INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigating the perception production link through perceptual adaptation and phonetic convergence

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

[1]Aix Marseille University, CNRS, Laboratoire Parole et Langage, Aix-en-Provence, France
[2]The University of Edinburgh, United Kingdom

lena-marie.huttner@univ-amu.fr, noel.nguyen-trong@univ-amu.fr, Martin.Pickering@ed.ac.uk

Abstract

Speech perception and production may be linked. In this pilot study, we aim to investigate the nature of this link by examining how perception and production of VOT adjust in social interaction. In an online experiment with 2x2 conditions participants were asked to categorize nine stimuli on a VOT continuum between /tin/ and /din/ and produce the words din and tin before and after playing a 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 two conditions participants alternated categorizing and producing the stimuli, while in the other two participants only categorized the stimuli. Significant differences in categorization between conditions occurred only during the game. Whereas significant changes in VOT can be observed between pre and post-test for three conditions. This could mean that perception and production do not adjust symmetrically over the course of an interaction.

Index Terms: Perceptual adaptation, phonetic convergence, perception production link

1. Introduction

Over the course of an interaction, interlocutors tend to converge in phonetic-acoustic realization with their partner. Similarly, interlocutors have to adjust their perception of the speech signal to their interlocutor's variable production. In speech production, plasticity of speech processing can be observed in the phenomenon of phonetic convergence – interlocutors becoming more similar in phonetic acoustic realization over time [1, 2]. Whereas, plasticity in speech sound perception can be observed in the phenomenon of perceptual learning [3, 4]: after exposure to an ambiguous stimulus, people adjust their perceptual boundaries to the input. According to the Interactive Alignment Account [5] interlocutors will tend to share their representations on every linguistic level. From this theoretical standpoint, perceptual learning and phonetic convergence can be regarded as manifestations of alignment on the level of speech sounds: A listener will hear their interlocutor's production, and adjust their perception to it. Following sensorimotor theories of speech [6] these two phenomena are not separate, but two sides of the same coin. The underlying cause of phonetic convergence is often hypothesized to be an inherent link between perception and production[5]: in order for people to produce speech more similarly to one another, they must first perceive the sounds and then fine tune their production to the perceived input [7]. While the existence of a link between perception and production is widely accepted, the nature of that link is still debated. It appears that though coordinated, speech perception and production are often regarded to be separate processes. A change in one domain may not suffice to elicit an adjustment in the other [8, 2].There is conflicting experimental evidence on the co–occurrence of perceptual adaptation and convergence in speech production. [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 was not altered. Whereas [9] 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 [2], [9] argue that a close link between speech perception and production is a requirement for category formation. In both of these studies, extralinguistic social factors can explain the participants' behavior. This highlights the overall importance of extralinguistic social knowledge in communication, but it does not necessarily help answer the question how and if speech perception and production are linked and how that link contributes to the shared representations created through language. We venture to explore the link between perception and production experimentally by combining experimental paradigms in phonetic convergence and perceptual learning in a single study, by investigating how a short-term social interaction influences acoustic-phonetic realisation and speech sound categorization.

2. Methods

To investigate whether an adaptation in peception was accompanied by a change in production, we designed an online experiment in Labvanced [10] with two by two conditions. Participants completed the experiment on their smartphones. While the use of smartphones as recording devices opens new doors for researchers to rapidly obtain large amounts of data from a diverse population. [11] studied the aptitude of data recorded on different devices (smartphone models and a head mounted condenser microphone)for phonetic research and found no significant influence of recording device on the data. For technological reasons, our experiment was limited to Android users. In this study we combined the paradigms used in perceptual learning and phonetic convergence experiments. We employed 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.

2.1. Stimuli

We chose VOT as the acoustic feature of interest. VOT has been used as an acoustic feature in both perceptual learning and phonetic convergence experiments[12, 13]. [4] and [14] further found that plosives on a VOT continuum were more likely to result in speaker non-specific perceptual learning than fricatives. We therefore created a 9-step VOT continuum of the words din and tin spoken by a female native speaker of English (dialect re-

10.21437/Interspeech.2023-2077

gion: Northern England). The continuum was created in Praat [15] using a script developed by [16]. The VOT of the stimuli ranged from 23 to 55 ms, the interval between stimuli was 3.6ms. A stimulus test was conducted to evaluate the fitness of the stimuli for the experiment. During the stimulus test, 30 participants (native English speakers form the UK) recruited on Prolific were asked to categorize the 9 stimuli on the continuum.

2.2. Experimental setup

We recruited 88 native English speakers from the UK between the ages of 18 and 40 on Prolific. Participants were required to wear headphones. Compliance with this requirement was ensured using a headphone test [17]. Participants were repeatedly asked to be in a quiet room to complete the experiment. Participants whose data included audio with noticeable background noise (i.e. conversations in the background, a child playing or crying, continued sneezing) were excluded from analysis and replaced. Participants were instructed to turn off any notifications for the duration of the experiment and to place their phones on a table in front of them. To ensure as realistic an interactive setting as possible, 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. Participants were first presented with a consent form followed by a basic demographic questionnaire. 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 on their screen 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 participants. 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 (interactive-t and interactive-d conditions). 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 (control-t and control-d condition). This was a between participant design, each participant only completed one condition. This study received ethics approval prior to data collection. Participants did not consent to the sharing of identifiable data, therefore the audio recordings cannot be made available. The authors will share the datasets without identifiable data upon request.

2.3. Hypotheses

We expect the following changes between pre- and post-test: In light of the findings by [9, 3, 4] we expect perceptual shifts to occur in all four conditions in the direction of the bias of the script. I.e. For the two conditions with a bias towards /tin/ we expect participants to categorize more of the 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. In production we expect the following shifts in VOT for each condition: We expect VOT for t stimuli to shorten in the /t/ bias conditions in the post test compared to the pre-test wheras we expect VOT of /d/ to lengthen in the d-bias conditions. We will further examine differences in categorization and production between interactive and control conditions. According to [6, 18] and [5], effects should be stronger in interactive conditions than in the control condition. However, according to [2], speech perception and production are coordinated but independent of one another, which in our experiment should result in different effects for interactive and control condition.

3. Results – Perception

We here compare the categorization curves within and between conditions. Figure 1 shows the proportion of categorizations of a stimulus as /tin/ per acoustic stimulus in all four conditions. Stimulus level 1 indicates the acoustic stimulus with the shortest VOT (22 ms) and 9 the one with the longest (55 ms). A missing data point, as seen in the interactive /t/ condition (bottom right panel), means that the stimulus was not categorized as /tin/ by any participants in that condition and task.

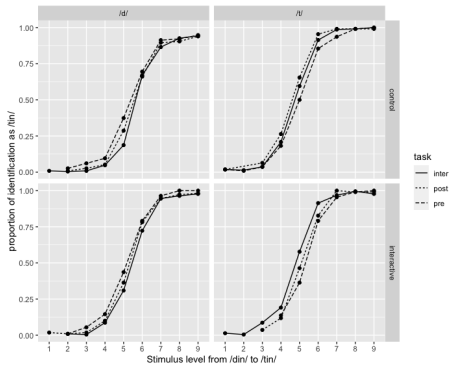

Figure 1: *Proportion of identification of the stimuli as /tin/ by condition. The columns show the bias of the condition, whereas the rows show the control and interaction conditions*

To compare between as well as within group differences, we ran a generalized mixed effects model using the lme4 package [19] in R. With task (pre-test, post- test, interaction) and condition as fixed effects and stimulus and subject as random effects (formula: glmer(response ~ condition * task + (1+ stimulus|subject), family = binomial(link = "logit")).We then ran a post-hoc Tukey test to compare the different conditions. Table 1 shows the significant differences for the between group contrasts:

Table 1: *Between group differences*

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

In this analysis a positive estimate indicates a shift of the means between groups towards /din/, a negative estimate a shift towards /tin/. For example, in the first row of table 1, more stimuli were categorized as /din/ in the control-d condition than in the control-t condition. There are further the following statistically significant within group differences in categorization:

Table 2: *Within group differences*

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

We would 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) to test if there was a difference. We again conducted a post-hoc analysis to compare contrasts. The statistically significant between group results are reported below in 3.

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

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

In this analysis the following within group differences can also be observed:

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

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

While 1 indicates that the pretest curves are significantly different from one another, the contrast analysis shows no such difference.

3.1. Results – Production

The differences in VOT between pre- to post test can be seen in 2; each dot represents a participant's mean VOT in seconds. The columns refer to the spoken stimulus, /din/ on the left, /tin/ on the right, the rows indicate the four different conditions. Within each panel, the violin plot on the left shows the pre-test data, the one on the right the post test data.

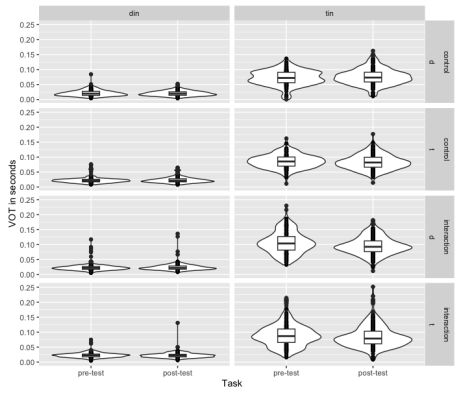

Figure 2: *Differences in VOT between pre and post-test of /din/ and /tin/ per condition.*

As can be seen in 2, the range of mean VOT is quite drastic. To account for the presence of outliers in the data, we ran a robust linear mixed effects model using the robustlmm package [20] in R with condition, task (pre vs post), and stimulus level (tin and din) as fixed effects and participant as random effect (rlmer(vot ~ condition * task * stimulus +(1|subject)). We again ran a post-hoc Tukey test to compare the different conditions. In this analysis, all between group pre-test comparisons were significant. This means that no inferences about between group differences can be made, as the baseline productions were too different for each group. Further, none of the differences for the VOT of /din/ were significant. We will therefore only present the within group comparisons for /tin/:

Table 5: *within group differences between VOT of /t/ in pre and post-test)*

condition	contrast	estimate	SE	z-ratio	p-value
interactive t	pre-post	0.00613	0.00143	4.293	<.0001
interactive d	pre–post	0.00811	0.00144	5.649	<.0001
control t	pre – post	0.00424	0.00143	2.974	0.0045
control d	pre – post	-0.000747	0.00143	0.521	0.7456

In this analysis a positive estimate means the VOT was shortened between pre and post test, whereas a negative estimate indicates a lengthening of VOT.

4. Discussion

The results show that the bias of the experiment script during the interactive task had the intended effect on the participants' categorization behavior between groups: in the post test the /d/ and /t/ bias conditions differed significantly, with exception of the two interactive conditions (see 1). Furthermore, the categorization patterns between interactive and control conditions do not differ significantly. Within group, there are statistically significant differences between the pre test and the interactive game for the interactive-t and control-d conditions. A significant difference between pre-and post test can only be found for the control-t condition (see 2). When only looking at three midpoint stimuli, there are significant differences between the bias conditions during interaction, as well as in the post test (see 3).

Within group, there are significant differences between pre-test and interaction for the two interactive conditions and between pre and post test for the control-t condition. The control-t condition diverts from a pattern the other three conditions exhibit: 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 sits in between pre-test and interactive task. It is unclear why the control-t condition diverts from the other three in this regard. We cannot rule out a Type I error. It is further possible that the number of stimuli repetitions used was not sufficient to elicit the desired effect. However, in online experiments, the participants' time and level of motivation must be considered. The experiment took 25 to 30 minutes in total. Doubling the stimuli would have added another 10 minutes to the experiment in which participants would have completed the same task over and over. Even though participants receive compensation for their participation, this alone cannot ensure that their interest in the study persists. Follow up studies should include more stimuli, along with fillers and attention checks to ensure participants stay engaged throughout the study. As for production, there is a statistically significant decrease in VOT of /t/ between pre and post test in both interactive conditions and the control-t condition, whereas no significant change could be observed for the control-d condition. As can be seen in 2 and 5, the differences in VOT between pre and post test are very small. It is possible that the observed shortening of VOT was not due to coordination with the stimuli, but rather an increase in overall speech rate as the experiment progressed. To account for this, we annotated the duration of speech using the voice detection function in Praat [15] with a threshold of -30db. The data show that there was no change in speech rate between pre and post test. Thus allowing us to infer that the results shown in 5 indicate phonetic convergence on VOT. We predicted a change in the VOT of /d/ as well, no such change could be found in analysis. Upon inspection of the data (see 2) this is not surprising: The VOT range we chose to construct our stimuli was guided by the ranges recommended in [16]. However, the participants regularly produced VOTs of 100ms and longer for /tin/ (see also 2). This means that the range in which the participants' VOT of /t/ could change was simply much larger than that of /d/ in all conditions. While the bias condition did not affect phonetic convergence, the interactive vs control condition may have. As can be seen in 5 VOT shortened in three conditions, both interactive ones and control-d. The estimates reported in able 5 further show that the change in VOT between pre and post test was slightly larger in the interactive conditions than in the control conditions. According to [8], whether an adjustment to the perceptual representation is made depends on the cause of the variation. In our experiment, the listener was still exposed to unambiguous stimuli during the training phase, exposure to the three stimuli in the middle of the continuum during the interaction may not have been sufficient to evoke a change in perception; by simply guessing the categorization of the midpoint stimuli and relying on correct categorization of the endpoint stimuli, participants could still ensure that understanding was good enough. As the within in group contrasts show (see 2 and 4), categorization of the stimuli during the interactive task differed significantly from the pre-test. This could imply that participants readily adjust their categorization when interacting, but that the adaptive effect interaction has on categorization subsides rapidly after the interaction has ended. This would be in line with arguments made by [8]. The adjustment of perceptual categories is costly, participants will readily employ other strategies to resolve the tension between the acoustic signal and the category. However, the change in speech production is observed after the interaction has ended. The lack of a significant difference in categorization between pre- and post test, compared with the phonetic convergence observed could mean that the perceptual system is quicker to adapt than the production system. Further, while both perception and production to an extent shifted throughout the experiment, they did so asymmetrically. As can be seen in figure 1 the categorization between t and d bias conditions shift in opposite directions, as expected. This indicates that adjustments in perception are not a necessary prerequisite for adjustments in production. This may further imply that adjustments in perception and production elicited through social interaction serve different purposes. While a significant lasting change in speech perception could have drastic consequences for an individual's ability to communicate, a slight lasting change in production has less of an impact on somebody's ability to speak. The difference in VOT between pre and post test are more pronounced for the interactive conditions than for the control group, while no such difference can be found in perception. This would be in line with [2], who argue that perception and production are independent processes that exhibit coordination, but aren't required for each other's function. This observed discrepancy between the perception and production data could also be an indication of perception adapting more readily than production, or even of adaptation in perception preceding a change in production.

5. Conclusion

Our results indicate that both speech perception and speech production change as a result of social interaction. However, they do not appear to change in synchrony with or symmetrically to one another. The discrepancy in statistically significant changes between the participants' production and the participants' categorization of the stimuli may mean that plasticity of the perception and production systems serve different ends. A quick adaptation in perception is beneficial to both the success of an interaction as well as the stability of a language system. Rather than adjusting their perception long term, participants may employ other strategies to resolve the tension between the acoustic stimulus and the intended category. The speech production system appears to be under less pressure to adjust quickly. Statistically significant changes in production, however slight, persist after the interaction has ended. Convergence in production occurred between the pre and the post test for three of the four conditions with a larger effect occurring in the interactive conditions. This may indicate that the phenomenon serves a social function. The changes in perception and production do not mirror one another. The bias of the experiment script affected the categorization of the stimuli but did not affect the speech production. Regardless of bias, the participant's production changed in the same manner. This may mean that speech perception and production, while related, are separate processes.

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] M. Natale, "Social desirability as related to convergence of temporal speech patterns," *Perceptual and Motor Skills*, vol. 40, no. 3, pp. 827–830, 1975.

[2] 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.

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

[4] 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.

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

[6] 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.

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

[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] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] ——, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] M. Koller, "robustlmm: An R package for robust estimation of linear mixed-effects models," *Journal of Statistical Software*, vol. 75, no. 6, pp. 1–24, 2016.