languages

Article

Exploring the Onset of Phonetic Drift in Voice Onset Time Perception

Jackson Kellogg * and Charles B. Chang

Department of Linguistics, Boston University, Boston, MA 02215, USA
* Correspondence: jkellogg@bu.edu

Abstract: Recent exposure to a second or foreign language (FL) can influence production and/or perception in the first language (L1), a phenomenon referred to as phonetic drift. The smallest amount of FL exposure shown to effect drift in perception is 1.5 h. The present study examined L1 perception at earlier timepoints of FL exposure, to determine whether the phonetic system is able to resist FL influence at an incipient stage. In a longitudinal pre-test/post-test design, L1 English listeners were exposed to Tagalog under different conditions varying in attention directed to the voice onset time (VOT) plosive contrast in the FL; they then completed an identification task on L1 tokens from VOT continua. In every condition, the likelihood of "voiceless" identifications decreased. This change indicates a shift towards a longer VOT crossover point between "voiced" and "voiceless", consistent with dissimilatory drift in perception. Listeners in a control condition, however, displayed a similar, albeit less lasting, change in L1 judgments, suggesting that the change arose partly from a task effect. We conclude by discussing directions for future research on phonetic drift in perception.

Keywords: phonetic drift; speech perception; VOT; Tagalog; English; laryngeal contrast; voicing; perceptual learning; crosslinguistic influence; attention

check for updates

Citation: Kellogg, Jackson, and Charles B. Chang. 2023. Exploring the Onset of Phonetic Drift in Voice Onset Time Perception. *Languages* 8: 78. https://doi.org/10.3390/languages8010078

Academic Editors: Olga Dmitrieva, Chiara Celata, Esther de Leeuw and Natalia Kartushina

Received: 20 August 2022
Revised: 28 February 2023
Accepted: 2 March 2023
Published: 8 March 2023

1. Introduction

The assumption that phonetic changes to the first language (L1) in adulthood would arise only out of extended contact with a second language (L2) (see, e.g., Major 1992) was challenged by the discovery that short-term, recent L2 experience can also influence the L1 (Chang 2012; Tice and Woodley 2012; Kartushina et al. 2016), a phenomenon referred to as phonetic drift (hereafter, "drift"; see Chang 2019b for a review). Apart from highlighting the need for a careful accounting of participants' language background (Chang 2019a), the occurrence of drift suggests that the L1 phonetic system may be more plastic and susceptible to crosslinguistic influence (CLI) than the other components of the L1 grammar (de Leeuw and Celata 2019). However, the extent of L1 phonetic plasticity remains poorly understood, pointing to the need for further research investigating changes in the L1 phonetic system at early stages of exposure to an additional language, such as an L2 that is actively being acquired or a foreign language (FL) that is not being acquired.

The present study sought to identify the earliest timepoint and most impoverished circumstances of FL exposure associated with drift in L1 perception (hereafter, "perceptual drift"). A small, but growing, literature has provided evidence of perceptual drift in various L1s, including English (Tice and Woodley 2012; Lev-Ari and Peperkamp 2013), Catalan (Mora and Nadeu 2012), French (Namjoshi et al. 2015), Mandarin (Gong et al. 2016), Brazilian Portuguese (Cabrelli et al. 2019), Russian (Dmitrieva 2019), Spanish (Gorba 2018, 2019), Japanese (Takahashi 2020), and Polish (Sypiańska and Cal 2022); however, most previous findings are based on cross-sectional data only. In this paper, we contribute the first longitudinal data on perceptual drift in English during initial exposure to Tagalog. These data, which include 20 tests of L1 perception over five days, position us to explore L1 phonetic plasticity in more detail than has been possible previously.

2. Background

2.1. Phonetic Drift in Perception

One of the earliest studies to observe perceptual drift is that of Tice and Woodley (2012), in which L1 English listeners were longitudinally tested on their perception of L1 plosives during the first six weeks of an introductory L2 French course. To observe the crossover point at which listeners would perceive a categorical shift between voiced and voiceless plosives, continua of plosive tokens were synthesized to incrementally span a range in voice onset time (VOT) of −85 to 90 ms in 12 steps. Following the third week of L2 instruction, participants displayed a significantly shorter VOT crossover point, meaning that a subset of tokens identified as voiced at the start of the study came to be identified as voiceless. This shift was consistent with a categorical difference between English and French: English is an "aspirating" language where, in word-initial or pre-tonic position, the voiced–voiceless VOT boundary occurs at a relatively long VOT value (around +30 ms) and short-lag VOT signals voiced plosives (Lisker and Abramson 1964), while French is a "true voicing" language where the voiced–voiceless VOT boundary occurs at a short VOT value (around 0 ms) and short-lag VOT signals voiceless plosives (Lein et al. 2016). The participants in Tice and Woodley's study were thus interpreted as displaying assimilatory drift in perception of L1 plosives towards the shorter VOT values of L2 French, due to an equivalence classification (Flege 1987) between the L1 and L2 voicing categories. A control group, conversely, displayed no significant change in L1 judgments.

Following Tice and Woodley (2012), there have been few subsequent studies of perceptual drift that focused specifically on L1 changes during short-term, recent L2 exposure, as opposed to L1 changes after extended immersion in an L2 environment (i.e., phonetic attrition; see de Leeuw and Chang 2023); however, the study of Gorba (2019) is one exception. This study tested perception of L1 VOT category boundaries in a cross-section of L1 Spanish–L2 English learners varying in L2 proficiency, including less experienced learners who had only engaged in non-immersive L2 classroom learning. Compared to monolingual peers, this less experienced learner group displayed an assimilatory shift in perception of L1 VOT towards the longer VOT crossover point of the L2, albeit to a lesser degree than the more experienced learners residing in the UK. The less experienced group was described as university students with no experience living in an English-speaking country, thus resembling the novice learners in Tice and Woodley's work; however, it is unclear how recently or for how long this group had been learning English at the time of the study, leaving open the question of the threshold of L2 or FL exposure needed for precipitating drift.

To our knowledge, the smallest amount of FL exposure leading to perceptual drift is provided in Gong et al. (2016), which examined whether the forced linkage of L1 and FL sounds would result in a destabilization of the L1 phonetic space. In this study, 20 L1 Mandarin Chinese speakers with no experience in Spanish were exposed to Spanish in a series of 10-min sessions.[1] Four sessions were held per day over a four-day period, for 16 sessions total. During these sessions, participants were played vowel-consonant-vowel tokens spoken with Spanish phonology and were instructed to map the medial sound to one of 18 Mandarin consonants presented orthographically on an onscreen keyboard. In addition, they completed an L1 perception task three times—once before (pre-task), once during (mid-task), and once following (post-task) the FL exposure period—to identify any change in listeners' consonant reception thresholds (CRTs) (see Plomp and Mimpen 1979) for various L1 sounds. The L1 perception task involved identifying Mandarin consonants played within speech-shaped noise, with the amplitude of the consonants incrementally raised in 2-dB steps to determine the lowest level at which they could be accurately identified (i.e., the CRT); a higher CRT indicates a lower tolerance for noise, implying a weaker phonetic representation of the target sound. At mid-task, with eight sessions having elapsed for roughly 1.5 h of FL exposure, participants displayed significantly higher CRTs for the L1 consonants /l/ and /w/. This result was attributed to the phonetic expansion of these categories via perceptual linkage with similar FL sounds. This study is notable for showing that just 1.5 h of FL exposure can lead to a significant change in L1 perception. However, it did not assess L1 perception before the eighth FL exposure session, raising the question of whether perceptual drift might have started even earlier.

As for the trajectory of perceptual drift, the literature presents a mixed picture. On the one hand, some studies show that drift is more prominent at early than late stages of L2 development, consistent with a novelty effect for an unfamiliar L2 (Chang 2013). Tice and Woodley (2012), for instance, observed perceptual drift in L1 plosive voicing judgments during the third and fourth weeks of an introductory L2 course, and then the restabilization of these judgments towards the baseline in subsequent weeks. Gong et al. (2016) observed a similar trajectory of drift in the perception of L1 /l/ before, during, and after the FL exposure period. On the other hand, other studies showed that a larger amount or greater intensity of L2 exposure correlates with more perceptual drift, suggesting that, if proceeding linearly, drift may not be observable at an incipient stage of L2 or FL exposure. For example, in the study of Gorba (2019), the greatest difference between L2 learners and monolingual controls was found in the most proficient and immersed learner group. Along similar lines, Sypiańska and Cal (2022) showed that L1 Polish–L2 English learners of Spanish in a non-immersive classroom context displayed less drift in a vowel discrimination task than learners of the same linguistic background living in Spain. One way of reconciling these conflicting findings is to posit that the trajectory of CLI reflected in drift is not linear but U-shaped: the early occurrence of drift in some studies may represent only an initial perturbation of the L1 phonetics due to L2/FL exposure, which subsides but is gradually followed by a more durable increase in L2 influence on the L1 (i.e., phonetic attrition) as L2 exposure continues.[2] Setting aside the later stages of CLI, the present study assumes that initial FL exposure can indeed precipitate perceptual drift, and investigates how quickly, and under what exposure circumstances, perceptual drift first arises.

2.2. The Present Study

The present study investigated perceptual drift in L1 English listeners exposed to Tagalog as a FL for the first time. The study addressed five research questions pertaining to perceptual drift:

Q1. What is the smallest amount of FL exposure leading to detectable perceptual drift?
Q2. Under what FL exposure circumstances does perceptual drift first arise? That is, what is the weakest condition of FL exposure leading to early perceptual drift?
Q3. Is early perceptual drift assimilatory or dissimilatory with respect to the FL?
Q4. How durable is perceptual drift following recent FL exposure?
Q5. How generalizable is perceptual drift beyond the specific details of FL exposure?

For the purposes of addressing Q1–Q5, we focused on perception of VOT—the primary feature distinguishing voiced and voiceless plosives in languages with voicing contrasts—because it is one of the most widely studied acoustic features in the literature on drift (e.g., Chang 2012; Tice and Woodley 2012), thus facilitating comparisons to previous findings. VOT is defined as the duration in milliseconds between the release burst of a plosive and the first emergence of periodicity in the waveform indicative of voicing (i.e., time at voicing onset minus time at release, meaning that VOT can be negative in cases where the voicing onset precedes the burst). As alluded to above, languages differ in how they carve up the VOT space: "true voicing" languages such as Tagalog (Kang et al. 2016) contrast plosives via negative vs. short-lag VOT, whereas "aspirating" languages such as English (Lisker and Abramson 1964) contrast plosives via short-lag vs. long-lag VOT. Voiced plosives in English, however, are also variably realized with negative VOT in different contexts, and voiceless plosives have shorter VOT when in unstressed syllables (Lisker and Abramson 1967); therefore, the "aspirating" distinction in English is prototypically observed in word-initial and pre-tonic contexts. In such contexts, short-lag VOTs would typically be perceived as voiceless in "true voicing" languages, but as voiced in "aspirating" languages. For L1 speakers of an "aspirating" language, this disparity could result in perceptual errors during L2 acquisition of a "true voicing" language, until short-lag VOTs in word-initial and pre-tonic contexts are reassociated to a phonologically voiceless category; the case of L1 English and FL Tagalog examined in the present study allows us to see whether this occurs.

We approached each of Q1–Q5 with a specific hypothesis. In regard to Q1, we hypothesized that very little FL exposure would be required to observe perceptual drift (H1). The logic behind this hypothesis is based on the proposal of the Speech Learning Model (Flege 1995) for the mechanism underlying bidirectional CLI in L2 learners—equivalence classification of similar L1 and L2 sounds—which should, in principle, apply at the onset of L2 exposure (Chang 2012) and potentially at the onset of FL listening as well. Together with a possible novelty effect augmenting CLI from a relatively unfamiliar L2 (Chang 2013), early-onset equivalence classification sets the stage for initial FL exposure to have a powerful effect on L1 perception. In regard to Q2, we hypothesized that only focused attention to the FL, not mere ambient FL exposure, would lead to perceptual drift (H2), because of previous findings suggesting that, at least once an L2 has become familiar, ambient exposure is not sufficient to promote continued L2 influence on L1 VOT (see Chang 2019a, pp. 104–5). In regard to Q3, we hypothesized that perceptual drift would be assimilatory (H3) on the basis of the findings of the previous study of VOT perception most comparable to the present one (i.e., Tice and Woodley 2012). Thus, we expected to see L1 English listeners perceiving more L1 plosive tokens as voiceless following FL exposure, due to a shift of their L1 voicing categories towards shorter VOT values. In regard to Q4, we hypothesized that, when perceptual drift occurred, it would last between consecutive FL exposures spaced several hours apart (H4) because of both the possible novelty effect mentioned above and the general effect of recency evident in previous studies of drift (e.g., Sancier and Fowler 1997). Finally, in regard to Q5, we hypothesized that, rather than being limited to L1 sounds corresponding to sounds in FL exposure, perceptual drift would occur in a generalizing manner (H5). The logic behind this hypothesis is based on the generalizing patterns reported in studies of drift in production (Chang 2012), studies of spontaneous imitation (Goldinger 1998; Nielsen 2011), and studies of selective adaptation in perception (Eimas et al. 1973; Eimas and Corbit 1973; Cooper 1974).

To test these hypotheses, we conducted a multi-session longitudinal study examining L1 English learners' perception of English VOT over the course of initial laboratory exposure to Tagalog. This study included multiple exposure sessions, multiple exposure conditions eliciting different degrees of attention to the FL speech stimuli, and a pre-test/post-test design examining the effect of a given FL exposure with no delay and with a delay of several hours. In addition, the study design included a disparity between the FL sounds that were included in the FL exposure sessions and the L1 sounds that were tested, allowing us to examine the scope of perceptual drift within the L1 phonetic system. We describe these features of the study in more detail below.

3. Materials and Methods

3.1. Participants

There were two eligibility criteria for participation in the study: (1) being a functionally monolingual L1 English speaker, and (2) having no prior exposure to Tagalog. For the purposes of determining eligibility, we considered individuals reporting experience with a non-English language (who comprised the majority of respondents to our call for participants) "functionally monolingual" in English if they said they would not consider themselves an advanced speaker of the other language.

A total of 65 participants were enrolled in the study, including students at Boston University (45) and residents of the Phoenix, AZ metropolitan area (20). All participants described themselves as speaking General American English; eight reported also speaking a second variety, either Southern American English (4), African American English (3), or Midwestern English (1). Most participants reported some experience with languages other than English. In particular, 52 participants reported either past (43) or current experience (9) in Spanish, French, or Portuguese, languages described as displaying a "true voicing" VOT contrast (see Section 2.2). Thus, it is possible that prior experience with a "true voicing" L2 may have influenced drift outcomes in this study, though this possibility is mitigated by the fact that the highest level of L2 knowledge reported by any participant was intermediate. Thirteen participants were ultimately removed from the study because they were unable to finish it (12) or their data were lost (1). This left a final participant sample of 52 (M_{age} = 26, range 18–73; 15 male, 35 female, 2 other).

Participants were split across four task conditions (see Section 3.2.2): crosslinguistic mapping (N = 15; M_{age} = 23; 4 male, 11 female), emotion identification (N = 11; M_{age} = 27; 3 male, 8 female), unrelated task (N = 9; M_{age} = 22; 1 male, 7 female, 1 other), and unrelated task with no FL exposure, a control condition (N = 17; M_{age} = 29; 7 male, 9 female, 1 other). Similar in gender distribution, these groups also did not significantly differ in age [F(3, 43) = 0.695, p = 0.560]. Furthermore, they were similar in terms of the majority of the group having prior experience with a "true voicing" language (i.e., only 2–4 participants in each group lacked such experience). Thus, we assume that any differences observed between conditions are attributable to the conditions themselves, as opposed to uncontrolled demographic differences between the participants assigned to these conditions.

3.2. Procedure and Stimuli

3.2.1. Study Design

Participation in the study took place over a period of five consecutive days, with two sessions a day (one in the morning and one in the evening) for 10 sessions overall. Consecutive sessions were separated by at least 6 hours, and also spread across different days, to approximate the methodology of Gong et al. (2016) and to allow for sleep consolidation, which has been shown to facilitate perceptual learning (Fenn et al. 2003). Participants completed all experiments in a quiet room using studio-quality binaural headphones.

Each session took about 10–15 min to complete, and comprised three experiments: an exposure task (involving FL exposure in most task conditions) and two L1 identification tasks (pre-test and post-test), one preceding and one following the exposure task. As mentioned above, there were four between-participant task conditions, to which participants were assigned randomly—three involving FL exposure (crosslinguistic mapping, emotion identification, unrelated task) and one not involving FL exposure (unrelated task with no FL exposure, which served as a control condition). The pre-test/post-test design allowed us to assess the durability of drift effects (i.e., whether or not they would be sustained in the interim between consecutive exposures). All experiments were built and administered in OpenSesame 3.3.12 (Mathôt et al. 2012). The overall study design is depicted in Figure 1.

3.2.2. Task Conditions

Participants were assigned to one of four conditions, which were intended to stimulate varying degrees of attention to the plosive contrast in the FL (Tagalog). Tagalog was chosen as the FL because of its "true voicing" plosive contrast, and because prospective participants were unlikely to have experience with it. The exposure task in all conditions involved responses made through a keyboard press and took about 8–10 min. At the recruitment stage, participants were told that the study involved "listening to words and providing responses". The exposure language (Tagalog) was introduced by name in the FL exposure conditions before the exposure interludes and tasks, but no details were given on the relation of the exposure tasks to the L1 tests.

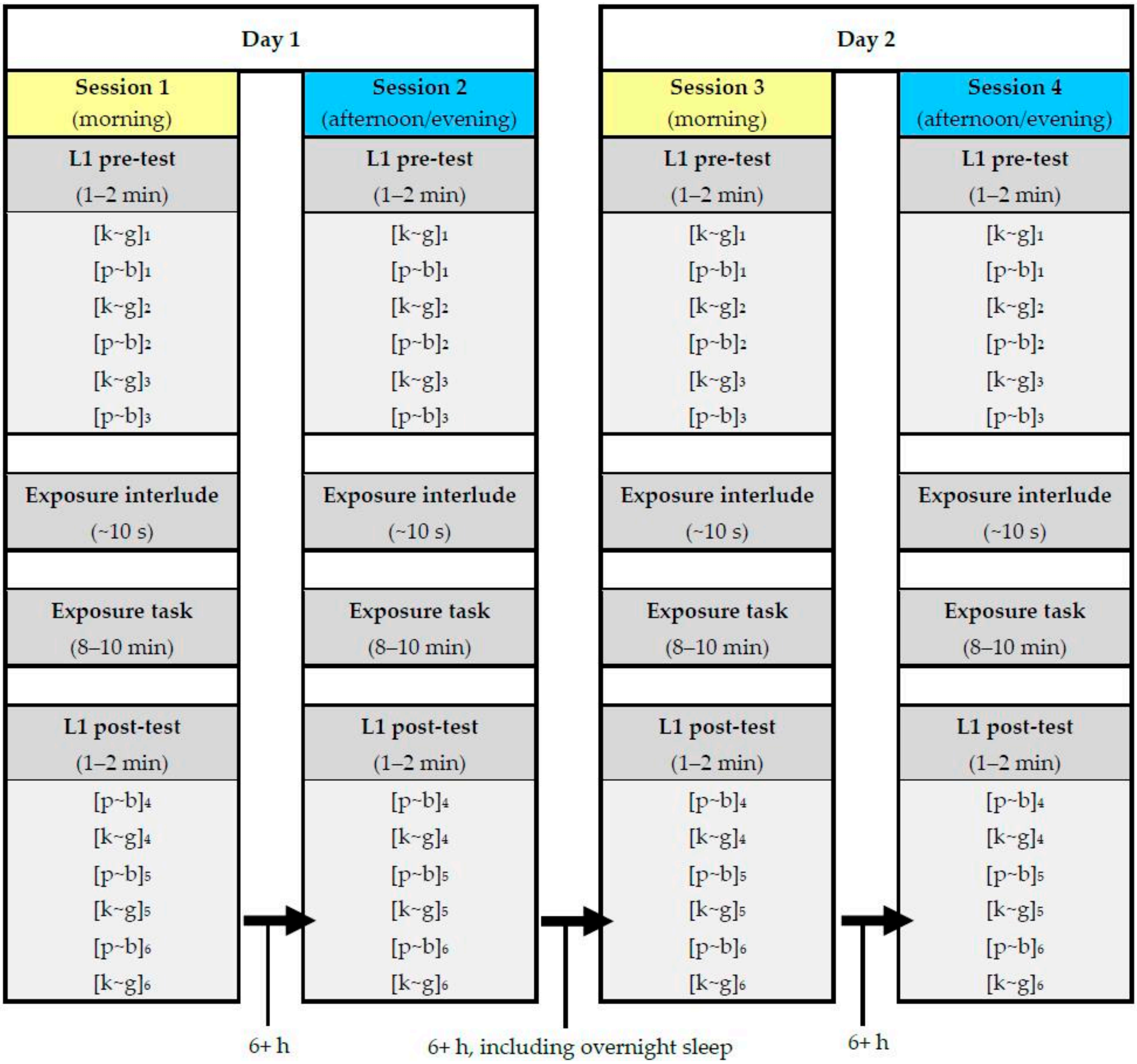

Figure 1. Schematic of the overall study design, showing the first two days of participation (of five total). Acoustically distinct VOT continua for a given place of articulation are denoted by different subscripts.

In condition 1 (crosslinguistic mapping), participants performed a crosslinguistic mapping task whereby they listened to plosive-initial Tagalog words and judged whether the initial plosive was voiced or voiceless in a two-alternative forced-choice (2AFC) identification paradigm (see Figure 2). By having participants make explicit linguistic judgments on the FL stimuli, this condition was designed to direct maximal attention to the phonetic details of the FL speech. Participants were instructed to answer as accurately as possible, and feedback was provided in two ways to promote improvement in the task as well as the association of short-lag VOT tokens with a phonologically voiceless category. First, a strident tone was played following an incorrect response. Second, an accuracy score was provided at the end of the exposure task.

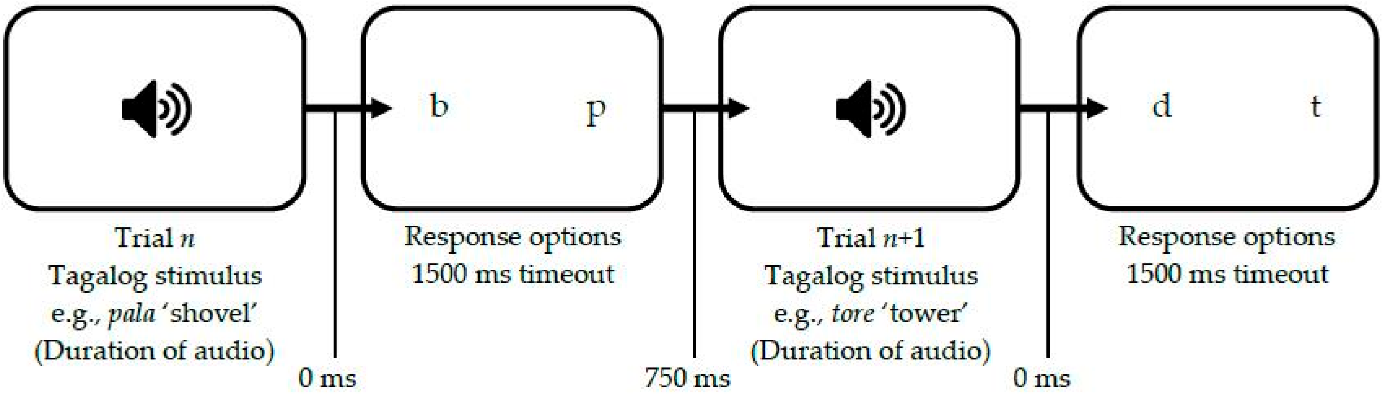

Figure 2. Trial design for condition 1 (crosslinguistic mapping).

In condition 2 (emotion identification), participants performed an emotion identification task whereby they listened to the same plosive-initial Tagalog words as in condition 1 and classified the emotion of the speaker (who varied across trials) as "positive", "negative", or "neutral" (see Figure 3). By having participants make auditory-based, but not explicitly linguistic, judgments on the Tagalog speakers, this condition was designed to direct less attention to the FL speech (in particular, without drawing specific focus to the FL plosive distinction). No feedback was provided in this condition.

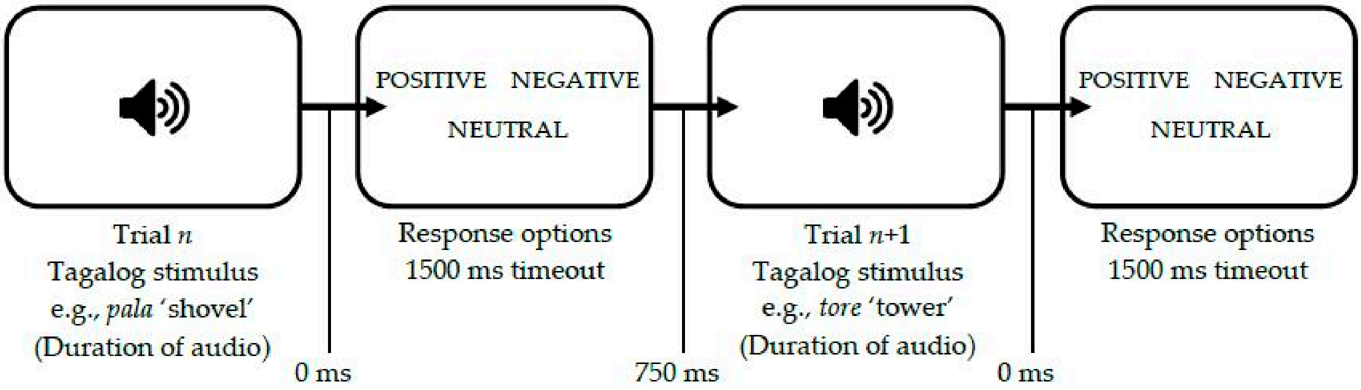

Figure 3. Trial design for condition 2 (emotion identification).

In condition 3a (unrelated task with interleaved FL exposure), participants performed an unrelated math task interleaved with exposure to the same plosive-initial Tagalog words as in conditions 1 and 2 (see Figure 4). A similar distractor task was used in Gordon et al. (1993) for the purpose of drawing attention away from acoustic cues in phoneme identification. On each trial in the math task, participants saw three numbers, each divisible by 10, displayed in a vertical stack, and identified whether they were numerically equidistant ("SAME") or not ("DIFF"). There were 180 trials per task, half "SAME" and half "DIFF". A list of the number combinations used is available in the Supplementary Materials. By having participants make non-linguistic judgments for which the auditory stimuli were irrelevant, this condition was intended to direct minimal attention to the FL speech. As in condition 1, participants were instructed to answer as accurately as possible, and were provided with an accuracy score at the end of the task.

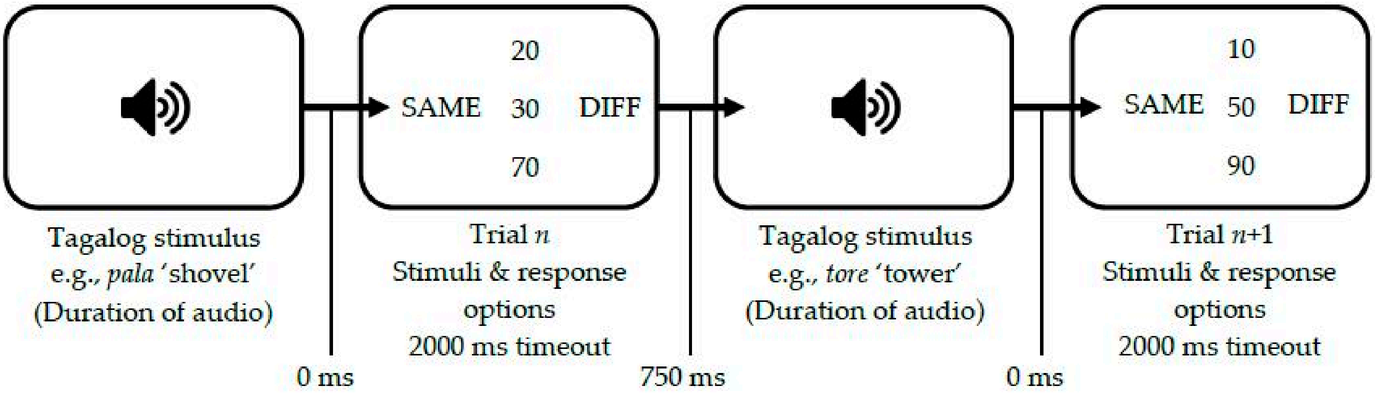

Figure 4. Trial design for condition 3a (unrelated task with interleaved FL exposure).

Because we realized that condition 3a, by virtue of presenting the FL stimuli on their own with no competing stimulus, may have encouraged participants to attend to the FL stimuli to some degree, we later modified this condition to present the FL stimuli simultaneously with the distractor task. In condition 3b (unrelated task with simultaneous FL exposure), participants performed the same math task as in condition 3a, except with the FL exposure proceeding continuously and simultaneously with respect to the numerical stimuli (see Figure 5). By having participants perform the distractor task at the same time as exposure to the auditory stimuli (which, again, were irrelevant to the task), this condition was intended to draw all attention away from the FL speech.

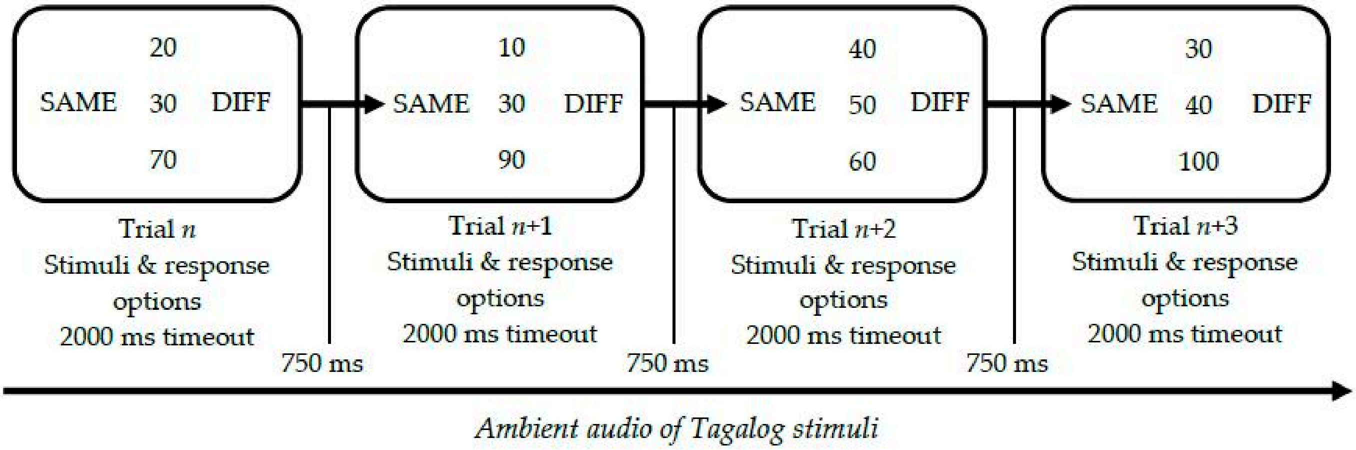

Figure 5. Trial design for condition 3b (unrelated task with simultaneous FL exposure).

The Tagalog stimuli used in conditions 1–3 were produced and distributed across exposure experiments as follows. First, a list of 180 plosive-initial Tagalog items was compiled, consisting of 90 minimal pairs contrasting in the voicing of the initial plosive (e.g., *pala* 'shovel' vs. *bala* 'bucket'). In some cases, the minimal contrast was with a nonce word (e.g., *tilay* 'scald' vs. *dilay*, a nonce word). The place of articulation of the initial plosive was limited to labial or coronal to position us to address Q5 concerning the generalizability of drift (i.e., whether drift would arise in L1 sounds not corresponding to sounds in FL exposure). Further, the initial plosive was made to be the only plosive in the word, meaning that any medial or final consonants were restricted to sonorants or /s/ and were thus prevented from possibly influencing VOT perception. Second, a list of 60 additional Tagalog items without plosives was compiled. These items contained only sonorants, /s/, or /h/ (e.g., *laho* 'eclipse', *yangasngas* 'teeth on edge'). The full list of items is provided in the Supplementary Materials. Both the plosive-initial and plosive-less Tagalog items were recorded by two L1 Tagalog speakers, one male (age 28) and one female (age 24), in a sound-attenuated booth at 44.1 kHz and 16-bit resolution, using an AKG C520 condenser microphone and Zoom H4n recorder. For each plosive-initial item, the speaker produced several tokens in each of four different emotional states (happy, sad, angry, neutral) to provide affectual variance in the tokens for the participants in condition 2 (emotion identification).[3] Each of the ten exposure experiments was then populated with a unique set of 180 tokens of plosive-initial items representing the range of emotional states, such that participants in conditions 1–3 were exposed to 1800 acoustically distinct tokens of Tagalog plosives by the end of the study. This exposure regimen was motivated by the study of Gong et al. (2016), where participants were also exposed to 180 tokens during the FL sessions. As for the plosive-less items, 12 unique tokens of these were played in random order (750 ms apart) as an exposure interlude between the (L1) pre-test and the FL exposure task in each session (see Figure 1). Paired with a white screen and lasting only a few seconds, the exposure interlude was intended to encourage participants into a 'foreign language mode', potentially priming an expectation of FL stimuli and thereby reducing the likelihood of their processing the FL stimuli as the L1. That is, we wanted any observed drift following from FL exposure to be interpretable as due to CLI, as opposed to (unintended) misparsing of the FL as the L1.

In addition to the three FL exposure conditions, an active control condition was included to examine effects of the experimental design (i.e., task effects). In condition 4 (unrelated task with no FL exposure), participants performed the same math task as in condition 3a/3b while receiving continuous auditory exposure to non-linguistic sounds, the sound of ocean waves (see Figure 6). This audio was spliced from a video of ocean wave ambience on YouTube (link in the Supplementary Materials). For the exposure interlude, a smaller clip of ocean waves audio was spliced from the same YouTube video, of around the same duration as the exposure interludes in the other conditions. Although there was no analogous motivation in the control condition to prime a particular language mode, we still included an exposure interlude in this condition for consistency with the FL exposure task conditions.

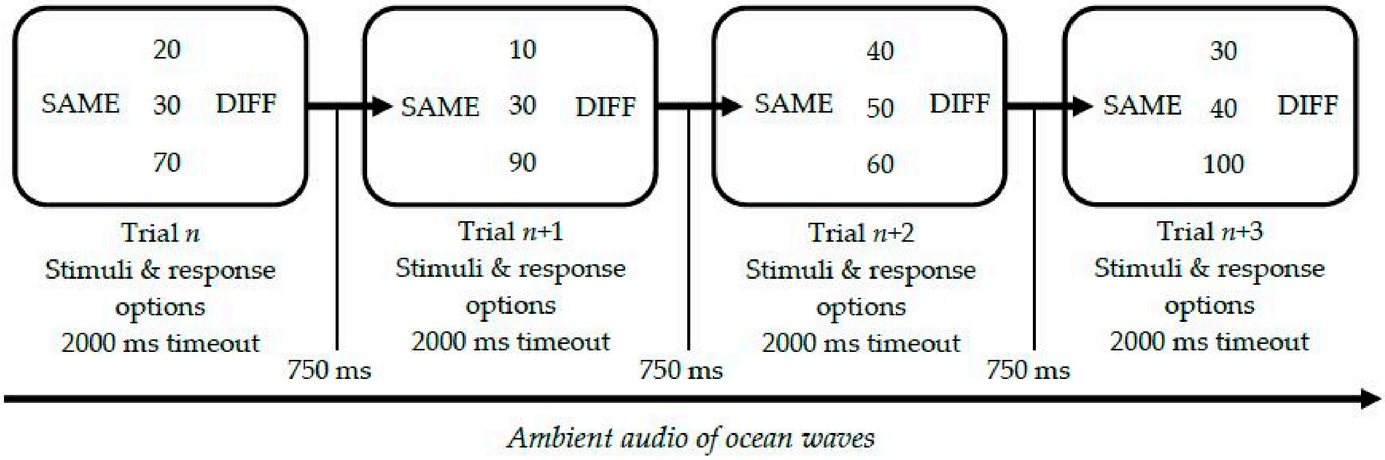

Figure 6. Trial design for condition 4 (unrelated task with no FL exposure; active control).

3.2.3. L1 Identification Experiments

Participants in all conditions completed a 2AFC identification task on L1 English plosive-initial tokens twice within each session, once before the exposure task (pre-test) and once after (post-test), for a total of 20 L1 identification experiments by the end of the study. The English tokens came from VOT continua generated for this study (described in further detail below). On each trial, participants heard an English token and indicated whether the initial sound was voiced or voiceless via the keyboard (see Figure 7). Each L1 identification experiment included six acoustically distinct continua, three bilabial and three velar; these two places of articulation were chosen to address Q5 concerning the generalizability of drift, as bilabial plosives, but not velar plosives, were present in the FL exposure. These continua were played in discrete sequence and associated with a fixed position in either the pre-tests or the post-tests. Each continuum contained 12 VOT steps, and the steps were presented once (in random order) within a continuum; thus, 72 responses were gathered per L1 identification experiment, which took about 1–2 min.

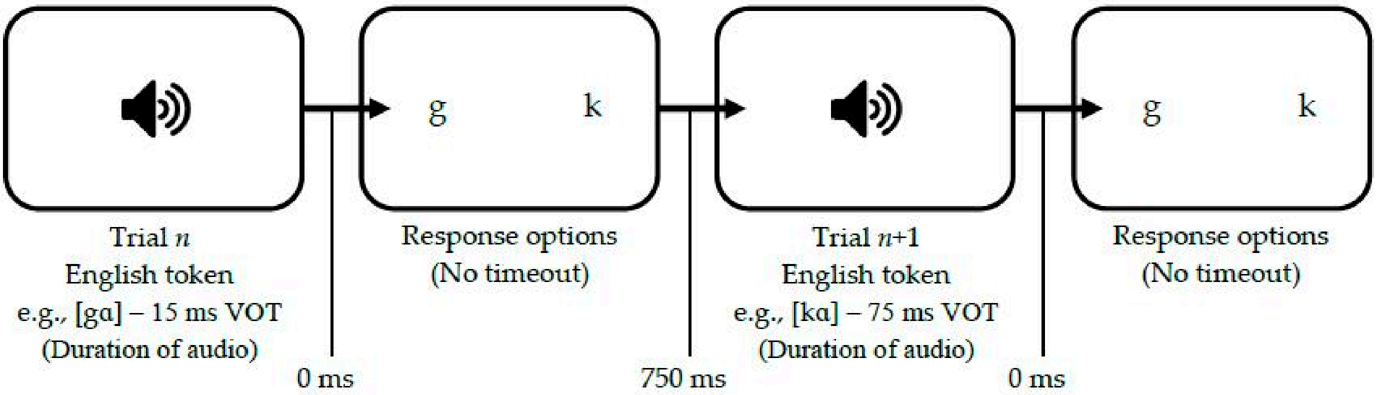

Figure 7. Trial design for the L1 English identification experiments.

The English stimuli used in the L1 identification experiments were produced and distributed across experiments as follows. All stimuli had the shape of a consonant–vowel syllable containing a low back unrounded vowel /ɑ/, as in the study of Tice and Woodley (2012). As above, the place of articulation of the initial plosive was limited to labial or dorsal; thus, there were four target syllables—/bɑ/, /pɑ/, /gɑ/, and /kɑ/. Multiple tokens of each target syllable were recorded by a 25-year-old male L1 English speaker in a quiet room on a smartphone (Samsung Galaxy S10) using its native microphone and voice recording app, at 44.1 kHz and 128 kbps. Due to the technology available to the speaker, the English syllables were recorded in MP3 format, later converted to WAV for use in OpenSesame. Using the script provided in Winn (2020) for the progressive-cutback method of VOT manipulation, 12 VOT continua were synthesized from English syllables in Praat (Boersma and Weenink 2021). Each continuum was based on a unique pair of base tokens (voiced and voiceless) and contained 12 equally-spaced steps. Following the advice of Winn (2020), the bilabial continua went from 3 ms to 60 ms VOT, and the velar continua from 15 ms to 70 ms VOT.

4. Results

As a first step in our analysis, we consulted the response times (RTs) for individual responses in the L1 identification experiments in order to exclude invalid responses that were unlikely to have been made to the audio stimuli. The duration of the stimuli was about 500 ms on average, and RTs were recorded from stimulus offset to the keypress response; therefore, under the assumption that participants started to make their identification judgment near stimulus onset (i.e., soon after hearing the initial stop and vowel onset), all responses would have been registered with ample time to process the stimulus (see, e.g., Bissiri et al. 2011, which uses a threshold RT of 150 ms from stimulus onset for excluding responses made before processing an auditory stimulus). As such, we focused on RTs that were overly long, using a threshold RT of 9500 ms given previous evidence that auditory memory traces last about 10 s (Böttcher-Gandor and Ullsperger 1992; Sams et al. 1993). Responses with RTs longer than 9500 ms were therefore deemed invalid, resulting in 25 identification responses (0.03%) being excluded from statistical analyses. The final dataset submitted to modeling thus consisted of 74,855 of the 74,880 (=52 participants × 20 experiments/participant × 72 trials/ experiment) total L1 identification responses.

Statistical analyses were conducted in R (R Development Core Team 2022) using logistic mixed-effects regression modeling with the 'lmerTest' package (Kuznetsova et al. 2017). Graphs were built with 'ggplot2' (Wickham 2016). We built three main models of responses in the 20 L1 identification experiments, evaluating the statistical significance of main effects and interactions with the Anova function in the 'car' package (Fox and Weisberg 2019); the outputs of the final models are provided in the Appendix A or in Supplementary Materials (model formulas specified in each table caption). In all models, the dependent variable, HeardVoiceless, was a binary, by-trial variable coding whether an L1 plosive token was identified as voiceless (1) or as voiced (0). The main independent variables were Exposure, the number of exposure sessions elapsed (0–10) for the given experiment, and Condition, the task condition the participant was assigned to (1: crosslinguistic mapping, 2: emotion identification, 3: unrelated task, or 4: control). We also tested two additional independent variables: Recency (i.e., whether HeardVoiceless was from a post-test immediately following an exposure task or from a pre-test done several hours after the last exposure) and Place (i.e., whether HeardVoiceless represented judgments on bilabial or velar plosives). All models included a random intercept for Participant to account for individual variability in drift (see the Supplementary Materials for graphs showing individual differences, which we do not discuss here due to space constraints). Random slopes for Exposure by Participant were explored, but did not consistently allow models to converge; further, when random slopes did allow a model to converge, they did not change the results we report below.

The three models were oriented toward addressing one or more of hypotheses H1–H5 (see Section 2.2). Model 1 was designed to detect the overall occurrence and directionality of perceptual drift in each condition (H1–H3) and its generalization (H5). Model 2 was designed to examine the durability of drift (H4) and tested for an effect of Recency (treatment-coded; reference level = pre-test/less recent exposure); therefore, Model 2 was built specifically on responses following 1–9 exposures, because this subset of the data allowed a balanced comparison of post-test responses and the pre-test responses for the next exposure (note that, by definition, there were no post-test responses for zero exposures, and there were no pre-test responses following the tenth exposure, which was the final exposure in the study). Thus, Model 1 included three simple fixed effects—Exposure (as a continuous variable; centered), Condition (treatment-coded; reference level = 4/control), and Place (treatment-coded; reference level = bilabial, i.e., the plosive place of articulation shared between FL exposure and L1 test stimuli)—while Model 2 included the simple fixed effects Exposure, Condition, and Recency (as above). We built four versions of each of these models, rotating the reference level of Condition to observe the simple-effect coefficient for continuous Exposure (Model 1) and that for Recency at the midpoint of the exposure range (Model 2) in each of the four exposure task conditions. Model 3 was designed to explore H1 more specifically—that is, to identify the earliest timepoint of exposure at which perceptual drift was significant overall within each condition. Thus, this model included two simple fixed effects: Condition (as above) and Exposure, which was treatment-coded as a categorical predictor (reference level = 0/baseline), allowing for comparisons of every exposure to the baseline. Because all of the models tested specific hypotheses, they were built using a "hypothesis testing" approach that allowed the above fixed effects to fully interact and thereby show a range of potential outcomes for H1–H5. Therefore, Models 1–3 additionally included all possible interactions among fixed predictors: the two-way Exposure × Condition, Exposure × Place, and Condition × Place interactions and the three-way Exposure × Condition × Place interaction in Model 1, the two-way Exposure × Condition, Exposure × Recency, and Condition × Recency interactions and the three-way Exposure × Condition × Recency interaction in Model 2, and the two-way Exposure × Condition interaction in Model 3.

4.1. H1–H3: Effects of Exposure and Task Condition

As shown in Figure 8, which displays the by-participant mean values of Heard-Voiceless for conditions 1–3 by number of FL exposures, exposure recency, and place of articulation, there was an overall trend for voiceless judgments to become less likely with more FL exposures, regardless of recency or place of articulation.

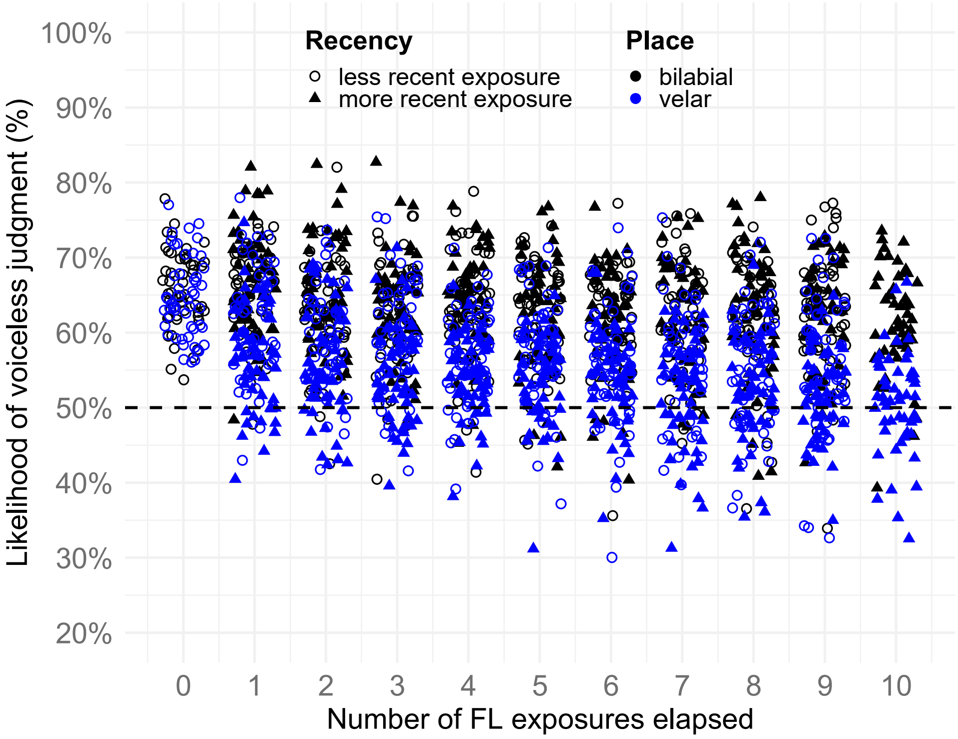

Figure 8. Likelihood of voiceless identification in the FL exposure task conditions, by number of FL exposures, recency of last FL exposure, and place of articulation of the plosive continua. Points (jittered) represent individual participants, averaging over the three plosive continua for the given place of articulation. The dashed line represents the chance value for the binary judgment.

By-condition means are plotted in Figure 9, which shows that the trend for voiceless judgments to become less likely with more exposures is evident in all task conditions, including the control condition (where the exposures were to non-linguistic sounds). However, compared to the control condition, the slope of the decrease is generally steeper in the FL exposure task conditions (conditions 1–3), particularly after exposure 3; this was true of both variants of condition 3 (3a and 3b), which did not noticeably differ from each other and were therefore combined in all visualizations and modeling.

Consistent with Figure 9, the results of Model 1 (see Table A1 in the Appendix A) showed a main effect of Exposure [$\chi^2(1) = 95.747$, $p < 0.001$] and Place [$\chi^2(1) = 431.294$, $p < 0.001$] but not of Condition [$\chi^2(3) = 3.670$, $p = 0.299$]. The Exposure × Place [$\chi^2(1) = 9.345$, $p = 0.002$] and Condition × Place [$\chi^2(3) = 15.068$, $p = 0.002$] interactions were significant, whereas the Exposure × Condition [$\chi^2(3) = 2.113$, $p = 0.549$] and Exposure × Condition × Place [$\chi^2(3) = 0.134$, $p = 0.987$] interactions were not. The Exposure effect was negative and significant for bilabials in the control condition [$\beta = -0.017$, $p = 0.010$], the emotion identification condition [$\beta = -0.024$, $p = 0.003$], and the unrelated task condition [$\beta = -0.019$, $p = 0.033$], and marginal in the crosslinguistic mapping condition [$\beta = -0.012$, $p = 0.092$]. In addition, at the midpoint of the Exposure range (i.e., exposure 5), the likelihood of bilabials being identified as voiceless was higher in the crosslinguistic mapping condition as compared to the control condition [$\beta = 0.160$, $p = 0.023$]. The Place effect showed that velars, as compared to bilabials, were overall less likely to be identified as voiceless in the control condition [$\beta = -0.233$, $p < 0.001$], and even less so at later exposures [$\beta = -0.018$, $p = 0.046$]. Furthermore, the Place effect was significantly enhanced in the crosslinguistic mapping and emotion identification conditions [β's ≤ -0.123, p's ≤ 0.001] and marginally enhanced in the unrelated task condition [$\beta = -0.077$, $p = 0.085$].

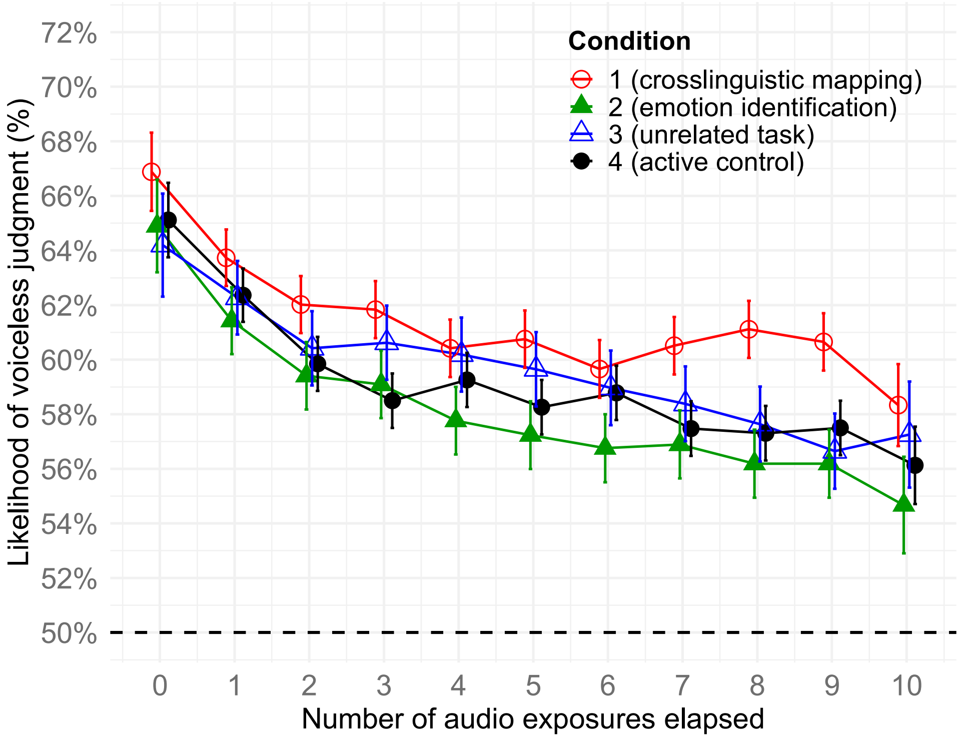

Figure 9. Likelihood of voiceless identification (averaged over bilabials and velars), by number of audio (FL or non-linguistic) exposures and task condition. Points represent the group mean at the given number of exposures in the given task condition. Error bars indicate standard error. The dashed line represents the chance value for the binary judgment.

The results of Model 1 therefore supported H1, but not H2 or H3. As expected, perceptual drift occurred even with the small amount of FL exposure in this study (H1). Contra H2 and H3, however, drift was evident in all task conditions (including the case of ambient FL exposure), and it was dissimilatory, implying a shift of L1 English voicing categories towards longer VOTs. However, given that a similar change in responses was found in the control condition, much of the change observed in the FL exposure task conditions appeared to be due to a task effect (discussed further in Section 5); nevertheless, the occurrence of a stronger Place effect in the FL exposure task conditions was consistent with a distinct effect of FL exposure. We discuss the Place effect, along with interactions with Place, in further detail in Section 4.2 below.

4.2. H5: Generalization of Perceptual Drift (Place Effects)

Perceptual drift by condition and place is plotted in Figure 10, which indicates that perceptual drift occurred in a generalizing fashion, for both bilabials and velars, thus supporting H5. The likelihood of voiceless identification tended to start lower, and consistently remained lower, for velars than for bilabials, reflecting an inherent difference between these places of articulation: velars show longer VOTs and a correspondingly longer VOT crossover point for the voiced–voiceless distinction (around 40–45 ms) than bilabials (around 20–30 ms) (Lisker and Abramson 1970; Christensen 1984; Winn 2020). Surprisingly, however, the likelihood of voiceless identification also tended to decrease more steeply for velars, at least in the control condition.

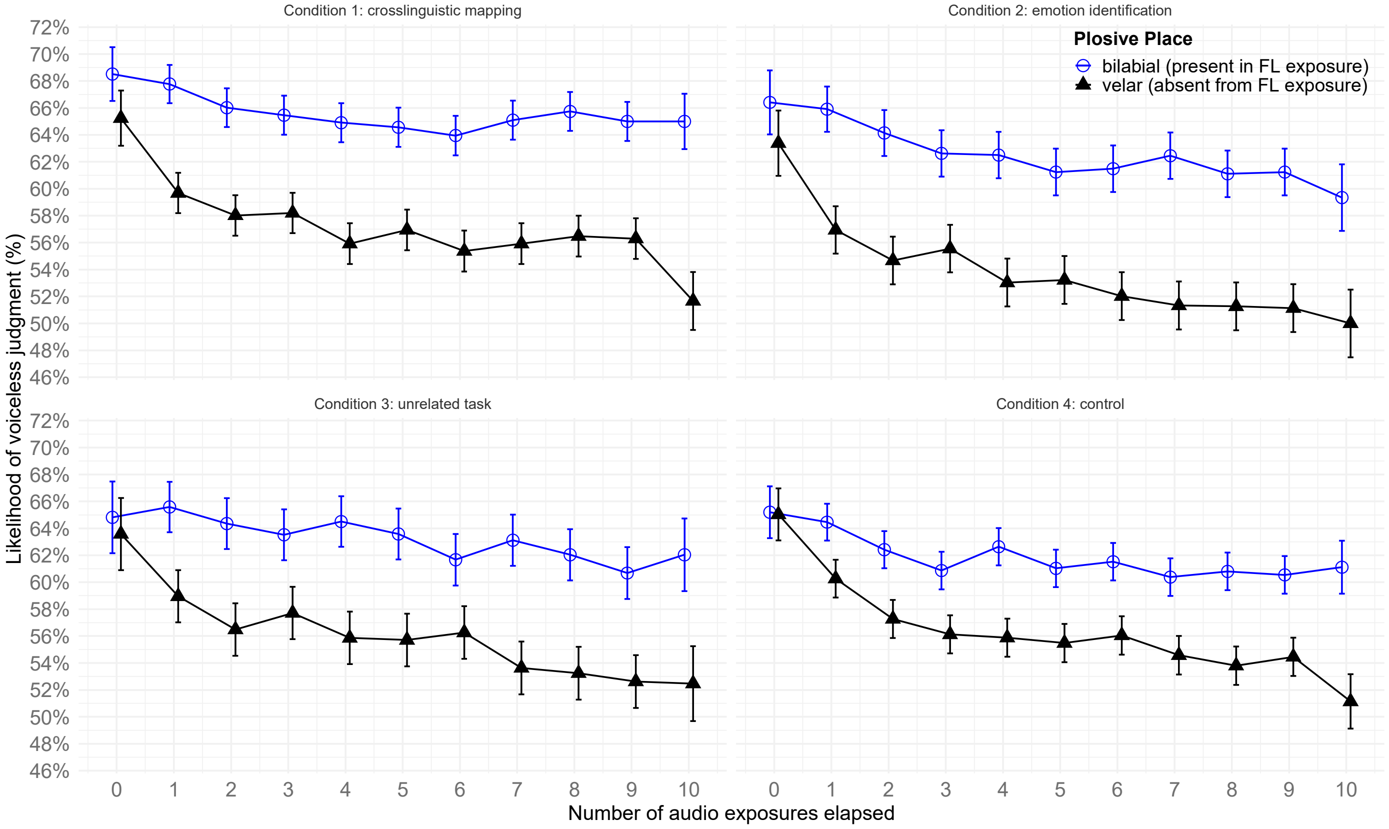

Figure 10. Likelihood of voiceless identification (averaged over post-test and pre-test), by task condition, number of audio exposures, and place of articulation. Error bars indicate standard error.

As mentioned in Section 4.1, Model 1 showed a significant or marginal negative effect of Exposure in every condition, as well as a significant Exposure × Place interaction in the control condition in which velars differed from bilabials in terms of a stronger Exposure effect. In the FL exposure task conditions, however, velars patterned more similarly to bilabials in terms of the Exposure effect. Notably, the Exposure effect for velars did not differ significantly from the Exposure effect for bilabials in any of the FL exposure task conditions: the crosslinguistic mapping condition [$\beta = -0.016$, $z = -1.645$, $p = 0.100$], the emotion identification condition [$\beta = -0.013$, $z = -1.145$, $p = 0.252$], or the unrelated task condition [$\beta = -0.015$, $z = -1.216$, $p = 0.224$].

Given that our hypothesis H5 did not predict a difference between bilabials and velars in the magnitude of drift, the finding of greater drift for velars in the control condition was unexpected. One potential explanation for this finding is inherently greater flexibility of the VOT crossover point for velars (at least in English), which may make them relatively more susceptible to drift. For instance, data on English suggest that, compared to bilabials, velars tend to vary more by vowel context in the perceived VOT crossover point (Nearey and Rochet 1994) and show more variable VOTs (Christensen 1984), which would be consistent with the flexibility account above. Further research is needed both to replicate this disparity in drift between places of articulation and to understand its sources.

Crucially, however, Model 1 indicated that the more negative effect of Exposure for velars vis-a-vis bilabials (i.e., the Exposure × Place interaction, which may be due to intrinsic differences between velars and bilabials as discussed above) was isolated to the control condition, meaning that the pattern of generalization in which velars drift similarly to bilabials was found in all and only the FL exposure task conditions. Recall that the simple Place effect (i.e., the reduction in likelihood of voiceless identification for velars vis-à-vis bilabials) was also enhanced in the FL exposure task conditions compared to the control condition, and significantly so in the crosslinguistic mapping and emotion identification conditions that encouraged more attention to the FL speech stimuli. Together with the disparity between the control condition and the FL exposure task conditions in regard to the Exposure × Place interaction, these results support the view that the perceptual drift observed in the FL exposure task conditions differed in kind from that observed in the control condition. To be specific, whereas the drift in the control condition can only be due to a task effect, we interpret the drift in the FL exposure task conditions as the joint outcome of a task effect and a FL exposure effect.

4.3. H4: Durability of Perceptual Drift (Recency Effects)

Model 2 showed, as Model 1 did, a significant main effect of Exposure [$\chi^2(1) = 43.700$, $p < 0.001$], no main effect of Condition [$\chi^2(3) = 3.580$, $p = 0.311$], and no Exposure × Condition interaction [$\chi^2(3) = 2.591$, $p = 0.459$]; in addition, it showed a significant main effect of Recency [$\chi^2(1) = 11.327$, $p < 0.001$] and a significant Condition × Recency interaction [$\chi^2(3) = 11.118$, $p = 0.011$]. The Exposure × Recency [$\chi^2(1) = 0.151$, $p = 0.698$] and Exposure × Condition × Recency [$\chi^2(3) = 0.726$, $p = 0.867$] interactions were not significant.

The coefficients of Model 2 (Table S1 in the Supplementary Materials) showed a significant negative effect of Recency in the control condition [$\beta = -0.126$, $p < 0.001$], but not in the FL exposure task conditions [$|\beta|$'s < 0.031, $|z|$'s < 1.008, p's > 0.1], supporting H4. In particular, the Condition × Recency interaction coefficients indicated that the negative Recency effect in the control condition was effectively canceled in the crosslinguistic mapping, emotion identification, and unrelated task conditions [β's > 0.096, z's > 2.362, p's < 0.05]. To test the absence of a Recency effect in the FL exposure task conditions further, we built an additional model (Model 2′; see Table S2 in the Supplementary Materials) with the same structure as Model 2 except with the FL exposure task conditions grouped together as the reference level of Condition. Model 2′ confirmed that there was no significant effect of Recency in the FL exposure task conditions even when analyzed together [$\beta = -0.017$, $z = -0.908$, $p = 0.364$]. Thus, we found no evidence that drift regressed toward baseline between FL exposures, meaning that drift was, instead, sustained several hours after the most recent FL exposure.

The non-effect of Recency in the FL exposure task conditions, juxtaposed against the negative Recency effect in the control condition, is visualized in Figure 11, which shows broad similarities between pre-test and post-test values (including a decreasing trajectory for both) and, crucially, no consistency in the direction of the numerical difference between pre-test and post-test values in the FL exposure task conditions. In contrast, in the control condition, the likelihood of voiceless identification was lower for the post-test (more recent audio exposure) than the pre-test (less recent audio exposure) for every number of exposures. These results converge with those concerning Place effects (Section 4.2) in suggesting that perceptual drift in the FL exposure task conditions was qualitatively different from that in the control condition. Whereas the pattern of drift in the control condition was consistent with a task effect that lowers the likelihood of voiceless identification but partially dissipates with temporal distance from the last audio exposure (or button-pressing experience), the pattern of drift in the FL exposure task conditions was not consistent with this type of task effect. This disparity between the control condition and the FL exposure task conditions provides additional evidence that the drift observed in the FL exposure task conditions was not solely due to a task effect. We return to this point in Section 5.

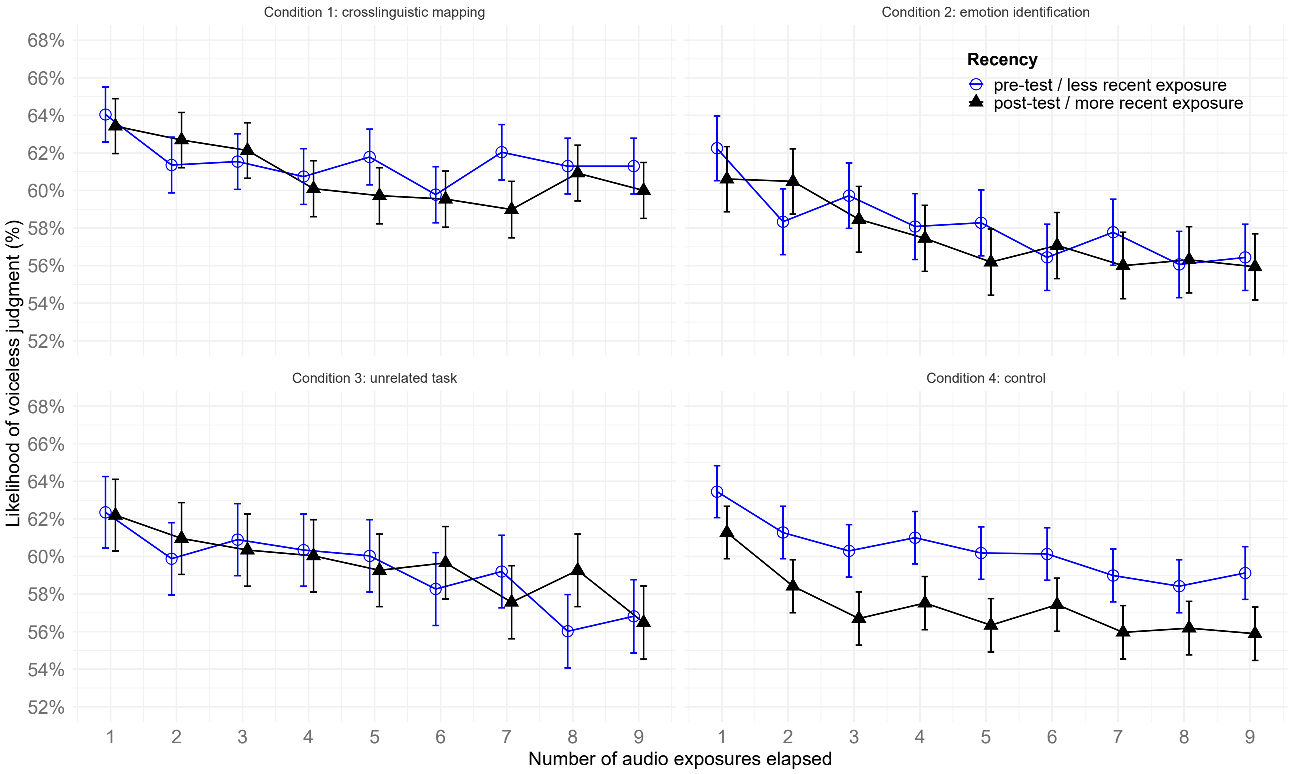

Figure 11. Likelihood of voiceless identification (averaged over bilabials and velars), by task condition, number of audio exposures (1–9), and recency of exposure. Error bars indicate standard error.

4.4. H1: Onset of Perceptual Drift

The results of Model 3 (see Table S3 in the Supplementary Materials) revealed that a significant change from baseline first occurred in most conditions well before the final exposure. Tukey-corrected planned comparisons using the 'emmeans' package (Lenth et al. 2021) showed that the change from baseline became significant following the sixth FL exposure (i.e., about 60 cumulative minutes of FL exposure) in the crosslinguistic mapping condition [*est.* (baseline/exposure 0 − exposure 6) = 0.313, z = 4.043, p = 0.034] and following the eighth FL exposure (80 cumulative minutes) in the emotion identification condition [*est.* = 0.369, z = 4.083, p = 0.029]. Further, change from baseline became significant following the fifth non-linguistic exposure (50 cumulative minutes) in the control condition [*est.* = 0.294, z = 4.050, p = 0.033]. In the unrelated task condition, however, none of the planned comparisons of FL exposure points to the baseline were significant.

The observation of a significant change at an early timepoint in two of three FL exposure task conditions provides additional support for H1; indeed, little FL exposure seems to be required for perceptual drift. However, the fact that the observed change is confounded with a task effect prevents us from identifying precisely the first point at which perceptual drift arises due to FL exposure per se. Thus, we cautiously conclude that perceptual drift can be detected before the accumulation of about 80–100 min of FL exposure (i.e., the final FL exposure in this study); however, we remain agnostic as to whether it may be detectable even earlier than this point.

5. Discussion

The present study investigated the dynamics of perceptual drift in L1 English listeners exposed to Tagalog for the first time. Employing a more frequent L1 testing regimen than in previous research (cf. Gong et al. 2016) with multiple FL exposure task conditions, we found that drift was detectable by the end of the exposure period (i.e., after little FL exposure overall, supporting H1), although it was not possible to pinpoint an earlier onset of drift with confidence. However, drift was not limited to the crosslinguistic mapping condition, the only one requiring L1-FL connections to be made (cf. H2); rather, it was found in all conditions, including a control condition, suggesting that the effect was due partly to an artifact of the study design. In all conditions, the pattern of drift was dissimilatory (cf. H3). Drift, moreover, lasted several hours after a FL exposure (supporting H4) and generalized to a plosive place of articulation that was not present in FL exposure (supporting H5). These findings provide further evidence of the plasticity of the L1 phonetic system (de Leeuw and Celata 2019) and the need to account for research participants' language background and recent language exposure (cf. Chang 2019a).

In regard to the task effect observed in our control condition, which was not expected (cf. Tice and Woodley 2012), we attribute this effect to two aspects of the L1 identification experiments that may have biased participants to shift their responses in the direction of fewer "voiceless" judgments. First, the 12-step L1 VOT continua did not represent prototypical "voiced" and "voiceless" VOTs evenly. As discussed in Section 3.2.3, the continua were constructed with a VOT range of 3–60 ms for bilabials and 15–70 ms for velars, but because the voiced-voiceless VOT boundaries in English typically fall closer to the beginning of these ranges (around 20–30 ms for bilabials and 40–45 ms for velars; Lisker and Abramson 1970; Christensen 1984; Winn 2020), more steps in each continuum favored a voiceless interpretation. At baseline, this asymmetry resulted in asymmetrical pressing of the two response keys in L1 identification (see Figure 8), in contrast to the symmetrical distribution of target (i.e., correct) binary responses in the exposure tasks in conditions 1, 3, and 4 (recall that condition 2 did not involve a binary judgment and also provided no feedback on responses). Second, the L1 identification experiments were completed adjacent to the exposure tasks. Together, the asymmetrical key-pressing in the L1 identification experiments and the juxtaposition of these experiments with the exposure tasks could have made participants gravitate toward symmetry (i.e., a 50–50 split) in their L1 identification responses, even in the control condition.

However, the task effect apparent in the control condition was only partly reflected in the FL exposure task conditions, where, unlike the control condition, *change* in L1 identification responses was not significantly affected by place of articulation or by recency of audio exposure. At the same time, the FL exposure task conditions also showed a larger effect of place on the *level* of L1 identification responses as compared to the control condition, consistent with more generalization from trained bilabials to untrained velars occurring after speech exposure. Together, these findings suggest that in the FL exposure task conditions, there was not only a task effect at work, but also a distinct effect of FL exposure, although it is ultimately unclear how a task effect and a FL exposure effect may have interacted in this study. In contrast to our hypothesis H3 and the vast majority of previously reported cases of perceptual drift, which have been assimilatory (e.g., Tice and Woodley 2012; Lev-Ari and Peperkamp 2013; Dmitrieva 2019; Gorba 2018, 2019; Takahashi 2020), the drift observed in this study was dissimilatory, converging with the results of Sypiańska and Cal (2022) for a long-term immersion context as well as studies reporting dissimilatory drift in production in short-term classroom learning contexts (Huffman and Schuhmann 2015a, 2015b; Huffman et al. 2017; Dmitrieva and Tews 2018; Dmitrieva et al. 2020). According to the (revised) Speech Learning Model (Flege 1995; Flege and Bohn 2021), which explains dissimilation in terms of separate L1 and L2 categories diverging in a shared phonetic space, the dissimilatory pattern would suggest that participants had established separate FL voicing categories even at an incipient stage of FL exposure, which is surprising. Could this have been facilitated by participants' prior L2 exposure to "true voicing" languages such as Spanish, French, and Portuguese (see Section 3.1)? Further research is needed to understand the role that prior exposure to "true voicing" languages, and its possible facilitation of forming distinct FL voicing categories, may have played in the trajectory of perceptual drift observed in this study.

Apart from the confounding of the FL exposure effect with a task effect, which prevents us from addressing our first research question precisely (see Section 4.4), there are some other limitations of the present findings, leaving several directions for future work.

First, to distinguish the FL exposure effect from task effects, future studies could incorporate a number of methodological modifications, such as different modes of response (e.g., transcription), different ranges for the speech continua (e.g., VOT continua extending into negative VOT values), and different presentation frequencies of steps from a given continuum. Second, we cannot guarantee that the FL exposure conditions mitigated participants' attention in the intended way; for example, it is possible that in the unrelated task condition, some degree of attention was paid to the FL speech even though it was irrelevant to the task at hand. This limitation could be addressed in future work by experimenting with different distractor tasks to minimize the likelihood of directed attention to the FL speech. Third, although we originally aimed to recruit monolingual English listeners, so as to ensure that their exposure to Tagalog would be their first extensive exposure to a "true voicing" VOT contrast, this turned out to be very difficult, and our participant sample mostly comprised English listeners who had significant prior exposure to a "true voicing" contrast in another language. Thus, it would be useful in future work to test other populations, including more monolingual-like language users but also proficient bilinguals and multilinguals. It would also be useful to observe drift in other, less familiar phonetic features, given that a "true voicing" plosive contrast may already be familiar to English speakers as an allophonic variant of their L1 "aspirating" contrast (Lisker and Abramson 1967), potentially resulting in easier adaptation to the VOT categories of a "true voicing" FL such as Tagalog without a restructuring of the L1 phonetic space. Finally, our results shed light only on changes in categorization and miss other aspects of change that may occur as part of perceptual drift. Further insight on drift would be provided by using alternative response paradigms, such as gradient judgments of category goodness.

6. Conclusions

The findings of the present study contribute an empirical basis for further work exploring the progression of perceptual drift in listeners exposed to a new language. Previous research on perceptual drift in the context of controlled exposure to an unfamiliar FL is scant, consisting of only one study to our knowledge (Gong et al. 2016). What the present analysis adds to that study, as well as to the broader literature on drift in contexts of initial laboratory exposure and extensive naturalistic exposure to an L2, is the central finding that perceptual drift in identification of L1 laryngeal (voicing) categories can be detected after less than an hour and a half of FL exposure, with little regard for attention to the FL speech. This finding is in line with previous results on L1 consonant reception thresholds (Gong et al. 2016) and on L1 VOT perception during classroom-based L2 exposure (Tice and Woodley 2012). Although further research is needed to replicate the current results, which may be underpowered, with a larger participant sample and to separate the magnitude and timing of the FL exposure effect from task effects, the present findings highlight the promise of future research on perceptual drift, which holds the potential to improve our understanding of the early stages of L1-L2 phonetic interaction and the connection of these early stages to later stages of bilingualism, language attrition, and language (re)learning.

Supplementary Materials: The following materials are available in an OSF repository at https://osf.io/3ph8r/: the language background questionnaire, the Tagalog word lists, the number sets used in conditions 3–4, links to the Praat script used for VOT manipulation (Winn 2020) and the ocean waves audio used in condition 4, the main dataset used for the analysis along with R analysis code, additional model outputs not included in the Appendix A, and graphs depicting individual participant trends.

Author Contributions: Conceptualization, J.K.; methodology, J.K. and C.B.C.; software, J.K.; validation, J.K. and C.B.C.; formal analysis, J.K. and C.B.C.; investigation, J.K.; resources, C.B.C.; data curation, J.K.; writing—original draft preparation, J.K. and C.B.C.; writing—review and editing, J.K. and C.B.C.; visualization, J.K. and C.B.C.; supervision, C.B.C.; project administration, J.K.; funding acquisition, J.K. and C.B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Internal funding from Boston University was received from the Department of Linguistics and the Associate Dean of the Faculty for the Humanities.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Boston University (Protocol 6136X, 31 August 2021).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The full dataset is available in the OSF repository at https://osf.io/pcywm/ (see Supplementary Materials above).

Acknowledgments: The authors gratefully acknowledge helpful comments and feedback from Kate Lindsey, Jon Barnes, two anonymous reviewers, the special issue editors, and audiences at Boston University and the 63rd Annual Meeting of the Psychonomic Society.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Fixed-effect coefficients in Model 1 of the log odds of voiceless identification (HeardVoiceless). Model formula: HeardVoiceless ~ Exposure * Condition * Place + (1 | Participant). Exposure was coded as continuous and centered; Condition as categorical, with four alternate reference levels (a–d); and Place as categorical. Significant effects are bolded.

(**a**) Condition 1 (crosslinguistic mapping) as reference level:

	β	*SE*	*z*-**value**	*p*-**value**
(Intercept)	0.646	0.052	12.533	<0.001
Exposure	−0.012	0.007	−1.685	0.091
Condition: 2	−0.128	0.079	−1.616	0.106
Condition: 3	−0.099	0.084	−1.169	0.242
Condition: 4	−0.160	0.071	−2.270	0.023
Place: velar	**−0.357**	**0.028**	**−12.666**	**<0.001**
Exposure × Condition: 2	−0.012	0.011	−1.132	0.257
Exposure × Condition: 3	−0.007	0.011	−0.636	0.525
Exposure × Condition: 4	−0.005	0.009	−0.508	0.612
Exposure × Place: velar	−0.016	0.010	−1.645	0.100
Condition: 2 × Place: velar	−0.017	0.043	−0.385	0.700
Condition: 3 × Place: velar	0.047	0.046	1.020	0.308
Condition: 4 × Place: velar	**0.123**	**0.038**	**3.209**	**0.001**
Exposure × Condition: 2 × Place: velar	0.003	0.015	0.210	0.833
Exposure × Condition: 3 × Place: velar	0.001	0.016	0.053	0.958
Exposure × Condition: 4 × Place: velar	−0.002	0.013	−0.157	0.875

(**b**) Condition 2 (emotion identification) as reference level:

	β	*SE*	*z*-**value**	*p*-**value**
(Intercept)	0.518	0.060	8.637	<0.001
Exposure	**−0.024**	**0.008**	**−2.964**	**0.003**
Condition: 1	0.128	0.079	1.616	0.106
Condition: 3	0.030	0.089	0.331	0.741
Condition: 4	−0.032	0.077	−0.422	0.673
Place: velar	**−0.373**	**0.033**	**−11.465**	**<0.001**
Exposure × Condition: 1	0.012	0.011	1.130	0.258
Exposure × Condition: 3	0.005	0.012	0.403	0.687
Exposure × Condition: 4	0.007	0.010	0.701	0.483
Exposure × Place: velar	−0.013	0.011	−1.145	0.252
Condition: 1 × Place: velar	0.017	0.043	0.385	0.700
Condition: 3 × Place: velar	0.063	0.049	1.301	0.193
Condition: 4 × Place: velar	**0.140**	**0.042**	**3.350**	**<0.001**
Exposure × Condition: 1 × Place: velar	−0.003	0.015	−0.209	0.834
Exposure × Condition: 3 × Place: velar	−0.002	0.017	−0.136	0.892
Exposure × Condition: 4 × Place: velar	−0.005	0.014	−0.361	0.718

Table A1. *Cont.*

(**c**) Condition 3 (unrelated task) as reference level:

	β	*SE*	*z*-value	*p*-value
(Intercept)	0.547	0.066	8.288	<0.001
Exposure	**−0.019**	**0.009**	**−2.130**	**0.033**
Condition: 1	0.098	0.084	1.173	0.241
Condition: 2	−0.030	0.089	−0.331	0.740
Condition: 4	−0.062	0.082	−0.760	0.447
Place: velar	**−0.310**	**0.036**	**−8.601**	**<0.001**
Exposure × Condition: 1	0.007	0.011	0.636	0.525
Exposure × Condition: 2	−0.005	0.012	−0.404	0.686
Exposure × Condition: 4	0.002	0.011	0.217	0.829
Exposure × Place: velar	−0.015	0.012	−1.216	0.224
Condition: 1 × Place: velar	−0.047	0.046	−1.020	0.308
Condition: 2 × Place: velar	−0.063	0.049	−1.303	0.193
Condition: 4 × Place: velar	0.077	0.045	1.724	0.085
Exposure × Condition: 1 × Place: velar	−0.001	0.016	−0.052	0.958
Exposure × Condition: 2 × Place: velar	0.002	0.017	0.136	0.892
Exposure × Condition: 4 × Place: velar	−0.003	0.015	−0.190	0.849

(**d**) Condition 4 (active control) as reference level:

	β	*SE*	*z*-value	*p*-value
(Intercept)	0.485	0.048	10.068	<0.001
Exposure	**−0.017**	**0.006**	**−2.577**	**0.01**
Condition: 1	**0.160**	**0.071**	**2.272**	**0.023**
Condition: 2	0.033	0.077	0.423	0.673
Condition: 3	0.062	0.082	0.757	0.449
Place: velar	**−0.233**	**0.026**	**−8.897**	**<0.001**
Exposure × Condition: 1	0.005	0.009	0.507	0.612
Exposure × Condition: 2	−0.007	0.010	−0.702	0.483
Exposure × Condition: 3	−0.002	0.011	−0.217	0.829
Exposure × Place: velar	**−0.018**	**0.009**	**−1.997**	**0.046**
Condition: 1 × Place: velar	**−0.123**	**0.038**	**−3.209**	**0.001**
Condition: 2 × Place: velar	**−0.140**	**0.042**	**−3.351**	**<0.001**
Condition: 3 × Place: velar	−0.077	0.045	−1.722	0.085
Exposure × Condition: 1 × Place: velar	0.002	0.013	0.158	0.875
Exposure × Condition: 2 × Place: velar	0.005	0.014	0.361	0.718
Exposure × Condition: 3 × Place: velar	0.003	0.015	0.19	0.849

Notes

[1] The duration of L2 sessions in Gong et al. (2016) was provided via personal correspondence with Jian Gong (September 2021).

[2] This would mirror other U-shaped patterns in acquisition, such as the development of the English past tense in both L1 learners (Jackson and Cottrell 1997) and late sequential L2 learners (Williams et al. 2022).

[3] An anonymous reviewer asked whether the emotionality of the speech tokens had been validated (e.g., by having a separate group of listeners rate the emotion they thought a given token conveyed). In short, we did not separately measure the perceptibility of the emotional states produced by the Tagalog speakers, because this was tangential to our research questions. Although the speakers were instructed to pronounce each word four times, with varying affect (happy, then angry, then sad, then neutral), this instruction was intended simply to produce a range of different speech qualities for the emotion identification condition, and the response options in this condition were not made to match the list of target emotional states. Speakers may have been inconsistent in how they interpreted this instruction, and the details of each speaker's specific performance of the target emotional states—including, for example, whether their 'happy' voice may have been 'misclassified' as 'negative'—were not our concern.

References

Bissiri, Maria Paola, Maria Luisa Lecumberri, Martin Cooke, and Jan Volín. 2011. The role of word-initial glottal stops in recognizing English words. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*. Edited by Piero Cosi, Renato De Mori, Giuseppe Di Fabbrizio and Roberto Pieraccini. Bonn: ISCA Archive, pp. 165–68.

Boersma, Paul, and David Weenink. 2021. *Praat: Doing Phonetics by Computer*, Version 6.2.09. Available online: http://www.praat.org/ (accessed on 2 March 2023).

Böttcher-Gandor, C., and Peter Ullsperger. 1992. Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval. *Psychophysiology* 29: 546–50. [CrossRef] [PubMed]

Cabrelli, Jennifer, Alicia Luque, and Irene Finestrat-Martínez. 2019. Influence of L2 English phonotactics in L1 Brazilian Portuguese illusory vowel perception. *Journal of Phonetics* 73: 55–69. [CrossRef]

Chang, Charles B. 2012. Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics* 40: 249–68. [CrossRef]

Chang, Charles B. 2013. A novelty effect in phonetic drift of the native language. *Journal of Phonetics* 41: 520–33. [CrossRef]

Chang, Charles B. 2019a. Language change and linguistic inquiry in a world of multicompetence: Sustained phonetic drift and its implications for behavioral linguistic research. *Journal of Phonetics* 74: 96–113. [CrossRef]

Chang, Charles B. 2019b. Phonetic drift. In *The Oxford Handbook of Language Attrition*. Edited by Monika S. Schmid and Barbara Köpke. Oxford: Oxford University Press, pp. 191–203. [CrossRef]

Christensen, Jens B. 1984. The perception of voice onset time: A cross language study of American English and Danish. *Annual Report of the Institute of Phonetics University of Copenhagen* 18: 163–84. [CrossRef]

Cooper, William E. 1974. Selective adaptation for acoustic cues of voicing in initial stops. *Journal of Phonetics* 2: 303–13. [CrossRef]

de Leeuw, Esther, and Charles B. Chang. 2023. Phonetic and phonological L1 attrition and drift in bilingual speech. In *The Cambridge Handbook of Bilingual Phonetics and Phonology*. Edited by Mark Amengual. Cambridge: Cambridge University Press, under review.

de Leeuw, Esther, and Chiara Celata. 2019. Plasticity of native phonetic and phonological domains in the context of bilingualism. *Journal of Phonetics* 75: 88–93. [CrossRef]

Dmitrieva, Olga. 2019. Transferring perceptual cue-weighting from second language into first language: Cues to voicing in Russian speakers of English. *Journal of Phonetics* 73: 128–43. [CrossRef]

Dmitrieva, Olga, and Alexis N. Tews. 2018. First language phonetic drift in second language instructional environment. *Journal of the Acoustical Society of America* 143: 1952. [CrossRef]

Dmitrieva, Olga, Allard Jongman, and Joan A. Sereno. 2020. The effect of instructed second language learning on the acoustic properties of first language speech. *Languages* 5: 44. [CrossRef]

Eimas, Peter D., and John D. Corbit. 1973. Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4: 99–109. [CrossRef]

Eimas, Peter D., William E. Cooper, and John D. Corbit. 1973. Some properties of linguistic feature detectors. *Perception & Psychophysics* 13: 247–52. [CrossRef]

Fenn, Kimberly M., Howard C. Nusbaum, and Daniel Margoliash. 2003. Consolidation during sleep of perceptual learning of spoken language. *Nature* 425: 614–16. [CrossRef] [PubMed]

Flege, James Emil. 1987. The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics* 15: 47–65. [CrossRef]

Flege, James Emil. 1995. Second language speech learning: Theory, findings, and problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winfred Strange. Timonium: York Press, pp. 223–72.

Flege, James Emil, and Ocke-Schwen Bohn. 2021. The Revised Speech Learning Model (SLM-r). In *Second Language Speech Learning: Theoretical and Empirical Progress*. Edited by Ratree Wayland. Cambridge: Cambridge University Press, pp. 3–83. [CrossRef]

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks: Sage Publications.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251–79. [CrossRef]

Gong, Jian, María Luisa García Lecumberri, and Martin Cooke. 2016. Can intensive exposure to foreign language sounds affect the perception of native sounds? In *Proceedings of Interspeech 2016*. Edited by Nelson Morgan, Panayiotis Georgiou, Shrikanth S. Narayanan and Florian Metze. Adelaide: Casual Productions, pp. 883–87. [CrossRef]

Gorba, Celia. 2018. The effect of L2 experience on the categorization of native and non-native stops by Spanish learners of English. In *Persistence and Resistance in English Studies: New Research*. Edited by Sara Martin, Owen David and Elisabet Pladevall-Ballester. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 164–73.

Gorba, Celia. 2019. Bidirectional influence on L1 Spanish and L2 English stop perception: The role of L2 experience. *Journal of the Acoustical Society of America* 145: EL587–92. [CrossRef]

Gordon, Peter C., Jennifer L. Eberhardt, and Jay G. Rueckl. 1993. Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology* 25: 1–42. [CrossRef]

Huffman, Marie K., and Katharina S. Schuhmann. 2015a. Effect of early L2 learning on L1 stop voicing. *Proceedings of Meetings on Acoustics* 23: 060007. [CrossRef]

Huffman, Marie K., and Katharina S. Schuhmann. 2015b. Individual differences in phonetic drift by English-speaking learners of Spanish. *Journal of the Acoustical Society of America* 137: 2268. [CrossRef]

Huffman, Marie K., Katharina S. Schuhmann, Kayla Keller, and Chanda Chen. 2017. Interaction of drift and distinctiveness in L1 English-L2 Japanese learners. *Journal of the Acoustical Society of America* 141: 3517. [CrossRef]

Jackson, Dan, and Garrison W. Cottrell. 1997. Attention and U-shaped learning in the acquisition of the past tense. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Edited by Michael G. Shafto and Pat Langley. Mahwah: Lawrence Erlbaum Associates, pp. 325–30.

Kang, Yoonjung, Sneha George, and Rachel Soo. 2016. Cross-language influence in the stop voicing contrast in heritage Tagalog. *Heritage Language Journal* 13: 184–218. [CrossRef]

Kartushina, Natalia, Alexis Hervais-Adelman, Ulrich Hans Frauenfelder, and Narly Golestani. 2016. Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics* 57: 21–39. [CrossRef]

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82: 1–26. [CrossRef]

Lein, Tatjana, Tanja Kupisch, and Joost van de Weijer. 2016. Voice onset time and global foreign accent in German–French simultaneous bilinguals during adulthood. *International Journal of Bilingualism* 20: 732–49. [CrossRef]

Lenth, Russell V., Paul Buerkner, Maxime Herve, Jonathon Love, Hannes Riebl, and Henrik Singmann. 2021. emmeans: Estimated marginal means, aka least-squares means [R package], Version 1.7.0. Available online: https://cran.r-project.org/web/packages/emmeans/index.html (accessed on 2 March 2023).

Lev-Ari, Shiri, and Sharon Peperkamp. 2013. Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics* 41: 320–31. [CrossRef]

Lisker, Leigh, and Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20: 384–422. [CrossRef]

Lisker, Leigh, and Arthur S. Abramson. 1967. Some effects of context on voice onset time in English stops. *Language and Speech* 10: 1–28. [CrossRef] [PubMed]

Lisker, Leigh, and Arthur S. Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*. Edited by Bohuslav Hála, Milan Romportl and Přmysl Janota. Prague: Academia, pp. 563–67.

Major, Roy C. 1992. Losing English as a first language. *The Modern Language Journal* 76: 190–208. [CrossRef]

Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44: 314–24. [CrossRef]

Mora, Joan C., and Marianna Nadeu. 2012. L2 effects on the perception and production of a native vowel contrast in early bilinguals. *International Journal of Bilingualism* 16: 484–500. [CrossRef]

Namjoshi, Jui, Annie Tremblay, Elsa Spinelli, Mirjam Broersma, Maria Teresa Martínez-García, Katrina Connell, Taehong Cho, and Sahyang Kim. 2015. Speech segmentation is adaptive even in adulthood: Role of the linguistic environment. In *Proceedings of the 18th International Conference on Phonetic Sciences*. Edited by The Scottish Consortium for ICPhS 2015. Glasgow: University of Glasgow, p. 0676.

Nearey, Terrance M., and Bernard L. Rochet. 1994. Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association* 24: 1–18. [CrossRef]

Nielsen, Kuniko. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39: 132–42. [CrossRef]

Plomp, Reiner, and A. M. Mimpen. 1979. Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology* 18: 43–52. [CrossRef]

R Development Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sams, Mikko, Riitta Hari, Josi Rif, and Jukka Knuutila. 1993. The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *Journal of Cognitive Neuroscience* 5: 363–70. [CrossRef]

Sancier, Michele L., and Carol A. Fowler. 1997. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics* 27: 421–36. [CrossRef]

Sypiańska, Jolanta, and Zuzanna Cal. 2022. Perceptual drift in L1 phonetic categories in multilinguals. In *Theoretical and Practical Developments in English Speech Assessment, Research, and Training: Studies in Honour of Ewa Waniek-Klimczak*. Edited by Veronica G. Sardegna and Anna Jarosz. Cham: Springer, pp. 299–313. [CrossRef]

Takahashi, Chikako. 2020. The Interaction Between L1 and L2 Phonetic Learning. Ph.D. thesis, Stony Brook University, Stony Brook, NY, USA.

Tice, Marisa, and Melinda Woodley. 2012. Paguettes and bastries: Novice French learners show shifts in native phoneme boundaries. *UC Berkeley Phonology Lab Annual Report* 8: 72–75. [CrossRef]

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. Available online: https://ggplot2.tidyverse.org (accessed on 12 May 2022).

Williams, Stefan, Pedro Guijarro-Fuentes, and Mila Vulchanova. 2022. U-shaped trajectories in an L2 context: Evidence from the acquisition of verb morphology. *Vigo International Journal of Applied Linguistics* 19: 223–66. [CrossRef]

Winn, Matthew B. 2020. Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *Journal of the Acoustical Society of America* 147: 852–66. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.