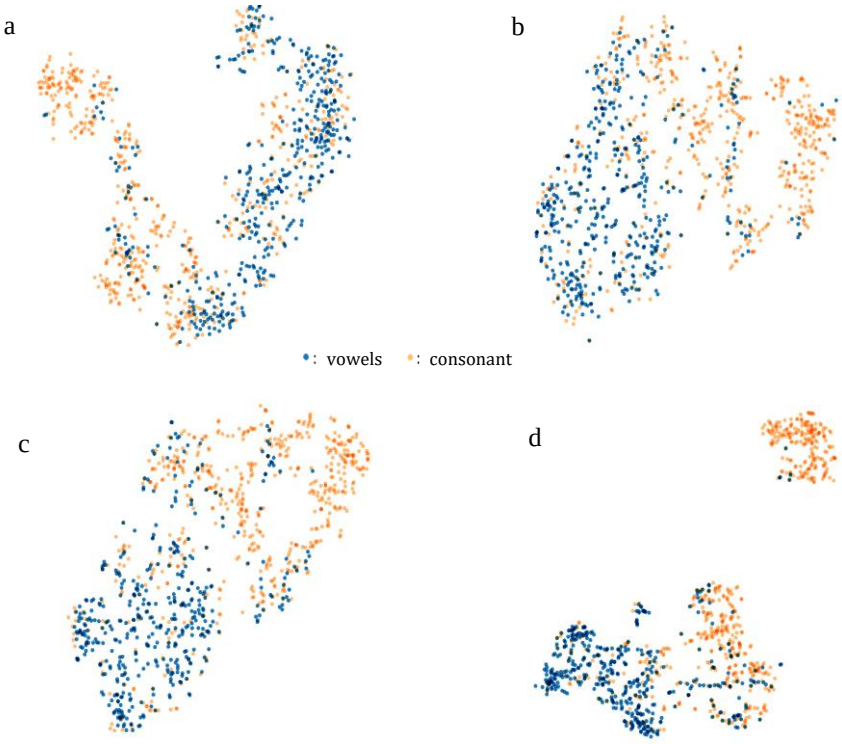

Fig. 1. Representation of the input frames at different layers of the model: a) input layer; b) hidden layer 1; c) hidden layer 2; d) hidden layer 3.

The intrusion of vowel labels in the consonant clusters and vice versa raise the question whether some of these frames are incorrectly classified. Fig. 2 shows the third hidden layer of the model again but now the correctly and incorrectly classified frames are indicated. The results clearly show that few of the frames in the cluster in the top right corner of the figure are misclassified. More errors are observed for the vowel and consonant clusters in the bottom of the figure, where indeed a number of the vowel frames which appeared in the consonant cluster are misclassified as consonant and vice versa.

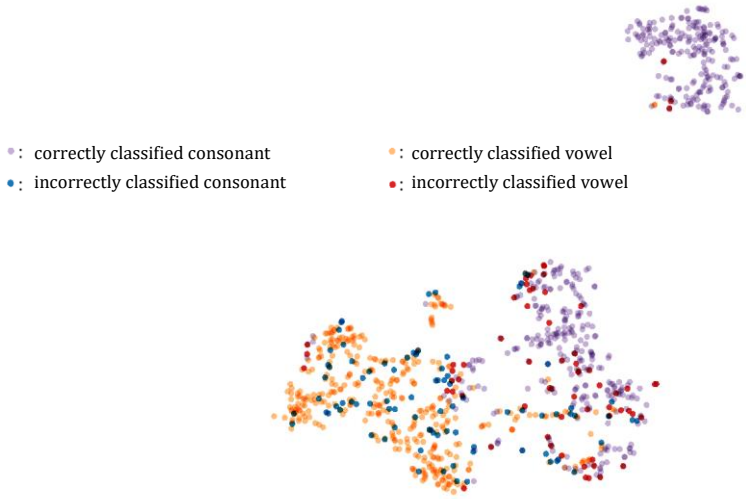

Fig. 2. Representation of the correctly and incorrectly classified vowel and consonant frames in the third layer of the model.

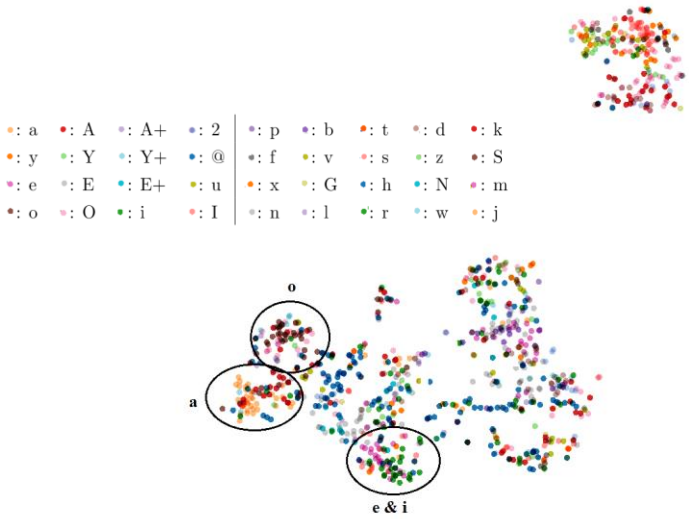

Fig. 3. Representation of the input frames labeled with their phoneme labels in the third hidden layer of the model.

Phoneme classification. Fig. 3 shows the third hidden layer of the model with the frames labeled with their ground-truth phoneme identity. Inspection of this figure sheds light onto the question how the network is learning. Despite the fact that the model was trained on the task of vowel/consonant classification, the network is implicitly learning clusters that are related to the way in which humans conceptualize speech. Specifically, we see that the phonemes are not randomly distributed within a larger cluster, but rather frames with the same phoneme label are clustered together. The cluster in the top right corner mainly consists of /p, t, k, z, s/. However, also in the larger cluster at the bottom of the figure, smaller clusters can be observed, with more consonants on the right of the big cluster and more vowels on the left of the big cluster. Three of these vowel clusters are indicated in the figure.

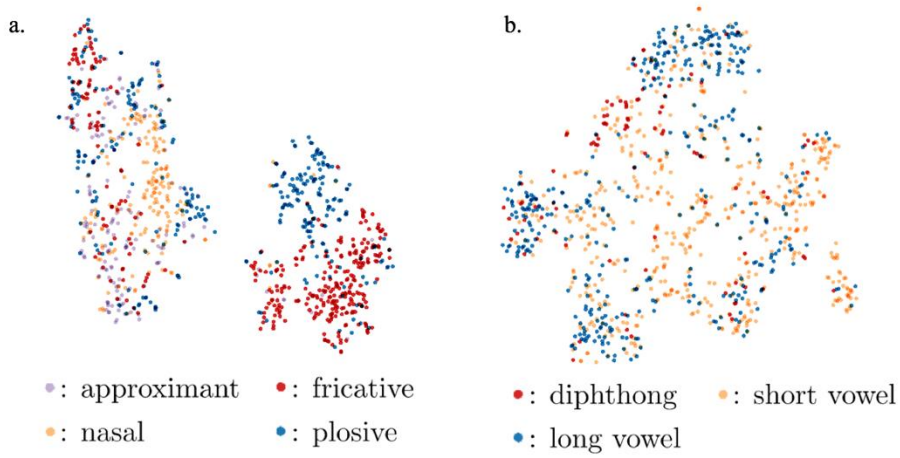

Fig. 4. Representation in the third hidden layer of the model of the input frames labeled with *manner of articulation*. Consonants are shown in a. and vowels in b.

Manner of articulation. Fig. 4 shows a visualization of the representations of the consonants (4a.) and vowels (4b.) at the third layer of the model. Each point represents a frame, and is color coded with the manner of articulation of that frame. Some of the frames are not correctly classified, but the numbers are relatively low (see Table 3), and we do not visualize the difference between correct and incorrectly classified frames in the figure.

The global picture arising in Fig. 3 regarding the smaller clusters in the larger cluster is confirmed by the visualization of the activations in the third hidden layer in terms of manner of articulation. With regards to the consonants in Fig. 4a., the frames in the right cluster almost exclusively belong to the fricative and plosive categories. From a linguistic perspective, these categories are clearly distinct in their production from the manner of articulation categories of the frames in the left cluster. Interestingly, within the plosive-fricative cluster on the right, there is a clear separation between the plosive (blue) and fricative (red) frames. So, even though the task of the DNN was to classify

vowels and consonants, underlyingly it represented the speech signal in similar speech clusters. We note that approximants and nasals share some vowel-like properties, and it is not surprising that they are less well separated than the consonants.

With regards to the vowels in Fig. 4b., we see that short and long vowels are not evenly distributed, but rather also have a tendency to group together. Diphthongs, however, show no particular pattern. Since diphthongs can be understood as a long vowel composed of a combination of two short vowels, their distribution is not surprising.

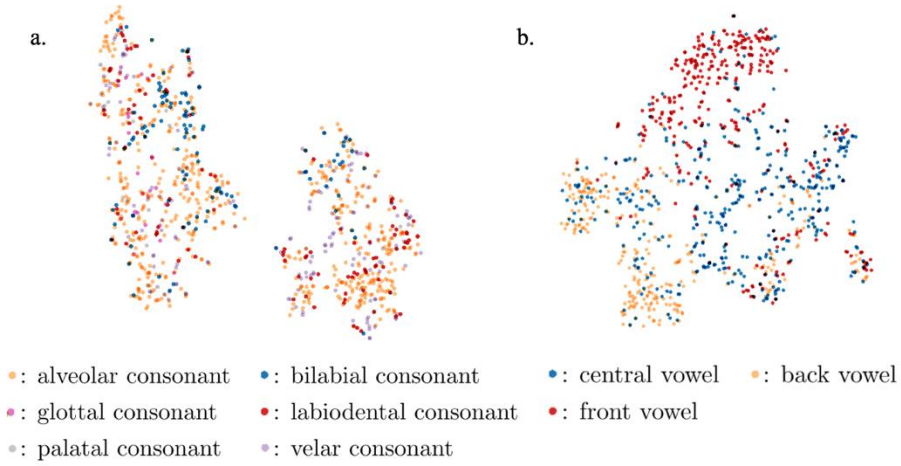

Fig. 5. Representation in the third hidden layer of the model of the input frames labeled with place of articulation. Left (4a.) are consonants and right (4b.) are vowels.

Place of articulation. Fig. 5 shows another visualization of the representations of the consonants (5a.) and vowels (5b.) at the third layer of the model. Each point represents a frame, and, this time, is color coded with the place of articulation of that frame. Compared to the labeling in terms of manner of articulation in Fig. 4a., fewer clear clusters can be observed for the consonants in Fig. 5a. This plot suggests that we cannot expect the DNN to learn all consonantal place distinctions that are relevant for human listeners. A human effortlessly distinguishes consonants that differ with respect to place of articulation. The DNN can achieve a relatively high classification rate for phonemes and place of articulation, while this is at the same time not well reflected in distinct clustering structure of the speech representations in the third hidden layer. In 5b. we see that the situation is different for vowels. Clusters corresponding to the back vowels (orange), the front vowels (red), and the central vowels (blue) are visible. Taking the analyses in terms of manner and place of articulation together, the visualizations seem to suggest that the model pays more attention to spectral information than to other information.

4. Discussion and conclusion

In this paper, we investigated whether a naïve, generic deep neural network-based ASR system can learn to capture the underlying speech representations by relating these speech representations to categories as defined in linguistics through the visualization of the activations of the hidden nodes. We trained a naïve feed-forward DNN on the task of vowel/consonant classification. Subsequently, we used different linguistic labels to visualize and investigate the clusters or speech representations at the different hidden layers.

There are two main findings. First, we established that our DNN was learning as we expected with the observation that the speech category representations became more abstract deeper into the model. So, like human listeners have been found to do [14], rather than storing all variation in the speech signal in the hidden layers, the variation was progressively abstracted away at subsequent higher hidden layers. Second, we moved beyond looking at categories which the DNN had been explicitly trained to recognize, to investigate whether the DNN would cluster the speech signal into linguistically-defined speech category representations (despite not explicitly been taught to do so) as are used during human speech processing. Indeed, underlyingly the model represented the speech signal in similar speech clusters, mostly grouping together consonants that have the same manner of articulation, while for vowels place of articulation seemed to be a good descriptor or explicator of the clusters.

A naïve DNN is thus not only able to deal with the large variability of the speech sounds in the speech stream but it does so by capturing the structure in the speech. In the future, we will move towards less naïve models to investigate whether these spontaneous emerging speech categories also emerge in other types of DNN architectures. If so, this would provide important insights into the optimal unit of representation of speech in automatic speech recognition. Specifically, we have seen that DNN models could possibly benefit from the incorporation of information on place of articulation for consonants, since this information does not seem to be implicitly learned during the training process. Moving forward, we expect that visualizations will continue to prove to be a useful tool in the investigation of how computers interpret multimedia signals and the connection between the ways in which computers learn to understand multimedia signals with the ways in which humans understand these signals.

Acknowledgements

This work was carried out by the second author as part of a thesis project under the supervision of the first, third, and fourth authors. The first author was supported by a Vidi-grant from NWO (grant number: 276-89-003).

References

1. Krizhevsky, A., Sutskever, I., and Hinton, G.E., ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12), Volume 1, 1097-1105 (2012).
2. van den Oord, A., Dieleman, S., and Schrauwen, B.: Deep content-based music recommendation. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13), Volume 2, 2643-2651 (2013).
3. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei Li: Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14), 1725-1732 (2014).
4. Ji Wan, Dayong Wang, Steven C.H. Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li: Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In Proceedings of the 22nd ACM International conference on Multimedia (MM '14), 157-166 (2014).
5. Juneja, A.: A comparison of automatic and human speech recognition in null grammar. Journal of the Acoustical Society of America, 131(3), EL256-261 (2012).
6. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In ECCV (1), 818-833 (2014).
7. Rauber, P.E., Fadel, S.G., Falcão, A.X., and Telea, A.C.: Visualizing the Hidden Activity of Artificial Neural Networks. IEEE Transactions on Visualization and Computer Graphics 23(1), 101-110 (2017). doi: 10.1109/TVCG.2016.2598838
8. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579-2605 (2008).
9. Mohamed, A.-R., Hinton, G., Penn, G.: Understanding how deep belief networks perform acoustic modelling. In: Proceedings of IEEE Acoustics, Speech and Signal Processing (ICASSP), pp. 4273-4276 (2012).
10. Scharenborg, O., Tiesmeyer, S., Hasegawa-Johnson, M., Dehak, N.: Visualizing phoneme category adaptation in deep neural networks. In: Proceedings of Interspeech, Hyderabad, India (2018).
11. Oostdijk, N.H.J., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., Baayen, H.: Experiences from the Spoken Dutch Corpus project. In: Proceedings LREC – Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, pp. 340-347 (2002).
12. Mohamed, A.-R., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing 20, 14-22 (2012).
13. Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Viet Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J.: On rectified linear units for speech processing. In: Proceedings of IEEE Acoustics, Speech and Signal Processing (ICASSP), pp. 3517-3521 (2013).
14. McQueen, J.M., Cutler, A., Norris, D.: Phonological abstraction in the mental lexicon. Cognitive Science 30(6), 1113-1126 (2006).