Immediate phonetic convergence in a cue-distractor paradigm

Stephen Tobin, Marc Hullebus, and Adamantios Gafos

Citation: The Journal of the Acoustical Society of America **144**, EL528 (2018); doi: 10.1121/1.5082984
View online: https://doi.org/10.1121/1.5082984
View Table of Contents: http://asa.scitation.org/toc/jas/144/6
Published by the Acoustical Society of America

Immediate phonetic convergence in a cue-distractor paradigm

Stephen Tobin,[a] **Marc Hullebus, and Adamantios Gafos**[b]
Linguistics Department, Universität Potsdam, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany
sjtobin@umich.edu, hullebus@uni-potsdam.de, gafos@uni-potsdam.de

Abstract: During a cue-distractor task, participants repeatedly produce syllables prompted by visual cues. Distractor syllables are presented to participants via headphones 150 ms after the visual cue (before any response). The task has been used to demonstrate perceptuomotor integration effects (perception effects on production): response times (RTs) speed up as the distractor shares more phonetic properties with the response. Here it is demonstrated that perceptuomotor integration is not limited to RTs. Voice Onset Times (VOTs) of the distractor syllables were systematically varied and their impact on responses was measured. Results demonstrate trial-specific convergence of response syllables to VOT values of distractor syllables.

[CCC]
Date Received: August 26, 2018 **Date Accepted:** November 23, 2018

1. Introduction

Well-entrenched motor habits are not straightforwardly modifiable in speech or other areas of motor skill. Yet habits can change. We know this because of studies that demonstrate change in speakers' phonetic characteristics to match those of ambient speech, referred to as phonetic convergence (Babel, 2010; Nielsen, 2011; Pardo, 2006; Sancier and Fowler, 1997; Sato *et al.*, 2013; Tobin *et al.*, 2017, among others). However, such effects have never been demonstrated at the shortest time scales of single perception-production loops. Rather, they are reported as the effects of trials distributed over the course of periods of typically at least 30 min in laboratory-based investigations and substantially longer periods in the case of investigations involving immersion in a different ambient language. For example, Fowler *et al.* (2003) tested whether the response Voice Onset Times (VOTs) of participants shadowing VCV sequences containing voiceless stops would be longer when the stimulus sequence stops had artificially extended VOTs compared with response VOTs to unmodified stops. Results indeed showed significantly longer VOTs in response to artificially extended VOTs than unmodified VOTs, that is, evidence for convergence toward the stimulus VOTs. However, assessment of convergence was based on distributions of deviations from the mean VOT for each condition (extended vs unmodified VOT) drawn from the full duration of the experiment. In our present study, we aim to uncover changes in VOT on a trial-specific basis. No previous study has demonstrated such changes at this scale of individual perception-production loops.

One line of work, using the so-called cue-distractor paradigm, has been particularly successful at demonstrating trial-specific perceptuomotor integration effects in speech (Kerzel and Bekkering, 2000; Galantucci *et al.*, 2009; Roon and Gafos, 2015). In a cue-distractor paradigm, participants utter syllables in response to visual cues ("if you see ** say /ka/, if you see ## say /ta/"). After a visual cue appears but before any response, participants are presented with a distractor syllable and are told to ignore it. This paradigm differs from that used in shadowing tasks, in that shadowing stimuli constitute both the cue to respond and the target of convergence, whereas in the cue-distractor paradigm, the distractor is presented after the visual cue and during speech planning. Roon and Gafos (2015) presented auditory distractors that never matched responses in having the same place of articulation, but had voicing that was either congruent (e.g., /ta/ response, /pa/ distractor with both syllable-initial consonants voiceless) or incongruent (e.g., /ta/ response, /ba/ distractor with /ta/ voiceless vs /ba/ voiced syllable-initial consonant) with the response. Response times (RTs) were slower in the incongruent case than in the congruent case. However, all previous studies using this paradigm employ distractors whose continuous phonetic parameters are constant: /pa/ distractors had a fixed value of VOT

[a]Present address: Department of Linguistics, University of Michigan, 440 Lorch Hall, 611 Tappan Avenue, Ann Arbor, MI 48109, USA. Author to whom correspondence should be addressed.
[b]Also at: Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA.

throughout the experiment. /pa/ was shown to speed up production of /ta/ (more than /da/) but no effect on the VOT of the /ta/ response could be assessed.

We present here an extension of the cue-distractor paradigm wherein we systematically varied the phonetic distance between the mean VOT of a particular speaker-hearer's speech and the distractor VOTs with which that speaker-hearer is presented. Our hypothesis is that perceptuomotor integration extends beyond RTs to the actual subcategorical phonetic properties of the produced responses. That is, we hypothesized that phonetic details of the produced response syllable, within each trial, would be subtly affected by those of the auditorily presented distractor within that trial. Testing this hypothesis required that we register participant-specific VOTs in a baseline block (without any distractors) and that we assign distractors to participants on the basis of participant-specific baseline phonetic properties. In order to assess changes in VOT on a trial-specific basis, we derived a dependent variable (response-baseline differential: δ^{rb}), in which every data point constitutes a measure of change from the baseline. This is in distinction to the standard approach of comparing overall measures of raw VOT between a baseline vs experimental block, wherein VOTs from all within-block trials are pooled. Furthermore, we presented the distractor stimuli very soon (150 ms) after the visual cue, so that participants' planning of the response would be ongoing by the time they heard the distractor. We thus intended to influence speech planning in progress (insofar as planning is distinct from execution, cf. Löfqvist, 2010). Our results provide, for the first time, evidence that participants' responses converge to these distractor VOTs.

2. Method

Twenty-two undergraduate students at Universität Potsdam were recruited to participate in the experiment and received course credit as compensation. All were native speakers of German with no history of speech or hearing disorders. Participants sat in a sound-treated booth facing a computer monitor, wearing headphones. Experimental sessions were divided into a baseline ($n = 100$) block followed by three experimental blocks yielding ($n = 3 \times 240$) 720 experimental trials. In the baseline block, participants were instructed to produce a syllable each time they saw a visual cue on the monitor; /ta/ for visual cue "##" and /ka/ for visual cue "**." On each trial, a fixation cross appeared for 500 ms at the center of the screen followed by the visual cue (## or **), which was presented in gray on a black background. Participants were instructed to respond as fast as possible once the cue appeared. The visual cue was kept on the screen until a response was detected by the built-in microphone of the presentation computer. Subsequently, the visual cue disappeared, and a new trial began after 800 ms. The 100 trials of the baseline block were evenly divided between the two cue syllables, yielding 50 tokens of each syllable. No auditory distractor syllables were presented during this baseline block. After the baseline, the experimenter ran an automatic acoustic landmark-detection algorithm to estimate the participant's baseline VOTs, which took 3–5 min. This allowed us to promptly assign participants to distractor VOT ranges near to or far from their baseline (see below). The baseline VOTs were also verified manually after the experiment with a semi-automatic algorithm. All measurements were carried out in software (Kuberski *et al.*, 2016) whose performance with respect to other landmark identification methods has been quantified. Response VOT and syllable duration were computed based on the stop release burst, phonation initiation, and cessation landmarks. VOT was the interval in milliseconds between release burst and phonation initiation. Syllable duration was the interval between stop release burst and phonation cessation. RT was the interval between visual cue onset and release burst.

After the baseline block and the estimation of the participant-specific baseline VOTs, the experimental blocks were administered. Trials during these blocks were identical to those of the baseline block with the exception that 150 ms after the presentation of the visual cue participants heard a distractor that phonemically matched (e.g., cue: "##" /ta/, distractor: /ta/) or mismatched (e.g., cue: "##" /ta/, distractor: /ka/) the intended response but varied in VOT. Participants were told to ignore everything they heard. Distractor syllable stimuli were drawn from 9-step VOT continua for /ta/ and /ka/, ranging from 45 to 85 ms in 5 ms steps. Continua were created in Praat by resynthesizing natural tokens of a female native German speaker. Stimuli were normalized to equal duration by trimming and fading the end of the vowel to 235 ms and to equal amplitude (73 dB). To ensure variation in phonetic distance between baseline and distractor VOT, half of the participants with short baseline VOTs (<65 ms) and half with long ones (>65 ms) heard the shorter five steps of the VOT continua (45–65 ms). The other halves of these two groups heard the longer five steps (65–85 ms). Experimental sessions lasted about 30 min.

3. Results

3.1 Baselines

A descriptive overview of our data is presented in Fig. 1. Density plots of individual participants' baseline VOTs for /ta/ and /ka/ appear on the left and boxplots of individual participants' response syllable durations in the baseline and distractor tasks, on the right. We see not only substantial speaker-specific variation in VOT distributions and syllable durations, but also considerable within-speaker variation among the responses—instances of lengthening and shortening of syllable duration between the baseline and distractor tasks can both be seen.

It is well known that syllable duration and VOT are correlated. Longer syllables yield longer VOTs (Allen *et al.*, 2003). In our data too, syllable duration is correlated with VOT both in the baseline ($r = 0.46$, $t = 2.34$, $p = 0.03$) and in the distractor task ($r = 0.64$, $t = 3.68$, $p = 0.001$; Fig. 2, left). Likewise, RT is also correlated with syllable duration, both in the baseline ($r = 0.57$, $t = 3.07$, $p = 0.01$) and in the distractor task ($r = 0.51$, $t = 2.65$, $p = 0.02$; Fig. 2, right).

In assessing convergence, our dependent measure uses the quotient of the (raw) response VOT and response syllable duration (VOT^r / σ^r), that is, the syllable duration-normalized response VOT. This allows one to maintain better data hygiene in two ways. It helps to identify cases in which raw response VOT happened to lengthen in the distractor task compared to baseline as a mere by-product of syllable duration lengthening, rather than genuine convergence of response VOTs to a longer distractor VOT. By way of illustration, consider participant 5 (P5, Fig. 3, top row). The boxplots in the top left panel indicate longer response syllables in the distractor task than in the baseline. The density plots in the top center and right panels compare distributions of response VOTs in the baseline and distractor tasks using raw VOTs centered on the distractor VOT mean (center) and syllable duration-normalized response VOTs (right). In the raw VOT plot, response VOTs from the distractor task are shifted upwards from baseline toward the mean distractor VOT (shown by the vertical line). However, this apparent increase in response VOTs disappears when the response VOTs are normalized by syllable duration (top right). Conversely, some participants whose syllable durations are substantially shorter in the distractor than in the baseline task do not appear to converge toward the mean distractor VOT if we only consider raw VOTs. Yet, they do show convergence to the distractor in the syllable duration-normalized response VOT measure. In the presence of distractors whose mean VOT is longer than that of the baseline, raw response VOTs may appear to go unchanged. This apparent lack of convergence is due to shortening of the response syllables to which the VOTs belong, with concomitant shortening of the response VOTs themselves as a result of the well-known relation between syllable duration and VOT. Participant 9 (P9) is an example (Fig. 3, bottom row). In contrast to P5, P9's syllable durations are shorter in the distractor task than in the baseline (Fig. 3, bottom left). Whereas centered raw response VOT distributions from baseline and distractor tasks largely overlap, suggesting no robust convergence, distributions for VOT^r/σ^r distributions indicate a robust VOT increase in response to distractors.

3.2 Relation between distractor and response VOT

We introduce some notation. We write VOT^b, VOT^d, and VOT^r for the VOTs of the baseline, distractor, and response, respectively. Our independent variable is the difference between distractor and mean baseline VOT, calculated separately for /t/ and for /k/: VOT^d - VOT^b. We refer to this as δ^{db} (distractor-baseline differential). Our aim is

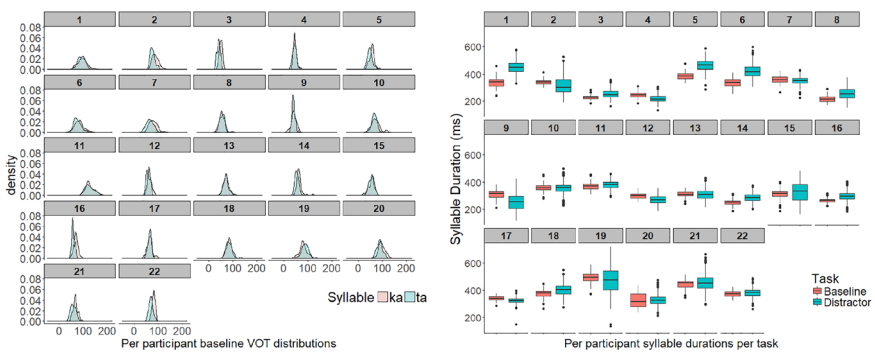

Fig. 1. (Color online) (Left) Baseline VOT distributions per participant, ordered by increasing mean VOT from top left to bottom right. (Right) Syllable durations per participant in baseline and distractor tasks.

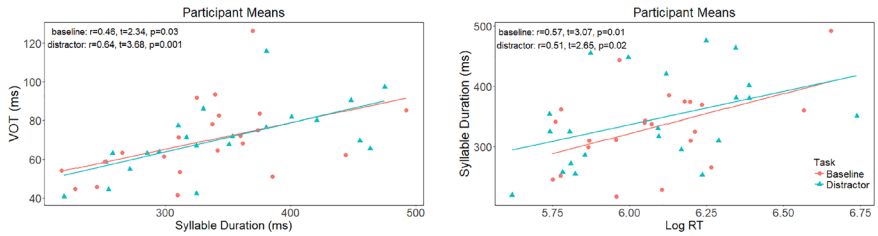

Fig. 2. (Color online) (Left) Correlation of VOT and syllable duration. (Right) Correlation of RT and syllable duration. Data are participant means in baseline and distractor tasks.

to scale δ^{db}, increasing it when the distractor has a VOT above baseline or decreasing it when the distractor has a VOT below baseline, and to observe whether δ^{db} has a systematic effect on response VOTs. For our dependent variable, we use the quantity $(VOT^r/\sigma^r) - (VOT^b/\sigma^b)$, the difference between syllable duration-normalized response VOT and syllable duration-normalized baseline VOT. We refer to this quantity as δ^{rb} (response-baseline differential).[1] We plot our data on the $\delta^{db} \times \delta^{rb}$ plane. In this plane, convergence is indicated by a line with a positive slope: the higher the distractor VOT relative to baseline VOT (greater positive difference as expressed by δ^{db}), the higher the δ^{rb}. Conversely, the lower the distractor VOT relative to baseline VOT (greater negative difference as expressed by δ^{db}), the lower the δ^{rb}.

We analyzed the data with linear mixed-effects regression in the R statistical environment (Bates *et al.*, 2015). The continuous independent variables δ^{db} (distractor-baseline differential) and RT ($\log_{10}$ transformed), and the dichotomous independent variables Cue Syllable (/ta/ vs /ka/, dummy coded with /ta/ as the reference level) and Congruency (*match* vs *mismatch*, dummy coded with *match* as the reference level) and their interactions were fixed factors. The continuous independent variables were centered to reduce multicollinearity. The model included random intercepts for Participant and by-Participant random slopes for δ^{db}. Our dependent variable is δ^{rb} (response-baseline differential). We used the lmerTest package to estimate *F*-statistics, denominator degrees of freedom, and *p*-values corresponding to the linear mixed-effects regression output for brevity (Kuznetsova *et al.*, 2015). Regression results are presented in Fig. 4. The regression lines in this figure are plotted in the δ^{db} (independent variable) $\times$ δ^{rb} (dependent variable) plane with 90% confidence intervals. The six panels in the figure show regression lines from our model at six representative points along the continuous variable of RT: early (200–300 ms), mid (400–500 ms), and late (600–700 ms). The top row corresponds to the /ta/ Cue Syllable condition, while the bottom row corresponds to the /ka/ Cue Syllable condition. Solid lines represent matching cue + distractor pairs (cue = /ta/, distractor = /ta/; cue = /ka/, distractor = /ka/) and dashed lines represent mismatching cue + distractor pairs (cue = /ta/, distractor = /ka/; cue = /ka/, distractor = /ta/).

The pattern of results shows a significant interaction of δ^{db} and RT [$F(1, 4679.4) = 46.5976$, $p < 0.01$]. The interplay of these effects can most clearly be seen in Fig. 4. At faster RTs, increases in δ^{db} yield increases in δ^{rb}, indicating convergence. In other words,

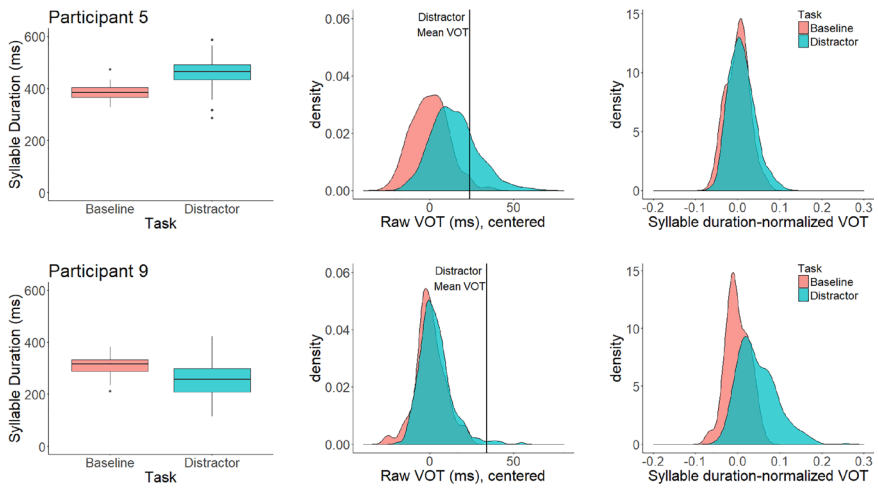

Fig. 3. (Color online) P5's syllable durations increase while P9's decrease in the distractor task relative to the baseline (leftmost top/bottom panels). Baseline-centered raw VOT densities suggest convergence for P5 (top center) but not for P9 (bottom center). With syllable duration-normalized VOT the evidence reverses. P5 does not show robust convergence but P9 does.

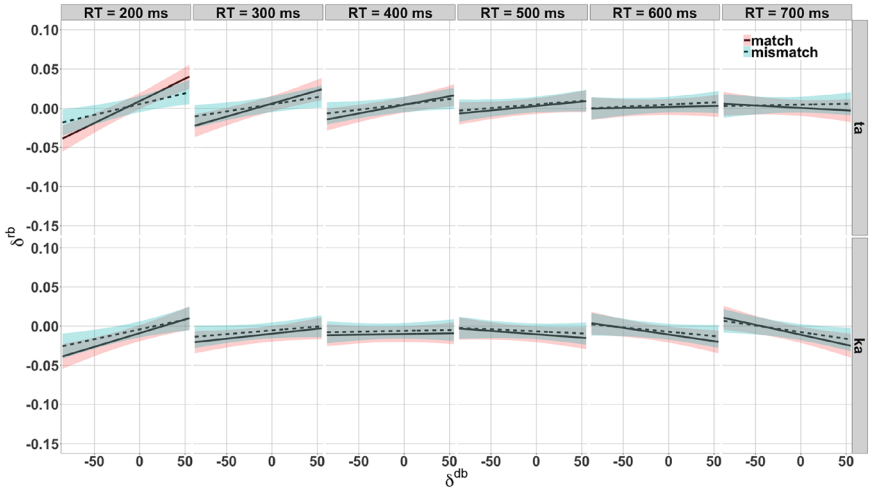

Fig. 4. (Color online) $\delta^{db} \times \delta^{rb}$ regressions with 90% confidence intervals. Positive slopes indicate convergence. Six representative points along the continuous variable RT are shown from left to right. Results from /ta/ vs /ka/ as Cue Syllables are shown in the top vs bottom row. Match and mismatch levels of Congruency are shown in solid vs dashed lines.

among faster responses, as distractor VOT moves away (either above or below) from a participant's baseline VOT, the response VOT converges toward the distractor VOT. This is visible in the positive slopes in Fig. 4. At slower RTs, however, the impact of δ^{db} is diminished. This is clear in the flattening of the regression lines in the right panels of Fig. 4.

We further observe a significant 3-way interaction of $\delta^{db} \times \text{RT} \times \text{Congruency}$ [$F(1, 14\,733.0) = 10.7266$, $p < 0.01$]: Congruency modulates the observed interaction of $\delta^{db} \times \text{RT}$: when cue and distractor *match*, the interaction is stronger, whereas when cue and distractor *mismatch*, the interaction is weaker. In Fig. 4, this is visible from the steeper and more variable slopes of the solid (*match*) lines vs the shallower and less variable slopes of the dashed (*mismatch*) lines. The stronger effect on δ^{rb} of $\delta^{db} \times \text{RT}$ among cue-distractor pairs matching in constriction location is consistent with the greater impact of matching/congruent tokens than mismatching/incongruent tokens on RT in prior cue-distractor investigations (Galantucci *et al.*, 2009; Roon and Gafos, 2015). Greater compatibility of distractor and cue syllable allows greater integration of the phonetic parameters of the distractor and the cued syllable in speech planning.

In order to test our fixed effects for multicollinearity, we calculated the Variance Inflation Factors of the model and found them all to be lower than 1.3. Though not of special experimental interest, we also note for completeness the two remaining significant effects: Cue Syllable /ka/ has a significantly higher intercept than Cue Syllable /ta/ [$F(1, 11\,222.7) = 189.3713$, $p < 0.01$], and cue-distractor pairs that *mismatch* have a significantly higher intercept than those that *match* [$F(1, 14\,731.4) = 17.3750$, $p < 0.01$].

The interactions suggest that effects of RT are tightly bound to the phonetic parameter effects we report. Remaining effects were not significant.

4. Discussion

Our main result is trial-specific convergence of response syllables to VOT values of distractor syllables. This convergence is modulated by reaction time. Why is convergence more evident at faster RTs? Recall the significant $\delta^{db} \times \text{RT}$ interaction: the impact of δ^{db} is greatest among shorter RTs, becoming weaker as RT increases. Consider a speaker's action upon presentation of the visual cue for a /ta/. In responding to the cue, the speaker must assemble a set of parameter values that specify the required vocal tract action. These include (but are not limited to) organ-specific parameters referring to the constriction location of the organ forming the consonant as well as a parameter specifying the VOT for that consonant and, crucially, metrification parameters relevant to the prosody of the response (Dell *et al.*, 1993). The latter include the duration of the syllable, the frame into which the segmental content must be fit. At the fastest RTs, metrification is completed at times comparable to those of the convergence effect. Once metric planning is fixed, VOT cannot be modified subsequently by changes in syllable duration (recall that syllable duration is correlated with VOT as shown in Fig. 2, left). It follows that convergence should surface most transparently at the fastest RTs, untainted by any additional effects deriving from syllable duration modification. We assume, as it is standard in phonetics and other areas of cognition and action, that planning is a stage during

which parameters (for us, VOT) settle to their chosen values for production or movement execution and that this stage takes place before movement execution (Catford, 1977, Erlhagen and Schöner, 2002). As RTs increase, syllable durations also increase (Fig. 2, right). This in turn affects the VOTs of the prolonged syllables to an extent that obscures any reliable effects of VOT change due specifically to convergence.

In addition to being consistent with prior work on (mis)matching cue-distractor pairs in this paradigm, the positive slopes among mismatching cue-distractor pairs suggest a non-articulator-specific level of planning. The observation is also consistent with Nielsen (2011), who found transfer of VOT convergence from shadowed /p/-initial to novel /k/-initial words.

Consider the evidence for divergence at the late RTs. Interpretation should be cautious here for at least three reasons. First, given that there is a positive correlation between VOT and RT (and thus between δ^{rb} and RT), as there is between σ duration and VOT, we are liable to find ceiling effects among longer RTs. Since longer RTs are associated with longer VOTs and higher values of δ^{rb}, participants' VOTs (and δ^{rb} values) may reach a VOT(δ^{rb}) ceiling range beyond which they are unlikely to go, in which these values are largely governed by RT. Second, as discussed in the preceding paragraph, when RTs lengthen, effects on our dependent measures orthogonal to convergence accumulate (such as the effect of syllable duration modification). Third, at large δ^{db} values (distractors with VOTs either far below or far above baseline VOTs), relevant observations become sparse. Less than 1.5% of the data come from RTs after 600 ms and large values of δ^{db}. Nevertheless, consider the quantity (VOT^r/σ^r) - (VOT^b/σ^b) at long RTs and as δ^{db} increases. At long RTs, σ durations of the responses also lengthen so that the first ratio in (VOT^r/σ^r) – (VOT^b/σ^b) decreases (because σ^r increases) while the second ratio remains constant. As δ^{db} increases (distractor VOT extends further and further above the baseline VOT), VOT^r does not follow suit: convergence may not be a simple linear function of the distance between distractor and baseline VOT. Consider findings from similar paradigms in other domains. In oculomotor and manual reaching studies, deviations toward distractors occur only when distractor and target are located close enough together (e.g., within 20° to 30° of the visual field) and saccade endpoints in these cases are usually in between target and distractor (Van der Stigchel and Theeuwes, 2005). Again, our datasets do not suffice to address these issues in our domain due to the paucity of relevant observations. More encompassing distractor-baseline differentials and more abundant observations at large values of δ^{db} are both required to explore these issues further. In sum, at longer RTs, there is more time than at fast RTs for other effects (additional to the effect that the distractor's VOT has on the VOT of the planned response) to change the ultimately observed VOT value.

What are the implications of our results for perception-production models? Past work on perceptuomotor effects sheds light either on phonetic parameter values (e.g., Sancier and Fowler, 1997; Nielsen 2011) or on how cue-response congruency affects the RTs (e.g., Kerzel and Bekkering, 2000; Galantucci *et al.*, 2009). Our results indicate that phonetic parameter values and RTs are related. Joint observations of phonetic parameter values and RTs thus offer prime data for developing perception-production models. Attending to both better constrains our understanding of the perception-production link than limiting focus to one of these aspects alone.

5. Conclusion

Imitating speech, whether intentionally or unintentionally, necessitates transforming perceptual input to vocal tract motor output. A cue-distractor task requires *concurrent* use of both perception and production, as participants hear distractors while planning responses. This differs from many tasks employed in investigations of convergence (e.g., shadowing, conversational interaction), where convergence occurs prior to speech planning for the response. Whereas previous cue-distractor studies have established perceptuomotor interactions in RTs, the phonetic characteristics of their distractor stimuli were kept constant. In an extension of this line of work, we varied the VOT of the distractors and asked whether perceptuomotor effects occur not just in RTs but also in VOT values. Participants' baseline VOTs were measured before delivery of distractor stimuli. This enabled quantification of changes in response VOTs for any given participant as a result of exposure to distractors with systematically different VOTs. Results indicate that within each trial, response VOTs were subtly attracted to VOTs of the auditorily presented distractors. Effects were more evident at the faster RTs. Overall, our results constitute the first demonstration of trial-specific phonetic convergence (seen most reliably at faster RTs) of response syllables to VOT values of distractor syllables.

What is the import of our results for change over longer time scales? Our general hypothesis was that perceptuomotor interactions subserve phonetic change. Crucially, the existence of convergence at the microchronic time scale, as demonstrated here in terms of VOT, does not necessitate macrochronic change; microchronic refers to, as in Catford (1977), the time scale of a few hundred milliseconds as in one trial of our experiment and macrochronic to the time scale that pertains to months or years as in, for example, the change implicated in the VOT effects reported in Sancier and Fowler (1997). It neither follows from our hypothesis nor is it a consequence of our demonstration that lasting change necessarily happens when a listener is exposed to speech from speakers whose production characteristics in terms of VOT differ from those of the listener. This is due to nesting of time scales. Perceptuomotor interactions take place at the microchronic time scale. When fast interactions are embedded in longer time scales, their effects are fragile. Specifically, effects at the fastest time scales can wash out due to the very same reasons giving rise to their existence. If a participant's production VOTs shorten as a result of being exposed to VOTs shorter than the participant-specific mean, listening to VOTs that are longer than the participant-specific mean would have the opposite effect. Issues of input variability, multiplicity of interactional partners, length, and consistency of exposure are all factors modulating change, or lack thereof, at longer time scales. While keeping these factors in mind, our study contributes to isolating the common denominator of any convergence, that is, the subtle effects that listening to speech has on the production of speech.

Acknowledgments

The authors gratefully acknowledge support by ERC Advanced Grant No. 249440 and by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 317633480–SFB 1287, Projects C03 and C04.

References and links

[1]Note that the δ^{db} measure cannot be normalized because the distractor duration (unlike response duration in the delta response-baseline measure) does not vary.

Allen, J. S., Miller, J. L., and DeSteno, D. (**2003**). "Individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **113**, 544–552.

Babel, M. (**2010**). "Dialect divergence and convergence in New Zealand English," Lang. Soc. **39**, 437–456.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**(1), 1–48.

Catford, J. (**1977**). *Fundamental Problems in Phonetics* (Edinburgh University Press, Edinburgh).

Dell, G. S., Juliano, C., and Govindjee, A. (**1993**). "Structure and content in language production: A theory of frame constraints in phonological speech errors," Cog. Sci. **17**, 149–195.

Erlhagen, W., and Schöner, G. (**2002**). "Dynamic field theory of movement preparation," Psychol. Rev. **109**, 545–572.

Fowler, C., Brown, J., Sabadini, L., and Weihing, J. (**2003**). "Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks," J. Memory Lang. **49**, 396–413.

Galantucci, B., Fowler, C. A., and Goldstein, L. (**2009**). "Perceptuomotor compatibility effects in speech," Attn., Percept., Psychophys. **71**, 1138–1149.

Kerzel, D., and Bekkering, H. (**2000**). "Motor activation from visible speech: Evidence from stimulus response compatibility," J. Exp. Psychol.: Human Percept. Perf. **26**, 634–647.

Kuberski, S. R., Tobin, S. J., and Gafos, A. I. (**2016**). "A landmark-based approach to automatic voice onset time estimation in stop-vowel sequences," in *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 60–65.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2015**). "lmerTest: Tests in linear mixed effects models. R package version 2.0–20," Vienna: R Foundation for Statistical Computing.

Löfqvist, A. (**2010**). "Theories and models of speech production," in *The Handbook of Phonetic Sciences*, edited by W. Hardcastle, J. Laver, and F. Gibbon (Wiley-Blackwell, Malden, MA).

Nielsen, K. (**2011**). "Specificity and abstractness of VOT imitation," J. Phonetics **39**, 132–142.

Pardo, J. (**2006**). "On phonetic convergence during conversational interaction," J. Acoust. Soc. Am. **119**(4), 2382–2393.

Roon, K. D., and Gafos, A. I. (**2015**). "Perceptuo-motor effects of response-distractor compatibility in speech: Beyond phonemic identity," Psych. Bull. Rev. **22**, 242–250.

Sancier, M. L., and Fowler, C. A. (**1997**). "Gestural drift in a bilingual speaker of Brazilian Portuguese and English," J. Phonetics **25**, 421–436.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., and Nguyen, N. (**2013**). "Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production," Front. Psychol. **4**, 422.

Tobin, S. J., Nam, H., and Fowler, C. A. (**2017**). "Phonetic drift in Spanish-English bilinguals: Experiment and a self-organizing model," J. Phonetics **65**, 45–59.

Van der Stigchel, S., and Theeuwes, J. (**2005**). "The influence of attending to multiple locations on eye movements," Vision Res. **45**, 1921–1927.