Beiträge zur allgemeinen und vergleichenden Sprachwissenschaft. Band 14
DOI: 10.23817/bzspr.14-2

Dominika Chojnacka (ORCID 0009-0003-2223-602X)
Rijksuniversiteit Groningen, the Netherlands

Statistical learning of tones and syllables in non-tonal speakers[1]

Abstract

Statistical learning is a cognitive mechanism that enables humans to identify and learn patterns from sensory input. Language acquisition is one of the processes influenced by statistical learning. This study investigates whether non-native speakers can simultaneously learn different aspects of a new language, specifically focusing on tones and syllables. With the aim of investigating the learning rate of tones and syllables, 26 non-tonal speakers were exposed to Mandarin tone-syllable combinations with varying frequencies. Although the results of the statistical analysis did not show that the difference between syllable and tonal learning was statistically significant, the raw results suggest that non-tonal speakers may be unable to learn tones through statistical learning. This finding contradicts previous research in the field and highlights the need for further investigation.

Keywords: statistical learning, tone perception, non-tonal speakers, language acquisition.

1. Introduction

In this paper, we investigate whether non-tonal language speakers can acquire both syllables and tones through statistical learning. Statistical learning refers to "learning on the basis of some aspect of the statistical structure of elements of the input, primarily their frequency, variability,

[1] This article is based on the author's Bachelor's thesis written under the supervision of Dr. Stephen Jones and PhD student Mi Tang at the University of Groningen.

distribution, and co-occurrence probability" (Erickson/Thiessen 2015). It enables cognitive systems to uncover underlying structures and distributional properties from input (Hoffmann/Hoffmann 2016, Frost/Armstrong/Siegelman/Christiansen 2015). Even though this mechanism applies to a number of different processes happening in humans, so far it has been investigated mainly in the context of human perception and cognition. In our paper, we apply statistical learning to language acquisition.

The paper is organised as follows. In section 2, we present the theoretical background, including the basics of statistical learning, the description of tonal languages, tone perception and its place in language acquisition. We also review previous research on the statistical learning of tonal contrasts by non-tonal language speakers. Finally, in section 2.5, we formulate the initial hypotheses and briefly review the related work. Section 3 describes the design, procedure, and implementation of an experiment conducted to extend existing research in this area. In section 4, the outcomes of the experiment are discussed, including statistical analyses that assess participants' ability to learn tones and syllables over time under different test conditions. This section focuses on patterns in accuracy, learning curves, and the relative difficulty of acquiring tones versus syllables. Finally, in section 5, we interpret these findings in light of current theoretical frameworks and consider their implications for our understanding of statistical learning and language acquisition.

2. Theoretical background

2.1. Statistical learning

This section introduces the theoretical framework for the study. We define key concepts such as statistical learning, tone perception, and language acquisition, and review relevant theories and previous findings, including the debate between statistical learning and generative grammar.

Humans possess extensive knowledge and abilities that go beyond what they have explicitly learned. For example, we can predict the rain forecast simply by looking at the clouds or guess someone's emotions based on their body language, although we have never studied it. This is caused by constantly receiving and processing vast amounts of sensory input from the environment. Despite our limited cognitive capacity, we make sense of information through patterns and connections that our brains identify. Several innate mechanisms enable humans to process and interpret sensory information efficiently (Whittlesea/Wright 1997). These mechanisms allow

us to make predictions and adapt to new situations with limited explicit instruction (Whittlesea/Wright 1997). Statistical learning is an example of this type of mechanism.

Simply speaking, by observing and analysing the world around us, we can track relationships in the world and predict future events. For instance, over time, people may unconsciously associate dark clouds with rain. We learn this relationship through repeated exposure without explicitly trying to memorise it (Brady/Oliva 2008).

Statistical learning is a domain-general cognitive mechanism (Thiessen 2011). Kirkham/Slemmer/Johnson (2002), meaning it can be applied to many different tasks, like recognizing faces or understanding social cues. It operates in a similar way across all the different areas (Frost/Armstrong/Siegelman/Christiansen 2015). For example, visual statistical learning allows humans to unconsciously learn statistical relationships between visual objects (Fiser/Aslin 2002), while auditory statistical learning is crucial for language acquisition (Seidenberg 1997).

Statistical learning is crucial because it operates subconsciously (Schapiro/Turk-Browne 2015). This means that our brains are constantly and effortlessly picking up on patterns and regularities from the environment without us being aware of them. This automatic nature allows us to efficiently process large amounts of information and adapt to new situations quickly. For instance, we can recognise familiar faces in a crowd without consciously thinking about it. Statistical learning enables us to navigate and understand the world more effectively, showing its importance in our daily lives.

Another reason for the importance of statistical learning in our daily functioning is its age invariance, meaning that children and adults can learn new patterns equally well. Arnon (2019) showed that this ability fully develops in young childhood and is present throughout our lives. This age-invariant nature of statistical learning highlights its fundamental role in human cognition, allowing us to continuously acquire new knowledge and adapt to changes at any stage of life.

2.2. Tones and tonal languages

A whisper is characterised by a soft and gentle tone, while a scream is high-pitched and sharp. These examples demonstrate how pitch influences how sound is perceived and interpreted. Pitch is created from the vibra-

tions of the vocal cords when speech is produced. The vibrations can have different frequencies (also known as fundamental frequencies) (Gussenhoven 2004). These frequencies determine the pitch level: the higher the frequency, the higher the pitch. Most languages use changes in pitch to convey meaning in different ways. The most common ways are stress and intonation, which is changing pitch to emphasise a certain part of a word or a sentence to convey meaning, indicate focus, or express emotion (Gussenhoven 2004). Specifically, stress can be used to highlight a part of a sentence or word, while intonation helps to convey different meanings and to highlight phrases (Gussenhoven 2004). For example, a rising pitch at the end of a sentence might indicate a question, while a falling pitch can indicate a statement or conclusion.

There is also a unique type of change in pitch, which is not present in all languages - tone. Tonal languages use pitch changes to distinguish words (Gussenhoven 2004). An example of a tonal language is Mandarin, which has four different tonal patterns (first tone: flat, second tone: rising, third tone: low-dipping, and fourth tone: falling). These tonal patterns are presented respectively in Figure 1. Depending on the tonal pattern, the same syllable, *ma* can have four distinguishable tone-syllable combinations with different meanings: *ma* + flat tone (ma1) = mother; *ma* + rising tone (ma2) = numb; *ma* + low-dipping tone (ma3) = horse; *ma* + falling tone (ma4) = to scold.

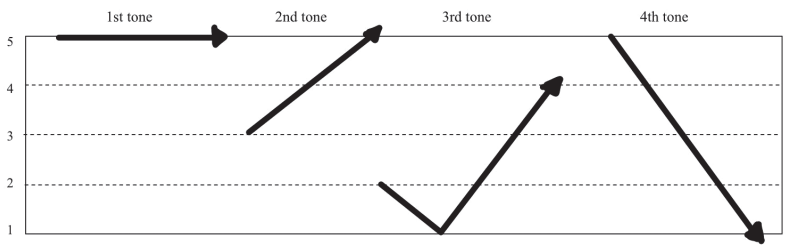

Figure 1: Four tones in Mandarin

2.3. Language acquisition

Infants and adults acquire languages in fundamentally different ways: infants rely solely on statistical learning, while adults can be influenced by the cumulative knowledge of all the languages previously acquired (Saffran/Aslin/Newport 1996, Flynn/Foley/Vinnitskaya 2004). Saffran/Aslin/Newport (1996) analysed how the acquisition of a native language differs

from the acquisition of a second, third, or subsequent language. They concluded that infants use statistical learning to acquire their native language. Through exposure to a language, infants can learn it unconsciously by detecting patterns within the speech they hear. This process is quite broad and involves several aspects of language acquisition. For instance, infants become more sensitive to frequently used syllables and words, develop the ability to segment continuous speech into distinct words, and learn to predict upcoming sounds based on their statistical regularities. Curtin/Zamuner (2014) explained how infants are capable of learning about syllables, rhythm, stress, sound distributions, and combinations, and much more simultaneously by experiencing the environment. They also mention an important characteristic of native language acquisition: initially, infants can distinguish speech sounds from any language. However, over time, they focus more on the sounds and contrast specific to their native language. This was also investigated by Colantoni/Steele/Neyra (2015), who established that second language learners struggle to distinguish novel sounds not encountered in their native language. This tendency makes the learning process of a new language considerably slower. At the same time, Flynn/Foley/Vinnitskaya (2004) showed that adults also rely on cumulative knowledge of previously known languages rather than just using statistical learning. Therefore, it differs from infants who depend on universal rules and early language acquisition stages, suggesting that an adult person who has encountered a tonal language will be able to distinguish tones and, consequently, learn a tonal language faster than a person who has never encountered tones.

These differences raise an intriguing question. Although statistical learning is an unconscious process that continues throughout life, irrespective of age or intent, is it influenced by the cumulative knowledge acquired by adults throughout their lives?

One possible explanation is that prior knowledge acts as a filter when processing new sounds. For example, speakers of non-tonal languages often struggle to distinguish between tones in tonal languages. The critical question is, then, whether this filtering happens before or after statistical learning occurs. If the filter is applied after statistical learning, adults could theoretically learn to differentiate tones simply through exposure, even if they initially struggle to perceive them. However, if the filter is applied before statistical learning, adults would be unable to learn tones through exposure alone, as these sounds would be filtered out from the outset. In essence, the order of processes that are involved in the perception and

learning of language will be investigated, as it determines the mechanisms used in language acquisition.

2.4. Statistical learning vs. generative theory

As mentioned earlier, language acquisition through statistical learning refers to the ability to learn patterns and regularities of a language from the surrounding environment. However, it is important to note that statistical learning is not the only theory of language acquisition. Generative theory argues that environmental input is too limited to account for the rapid language acquisition (Chomsky 1978). It posits that all human languages share fundamental similarities, known as Universal Grammar, which are considered to be innate (Chomsky 1975, Crain/Thornton/Murasugi 2009). This perspective is further supported by the Poverty of the Stimulus argument, which claims that children exhibit knowledge of language that exceeds what could be learned from the input alone (Crain/Lilo Martin 1999). Although the generative theory was initially described as "unassailable" (Smith/Tsimpli 1995), it has since faced criticism due to the lack of conclusive empirical evidence (Dąbrowska 2015). In contrast, the theory of statistical learning challenges generative assumptions by proposing that language structures can emerge simply through exposure to input. Even though this paper is grounded in statistical learning, it involves broader debates on language acquisition. More specifically, whether domain-general learning mechanisms are sufficient or whether innate, domain-specific constraints are necessary.

2.5. Related work

Wiener/Ito/Speer (2021) investigated whether adult listeners unfamiliar with tonal languages can learn syllable-tone statistical regularities through exposure and statistical learning. The study involved 80 American participants with no prior experience with tonal languages. Over four consecutive days, which allowed for overnight knowledge consolidation, participants were exposed to a series of sounds from an artificial language resembling Mandarin. Different syllable frequencies and tonal probabilities were used to test if participants were able to discriminate which sound they heard.

Participants performed a series of tasks each day, including listening to the sounds, repeating them, naming critical items, and selecting symbols corresponding to the sounds they heard. Feedback was provided for the

last task after each response to help with learning. Results indicated that participants obtained above-chance scores on all four days, with recognition performance improving over time. This improvement suggests that adults can learn syllable-tone regularities through statistical learning. Even though the accuracy peaked on the fourth day, the fact that learners continued to make frequent tonal mistakes suggests that the extent to which they learned the tonal patterns remains unclear. Moreover, the role of feedback provided to participants makes it uncertain whether the learning was primarily driven by statistical learning or error-driven learning, which involves noticing mistakes and adjusting responses based on those errors, rather than learning patterns (Nixon 2020).

To address the limitations of the study by Wiener/Ito/Speer (2021) and eliminate error-driven learning, Tang/Spenader/Jones (2024) conducted two experiments. In Experiment 1, four syllables and four tones were used to create 16 different syllable-tone combinations. During the training phase, participants were exposed to half of these combinations (referred to as "legal" items) over three consecutive days. After each day, participants were tested on all 16 original combinations, along with an additional eight combinations created from two new syllables. The sounds not presented during the training phase were termed "illegal" items. The aim was to determine whether participants could accurately assess whether a sound had been encountered during the training phase. For the original 16 items, participants needed to recognise both the tone and the syllable to respond correctly, whereas for the 8 new items, identifying the syllable alone was sufficient. The results indicated that participants assessed legal and illegal items with familiar syllables as being heard during the training phase, but frequently rejected illegal items with new syllables. This pattern suggests that participants primarily learned the syllables rather than the conditional tone-syllable patterns, providing no conclusive evidence for tone acquisition through statistical learning.

The introduction of new syllables appeared to facilitate syllable learning rather than lexical tone learning, leading to the design of a second experiment to investigate these findings further. In this experiment, Tang/Spenader/Jones (2024) aimed to control syllable learning while eliminating the influence of Wiener's error-driven learning model. Participants were exposed to 16 syllable-tone combinations, created from four distinct tones and four syllables. Among these, four sounds appeared with notably higher frequency, each combining a specific syllable with a unique tone. During the training phase, all sounds were presented to the participants. In

the testing phase, pairs consisting of a high-frequency sound and a regular sound were presented, and participants were asked to identify which sound was more familiar. The experiment was also conducted over four days to facilitate the consolidation of the language. Results showed above-chance accuracy from the first day, with accuracy increasing throughout the training period, indicating that participants effectively learned the tone-syllable patterns. The results confirmed Wiener's findings while eliminating error-driven learning, demonstrating that participants were able to learn the tones through statistical learning.

2.5.1. The present study

Wiener/Ito/Speer (2021) and Tang/Spenader/Jones (2024) demonstrated that people, including non-tonal speakers, are capable of learning elements of language previously unfamiliar to them, such as tones. Tang/Spenader/Jones (2024) investigated this learning ability by assigning four tone-syllable high-frequency sounds, while retaining 12 low-frequency sounds. A significant limitation in the design of the experiment is that it only focused on tone-syllable combinations. No isolated syllables or tones had varying frequency. As a result, it remains unclear how well individuals can learn tones or syllables in isolation, and the speed at which these distinct learning processes occur. This gap in the research leaves unanswered questions about the specific mechanisms and rates of learning for different elements. To investigate the contribution of statistical learning to the adaptation to various elements of language, the present paper poses the following question: Are novel and non-novel aspects of sound in a new language learned simultaneously?

2.5.2. Hypothesis

Based on the findings of Wiener/Ito/Speer (2021) and Tang/Spenader/Jones (2024), which demonstrated that tone-syllable combinations can be learned statistically, I expect that participants will be able to acquire both tones and syllables over time. Additionally, as shown by Colantoni/Steele/Neyra (2015), distinguishing novel elements, such as unfamiliar tones, poses a greater challenge. Therefore, tones, being less familiar than syllables for non-tonal speakers, are likely to be more difficult to learn. This should result in a longer learning time for tones compared to syllables. Furthermore, participants are expected to show a higher initial accuracy in recognising syllables than tones throughout the experiment.

We use two types of test cases in the experiment to allow the discrimination of the learning rate between two different dimensions. The Distractive-Tone case will use tones as distractors to determine the learning rate of syllables, while the Distractive-Syllable case will use syllables as distractors to determine the learning rate of tones using statistical learning. Based on the hypothesis, two main effects are anticipated. First, it is expected that participants will be able to learn both tones and syllables. To support this, the percentage of correct responses in both Distractive-Tone and Distractive-Syllable conditions should increase significantly over time. Second, it is hypothesised that learning tones will be more challenging than learning syllables. Specifically, participants are expected to show slower progress in the Distractive-Syllable condition compared to the Distractive-Tone condition, as syllables are more familiar elements of their native language and therefore more likely to be distracting. In contrast, tones, being novel elements, should be less distracting. We accept this hypothesis if the percentage of correct answers in the Distractive-Syllable condition shows a significantly slower rate of increase compared to the Distractive-Tone condition, and if the initial percentage of correct answers in the Distractive-Syllable condition is lower.

3. Methods

In this section, we describe the experimental design used to assess statistical learning of tones and syllables. We outline the participant demographics, procedure, stimuli, and test structure in detail to ensure replicability and transparency.

In the second experiment of their paper, Tang/Spenader/Jones (2024) demonstrated that non-tonal language speakers can learn tone-syllable combinations through exposure. Participants were exposed to various tone-syllable combinations presented at different frequencies and were then asked to identify the combinations that occurred more frequently. In the present study, the variation in tones was independent of the variation in syllables, allowing us to investigate the learning of each element separately.

3.1. Participants

A total of 26 students of the University of Groningen participated in the experiment (median age: 21; range: 18-26; 15 male, and 11 female). All participants had no experience with tonal languages and reported no hear-

ing or language deficits. Participants came from 11 different countries and exhibited considerable linguistic diversity. In total, they reported speaking 10 different native non-tonal languages, along with 13 additional languages acquired later in life. All participants signed a consent form before taking part in the study and were allowed to withdraw at any time, either during or after the experiment. Finally, upon completing the full study, participants received compensation, consisting of a base payment of 15 euros plus an additional performance-based accuracy bonus.

3.2. Procedure

The experimental procedure followed the same design as in Tang/Spenader/Jones (2024). Participants took part in four on-site sessions over four consecutive days. Each session consisted of a 10-minute training phase followed by a 5-minute test phase, conducted using OpenSesame (version 3) (Mathot/Schreij/Theeuwes 2012).

The training phase was introduced as a counting task in which participants listened to speech stimuli composed of Mandarin syllables and tones. The stimuli were randomly interrupted with beep sounds. The participants were instructed to count and record the number of beeps they heard. This task was introduced to ensure that the participants were paying attention, while also preventing them from explicitly focusing on the sounds they heard. To minimise fatigue, the set of 600 stimuli that were played to participants on a study day was divided into five training blocks. After the five training blocks, the test phase started.

During the test phase, participants were presented with 32 pairs of tone-syllable combinations, with 16 pairs spoken by a female speaker and the remaining 16 by a male speaker. This variation was introduced to check if participants could generalise their learning in different voices of speakers. They were asked to identify which sound in each pair had been presented more frequently during the training phase. They had 5 seconds to respond to each test question, following the parameters used in the experiment by Tang/Spenader/Jones (2024). After completing the test phase, participants were shown their accuracy scores for both the counting task and the test phase. A diagram of the procedure is presented in Figure 2.

Data for each participant were stored in four separate CSV files: one file for each day of the experiment. Each participant received a unique ID number to ensure anonymity. The experiment was conducted on a laptop, and the sound was played using headphones to avoid possible distractions.

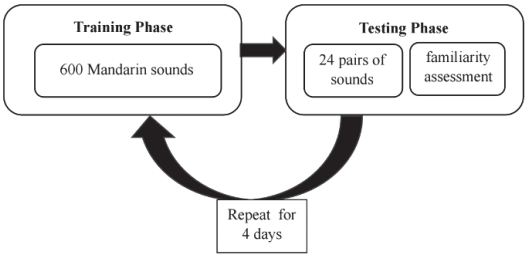

Figure 2: A diagram of the procedure

3.3. Stimuli

Sixteen tone-syllable combinations were used in the experiment, created by combining four Mandarin tones (flat [tone1], rising [tone2], low-dipping [tone3], falling [tone4]) with four syllables (/ge/, /bi, /du/, /kou/). The stimuli were created for Tang/Spenader/Jones (2024), and the same stimuli were used in this experiment. The recordings were done by a voice actor and actress. "The recording was done in a studio in China, using an Audio-Technica AT2020 microphone. All of the files were normalised and processed with Adobe Audition CS6: the sounds were set to an average of 68.92 (standard deviation: 4.79) dB SPL and an average of 416.88 (standard deviation: 111.34) ms in duration" (Tang/Spenader/Jones 2024).

3.4. Training set

All 16 combinations of the tone-syllable combination were used in the training phase. The number of occurrences was calculated using the following calculations.

3.4.1. Frequency of tones and syllables

Each training phase consisted of 600 sounds, consistent with the original experiment (Tang/Spenader/Jones 2024). Two syllables and two tones were assigned as High-Frequency (HF), while the remaining two syllables and two tones were set as Low-Frequency (LF). HF sounds were intended to occur significantly more often than LF sounds, but the frequency of LF sounds had to be sufficient to be noticeable within each training block. Consequently, each HF item accounted for 40% of the total sound count (240 occurrences each), while each LF item made up 10% (60 occurrences

each). Table 1 shows the frequency of occurrence for tones and syllables, with LF items highlighted in light grey and HF items in black.

3.4.2. Frequency of tone-syllable combinations

Among the 16 tone-syllable combinations, two were designated as medium-frequency, and another two as high-frequency, ensuring that each isolated syllable and each isolated tone appeared with exactly one medium or high-frequency item. Their frequency equalled 90% of the total count of the isolated tone or syllable that they are made out of, resulting in a frequency of 216. Similarly, the frequency of each medium-frequency item equalled 90% of the total count of the isolated tone or syllable they are made out of, resulting in a frequency of 54. The remaining tone-syllable combinations were categorised as low-frequency items, with frequencies of either 2 or 20 occurrences.

Each block contained approximately the same number of medium and high-frequency sounds, while low-frequency sounds were distributed randomly.

	tone A	tone B	tone C	tone D	sum
syllable 1	216	2	2	20	240
syllable 2	2	2	54	2	60
syllable 3	20	2	2	216	240
syllable 4	2	54	2	2	60
sum	240	60	60	240	600

Table 1: Frequency count

All participants were selected for their non-tonal language backgrounds, meaning that in their native languages, tones do not convey syntactic information. However, tones can still serve to convey prosodic information (Li/Tang/Lu/Wu/Chang 2021). As a result, certain Mandarin tones might be more recognisable to non-tonal speakers. To address this potential bias, six variations of Table 1 were created by combining the available tones into every possible pairing of High-Frequency tones. All possible tone combinations are shown in Table 2.

Similarly, all syllables used in the experiment were unfamiliar to the participants, but some may have been easier to identify than others. To account for this variability, six variations of Table 1 were generated by combining the available syllables into every possible pairing of High-Frequency syllables. All syllable combinations are presented in Table 3.

Finally, the tone and syllable variations were combined, resulting in six different matrices with varying HF tones and HF syllables.

	tone A	tone B	tone C	tone D
matrix 1	tone1	tone2	tone3	tone 4
matrix 2	tone1	tone3	tone4	tone 2
matrix 3	tone2	tone1	tone4	tone 3
matrix 4	tone3	tone4	tone2	tone 1
matrix 5	tone4	tone3	tone1	tone 2
matrix 6	tone3	tone1	tone2	tone 4

Table 2: Tone variation of the matrices

	syllable 1	syllable 2	syllable 3	syllable 4
matrix 1	/ge/	/bi/	/du/	/kou/
matrix 2	/ge/	/du/	/bi/	/kou/
matrix 3	/ge/	/bi/	/kou/	/du/
matrix 4	/bi/	/ge/	/du/	/kou/
matrix 5	/bi/	/ge/	/kou/	/du/
matrix 6	/kou/	/bi/	/du/	/ge/

Table 3: Syllable variation of the matrices

3.5. Test set

During the test phase, participants were presented with 32 pairs of tone-syllable sounds. Their task was to assess which sound had been presented more frequently during the training phase. The test pairs were categorised into three types: Distractive-Syllable, Distractive-Tone, and Non-Distractive.

3.5.1. Distractive-Syllable

This type of test case consists of 8 pairs of sounds. Each pair includes one medium-frequency sound (items with the frequency of 54 in Table 1) and one low-frequency sound (items with the frequency of 2 in Table 1). For example, syllable 4 with tone B versus syllable 3 with tone C (look in Table 1). These test cases are referred to as Distractive-Syllable because the low-frequency sounds are composed of an HF syllable, which may distract participants from selecting the correct medium-frequency sound.

These test cases are designed to evaluate predictions derived from three hypotheses:

- Hypothesis 1: If tone and syllable contribute equally to the perception of sound, participants are more likely to choose the medium-frequency item, as its frequency is higher than that of the low-frequency item.
- Hypothesis 2: If participants cannot perceive tones and rely solely on syllable recognition, they are more likely to choose the low-frequency item, as its syllable frequency is higher than that of the medium-frequency item.
- Hypothesis 3: If participants can only perceive tones without recognising syllables, their choice will be random (50%), as the tone frequency of both the medium-frequency and low-frequency items is the same.

3.5.2. Distractive-Tone

Similar to the previous type, Distractive-Tone test cases consist of 8 pairs of sounds, each comparing a medium-frequency sound with a low-frequency sound. In this case, the low-frequency sound is composed of an HF tone, which serves to distract participants. For instance, syllable 4 with tone B versus syllable 4 with tone A.

These test cases are designed to test 3 hypotheses, which contrast with the hypothesis in Distractive-Syllable:

- Hypothesis 1: If participants can perceive both tones and syllables, they are more likely to choose the medium-frequency item. It aligns with Distractive-Syllables, where the ability to distinguish the features results in the same choice.
- Hypothesis 2: If participants cannot perceive syllables and rely solely on tonal recognition, they are more likely to choose the low-frequency item. In the Distractive-Syllable, the low-frequency item would be chosen in the contrasting situation when participants rely solely on syllables.
- Hypothesis 3: If participants can only perceive syllables without recognising tones, their choice would be random (50%). In Distractive-Syllable, this would be the case if participants relied only on tones.

3.5.3. Non-Distractive

The last type of test cases ensures an even distribution of syllables and tones within the test set. Distractive-Syllable and Distractive-Tone result in two syllables and two tones appearing with a frequency of 12, while the remaining two tones and syllables have a frequency of 4. To avoid bias from an uneven distribution, 16 additional pairs of sounds were included in the test set. This adjustment brings the total occurrences of each syllable and tone to 16 in the test set. These additional pairs are only used to balance the stimuli and do not contribute any further data for analysis.

The complete list of test cases for Table 1 is provided in Appendix A.

4. Results

This section presents the empirical findings of the study. We show how participants' performance evolved across four sessions, compare the effects of different distractor conditions, and analyse the data using statistical models to test our hypotheses.

4.1. Data visualisation

Figure 3 illustrates the variation in the percentage of correct answers for both Distractive-Tone and Distractive-Syllable over four days. The trends highlight contrasting patterns in participant performance between the two conditions.

In the Distractive-Syllable test cases, the initial percentage of correct answers was relatively low (around 30%), indicating that participants were significantly distracted by the high-frequency syllables. As exposure increased, the percentage of correct answers decreased further. This suggests that participants became progressively more affected by the distracting effect of the syllables.

In contrast, the Distractive-Tone test cases showed accuracy levels remaining around the chance level (41-48%) across all days. This trend implies that participants primarily relied on syllables rather than tones when making decisions, showing little evidence of statistical learning for tone. The lack of change in the percentage suggests that participants relied primarily on the syllables and were not acquiring tones through statistical learning.

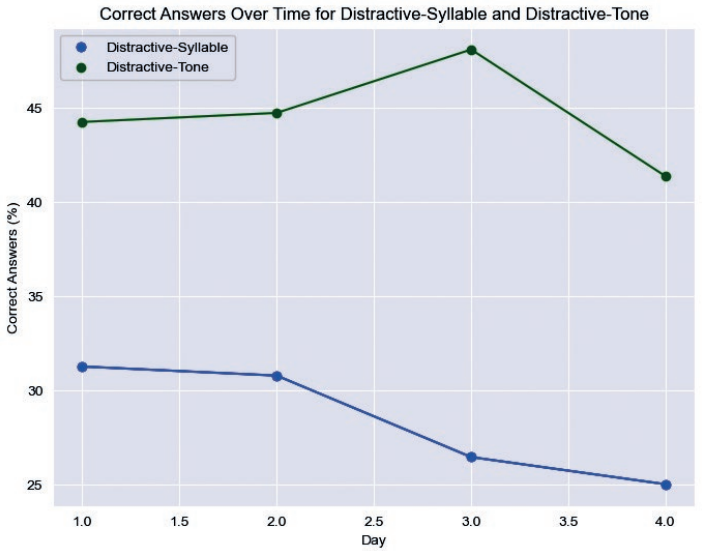

Figure 3: Percentage of correct answers over days

4.2. Statistical analysis

The visual representation showed that non-tonal speakers were learning syllable differences. Simultaneously, they were not learning tone differences, suggesting that it is impossible for native non-tone speakers to statistically learn tones. To provide more confidence about the inferences found in the raw data, statistical models were built. More specifically, Generalised Additive Models were constructed in order to test the differences between test conditions.

4.2.1. Introduction to GAMs

Generalised Additive Models (GAMs) (Hastie/Tibshirani 1987) are used to uncover linear or non-linear relationships between independent and dependent variables. The model builds upon an assumption that a function has an underlying additive structure. In other words, each function can be represented as an addition of multiple functions. It is a very flexible and universal model since it approximates the function by adding multiple individual components with a smoothing term (Xiang 2001). The general formula of a GAM model is shown in Equation 1.

$$\eta(x) = \alpha + s_1(x_1) + ... + s_p(x_p) \qquad \text{Equation 1}$$

In Equation 1, is the dependent variable, s_1, ... , s_p are smooth non-parametric functions, x_1, ... , x_p are predictors of $\eta(x)$, α is the intercept.

There is an important disadvantage of GAMs that has to be considered. Each smoothing term results in more degrees of freedom, which makes the models extremely flexible. The flexibility allows us to build models that learn complex patterns. However, if degrees of freedom are added carelessly, it might result in overfitting. The model could then find patterns that fit the data, which do not represent real-life patterns. Therefore, it is vital to add smoothing terms and random effects of variables reasonably. This means that before adding a new term to the equation, it must be carefully considered whether it represents a real phenomenon.

4.2.2. Base Model

The Base Model was built with the simplest possible structure, using |correct|, |distractor|, and |day| as the key variables, serving as a baseline for comparison. Any future models that do not exceed its performance will be rejected, as it would suggest that the simplest model performs better. This model checks the relationship between the correctness of answers based on the distraction type and days and is given by Equation 2.

$$\text{correct} \sim \text{distractor} + s(\text{day}, \text{by} = \text{distractor}, \text{bs} = \text{'fs'}) \qquad \text{Equation 2}$$

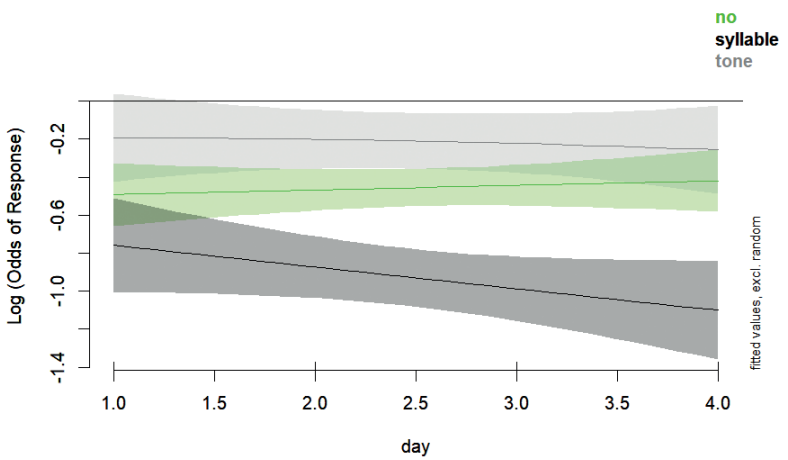

Figure 4: Results from the Base Model

```
Parametric coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.45488    0.05031  -9.042  < 2e-16 ***
distractorsyllable -0.47503    0.09211  -5.157  2.5e-07 ***
distractortone      0.23765    0.08600   2.763  0.00572 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                              edf Ref.df Chi.sq p-value
s(day):distractorno       1.001  1.001  0.292  0.5899
s(day):distractorsyllable 1.001  1.002  2.718  0.0993 .
s(day):distractortone     1.087  1.167  0.078  0.8266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.014   Deviance explained =  1.2%
-ML = 2177.9  Scale est. = 1         n = 3328
```

Figure 5: Numerical results from the Base Model

The visual and numerical results from the Base Model are shown in Figure 4 and Figure 5, respectively. The results show that for all test cases, the probability of a correct response was significantly different from chance (log – odds $\neq 0$, $p < 0{,}55$). Additionally, the type of distractor influenced accuracy: the syllable distractor significantly decreased the probability of answering correctly ($p < 0{,}55$), while the tone distractor increased it ($p < 0{,}55$). The numerical results show that there were no changes in learning in any of the distractor types ($p < 0{,}55$ for all types). Lastly, the model explains only 1.2% of the deviance, suggesting that additional variables might be added in order to capture more complex relationships.

4.2.3. Model 2

The second GAM model was created to predict the correctness of an answer based on several predictors. The predictors include |distractor|, |day|, |participant|, |cohort|, and |item|. The |distractor| refers to the type of test case. As mentioned previously, there are three types of test cases: tone, syllable, and none. It is expected that different types of test cases result in varying trends in the correctness of answers. The distractor type is treated as the main predictor, modelled as a linear term, and it is assumed that the effect of |distractor| on |correct| is not affected by other predictors.

Next, three smooth terms were added to model non-linear relationships between the predictors and the correctness of answers. The first is |day|, which is modelled as a smooth term to capture the relationship between correctness and the day participants were tested. Importantly, it is assumed that random effects vary across individual participants, accounting for individual pre-

dispositions. Additionally, the effect of |day| depends on the distractor type, with potentially different effects for tones, syllables, and none.

The second smoothing term is |cohort|, which represents variability among different matrices, as explained in Section 3. Similar to the previous smooth term, this one also accounts for individual participant effects and separate effects for each distractor type.

Lastly, the predictor |item| accounts for variability in correctness due to specific items. For example, some items may be inherently more difficult to learn or recognise. The model is shown in Equation 3.

correct~distractor +
s(day, participant, by = distractor, bs = 'fs') +
s(cohort, participant, by = distractor, bs= 're') +
s(item, by = distractor, bs = 're') Equation 3

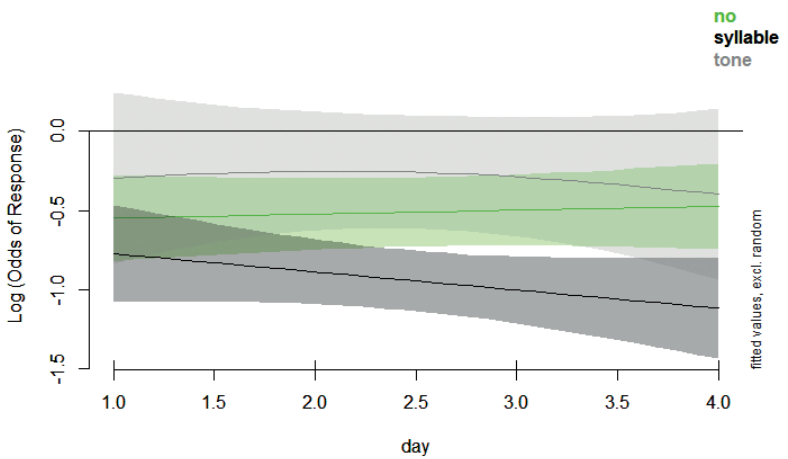

Figure 6: Results from the GAM Model

Figure 6 provides a visual representation of the results of the GAM model shown in Equation 3. The graph illustrates the relationship between days and the log odds of correctness, with separate smooth curves for different distractors (tone, syllable, and none).

The green and light grey curves (representing no distractor and tone distractor, respectively) are relatively flat, indicating minimal change over time. In contrast, the black curve (representing the syllable distractor)

decreases over time, suggesting that participants perform worse as time progresses with this distractor. The shaded areas represent confidence intervals, reflecting the uncertainty in the estimates. In some regions, the confidence intervals overlap, indicating that the differences between distractors are not statistically significant.

The following conclusions can be drawn from the statistical analysis. The confidence spread shows that for no distractor and syllable distractor, learning occurred from day 1. However, over time, there was a difference between both two conditions, and syllables were progressively becoming more distracting. For the tone distractor, learning remained at chance for all four days (log odds = 0). The numerical results are presented in Figure 7 and show that when the distractor is a syllable, it significantly reduces the log odds of correctness compared to the absence of a distractor ($p < 0.05$). However, the tone distractor does not differ significantly. For the smooth terms, there is no significant effect of days by distractor ($p < 0.05$), suggesting that there is no clear trend in correct answers over days within distractor groups. However, when accounting for the variability of individual participants, statistically significant effects are found for days by distractor in both syllables and tones ($p < 0.05$). This indicates that the impact of distractions varies across individuals, with both syllables and tones having meaningful effects over days. Additionally, a significant effect was found between cohorts by distractor in tonal distractors ($p < 0.05$), while accounting for individual differences of participants. Lastly,

```
Parametric coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.5325     0.1337  -3.982 6.83e-05 ***
distractorsyllable -0.4553     0.1953  -2.331   0.0198 *
distractortone      0.2348     0.2367   0.992   0.3212
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                          edf Ref.df  Chi.sq  p-value
s(day):distractorno                  1.000100  1.000   0.200  0.65505
s(day):distractorsyllable            1.000111  1.000   1.821  0.17721
s(day):distractortone                1.431035  1.675   0.656  0.71815
s(day,participant):distractorno     18.260227 51.000  80.016  0.03834 *
s(day,participant):distractorsyllable 17.765719 51.000  28.304  0.01242 *
s(day,participant):distractortone   33.148320 51.000 138.672  0.00461 **
s(cohort,participant):distractorno   9.522531 25.000  20.989  < 2e-16 ***
s(cohort,participant):distractorsyllable 0.007081 25.000   0.006  0.16752
s(cohort,participant):distractortone 5.714850 25.000   7.769  < 2e-16 ***
s(item):distractorno                46.082956 85.000  95.670  < 2e-16 ***
s(item):distractorsyllable          27.498552 47.000  64.637 4.38e-07 ***
s(item):distractortone              19.391694 44.000  33.848  0.00183 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.164   Deviance explained = 16.8%
```

Figure 7: Numerical results from the GAM Model

there was an effect on correctness from items for all distraction types ($p < 0.05$ for all distraction types). Despite these significant terms, the model explains only 16.8% of the data, which limits its predictive power.

4.2.4. Bonus Model

Since the deviance explained in both models was relatively low, a new model was created in order to see if it is possible to achieve a high deviance explained using the GAM model in this task. In this model, all variables collected in the experiment were added as smoothing terms, regardless of their real-world implications. It resulted in using the terms used in the Model 2, while adding |gender|, |gender_train|, and |response_time|. The variables |gender|, |gender_train|, and |response_time| indicate the gender that read a test case, the gender that read the whole training set, and the response time needed to respond, respectively. The full model is shown in Equation 4.

```
correct~distractor + s(day, participant, by = distractor, bs = 'fs') +
        s(cohort, participant, by = distractor, bs = 're') +
      s(gender, gender_train, by = distractor, bs = 're') +
        s(response_time, by = distractor, bs = 're') +
            s(item, by = distractor, bs = 're')
```

Equation 4

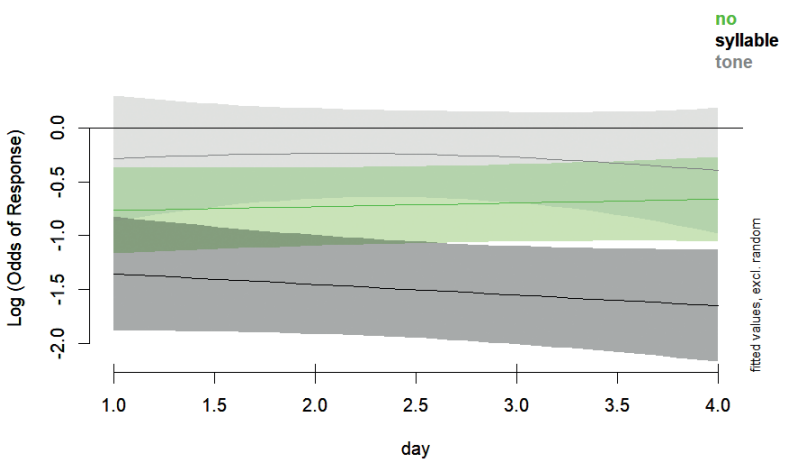

Figure 8: Results from the Bonus GAM Model

Figure 9 presents the visual results of the Bonus GAM Model. Notably, the addition of new terms increased uncertainty in the predictions, resulting in wider confidence intervals compared to the previous models. This causes the grey (tone distractor) and green (no distractor) areas to overlap around the 50% mark, suggesting a lack of clear distinction between them. Similarly to the previous model, there is a significant difference between the no distractor and syllable distractor ($p < 0.05$). Moreover, there is a significant relationship between correctness and days as well as correctness and cohort (while assuming individual differences of participants) for both syllables and tones. Furthermore, the model found other significant dependencies within the newly added variables, between the correctness and the response time when syllables are distractors ($p < 0.05$). Finally, the deviance explained equalled 17.9%, which is a slight improvement compared to Model 2.

```
Parametric coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.7139     0.1813  -3.937 8.25e-05 ***
distractorsyllable -0.7861     0.2913  -2.698  0.00697 **
distractortone      0.4161     0.2665   1.562  0.11837
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                                    edf Ref.df  Chi.sq  p-value
s(day):distractorno                           1.000e+00  1.000   0.359  0.54923
s(day):distractorsyllable                     1.001e+00  1.002   1.121  0.28985
s(day):distractortone                         1.431e+00  1.675   0.656  0.71815
s(day,participant):distractorno               1.741e+01 51.000  70.058  0.03821 *
s(day,participant):distractorsyllable         2.239e+01 51.000  45.157  0.00246 **
s(day,participant):distractortone             3.315e+01 51.000 138.672  0.00461 **
s(cohort,participant):distractorno            8.935e+00 25.000  19.084  < 2e-16 ***
s(cohort,participant):distractorsyllable      7.452e-03 25.000   0.006  0.04602 *
s(cohort,participant):distractortone          5.715e+00 25.000   7.769  < 2e-16 ***
s(gender,gender_training):distractorno        1.894e+00  3.000   9.713  0.04527 *
s(gender,gender_training):distractorsyllable 1.483e+00  3.000   5.378  0.07344 .
s(gender,gender_training):distractortone      8.329e-04  3.000   0.001  0.43585
s(response_time):distractorno                 8.645e-01  1.000   7.185  0.00709 **
s(response_time):distractorsyllable           9.478e-01  1.000  19.646 1.57e-05 ***
s(response_time):distractortone               5.041e-05  1.000   0.000  0.42569
s(item):distractorno                          4.633e+01 85.000  97.300 1.05e-06 ***
s(item):distractorsyllable                    2.596e+01 47.000  59.696 4.07e-06 ***
s(item):distractortone                        1.939e+01 44.000  33.848  0.00183 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.174   Deviance explained = 17.9%
```

Figure 9: Numerical results from the Bonus GAM Model

4.2.5. Linear Model

The GAM models in the previous sections uncovered complex non-linear relationships between variables. However, as seen in Figure 6 the lines for no distractor and syllable distractor appear linear, while a slight nonlinear trend is visible for the tone distractor type. To investigate how well a linear model could describe the data, a Generalised Linear Mixed Model (GLMM) was built. The model included the same variables as used in

Model 2, as the linear model will be compared with it. There were two fixed effects in the model: distractor, and day: distractor. They were checking the effect of the distractor on correctness, and if the effect of the distractor changes over days, respectively. All the other variables were modelled as random effects. They were accounting for variations between the participants, items, and cohorts. The model is shown in Equation 5.

$$\text{correct} \sim \text{distractor} + \text{day: distractor} + (1|\text{participant: distractor}) + (1|\text{day: participant: distractor}) + (1|\text{cohort: participant: distractor}) + (1|\text{item: distractor}) \quad \text{Equation 5}$$

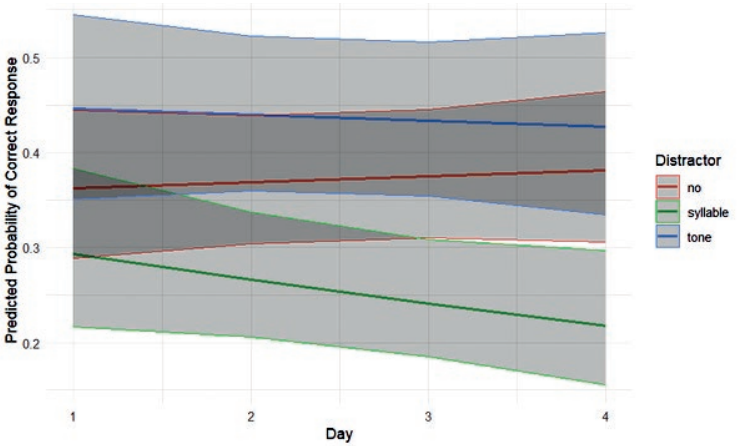

Figure 10: Results from the Linear Model

```
Random effects:
 Groups                          Name        Variance Std.Dev.
 day:participant:distractor      (Intercept) 0.2188   0.4677
 item:distractor                 (Intercept) 0.3395   0.5827
 cohort:participant:distractor (Intercept) 0.1721   0.4148
 participant:distractor          (Intercept) 0.1355   0.3681
Number of obs: 3328, groups:  day:participant:distractor, 312;

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.58943    0.21515  -2.740  0.00615 **
distractorsyllable    -0.15675    0.34212  -0.458  0.64683
distractortone         0.39698    0.33713   1.178  0.23899
distractorno:day       0.02654    0.06345   0.418  0.67573
distractorsyllable:day -0.13274    0.08424  -1.576  0.11507
distractortone:day    -0.02505    0.07981  -0.314  0.75367
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Numerical results from the Linear Model

The visual results from the Linear Model are shown in Figures 10. Due to the complexity of the model, the confidence intervals are wide, causing different types of distractors to overlap. The overlaps suggest that different types do not differ from each other significantly. The numerical results shown in Figure 11 indicate that neither syllable distractors nor tone distractors across days had a significant effect on the correctness. The high variance and standard deviations of random effects indicate a high variability of these variables. This suggests that the effects of distractors vary significantly across different participants, items, and cohorts. This high variability indicated that individual differences play a crucial role in the correctness of answers. Lastly, the r^2 value equalled 0.226, showing that 0.266 of the dependent variable was explained by the independent variable.

5. Discussion

In the discussion, we interpret the experimental results in light of previous research and theoretical models. We argue that while syllables were acquired through statistical learning, tones proved more difficult for non-tonal speakers, which challenges earlier findings.

5.1. General discussion

The purpose of the study was to examine the acquisition of tones and syllables by non-tonal speakers using statistical learning.

The experiment contained various types of tone-syllable combinations and distractors to isolate certain tones and syllables, allowing for the investigation of the effects of statistical learning across four study days. The visual representation of the results suggested that participants based their answers solely on syllables, ignoring lexical tones. Specifically, participants were increasingly misled over time in the case of syllable distractors, whereas with tone distractors, the responses remained consistent across days. As predicted in the hypothesis, participants struggled more with syllable distractors than with tones, leading to lower correctness in responses when syllable distractors were present. These findings suggest that participants did not learn lexical tones through statistical learning during the experiment, contradicting the initial hypothesis, which proposed that participants could learn both tones and syllables through statistical learning.

The results of this study also contradict the findings of Tang/Spenader/Jones (2024) and Wiener/Ito/Speer (2021), both of whom investigated the

role of statistical learning in acquiring lexical tones. However, both studies contained potential flaws that may have hindered their ability to study this phenomenon appropriately. Wiener/Ito/Speer (2021) did not account for error-driven learning, which made it significantly easier for participants to learn syllable-tone combinations. Tang/Spenader/Jones (2024), on the other hand, removed the factor of error-driven learning but focused solely on syllable-tone combinations while neglecting individual variables. By incorporating distractors and eliminating feedback during the phases, this study avoided the flaws of these previous investigations.

The visual representation of the findings provided insights into the inability to learn novel aspects using statistical learning. To provide more certainty, three GAM models were created. The Base Model was created to compare the performance of future models. Nevertheless, it showed that the answer correctness is significantly different depending on the distraction type. However, its simplicity did not manage to uncover more complex dependencies. The intended model, Model 2, included all variables with real-world justifications for influencing correctness. This model uncovered complex dependencies between correctness and variables such as item and cohort. Notably, it demonstrated that when accounting for individual variability, both tone and syllable distractors significantly impacted correctness over time. Moreover, the statistical analysis confirmed that the chances of participants answering correctly when distracted by a high-frequency syllable were lower compared to when there were no distractions. There was no significant difference when distracted by tones compared to no distractions, suggesting that isolated tones were not perceived by participants enough to distract them. However, despite successfully identifying significant dependencies, the model lacked strong predictive power, explaining only 16.8% of the deviance. Moreover, the statistical analysis did not confirm a significant change in correctness across the days without accounting for individual variability for participants. The Bonus Model extended Model 2 by incorporating additional terms without clear real-world implications, aiming to assess whether more deviance could be explained. However, its performance was underwhelming, as it only accounted for 17.9% of the deviance, just a minor improvement over Model 2. Finally, the Linear Model was created to see if using linear relationships only can match the performance of complex, non-linear GAM models. The model was not capable of uncovering any significant findings with regard to distractors and days. Moreover, the big variation in deviance in random effects caused the confidence intervals of the models to be too

large to have conclusive interpretation. GAM models found more complex interactions between variables, which were not possible in the Linear Model, suggesting that the initial choice of GAM was appropriate.

5.2. Limitations and improvements

Various factors have influenced the present results, potentially reducing their validity. First, only 26 participants took part in the study. There were six different matrices containing test and training cases, with approximately four participants per matrix. Such a small sample size leaves substantial room for random errors. Additionally, each matrix varied both syllables and tones, introducing a high degree of variability. These differences may have been too large to reliably evaluate the results. The data collection phase in this experiment limited the number of participants, restricting the generalizability of the findings. Moreover, individual differences in perception and language acquisition could have further influenced the results. For future research, a power analysis should be conducted to determine an appropriate sample size, ensuring more reliable results. Increasing the number of participants could improve the robustness of the findings and their reliability (Cohen 1988). Additionally, the lack of statistical power may have prevented the detection of smaller effects that potentially had meaningful effects. For future research, it might be better to create more matrices, incorporating all combinations of syllables and tones. Lastly, the duration of the experiment (4 days) may have been too short given the task's difficulty. In the experiments conducted by Tang/Spenader/Jones (2024) and Wiener/Ito/Speer (2021), the tasks were easier than those in this study, as the total accuracy in this experiment was lower compared to the other experiments. Moreover, in this paper, participants were deliberately misled into choosing incorrect answers (distracted by high-frequency syllables or tones) rather than simply identifying high-frequency sounds. Consequently, the four-day duration may have been insufficient for participants to adequately learn the sounds. Extending the experiment duration could yield more conclusive results.

Another limitation of the study was the model, which explained only 16.8% of the data. This suggests that the predictors of interest did not fully account for the variability in the dataset. The low explained deviance could be due to two factors: the use of binomial data and the lack of trial-by-trial time series. Binomial data only has two possible outcomes (0 or 1), so there is less variation for the model to work with compared to continuous data, making it more difficult to explain a lot of variance. The

choice of the binomial data was made based on the nature of the dependent variable in the study, and is appropriate, but it inevitably constrained the proportion of deviance that could be explained. Secondly, using each trial as a time series predictor instead of days could help capture more detailed changes in participants' performance, potentially increasing the deviance explained. Days as predictors account for broader trends, but modelling performance using smaller time steps (trial-by-trial) might allow the model to detect smaller variations more effectively. This could be explored in future research to investigate the influence of different time step sizes in the statistical learning of tones and syllables. Including these two changes could significantly enhance the model's explanatory power.

Finally, a greater variance explained could theoretically be achieved by adding more predictors to the model, like response time or the gender of the speaker. However, this approach has two key drawbacks. First, new terms and predictors must be added cautiously to avoid overfitting. Even if adding multiple new terms increased the variance explained, the results might not be reliable. Second, an additional model was tested to explore the potential for improvement. However, even after incorporating new terms in the Bonus Model, the variance explained was only 17.8%, which still provides limited predictive power.

The final limitation that could have influenced the results was the experiment design. Participants had varying availability across the four training days, often leading to testing at different times of the day. Additionally, numerical feedback was provided after each testing session, which may have influenced their responses in subsequent days. Standardising testing times and eliminating feedback could help produce more consistent results in future research.

5.3. Practical relevance

A key practical relevance of studying statistical learning of tones lies in its implications for second language acquisition. Many learners of tonal languages encounter many difficulties in distinguishing between different pitch patterns. This study investigated how non-tonal speakers perceive tones, providing a foundation for the development of more effective teaching methods and training tools. The findings suggest that learning by exposure to tonal languages may not be sufficient for learning. Instead, other methods, such as exercises to strengthen learners' sensitivity to tonal contrast, could be used to accelerate acquisition.

5.4. Final conclusion

This study can be considered an introduction to a novel perspective on statistical learning, particularly regarding the inability to learn novel aspects of sounds. The results have shown that there is a significant difference in correctness if syllables act as distractions. However, the main objective of the study was to investigate whether people can learn tones and syllables through statistical learning. The visualisation showed that it was not, as participants based their answers solely on syllables. This finding would have contradicted the results of Tang/Spenader/Jones (2024) and Wiener/Ito/Speer (2021), who concluded that learning lexical tones through statistical learning is possible. However, a clear relationship between correct answers across days and different distraction types was not statistically proven. A statistically significant relationship between days and correctness was only found when accounting for individual differences among participants. This suggests that further studies are needed in this field to clarify the role of statistical learning in acquiring novel elements of speech. Despite the limitations of the current experiment, it is plausible that replicating the study with improvements could yield more satisfactory and meaningful findings.

6. Acknowledgements

I would like to thank Dr. Stephen Jones and doctoral student Mi Tang for their support and supervision during the development of my Bachelor's thesis, which served as the foundation for the present article.

Bibliography

Arnon Inbal, 2019, Statistical learning, implicit learning, and first language acquisition: A critical evaluation of two developmental predictions, in: Topics in cognitive science 11 (3), pp. 504-519.

Brady Timothy F. / Aude Oliva, 2008. Statistical Learning Using Real-World Scenes: Extracting Categorical Regularities Without Conscious Intent, in: Psychological Science 19 (7), pp. 678-685.

Chomsky Noam, 1975, Reflections on Language, New York, NY: Pantheon.

Chomsky Noam, 1978, Topics in the Theory of Generative Grammar, Berlin/Boston: De Gruyter Mouton.

Cohen Jacob, 1988, Statistical power analysis for the behavioral sciences (2nd ed), New York: Lawrence Erlbaum Associates.

Colantoni Laura / Steele Jeffrey / Neyra Paola R.E., 2015, Second language speech, Cambridge: Cambridge University Press.

Crain Stephen / Lillo-Martin Diane C., 1999, An Introduction to Linguistic Theory and Language Acquisition, Malden, MA: Blackwell.

Crain Stephen / Thornton Rosaling / Murasugi Keiko, 2009, Capturing the Evasive Passive, in: Lang. Acquis 16, pp. 123-133

Curtin Suzanne / Zamuner Tania S., 2014, Understanding the developing sound system: interactions between sounds and words, Wiley Interdisciplinary Reviews, in: Cognitive Science 5 (5), pp. 589-602.

Dąbrowska Ewa, 2015, What exactly is Universal Grammar, and has anyone seen it?, in: Frontiers in Psychology 6 (Art. 852), pp. 1-17.

Erickson Lucy C. / Thiessen Erik D., 2015, Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition, in: Developmental Review 37, pp. 66-108.

Fiser József / Aslin Richard N., 2002, Statistical learning of higher-order temporal structure from visual shape sequences, in: Journal of Experimental Psychology: Learning, Memory and Cognition 28 (3), pp. 458-467.

Flynn Suzanne / Foley Claire / Vinnitskaya Inna, 2004, The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development in First, Second and Third Language Acquisition of Relative Clauses, in: International Journal of Multilingualism 1 (1), pp. 3-16.

Frost Ram / Armstrong Blair C. / Siegelman Noam / Christiansen Morten, 2015, Domain generality versus modality specificity: the paradox of statistical learning, in: Trends in Cognitive Sciences 19, pp. 117-125.

Gussenhoven Carlos, 2004, Pitch in Language I: Stress and Intonation. In: The Phonology of Tone and Intonation, Cambridge: Cambridge University Press, (Chapter 2) pp. 1-25.

Hastie Trevor / Tibshirani Robert, 1987, Generalized Additive Models: Some Applications, in: Journal of the American Statistical Association (American Statistical Association, Taylor & Francis, Ltd.) 82 (398), pp. 371-386.

Hoffmann Michael, 2016, Vision: Elementary and Complex Visual Processing, in: Hoffmann M. (ed.), Cognitive, Conative and Behavioral Neurology: An Evolutionary Perspective, New York, NY: Springer, pp. 51-82.

Kirkham Natasha Z. / Slemmer Jonathan A. / Johnson Scott P., 2002, Visual statistical learning in infancy: evidence for a domain general learning mechanism, in: Cognition 83 (2), pp. B35-B42.

Li Yuanning / Tang Caroline / Lu Junfeng / Wu Jinsong / Chang Edward F., 2021, Human cortical encoding of pitch in tonal and non-tonal languages, in: Nature communications 12 (1), Article 1161.

Mathot Sebastriaan / Schreij Daniel / Theeuwes Jan, 2012, OpenSesame: An open-source, graphical experiment builder for the social sciences, in: Behavior research methods 44 (2), pp. 314-324.

Nixon Jessie S., 2020, Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking, in: Cognition 197 (104081).

Saffran Jenny R. / Aslin Richard N. / Newport Elissa L., 1996, Statistical Learning by 8-Month-Old Infants, in: Science (American Association for the Advancement of Science) 274 (5294), pp. 1926-1928.

Schapiro Anna / Turk-Browne Nicholas, 2015, Statistical learning, in: Brain mapping 3 (1), pp. 501-506.

Seidenberg Mark S., 1997, Language acquisition and use: Learning and applying probabilistic constraints, in: Science 275 (5306), pp.1599-1603.

Smith Neilson V. / Tsimpli Ianthi-Maria, 1995, The mind of a savant: language learning and modularity, Oxford: Blackwell.

Tang Mi / Spenader Jennifer / Jones Stephen, 2024, Learning lexical tone through statistical learning in non-tone language speakers, in: Frontiers in Education 9 (online: https://doi.org/10.3389/feduc.2024.1393379).

Thiessen Erik D, 2011, Domain General Constraints on Statistical Learning, in: Child Development 82 (2), pp. 462-470.

Whittlesea Bruce W. / Wright Richard L., 1997, Implicit (and explicit) learning: acting adaptively without knowing the consequences, in: Journal of Experimental Psychology Learning Memory and Cognition, pp.181-200.

Wiener Seth / Ito Kiwako / Speer Shari R., 2021, Effect of multitalker input and instructional method on the dimension-based statistical learning of syllable-tone combinations; an eye tracking study, in: Studies in Second Language Acquisition 43 (1), pp. 155-180.

Xiang Dong, 2001, Fitting generalized additive models with the GAM procedure, in: SUGI Proceedings, pp. 256-326.

Appendix A

Sound 1					Sound 2				
syllable	tone	total frequency	syllable frequency	tone frequency	syllable	tone	total frequency	syllable frequency	tone frequency
Distractive-Syllable									
syllable 2	tone C	54	60	60	syllable 1	tone C	2	240	60
syllable 2	tone C	54	60	60	syllable 1	tone B	2	240	60
syllable 2	tone C	54	60	60	syllable 3	tone C	2	240	60
syllable 2	tone C	54	60	60	syllable 3	tone B	2	240	60
syllable 4	tone B	54	60	60	syllable 1	tone C	2	240	60
syllable 4	tone B	54	60	60	syllable 1	tone B	2	240	60
syllable 4	tone B	54	60	60	syllable 3	tone C	2	240	60
syllable 4	tone B	54	60	60	syllable 3	tone B	2	240	60
Distractive-Tone									
syllable 2	tone C	54	60	60	syllable 2	tone D	2	60	240
syllable 2	tone C	54	60	60	syllable 2	tone A	2	60	240
syllable 2	tone C	54	60	60	syllable 4	tone D	2	60	240
syllable 2	tone C	54	60	60	syllable 4	tone A	2	60	240
syllable 4	tone B	54	60	60	syllable 2	tone D	2	60	240
syllable 4	tone B	54	60	60	syllable 2	tone A	2	60	240
syllable 4	tone B	54	60	60	syllable 4	tone D	2	60	240
syllable 4	tone B	54	60	60	syllable 4	tone A	2	60	240
Non-Distractive									
syllable 1	tone D	-	-	-	syllable 2	tone D	-	-	-
syllable 1	tone D	-	-	-	syllable 2	tone A	-	-	-
syllable 1	tone D	-	-	-	syllable 4	tone D	-	-	-
syllable 1	tone D	-	-	-	syllable 4	tone A	-	-	-
syllable 3	tone A	-	-	-	syllable 2	tone D	-	-	-
syllable 3	tone A	-	-	-	syllable 2	tone A	-	-	-
syllable 3	tone A	-	-	-	syllable 4	tone D	-	-	-
syllable 3	tone A	-	-	-	syllable 4	tone A	-	-	-
syllable 1	tone D	-	-	-	syllable 1	tone C	-	-	-
syllable 1	tone D	-	-	-	syllable 1	tone B	-	-	-
syllable 1	tone D	-	-	-	syllable 3	tone C	-	-	-
syllable 1	tone D	-	-	-	syllable 3	tone B	-	-	-
syllable 3	tone A	-	-	-	syllable 1	tone C	-	-	-
syllable 3	tone A	-	-	-	syllable 1	tone B	-	-	-
syllable 3	tone A	-	-	-	syllable 3	tone C	-	-	-
syllable 3	tone A	-	-	-	syllable 3	tone B	-	-	-