Visualizing Phoneme Category Adaptation in Deep Neural Networks

Scharenborg, Odette; Tiesmeyer, Sebastian; Hasegawa-Johnson, Mark; Dehak, Najim

DOI
10.21437/Interspeech.2018-1707

Publication date
2018

Published in
Proceedings of Interspeech 2018

Visualizing Phoneme Category Adaptation in Deep Neural Networks

Odette Scharenborg[1,2], *Sebastian Tiesmeyer*[1], *Mark Hasegawa-Johnson*[3], *Najim Dehak*[4]

[1]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[2] Multimedia Computing Group, Delft University of Technology, Delft, the Netherlands
[3] ECE Department & Beckman Institute, University of Illinois, Urbana-Champaign, IL, USA
[4]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
o.scharenborg@let.ru.nl

Abstract

Both human listeners and machines need to adapt their sound categories whenever a new speaker is encountered. This perceptual learning is driven by lexical information. The aim of this paper is two-fold: investigate whether a deep neural network-based (DNN) ASR system can adapt to only a few examples of ambiguous speech as humans have been found to do; investigate a DNN's ability to serve as a model of human perceptual learning. Crucially, we do so by looking at intermediate levels of phoneme category adaptation rather than at the output level. We visualize the activations in the hidden layers of the DNN during perceptual learning. The results show that, similar to humans, DNN systems learn speaker-adapted phone category boundaries from a few labeled examples. The DNN adapts its category boundaries not only by adapting the weights of the output layer, but also by adapting the implicit feature maps computed by the hidden layers, suggesting the possibility that human perceptual learning might involve a similar nonlinear distortion of a perceptual space that is intermediate between the acoustic input and the phonological categories. Comparisons between DNNs and humans can thus provide valuable insights into the way humans process speech and improve ASR technology.

Index Terms: phoneme category adaptation, human perceptual learning, deep neural networks, visualization

1. Introduction

Although the algorithms and functions computed by humans and machines in order to recognize speech are most likely different due to differences in the hardware between humans and machines, at the top level (also referred to as the *computational* level by [1]) both humans and machines carry out the same process, i.e., the recognition of words from the speech signal [2]. Consequently, comparing humans and machines on the same task can provide valuable insights into the way humans process speech and can improve automatic speech recognition (ASR) technology [3][4][5]. Knowledge about human speech processing has been used to improve ASR technology, e.g., in the creation of the acoustic features used in ASR systems (e.g., MFCCs [6], PLPs [7]), and to change the underlying approach to ASR, e.g., template-based approaches to ASR (e.g., [8]) are based on the episodic theory of human speech processing [9]. Conversely, ASR techniques have been used to investigate psycholinguistic questions regarding human speech processing: they have been successfully used in a speech-based computational model to test a theory about the use of fine-grained durational information in human speech processing [10], and to verify the hypothesis that infant-directed speech is easier to learn than adult-directed speech [11]. Here we aim to pursue both directions, and focus on the process of perceptual learning.

Perceptual learning is the temporary or more permanent adaptation of sound categories after exposure to deviant speech, in a manner which includes the deviant sounds into pre-existing sound categories, thereby improving intelligibility of the speech (see for a review [12]). Perceptual learning is crucial for both human and automatic speech processing as the listener, whether human or machine, needs to adapt its sound categories whenever, e.g., a new speaker, channel condition, dialect or speaking style is encountered. Humans have been shown to rapidly change sound category boundaries to include deviant sounds [13-16] after hearing only a few instances of the deviant sounds [17][18]. ASR systems, on the other hand, adapt to a new speaker and new listening conditions using both short-time adaptation algorithms (e.g., fMLLR [19]) and longer-term adaptation techniques (e.g., DNN weight training [20]). In both cases, lexical knowledge about the word in which the deviant sound occurs is crucial [12][13] to guide the listener in interpreting the ambiguous sounds. ASR adapts correctly if it correctly recognizes the word in which the deviant sound occurs (or if a correct word label is provided), and incorrectly otherwise. Humans adapt if there is only one interpretation of the deviant sound that renders the word intelligible; if multiple interpretations are possible, then adaptation does not occur.

We aim to answer the question whether DNNs can adapt to ambiguous speech as rapidly as human listeners; if they are shown to have this ability, then we hope to visualize the mechanism by which the DNNs perform this adaptation, in search of intermediate representations that might have correlates in human perceptual adaptation. Specifically, we investigate the adaptation of phoneme categories to include an ambiguous sound in between an [l] and [ɹ] sound after exposure to this ambiguous sound [l/ɹ], a process also referred to as lexical retuning [13] in the psycholinguistic literature. We use an experimental set-up that mimics that of the human listening experiment [16] to which we will compare the DNNs' performance. Crucially, when investigating the DNNs' ability of lexical retuning, we will not only look at the output level but rather at intermediate levels of phoneme category adaptation in the DNNs by calculating the average difference between the activations of the hidden nodes in response to [l], [ɹ], and ambiguous [l/ɹ] sounds, respectively, and by visualizing the activations of the hidden nodes in response to each training token. The proposed methodology opens up the 'black box' of the DNNs [21].

Many factors about the mechanisms underlying perceptual learning are clear (for a review, [12]). However, what is still unclear is what actually happens during perceptual learning to the phoneme categories in the brain. This is not easily observed (neuroimaging experiments have not yet demonstrated sufficient spatial resolution to monitor changes correlated with lexical retuning; psycholinguistic experiments can demonstrate the classification effects of lexical retuning, but have not yet been able to discriminate among hypotheses about the underlying cognitive representations). As a second aim, therefore, we propose to use the visualizations of the hidden node activations of DNNs carrying out the same task as human listeners to make suggestions about what might be happening in the human brain during perceptual learning.

2. Methodology

2.1. Lexical retuning in human speech processing

In a typical human perceptual learning experiment, listeners are first exposed to deviant, new speech (sounds), after which a test phase follows to investigate the influence of the deviant, new speech on the sound categories. Here, we base our experimental set-up on the lexical retuning paradigm [12][13][16]. In a lexical retuning paradigm, two groups of listeners are tested. Using the experiment from which we take our stimuli, [16], as an example: one group of Dutch listeners was exposed to an ambiguous [l/ɹ] in [l]-final words such as *appel* (Eng: *apple*; where *appel* is an existing Dutch word and *apper* is not; ambiguous L (AmbL) group). Another group of (Dutch) listeners was exposed to the exact same ambiguous [l/ɹ] sound but then in [ɹ]-final words, e.g., *wekker* (Eng: *alarm clock*; where *wekker* is a Dutch word, *wekkel* is not; ambiguous R (AmbR) group). After exposure to words containing the ambiguous sound, both groups of listeners are tested on the same continuum of ambiguous sounds from more [l]-like sounds to more [ɹ]-like sounds for which they have to indicate whether the heard sound is an [l] or an [ɹ]. Percentage [ɹ] responses for the continuum of ambiguous sounds for the two listener groups are measured. The results consistently show that listeners exposed to [l/ɹ] in [ɹ]-final words give significantly more [ɹ] responses than listeners exposed to the exact same [l/ɹ] sounds in [l]-final words. This difference between the group is referred to as the lexical retuning effect, and shows that listeners have retuned their phoneme category boundaries to include the deviant sound into their pre-existing phone category of [ɹ] or [l], respectively.

2.2. Experimental set-up

To mimic or create a Dutch listener, we first train a baseline DNN using read speech from the Spoken Dutch Corpus (CGN; [22]), which amounts to 551,624 words spoken by 324 unique speakers for a total duration of approximately 64 hours of speech. The training data was split into a training (80% of the full data set), validation (10%) and test set (10%) with no overlap in speakers.

Subsequently, this baseline model is retrained using the acoustic stimuli from the human perception experiment [16]. To mimic the two listener groups, we used two different configurations of the stimulus/training set (also referred to as lexical retuning set), resulting in two different retrained models. The stimuli consist of 200 Dutch words produced by a female Dutch speaker in isolation: 40 words with final [ɹ], 40 words with final [l], and 120 'distractor' words with no [l] and [ɹ]. For the 40 [l]-final words and the 40 [ɹ]-final words, versions also existed in which the final [l] or [ɹ] was replaced by the ambiguous [l/ɹ] sound. To mimic the two human listener groups, we trained:

- Amb(iguous)L model: trained on the 120 distractor words, the 40 [ɹ]-final words, and the 40 [l]-final words in which the [l] was replaced by the ambiguous [l/ɹ].
- Amb(iguous)R model: trained on the 120 distractor words, the 40 [l]-final words, and the 40 [ɹ]-final words in which the [ɹ] was replaced by the ambiguous [l/ɹ].

We expect to see differences specifically in the [l] and [ɹ] categories between the AmbL and AmbR models. In order to investigate the effect of retraining with the additional speech and particularly the ambiguous sounds, the AmbL and AmbR models will also be compared with a new baseline model, which is trained on all 200 natural words, without ambiguous sounds.

2.3. Model architecture

All experiments used a fully-connected, feedforward DNN with five hidden layers, 1024 nodes/layer. The output softmax layer had a dimension of 39 nodes, corresponding to the 39 phone categories of CGN. The network was trained on CGN using rectified linear (ReLU) nonlinearities on all hidden nodes, then re-trained for one epoch on CGN using logistic sigmoid nonlinearities. The sigmoid nonlinearity was retained during lexical retuning. CGN training used the Adam optimizer; the network was trained to a validation set frame error rate of 65.10%. Lexical retuning was performed using 50 epochs of stochastic gradient descent, with a learning rate of 0.00001, no momentum, and no decay.

2.4. I-vector representation for visualization

In order to model and visualize the information present in the hidden layers, we applied a discrete version of I-vector representation to the hidden node activations. The purpose of using the I-vector framework is to capture the acoustic variability and model the behavior of the neural responses of the DNNs. This modeling will allow us to visualize the different clusters the network learned. The goal is to make a link between the representations learned by the I-vector based DNN model and the perceptual learning experiment. The discrete I-vector is computed by first normalizing the hidden node activations (after the sigmoid nonlinearity), so that for any given frame, in any given hidden layer, all of the node activations sum to one [23]. The normalized activations are then summed over each phone segment determined by the forced alignment, resulting in a non-negative 1024-dimensional vector whose L1 norm is equal to the number of frames in the segment. The I-vector representation is a non-negative factor analysis (NFA) of the 1024-dimensional segment summary vectors [23][24]. The purpose of this modeling is to represent the DNN responses for a given phone (or speech segment) as a shift from the average responses of all the phones for a given hidden layer. This shift can be modeled as follow:

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{w}$$

where $\boldsymbol{M}$ is the 1024-dimensional segment summary vector, and $\boldsymbol{m}$ is the average across all segments in the corpus. The matrix $\boldsymbol{T}$ models the most important non-negative factors of variability in the DNN's reactions to the set of phone segments. The I-vector $\boldsymbol{w}$ describes the best segment dependent offset within the span of the subspace defined by matrix $\boldsymbol{T}$. The matrix $\boldsymbol{T}$ is trained using an EM-Like algorithm [24]. This framework has been used to visualize the language clusters that emerge in

the hidden-layer I-vectors of a language identification DNN [23]. In this paper, we used the same modeling technique to visualize the clusters of the studied phones.

3. Results

We investigated the DNN's perceptual learning ability using three measures. First, we investigate the DNNs' classification accuracy, and particularly we investigate whether the AmbR model gave more [ɹ] responses than the AmbL model, similar to the lexical retuning effect found in human listeners. Second, to check whether retraining with the ambiguous sounds resulted in a shift of the phoneme categories in the hidden layers, rather than a shift in the output layer (which would be comparable to the decision making step in humans), we investigate whether the DNNs show adaptation of the phone categories, similar to what has been found for humans [14]. We do so by calculating the average difference between the activations of the hidden nodes to the [l] and [ɹ] sounds and the ambiguous sounds, respectively, and by visualizing the hidden nodes' activations.

3.1. Phoneme classification

Table 1 shows the classification results. There is no separate test set; these are classification rates on the same dataset that was used to adapt the network, therefore it is not surprising that the network correctly labels almost all of the sounds on which it has been adapted. Since we are primarily interested in the [l], [ɹ], and ambiguous [l/ɹ] sound, we only report those. The baseline model showed high performance in the classification for the [l] and [ɹ] sounds. Interestingly, the baseline system classified the [l/ɹ] sound as [l] almost half of the time and also as [ɹ] a few times. The AmbL model had a high accuracy in the classification of the [ɹ]; however, it did not classify the [l] correctly. The lexical retuning dataset contains no labeled examples of a natural [l]; apparently, in this case, the model has learned the retuning data so well that it has forgotten what a natural [l] sounds like. On the other hand, the network has correctly learned to label the [l/ɹ] sounds as [l], indicating 'perceptual learning' by the AmbL system. The AmbR model had perfect classification of [l], and it has correctly learned to classify the [l/ɹ] sound as [ɹ], so we can say that the AmbR model also shows perceptual learning. Unlike AmbL, the AmbR model has not forgotten what a natural [ɹ] sounds like: it classifies natural [ɹ] tokens correctly in 29 out of 40 cases.

Table 1. Classification results on the [l], [ɹ], and [l/ɹ] training tokens for the three models. Correct classification in boldface.

Sound presented	**Sound(s) classified (%)**
Baseline, retrained model	
[l]	**l(97.6)**, m(3.4)
[ɹ]	**ɹ(95.0)**, sil(5.0)
[l/ɹ]	l(46.9), sil(23.5), ə(19.8), ɹ(8.6), ɛi (1.2)
AmbL model	
[l]	o(78.0), ɔ(10.4), sil(2.4), e(2.4), ɛi (2.4), Øː(2.4)
[ɹ]	**ɹ(87.5)**, e(7.5), ʌu(2.5), ɛi(2.5)
[l/ɹ]	l(81.5), ə(12.3), ə(2.5), ɛi(2.5), t(1.2)
AmbR model	
[l]	**l(97.6)**, sil(3.4)
[ɹ]	**ɹ(72.5)**, ə(15.0), sil(10.0), t(2.5)
[l/ɹ]	ɹ(88.9), sil(6.2), ə(4.9)

3.2. Average distances between phoneme categories

In order to explore the effect of lexical retuning on the hidden layers of the neural network, we began by calculating the inter-category distances at each hidden layer of the DNNs. The 1024-dimensional segment vectors were first re-normalized, so that each vector sums to one. The Euclidean distance was calculated between each ambiguous segment and each [l] segment, and was averaged over all pairs to compute an average [l]-to-[l/ɹ] distance; the same procedure was used to calculate an average [ɹ]-to-[l/ɹ] distance. The ratio of these two distances, then, was taken as a single measure of the degree to which lexical re-tuning has modified the feature representations at each of the five hidden layers and the soft-max layer (layer 5) of the DNN. Reasonable expectations include:

a. This ratio should be similar across all three models at layer 0. Layer 0, in our notation, is the first hidden layer of the network, not the acoustic input spectrum; but as the first hidden layer, it is closest to the spectrum, and likely to be least affected by lexical retuning.

Table 2. Ratio of distance(learned,l)/distance(learned,r) for the three models calculated using the posterior probabilities of the hidden nodes.

Model	**Layer number**					
	0	1	2	3	4	5
Baseline	1.112	1.135	1.085	1.085	1.031	0.931
AmbL	1.110	1.124	1.118	1.114	1.003	0.963
AmbR	1.149	1.167	1.214	1.274	1.284	2.280

b. For model 'L', this ratio should become smaller for higher layers (the distance from [l/ɹ] to [l] should be smaller than the distance from [l/ɹ] to [ɹ], because the model has been trained to recognize ambiguous tokens as [l]).
c. For model 'R', this ratio should become larger for higher layers (the distance from [l/ɹ] to [l] should be larger than the distance from [l/ɹ] to [ɹ], because the model has been trained to recognize the ambiguous tokens as [ɹ]).

Table 2 shows that indeed the average distances between the ambiguous sound tokens and the [l] and [ɹ] tokens, respectively, are fairly similar across all three models at layer 0. In the AmbL model, for higher layers, indeed the ratio decreases, indicating that the average distance between the [l/ɹ] tokens and the [l] tokens decreases, although the effect is somewhat small. In the AmbR model, the ratio increases substantially, indicating that the average distance between the [l/ɹ] tokens and the [ɹ] tokens decreases substantially.

3.3. Visualizations

Figures 1-3 show 3D Principal Component Analysis (PCA) visualizations of the I-vectors. These vectors are trained on the activations of the nodes of hidden layer 4 to the input sounds for the baseline model, the AmbL model, and AmbR model, respectively. The I-vector was trained on the activations of all the sounds, but PCA was only trained on the sounds of interest in this study. Green bullets denote the representation of the activations to [l/ɹ] tokens, red to the [l] tokens, and blue to the [ɹ] tokens. Figure 1 shows that the hidden nodes which are

activated when the input contains [l/ɹ] are positioned right in between the hidden nodes which are activated for input [l] and those for input [ɹ]. From an acoustic point of view this makes sense as the ambiguous sound [l/ɹ] is midway between a natural [l] and a natural [ɹ]. Retraining with the ambiguous sounds in a context that favors an [l] interpretation, i.e., mapping the ambiguous sound onto the [l] phoneme category, causes the network weights to be retuned in such a way that the I-vectors corresponding to ambiguous sounds are closer to those of [l] sounds, while the distance to the [ɹ] sounds increases (see Figure 2). Retraining such that the ambiguous sound is mapped onto the [ɹ] phoneme category results in the opposite pattern (see Figure 3): the I-vectors for [l/ɹ] segments are closer to the I-vectors for [ɹ] segments compared to the baseline model, while the distance to the [l] segments has increased. These results, and the results of Table 2, show that lexical retuning is implemented not just by changing the classification boundary between [l] and [ɹ], but also by changing the internal hidden-layer representation of each phonetic segment. When the network is trained to classify deviant sounds as [l], it does so by changing the hidden layers of the network so that the deviant sounds are shifted closer to [l]; when the network is trained to classify deviant sounds as [ɹ], it implements the opposite shift.

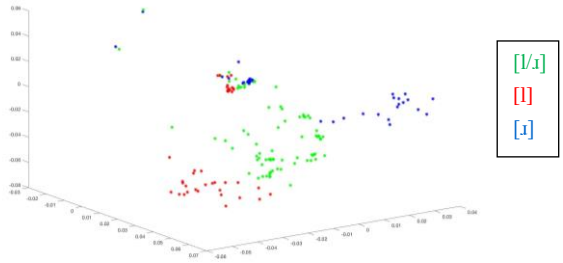

Figure 1. PCA visualization of the activations of the 4th hidden layer to input /l/, /r/, and the ambiguous sounds in the baseline model.

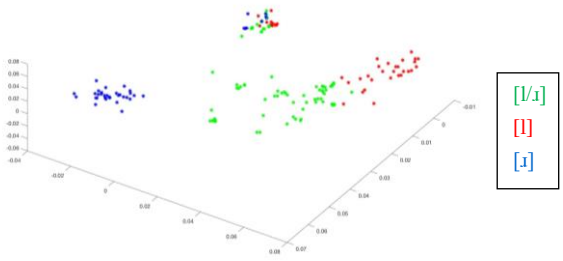

Figure 2. PCA visualization of the activations of the 4th hidden layer to input /l/, /r/, and the ambiguous sounds in ambL model.

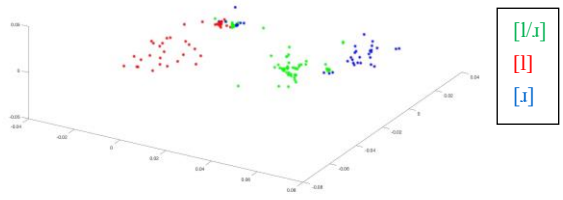

Figure 3. PCA visualization of the activations of the 4th hidden layer to input /l/, /r/, and the ambiguous sounds in the ambR model.

4. Discussion and concluding remarks

We investigated the adaptation of phoneme categories after exposure to ambiguous speech, a process also referred to as perceptual learning. We retrained a DNN model with additional material from a new speaker who had an (artificially created) ambiguous sound [l/ɹ], in between natural [l] and [ɹ] in a set-up similar to that of lexical retuning experiments with human listeners [16]. The DNNs retrained with the [l/ɹ] sounds indeed showed perceptual learning: while a model not exposed the [l/ɹ] sound classified [l/ɹ] during a subsequent test phase as [l] half of the time and also as [ɹ] a few times, the models retrained with the [l/ɹ] sound classified these [l/ɹ] sounds as either [l] or [ɹ] depending on the labels used during training. Calculations of the distances between the average activations to the natural sounds and the ambiguous sound and the visualizations of the activations of the hidden layers clearly indicated that the DNNs showed perceptual learning at the intermediate levels, not just at the output level. Future work will include more detailed analyses of the effect of the number of ambiguous training items on (the time-course of) retraining.

The results are in line with a plethora of lexical retuning experiments in human listeners, e.g., [12-18]. Upon hearing an ambiguous sound, humans have been suggested to change their internal representation of the sound category [14]. The results of our visualizations corroborate this suggestion, but furthermore suggest that perceptual learning does not simply result in the redrawing of phoneme category boundaries to include the ambiguous sound into an existing phoneme category [14], but rather that the phoneme category space is warped such that the representation of the ambiguous sound moves closer to the sound category as which it was classified. These results line up with other findings: Although human listeners have been shown to treat the ambiguous sound as if they are natural versions of the particular sound [16], they nevertheless remain able to distinguish between the ambiguous sound and examples of the natural phoneme category (but they are less good at it compared to before exposure) [14].

The experiment consisted of a carefully controlled, though restricted, set-up with only one ambiguous sound – a situation that might not often occur in everyday speech. This set-up allowed us, on the one hand, to focus specifically on one process, i.e., perceptual learning, without interference of other factors, and on the other hand, to explore the effects of that process on the hidden layers of DNNs. The success of the approach and the success of the DNNs to mimic human perceptual learning pave the way for further investigations of perceptual learning, in both human and automatic speech processing, to other types of speech, including naturally ambiguous speech, and other types of acoustic deviances from 'normal' speaking and listening conditions, such as the effect of (non-native) accents, dysarthric speech, or the presence of background noise. This work brings us one step closer to our ultimate goal of building human-speech processing inspired ASR systems that, similar to human listeners, can adjust flexibly and fast to all kinds of new input, and show that DNNs can be used as a way to investigate human speech processing.

5. Acknowledgements

O.S. was partly supported by a Vidi-grant from The Netherlands Organization for Scientific Research (NWO; grant number: 276-89-003). The authors would like to thank Raghavedra Pappagari for writing code to accumulate feature vector activations within phonetic segments.

6. References

[1] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman, 1982.

[2] O. Scharenborg, D. Norris, L. ten Bosch, J.M. McQueen, "How should a speech recognizer work?" *Cognitive Science*, vol., 29, no. 6, pp. 867-918, 2005.

[3] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research", *Speech Communication*, vol. 49, pp. 336-347, 2007.

[4] M.H. Davis & O. Scharenborg, "Speech perception by humans and machines", In: M.G. Gaskell & J. Mirkovic (Eds.) *Speech Perception and Spoken Word Recognition*, part of the series "Current Issues in the Psychology of Language", Routledge: London and New York, pp.181-203, 2017.

[5] S. Dusan & L.R. Rabiner, "On integrating insights from human speech recognition into automatic speech recognition. *Proceedings of Interspeech*, pp. 1233-1236, 2005.

[6] S. Davis & P. Mermelstein, "Comparison of the parametric representation for monosyllabic word recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[7] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, pp. 3-27, 1998.

[8] M. De Wachter, K. Demuynck, D. van Compernolle, & P. Wambaq, "Data driven example based continuous speech recognition", *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 1133-1136, 2003.

[9] S.D. Goldinger, "Echoes of echoes? An episodic theory of lexical access", *Psychological Review*, vol. 105, pp. 251-279, 1998.

[10] O. Scharenborg, "Modeling the use of durational information in human spoken-word recognition", *Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3758-3770, 2010.

[11] B. de Boer & P.K. Kuhl, "Investigating the role of infant-directed speech with a computer model", *ARLO*, vol. 4, pp. 129-134, 2003.

[12] A.G. Samuel & T. Kraljic, "Perceptual learning in speech perception", *Attention, Perception & Psychophysics*, vol.71, pp. 1207-1218, 2009.

[13] D. Norris, J.M. McQueen, & A. Cutler, "Perceptual learning in speech", *Cognitive Psychology*, vol. 47, no. 2, pp. 204-238, 2003.

[14] C. Clarke-Davidson, P.A. Luce, & J.R. Sawusch, "Does perceptual learning in speech reflect changes in phonetic category representation or decision bias?" *Perception & Psychophysics*, vol. 70, pp. 604-618, 2008.

[15] P. Drozdova, R. van Hout, & O. Scharenborg, "Lexically-guided perceptual learning in non-native listening," *Bilingualism: Language and Cognition*, vol. 19, no. 5, pp. 914-920, 2016. doi:10.1017/S136672891600002X.

[16] O. Scharenborg & E. Janse, "Comparing lexically-guided perceptual learning in younger and older listeners", *Attention, Perception, and Psychophysics*, vol. 75, no. 3, pp. 525-536, 2013. doi: 10.3758/s13414-013-0422-4.

[17] P. Drozdova, R. van Hout & O. Scharenborg, "Processing and adaptation to ambiguous sounds during the course of perceptual learning," *Proceedings of Interspeech*, pp. 2811-2815, 2016.

[18] K. Poellmann, J.M. McQueen, & H. Mitterer, "The time course of perceptual learning", *Proceedings of ICPhS*, 2011.

[19] M.J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech & Language*, vol. 12, no. 2, pp. 75-98, 1998.

[20] H. Liao, "Speaker adaptation of context dependent deep neural networks", *Proceedings of ICASSP*, pp. 7947-7951, 2013.

[21] D. Castelvecchi, "Can we open the black box of AI?", *Nature*, vol. 538, pp. 20-23, 2016.

[22] N.H.J. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.-P. Martens, M. Moortgat, & H. Baayen, "Experiences from the Spoken Dutch Corpus project", *Proc. LREC – Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, pp. 340-347, 2002.

[23] N. Dehak, "I-vector representation based on GMM and DNN for audio classification", *Keynote speech at Odyssey 2016 Speaker and Language Workshop*, 2016.

[24] M.H. Bahari, N. Dehak, H. Van hamme, L. Burget, A.M. Ali, & J. Glass, "Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1117-1129, 2014.