# THE ROLE OF PRODUCTION IN PERCEPTUAL LEARNING

Lena-Marie Huttner[1], Noël Nguyen[1], Martin J. Pickering[2]

[1]Aix Marseille University, CNRS, LPL, Aix-en-Provence, France, [2] The University of Edinburgh
lena-marie.huttner@univ-amu.fr, noel.nguyen-trong@univ-amu.fr, martin.pickering@ed.ec.uk

## ABSTRACT

Speech perception and production are often said to be intrinsically linked, though the nature of the link is contested. In this pilot study, we investigated whether the production of a VOT contrast would influence changes in perception. We designed an online experiment with 2x2 conditions, in which participants were asked to categorize nine stimuli on a voice onset time continuum between /tin/ and /din/ before and after playing a categorization game with a bot. During the game participants were trained on a sound to word correspondence with a bias either towards /din/ or /tin/. In one condition participants alternated categorizing and producing the stimulus, while in another participants only categorized the stimuli. Significant differences in categorization between conditions occurred only during the interactive phase of the experiment. This could mean that while perception and production may be related, production does not appear to mitigate a long-term adjustment in perception.

**Keywords:** Speech perception, perceptual learning, speech production.

## 1. INTRODUCTION

In any interaction people have to adjust their perception of the speech signal to their interlocutor's variable production. In laboratory settings this plasticity in speech sound perception can be observed in the phenomenon of perceptual learning [1, 2]: After exposure to an ambiguous stimulus, people will tend to adjust their perceptual boundaries to the input. Theories of speech perception offer several explanations for this observation. However, the nature of the underlying mechanisms that allow for this plasticity are still debated. Yet, it serves a fundamental communicative purpose without which communication would be impossible. In a conversation, interlocutors will come to share their representations on every linguistic level [3]. Perceptual learning may be a manifestation of this alignment on the level of speech sounds: A listener will hear their interlocutor's production, supposedly a manifest of their own representation of a sound, and adjust their perception to it. Similarly, social interaction does not only influence a speaker's perception of speech sounds but also their production. Experiments on phonetic convergence have found that interlocutors will become more similar in their acoustic phonetic realization over time [4, 5]. It is assumed that this phenomenon is caused or accompanied by a change in perception: Speakers are assumed to adjust their production to the perceived input [6]. [3] further stipulate that in order for this process of alignment to function, perception and production must be intrinsically linked. This idea is also iterated in sensorimotor theories of speech perception and production [7]. Yet, it is important to note that though coordinated, speech perception and production are still regarded to be separate processes. A change in one domain may not suffice for adjustment in the other [8, 9]. There is also conflicting experimental evidence on the co-occurrence of perceptual adaptation and convergence in speech production: On the one hand [8] found that native English speakers' perception of a contrast adapted to the speaker when it was presented as an idiolect rather than a dialect, however, production of the contrast was unaffected by this training. On the other hand[10] found that training native Japanese learners of English in the perception of a novel phonemic contrast lead to the production of that contrast. Contrary to [8] and [9], [10] argue that a close link between speech perception and production is a requirement for category formation. These studies ([8, 10] are just two examples of experiments examining the mechnanisms that may govern perceptual adaptation. A common denominator of these studies is the inclusion of social factors that inevitably influence the participant's behavior. Language is a social practice and linguistic behaviors cannot always be isolated from the social environment in which they occur. Yet, when trying to understand the perception production link, it can be useful to break down social interaction to a minimum and exclude any extralinguistic information. It

appears to be debated how and if perception and production influence one another. We here ask what role does the production of a contrast throughout a conversation play in the adjustment of perception? An influence of speech production on speech perception has been observed in [11]. Their ([11]) findings are in line with the assumptions made by [7], which stipulate that a speech sound representation is composed of both perceptual and motor information. In this pilot study we aim to shed a light on the link between perception and production by examining role of speech production on perceptual learning.

## 2. METHODS

To investigate whether an adaptation in perception was influenced by the production of a contrast, we designed an online experiment in Labvanced [12] with two by two conditions. The experiment was run on participants' smartphones. [13] compared the quality of speech data recorded on smartphones and a head mounted condenser microphone and found no significant influence of recording device on fitness the data for phonetic research. In this study we combined the experimental paradigms of perceptual learning and phonetic convergence studies; using a pre-post paradigm to measure changes in categorization as well as changes in speech production: We measured participants categorization of the stimuli as well as their production before and after playing a categorization game with a bot. We here present and discuss the results of the perceptual side of the experiment.

#### *2.0.1. Stimuli*

We measure phonetic adaptation and convergence on voice onset time (VOT). Perceptual adaptation and phonetic convergence on VOT has been shown by [14] and [15] respectively. We created a 9-step VOT continuum ranging form 23 to 55 ms of the words din and tin spoken by a female native speaker of English (dialect region: Northern England). The continuum was created in Praat [16] using a script [17]. Prior to the experiment, we conducted a stimulus test in which 30 participants (native English speakers from the UK) were asked to categorize the stimulus continuum.

### 2.1. Experimental setup

We recruited 88 native English speakers from the UK between the ages of 18 and 40 on Prolific. Participants were first presented with a consent form followed by a basic demographic questionnaire. Prior to the experimental tasks, we conducted a headphone screening test[18]. Participants were also repeatedly told they should be in a quiet room to participate in the experiment. Participants whose audio data included noticeable background noise that indicated otherwise were excluded from analysis and replaced. Participants were instructed to turn off notifications for the duration of the experiment and to place their phones on a surface in front of them. Participants were told that they would be interacting with a partner whose voice they would be hearing throughout the experiment. To make this claim more believable participants exchanged a small greeting with the experiment script before beginning the first task. The experiment began with the pre-test which consisted of a production and categorization task. In the production task participants were first asked to produce 10 instances of each tin and din; the words were shown in randomized order along with an image of a red microphone to indicate that recording was taking place. In the categorization task, participants first heard one of the nine stimuli before the words tin and din appeared on their screens. Participants were asked to tap on the word they had perceived. The perception task ran for 45 trials in which 5 iterations of each acoustic stimulus were presented. Participants were then told they would be playing an interactive game with another participant. During this interactive phase of the experiment, participants again heard one of the acoustic stimuli and were asked to tap on the word they just heard. They were then shown the word their supposed partner had read on their screen, i.e. they were given feedback on the intended categorization of the stimulus they heard. Here the experiment script was biased towards one of the endpoints. In two of the conditions, VOT-steps 1 through 6 were identified as /din/ (d-bias conditions) whereas in the other two, steps 4 through 9 were identified as /tin/ (t-bias condition). Participants in the interactive conditions were then asked to produce a word (either tin or din) which they read on their screen. In the control condition, participants did not speak and went on to the next categorization task. The interactive game consisted of 90 trials in which each acoustic stimulus was presented 10 times. Participants then proceeded to the post-test in which the two pre-test tasks were repeated in randomized order.

### 2.2. Hypotheses

In light of the findings by [10, 1, 2] we expect perceptual shifts to occur in all four conditions in

the direction of the bias of the script. For the two conditions with a bias towards /tin/ we expect participants to categorize more stimuli as /tin/ and in the conditions with a bias towards /din/ we expect participants to categorize more stimuli as /din/ in the post-test in comparison to the pre-test. We further expect there to be a difference in categorization between interactive and control condition.

## 3. RESULTS

We compare the categorization curves between and within conditions. The curves are depicted below 1.

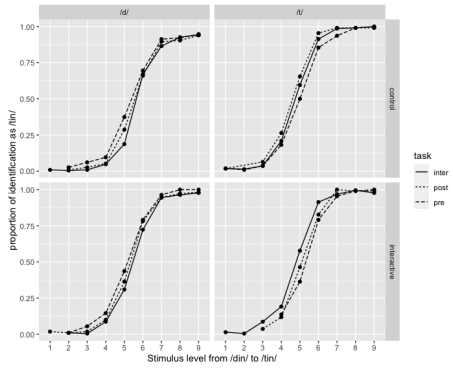


**Figure 1:** Proportion of identification of the stimuli as /tin/ by condition.

To compare between as well as within group differences, we ran a generalized mixed effects model using the lme4 package in R [19]. With task (pre-test, post-test, interaction) and condition as fixed effects and subject as a random effect (formula: response ∼ condition * task + (1+ stimulus|subject), family = binomial). In this model, the effect of task (interaction) is statistically significant and negative (95% CI [-1.08, -0.44], p < .001) as well as the effect of task (post) (95% CI [-0.90, -0.17], p = 0.004). We conducted a post-hoc Tukey test to examine the differences within and between conditions. Table 1 shows the significant differences between conditions.

| task | contrast | estimate | SE | Z-score | p-value |
|---|---|---|---|---|---|
| interaction | control t - control d | 1.6288 | 0.338 | 4.825 | 0.0001 |
| interaction | interactive t - control d | 1.6471 | 0.334 | 4.928 | 0.0001 |
| interaction | control t - interactive d | 1.3729 | 0.316 | 4.341 | 0.0009 |
| interaction | interactive t - interactive d | 1.3912 | 0.316 | 4.405 | 0.0006 |
| post | control t - control d | 1.7238 | 0.359 | 4.797 | 0.0001 |
| post | control t - interactive d | 1.3676 | 0.339 | 4.033 | 0.0032 |

**Table 1:** Between group differences

There are the following statistically significant within group differences in categorization:

| condition | contrast | estimate | SE | Z-score | p-value |
|---|---|---|---|---|---|
| control d | pre - interaction | 0.7458 | 0.160 | 4.653 | 0.0002 |
| control t | pre - post | -0.8035 | 0.188 | -4.278 | 0.0011 |
| interactive t | pre - interaction | -0.9798 | 0.165 | -5.939 | <.0001 |

**Table 2:** Within group differences

We expect any differences in categorization to be more pronounced for the three midpoint stimuli than the endpoint stimuli. We therefore ran the same model as above, but only on the three midpoint stimuli (points 4, 5, and 6 in 1). In this model, with the intercept corresponding to control d-bias and pre-test, The effect of task (interaction) is statistically significant and negative (95% CI [-1.12, -0.38], p < .001) as well as the effect of task (post) (95% CI [-0.89, -0.04], p = 0.033) We again conducted a post-hoc Tukey test to compare contrasts. The statistically significant between group results are reported below in table 3:

| task | contrast | estimate | SE | Z-score | p-value |
|---|---|---|---|---|---|
| interaction | control t - control d | 2.1335 | 0.403 | 5.298 | <.0001 |
| interaction | interactive t - control d | 1.9063 | 0.401 | 4.757 | 0.0001 |
| interaction | control t - interactive d | 1.6017 | 0.408 | 3.930 | 0.0048 |
| interaction | interactive t - interactive d | 1.3745 | 0.400 | 3.434 | 0.0294 |
| post | control t - control d | 2.2412 | 0.429 | 5.220 | <.0001 |
| post | control t - interactive d | 1.6837 | 0.433 | 3.892 | 0.0056 |

**Table 3:** Between group differences in the categorization of the three midpoint stimuli

In this analysis the within group comparisons also yielded statistically significant results:

| condition | contrast | estimate | SE | Z-score | p-value |
|---|---|---|---|---|---|
| control t | pre - post | -0.8496 | 0.213 | -3.994 | 0.0038 |
| interactive d | pre - interaction | 0.6116 | 0.179 | 3.423 | 0.0305 |
| interactive t | pre - interaction | -0.9744 | 0.185 | -5.268 | <.0001 |

**Table 4:** Within group differences in the categorization of the three midpoint stimuli

## 4. DISCUSSION

The results show that the categorization behavior between bias conditions differed significantly in the predicted direction: in the post test the /d/ and /t/ bias conditions differed significantly both in the overall analysis and the analysis of the three midpoint stimuli (see tables 1 and 3). There are no statistically significant differences between the categorizations of interactive and control conditions (of the same bias). Within condition there are significant differences between the categorization during the pre-test and the interactive game, a significant difference between pre- and post-test categorization could only be found for the control-t condition. The within condition comparisons of

the overall analysis and the analysis of the three midpoint stimuli show slightly different results: Overall, there are significant differences between pre-test and interactive task categorization in one control and one interactive condition, whereas analysis of the midpoint stimuli shows a significant difference between pre-test and interaction only for the two interactive conditions. A significant difference between pre- and post test categorization can be found in the control-t condition. As shown in figure 1, the control t-bias condition diverts from a pattern the other three exhibit: in the control-t condition, more stimuli are identified as /tin/ in the post-test than in the interactive task. Whereas in the other three conditions the categorization of the post-test lied in between pre-test and interactive task. It is unclear why the control-t condition differs in this regard. It could be that there is an issue of power and that a repetition of the experiment with more participants would result in clearer picture. Further, we only presented five iterations of each acoustic stimulus in the pre- and post-test perception tasks, keeping in mind that in an online experiment, the participants' attention may wane. Follow-up experiments should include attention checks and perhaps more varied stimuli. While interaction does not appear to elicit a robust effect on categorization, the bias condition lead to a significant contrast between groups. Perceptual learning experiments do not always yield a robust effect. For example [20] discuss that some acoustic properties may be easier to learn than others, noting that learning of VOT contrasts by native speakers may be more pronounced than other contrasts. According to [8], adjustments to perceptual representations occur as a function of their cause. In our experiment, participants were asked to categorize non-ambiguous stimuli during the interactive task. Perhaps this impeded participants from adapting their perceptual categories. By simply guessing the categroization of the midpoint stimuli, the participant could still ensure that the understanding was good enough. As the within group contrasts show, categorization of the stimuli was more likely to divert from the pre-test pattern in the interactive task. This could imply that participants adjust their categorization when interacting, but the interaction has no long lasting effect on the perception. This is in contrast to arguments made by [1], who state that changes elicited by perceptual learning are long lasting, but would be similar to what [8] state: Adjustment of perceptual categories is costly, if there is another way to resolve the tension between the acoustic signal and the category, participants will readily do so. In [10] participants further have an external motivation - learning a foreign language - to adapt their perception of a contrast. In our experiment there was no such motivation. Perhaps long lasting changes to perceptual categories are less likely to be elicited if participants do not have a social reason that would yield such changes beneficial. It should further be noted that [17] recommend VOT ranges for stimulus construction which we took into account. However, we noticed that the talent producing the unaltered recordings of our stimuli regularly produced VOTs of more than 100ms for /t/-initial words. Further, initial inspection of the production data showed that the participants themselves regularly produced VOTs of more than 100ms. It is possible that our stimuli were therefore perceived as so extreme that the participant attributed an origin to the stimulus that would inhibit perceptual learning, which would be similar to the findings by [8]. We did however not ask the participants any debrief questions, these interpretations of the participants' possibly socially motivated behavior are therefore speculative. It is also possible that the perceptual system treats the participant's own production as corollary discharge. [21] have shown that self vocalizations are attenuated by the auditory system. Such a suppression of the signal could mean that experimental condition of "interaction" had little influence on the perceptual changes. [9] further argue that perception and production are independent processes that exhibit coordination, but aren't required for each other's function. The results presented here are in line with that perspective.

## 5. CONCLUSION

Overall our results suggest that adjustments in perception are independent of the speaker's own production. How and whether the participants' production changed throughout the experiment and how that relates to the changes in perception will have to be determined in further analysis. Rather than adjusting their perception long term, participants may have opted to employ another strategy to resolve the tension between the acoustic stimulus and the intended category. Future experiments may further explore the social factors that influence the participant's behavior.While this experiment still raises many questions, it offers a starting point to explore the link between speech sound perception and production.

## 6. ACKNOWLEDGEMENTS

This work has been conducted in the framework of the Conversational Brains (COBRA) Marie Sklodowska-Curie Innovative Training Network and has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 859588.

## 7. REFERENCES

[1] R. L. Goldstone, "Perceptual learning," *Annual review of psychology*, vol. 49, no. 1, pp. 585–612, 1998.

[2] T. Kraljic and A. G. Samuel, "Perceptual learning for speech: Is there a return to normal?" *Cognitive psychology*, vol. 51, no. 2, pp. 141–178, 2005.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[4] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 790, 1975.

[5] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[6] N. Nguyen and V. Delvaux, "Role of imitation in the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 46–54, 2015.

[7] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The perception-for-action-control theory (pact): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.

[8] T. Kraljic, S. E. Brennan, and A. G. Samuel, "Accommodating variation: Dialects, idiolects, and speech processing," *Cognition*, vol. 107, no. 1, pp. 54–81, 2008.

[9] J. S. Pardo and R. E. Remez, "On the relation between speech perception and speech production," *The Handbook of Speech Perception*, pp. 632–655, 2021.

[10] R. Akahane-Yamada, Y. Tohkura, A. R. Bradlow, and D. B. Pisoni, "Does training in speech perception modify speech production?" in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE, 1996, pp. 606–609.

[11] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1245–1248, 2009.

[12] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified javascript framework for online studies," in *International conference on computational social science (cologne)*, 2017.

[13] E. U. Grillo, J. N. Brosious, S. L. Sorrell, and S. Anand, "Influence of smartphones and software on acoustic voice measures," *International journal of telerehabilitation*, vol. 8, no. 2, p. 9, 2016.

[14] T. Kraljic and A. G. Samuel, "Generalization in perceptual learning for speech," *Psychonomic bulletin & review*, vol. 13, no. 2, pp. 262–268, 2006.

[15] K. Nielsen, "Specificity and abstractness of vot imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, 2011.

[16] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[17] M. B. Winn, "Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 852–866, 2020.

[18] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/rl/by japanese adults learning english," *Perception & psychophysics*, vol. 36, no. 2, pp. 131–145, 1984.

[21] J. D. Greenlee, A. W. Jackson, F. Chen, C. R. Larson, H. Oya, H. Kawasaki, H. Chen, and M. A. Howard III, "Human auditory cortical activation during self-vocalization," *PloS one*, vol. 6, no. 3, p. e14744, 2011.