NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF PHILOSOPHY AND HISTORY OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS
DEPARTMENT OF NURSING
DEPARTMENT OF PHILOSOPHY, PEDAGOGY AND PSYCHOLOGY
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
DEPARTMENT OF INFORMATICS

"The role of discriminant reinforcement in task-irrelevant perceptual learning of acoustic-phonetic categories"

by

Theodora I. Dimopoulou

Student registration number: 09M03

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Basic and Applied
Cognitive Science at the Department of Philosophy and
History of Science

Athens
September 2014

This thesis by Theodora Dimopoulou is accepted in its present form by the supervisory committee as satisfying the thesis requirement for the degree of Master of Science.

Athanasios Protopapas, Associate Professor, University of Athens ……

Aaron R. Seitz, Professor, University of California Riverside ……

The author asserts that the content of this project is the result of personal work that has been conducted, and the appropriate reference to the work of others, whenever necessary, has been done, in accordance with the rules of academic ethics.

Acknowledgements

I would like to thank my thesis supervisor and Director of the graduate program in “Basic & Applied Cognitive Science”, Athanasios Protopapas for his guidance, trust and patience throughout the past years. Special thanks to Professor Aaron R. Seitz for the honor of becoming member of the supervisory committee. Many thanks to fellow student of the graduate program Laoura Ziaka and doctoral candidates Sofia Loui and Fotis Fotiadis for their constructive comments and assistance completing the present study. Last but not least, I would like to thank all of the participants, without whom this study would not have been possible.

This study is dedicated to my family, friends and Sotiris Michalos, who accompanied me along this journey.

Summary

Perception of phonetic categories that are not included in a listener's native speech environment is an effortful and often unsuccessful task. In the present thesis we examined phonetic learning of non-native categories from the view of the theoretical model of Task-Irrelevant Perceptual Learning (TIPL), in an attempt to train native Greek listeners to perceive the dental-retroflex phonetic contrast of Hindi. In a series of three experiments we aimed to study the role of reinforcement during training on the phonetic learning under TIPL. For the first time in phonetic learning paradigms the innovative software Mouse Tracker was used, allowing us to record hand movements in terms of a computer's mouse trajectories. Via mouse tracking we were able to record ample data in order to study on-line, competing mental processes for a closer examination of the effects of training on phonetic learning.

Keywords: phonetic learning, mouse tracking, task-irrelevant perceptual learning, second language learning, implicit learning.

Περίληψη

Η αντίληψη φωνητικών κατηγοριών που δεν περιέχονται στην μητρική γλώσσα είναι ένα δύσκολο και συχνά ανεπιτυχές έργο. Στην παρούσα εργασία εξετάσαμε τη μάθηση ξενόγλωσσων φωνητικών κατηγοριών υπό το πρίσμα του θεωρητικού μοντέλου της Άρρητης Μάθησης Άσχετου Έργου (TIPL), σε μία προσπάθεια να εκπαιδεύσουμε Έλληνες ακροατές στη διάκριση του οδοντικού και του ανακεκαμμένου φθόγγου της ινδικής. Μέσω μίας σειράς τριών πειραμάτων επιχειρήσαμε να μελετήσουμε το ρόλο της ενίσχυσης στο είδος της μάθησης που επιτυγχάνεται μέσω της Άρρητης Μάθησης Άσχετου Έργου. Για πρώτη φορά χρησιμοποιήθηκε το λογισμικό Mouse Tracker σε παραδείγμα φωνητικής μάθησης, μέσω του οποίου παρέχεται η δυνατότητα

μελέτης της κίνησης του χεριού μέσω της καταγραφής της τροχιάς του ποντικιού ενός υπολογιστή. Μέσω αυτής της καταγραφής έγινε προσπάθεια μελέτης των νοητικών διεργασιών που λαμβάνουν χώρα πριν την τελική απόκριση, προσπαθώντας να μελετήσουμε περισσότερο λεπτομερώς την επίδραση της εξάσκησης στη φωνητική μάθηση.

Λέξεις-κλειδιά: φωνητική μάθηση, καταγραφή ποντικιού, άρρητη μάθηση άσχετου έργου, εκμάθηση δεύτερης γλώσσας, άρρητη μάθηση.

Table of Contents

Part I

Introduction

We are born with the ability to perceive a wide range of phonemes, equipped with perceptual processes that allow as to be sensitive to phonetic categories even after minimal exposure (Eimas, Siqueland, Jusczyc, & Vigorito, 1971). Our perceptual system is then neurally organized as a result of experience to make language-specific phonetic distinctions that are critical to our native language. The effects of this developmental perceptual "tuning" become obvious when, as adults we come before new phonetic categories while learning a second language (Best & Tyler, 2007). It turns out we are bound by the phonological system of our native language (Flege, 2003). Learning to differentiate between non-native phonemes is a rather challenging task and the degree of interference for each individual is dependent on the phonetic perceptual space that is shaped by his/her own native speech environment (Iverson et al., 2003). A few examples of this interference are the discrimination of the English /r/ and /l/ by native Japanese listeners (Miyawaki et al., 1975), the discrimination of the Hindi dental and retroflex consonants by native English listeners (Werker & Tees, 1984) and the perception of the English tense and lax vowels by native Spanish listeners (Fox, Flege, & Munro, 1995).

In order to form new phonetic representations it is important to enhance attention to critical dimensions of the acoustic stimuli and at the same time divert attention away from the unimportant ones concerning the target phonetic contrast. Many different training methods have been used to study the necessary conditions leading to improvement in identification and discrimination of non-native phonemes, as well as generalization of learning in novel speakers and/or complex linguistic structures (Kondaurova & Francis, 2008; McCandliss et al., 2002; Seitz et al., 2010; see also

review from Bradlow, A., 2008). Experimental parameters that could vary during phonetic training are stimulus variability and presentation schedules, attention, intention and performance feedback. McCandliss et al. (2002) were the first to study the role of feedback during training, in their attempt to examine phonetic learning from a Hebbian point of view. In their research they used stimuli from a rock-lock continuum, asking Japanese listeners to identify the initial phoneme as /r/ or /l/. There were four training groups, one with a fixed rock-lock pair throughout training and another one with adaptively less distinguishable stimuli, each group either received feedback on their performance or not. In agreement with past research the presence of feedback led to robust training for both training groups, confirming the effectiveness of traditional training approaches. Interestingly, in the case of adaptive training substantial learning was also found after only three days in the absence of feedback.

Another study amongst the novel training approaches is the one of Seitz et al. (2010), who examined the perceptual learning of single formant transitions. In their experiment participants were not aware of any kind of phonetic training and their intention was to identify the two loudest amongst eight different animal sounds. Their goal was to assess the effects of training on an irrelevant explicit task for the participants, such is the identification of the loudest animal sound, on the detection of formant transitions that are critical in speech perception. Such an approach falls into the recent model of task-irrelevant perceptual learning (TIPL) (Seitz & Watanabe, 2009), which posits that learning can occur for features of a given stimulus that are not necessarily related to the task at hand. According to TIPL perceptual learning can take place as long as two conditions are being fulfilled within a small temporal window: First an incoming stimulus is neurally registered (regardless of attention or even awareness) and second positive reinforcement is given as a result of successful performance on any given task. Positive reinforcement could be explicitly

given by the training task or received as internally generated signals, on the basis of one's realization of successful performance. TIPL has been mostly implemented in the visual modality and has proven to be effective even when stimulus' features are presented in subthreshold level (Watanabe, Náñez, & Sasaki, 2001; Tsushima, Seitz, & Watanabe, 2008), which has been often interpreted as "sensory plasticity in the absence of attention" (Seitz & Watanabe, 2005).

Studying phonetic learning with mechanisms of task-irrelevant perceptual learning has been recently tried (Vlahou, Protopapas, & Seitz, 2012). In this research experimenters contrasted various phonetic training methods to see which would lead to better identification and discrimination of the Hindi dental and retroflex consonant by Greek listeners. Vlahou et al. (2012) used consonant-vowel syllables starting with the dental /t/ or the retroflex /ʈ/, phonetic categories that correspond to only one phonetic representation for native Greek listeners, in a three training days' experiment. In Experiment 1 the traditional approach of explicit learning with immediate feedback was used. In Experiments 2 and 3 participants trained in another explicit task; after listening to two pairs of identical sounds, one with a dental syllable and - before or after - another one with retroflex, participants were asked to decide which one of the two pairs differed in *intensity*. Unbeknownst to participants the pair that differed in intensity was always the retroflex one, so correct detection of the pair whose intensity differed was equivalent to correct identification of the retroflex pair. The difficulty of the task was adjusted adaptively in order to always be just hard enough, ensuring high levels of positive reinforcement, either explicitly (Experiment 2) or implicitly (Experiment 3 – Implicit training without feedback). Improvement in phonetic identification and discrimination was found in all three experiments, with the most robust improvement being in the implicit training group without any feedback. Even though these are strong evidence confirming that implicit mechanisms of learning can effectively improve phonetic

discrimination, the question of the nature of what was learned remains open. Did participants manage to learn two distinctive phonetic categories or is it the refinement of perceptual representation for the retroflex tokens, which led to such results? Would differential reinforcement of both phonetic categories lead to more efficient training?

The present study aims to examine the efficiency of task-irrelevant learning without feedback in phonetic training and using different reinforcement regimes to search for signs of categorical or perceptual learning. For the first time in phonetic learning designs the Mouse Tracker software is used (Freeman & Ambady, 2010) in order to take advantage of the recordings of hand movements reflecting on-line processes. Detailed analysis of the Mouse Tracker software and the data we collected is presented in Part II.

Experiment 1: Source Detection Training with Pairing of Source and Phoneme type

The purpose of this experiment was mainly to study the effect of TIPL on phonetic learning and compare our findings with Experiment 3 of Vlahou e al., 2012. The general paradigm of implicit training without feedback on an irrelevant, explicit task and post-training testing on a relevant stimulus' parameter, such as successful discrimination of the phonetic contrast, was maintained. We altered this Experiment in four ways: first, because of the use of the Mouse Tracker software, in all the phases of the experimental process responses were given via mouse-clicking. Second, the explicit task was an adaptive psychophysical procedure of acuity in sound source localization (left or right). Third in our experiments reinforcement was paired with both consonant types (i.e. both sources in Experiment 1), whereas in the experimental design of Vlahou et al. (2012) reinforcement was only provided for the retroflex consonants. Last, in each trial participants heard a given syllable only once, while Vlahou et al. (2012) presented pairs consisted of the same

syllable and both syllable pairs were presented in each trial.

Method

Participants. Fifteen adult native Greek speakers (8 females, 7 males, 21-34 years old, mean 26.9) participated in Experiment 1. One participant was excluded due to failure to comply with the instructions during training, so data from only fourteen participants are reported for this Experiment. Most were students who either volunteered or were given course credit for their participation and the rest were adults of higher Education. None of the participants reported any hearing or speech impairments. None reported any previous experience with Hindi sounds when asked at the end of the experiment.

Stimuli. Stimuli were natural recordings of Hindi syllables spoken by two native Hindi speakers, the ones used by Vlahou et al. (2012). There were ten syllables starting with the dental consonant ([ta:]) and ten syllables starting with the retroflex consonant ([ʈa:]) for each speaker, that is a total of 20 [ta:] and 20 [ʈa:] tokens, all resembling the Greek /ta/ syllable. The initial audio files were sampled at 22050 Hz (16-bit mono) of 350 ms total duration each. We further manipulated all audio files into having the same mean intensity (set at 77 dB). For the purposes of the first two experiments we additionally resampled the 20 tokens (ten [ta:] and ten [ʈa :]) of the speaker used in the training phase (Speaker 1) to 96000 Hz (16-bit mono). Because of this manipulation, when the syllables were turned into asynchronous stereo sounds, a minimum of 0.010 ms asynchrony (time difference) between the two audio channels could be achieved, reaching the interaural delay threshold of 0.010-0.020 ms between the two ears (Skottun, Shackleton, Arnott, & Palmer, 2001). All manipulations of the stimuli were done using Praat (Boersma, 2001).

For the psychophysical source detection task a simple 1-kHz harmonic tone was used matching the characteristics of the Hindi syllables (sampling rate 96000 Hz, output resolution 16-bit mono, duration 350 ms, mean intensity 77 dB).

Procedure. The experiment consisted of three phases. A pre-training testing phase, a training phase and a post-training testing phase. During all phases stimulus presentation was controlled by the Mouse Tracker software interfacing on-line with an external environment to determine stimulus computation and presentation (more details on the procedure regarding Mouse Tracker are presented in Part II). Pre- and post-training phases were carried out on the experimenter's computer under direct supervision, while the training phase took place on the participant's own personal computer or laptop at home, which was accordingly set by the experimenter to match the conditions of the pre-training and post-training phases. In detail, we provided each participant with headphones and we adjusted the sound volume at an individually comfortable level, which was maintained throughout all the phases of the experiment. We also gave participants printed instructions in order to carry out the training phase by themselves and we monitored their progress by daily communication.

Pre-training testing. Participants were informed that the purpose of this experiment was to study an individual's ability to detect the source of a stereo sound, namely if it comes from the left or the right. For this task we used a 1-kHz harmonic tone at the pre-training and post-training testing phase. In each trial the harmonic tone was turned into a stereo sound, in which silence was added in the beginning of one of the two channels, creating the impression that the tone was heard either coming from the left (if the silence was added on the right channel) or coming from the right (if the silence was added on the left channel). The time difference between the two channels (i.e. the duration of the silence) was starting at 0.5 ms, which is an easy detectable time difference, and

was further adjusted adaptively according to the participant's performance with minimal step set at 0.01 ms. Stimulus' asynchrony was adaptively computed so as to maintain participants' attention and at least 75% correct responses for the normal trials. The adaptive procedure we used was based on the Accelerated Stochastic Approximation (Kesten, 1958 as cited in Treutwein, 1995). Separate adaptive procedures were applied for each direction (left and right).

We informed participants that they would hear a simple tone which would either come from the left or the right and they had to *respond as quickly and as correctly as possible about the source of it*, by clicking to the corresponding box on the screen (*Figure 1*).

Figure 1. Image of the experimental environment for the explicit tasks and training phases in all Experiments. For Experiments 1 and 2 each button corresponds to a sound coming either from the left or the right respectively and for Experiment 3, a sound of either lower (left button) or higher (right button) intensity difference. For each trial, stimulus presentation was initiated by left-clicking on the START button.

Participants were also informed that the difficulty level of the task would be adjusted according to their performance. The onset of each trial was controlled by the participants via left-clicking on a 'START' button at the bottom center of the screen, after which they had to left-click on one of two black boxes with a white cross ('+') at the top corners of the screen, indicating the source of the sound. The source of the harmonic tone was pseudorandomly computed so as to not be the same for more than five consecutive trials. Each threshold measurement was repeated three

times by each participant and ended either when the participant had reached seven reversals of response for each source, or had completed 70 trials in total. The three sessions lasted approximately ten minutes.

The purpose of this source detection task was both to convince participants of the experiment's purpose and divert their attention away from the Hindi sounds during training, and to assess the effect of training on the task-relevant aspect of the experiment.

Training. The structure of training followed the one in pre-training testing phase. The only difference was that during this phase the stimuli were the 20 recordings (ten [ta:] and ten [ʈa:]) of one of the Hindi speakers. We asked participants to listen to each (asynchronous) /ta/ syllable and respond via left clicking on the corresponding top-corner of the screen, indicating the source of the sound. Training included six sessions conducted over a period of three days. Each day participants completed two sessions that lasted approximately 15 minutes and consisted of 400 "normal" trials and 40 "probe" trials, which sums up to a total 1320 trials. After the first session of each training day participants were given the option for a break time.

Normal trials. In each normal trial a /ta/ syllable was randomly presented that started either with a dental consonant ([ta:]) or a retroflex consonant ([ʈa:]). In this experiment the syllables with dental consonants were always paired with one source location (for example always coming from the left) and syllables with retroflex consonants were always assigned to the other source location (i.e. coming from the right). This way each participant was presented with 200 syllables of each

Consonant type	**Normal trials**	**Probe trials**	
Dental /ta/	5x40=200	5x4=20	220
Retroflex /ʈa/	5x40=200	5x4=20	220
	400	40	**440**

Table 1. Training stimuli distribution in all of the Experiments.

consonant type that were paired with one source location (*Table 1*).

With this manipulation correct response about the direction of a sound was equivalent to successful identification of the corresponding consonant type. The source of each consonant type was pseudorandomly assigned for each participant, so that for half of the participants the syllables with retroflex consonants were coming from the left whereas the syllables with dental consonants were coming from the right and vice-versa. As in the pre-training testing phase the time difference between the two channels of the asynchronous stereo sound was ranging between 0.01-0.5 ms, starting at 0.5 ms and adjusted adaptively for each one of the sources (in this experiment for each one of the consonant types as well). The same adaptive procedure was implemented as in the source detection task. During this training phase the decreasing and increasing time difference was equivalent to positive and negative feedback for each consonant type respectively. For the normal trials only five of ten tokens for each consonant type were used during training.

Probe trials. Within the 440 trials of each day of training there were 40 probe trials randomly interspersed. For the probe trials we used the other five tokens of the same speaker for each consonant type that were not used in the normal trials, so the participants heard each one of the ten tokens only four times each day (*Table 1*). Further, for these trials there was no time difference added between the two channels, so they were actually heard as coming from the middle. This way there was no task-relevant correct response, so participants never received feedback for their performance at these trials. A probe trial was considered correct if the same direction was chosen, as the one that all the normal trials of the same consonant type would come from. Because of the fact that probe trials had no asynchrony added, to respond correctly participants would have to base their judgment only on the (task-irrelevant) phonetic contrast. This way possible signs of early learning of the phonetic contrast because of the direction-phoneme

type pairing could be visible.

On the same day as the the pre-training phase we instructed participants to start the training phase at home and we made an appointment to meet with them again after three days. Participants were specifically asked not to have any training on the day of the post-training phase (on the fourth day), on which they only knew that they would be presented with the source detection task with the harmonic tone.

Testing. Testing consisted of two parts. During the first part the same source detection task as in the pre-training phase was held, in order to assess the effects of training on the task-relevant aspect of this experiment. This part lasted as in the first day approximately ten minutes. During the second part of testing we focused on the effects of training on the task-irrelevant aspect of the experiment, which is the discrimination of the phonetic contrast between the retroflex and dental consonant of Hindi. At the beginning of this phase we informed participants that "*in Hindi there are two different groups of sounds, both sounding like the Greek /t/".* These two groups were labeled 'T1' and 'T2' and each label was assigned to one of the consonant types. The testing phase preceded a brief familiarization phase, where each participant heard the ten tokens of each category (five normal and five probes). They were informed to pay close attention, since they would hear the tokens only once. We informed participants that they would "*hear ten tokens of each category in a /ta/ syllable each and they would have to focus solely on the /t/ phoneme with which the syllable was starting; other stimulus' characteristics such as intensity or duration would not be of any assistance to the tasks that followed"*. These detailed instructions were given because pilot administration of the experiment showed a bias towards the duration of the syllables as the key distinction, reported by the participants themselves. For the same reason we also informed participants that in all of the following tasks the sounds would be coming from the center.

This part of the experiment was also held by the Mouse Tracker software, with the only difference that the two response buttons on the top-corners of the screen were labeled with the name of each group, T1' on the top-left corner and 'T2' on the top-right corner (*Figure 2*). The label of the group was consistent with the consonant type coming from the same source during training. For example, for a participant who heard the dental consonants coming from the left during training, 'T1' group would be the syllables with dental consonants, while 'T2' the retroflex ones.

Figure 2. Image of the experimental environment for the identification task (top) and discrimination task (bottom).

Identification. This was a standard identification task. In each trial, one Hindi syllable was presented and the participant had to categorize it as 'T1' or 'T2', as shown in the familiarization phase. This way each participant had to identify each token as a dental or a retroflex sound. There

were 100 trials, half starting with a dental consonant and half with retroflex. Half of the dental sounds were tokens used for the normal trials, that is tokens for which the participant had received feedback during training, whereas the other half were tokens used for the probe trials, for which no feedback was received. For the other 50 retroflex sounds the same proportion was kept.

Discrimination. This was a standard categorical AX discrimination task. In each trial two tokens were presented that either belonged to the same phonetic type (both dental or both retroflex) or to different types (one dental and one retroflex of random order) and the participants had to respond by clicking on the corresponding response button. There were 80 trials, half of which contained tokens from normal trials and half from probe trials. The number of the stimuli (the number of different pairs) were 40 for each voice, paired randomly. For the discrimination task we altered the configuration of the response buttons on the screen so that it would in no way resemble the training conditions, thus avoiding any bias of the training phase and the identification task. Each trial onset was controlled via left-clicking on the START button in the center of the screen and the two response buttons were set on the top and bottom centers of the screen (*Figure 2*). The participant had to click on the top button with the indication '= =' if the two tokens belonged to the same group or click on the button at the bottom with the indication '< >' otherwise.

Before both tasks we also gave participants some trials (5 before the identification task and 8 before the discrimination task) in order get familiarized with the testing environment and process. After completing the Identification and Discrimination tasks with tokens from the trained voice, the same tasks were administered with tokens from an untrained voice (the second Hindi speaker). No feedback was given in any of the testing tasks. The presentation order was randomly computed for each participant in all of the tasks. There were no trials for practicing before the tasks with the untrained voice as we believed that the participants would be familiar with testing process at this

point. The second part of the testing phase lasted approximately 20-25 minutes and the entire testing phase approximately 35 minutes.

In order to assess the effects of training we also administered both the identification and discrimination tasks to 15 other participants (7 females, 8 males, 22-35 years old, mean: 28.4) that did not undertake any prior training. The participants were also native Greek speakers, and reported no hearing or speech impairments as well as no experience with Hindi sounds. The participants were mostly students who participated voluntarily. The purpose of this naïve group was to be used as a baseline reference to this and all of the following experiments.

Data analysis. For the analyses of accuracy we report below we employed generalized mixed-effects logistic regression models for binomial distributions (Dixon, 2008), via logit transformation (Jaeger, 2008) with participants and tokens as random factors (Baayen, Davidson, & Bates, 2003) using package lme4 (Bates, Maechler, Bolker, & Walker, 2014) in R. Effect sizes (β) are estimated log-odds regression coefficients with zero corresponding to no effect. Asynchrony and intensity differences, as continuous dependent variables, were analyzed using lmer() for Gaussian family.

Results and Discussion

Source Detection task. We should note at this point that participants were under the impression of taking part into an experiment whose purpose was to examine the differential time threshold between two sources. Participants were given an explicit psychophysical task before and after training with a simple harmonic tone, whereas during training they had a similar task with a /ta/ syllable in each trial. We implemented two separate adaptive procedures, one for each source (left and right). For each of the three sessions, interaural discrimination threshold was estimated

as the average of the stimulus' asynchrony for the last five reversals of response. In cases where the participant had not completed a total of five changes of response, all of the reversals where used. Since two separate adaptive procedures were used for each one of the sources, 12 thresholds were computed for each participant, three before and three after training, for the left and the right source.

First, we wanted to investigate if the source of the stimulus had any effect on the estimated thresholds so we regressed threshold, as a continuous variable, onto the time on which the task was given ("before" or "after") and the direction of the sound ("left" or "right") with a linear model described in R notation of the form:

```
threshold ~ time * direction + (1|sID),
```

with participants as random factor. The effect of direction was not significant (*right vs. left: β=0.06, t=1.865*), suggesting that the two adaptive procedures used did not affect the estimated thresholds, so there was no bias of direction. Significance was also not found concerning the time of the task (*after vs. before: β =0, t=−0.004*), suggesting that there was no effect of the explicit training task on source detection performance for both trial types (no interaction between time and direction, (χ^2 *=1.26, df=1, p=0.262*)).

Training. *Figure 3* shows the mean absolute asynchrony for normal trials (in ms), through the 440 trials of each training day. If an asynchronous sound was coming from the left, asynchrony (i.e. the step of the adaptive procedure) was negative whereas a positive sign was given for sounds coming from the right. Here we present asynchrony as an absolute value (regardless of direction).

We noticed an initial abrupt dropping of trajectories from the high starting level of asynchrony each day and a settling after the first trials, so we first wanted to examine if there was

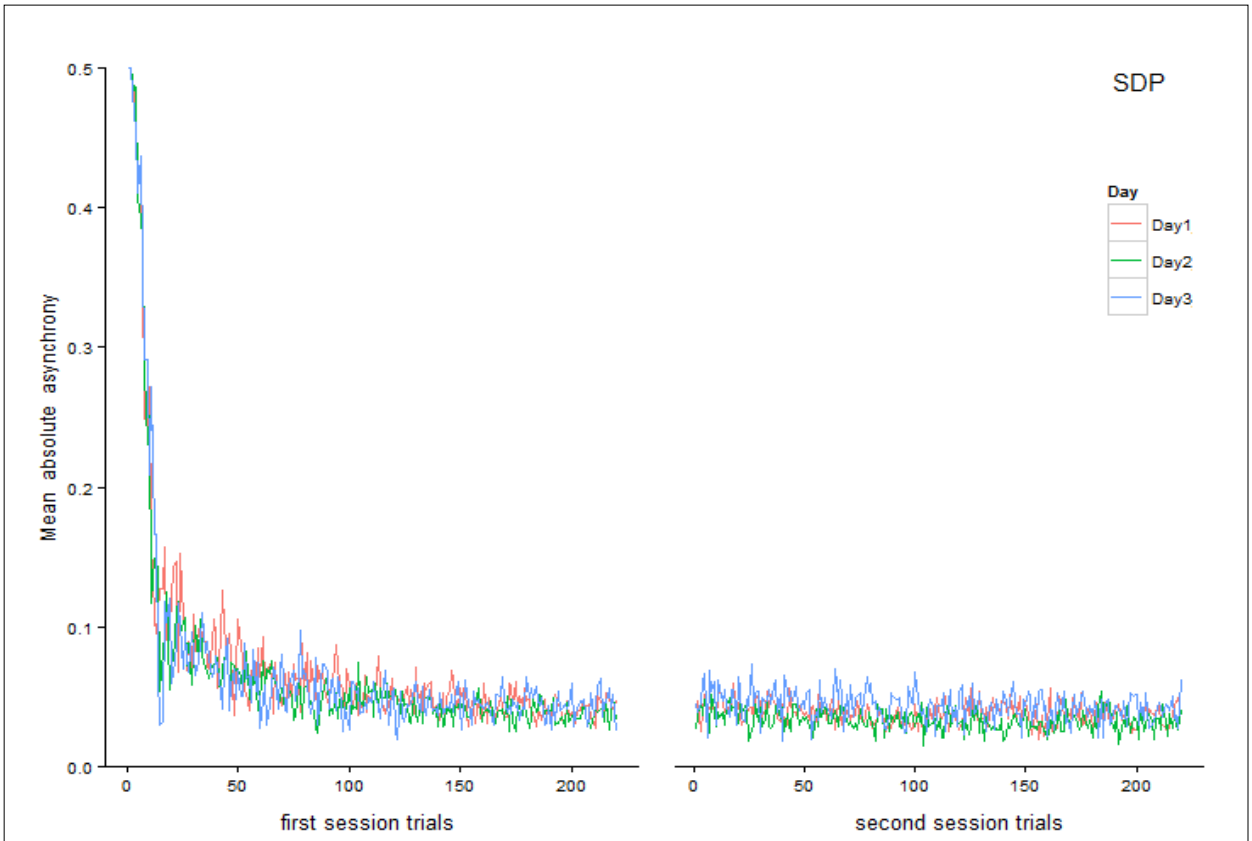


Figure 3. Mean absolute asynchrony of stereo sound as a function of trials, grouped by day. Only normal trials, which were the asynchronous sounds, are presented here. Interaural time difference started at 0.5 ms and was adaptively adjusted.

an effect of session ("first" or "second") on participants' performance. With asynchrony as a continuous dependent variable, we used a linear model of the form:

```
asynchrony ~ trial * Day * Session + (1|subject) + (1|token),
```

with trial (1-440), Day (1-3) and Session ("first" or "second") as fixed effects and participants and tokens as random effects. Results showed a significant effect of session ($\beta = -0.12$, $t = -10.957$), suggesting a large difference on performance within each day. In order to focus on the possible improvement during training, the first 220 session trials were excluded, so only the second session trials of each day are used in further analysis. In order to examine the linear effect of trial and training day, as well as their possible interaction on participants' performance, we regressed mean absolute asynchrony onto trial and day with participants and tokens as random factors using a

model of the form:

```
asynchrony ~ trial * Day + (1|subject) + (1|token).
```

There was a significant effect of day on asynchrony (*β=0, t=2.687*) with an increase of asynchrony values (higher values of asynchrony represent larger time differences and therefore worse performance) and no significant effect of trial (*β=0, t=−0.552*), suggesting that participants' performance remained stable within the second session of each training day and that across days it improved slightly (note that the effect size of day is really low). Results are consistent with no improvement during training which can be explained focusing on the individual performances (*see Appendix A for the individual training performance for the 3 training days*). Individual asynchrony very often reached threshold levels, making further improvement almost impossible.

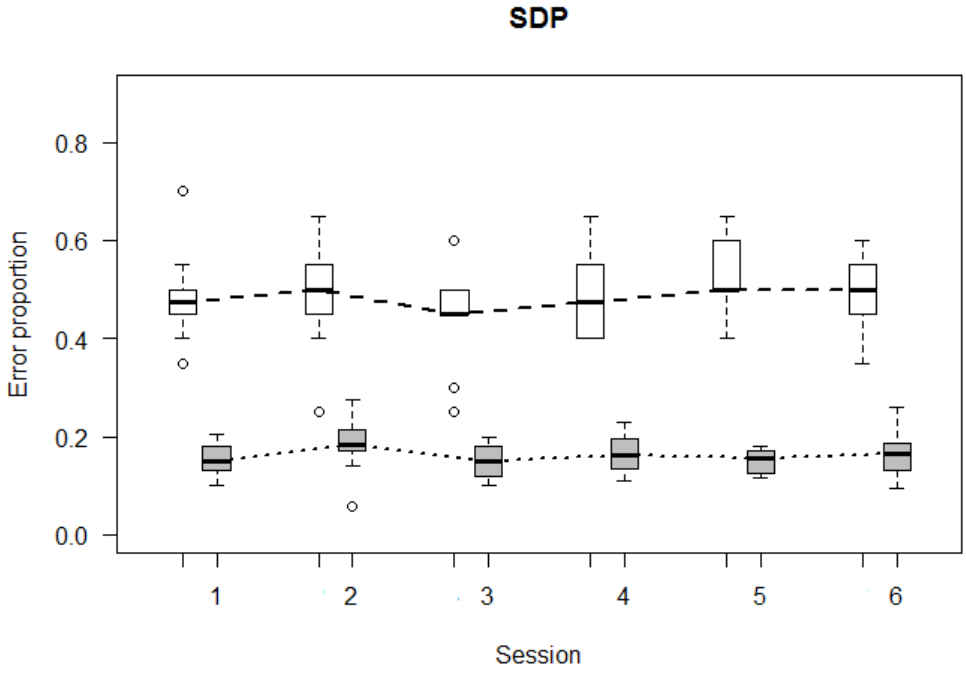


Figure 4. Mean error proportion in each training session (2 per training day) for normal trials (dark grey bars) and probe trials (white bars). Boxes enclose the middle 50% of error proportions and medians are represented with thick lines. White circles represent extreme values.

Figure 4 shows error proportion during the 6 sessions of training (two every day, 1 is the first session of the first day and 6 the last session of the third day) for normal and probe trials.

Participants heard 440 asynchronous /ta/ syllables (400 normal and 40 probe trials) and they had to respond about the direction of each sound. Probe trials were always coming from the center so there was no feedback in the form of asynchrony alteration. One participant completed the 1320 trials over a period of four days due to software problems. For him performance for only the last three training days was used in the analysis, for the 1053 trials available.

First, as for mean absolute asynchrony, we wanted to examine if there was significant difference between the first and second session of each day. We regressed responses onto session ("first" or "second") and trial type ("normal" or "probe") via GLMM with participants and tokens as random factors. This model in R notation was of the form:

```
response ~ Session * trialtype + (1|subject) + (1|token).
```

Since the main effect of session was not significant (*β=-0.02, z=-0.19, p=0.845*), all trials from each training day are included in the following analysis. The effect of trial type on performance was significant (*β=-1.71, z=-21.69, p<.001*), with error proportion for probe trials at chance level (*β=-0.014, z=-0.17, p=0.863*). This shows that the pairing of direction and phoneme type, which was for the probe trials the only available cue, did not provide any facilitation on the explicit task for those trials, showing no signs of early phonetic learning due to the explicit task. Low error proportion for the normal trials (*Figure 4, see also Appendix D for details on error proportion scores for all three experiments*) is consistent with the adaptive nature of the task. The linear effect of session was not significant (*β=0.019, z=-0.66, p=0.508*) and there was no interaction between trial type and session (*β=-.02, z=-0.84, p=0.402*), suggesting that for both trial types error proportion maintained on the same levels throughout training.

Testing. In this Experiment, for both the identification tasks an additional parameter was taken into consideration. If a participant was able to identify the phonetic contrast successfully but

had assigned each phoneme type to the opposite label after familiarization, then error rates would exceed 50%. This would be more likely for the naive group, but this rationale applies in general, and therefore was extended, to all of the identification tasks. We considered that any participant who had achieved total error rate for normal and probe trials over 50% could have drawn a distinction between the two phonemes but they were mislabeled, so we reversed all of the responses leading to a total error rate of less than 50%. This conservative transformation was also applied by Vlahou et al. (2012) and we decided to also adjust our analyses this way so that direct comparison was possible.

For the post-training analysis we first examined the performance of the naive group that didn't have any training with the explicit task, in order to use as baseline for the trained group in this and following experiments. To this group only the identification and discrimination tasks for both voices were administered. The structure of the stimuli was exactly the same as for the participants that underwent training. Of course for this group there was no actual distinction between normal and probe trials, since participants had no previous experience with either stimulus subgroup. There was also no actual distinction between the two speakers. For the naive group we examined performance in regard to the speaker and compared to chance level. The model here was of the form:

```
response ~ speaker + (1+speaker|subject) + (1|token)
```

with two types of responses (“correct” or “incorrect”), regressed onto two speakers (“trained” or “novel”) with participants and tokens as random factors. Error proportion was significantly below chance for the trained voice (Speaker 1) ($\beta=-0.32$, $z=-2.03$, $p=0.042$) which shows that native Greek listeners were to some extent able to detect the phonetic contrasts of the two Hindi consonant types even without any training. Error proportion was still high enough (44.13% for normal and

41.07% for probe trials for the trained voice and 41.33% for the untrained voice) allowing plenty of room for improvement. Further, no significant difference of performance was found between speakers (*β=−0.06, z=−0.28, p=0.781*), which suggests that participants' performance was not based on any specific characteristic of the speaker used during training. For the discrimination task the model was of the form: `response ~ speaker+ (1+speaker|subject) + (1|pair)` as mentioned above, with participants and pairs as random effects. Here too, there was no difference between speakers (*β=0.10, z=0.58, p=0.559*) with performance for Speaker 1 at chance level (*β=−0.15, z=1.14, p=0.255*).

To examine the effects of training on identification performance for the trained voice, we used a model of the form:

`response ~ group * trialtype + (trialtype|subject) + (1|token),`

regressing the two types of responses (“correct” or “incorrect”) onto two testing groups (“naive” or “trained”) and two trial types (“normal” or “probe”), with participants and tokens as random factors. For the discrimination task a similar model was used with participants and pairs (the tokens of the discrimination task) as random factors. The linear model in discrimination task was of the form: `response ~ group * trialtype + (trialtype|subject) + (group|pair).` For the untrained voice there was no fixed effect of trial type in either task, since there were no probe trials. *Figure 5* (groups N and SDP) shows participants' performance (error proportion) for the naïve group (N) and the trained group (SDP) of Experiment 1. We found no significant effect of group (*β=−0.15, z=−0.74, p=0.458*) in phonetic identification, suggesting that the explicit training did not improve the overall performance. Error proportion at the discrimination task was also not significantly different between naïve and trained listeners (*β=−0.04, z=−0.21, p=0.836*).

Analyses for the untrained voice have mainly the purpose of searching for evidence of

learning generalization to different speakers, since it has been shown to be a rather difficult task and highly dependent on stimulus variability either for natural recordings (Lively, Logan, & Pisoni, 1993) or synthetic stimuli (Protopapas & Calhoun, 2000). Given the limited variability of our stimuli we did not expect to find improvement of performance for the novel speaker, as shown by Vlahou et al. (2012). We have already seen that participants in the naïve group could, at least to some extent, discriminate between the phoneme types. For the novel speakers performance of the trained group of Experiment 1 was not significantly different than participants' performance in naïve group (*β=0.15, z=0.67, p=0.504*), being in fact slightly worse (*44.29% mean error proportion*). Only in the discrimination task performance was significantly better than naïve listeners' (*β=−0.32, z=−2.18, p=0.029*), although error rates were still rather high (*41.88% mean error proportion*).

In sum, the results show no improvement of performance across training days on the explicit task, as well as no improvement on the psychophysical Source Detection task. Further, we found no difference of performance in identification both for the trained tokens and for probe tokens (consisted only 10% of all trials and participants were given no feedback) compared to the naïve group. Discrimination of the tokens for the trained voice was not improved, while there was some -although rather small - improvement in discrimination for the tokens of the novel speaker.

One of the purposes of this experiment was to try to confirm the effectiveness of TIPL in learning the phonetic contrast, as observed in previous experiments (Seitz et al., 2010; Experiment 3 of Vlahou et al.,2012), which we did not achieve. Assumptions of the reasons why such a difference was found are discussed extensively after Experiment 2 and in General Discussion. The second purpose of this experiment was to examine the nature of phonetic learning achieved through TIPL and the role of reinforcement. To examine our latter question we run at the same

time Experiment 2 changing the reinforcement conditions.

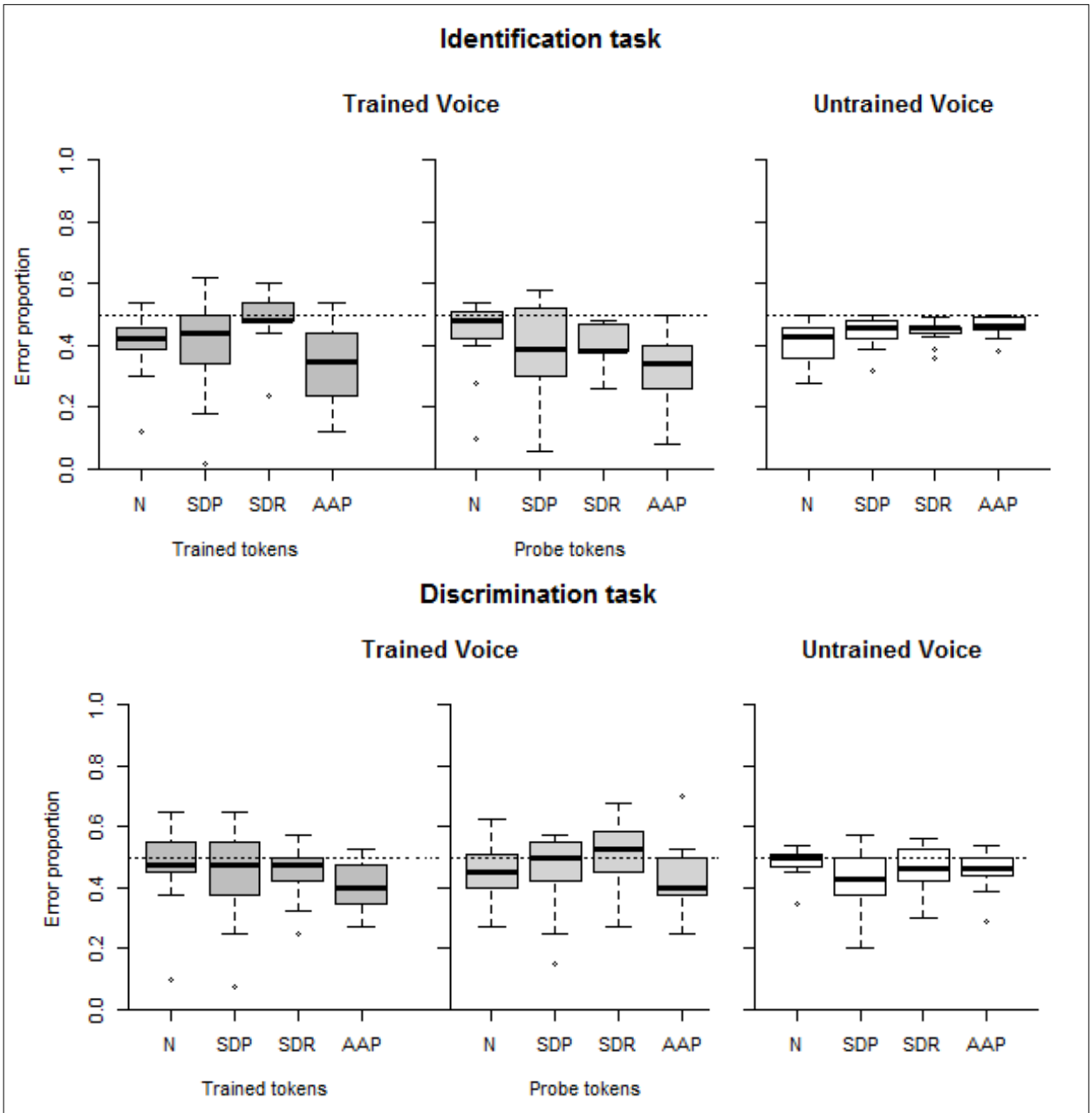


Figure 5. Group performance in each of the posttraining tests, for the trained voice (dark grey boxes) and the untrained voice (white boxes). Bars enclose the middle 50% of error proportion and the median is marked with a thick black line. The dotted line shows chance performance (50% error proportion). N: naïve listeners (group without training), SDP: Source Detection Task with Pairing of source and phoneme type (Experiment 1), SDR: Source Detection Task with Random source of phoneme type, AAP: Amplitude Adjustment Task with Pairing of intensity and phoneme type.

Experiment 2: Source Detection Training with Random Source of Phoneme type

In parallel to Experiment 1 we administered Experiment 2. The explicit training task was again the source detection task. Like Experiment 1, participants had to respond to the direction of a /ta/ syllable but, unlike Experiment 1, both consonant types could be coming from either direction. This way, there could be still reinforcement signals on the basis of success in the explicit task, although no additional categorical information was paired with those signals as in Experiment 1. We thought that in Experiment 1 participants could categorize the tokens based on their direction (the "left" ones and the "right" ones), but at the same time that would mean categorizing them by phonetic contrast. In Experiment 2 this was not the case because the direction of each phonetic category was randomized. If there was significant learning in both Experiments, then same learning effect between the two groups would mean that through TIPL perceptual refinement of the phonetic representations for each token is accomplished, while if performance in Experiment 1 was better, then some categorical learning would have been achieved.

Method

Participants. Fifteen adult Greek speakers (10 females, 5 males, 21-35 years old, mean 27.5) participated in this experiment. Most were students who either volunteered or were given course credit for their participation. None of the participants reported any hearing or speech impairments, or any previous experience with Hindi sounds when asked at the end of the experiment.

Stimuli. In Experiment 2 the same 20 [ta:] and [ʈa:] Hindi syllables were used spoken by two native Hindi speakers, as manipulated for Experiment 1.

Procedure. Experiment 2 consisted of the same three phases as in Experiment 1 with the

same source detection task prior Training as well as the same Training and Testing phases.

Training. Training phase consisted of the same 400 normal and 40 probe trials, with the only difference that in this case the source location of a given consonant type was presented in random order, so that both dental and retroflex sounds could be heard either from the left or the right during training independently from the phonetic category. As there was no consistent pairing between direction and phoneme type correct response on a syllable's direction was did not necessarily mean correct identification of one phoneme type. In addition, for the probe trials there could be no correct response either on regards to the task-relevant parameter (there was no asynchrony) or to the task-irrelevant one (each consonant type would not correspond to a specific direction). So no error proportion scores were computed for the probe trails in Training for Experiment 2.

Post-training. The exact same procedure was implemented as in Experiment 1, with the same two phases, tasks and duration.

Results and Discussion.

Source Detection task. In Experiment 2, two separate adaptive procedures during the explicit training task were also used, one for each source (left and right). There were 3 sessions of the psychophysical task before and after training, yielding 12 thresholds for each participant, six for each direction (three before and three after training). As in Experiment 1, we used a model of the form: `threshold ~ time*direction + (1|sID),` to examine the effect of time and direction on our estimated thresholds. Again, no significant bias of direction was shown (*right vs. left: β=0.03, t=0.476*). As in Experiment 1, we found no significant difference on the estimated thresholds before and after training (*after vs. before: β=0.09, t=1.202*), suggesting no effect of the

explicit training task on source detection performance. There was no significant interaction between time and direction ($\chi^2 = 1.08$, $df = 1$, $p = 0.299$).

Training. Mean absolute asynchrony for normal trials in Experiment 2 through the trials of each training day is shown on *Figure 6*. We found that performance for the first session trials

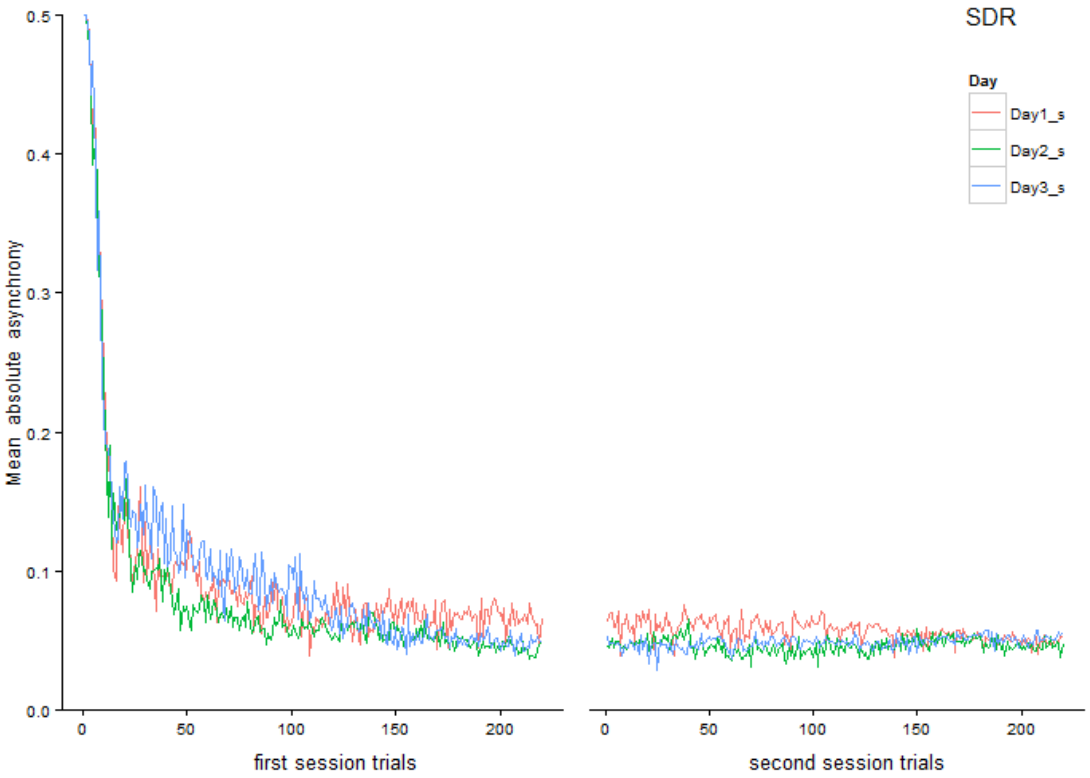


Figure 6. Mean absolute asynchrony of stereo sound as a function of trials, grouped by day for Experiment 2. Only normal trials, which were the asynchronous sounds, are presented here. Interaural time difference started at 0.5 ms and was adaptively adjusted.

was significantly different than performance for the second ones ($\beta = -0.05$, $t = -4.85$), so here too, further analysis includes only the second session trials of each day. Linear effect of trial ($\beta = 0$, $t = -8.633$) as well as linear effect of day ($\beta = -0.02$, $t = -10.42$) were significant, suggesting that there was improvement of performance at the psychophysical training task. Interaction between trial and day was also significant ($\chi^2 = 70.45$, $df = 1$, $p < 0.001$), due to the fact that the linear effect of trial had negative slope ($\beta = 0$, $t = -11.61$) on the first day, whereas positive beta coefficient on

the second and third day ($\beta=0$, $t=5.232$ for the second, $\beta=0$, $t=7.454$ for the third day). Analysis showed that in Experiment 2 there was overall improvement in the psychophysical task, which was mainly dependent on the first day, and that during the rest of the training performance remained on the same levels as in Experiment 1 (*see appendix B, individual training performance on Experiment 2). Figure 7* shows group error proportion during the six sessions of training only for normal. For one participant, initial analysis of training showed an increase of asynchrony during the last trials on the second day of training, possibly because of fatigue. We excluded of further analysis the last 180 trials of that day, keeping a total 1140 trials for this participant.

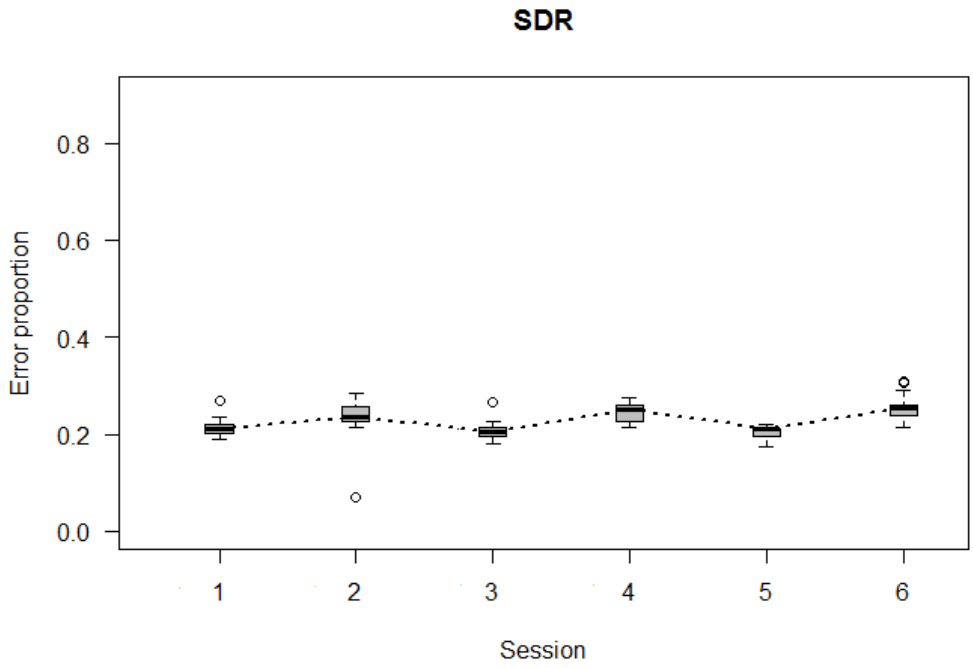


Figure 7. Mean error proportion in each training session (2 per training day) only for normal trials in Experiment 2. Boxes enclose the middle 50% of error proportions and medians are represented with thick lines. White circles represent extreme values.

To examine the linear effect of session throughout all training days, we used a model of the form: `response ~ Session + (1|subject) + (1|token)`, regressing response onto six sessions with participants and tokens as random factors. The linear effect of session was significant for the

normal trials (*β=0.03, z=2.65 p=0.008*) to the opposite direction than expected, showing in fact an increase of error proportion.

Testing. Responses on both identification tasks were reversed for participants who exceeded 50% error proportion for the reasons mentioned in Experiment 1. *Figure 5* shows post-training performance in all four tasks (group SDR). To examine the effects of training we compared performance on the post-training task with performance of the naive group via GLMM as in Experiment 1. There was no significant effect of group (*β=0.1, z=0.78, p=0.44*), suggesting that explicit training did not affect participants' performance on the phonetic contrast discrimination. There was though significant interaction between group and trial type on account of higher error proportion in the normal trials (*β=0.33, z=2.43, p=0.015*) and no difference in the probe trials (*β=−0.14, z=−0.98, p=0.325*) compared to the naive group. Error proportion was still rather high in both trial types (49.2% for the normal trials and 40.92% for the probe trials). Performance in the discrimination task for the trained voice did not differ significantly compared to the naive group (*β=0.08, z=0.61, p=0.541*).

Results also showed no effect of training on identification and discrimination tasks for the untrained voice (*β=0.17, z=0.85, p=0.398* in identification, *β=-0.11, z=-1.06, p=0.288* in discrimination).

Experiments 1 and 2. Results of Experiments 1 and 2 showed no learning of the phonetic contrast with no difference in performance compared to naive participants. The two Experiments presented above differed essentially only in the reinforcement provided during training. In Experiment 2 participants received positive feedback only in relevance to the explicit task, whereas in Experiment 1 positive feedback provided underlying categorical information. Our original hypothesis was that: a) this type of (intrinsic) reinforcement signals could lead to perceptual

sensitivity to the relevant phonetic cues and b) if this effect was equivalent between Experiments 1 and 2, then the extra categorical information would not provide assistance through task irrelevant perceptual learning. The first hypothesis was not verified since there was no significant effect of training on post-training performance. The second hypothesis cannot be examined because of the lack of phonetic learning. This way, even if the comparison between the two experiments showed no significant difference (Trained voice: $\beta=0.26$, $z=1.47$, $p=0.143$ for identification, $\beta=0.12$, $z=0.71$, $p=0.48$ for discrimination. Untrained voice: $\beta=0.02$, $z=0.28$, $p=0.779$ for identification and $\beta=-0.14$, $z=-1.64$, $p=0.102$ for discrimination), these results do not allow us to draw safe conclusions on the nature of phonetic learning through TIPL.

An experimental parameter that could explain the afore mentioned results is the nature of the explicit source detection task. In both experiments we noticed an initial settling down of asynchrony to rather low levels, consistent with source detection being a rather easy task. Even if that was only the case, the adaptive nature of the task could be enough to ensure motivation and therefore, continuous reinforcement. Another more likely hypothesis could be that – unlike what the TIPL model posits - in task irrelevant perceptual learning the explicit task must ensure attention to the stimulus (and its "irrelevant" features) so that the nonspecific, global reinforcement signals can be effective. To test the latter assumption we changed the nature of the explicit training task as presented in Experiment 3.

Experiment 3: Amplitude Adjustment Training with Pairing of Intensity and Phoneme type

Experiments 1 and 2 showed that the source detection task did not help participants improve their performance on the phonetic contrast distinction. We hypothesized that the lack of improvement was due to inadequate attention on the acoustic stimuli. It can be, that deciding

successfully about the direction of the sound does not necessarily demand attention on the stimulus' characteristics and therefore does not lead to learning of the phonetic contrast. In Experiment 3 we manipulate the stimuli (the /ta/ syllables) during training on regard to a different, more substantial, acoustic parameter, which is the amplitude level, in order to test this hypothesis.

Method

Participants. Fourteen adult Greek speakers (6 females, 8 males, 24-35 years old, mean 28.5) participated in this experiment. Most were students who either volunteered or were given course credit for their participation. None of the participants reported any hearing or speech impairments, or any previous experience with Hindi sounds when asked at the end of the experiment.

Stimuli. In Experiment 3 the same 20 [ta:] and [t̪a:] Hindi syllables were used spoken by two native Hindi speakers, as manipulated in Experiments 1 and 2.

Procedure. Experiment 3 consisted again of three phases. Schedule, equipment, software and participants' supervision were identical to those of Experiments 1 and 2. The Post-training phase was also exactly the same as in the other experiments.

Pre-training testing. The difference in Experiment 3 was that the psychophysical explicit task was the amplitude adjustment of a given sound to match a subjectively set, comfortable level. Participants were informed that the purpose of Experiment 3 was to study an individual's intensity differential threshold. For this task we used the same simple 1-kHz harmonic tone as in the previous experiments. In the beginning we asked participants to listen to that tone and adjust the intensity to a level that to them would be considered *"neither high nor low"*. They could listen to the tone as many times as they wanted and after the comfortable level was found, they would listen

to it three more times. After that, in each trial the harmonic tone was turned into an amplitude modulate stereo sound using the audioop module of python environment (https://docs.python.org/2/library/audioop.html), with an - adaptively adjusted - multiplication factor. In each trial participants would have to decide if they considered the tone to be relatively high or relatively low in intensity. In order to maintain the same experimental environment, if the sound was relatively high participants had to left-click on the cross on the right corner of the screen and if it was relatively low on the left corner button (*Figure 1*). The difficulty of the task was analog to their performance. Here only one adaptive procedure was use (also the Accelerated Stochastic Approximation) with the multiplication factor's starting value at 0.500 and minimal step at 0.01. Starting intensity of the tone was 77 dB, (0.141589716 Pa RMS value, as reported by Praat**)** and in each trial a new file would be created with RMS value multiplied with the adaptively adjusted factor. The starting multiplication factor 0.500 corresponds to around 4dB difference for higher intensity and 6dB for lower intensity and the minimal step 0.01 corresponds to around 0.9dB difference (*see Appendix E for detailed audio information*). We repeated each threshold measurement three times with each session's ending criterion at 14 reversals of response or 70 trials completed. For the threshold estimation we used the multiplication factor values at the last six reversals of response. Intensity differential threshold, as depicted in the multiplication factor threshold was the mean of those three estimates.

Training. The same Amplitude Adjustment task was used with the same 20 Hindi syllables. The normal trials in Experiment 3 were sounds that always differed in intensity from the target level and each consonant type was paired with one intensity difference (higher or lower). For example, for half of the participants the syllables starting with dental consonant would always be higher while the retroflex ones would be lower. This way correct response on the intensity

difference of a sound would be equivalent to correct identification of the corresponding consonant type. Probe trials did not differ in intensity so there was no correct response as far as the task-relevant aspect is concerned. A probe trials was considered correct if the response matched the paired consonant type. Post-training phase consisted of a repetition of the Amplitude Adjustment task and the same Identification and Discrimination tasks as in the previous experiments.

Results and Discussion.

Amplitude Adjustment task. In Experiment 3 only one adaptive procedure was used to estimate the intensity difference threshold. There were again three sessions of the psychophysical task before and after training, yielding this time 6 thresholds for each participant, three before and three after training. We regressed threshold, as a continuous variable, onto time using a model of the form: `threshold ~ time + (1|sID),` to examine the effect of time ("before" or "after") on our estimated thresholds. Results showed no significant difference on the estimated thresholds before and after training (*after vs. before: β=−0.05, t=−1.818*), suggesting that training did not improve performance on detecting intensity differences.

Training. *Figure 8* shows mean absolute intensity difference (relative to baseline intensity, as indicated by the multiplication factor) along the trials of each training day, only for normal trials. The line graph on figure 8 shows the absolute difference from the target level, i.e. the step of the adaptive procedure. Low values of the absolute difference suggest approaching of the baseline "comfortable" intensity level.

First we examined the effect of trial (1-440), Day (1-3) and Session ("first" or "second), on intensity difference, as a continuous variable using lmer() in R with a linear model of the form:

```
difference ~ trial * Day * session + (1|subject) + (1|token).
```

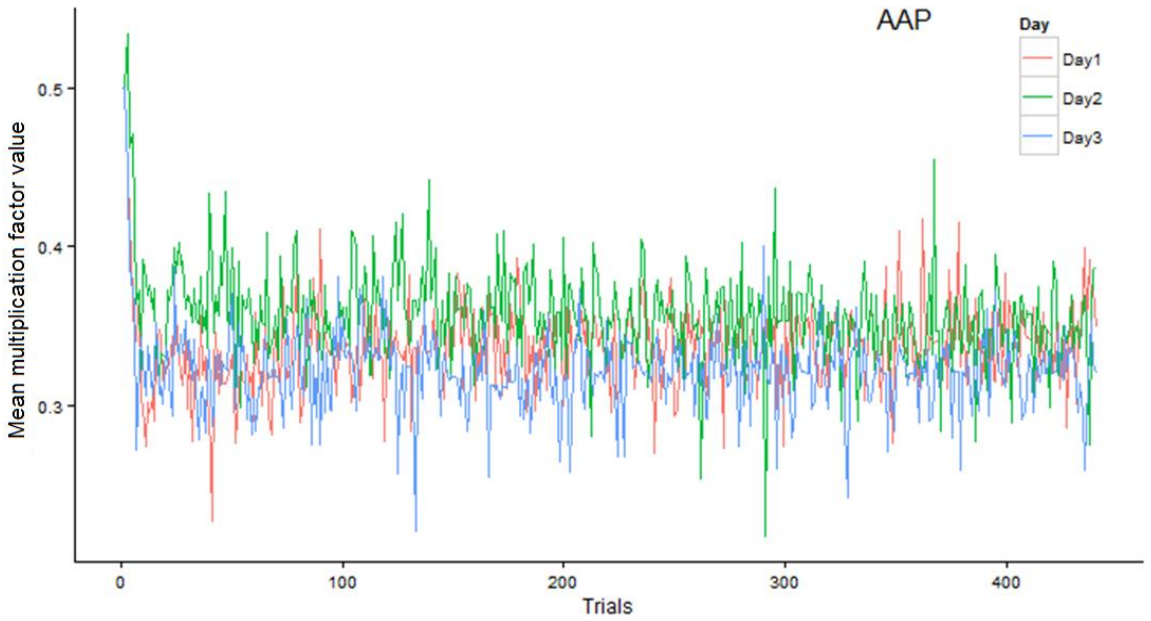


Figure 8. Mean absolute multiplication factor value of the audioop module as a function of trials, grouped by day for Experiment 3. Only normal trials, which differed in intensity, are presented here. The multiplication factor value started at 0.5 and was adaptively adjusted for each participant. This value was multiplied with the RMS value of the initial acoustic stimulus.

The main effect of session was not significant ($\beta=0$, $t=-0.394$) since here the initial settling down of trajectories observed in previous experiments was not present. Because of that and the fact that in Experiments 1 and 2 we used only the second session trials, we excluded here as well the first session of each day, so results could be comparable. To examine the effect of trial and day on intensity difference, as well as their interaction, we used a model of the form:

```
difference ~ trial * Day + (1|subject) + (1|token).
```

The linear effect of trial on participants' performance was significant ($\beta=0$, $t=3.391$), suggesting a slight increase of intensity difference along the trials of training. There was no significant linear effect of day ($\beta=0$, $t=-0.046$). The interaction between trial and day was significant ($\chi^2=12.8$, $df=1$, $p<0.001$), which is due to the fact that the linear effect of trial was always significant, although it had a positive slope on the first training day ($\beta=0$, $t=11.47$) and

negative slope on second and third day (*β=0, t=−9.646 for the second day, β=0, t=−2.494 for the third*). Results showed that there was no linear effect of day and very small effect sizes of trial, indicating very limited within-day learning for the second and third day and no between-day learning.

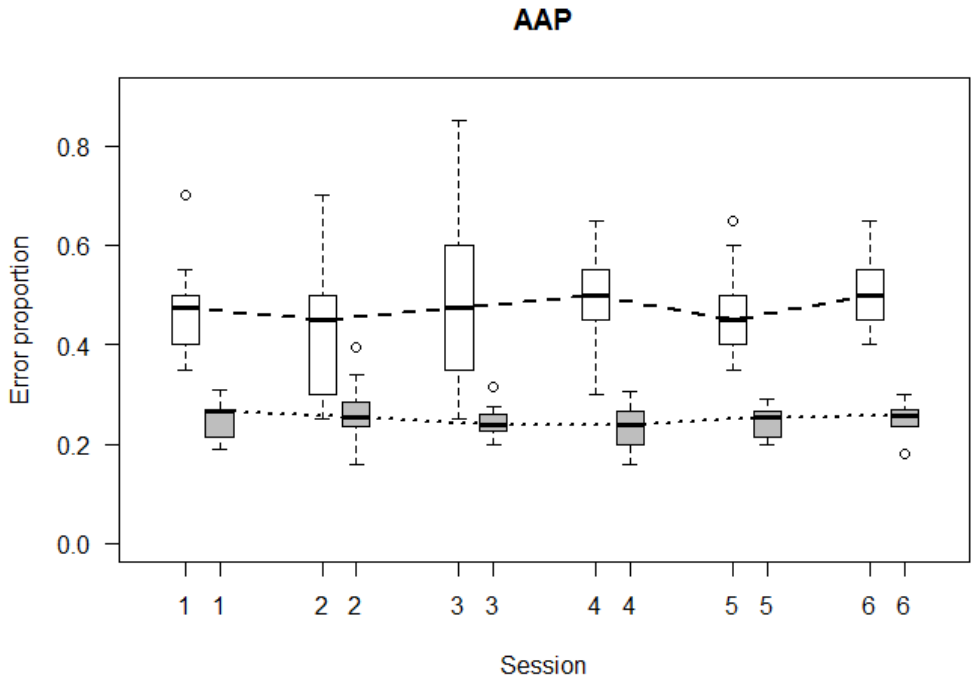


Figure 9. Mean error proportion in each training session (2 per training day) for normal trials (dark grey bars) and probe trials (white bars) in Experiment 2. Boxes enclose the middle 50% of error proportions and medians are represented with thick lines. White circles represent extreme values.

Figure 9 shows mean error proportion for normal and probes trials separately through the total six sessions of training. There was no difference between the first and second session trials (*β=0, z=−0.07, p=0.948*), so we included all of the trials in further analysis. There was again a significant effect of trial type (*β=−1.03, z=−8.12, p<0.001*) because of the fact that for the probe trials error proportion was at chance level (*β=−0.09, z=−0.89, p=0.377*) while for the normal ones it was much lower due to the adaptive nature of the task. As is Experiment 1, for the probe trials there was no facilitation due to the phoneme and intensity (higher or lower) pairing.

To examine the linear effect of session on error proportion we use a model of the form:

```
response ~ Session * trialtype + (1|subject) + (1|token),
```

with phoneme type being "dental" or "retroflex". There was no linear main effect of session (β=0.03, z=1.14, p=0.252), as well as no significant interaction between trial type and session (β=-0.05, z=-1.45, p=0.149) suggesting no change in participants' performance throughout the training phase both for the normal and the probe trials.

Testing. Because of the same rationale as in previous experiments we will report results based on reversed responses in both identification tasks (only for participants with more than 50% error proportion). *Figure 5* shows post-training performance in all four tasks (group AAP). To examine the effects of training we compared performance on the post-training tasks with performance of the naïve group, regressing response onto group ("Naïve" or "AAP") and trials type ("normal" or "probe") via GLMM as before.

The results for the trained voice showed a significant effect of group (*β=−0.43, z=−2.25, p=0.025*), suggesting that explicit training on the amplitude adjustment task improved identification accuracy of the non-native phonetic contrast *(mean error rate 33.71% in normal trials, 32.29% in probe trials*). Improvement was observed both in normal and probe trials as there was no significant interaction between trial type and group (*β=−0.1, z=−1.27, p=0.204*). Performance in the discrimination task however, did not differ significantly from naïve listeners' *(β=−0.18, z=−1.38, p=0.167*).

In regard to performance for the untrained voice, results showed a significant effect group in the identification task (*β=0.20, z=2.34, p=0.019*) to the opposite direction as indicated by the positive beta coefficient, suggesting that error proportion was actually higher than for the naive group and, therefore there was no generalization of learning. In the discrimination task the effect

of group was also non-significant ($\beta=-0.14$, $z=-1.64$, $p=0.102$).

Results suggest that the choice of the training task plays a vital role on TIPL, at least in this phonetic learning paradigm. Compared to an acoustic stimulus' direction, it appears that intensity orients attention to the non-relevant characteristic to be learned as well, while source detection does not require any attention on the phonetic contrast – even if it can be perceived or not. In addition, even if identification is more indicative of phonetic training than discrimination, the lack of improvement in the discrimination task, could suggest weak learning effects.

Part II – Mouse Tracker

Introduction

In experimental psychology the vast majority of researches relies on accuracy rates and response times when trying to address cognitive phenomena or contrast existing theoretical models. It has been constantly pointed out though, that decision and response are not two independent processes but rather they overlap. Either in low-level or in more complex tasks the dynamics of the hand while reaching a decision could reflect the dynamics of the mind (Freeman, Dale, & Farmer, 2011). An innovative software package under the name Mouse Tracker has been designed in order to study on-line cognitive processes unfolding over time, recording hand movements in terms of a computer mouse trajectory (Freeman & Ambady, 2010). With the Mouse Tracker software responses are given via mouse-clicking instead of keyboard buttons. It allows researchers to design and run experiments, recording with high temporal resolution the entire hand movement up to the final response and therefore track antagonizing active representations and their interaction. Mouse Tracker has been used in many different scientific fields as a means to compare contradicting theoretical approaches or study in depth cognitive processes. It has been

used in language research (Barca, & Pezzulo, 2012; Spivey, Grosjean, & Knoblich, 2005), social cognition (Freeman et al., 2011), spatial cognition (Tower-Richardi et al., 2012) and in high-level cognition (Papesh, & Goldinger, 2012). Most researches concern the visual modality but it has been used in auditory tasks as well (Krestar et al., 2013).

For all experiments (during all three phases) we used Mouse Tracker. We wanted to study possible signs of early phonetic learning during training and the competition of active phonetic representations during identification. Because of the adaptive methods used to compute different stimulus parameters, we took advantage of the opportunity provided by the developers to interface on-line with an external programming environment. Using the Mouse Tracker software we were able to record ample data such as accuracy values, reaction and initiation times, and most important trajectory data. We present here trajectory analysis for Experiment 3 (Amplitude Adjustment with pairing of phoneme and intensity difference), for which significant effect of training on post-training performance was found.

Procedure (additional information). Standard Windows mouse-sensitivity settings were used with an 800 DPI mouse. 'START' button was the Mouse Tracker default choice, in Arial font, white print on gray background. Response buttons were black with a cross ('+') in printed in white. Both response buttons were rectangle (0.3 units long, 0.2 units wide). *Figure 10* displays a screenshot of the *Designer* environment.

Participants controlled stimulus presentation by clicking on the 'START' button and were asked to respond via clicking on the corresponding button as quickly and as correctly as possible. There was no message appearing on the screen in case of late initiation time, because pilot administration of the experiment showed interference with the external environment. During the time we designed the experiments, communication with the external environment was

accomplished through Windows Clipboard. Now communication is accomplished through UDP (User Datagram Protocol. *See User's Manual for more information*: http://psych.nyu.edu/freemanlab/mousetracker/help/). For the same reason there was also no limitation of response time, or a 'TIME OUT' message as in previous studies (Barca & Pezzulo, 2012; Freeman, 2013).

Mouse Tracker Data Analysis

Following previous studies using the Mouse Tracker software we performed several steps importing trajectory data prior to analysis. First, we rescaled trajectories into a standard coordinate space represented by a 2x1.5 rectangle. The location of the 'START' button was assigned to coordinates (0, 0), top left corner to (-1, 1.5) and bottom right corner to (1, 0). Thus, positive sign of the x-coordinate responds to the right half of the screen while negative sign to the left half

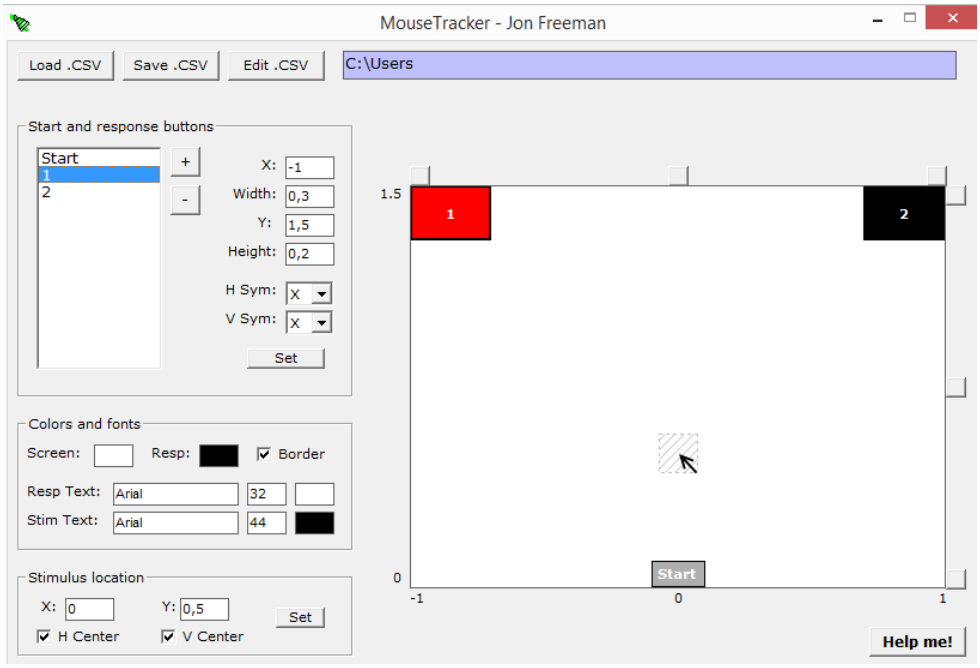


Figure 10. Screenshot of the Designer environment of the Mouse Tracker software.

(*Figure 10*). Second, in order to retrieve trajectory information on Maximum Deviation from the

ideal trajectory (MD), Area Under Curve (AUC), x-flips and y-flips all durations were normalized by resampling into 101 time steps, thus allowing averaging across trials. Third, we excluded all trials exceeding 3000 ms response time (since no such screening was already available in the procedure) and last but not least, we excluded all incorrect responses, analyzing trajectory information only for the correct ones. After importing the trajectories in Mouse Tracker's *Analyzer* all data can be exported into a comma separated value (.csv) file and be further analyzed. All analyses we report below were done for the exported data files, using generalized mixed-effects regression models with package lme4 as in Part I, with participants and tokens as random factors.

First, we will examine a possible change of trajectories through the training days, and study performance for the probe trials in terms of spatial attraction and complexity of the hand movements. In the case of approximation of the trajectories to the ideal trajectory on the third training day compared to the first one would suggest early signs of learning. Second, we will compare trajectories of normal versus probe trials in the identification task with the trained voice, for which no significant difference was found to see if our results regarding error proportion will be reflected on the participants' movements (in terms of mouse trajectories) as well. Probe trials' concentration was only 10% of the normal ones and for those tokens participants did not receive any feedback, so we would assume that that for those trials attraction of the alternative phonetic category during identification would be stronger than the one observed for the normal trials. Last, we will contrast identification performance of the trained group to the one of the Naïve group for a closer examination on the effects of training.

Training. Intensity difference analysis for the normal trials showed no difference between days of training. Error proportion analysis showed low error scores for the normal trials, which is consistent with the adaptive procedure, maintaining difficulty levels so that the training task would

never be for each participant too difficult, maintaining high degree of positive reinforcement through the adaptive procedure. Error proportion for the probe trials showed performance at chance level. We examine below only trajectories for the correct responses in order to study the tendency of participants to choose the opposite (incorrect) intensity difference and see if this behavior changed along training. Based on the afore mentioned results we expected **a)** no difference in trajectories between the first and last training day for the normal trials since participants did not get better on the amplitude adjustment task, **b)** trajectories for normal trials being closer to the ideal trajectory than the probe ones, since the lack of intensity difference for the latter ones would lead to more ambiguous judgment. Training trajectory data from two participants were excluded from this analysis due to software problems. We report analysis for 12 participants.

We loaded trajectories from the first and third day of training (12x2x440=10560 trajectories). We excluded the ones with reaction times over 3000 ms (71 trajectories) and we focused solely on the correct responses, excluding 2896 incorrect trials. The remaining trials were 7593; 3373 of the first day and 3820 of the last day. First we regressed reaction and initiation times, as continuous dependent variables, onto day ("first" or "last") and trial type ("probe" or "normal), with participants and trial codes as random factors, with a model in R notation:

```
time ~ day * trialtype + (1|subject).
```

Results showed no significant effect of day on reaction times ($\beta=-10.23$, $t=-0.767$) as well as initiation times ($\beta=8.83$, $t=1.217$). There was also no significant interaction between trial type and day (*reaction times:* $\chi^2=3.28$, $df=1$, $p=0.07$, *initiation times:* $\chi^2=0.13$, $df=1$, $p=0.712$), suggesting that reaction and initiation times were at the same levels both for the trained tokens and the probe ones through the training days.

Trajectory analysis. Using 'between subjects" analysis, the *Analyzer* environment assigned trajectories for the first day on "condition 1" and for the last day on "condition 2". While importing trajectory data, we used the 'Add Remap' option, which allowed us to flip horizontally all trajectories towards the left response button to the right. With this option we could average and contrast all responses by training day and visualize the mean trajectories in a more direct way. *Figure 11* shows mean trajectories for the first ("day 1") and the last training day ("day 3") for the normal trials.

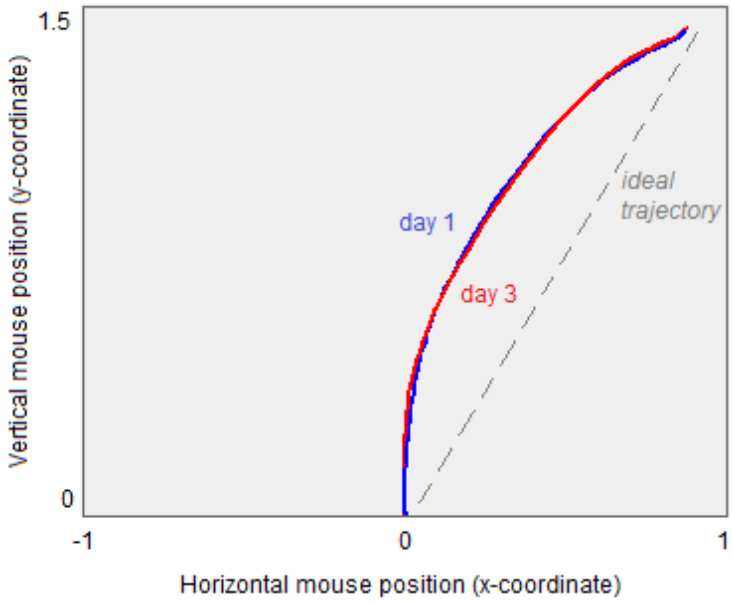


Figure 11. Mean trajectories only for the normal, correct responses during training in Experiment 3. Blue color refers to the first training day and red to the last (the third) training day trials. The gray dashed line is the ideal trajectory, connecting the 'START' button with coordinates (0,0) with each response button.

We wanted to examine spatial attraction of the opposite intensity difference (for example the tendency to respond that a token was of rather low intensity when a participant responded that it was actually rather high). This tendency is depicted in MD and AUC values calculated by the Mouse Tracker. The Maximum Deviation (MD) of each trajectory is the largest perpendicular deviation from the ideal trajectory (a straight line connecting the 'START' button and the response button). High MD values mean larger attraction to the response alternative. For each trajectory the

Area Under the Curve (AUC) is also computed, which is the geometric area between the actual and the ideal trajectory. Negative AUC values represent curves below the ideal trajectory. Higher attraction to the opposite response leads to increased AUC values as well. In order to compare MD and AUC values we used a linear mixed effect model of the form:

```
MD/AUC ~ day * trial type + (1|subject).
```

Results showed no significant main effect of day on MD values (*β=0.02, t=0.950*) and AUC values (*β=0.08, t=1.302*), suggesting overall no difference in attraction towards the alternative intensity along the training. There was also no significant effect of trial type on MD (*β=0.02, t=1.060*) or AUC (*β=0.06, t=1.567*) as well as no interaction of the two factors both on MD (χ^2*=0.48, df=1, p=0.49*) and AUC (χ^2*=0.93, df=1, p=0.336*), showing additionally no difference between tokens with intensity difference for normal and the probe ones. As far the measurement of a movement complexity, we regressed the number of x-flips onto day and trial type using a linear model of the same form. As for MD and AUC, there was no effect of day on x-flips (*β=−0.02, t=−0.157*). *Table 2* shows mean Mouse Tracker values for normal and probe tokens. In sum, results confirmed our first prediction, namely that there would be no difference in participants' movements while responding on the explicit task because of training. That was the fact for the trained tokens, in which neither intensity threshold, nor accuracy or spatial attraction and movement complexity changed, and also for the probe tokens, which could be different only in the case of phonetic learning categorization. Interestingly, even though accuracy between trial types was different for obvious reasons, this did not reflect on terms of spatial attraction. It seems that participants demonstrated the same competition of the alternative intensity level even for the probe tokens, for which accuracy analysis (*Figure 9*) points to answering by luck. This last result shows that maybe there was in fact a generalization of learning of some kind that could not be reflected on accuracy

rates as well.

	Trial type	Day 1		Day 3	
		Mean	**St. Deviation**	**Mean**	**St. Deviation**
Reaction time (ms)	**Normal**	190.56	246.90	1133.71	236.48
	Probe	1133.71	291.93	1218.35	282.26
Initiation time (ms)	**Normal**	196.28	74.08	203.78	91.95
	Probe	202.51	102.30	225.41	90.97
MD	**Normal**	0.38	0.28	0.42	0.30
	Probe	0.27	0.21	0.49	0.31
AUC	**Normal**	0.83	0.74	0.89	0.77
	Probe	0.45	0.38	1.11	0.97
x-flips	**Normal**	6.79	1.38	6.82	1.33
	Probe	6.33	1.82	6.44	1.37

Table 2. Mean and standard deviation values from trajectory analysis for the first and last day of training in Experiment 3.

Identification task with the trained voice. As mentioned above we performed trajectory analysis only for the correct responses given within 3000 ms. Through this screening process we excluded 24 out of 1400 trajectories that were out of time and then 463 incorrect responses out of the remaining 1376. As in previous analysis initial responses were reversed when a participant exceeded 50% total error proportion. Analysis of error proportion has already shown no significant difference between normal and probe trials. We further regressed reaction and initiation times onto trial type ("probe" or "normal) and phoneme type ("dental" or "retroflex"), with participants and tokens as random factors, with a model in R notation:

```
time ~ trialtype * phonemetype + (1|subject) + (1|token).
```

Results showed no significant effect of trial type both on reaction times (*probe vs normal: β=2.85,*

t=0.109) and initiation times (*β=−9.95, t=−0.631*), consistent with no difference on performance between the trained and untrained tokens. There was also no significant effect of phoneme type (*Retroflex vs Dental: β=−4.26, t=−0.157 for reaction times, β=10.26, t=0.626 for initiation times*) as well as no interaction (χ^2*=0.53, df=1, p=0.466 for reaction times,* χ^2*=0.65, df=1, p=0.42 for initiation times*), suggesting overall the same performance amongst trials.

Trajectory analysis. The experimental environment for the Identification task had a standard Mouse Tracker configuration with the 'START' button on the bottom center of the screen

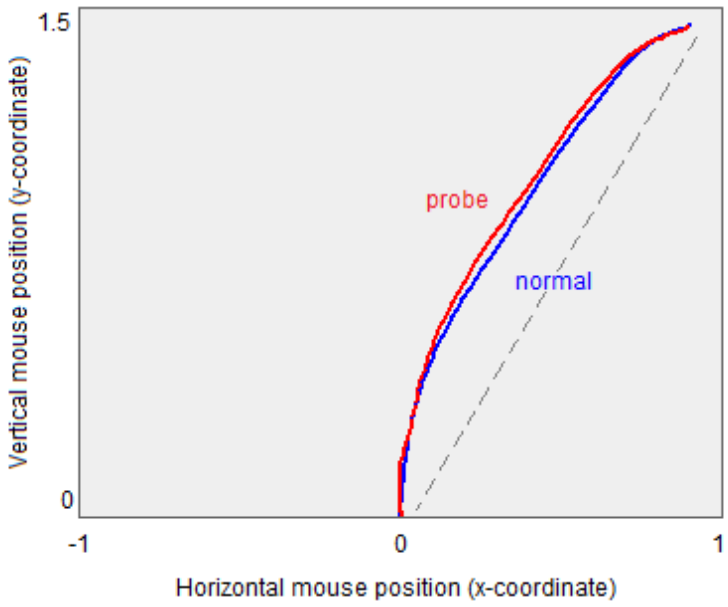


Figure 12. Mean trajectories only for the correct responses in identification task with the trained voice in Experiment 3. Blue color refers to the normal trials and red to the probe trials. The gray dashed line is the ideal trajectory.

and two response button on the top corners (*see Figure 1 top, Part I*). The response buttons were always named 'T1' on the left and 'T2' on the right, with the corresponding phonetic category of each group name counterbalanced to match the training conditions. We used the 'Add Remap' option, flipping horizontally all trajectories towards the left response button to the right. *Figure 12* shows mean trajectories for normal and probe trials in identification task.

In order to examine spatial attraction between the two trial types, we used a linear mixed effect model of the form:

```
MD/AUC ~ trialtype * phonemetype + (1|subject) + (1|token).
```

In agreement with previous analysis, results showed no significant effect of trial type on MD values (*β=−0.04, t=−0.972*) as well as AUC values (*β=−0.06, t=−0.601*), consistent with the same degree of spatial attraction to the opposite trial type. There was also no significant effect of phoneme type (*Retroflex vs Dental: β=−0.05, t=−1.214 for md values, β=−0.12, t=−1.105 for auc values*) as well as no interaction (χ^2*=0.94, df=1, p=0.333 for md values,* χ^2*=1, df=1, p=0.317 for auc values*), suggesting the same degree of spatial attraction for all of the tokens, trained and untrained of both phonetic categories. *Table 3* shows mean and standard deviation values taken from the trajectory analysis for normal and probe trials.

	Normal trials		**Probe trials**	
	Mean	**St. Deviation**	**Mean**	**St. Deviation**
Reaction time (ms)	1210.12	227.68	1225.86	220.99
Initiation time (ms)	224.64	100.61	220.54	85.97
MD	0.34	0.18	0.32	0.14
AUC	0.63	0.43	0.63	0.35
x-flips	7.56	1.74	7.39	1.76

Table 3. Mean and standard deviation values from trajectory analysis for the identification task with the trained voice of Experiment 3.

Another insightful indicator of a trajectory's complexity is the value of x-flips, which is the number of fluctuations along the x-axis. Results showed no significant effect of trial type onto x-flip value (*β=−0.2, t=−0.544*), indicating overall the same way of reaching to the final response. All of the results mentioned above confirm what was shown by the accuracy analysis in Part I. There was no difference in performance for the trained tokens compared to the probe ones, which

were within the trial set in 10% concentration and for which participants never received feedback.

Effects of training on identification performance. Analysis of the Mouse Tracker data for Experiment 3 was in agreement with our previous findings concerning performance for normal and probe trials. Here, we will compare identification performance in Experiment 3 with the one of naïve participants. Based on analysis in Part I that showed a significant effect of training on identification scores, we predict grater MD, AUC and x-flip values for the naïve group in contrast to trajectories of the trained group with the Amplitude Adjustment task. For this analysis we imported trajectory data using the "between subjects" analysis option (with "condition 1" being the trajectories of the naïve participants and "condition 2" the trajectories of the trained group) including both normal and probe trials. From the total 2900 trials (100 trials for 15+14=29 participants) we excluded 39 responses that exceeded the 3000ms limit and 1235 incorrect responses, leaving a total 1626 trajectories (713 of the naïve group and 913 of the trained group).

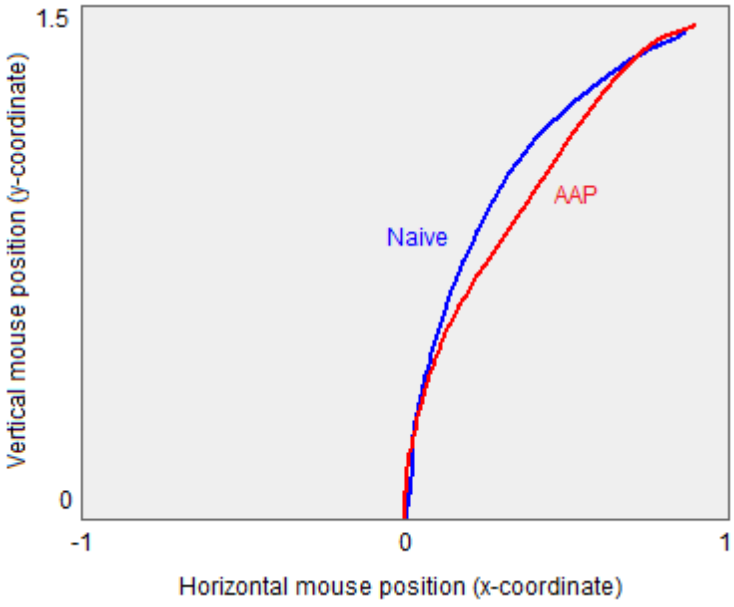


Figure 13. Mean trajectories only for the correct responses in identification task with the trained voice in Experiment 3 (red curve) compared to the Naïve group (blue curve).

Although reaction times were smaller for the trained group than for naïve participants, the difference failed to reach significance (*β=−177.5, t=−1.784*). Initiation times were also not significantly different (*β=−78.39, t=−1.595*). Figure 13 shows mean trajectories for the two groups.

We employed linear mixed-effects model to assess the effect of group ("naïve" or "trained"), trial type ("normal" or "probe") and phoneme type ("dental" or "retroflex"), on indicators of spatial attraction (MD and AUC values). Results showed no significant main effect of group (*Naive vs trained: β=0.01, t=0.195*) on MD values. We also found no significant effect of trial type (*β=0.08, t=1.576*), phoneme type (*β=0.07, t=1.033*) as well was no triple interaction (χ^2*=5.91, df=1, p=0.206*). Analysis for AUC values also showed no significant effect of training (β=−0.03, t=−0.256). AUC values were comparable on regard to the effects trial and phoneme type. In contradiction to our hypothesis, participants' responses did not differ in spatial attraction, even though they differed significantly in accuracy scores, indicating that the improvement in performance may be rather weak. Comparable MD and AUC values are consistent with the same degree of spatial attraction to the opposite phoneme category between the two groups, suggesting

	Naïve		**AAP**	
	Mean	**St. Deviation**	**Mean**	**St. Deviation**
Reaction time (ms)	1386.77	322.42	1217.68	222.32
Initiation time (ms)	290.1	167.02	222.25	92.18
MD	0.35	0.12	0.33	0.15
AUC	0.69	0.32	0.63	0.38
x-flips	6.45	1.51	7.46	1.72

Table 4. Mean and standard deviation values from trajectory analysis for the identification task with the trained voice of the Naive group and trained participants with the Amplitude Adjustment task in Experiment 3.

that even if participants were more accurate, they were not much more certain of their choice. The complexity of trajectories as depicted in the x-flip value also failed to reach significance (*β=0.98, t=1.619*). Mouse Tracker mean and standard deviation values of time, spatial attraction and complexity are presented in *Table 4*.

In sum, analysis of the Mouse Tracker data gave as a more detailed insight into participants' performance in Experiment 3. Results confirmed that participants who underwent amplitude adjustment training were more successful in the phonetic category identification both for trained and probe tokens. Trajectory analysis showed though, that their responses were still attracted to the alternative choice, consistent with rather high error proportion scores (*33.71% for normal and 32.29% for probe tokens*) compared to the INF group (Implicit Training without Feedback) of Vlahou et al. (2012) and no effect of training on the discrimination task performance.

General Discussion

In sum, results show that the choice of the explicit training task can play a significant role on task-irrelevant learning of phonetic categories. In particular, it appears that certain amount of orienting attention on the critical (even though irrelevant to the task) stimulus' characteristics is necessary in order for the intrinsic rewarding signals to be effective. The role of attention in the TIPL model has been mentioned in reviews (Seitz & Watanabe, 2005; 2009) and experimentally tested (Leclercq & Seitz, 2012) – mostly in that it restricts learning (Choi, Seitz, & Watanabe, 2009; Tsushima et al., 2008). Interestingly enough, it appears that too much attention during task-irrelevant learning suppresses the "irrelevant" features to be learned and reduces the effectiveness of TIPL. In all of the studies though the relevant and irrelevant features do not constitute member of the same stimulus, as in our case. For example in the experiment of Seitz et al. (2010) the

relevant feature was an animal sound, while the irrelevant one was a paired formant transition at subthreshold level. In Tsushima et al. (2008) the same separation was present for visual stimuli, with digits in the center of the screen as target features and coherent motion stimuli as the irrelevant parameter. In our case the irrelevant (consonant type) and the relevant (direction/intensity) features were parts of one whole. In such cases it is possible that orienting attention through the demands of the explicit task could be necessary for the neural enhancement of the stimulus' characteristics, while in the case of other explicit tasks (as was the source detection task in Experiments 1 and 2) the lack of sufficient attention could mean that stimulus' features are not sufficiently enhanced (i.e. in our case the initial consonant of the syllable was not neurally registered at a level appropriate for phonetic learning).

On the initial interest of our study, namely if participants would learn two phonetic categories or enhance their phonetic representation of each individual token, we cannot draw any conclusion since no learning was found. To this question we could be able to answer repeating Experiment 3 (with the Amplitude Adjustment task) but in this case a token of a given phonetic category could be heard either louder or softer. Especially in this experimental design, where the nature of the task itself does not demand categorization or promotes any kind of comparison between tokens, results would be at least informative on the role of discriminant reinforcement on task-irrelevant phonetic learning.

In addition, the lack of robustness in learning compared to implicit learning with feedback in Vlahou et al. (2012) could be due to the changes in the reinforced tokens. In our experiment tokens from both phonetic categories received equal amount of positive reinforcement. On one hand reinforcing both, rather than only one phonetic category could mean perceptual refinement of more acoustic representations, leading to more accurate identification of the phonetic contrast.

On the other hand, if some categorical learning is present in phonetic learning, reinforcing only one of phonetic categories could induce the proper distance between categories and lead to more effective training. It would be also interesting to examine if the reinforcement of only one phonetic category is symmetrical. Would it have the same effect reinforcing only the retroflex tokens (which are the ones that don't exist in Greek phonetic inventory) compared to reinforcing the dental ones? Future research could address this question.

Using the Mouse Tracker software we took a closer look into performance of the trained group of Experiment 3. Accuracy analysis had shown that while there was no significant improvement during training, there was significantly more accurate identification of the phonetic contrast both for the trained and the untrained tokens. Specifically for the training phase, trajectory analysis confirmed the same behavior in categorizing the tokens by intensity in the last training day compared to the first day, indicating that the training task was a rather effortful one, as participants were constantly drawn to the alternative intensity sign. Interestingly enough, the tendency to choose the wrong intensity sign was not greater for the probe tokens. This finding combined with the same error proportion and spatial attraction values during identification of both trained and untrained tokens, as well as significant better identification for both, could lead to two conclusions. Either that the probe trials were in some way easier to learn than the normal ones, or that participants were able to generalize what they learned through training to novel tokens as well. Future research should address this phenomenon counterbalancing the normal and probe trials across participants.

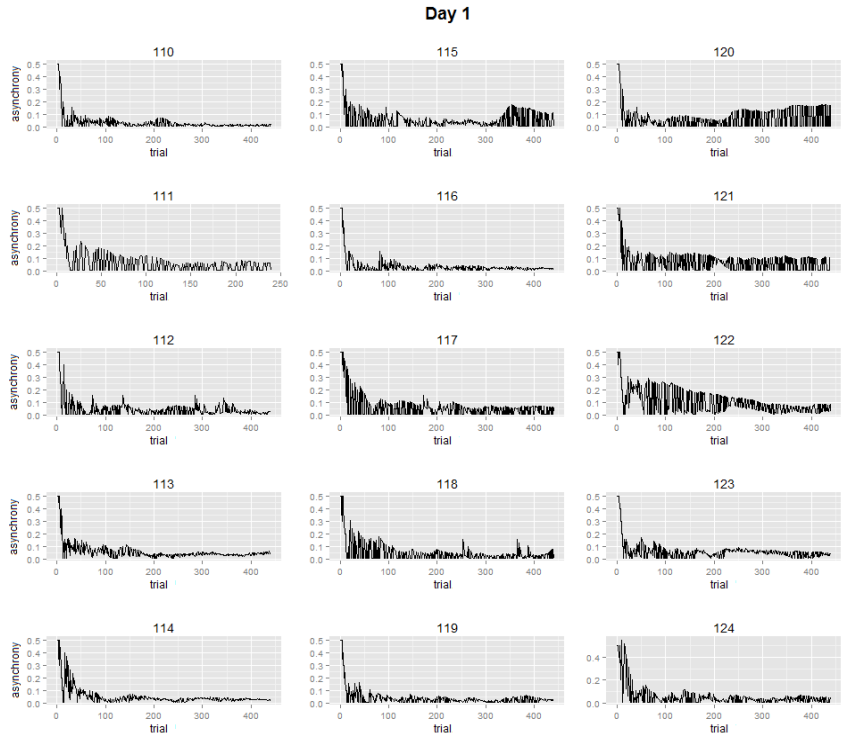
Concerning the effect of training on the identification task, improvement in performance was not mirrored in participants' hand movements. This inconsistency could be due to weak learning through the training method that could not appear on participants' hand movements. Error

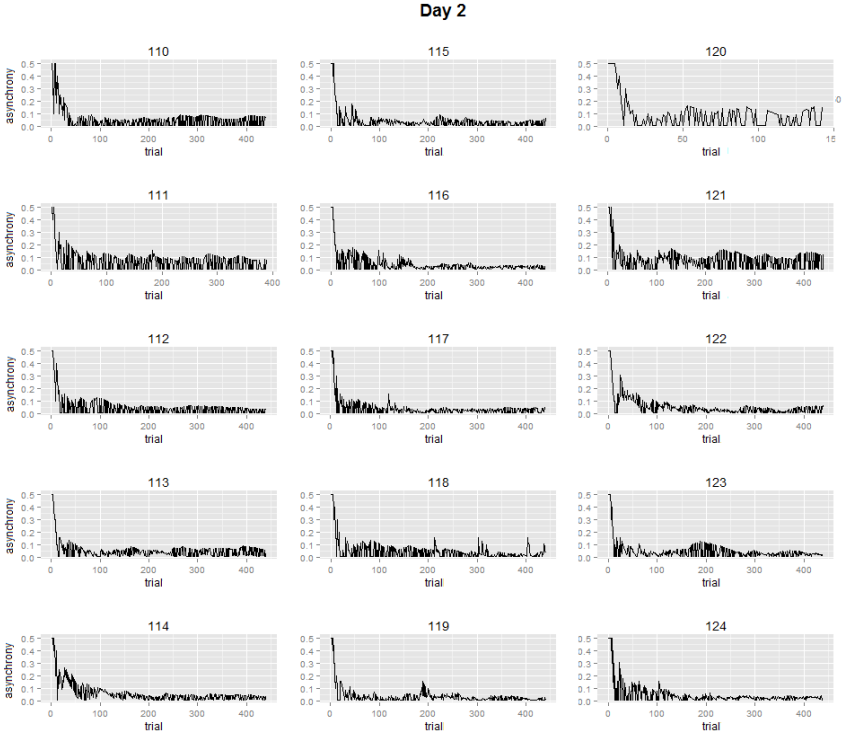
proportion in phoneme identification was close to the one of the Explicit Training group of Vlahou et al. (2012) but still not close enough to performance of the Implicit Learning without Feedback group (*16.27% for normal and 21.20% for probe tokens*). Another way to explain participants' behavior relies on the value of x-flips. Previous research has shown the possibility of dual processes in categorization (Freeman, 2013); one through constant competition of active representations leading to smooth trajectories towards the correct response and another one with more abrupt changes, present with increasing stimulus ambiguity. The value of x-flips taken from our data was the same for both groups (*6.45 for the Naïve and 7.5 for the trained group*) but one could say that the high x-flips value for the naïve listeners was due to ignorance of the two phonetic categories, while participants in the trained group did in fact learn the phonetic contrast but they corrected their responses more abruptly. Such behavior would produce similar trajectories after averaging but does not necessary reflect lack of improvement. On the other hand, this assumption is at this time at least premature because there is only one study (to our knowledge) using Mouse Tracker with acoustic stimuli (Krestar et al., 2013). The majority of studies using Mouse Tracker focus on the visual modality, with the visual stimuli presented in the center or the bottom of the screen after a given time. Such experimental design, as the one used by Freeman (2013), leaves less room for participants to "wander around" in the screen and leads to significantly lower x-flips value. In the auditory lexical decision task by Krestar et al. (2013) the size of x-flips value resembled ours, because as in our study stimulus onset was initiated directly after clicking on the 'START' button. It is an interesting question to test further if the use of the Mouse Tracker software is general appropriate to experiments using auditory stimuli, and which experimental adjustments are needed to lead to safe conclusions.

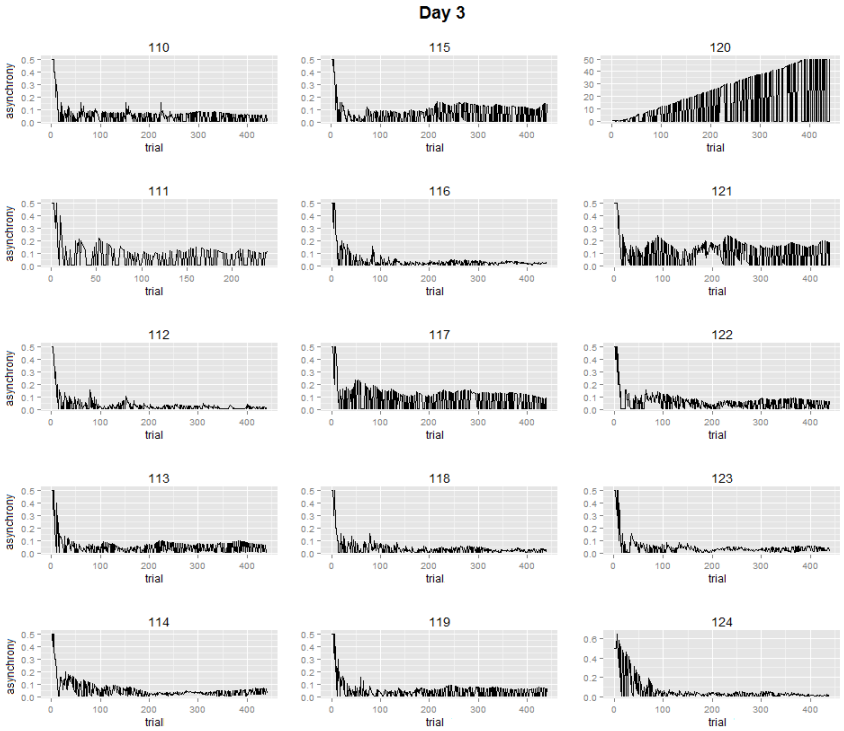
Last but not least, if participants did learn, what is it exactly they learned? Apart from the

stimulus-specific or stimulus general question, lexical training experimental designs using Mouse Tracker should take into account the contribution of the motor system during the transformation of acoustic speech stimuli to a phonetic code. Wilson and Iacoboni (2006) found that speech motor areas such as the superior part of ventral premotor cortex (svPMC) play an important role during non-native speech stimuli, with different activation for native and non-native tokens. Accepting that speech perception is a sensorimotor process we should ask ourselves what is the contribution from the use of Mouse Tracker. Asking participants to respond via clicking with the mouse for the total 1320 trials of training, could we expect significant activation of motor areas? And if so, could it be that participants learned only a motor response to the phonetic contrast? Further research should examine in detail the effects of Mouse Tracker use during training. In our case it would be elucidating to repeat this experiment with participants responding with keyboard during training and performing the identification and discrimination tasks using Mouse Tracker and vice-versa. The methodological ramifications of the Mouse Tracker software in the auditory modality, as well as in learning experimental designs should, be anyway further examined in order be more confident to interpret our results.

Appendix A


Day 1
110
115
120
111
116
121
112
117
122
113
118
123
114
119
124
asynchrony
trial

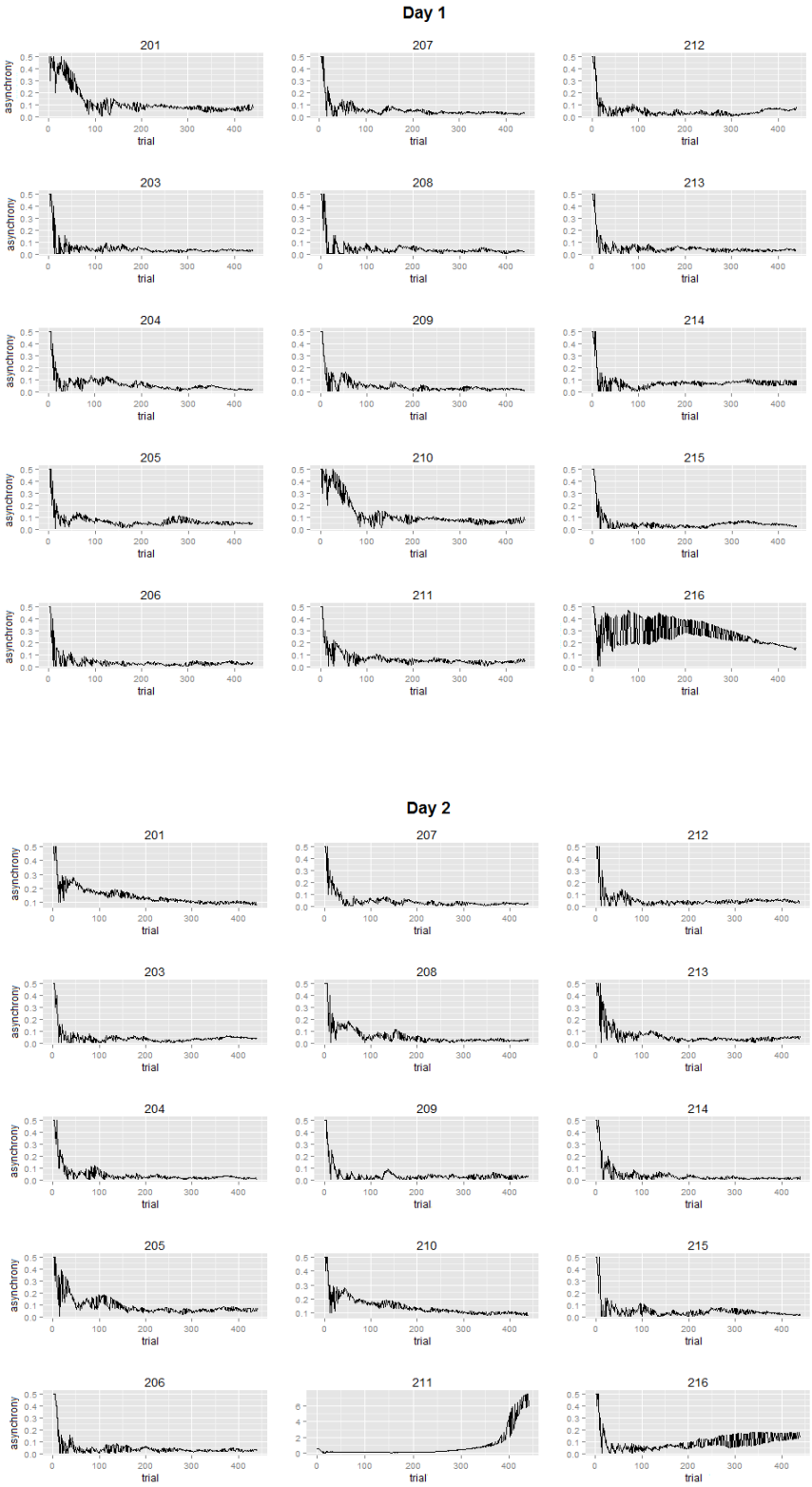

Day 2
110
115
120
111
116
121
112
117
122
113
118
123
114
119
124
asynchrony
trial

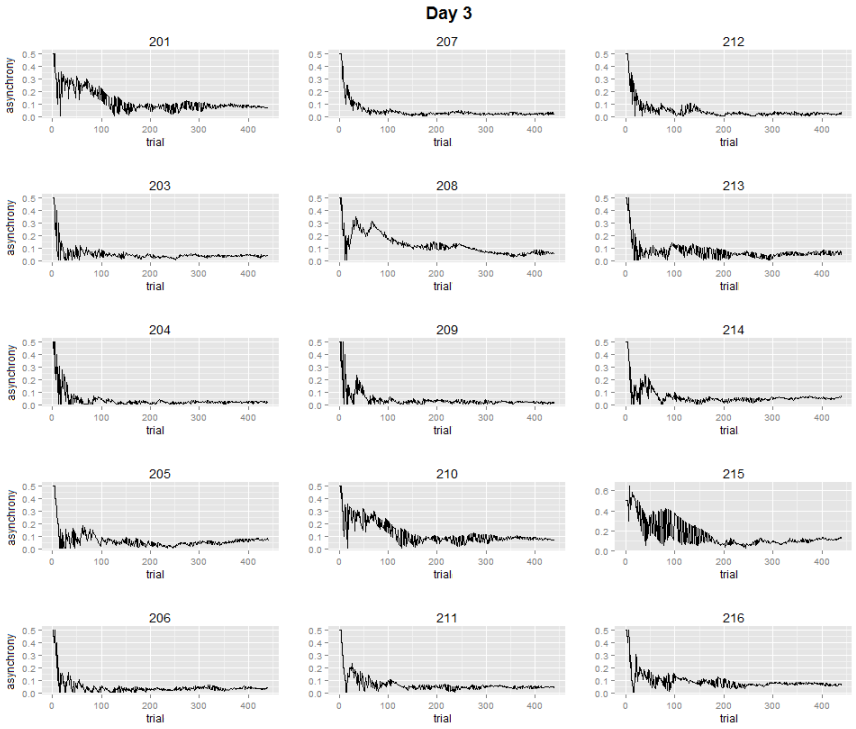


Appendix A. Individual graphs for the participants in Experiment 1 for all training days. On the y-axis is the absolute asynchrony of the stereo sound in ms. Participants whose ID is an odd number were always listening the syllables starting with dental /t/ from the right and the retroflex tokens from the left (and vice versa). For those participants group 'T1' during post-training was the retroflex consonant type and 'T2' was the dental.

Participant '120' was excluded because during the third training day all responses were 'left', which led to increasing asynchrony for the retroflex trials (the sounds coming from the right), suggesting no compliment with the instructions.

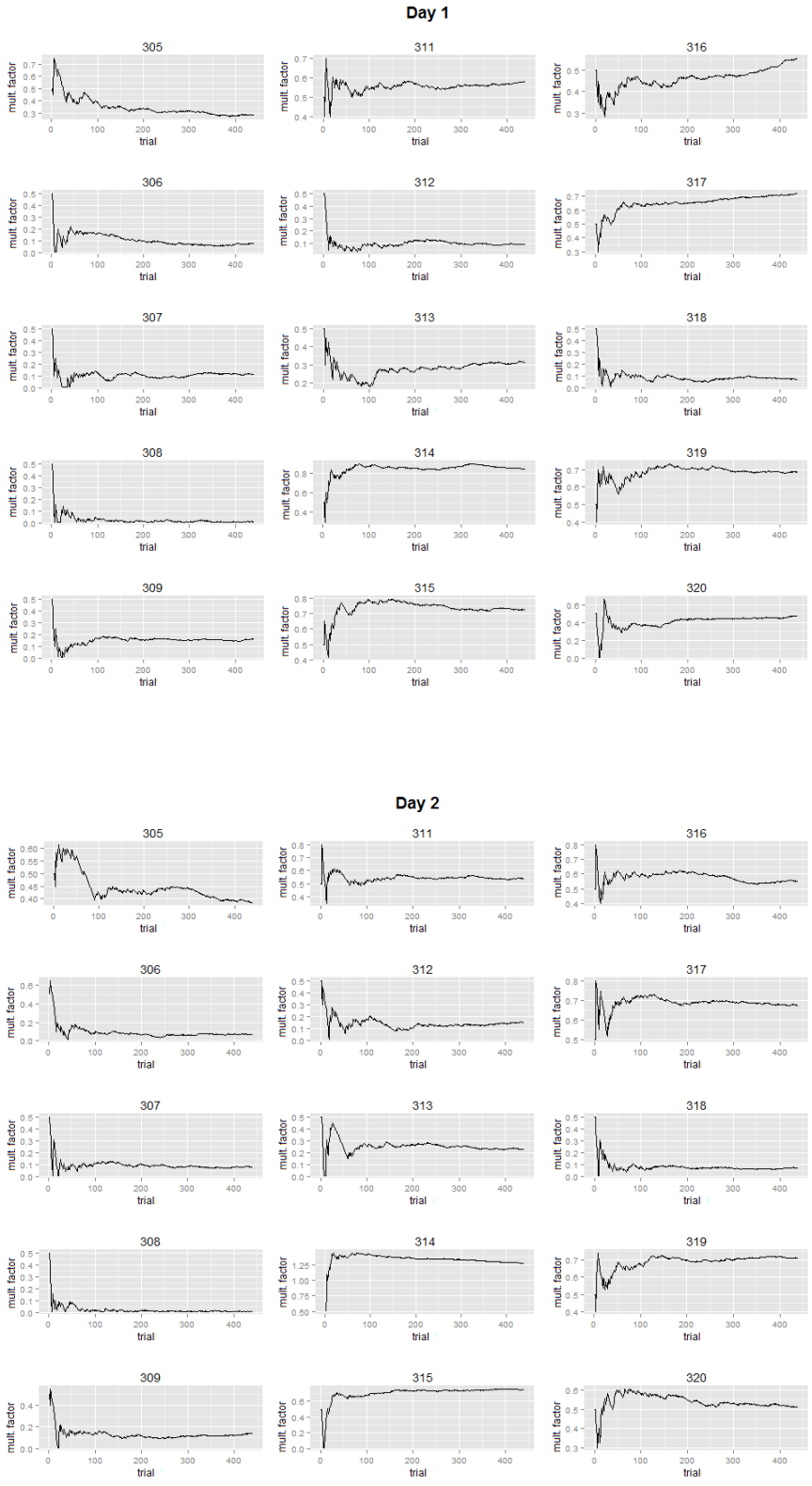
Appendix B

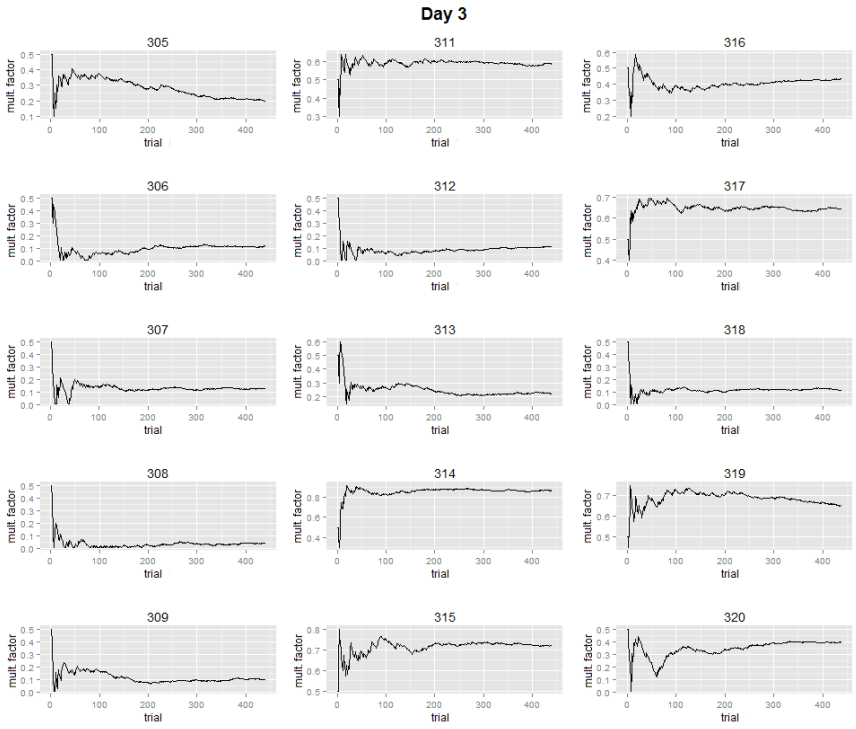

Day 1
201
207
212
203
208
213
204
209
214
205
210
215
206
211
216
asynchrony
trial
Day 2
201
207
212
203
208
213
204
209
214
205
210
215
206
211
216
asynchrony
trial



Appendix B. Individual graphs for the participants in Experiment 2 (Random direction of each consonant type) for all training days. On the y-axis is the absolute asynchrony of the stereo sound in ms. For participant '211' the last 180 trials were excluded from training data analysis because asynchrony values steadily increased (possibly due to fatigue).

Appendix C


Day 1
305
311
316
306
312
317
307
313
318
308
314
319
309
315
320
mult factor
trial
Day 2
305
311
316
306
312
317
307
313
318
308
314
319
309
315
320
mult factor
trial



Appendix C. Individual graphs for the participants in Experiment 1 for all training days. On the y-axis is the absolute asynchrony of the stereo sound in ms. Participants whose ID is an odd number were always listening the syllables starting with dental /t/ from the right and the retroflex tokens from the left (and vice versa). For those participants group 'T1' during post-training was the retroflex consonant type and 'T2' was the dental.

Participant '120' was excluded because during the third training day all responses were 'left', which led to increasing asynchrony for the retroflex trials (the sounds coming from the right), suggesting no compliment with the instructions.

Appendix D

Experiment 1	**Session**	**1**	**2**	**3**	**4**	**5**	**6**
NORMAL	**Dental**	0.14	0.17	0.15	0.17	0.16	0.18
	Rertroflex	0.17	0.20	0.16	0.16	0.14	0.15
PROBE	**Dental**	0.42	0.42	0.46	0.50	0.58	0.54
	Rertroflex	0.57	0.58	0.45	0.46	0.50	0.49

Experiment 2	**Session**	**1**	**2**	**3**	**4**	**5**	**6**
NORMAL	**Dental**	0.21	0.24	0.20	0.29	0.20	0.26
	Rertroflex	0.22	0.23	0.21	0.29	0.20	0.25

Experiment 3	**Session**	**1**	**2**	**3**	**4**	**5**	**6**
NORMAL	**Dental**	0.26	0.23	0.21	0.22	0.24	0.24
	Rertroflex	0.25	0.30	0.27	0.25	0.25	0.26
PROBE	**Dental**	0.39	0.37	0.37	0.36	0.36	0.43
	Rertroflex	0.54	0.50	0.61	0.61	0.58	0.58

TableD1. Error proportion for normal and probe trials presented by phoneme type through the 6 sessions of training. There were 2 sessions in each training day consisted each of 220 trials (200 normal and 20 probe ones). For Experiments 1 and 3 a probe trial was considered correct if the phonetic category of a given token was in agreement with the direction/intensity of the rest of the normal trials from this category respectively. For Experiment 2 there are no error proportions computed for the probe trials, since there was no correct response regarding either the explicit task or the phonetic category.

Appendix E

ASA step	Multiplication factor (audioop)	Sign	RMS (Pa)	dB (absolute value)	Intensity difference (dB)
0 (basic tone)	**1.000 (basic tone)**		**0.141589716**	**77**	**0**
0.700	1.700	+	0.240696172	81.61	4.61
	0.300	-	0.0424692021	66.54	10.46
0.600	1.600	+	0.226537968	81.08	4.08
	0.400	-	0.0566287092	69.04	7.96
0.500	**1.500**	+	0.212381837	80.52	3.52
	0.500	-	0.0707921213	70.98	6.02
0.400	1.400	+	0.198218228	79.92	2.92
	0.600	-	0.084948252	72.56	4.44
0.300	1.300	+	0.184058917	79.28	2.28
	0.700	-	0.0991063606	73.9	3.1
0.200	1.200	+	0.169902431	78.58	1.58
	0.800	-	0.113264246	75.06	1.94
0.100	1.100	+	0.155742453	77.83	0.83
	0.900	-	0.127422921	76.08	0.92

TableE1. Presents values of the multiplication factor computed through the adaptive process, the Root-Mean-Square of each stereo audio file created (RMS of the basic tone multiplied with the multiplication factor) and the intensity difference (in dB) to which it translates. Initial step of the adaptive process was 0.500. Audio file information presented here are taken from Praat.

Analysis

As we see in *Table E1* the step computed through the adaptive procedure translates to different dB levels if the audio file of a given value was of higher or lower intensity. We will examine here if this difference led to different performance (and therefore there was a different effect of learning) in the Identification task for the trained voice, in which significant effect of

training was found. In Experiment 3 there were two subgroups of training, one with pairing of dental syllables with higher intensity and retroflex syllables with lower intensity (8 out of 14 participants). The remaining 6 participants had the opposite pairing, i.e. retroflex syllables with higher intensity and dental syllables with lower intensity. First, we wanted to examine if performance on the identification task (mirrored in error proportion) was different between the two phoneme types (dental and retroflex) for each one of these subgroups. We regressed response ("correct" or "incorrect") onto phoneme type ("dental" or "retroflex") and trial type ("normal" or "probe") with participants and tokens as random factor with a model of the form:

```
response ~ phontype*trialtype + (1|subject) + (1|token).
```

There was no significant effect of phoneme type ($\beta=0.32$, $z=1.26$, $p=0.209$) as well as trial type ($\beta=-0.13$, $z=0.95$, $p=0.49$) on performance for the subgroup with dental syllables of higher intensity. Non significance was also found for the subgroup with dental syllables of lower intensity ($\beta=0.07$, $z=-1.79$, $p=0.65$ for the effect of phoneme type, $\beta=-0.21$, $z=1.78$, $p=0.073$ for the effect of trial type), suggesting that the difference in dB did not affect the overall performance in the identification task.

Second, we wanted to examine across group effect of training on the identification of the phonetic contrast, with a model of the form:

```
response ~ PhonType*subgroup + (1|subject) + (1|token),
```

with subgroups being "DentalHigh" or "DentalLow". Results showed no significant effect of subgroup ($\beta=0.46$, $z=1.54$, $p=0.124$), as well as no significant interaction between phoneme type and subgroup ($\beta=0.13$, $z=1.07$, $p=0.284$), suggesting that both subgroups had the same improvement of performance in the identification task compared to the naive participants.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. doi:10.1016/j.jml.2007.12.005

Barca, L. & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *Proceedings of the National Academy of Sciences, 7(4):* e35932. doi:10.1371/journal.pone.0035932

Bates D, Maechler M, Bolker B and Walker S (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In O. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning* (pp. 13-34). Philadelphia, PA: John Benjamins.

Boersma, P. (2001). Praat, a system of doing phonetics by computer. *Glot International, 5*, 341-345.

Bradlow, A. (2008). Training non-native language sound patterns. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition (pp. 287-308)*. Philadelphia: PA: John Benjamins.

Choi, H., Seitz, A. R., & Watanabe, T. (2009). When attention interrupts learning: Inhibitory effects of attention on TIPL. Vision Research, 49, 2586-2590. doi:10.1016/j.visres.2009.07.004

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*, 447–456. doi:10.1016/j.jml.2007.11.004

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303-306.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and

perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production, differences and similarities* (pp. 319-335). Berlin: Mouton de Druyter.

Fox, R.A., Flege, J.E., & Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of the Acoustical Society of America 97, 4,* 2540-2550.

Freeman, J.B. (2013). Abrupt category shifts during real-time person perception. *Psychonomic Bulletin & Review,* 21, 85-92. doi:10.3758/s13423-013-0470-8

Freeman, J.B. & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods, 42*, 226-241. doi:10.3758/BRM.42.1.226

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2(59)*. doi:10.3389/fpsyg.2011.00059

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the Part: Social Status Cues Shape Race Perception. *Proceedings of the National Academy of Sciences, 6(9)*: e25107. doi:10.1371/journal.pone.0025107

Iverson, P., Kuhl, P. K., Akahane-Yamada, R. Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* B47-B57. doi:10.1016/S0010-0277(02)00198-1

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446. doi:10.1016/j.jml.2007.11.007

Kesten, H. (1958). Accelerated stochastic approximation. *Annals of Mathematical Statistics, 29,*

41-59.

Kondaurova, M.V., & Francis, A., L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. Journal of Phonetics, 38, 569-587. doi:10.1016/j.wocn.2010.08.003

Krestar, M. L., Incera, S., & McLennan, C. T. (2013). Using mouse-tracking to examine the time course of an auditory lexical decision task. *The Ohio Psychologist, 60,* 29-32.

Leclercq, V., & Seitz, A. R. (2012). The impact of orienting attention in fast task-irrelevant perceptual learning. *Attention, Perception and Psychophysics, 74*, 648-60. doi:10.3758/s13414-012-0270-7.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242-1255. doi:10.1121/1.408177

McCandliss, B., Conway, M., Protopapas, A., & McClelland, J. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2,* 89-108. doi:10.3758/CABN.2.2.89

Protopapas, A, & Calhoun, B. (2000). Adaptive phonetic training for second language learners. In P. Delcloque (Ed.), *Integrating speech technology in language learning and the assistive interface (Proceedings of Instil 2000)* (pp. 31-38). Dundee, UK: InSTIL Publications.

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition, 115*, 435-43. doi:10.1016/j.cognition.2010.03.004

Seitz, A. R., & Watanabe, T. (2005). A unified model of perceptual learning. *Trends in Cognitive Sciences, 9,* 329-334. doi:10.1016/j.tics.2005.05.010

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research, Vol. 49,* No. 21, 2604-2610. doi:10.1016/j.visres.2009.08.003

Skottun B.C., Shackleton T.M., Arnott R.H., & Palmer A.R. (2001). The ability of inferior colliculus neurons to signal differences in interaural delay. *Proceedings of the National Academy of Sciences, USA, 98,* 14050–14054. doi:10.1073/pnas.241513998

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102*, 10393-10398. doi:10.1073/pnas.0503903102

Tower-Richardi, S. M., Brunyé, T. T., Gagnon, S. A., Mahoney, C. R., & Taylor, H. A. (2012). Abstract spatial concept priming dynamically influences real-world actions. *Frontiers in Psychology, 3:361*. doi:10.3389/fpsyg.2012.00361

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*, 2503-2522.

Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the relevant feature is weak. *Current Biology, 18,* 516-517. doi*:*10.1016/j.cub.2008.04.029

Vlahou, E., Protopapas, A., & Seitz, A. R. (2012). Implicit training of non-native speech stimuli. *Journal of Experimental Psychology: General 141 (2)*, 363–381. doi:10.1037/a0025014

Watanabe, T., Náñez, J.E., & Sasaki, Y. (2001). Perceptual learning without perception. Nature, 413, 844-848. doi:10.1038/35101601

Werker J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. Journal of the Acoustical Society of America, 75, 1866-1878.

Wilson, S. W., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in

producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage, 33,* 316-325. doi:10.1016/j.neuroimage.2006.05.03