university of groningen / faculty of science and engineering

An EEG Study of Statistical Tone Learning: Results from an ITC Analysis

Bachelor's Project Thesis

Ronald Musch, s3347508, r.m.musch@student.rug.nl,
Supervisors: J.K. Spenader, Dr & M. Tang, PhD candidate

Abstract: This bachelor's thesis investigates whether memory traces, that would be evidence of tonal learning, can be found for adult non-tonal language speakers during a short exposure. Eighteen participants were presented with two multi-feature oddball tasks and a short behavioral test, while EEG data was recorded. Each participant was exposed to four Mandarin tones while being distracted by a nature documentary. The EEG data were analyzed using a time-frequency analysis, looking at the difference in inter-trial coherence (ITC) in particular. In both the EEG and behavioral data, it was found that participants show signs of statistical syllabic and tonal learning, within five to ten minutes of exposure. This suggests that tones are statistically learnable and that the process of learning tones is visible in memory traces.

1 Introduction

Most languages use tone to give an emotional meaning to words or sentences. For speakers of these languages it is difficult to understand a language that uses tone to change the meaning of a word. To investigate if these tonal languages can be statistically learned by non-tonal language speakers, we will expose them to four Mandarin tones. For this study, EEG data is gathered and analyzed. Previous research mainly focuses on analyzing mismatch negativity from an event-related potential (ERP) analysis for similar tasks. A limitation of ERP analyses, is that the domain is one-dimensional. Other research indicates that inter-trial coherence (ITC), as a result of a time-frequency analysis, could also indicate signs of learning. Time-frequency analyses have a two-dimensional domain (time & frequency) and may give a fuller picture of the underlying brain activity.

In 1.1 we will introduce tones and tonal languages. Then in 1.2, an overview of statistical learning, the mechanism that drives implicit language learning, is given. Next, in 1.3, the most common way to observe statistical learning, analyzing mismatch negativity, is explained. Lastly, in 1.4, we will introduce inter-trial coherence as an alternative method of determining whether non-tonal language speakers are able to learn tones statistically.

1.1 Tone

A verbal language uses properties of sounds to distinguish the meaning of words. One of these properties is tone, the use of pitch in a language. Tone is used emotionally in all languages, however, Yip (2002) estimates that in 60-70 percent of all languages tone is more than just a subtlety. For over a billion speakers worldwide, tone is phonetic. A word, can have a different meaning based on the tone. One of these tonal languages is Mandarin, the Chinese language. According to Chao (2013) Mandarin is one of the oldest and most spoken languages in the world. Mandarin knows four lexical tones, therefore a single syllable combination can have up to four different meanings in Mandarin. An overview of the four Mandarin tones can be found in *Table 1.1.*

(1) Flat
(2) Rising
(3) Low Dipping
(4) High Falling

Table 1.1: The four Mandarin tones

1.2 Statistical Learning

Shen (1989) and Bluhme (1971) noted early on that tonal differences of the Mandarin language were difficult to learn for adult non-tonal language speakers. This can partly be attributed to the lesser language learning ability adults have when compared to children. (Clahsen and Muysken, 1989; Marinova-Todd, Marshall, and Snow, 2000; McLaughlin, 2013). Children from the age of three are able to acquire new languages quicker and more easily than adults. This period of childhood is known as the critical period. (Johnson and Newport, 1989).

However, age does not affect the capacity to learn, but only the process of learning. Adults learn languages explicitly, meaning that they use conscious processes to acquire the structure and vocabulary of a new language. Children acquire a new language implicitly, meaning that they learn unconsciously (Lichtman, 2013). It is unknown to what extend the implicit learning mechanisms from the critical period remain in adulthood. Previous research has found that traces of these mechanisms can still be found in adults (See Kittleson, Aguilar, Tokerud, Plante, and Asbjørnsen (2010) for implicit word segmentation in adults or Conway, Karpicke, and Pisoni (2007) for verbally easy to encode sequence learning in adults).

An example of implicit learning is statistical learning (SL). SL is one of the most researched components of language learning (Frost, Armstrong, and Christiansen, 2019). Statistical learning is the unconscious recognition of patterns that happens within a few minutes of exposure. SL is an important mechanism for the previously mentioned critical period (Thiessen, Girard, and Erickson, 2016). Saffran, Aslin, and Newport (1996) first discovered the possibility of SL playing a role in word segmentation of eight-month-old infants. Within two minutes of exposure, infants were able to segment words from fluent speech.

SL is difficult to test with a behavioral test, because it is implicit. Asking an explicit question makes the learning goal explicit, which corrupts the results. Therefore, signs of SL are investigated using equipment to analyse brain activity, for example; fMRI or EEG. With this equipment a participant does not have to actively respond to the stimuli to provide data. For our experiment we will use EEG equipment, since it is the least invasive method of recording brain activity.

1.3 Mismatch Negativity

A common method in the field of language learning for determining whether SL is present, is by looking at the mismatch negativity (MMN). MMN is a demonstration of the brain learning the statistical structure of its environment, i.e. statistical learning (Lieder, Daunizeau, Garrido, Friston, and Stephan, 2013). MMN has been widely used to analyse the brain's responsiveness to auditory stimuli (Näätänen, Pakarinen, Rinne, and Takegata, 2004; Stefanics, Kremláček, and Czigler, 2014). MMN is calculated by subtracting the brain's response to a deviant stimulus from the response to a standard stimulus. If there is a significant difference, it means that the brain is able to distinguish the stimuli.

MMN is observed 100-250ms after the presentation of the stimuli. It is best visible in the frontal lobe (Zhang, Yan, and Huang, 2018). We will follow the methods of Näätänen, Paavilainen, Rinne, and Alho (2007) and look at the activity captured by the Fz-electrode, located at the middle of the frontal lobe. The most common way to visualize MMN is by doing an ERP analysis. But other research suggests a strong relation between MMN and a difference in inter-trial coherence (ITC).

1.4 Inter-Trial Coherence

ITC is a score for the phase similarity of brain waves among participants. When the ITC is high, the phases of the brain waves among participants are within a small margin of each other. Bishop and Hardiman (2010) concluded that for their participants, a significant or above average difference in ITC was visible, when they observed MMN with an ERP analysis. This suggests that a significant difference in ITC can also be used as an indicator of the ability of the participant to distinguish stimuli unconsciously.

Because ITC is calculated by doing a time-frequency analysis, we keep the data two-dimensional. This makes it possible to draw a conclusion about which frequencies show notable activity during the MMN time window. This allows us to

analyze in which brain waves the responses to stimuli occur, which gives additional information about the mental state of the participant. This is additional information that cannot be revealed with an ERP analysis.

In conclusion, we will be looking for signs of statistical learning when we expose adult non-tonal language speakers to four Mandarin tones. The presence of these signs will be determined by analyzing the significant differences in ITC as a response to the stimuli.

2 Methods

The experiment took place at the EEG lab in the Bernoulliborg of the University of Groningen, over the course of three weeks.

2.1 Participants

For our experiment we recruited eighteen participants (44% female, mean age = 23.8 with a SD of 4.3). We recorded their spoken languages and musical experience, since these factor might influence their ability to learn tones. We determined that none of the participants spoke, or were learning, a tonal language. Furthermore, the participants varied in musical experience.

The participants signed an informed consent form before participating. They were paid in euros for their participation, regardless of their performance. They were not instructed on the goal of the experiment beforehand.

2.2 Stimuli

The creation of our experiment was done using OpenSesame 3.3.11. All stimuli were recorded by a voice actress. The recording was done in a studio in China, using Audio-Technica AT2020. Stimuli were normalized and processed with Adobe Audition CS6. In total, we presented our participants with six different tone-syllable combinations to account for all four tones in Mandarin: two standard stimuli (SS) and four deviant stimuli (DS). These stimuli are distributed into two groups with one SS and two DS per group. See *Table 2.1* for an overview.

Stimuli	**Syllable**	**Tone**
SS1	/bi/	flat
DS1	/bi/	rising
DS2	/du/	flat
SS2	/kou/	low dipping
DS3	/kou/	high falling
DS4	/pei/	low dipping

Table 2.1: All tone-syllable combinations used for our experiment. SS = standard stimuli, DS = deviant stimuli

2.3 Multi-Feature Oddball

For the presentation of our stimuli we opted for a multi-feature oddball task (MO) as presented by Näätänen et al. (2004). An oddball task consists of a stream of standard stimuli, occasionally interrupted by a deviant stimuli (the oddball). We use multiple deviant stimuli, hence it becomes a multi-feature oddball task. Because we grouped our stimuli, we were able to present the participant with two MO tasks (MO1 & MO2). See *Table 2.2* for an overview. Besides tonal learning, we also included syllabic learning (the learning of new syllable combinations). Syllabic learning is something the participant is familiar with, since all languages use different syllable combinations to distinguish meaning.

	Stimuli pair:	**Type**
MO1:	SS1 vs. DS1	Tonal learning
	SS1 vs. DS2	Syllabic learning
MO2:	SS2 vs. DS3	Tonal learning
	SS2 vs. DS4	Syllabic learning

Table 2.2: Stimuli grouped in two MO's and the associated type of learning

At the start of a MO, the SS is presented fifteen times to familiarize the participant with the experiment. Next, one of the DS is presented, alternating with the SS for a total of 300 times. The result is a MO with 315 iterations of SS and 150 iterations of each DS. Each stimulus is presented for a total of one second. The total of both MO's takes roughly twenty minutes to complete. An example presentation of the standard and deviant stimuli is given in *Table 2.3*. Note that DS1 and DS2 are interchange-

able and chosen randomly. A similar presentation is used for MO2, but with SS2, DS3 & DS4 from *Table 2.1.*

15 * SS1	DS1	SS1	DS2	SS1	DS2	SS1
Time →						

Table 2.3: Example presentation of MO1

Half of the participants were presented with MO1 first and MO2 second. The other half of the participants were first presented with MO2 followed by MO1, to counteract potential ordering effects.

2.4 EEG Recording

During the MO's the EEG data of the participant was recorded. We recorded EEG from 32 positions using active Ag-AgCl electrodes (BioSemi ActiveTwo system) digitized with a sampling rate of 1024 Hz. The electrodes were placed using the international 10–20 system layout including two "ground" channels: Common Mode Sense and Driven Right Leg. Six additional electrodes were attached, with two horizontal and two vertical electrodes used to detect eye movement and two electrodes to measure mastoid signals. Scalp impedance for each electrode was kept under 30kΩ for all participants. During the experiment, participants watched a nature documentary of their choice to keep their mind occupied.

2.5 Processing Pipeline

For the EEG data (pre-)processing and analysis, the open-source toolbox EEGLAB (v2021.1) was used along with custom-made scripts in MATLAB R2021b, update 3. The processing pipeline consists of seven steps:

(1) EEG data were applied to a Notch filter of 50 Hz and then filtered with a high-pass filter of 0.01 Hz and a low-pass filter of 30 Hz.
(2) The data were split into two groups named 'Early' and 'Late'. Group early consists of the first half of both MO's and group late consists of the second half of both MO's, resulting in two data sets of roughly ten minutes.
(3) Both data sets were post hoc referenced to the average of the whole-head electrodes.
(4) Bad channel removal and artifact rejection was done by using the Clean Rawdata & ASR algorithm from EEGLAB's build-in tools. Artifact rejections resulted in 16.2% and 20.8% of data loss for the early and late data set respectively. Bad channel removal resulted in 7.3% and 6.6% of channels removed for the early and late data set respectively. (See *Appendix A* for the detailed rejection and removal rates per data set).
(5) Data sets were decomposed into ICA components and all components that are not labeled 'brain' were removed.
(6) One second epochs from [-100ms, 900ms] were created for both data sets.
(7) ITC images (from 3-30Hz and 100 to 700ms) are generated as a result from a time-frequency analysis with wavelet cycles set to 1.

2.6 Behavioral Test

After both multi-feature oddball tasks we ended the EEG recording and asked the participant to make a short behavioral test. In each question we played a pair of sounds; one of the SS and one of the DS. We then asked the participant the following question: *"Which one sounds more familiar to you?"*, which they could answer by pressing *1* or *2* on the keyboard. If the participant correctly identifies the SS, they score a point. Because the SS had the highest frequency during the EEG experiment, we expect this to be more familiar to the participant. We paired all possible combinations of SS and DS and presented them in two orders (SS first & DS first) twice, resulting in a total of 32 questions.

From these questions, only half were relevant for the goal of the experiment: Questions that presented the stimuli of MO1 (SS1 vs. DS1 or DS2) as a sound pair, and the questions that presented the stimuli of MO2 (SS2 vs. DS3 or DS4) as a sound pair. This leaves us with sixteen target questions. The other sixteen questions are fillers, and are used to obscure the learning goal of the experiment.

From the sixteen target questions, we post hoc distributed them into two groups. The first group is considered tonal learning, and consists of the questions where the participant has to correctly identify the SS from the deviant tone stimuli (SS1 vs. DS1 or SS2 vs. DS3). The second group is syllabic learning, which consists of the questions where the participant has to correctly identify the SS from the de-

viant syllable combination (SS1 vs. DS2 or SS2 vs. DS4). We calculated the accuracy per participant for both tonal and syllabic learning by dividing the number of correct answers by the total number of questions from that group. (See *Appendix B* for an overview of the questions).

The behavioral test started with a set of instructions and three trial questions to familiarize the participant with the task. If the participant did not respond within five seconds they would score a 0 for that question.

3 Results

In this part of the paper the results from the EEG data and the behavioral test are presented. For the statistical analysis of the behavioral results, R-Studio 2022.02.3, build 492 with R version 4.2.1 was used. The statistical analysis of the EEG data was done with the build-in tools of EEGLAB

3.1 ITC Results

For the early and late data sets the ITC for each point from 3 to 30Hz and 100 to 700ms after stimulus presentation was calculated. The mean ITC among all participants for the SS is plotted in figure 3.1a for the early data set and in figure 3.2a for the late data set. The mean ITC for both DS are plotted in figure 3.1b and 3.1c for the early data set and in figure 3.2b and 3.2c for the late data set.

The black arrow at 290ms marks the point of voice onset time (VOT) of the stimuli. In all figures, the response to stimuli is visible from about 330ms to 700ms, with an ITC peak around 550ms. The yellow area before VOT and above 15Hz that is visible in figure 3.1a and 3.2a is considered noise. The same goes for figure 3.1b, 3.1c, 3.2b and 3.2c. However in these figures this area is noisier because the data were averaged and there were less trials for the deviant stimuli.

3.2 ITC Statistical Analysis

For the statistical analysis, a paired t-test is performed for each ITC score of SS and DS. The results are plotted in four figures where a significant difference is marked with a red dot. The p-values differ from 0.05 to 0.001, and a darker shade of red corresponds to a lower p-value. The results are in figure 3.3.

The vertical dotted line marks the VOT. The horizontal dotted line is the lower bound of the beta band (12.5Hz). Activity in the beta waves is considered to be a part of active thought processes or muscle movement, and therefore it is irrelevant for statistical learning. All significant differences in ITC that are within the gray area, are disregarded. The red dotted rectangle marks the time window of MMN (100-250ms after VOT).

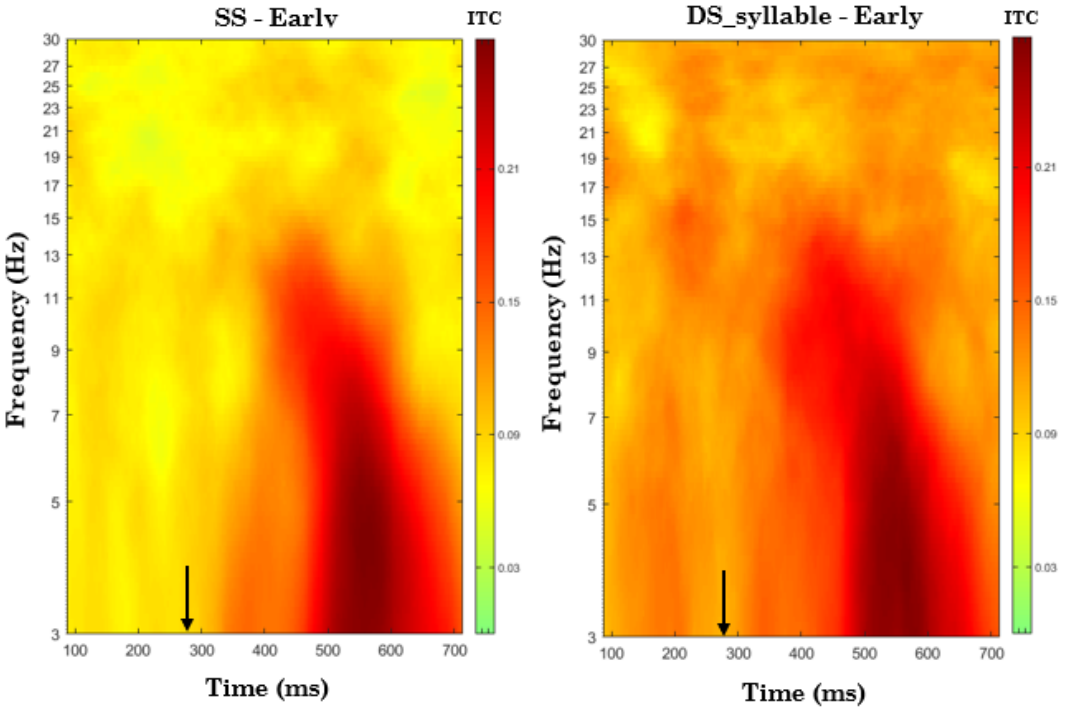
Early Syllabic Learning (figure 3.3a):
The statistical analysis indicates a significant difference in ITC at the 7-11Hz and 300-400ms window for early syllabic learning with a p-value range of 0.001-0.0025. This area falls before the MMN time window.

Early Tonal Learning (figure 3.3b):
The statistical analysis does not indicate any areas of significant difference in ITC for early tonal learning, within the area of interest. There is a significant ITC difference at 9-11Hz at 700ms, but this is on the edge of what is considered a response to stimuli and is most likely noise.

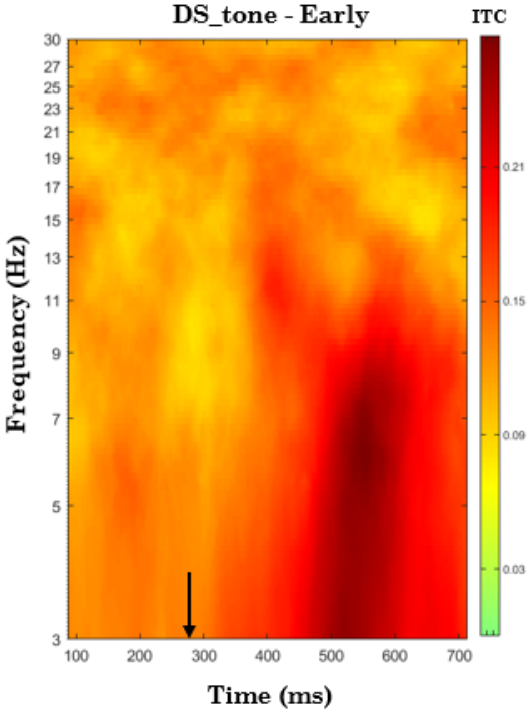
Late Syllabic Learning (figure 3.3c):
The statistical analysis indicates a significant difference in ITC at the 5-9Hz and 400-550ms window for late syllabic learning with a p-value range of 0.01-0.05. This area falls within the MMN time window.

Late Tonal Learning (figure 3.3d):
The statistical analysis indicates a significant difference in ITC at the 7-11Hz and 300-400ms window, as well as, the 6-9Hz and 400-475ms window, and, the 3-6Hz and 600-700ms window, for late tonal learning with a p-value range of 0.0025-0.01. The second area falls within the MMN time window.

In sum, we observed multiple areas of interest in these figures. For the late data sets these areas fall within the MMN time window. In some figures we observe areas of interest before and/or after the MMN time window. These areas will be discussed in the final section of this paper.

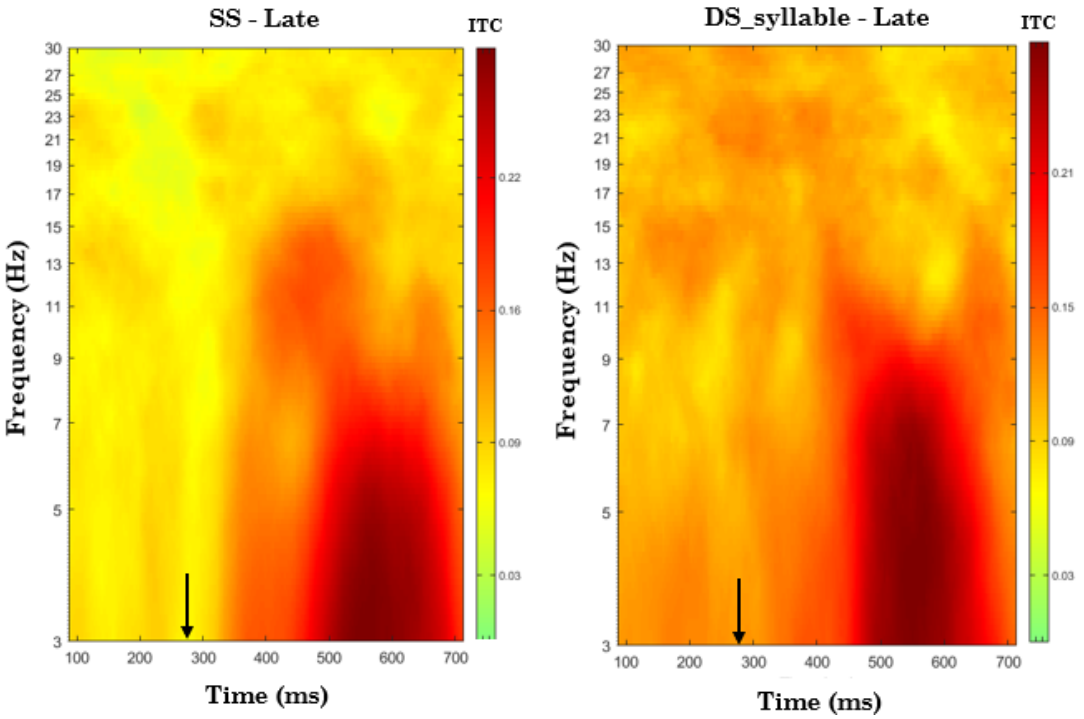


(a) (b)

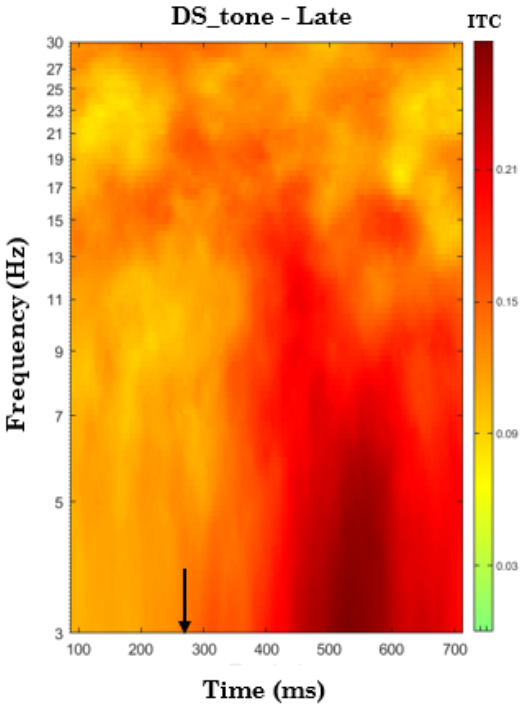


(c)

Figure 3.1: ITC results for the early data set with: (a) ITC response to the standard stimuli. (b) ITC response to the deviant stimuli that differ in syllable combination. (c) ITC response to the deviant stimuli that differ in tone.



(a) (b)



(c)

Figure 3.2: ITC results for the late data set with: (a) ITC response to the standard stimuli. (b) ITC response to the deviant stimuli that differ in syllable combination. (c) ITC response to the deviant stimuli that differ in tone.

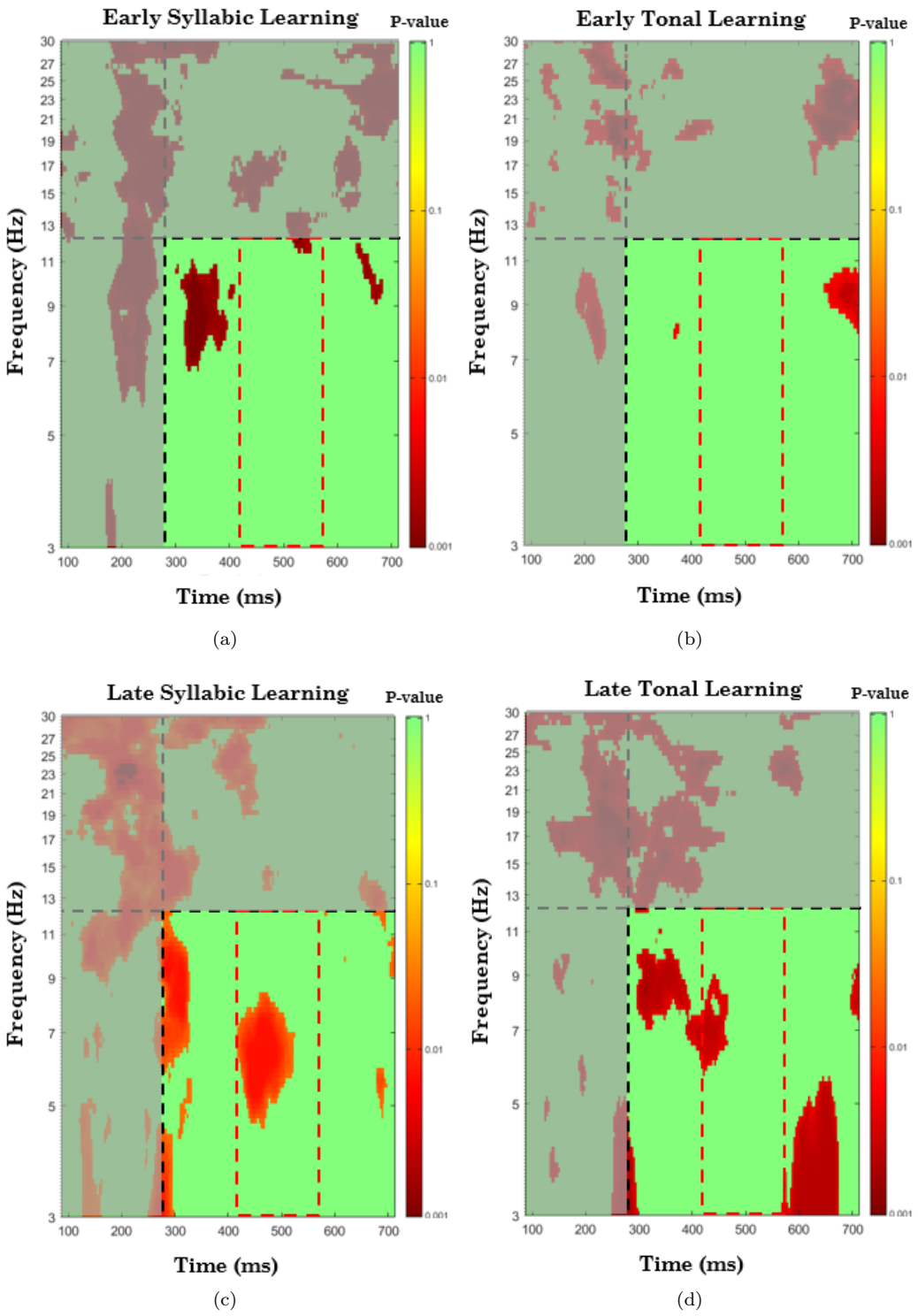


Figure 3.3: Results from a paired t-test performed on: (a) SS and DS_syllable from the early data set. (b) SS and DS_tone from the early data set. (c) SS and DS_syllable form the late data set. (d) SS and DS_tone from the late data set.

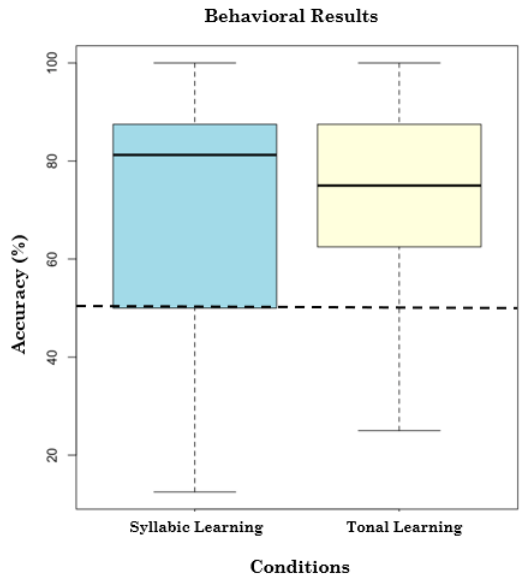


Figure 3.4: A box plot of the accuracy for syllabic and tonal learning from the behavioral test

3.3 Behavioral Results

For the behavioral results we plotted the mean accuracy for syllabic and tonal learning in figure 3.4. The mean accuracy for syllabic learning is 71.53% (SD = 27.58%). The mean accuracy for tonal learning is 69.14% (SD = 27.58%). The horizontal dotted line marks the chance level (50% accuracy).

A one-sample t-test on tonal learning and chance level (mu=50), reveals that participants are significantly more likely to distinguish tones above chance level (t = 4.05, df = 17, p-value = 0.0008). Furthermore, another one-sample t-test on syllabic learning and chance level, reveals that participants are significantly more likely to distinguish syllable combinations above chance level (t = 3.22, df = 17, p-value 0.005). Finally, a paired t-test reveals that there is no significant difference between syllabic and tonal learning (t = -0.41, df = 17, p-value = 0.685).

4 Conclusions & Discussion

From our results we are able to conclude that tones are statistically learnable for adult non-tonal language speakers, within five to ten minutes of exposure. This is confirmed by the significant difference in ITC that is visible in the 400-475ms (100-175ms after VOT) and 6-9Hz window for late tonal learning (see *Figure 3.3d*). This means that the activity takes place in the lower alpha and upper theta waves. The existence of this area shows that the phases of the brain waves of the participants are significantly different when the standard stimuli is presented versus when the deviant tone stimuli is presented. This shows that memory traces are able to reflect the process of statistically learning tones. By analyzing this area in figure 3.2a and 3.2c, it is visible that the ITC for SS is significantly lower than the ITC for DS_tone. Our behavioral results confirm this conclusion, by showing that participants have a high accuracy on identifying the deviant tone from the standard stimuli after ten minutes of exposure.

In addition to tonal learning, we also observe syllabic learning within five to ten minutes of exposure. This can be logically explained, because syllable combinations are used to differentiate the meaning of words in every language. Non-tonal language speakers are likely to learn the difference between syllable combinations, because they are familiar with this.

Furthermore, we conclude that both new syllable combinations and tones are not learnable within five minutes of exposure. In the early data sets we see no significant difference in ITC among the participants, suggesting their brains treat SS and both DS similarly. They were not prompted beforehand to pay attention to the stimuli nor did they know their meaning, therefore it is logical that at this moment the brain is treating the stimuli as noise.

4.1 Endogenous Activity

Endogenous activity is a response that lasts longer than 100ms and take place before or after the brain's response to stimuli (Başar, Başar-Eroglu, Rosen, and Schütt, 1984). Some of these are known, such as the P300 component, which is associated with context updating (Donchin and Coles, 1988). This response is usually visible 300ms after VOT. Presumably the significant ITC difference for late tonal learning, observed at 600-700ms in the lower theta and delta brain waves (see *Figure 3.3d*) is a result from the P300 component.

In some cases we observed premature significant ITC differences. This was the case for early syllabic

learning (see *Figure 3.3a*) and late tonal learning (see *Figure 3.3d*). In both cases a significantly different ITC at 300-400ms in the alpha brain waves was detected. According to Toscani, Marzi, Righi, Viggiano, and Baldassi (2010), phase coherence in alpha waves can be a consequence of the lack of stimulation of sensory areas, in particular when the eyes are closed. We presented auditory stimuli for the experiment and visual stimuli as a distraction. Lack of sensory stimulation should not be the cause of this. It could be that multiple participants chose to close their eyes for longer periods of time during the experiment. If this period extends for multiple seconds, this would not have been filtered out by artifact rejection or ICA component labelling.

4.2 Duration

Several aspects of our experiment show that the duration of the experiment is a relevant factor. The duration of the EEG part of the experiment was a little over twenty minutes. According to Craciun, Gardella, Alving, Terney, Mindruta, Zarubova, and Beniczky (2014) this is at the upper bound for awake EEG trials. The first argument for a shorter duration is to reduce artifacts. We see that the late data set has a 4.5 percentage point higher rejection rate than the early data set. Furthermore, we see a decrease in p-value range during syllabic learning. We see p-values nearing the 0.001 for early syllabic learning (see *Figure 3.3a*), while the p-value are around 0.01 for late syllabic learning (see *Figure 3.3c*). This suggests that the phases of brain waves as a response to a different syllable combinations become generally more random as the experiment goes on. It could be the case that the brain is blocking out these stimuli due to over-stimulation. On the other hand it could be the case that participants have consciously figured out that the deviant syllable combination is less important for the experiment. We did not ask the participants this explicitly, but informally none could correctly guess the goal of the experiment afterwards. Notably we do not observe this effect when comparing early and late tonal learning. This suggests that the difference in tones still remained interesting enough for the brain to not reduce it to noise.

A suggestion for further research is an experimental setup with a shorter duration. From our results it is evident that tones are learnable after five to ten minutes of exposure. This sets five minutes as the lower bound for the duration of a similar experiment. The total duration of our experiment exceeded the twenty minute mark, because we presented two MO tasks. Further research could opt for a modified oddball paradigm that includes multiple SS and DS. Oddball tasks with multiple SS and DS have known to also evoke MMN and ITC differences for auditory stimuli (Psiouri, Stavrinou, Koupparis, Kokkinos, and Kostopoulos, 2009). This could potentially cut the experiment duration in half.

4.3 Other Limitations

Besides duration, there are some other limitations and improvements for the experiment. First, our recording lab was not completely sound-proof and the door could not lock, causing some interruptions for a few participants. The most extreme deviations in EEG data were filtered out during the data processing, but general noise from outside during the experiment causes brain responses that are unlikely to be filtered. This results in noisier EEG data in general which could interfere with scientific conclusions. For further research a sound-proof lab, a door lock or sign on the door would be beneficial.

Second, while MMN is a proven determinant of the brain's ability to differentiate stimuli, ITC difference is not. Research has proven that ITC differences occur almost always when MMN is observed (Bishop and Hardiman, 2010). However this does not prove that MMN can be derived from significant ITC differences. Therefore the conclusions from this research are limited to assuming MMN is present when ITC differences are observed within the MMN time window. Alternatively this research could be a stepping stone for looking at ITC differences to determine if the brain is differentiating stimuli. Or this research can be used as an argument to include ITC alongside of ERP results, to still be able to include frequencies in the results and discussion. Our behavioral results somewhat support that ITC is a valid indicator of statistical learning. Apparently our participants have learned tonal differences significantly above chance level while we observed significant ITC differences in the MMN time window during the learning process.

Lastly, the usable questions from the behavioral test should be doubled in size. For our behavioral

results we only had sixteen target questions (eight for syllabic learning and eight for tonal learning). Looking at the fairly large standard deviations, these results are not very strong. Stronger results could be achieved by replacing the filler questions with target questions, but this will increase the risk of participants discovering the learning goal, making their answer explicit. Alternatively the duration of the behavioral test could be doubled. This last argument is especially valid because the behavioral test is a short task and the duration of the EEG experiment can be reduced.

4.4 Conclusion

In conclusion, we have found reasonable evidence of tonal learning for non-tonal languages speakers within five to ten minutes of exposure. Furthermore, our EEG results strongly suggest that ITC could also be used as an indicator of statistical learning.

Further EEG research in statistical learning of tones could focus on a different presentation of stimuli and a shorter duration of the EEG experiment. It could also be interesting to compare ITC and ERP results from the same data set, to strengthen the claim that ITC is a valid indicator of statistical learning. And finally, a more robust behavioral test in future research could lead to a stronger comparison between the participant's conscious and unconscious learning.

References

E Başar, C Başar-Eroglu, B Rosen, and A Schütt. A new approach to endogenous event-related potentials in man: relation between eeg and p300-wave. *International Journal of Neuroscience*, 24 (1):1–21, 1984.

Dorothy Vera Margaret Bishop and Mervyn James Hardiman. Measurement of mismatch negativity in individuals: a study using single-trial analysis. *Psychophysiology*, 47(4):697–705, 2010.

Hermann Bluhme. An audio-visual display of pitch for teaching chinese tones. *Studies in Linguistics*, 22:51–57, 1971.

Yuen Ren Chao. Mandarin primer. In *Mandarin Primer*. Harvard University Press, 2013.

Harald Clahsen and Pieter Muysken. The ug paradox in l2 acquisition. *Interlanguage studies bulletin (Utrecht)*, 5(1):1–29, 1989.

Christopher M Conway, Jennifer Karpicke, and David B Pisoni. Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with hearing adults. *Journal of deaf studies and deaf education*, 12(3): 317–334, 2007.

L Craciun, Elena Gardella, Jørgen Alving, D Terney, I Mindruta, J Zarubova, and S Beniczky. How long shall we record electroencephalography? *Acta Neurologica Scandinavica*, 129(2):e9–e11, 2014.

Emanuel Donchin and Michael GH Coles. Is the p300 component a manifestation of context updating? *Behavioral and brain sciences*, 11(3): 357–374, 1988.

Ram Frost, Blair C Armstrong, and Morten H Christiansen. Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145(12):1128, 2019.

Jacqueline S Johnson and Elissa L Newport. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive psychology*, 21(1):60–99, 1989.

Megan M Kittleson, Jessica M Aguilar, Gry Line Tokerud, Elena Plante, and Arve E Asbjørnsen. Implicit language learning: Adults' ability to segment words in norwegian. *Bilingualism: Language and Cognition*, 13(4):513–523, 2010.

Karen Lichtman. Developmental comparisons of implicit and explicit language learning. *Language Acquisition*, 20(2):93–108, 2013.

Falk Lieder, Jean Daunizeau, Marta I Garrido, Karl J Friston, and Klaas E Stephan. Modelling trial-by-trial changes in the mismatch negativity. *PLoS computational biology*, 9(2):e1002911, 2013.

Stefka H Marinova-Todd, D Bradford Marshall, and Catherine E Snow. Three misconceptions about age and l2 learning. *TESOL quarterly*, 34 (1):9–34, 2000.

Barry McLaughlin. *Second language acquisition in childhood: Volume 2: School-age Children*. Psychology Press, 2013.

Risto Näätänen, Satu Pakarinen, Teemu Rinne, and Rika Takegata. The mismatch negativity (mmn): towards the optimal paradigm. *Clinical neurophysiology*, 115(1):140–144, 2004.

Risto Näätänen, Petri Paavilainen, Teemu Rinne, and Kimmo Alho. The mismatch negativity (mmn) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118 (12):2544–2590, 2007.

G Psiouri, ML Stavrinou, A Koupparis, V Kokkinos, and GK Kostopoulos. Demonstration of mmn in human subjects after stimulation with two complex deviant sounds. In *Conference Abstract: 41st European Brain and Behaviour Society Meeting. doi: 10.3389/conf. neuro*, volume 8, 2009.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

Xiaonan Susan Shen. Toward a register approach in teaching mandarin tones. *Journal of the Chinese Language Teachers Association*, 24(3):27–47, 1989.

Gábor Stefanics, Jan Kremláček, and István Czigler. Visual mismatch negativity: a predictive coding view. *Frontiers in human neuroscience*, 8: 666, 2014.

Erik D Thiessen, Sandrine Girard, and Lucy C Erickson. Statistical learning and the critical period: how a continuous learning mechanism can give rise to discontinuous learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(4): 276–288, 2016.

Matteo Toscani, Tessa Marzi, Stefania Righi, Maria Pia Viggiano, and Stefano Baldassi. Alpha waves: a neural signature of visual suppression. *Experimental brain research*, 207(3):213–219, 2010.

Moira Yip. *Tone*. Cambridge University Press, 2002.

Yun Zhang, Qiang Yan, Wang, and Liyu Huang. Cortical areas associated with mismatch negativity: A connectivity study using propofol anesthesia. *Frontiers in Human Neuroscience*, 12:392, 2018.

A Appendix - Artifact Rejection & Channel Removal Rates

Dataset name	Length (s)	# of channels	Post-artifact rejection length (s)	Post-artifact rejection # of channels	% of length (kept)	% of channels (kept)
S0_early	595	32	378	29	63.53%	90.63%
S1_early	594	32	554	30	93.27%	93.75%
S2_early	597	32	591	29	98.99%	90.63%
S3_early	596	32	450	29	75.50%	90.63%
S4_early	595	32	486	29	81.68%	90.63%
S5_early	594	32	412	29	69.36%	90.63%
S6_early	594	32	479	28	80.64%	87.50%
S7_early	594	32	554	31	93.27%	96.88%
S8_early	593	32	498	32	83.98%	100%
S9_early	600	32	587	30	97.83%	93.75%
S10_early	595	32	432	30	72.61%	93.75%
S11_early	593	32	533	28	89.88%	87.50%
S12_early	594	32	418	30	70.37%	93.75%
S13_early	595	32	563	29	94.62%	90.63%
S14_early	689	32	489	30	70.97%	93.75%
S15_early	593	32	576	30	97.13%	93.75%
S16_early	594	32	508	32	85.52%	100%
S17_early	601	32	536	29	89.18%	90.63%
Mean:	600.33	-	502.44	29.67	83.80%	92.71%
SD:	21.61	-	62.81	1.11	10.91%	3.45%

Figure A.1: Artifact rejection & channel removal for the early data set

Dataset name	Length (s)	# of channels	Post-artifact rejection length (s)	Post-artifact rejection # of channels	% of length (kept)	% of channels (kept)
S0_late	643	32	389	26	60.50%	81.25%
S1_late	640	32	575	31	89.84%	96.88%
S2_late	683	32	646	29	94.58%	90.63%
S3_late	644	32	414	29	64.29%	90.63%
S4_late	652	32	562	30	86.20%	93.75%
S5_late	639	32	472	29	73.87%	90.63%
S6_late	659	32	506	29	76.78%	90.63%
S7_late	632	32	490	32	77.53%	100%
S8_late	642	32	488	32	76.01%	100%
S9_late	635	32	604	30	95.12%	93.75%
S10_late	682	32	446	29	65.40%	90.63%
S11_late	627	32	559	30	89.15%	93.75%
S12_late	628	32	402	30	64.01%	93.75%
S13_late	640	32	603	29	94.22%	90.63%
S14_late	639	32	420	27	65.73%	84.38%
S15_late	640	32	572	32	89.38%	100%
S16_late	649	32	502	32	77.35%	100%
S17_late	642	32	552	32	85.98%	100%
Mean:	645.33	-	511.22	29.88	79.22%	93.41%
SD:	15.14	-	74.91	1.70	11.43%	5.30%

Figure A.2: Artifact rejection & channel removal for the late data set

B Appendix - Behavioral Test Questions

	Stimuli pair	Order of appearance	MO	Type of question
1	SS1 vs. DS1	SS first	MO1	Tonal learning
2	SS1 vs. DS2	SS first	MO1	Syllabic learning
3	SS1 vs. DS3	SS first	-	Filler
4	SS1 vs. DS4	SS first	-	Filler
5	SS2 vs. DS1	SS first	-	Filler
6	SS2 vs. DS2	SS first	-	Filler
7	SS2 vs. DS3	SS first	MO2	Tonal learning
8	SS2 vs. DS4	SS first	MO2	Syllabic learning
9	SS1 vs. DS1	DS first	MO1	Tonal learning
10	SS1 vs. DS2	DS first	MO1	Syllabic learning
11	SS1 vs. DS3	DS first	-	Filler
12	SS1 vs. DS4	DS first	-	Filler
13	SS2 vs. DS1	DS first	-	Filler
14	SS2 vs. DS2	DS first	-	Filler
15	SS2 vs. DS3	DS first	MO2	Tonal learning
16	SS2 vs. DS4	DS first	MO2	Syllabic learning

Figure B.1: An overview of the questions that were presented in random order during the behavioral test. This list was presented twice, which equals to 32 questions in total for the behavioral test.