The Time-Course of Phoneme Category Adaptation in Deep Neural Networks

Junrui Ni[1], Mark Hasegawa-Johnson[1,2], and Odette Scharenborg[3(✉)]

[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, USA
[2] Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL, USA
[3] Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
o.e.scharenborg@tudelft.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. In previous work, we have shown that deep neural network-based (DNN) ASR systems can learn to adapt their phoneme category boundaries from a few labeled examples after exposure (i.e., training) to ambiguous sounds, as humans have been found to do. Here, we investigate the time-course of phoneme category adaptation in a DNN in more detail, with the ultimate aim to investigate the DNN's ability to serve as a model of human perceptual learning. We do so by providing the DNN with an increasing number of ambiguous retraining tokens (in 10 bins of 4 ambiguous items), and comparing classification accuracy on the ambiguous items in a held-out test set for the different bins. Results showed that DNNs, similar to human listeners, show a step-like function: The DNNs show perceptual learning already after the first bin (only 4 tokens of the ambiguous phone), with little further adaptation for subsequent bins. In follow-up research, we plan to test specific predictions made by the DNN about human speech processing.

Keywords: Phoneme category adaptation · Human perceptual learning · Deep neural networks · Time-course

1 Introduction

Whenever a new speaker or listening situation is encountered, both human listeners and machines need to adapt their sound categories to account for the speaker's pronunciations. This process is called perceptual learning, and is defined as the temporary adaptation of sound categories after exposure to deviant speech, in a manner such that the deviant sounds are included into pre-existing sound categories, thereby improving intelligibility of the speech (e.g., [1–6]). A specific case of perceptual learning is lexically-guided perceptual learning [2], in which the adaptation process is driven by lexical information. Human lexically-guided perceptual learning has been shown to be fast, and requires only a few instances of the deviant sound [5, 6]. Automatic speech recognition (ASR) systems typically adapt to new speakers or new listening conditions

C. Martín-Vide et al. (Eds.): SLSP 2019, LNAI 11816, pp. 3–15, 2019.
https://doi.org/10.1007/978-3-030-31372-2_1

using both short-time adaptation algorithms (e.g., fMLLR [7]) and longer-term adaptation techniques (e.g., DNN weight training [8]). In previous work [9], we showed that Deep Neural Networks (DNNs) can adapt to ambiguous speech as rapidly as a human listener by training on only a few examples of an ambiguous sound. Here, we push this research further and ask the following questions: Are ambiguous sounds processed in the same way as natural sounds; and, how many examples of the ambiguous sound are needed before the DNN adapts?

In short, the aim of this paper is two-fold: (1) we investigate the time-course of phoneme category adaptation in a DNN in more detail focusing on the amount of deviant speech material and training needed for phoneme category retuning to occur in a DNN, (2) with the larger aim to investigate the DNN's ability to serve as a model of human perceptual learning. In order to do so, we base our research on the experimental set-up and use the stimuli of a human perceptual learning experiment (see for other examples, e.g., [9–11]).

In a typical human lexically-guided perceptual learning experiment, listeners are first exposed to deviant phonemic segments in lexical contexts that constrain their interpretation, after which listeners have to decide on the phoneme categories of several ambiguous sounds on a continuum between two phoneme categories (e.g., [1–6]). This way the influence of exposure to the deviant sound can be investigated on the phoneme categories in the human brain. In this paradigm, two groups of listeners are tested. Using the experiment from which we take our stimuli [4] as an example: one group of Dutch listeners was exposed to an ambiguous [l/ɹ] sound in [l]-final words such as *appel* (Eng: *apple*; *appel* is an existing Dutch word, *apper* is not). Another group of Dutch listeners was exposed to the exact same ambiguous [l/ɹ] sound, but in [ɹ]-final words, e.g., *wekker* (Eng: *alarm clock*; *wekker* is a Dutch word, *wekkel* is not). After exposure to words containing the [l/ɹ], both groups of listeners were tested on multiple steps from the same continuum of [l/ɹ] ambiguous sounds from more [l]-like sounds to more [ɹ]-like sounds. For each of these steps, they had to indicate whether the heard sound was an [l] or an [ɹ]. Percentage [ɹ] responses for the continuum of ambiguous sounds were measured and compared for the two groups of listeners. Lexically-guided perceptual learning shows itself as significantly more [ɹ] responses for the listeners who were exposed to the ambiguous sound in [ɹ]-final words compared those who were exposed to the ambiguous sound in [l]-final words. A difference between the groups is interpreted to mean that listeners have retuned their phoneme category boundaries to include the deviant sound into their pre-existing phone category of [ɹ] or [l], respectively.

We base our research on the time-course of adaptation found in human listeners in the experiment in [5]. Their question was similar to ours: Are words containing an ambiguous sound processed in the same way as "natural" words, and if so, how many examples of the ambiguous sound are needed before the listener can do this? Participants had to listen to nonsense words, natural words, and words containing an ambiguous sound, and were instructed to press the 'yes' button as soon as possible upon hearing an existing word and 'no' upon hearing a nonsense word. Yes/no responses and reaction times to the natural and "ambiguous" words were analyzed in bins of 5 ambiguous words. They found that words containing an ambiguous sound were accepted as words less often, and were processed slower than natural words, but this difference in acceptance disappeared after approximately 15 ambiguous items.

2 Methods

In our DNN experiment, we follow the set-up used in [9]. To mimic or create a Dutch listener, we first train a baseline DNN using read speech from the Spoken Dutch Corpus (CGN; [12]). The read speech part of the CGN consists of 551,624 words spoken by 324 unique speakers for a total duration of approximately 64 h of speech. A forced alignment of the speech material was obtained using a standard Kaldi [13] recipe found online [15]. The speech signal was parameterized using a 64-dimensional vector of log Mel spectral coefficients with a context window of 11 frames, each has a segment length of 25 ms with a 10 ms shift between frames. Per-utterance mean-variance normalization was applied. The CGN training data were split into a training (80% of the full data set), validation (10%), and test set (10%) with no overlap in speakers.

Because we aim to investigate the DNN's ability to serve as a model of human perceptual learning, we used the same acoustic stimuli as used in the human perception experiment [4] for retraining the DNN (also referred to as retuning). The retraining material consisted 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. Forced alignments were obtained using a forced aligner for Dutch from the Radboud University. For four words no forced alignment was obtained, leaving 196 words for the experiment.

2.1 Model Architecture

All experiments used a simple fully-connected, feed-forward network with five hidden layers, 1024 nodes per layer, with logistic sigmoid nonlinearities as well as batch-normalization and dropout after each layer activation. The output layer was a softmax layer of size 38, corresponding to the number of phonemes in our training labels. The model was trained on CGN for 10 epochs using an Adam optimizer with a learning rate of 0.001. After 10 epochs, we reached a training accuracy of 85% and a validation accuracy of 77% on CGN.

2.2 Retuning Conditions

To mimic the two listener groups from the human perceptual experiment, and to mimic a third group with no exposure to the ambiguous sound (i.e., a baseline group), we used three different configurations of the retuning set:

- Amb(iguous)L model: trained on the 118 distractor words, the 39 [ɹ]-final words, and the 39 [l]-final words in which the [l] was replaced by the ambiguous [l/ɹ].
- Amb(iguous)R model: trained on the 118 distractor words, the 39 [l]-final words, and the 39 [ɹ]-final words in which the [ɹ] was replaced by the ambiguous [l/ɹ].
- Baseline model: trained on all 196 natural words (no ambiguous sounds). This allows us to separate the effects of retuning with versus without the ambiguous sounds.

In order to investigate the time-course of phoneme category adaptation in the DNNs, we used the following procedure. First, the 196 words in the three retuning sets were split into 10 bins of 20 distinct words, except for the last two bins, which each contained only 18 words. In order to be able to compare between the different retuning conditions, the word-to-bin assignments were tied among the three retuning conditions. Each word appeared in only one bin. Each bin contained: 4 words with final [r] (last bin: 3 words) + 4 words with final [l] (penultimate bin: 3 words) + 12 'distractor' words with no [l] or [r] (last two bins: 11 words). The difference between the retuning conditions is:

- AmbL: the final [l] in the 4 [l]-final words was replaced by the ambiguous [l/ɹ] sound.
- AmbR: the final [ɹ] in the 4 [ɹ]-final words was replaced by the ambiguous [l/ɹ] sound.
- Baseline: only natural words.

The [l]-final, [ɹ]-final, and [l/ɹ]-final sounds of the words in bin t from all three retuning sets, combined, functioned as the test set to bin $t - 1$. As all the acoustic signals from the test bin were unseen during training at the current time step, we denote this as "open set evaluation". Figure 1 explains the incremental adaption. Note that the final bin was only used for testing; because at $t = 10$, there is no subsequent bin that could be used for testing.

Retuning was repeated five times, with five different random seeds for permutation of data within each bin, for each retuning condition/model. Each time, for every time step of incremental adaptation, we retrained the baseline CGN-only model using bin 0 up to bin $t - 1$ of the retraining data for 30 epochs using an Adam Optimizer with a learning rate of 0.0005. The re-tuning accuracy on the training set after 30 epochs always reached an accuracy of 97.5–99%.

```
for each retuning set from {Baseline, AmbL, AmbR}
    Test the CGN-only model using bin 0 from the test set

    for t in [1,9]:
        Retrain the CGN-only model using bin 0 up to bin t-1
        Test the retrained model from bins 0 through t-1 using test set bin t
```

Fig. 1. Incremental retuning procedure for the open set evaluation.

3 Classification Rates

In the first experiment, we investigated the amount of training material needed for perceptual learning in a DNN to occur. Classification accuracy was computed for all frames, but since we are primarily interested in the [l], [ɹ], and the ambiguous [l/ɹ] sound, we only report those. Figure 2 through 4 show the proportion of correct frame classifications as solid lines, i.e., [l] frames correctly classified as [l] and [ɹ] frames correctly classified as [ɹ], for each of the 10 bins ($0 \leq t \leq 9$). Dashed lines show, for example, the proportion of [l] frames incorrectly classified as [ɹ], and of [ɹ] frames

incorrectly classified as [l]; the rate of substitutions by any other phone is equal to 1.0 minus the solid line minus the dashed line. The interesting case is the classification of the [l/ɹ] sound (see triangles), which is shown with a dashed line when classified as [l] and with a solid line when classified as [ɹ]. Note, in the legend, the capital letter denotes the correct response, lowercase denotes the classifier output, thus, e.g., L_r is the percentage of [l] tokens classified as [ɹ].

Figure 2 shows the results for the baseline model retrained with the natural stimuli. The baseline model shows high accuracy in the classification of [ɹ]. The [l] sound is classified with high accuracy at $t = 2$, then drops for increasing t, up to $t = 8$. The [ɹ] sound, on the other hand, is classified with very high accuracy after seeing a single bin of retuning data, with very little further improvement for subsequent bins. The [l/ɹ] sound (not part of the training data for this model) is classified as [ɹ] about 70% of the time, and as [l] about 10% of the time, with the remaining 20% of instances classified to some other phoneme.

Figure 3 shows the results for the model retrained with the ambiguous sound bearing the label of /l/. The AmbL model has a high accuracy in the classification of the [ɹ]; however, the accuracy of natural [l] is less than 50% after the first bin and continues to worsen as more training material is added. The lexical retuning dataset contains no labeled examples of a natural [l]; apparently, in this case, the model learns the retuning data so well that it forgets what a natural [l] sounds like.

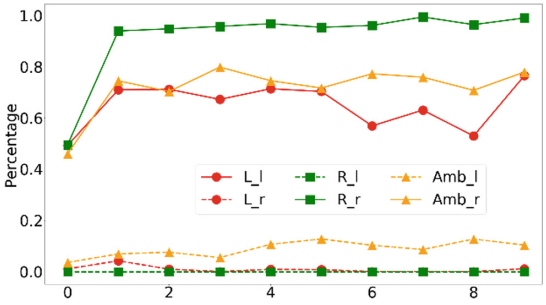

Fig. 2. Proportion of [l] and [ɹ] responses by the baseline model, retrained with natural stimuli, per bin.

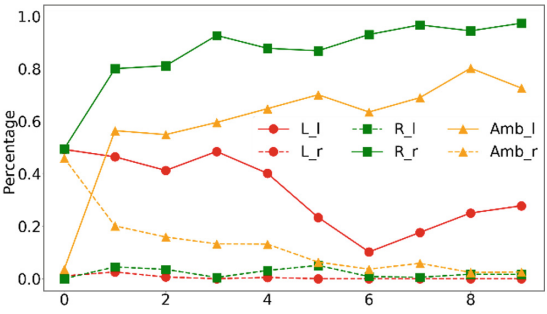

Fig. 3. Proportion of [l] and [ɹ] responses by the AmbL model, retrained with [l/ɹ] labeled as [l], per bin.

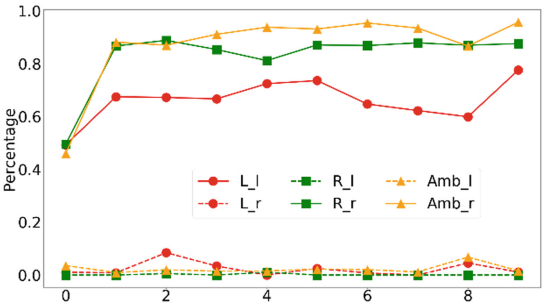

Fig. 4. Proportion of [l] and [ɹ] responses by the AmbR model, retrained with [l/ɹ] labeled as [ɹ], per bin.

Importantly, after the first bin, the network has correctly learned to label the [l/ɹ] sounds as [l], indicating 'perceptual learning' by the AmbL system. The classification of [l/ɹ] as [l] continues to rise slightly for subsequent bins. While the AmbL model already correctly recognizes most [l/ɹ] sounds as [l] after the first bin, recognition further improves for subsequent bins.

Figure 4 shows the results for the model retrained with the ambiguous sound labeled as /r/. The AmbR model has high accuracy for both the [l] and [ɹ] sounds. So, unlike the AmbL model, the AmbR model did not forgot what a natural [ɹ] sounds like. Moreover, after the first bin, this model has learned to classify [l/ɹ] as [ɹ] more than 85% of the time, which is a 10% increase over the model trained on the natural sounds at the same time step, thus showing perceptual learning. Unlike the AmbL model, additional [l/ɹ] training examples show little tendency to further increase the classification of [l/ɹ] as [ɹ], up to and including the last point.

Interestingly, all phones, including the natural [l] and [ɹ] as well as the ambiguous phone, show classification accuracy of around 50% prior to retraining. This rather low accuracy is most likely due to the differences in recording conditions and speaker between the CGN training set and the retraining sets. After retraining with the first bin, the classification accuracies make a jump in all models, with little further adaptation for subsequent bins, although the AmbL shows a small increase in adaptation for later bins, while this is not the case for the baseline and AmbR models. This adaptation suggests that the neural network treats the ambiguous [l/ɹ] exactly as it treats every other difference between the CGN and the adaptation data: In other words, exactly as it treats any other type of inter-speaker variability. In all three cases, the model learns to correctly classify test tokens after exposure to only one adaptation bin (only 4 examples, each, of the test speaker's productions of [l], [r], and/or the [l/ɹ] sound).

All three models show little tendency to misclassify [l] as [ɹ], or vice versa. This indicates that the retraining preserves the distinction between the [l] and [ɹ] phoneme categories.[1]

To investigate where the retuning takes place, we examined the effect of increasing amounts of adaptation material on the hidden layers of the models using the inter-category distance ratio proposed in [9]. This measure quantifies the degree to which lexical retuning has modified the feature representations at the hidden layers using a single number. First, the 1024-dimensional vector of hidden layer activations is re-normalized, so that each vector sums to one, and averaged across the frames of each segment. Second, the Euclidean distances between each [l/ɹ] sound and each [l] segment are computed, after which the distances are averaged over all [l/ɹ]-[l] token pairs, resulting in the average [l]-to-[l/ɹ] distance. Third, using the same procedure the average [ɹ]-to-[l/ɹ] distance is computed. The inter-category measure is then the ratio of these two distances, and is computed for each of the ten bins.

Figures 5 through 7 show the inter-category distance ratio ([l/ɹ]-to-[l] over [l/ɹ]-to-[ɹ]) for the baseline model, the AmbL model, and the AmbR model, respectively, for each of the 5 hidden layers, for each of the bins.

Figure 5 shows that for earlier bins in the baseline model, the distance between the ambiguous sounds and the natural [l] category and natural [ɹ] category is approximately the same for the different layers, with a slight bias towards [ɹ] (the ratio is >1); the lines for the five layers are close together and do not have a consistent ordering. From bin 5 onwards, and particularly for the last 3 bins, the distance between [l/ɹ] and the natural [l] category decreases from the first (triangles) to the last layer (diamonds), suggesting that [l/ɹ] is represented closer to the [l] category. However, this cannot be observed in the classification scores: Fig. 2 shows that [l/ɹ] is primarily classified as [ɹ]. The adaptation of [l/ɹ] towards natural [l] for the later bins suggests that adding training material of the speaker improves the representation of the natural classes as well, because the distance between [l/ɹ] and the natural classes changes without the model being trained on the ambiguous sounds.

Figure 6 shows that, for the AmbL model, the distance between [l/ɹ] and the natural [l] category becomes increasingly smaller deeper into the network: The line showing hidden layer 1 (triangles) is almost always on top, and the line showing layer 5 (diamonds) is almost always at the bottom. Interestingly, there is a downward slope from the first to the last bin, indicating that with increasing numbers of [l/ɹ] training examples labeled as [l], the distance between [l/ɹ] and natural [l] continues to decrease, even though there are no natural [l] tokens in the retuning data. This continual decrease

[1] We repeated this experiment using a Recurrent Neural Network (RNN) model trained under the Connectionist Temporal Classification (CTC) [14] criterion. The network architecture was different from the DNN architecture used in this paper, and consisted of two convolutional layers on the raw spectrogram, followed by six layers of stacked RNN. Despite the vastly different architecture, our new model showed highly similar behavior in terms of classification rate over the time course of incremental retuning. Most interestingly, both models seemed to have forgotten what a natural [l] sounds like.

in distance between [l/ɹ] and natural [l] seems to be correlated with the continual increase in classification of the ambiguous sound as [l] for the later bins in Fig. 3, and might indicate further adaptation of the representation of the ambiguous sound towards the natural [l].

In the AmbR model (Fig. 7), the ratio of distance([l/ɹ],[l]) over distance([l/ɹ],[ɹ]) increases from layer 1 to layer 5, indicating that the neural embedding of [l/ɹ] becomes more [ɹ]-like deeper in the network. So, like the AmbL model, the AmbR model also shows lexical retuning: The speech representation of [l/ɹ] becomes increasingly closer to that of the natural [ɹ] deeper into the model. The effect of increasing amounts of adaptation material is however not as clear-cut as for the AmbL model. The distance ratio rises until bin 2 (8 [l/ɹ] training examples), then falls until bin 5, then rises until bin 7, then falls again. This inconsistency is also found in the classification scores of [l/ɹ] as [ɹ] in Fig. 4 but to a lesser extent, which suggest that the increase in the distance between the [l/ɹ] and [ɹ] categories is not large enough to substantially impact classification results.

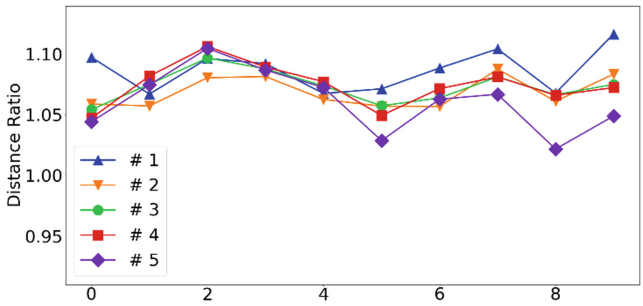

Fig. 5. Ratio of distance([l/ɹ],[l])/distance([l/ɹ],[ɹ]) for the Baseline model.

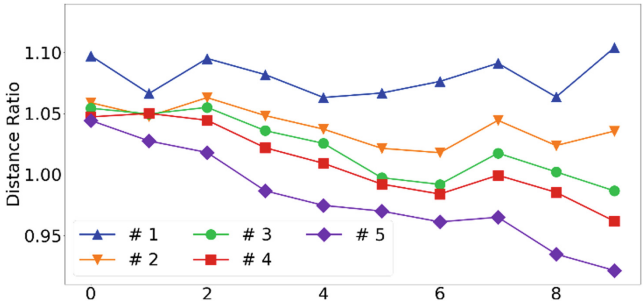

Fig. 6. Ratio of distance([l/ɹ],[l])/distance([l/ɹ],[ɹ]) for the AmbL model.

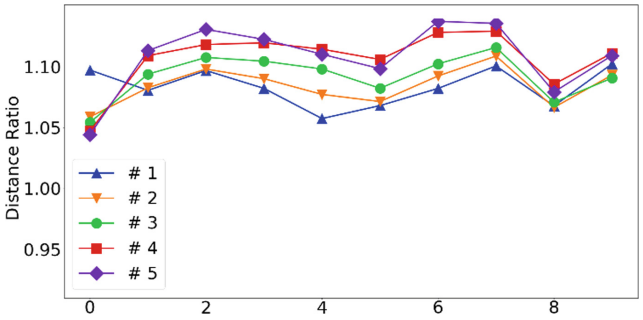

Fig. 7. Ratio of distance([l/ɹ],[l])/distance([l/ɹ],[ɹ]) for the AmbR model.

As the classification rates make a significant jump after just seeing the first bin of words for all three experimental sets, which indicates very fast adaptation, in the second experiment, we investigate how the CGN-only model adapts to a single bin of retuning data over the training course in the very first time step. Similar to the procedure above, we evaluate the classification rates by training the CGN-only model using the first training bin (training bin 0) from each experiment set (natural, AmbL, AmbR) for 30 epochs. Before the first epoch of training ($t = 0$), and after each epoch of training ($1 \leq t \leq 30$), we record the percentage of [l], [ɹ], and ambiguous [l/ɹ] sounds from the second test bin (test bin 1) that are classified as either [l] or [ɹ] (a total of 31 time points, $0 \leq t \leq 30$). Figure 8 shows the classification rates for the Baseline model: both [l] and [ɹ] sounds show immediate adaptation after the first epoch (correct response rate increases by about 20% from $t = 0$ to $t = 1$). The [ɹ] sound shows the highest accuracy over 30 epochs, but the number of [ɹ]'s correctly recognized only increases very slightly after the fifth epoch. After reaching a peak by the first epoch, the classification rate for [l] decreases until the third epoch, and then flatlines (with some small oscillations). Interestingly, while ambiguous [l/ɹ] sounds are not present in the training data, more and more [l/ɹ] get classified as [ɹ] as training progresses, meaning

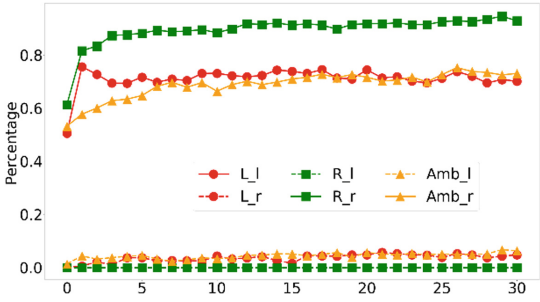

Fig. 8. Proportion of [l] and [ɹ] responses by the baseline model over 30 epochs for the first bin.

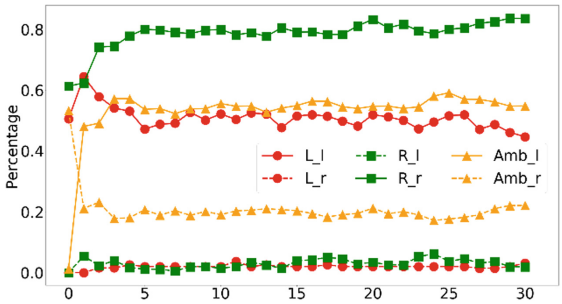

Fig. 9. Proportion of [l] and [ɹ] responses by the AmbL model over 30 epochs for the first bin.

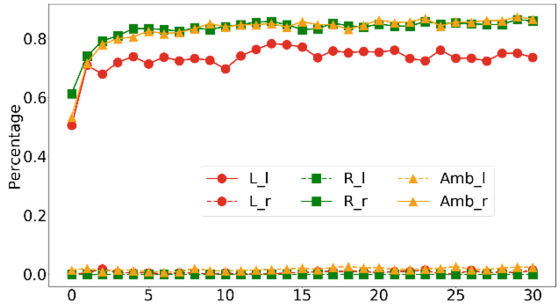

Fig. 10. Proportion of [l] and [ɹ] responses by the AmbR model over 30 epochs for the first bin

that the bias of [l/ɹ] toward [ɹ] somehow increases without the model seeing any ambiguous sounds.

Figure 9 shows the classification rates over 30 epochs for the AmbL model using stimuli from the first training bin with ambiguous sounds labeled as [l]. The classification rates at $t = 0$ are the same in Figs. 8 and 9, because they are based on the same model; it is only after the first training epoch ($t = 1$) that their rates diverge. Similar to Fig. 8, the accuracy for [ɹ] reaches 80% within 5 epochs, with a large jump at the second epoch. The accuracy for natural [l] also jumps up after the first epoch, even though there are no [l] tokens in the training data, but beginning with the second epoch, the model starts to forget how to correctly classify natural [l] tokens. The most important observation comes with the ambiguous [l/ɹ] sound. After just a single epoch on a single bin of data, the percentage of [l/ɹ] sounds classified as [l] goes from 0% to a little below 50%. However, after 5 epochs, the accuracy for [l/ɹ] as [l] flatlines around 50%, meaning that the model has reached its limit of perceptual learning by seeing only one training bin.

Figure 10 shows the classification rates over 30 epochs for the AmbR model using stimuli from the first training bin with ambiguous sounds labeled as [ɹ]. While no natural [ɹ] is present in this experiment set, the accuracy for natural [ɹ] gradually increases until the fifth epoch, meaning that perceptual learning on ambiguous sounds as [ɹ] also helps the model learn a natural [ɹ]. The ambiguous [l/ɹ] sound is classified as

[ɹ] 50% of the time at $t = 0$, i.e., with no training; the $t = 0$ case is identical to those shown in Figs. 8 and 9. After just one epoch of training, using one bin of ambiguous sounds labeled as [ɹ], the model learns to perform this classification with 70% accuracy, and accuracy increases until the fifth epoch.

It is worthwhile, at this point, to remind the reader what is meant by "one epoch" in the training of a neural net. Each epoch of training consists of three stages: (1) a direction vector d is chosen; in the first epoch, this is just the negative gradient of the error; (2) a search procedure is used to choose the scale, g; (3) the neural network weights are updated as $w = w + gd$. Each epoch of training can only perform a constant shift of the previous network weights. Figures 2, 3, 4, 8, 9 and 10 show that most of the DNN adaptation occurs in the first epoch on the first bin of the adaptation material, i.e., on the first update of the direction, therefore most of the DNN adaptation can be characterized as a constant shift in the network weights. This makes sense since the model is just learning about 4 additional training tokens (one adaptation bin) — with only 4 tokens, while it is not possible to learn a very complicated modification of the boundary, learning a boundary shift is indeed possible and very likely the case here.

In a deep neural network, a constant shift of the network weights is not the same thing as a constant shift of the classification boundary, but in practice, the revision of w after the first epoch is usually not much more complicated than a shifted boundary. The finding that inter-talker adaptation can be accomplished by a constant shift in cepstral space is not new; it has previously been reported by [16]. The finding that a comparable constant shift is sufficient to learn distorted sounds, like the ambiguous [l/ɹ] sound, has never previously been reported.

4 Discussion and Concluding Remarks

Inspired by the fast adaptation of human listeners to ambiguous sounds (e.g., [1–6]), we investigated the time-course of phoneme category adaptation in a DNN, with the ultimate aim to investigate the DNN's ability to serve as a model of human perceptual learning. We based our investigation on the time-course of adaptation of the human perceptual learning experiment in [5]. In the first experiment, we provided the DNN with an increasing number of the original ambiguous acoustic stimuli from [5] as retraining tokens (in 9 bins of 4 ambiguous items), compared classification accuracy on the ambiguous items in an independent, held-out test set for the different bins, and calculated the ratio of the distance between the [l/ɹ] category and the natural [l] and [ɹ] categories, respectively, for the five hidden layers of the DNNs and for the 9 different bins. In the second experiment, the amount of training was investigated by calculating the classification rates over 30 epochs when only one bin is used for retuning.

Results (both presented here and the unpublished results with a CTC-RNN model) showed that, similar to human listeners, DNNs quickly learned to interpret the ambiguous sound as a "natural" version of the sound. After only 4 examples of the ambiguous sound, the DNN showed perceptual learning, with little further adaptation for subsequent training examples, although a slight further adaptation was observed for the model which learned to interpret the ambiguous sound as [l]. In fact, perceptual learning could already clearly be seen after only one epoch of training on those 4

examples, and showed very little improvement after the fifth epoch. This is in line with human lexically-guided perceptual learning; human listeners have been found to need 10–15 examples of the ambiguous sound to show the same type of step-like function [5, 6]. We should note, however, that it is not evident how to compare the 4 examples needed by the DNN with the 10–15 examples of the human listener. We know of no way to define the "learning rate" of a human listener other than by adjusting the parameters of a DNN until it matches the behavior of the human, which is an interesting avenue for further research into the DNN's ability to serve as a model of human perceptual learning. Nevertheless, both DNNs and human listeners need very little exposure to the ambiguous sound to learn to normalize it.

Retuning took place at all levels of the DNN. In other words, retuning is not simply a change in decision at the output layer but rather seems to be a redrawing of the phoneme category boundaries to include the ambiguous sound. This is again exactly in line with what has been found for human listeners [17].

This paper is the first to show that, similar to inter-talker adaptation, adaptation to distorted sounds can be accomplished by a constant shift in cepstral space. Moreover, our study suggests that DNNs are more like humans than previously believed: In all cases, the DNN adapted to the deviant sound very fast and after only 4 presentations, with little or no adaptation thereafter. Future research will aim to test, in perceptual experiments with human listeners, the prediction of the DNN that the speech representations of the ambiguous sound and the natural [l] and [ɹ] change very little once the category adaptation has taken place.

Acknowledgements. The authors thank Anne Merel Sternheim and Sebastian Tiesmeyer with help in earlier stages of this research, and Louis ten Bosch for providing the forced alignments of the retraining material. This work was carried out by the first author under the supervision of the second and third author.

References

1. Samuel, A.G., Kraljic, T.: Perceptual learning in speech perception. Atten. Percept. Psychophys. **71**, 1207–1218 (2009)
2. Norris, D., McQueen, J.M., Cutler, A.: Perceptual learning in speech. Cogn. Psychol. **47**, 204–238 (2003)
3. Scharenborg, O., Weber, A., Janse, E.: The role of attentional abilities in lexically-guided perceptual learning by older listeners. Atten. Percept. Psychophys. **77**(2), 493–507 (2015). https://doi.org/10.3758/s13414-014-0792-2
4. Scharenborg, O.: Janse, E: Comparing lexically-guided perceptual learning in younger and older listeners. Atten. Percept. Psychophys. **75**(3), 525–536 (2013). https://doi.org/10.3758/s13414-013-0422-4
5. Drozdova, P., van Hout, R., Scharenborg, O.: Processing and adaptation to ambiguous sounds during the course of perceptual learning. In: Interspeech 2016, San Francisco, CA, pp. 2811–2815 (2016)
6. Poellmann, K., McQueen, J.M., Mitterer, H.: The time course of perceptual learning. In: Proceedings of ICPhS (2011)

7. Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
8. Liao, H.: Speaker adaptation of context dependent deep neural networks. In: Proceedings of ICASSP, pp. 7947–7951 (2013)
9. Scharenborg, O., Tiesmeyer, S., Hasegawa-Johnson, M., Dehak, N.: Visualizing phoneme category adaptation in deep neural networks. In: Proceedings of Interspeech, Hyderabad, India (2018)
10. Scharenborg, O.: Modeling the use of durational information in human spoken-word recognition. J. Acoust. Soc. Am. **127**(6), 3758–3770 (2010)
11. Karaminis, T., Scharenborg, O.: The effects of background noise on native and non-native spoken-word recognition: a computational modelling approach. In: Proceedings of the Cognitive Science conference, Madison, WI, USA (2018)
12. Oostdijk, N.H.J., et al.: Experiences from the spoken Dutch Corpus project. In: Proceedingd of LREC – Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, pp. 340–347 (2002)
13. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding, Hawaii, US (2011)
14. Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA (2006)
15. https://github.com/laurensw75/kaldi_egs_CGN
16. Pitz, M., Ney, H.: Vocal Tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech Audio Process. **13**(5), 930–944 (2005)
17. Clarke-Davidson, C., Luce, P.A., Sawusch, J.R.: Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? Percept. Psychophys. **70**, 604–618 (2008)