Neural Indices of Online Statistical Learning in Visual Speech

Kateřina Kynčlová (kynclova@praha.psu.cas.cz)
Institute of Psychology, Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 00, Praha, Czechia
Department of English Language and ELT Methodology, Faculty of Arts, Charles University, Nám Jana Palacha 2, 116 38 Praha, Czechia

Kateřina Chládková (chladkova@praha.psu.cas.cz)
Institute of Psychology, Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 00, Praha, Czechia
Department of Medical Biophysics, Faculty of Medicine in Hradec Králové, Charles University, Šimkova 870, 500 03, Hradec Králové, Czechia

Abstract

The present study investigated online neural indices of statistical learning of silent speech. Adult participants were exposed to naturally mouthed, silent syllable streams in an artificial language in two conditions. In one condition, 12 syllables occurred randomly, and in the other condition the syllables were structured into four syllable triplets, i.e. statistical words. In the recorded EEG signal, phase synchronisation in neural oscillations was assessed at the rate of syllables and at the rate of words occurring in the exposure streams. Largest phase synchronisation was detected for the word rate during exposure to the structured stream. Moreover, the neural synchronisation to word rate increased throughout the exposure within the structured stream. In a behavioural post-test, however, no learning effects were detected. The EEG results demonstrate sensitivity to statistical regularities in viewed silent speech. These findings indicate that statistical learning in speech and language can be effectively measured online even in the absence of auditory cues.

Keywords: statistical learning; speech processing; silent speech; visual speech; neural tracking.

Introduction

Segmenting words from continuous speech is one of the key abilities that language learners need to acquire in order to master the ambient language. Unlike in writing, in spoken language words are rarely bounded by pauses (Cole, Jakimik, and Cooper, 1980). Instead, word boundaries in continuous speech are cued mainly by prosodic information and/or sequential probability cues. Since the seminal study by Saffran, Aslin, and Newport (1996), it has been repeatedly demonstrated that infants as well as adults track the statistical cues to detect where words begin and end, although with varying success (see Christiansen, 2019; and Frost, Armstrong and, Christiansen, 2019, for review). It has been suggested that even when the speech signal contains prosodic cues to word boundaries, such as word stress placement or word-final lengthening, the statistical information might still be needed (at least) for (some types of) learners to successfully segment words (Saffran et al., 1996a). Statistical learning (SL) thus appears to be a key mechanism in information processing and can be understood as unintentional adaptation to the regularities in the environment (Frost et al., 2019).

SL is not limited only to language, nor to the auditory modality. There is numerous evidence suggesting that SL is to a certain extent available to learners across modalities and domains, and perhaps even across species (Fiser & Aslin, 2001; Lu & Vicario, 2014). Modified versions of Saffran et al.'s embedded pattern paradigm have been used to investigate SL in the auditory and the visual domain, often using shapes as a non-linguistic substitute for linguistic segments. Fiser et al. (2001) demonstrated that replacing artificial word structures with fixed combinations of non-linguistic shapes can yield a similar effect. In their study, passive viewing of complex visual scenes of 12 shapes arranged on a grid resulted in statistical learning in the form of learned single-shaped frequency, absolute shape positions, and shape-pair arrangement statistics, similar to artificial word structures. Similarly, Kirkham, Slemmer, and Johnson (2002) demonstrated that when exposed to statistically predictable patterns, specifically a fixed sequence of coloured shapes, infants habituate and show greater response to novel sequences violating the statistical pattern.

In speech and language, statistical learning is traditionally tested using an exposure-posttest design. Participants are typically exposed to one of two continuous streams of syllables: one stream is structured into repeating trisyllabic words, and the other is unstructured with syllables occurring in a random order (Aslin, Saffran, and Newport, 1998; Batterink et al., 2015; Batterink & Paller, 2017; Choi et al., 2020; Saffran et al., 1996a; Saffran, Newport, and Aslin, 1996). In the structured syllable stream, the only cue to the word boundaries are the transitional probabilities (TPs) between syllables. The TPs are higher between syllables that occur within words than between syllables that occur across word boundaries. After several minutes of exposure, participants are tested on whether they recognize the exposed statistical words (i.e. the word-like units defined by TPs in the structured stream). The test is usually a forced-choice identification task (or a looking paradigm with infants) with only a few trials, in which participants are presented with syllable strings (statistical words from the exposure, part-words, or non-word foils) and have to respond whether they recognize them from the exposure phase (or, rate their familiarity on a likert scale). But what if participants fail to recognize the statistical words over nonwords (i.e. the statistically unrelated foils) in the few posttest trials? Does it mean that they did not track the statistics in the exposure stream?

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

A recent study suggests the effectiveness of statistical learning, at least in adults, is influenced by native-language phonotactic constraints. Dal Ben, Souza, and Hay (2021) reviewed the statistical learning literature and pointed out that in many prior designs, some stimulus items had more native-like phonotactic probabilities than others, and might thus have served as anchors, bootstrapping statistical learning effects. Using an empirical test, the authors demonstrated that learning effects are indeed larger in syllable streams that contain phonotactically plausible anchors over those that do not. Yet, the question remains whether such a native-language driven, anchoring effect, applies only at the level of behavioural, voluntary word decision task, or whether it also affects the preattentive processing of novel syllable streams. Besides the above-described, native-language specific, and perhaps somewhat automatic effect of native phonotactics, performance in statistical learning post-test tasks may be affected by fatigue, lack of attention, or cognitive load associated with having been exposed to rather unnaturally sounding syllable streams.

In order to test the effects of statistical learning directly, without the biases associated with behavioural decisions and task complexity, researchers have come up with methods that can assess statistical learning online, during the exposure phase. In the auditory domain, Buiatti, Peña, and Dehaene-Lambertz (2009) used electroencephalography (EEG) to monitor the learning process during continuous auditory speech, by assessing the correlation between the elicited steady-state response at the word rate and behavioural detection of words (see also Abla, Katahira, and Okanoya, 2008, for an event-related potential, ERP, study on statistical learning in adults). Kabdebon et al. (2015) showed that phase-locking of the neural activity to the artificial speech stream at syllable and word frequencies can be used to test statistical learning in 8-month old infants. Batterink and Paller (2017) used a similar phase-locking paradigm to further assess statistical learning online, throughout exposure to auditorily presented syllable streams. They showed that a measure of phase-locking, the intertrial phase coherence, (ITPC), also referred to as phase-locking value, indexes the brain's tracking of syllables and words, and reflects learning effects online throughout exposure (Batterink & Paller, 2017).

ITPC offers means to quantify the synchronisation of the intrinsic neural oscillations to the external stimulus, such as speech, over several trials. During exposure to (quasi) rhythmic stimuli, neural oscillations temporally align to the periodic or quasi-periodic properties of the stimulus, and phase-reset with the prominent points in the signal (Giraud & Poeppel, 2012; Obleser & Kayser, 2019). In speech, the quasi-periodicity stems from fluctuating linguistic units, most prominently syllables and words, which are tracked simultaneously by neural oscillations at distinct hierarchically organised frequency bands. Apart from other domain-general functions, the frequency bands retain certain specific linguistic functions. The delta band (1-4 Hz) corresponds to speech processing at the word rate, theta band (4-8 Hz) to the syllabic rate, and the gamma band (> 30 Hz) to semantic, syntactic and phonological integration (Tune & Obleser, 2022). The closer the intertrial phase coherence (ITPC) value to 1, the more synchronised, or phase-locked the neural activity.

Neural evidence of SL in the visual domain is provided e.g. by the ERP study of Abla & Okanoya who showed that fixed visual shape triplets were detected by adults in a continuous stream of shapes based on statistical cues only, as demonstrated by elicited larger N400 amplitude at the triplet onset (Abla & Okanoya., 2009). Jost et al. further demonstrated in their ERP study, that after a sufficient exposure to the statistical probabilities of consecutive coloured circles, both adults and children show larger P300 values to highly predictable items (Jost et al., 2015).

The visual modality is exploited to a various extent during speech processing by both hearing and hearing-impaired individuals. Visual cues can immensely improve comprehension, e.g. in a noisy environment (Chládková et al., 2021; Basu Mallick, Magnotti, & Beauchamp, 2015; McGurk & McDonald, 1976). Still, to our knowledge, statistical learning has not yet been tested on spoken visual stimuli. Yet, testing whether SL operates on visual speech alone is needed to better understand the processes underlying the perception, learning, and comprehension of speech in naturalistic – i.e. varying – environments where auditory cues may not always be available as well as in individuals with varying visual-auditory capacities. It has been demonstrated that observing visual speech evokes a similar neural response as listening to auditory speech (Calvert et al., 1997; Hall, Fussell, & Summerfield, 2005; MacSweeney et al., 2000; Park et al., 2016) and is processed differently than non-linguistic lip movements (Calvert et al., 1997, Muthukumaraswamy et al., 2006). Studies investigating the processing of visual speech predominantly focused on processing of existing languages (Bourguignon et al., 2020, Crosse et al., 2015) providing the observer with important context cues, or on the activated brain areas (Muthukumaraswamy et al., 2006, Bourguignon et al., 2020) rather than on the process of perception and segmentation itself.

The present study investigates whether there is evidence of gradual learning of word structures in visual artificial speech in the absence of auditory cues and their possible bootstrapping effect. To this end, we assessed the synchronisation of the neural oscillations to silent visual speech in hearing adults, similarly to what Batterink and Paller (2017) did for auditory speech. In line with Batterink and Paller's (2017) findings for auditory speech, we predicted that the phase-locked neural activity at the word, or syllable triplet, rate of our silent-speech exposure materials will be greater in the structured stream of visual artificial speech than in the random stream with no underlying structure. We further predicted that the sensitivity to the trisyllabic structure will increase during the exposure as a result of statistical learning and will be greater

in the structured speech stream. This paper reports an experiment in which SL in silent speech was assessed in adults with normal hearing. It is a part of a larger study comparing SL in silent speech in hearing, hard-of-hearing and deaf adults, which will be reported elsewhere.

Method

Participants

16 adult participants (*n*=16, 14 women, mean age=29 y, SD=9.41) with normal hearing and no significant prior knowledge of any sign language were included in the present study. All participants reported normal or corrected to normal vision, no neurological or psychiatric disorders, and were not taking any medication or substances affecting the nervous system. Furthermore, there was no evidence of dyslexia or familial risk of dyslexia in any of the participants included in the study. The native language of all participants was Czech. The experiment was approved by the ethics committee of the Institute of Psychology, Czech Academy of Sciences..

Stimuli

For the purposes of the present study, we adapted the stimulus materials used in previous studies on statistical learning in auditory speech (Batterink & Paller, 2017; Podlipský et al., 2022). The syllable set was minimally altered to achieve a visually distinguishable syllable and word sets. Sets of consonants which cannot be visually sufficiently distinguished, such as /m/ and /b/, were avoided. The resulting syllable material consisted of 12 CV syllables, namely *pa, be, ku, da, fo, pi, ti, bu, fe, go, la, tu*. In the structured stream the 12 syllables were arranged into trisyllabic sequences *pabeku*, *dafopi*, *tibufe, golatu*, which were presented in pseudorandom order with no pauses that would indicate a boundary between the triplets. Boundaries between the words were cued solely by transitional probabilities between the syllables, the TPs were 1 for syllables within a triplet, and 0.33 for syllables across triplets. The random stream consisted of the same 12 CV syllables in a pseudorandom order (with occasional occurrences of mispronounced syllables). The order of the syllables in the random stream could not be predicted based on the transitional probabilities or any other cues.

The stimuli were presented in a video-only format showing the head and part of the upper body of a speaker silently mouthing the syllables. The stimuli were recorded by a young woman, a native speaker of Czech who read the syllable streams from a computer stream while listening to a metronome (at the rate of 200 bpm corresponding to a 3.3-Hz syllable rate). The speaker was instructed to read the presented stimuli set with natural articulation and in alignment with the rhythm set by a metronome. Parts in which the speaker stumbled or paused were extracted offline. The audio track was subsequently removed from the recordings.

Procedure

Exposure Phase In the exposure task, participants were presented with videos of the structured and the random stream, with the order of conditions counterbalanced across participants. Each condition consisted of 6 minutes of continuous visual speech stream divided into 3-minute blocks separated by a short break. A post-exposure word recognition and rating task was always administered immediately after the structured condition. Participants were instructed to attend to the silent video and focus on the speaker's mouth. During the brief break between the blocks, participants were advised to close their eyes or change the centre of their visual focus to avoid strained eyes. The two conditions were separated by a longer 4-minute break (in case of structured condition first, the break followed after the word recognition task), during which the participants were instructed to disassemble and reassemble a wooden puzzle to further rest their eyes from viewing the computer screen and to keep them alert.

Post-exposure Task The post-exposure task tested participants' word recognition accuracy and collected word rating scores. In the post-test, participants were presented with video clips of the same speaker as in the exposure materials. On each trial they saw a video of a silently mouthed trisyllabic statistical word that repeatedly appeared in the structured stream, or a trisyllabic non-word sequence that did not appear in the exposure stream. Prior to viewing the video clips, participants were instructed via a prompt on the screen to indicate for each clip whether the speaker has said it before or not. The word recognition task consisted of 12 trials.

EEG Recording During the exposure phase, EEG was recorded at a sampling rate of 200 Hz from 19 scalp electrodes attached to an electrode cap (with an additional FCz channel serving as an online reference). Additional five electrodes were placed on the participant's nose, on the outer canthus of the right eye, under the right eye, and on the mastoid bones behind the right and the left ear. Recordings were made with the TruScan software (Deymed diagnostic s.r.o.).

Analysis

The EEG data analyses follow the procedures reported by Batterink and Paller (2017), with a few adaptations (e.g. in the time-locking syllable used for the EEG epoching, in us non-transforming ITPC values to logarithms).

EEG Data Analysis The EEG data analysis was carried out in Matlab (version R2023a, The MathWorks Inc., 2022) using the EEGLAB toolbox (Delorme & Makeig, 2004). The EEG data was band-pass filtered from 0.1 Hz to 30 Hz and epoched into 10.8-sec long segments relative to the onset of each word in the structured condition, and to every first syllable of a syllable triplet in the random condition.

Epochs containing large artefacts (exceeding absolute amplitude of 350 μV) were removed. Further analyses were done for seven channels: three posterior/occipital channels, Pz, O1, O2, and four temporal channels, T3, T4, T5, T6, as previous literature suggests these areas may most strongly reflect statistical regularity learning in the visual speech domain (Capek et al., 2008; MacSweeney et al., 2002).

Inter-trial phase coherence (ITPC) was computed using a continuous Morlet wavelet transform in 200 sliding windows across the 10.8-s epoch. The transform was run for 0.1-Hz frequency bins from 0.2 to 20.2 Hz, with 1 cycle at the lowest frequency and increasing by a factor of 0.5 for the higher frequency bins (with 25.5 cycles at the highest frequency). Per-participant average ITPC values at 3.3 Hz and at 1.1 Hz, corresponding to the syllable and word rates, respectively, were then extracted for further analyses.

We further computed the word learning index (WLI), quantified as the ITPC at the word frequency divided by the ITPC at the syllable frequency. A WLI value larger than 1 would indicate more robust tracking of words over syllables, and vice versa for WLI below 1. Also, the higher the WLI the higher the neural sensitivity to tracking words over syllables.

Behavioural Data Analysis Data from the post-exposure rating task were used to compute each participant's word recognition accuracy. The recognition accuracy was computed as the percentage of trials in which the participant correctly recognized a statistical word or correctly rejected a nonword.

Statistical Analysis The ITPC and the WLI were analysed with linear mixed-effects models using the package *lme4* (Bates et al., 2015) in R (version 4.3.1, R Core Team, 2021). In all models, fixed factors were estimated with sum-to-zero coded contrasts.

In the model for ITPC, the fixed factors were condition (random vs structured), chunking rate (syllable vs word), block (1st vs 2nd), and their interaction; a random intercept for participant was included (including also a random intercept for channel lead to a singular fit). In the model for WLI, the fixed factors were condition (random vs structured) and block (1st vs 2nd), and their interaction; per-participant and per-channel random intercepts were estimated. For pairwise comparisons, package ggeffects (Lüdecke, 2018) was used to estimate means and confidence intervals.

Behavioural rating accuracy scores were analysed using a one-sample t-test against chance level, i.e. against μ = 0.5.

Results

The ITPC model detected a significant main effect of condition, a main effect of block, a non-significant interaction of condition and block, and a significant triple interaction of condition, block, and chunking rate. This triple interaction is visualised in Figure 2 (left). Pairwise comparisons show that ITPC to words in the second block of structured condition was larger than ITPC for all other factor levels. The WLI model yielded a significant interaction of condition and block. Pairwise comparisons reveal that in the structured condition WLI was greater in the second than in the first exposure block, and that the WLI in the second exposure block of the structure condition was also larger than the WLI in either block of the random condition; see Figure 2 (right). The results of WLI are in line with the effects seen for ITPC for the two different chunking rates.

The grand average ITPC per condition, per chunking rate, and per exposure block is plotted in Figure 1. The fixed-effect model summaries are shown in Table 1 and 2.

Table 1: ITPC fixed-effect model summary.

	Est	SE	df	t	p
(Intercept)	0.08	0.004	15	17.96	<.001
condition (-rnd+str)	0.002	0.001	873	2.466	.014
chunking (-syl+wd)	0.001	0.001	873	1.293	.196
block(-1+2)	0.003	0.001	873	2.746	.006
cond:chunk	0.002	0.001	873	1.926	.054
cond:block	0.002	0.001	873	0.151	.880
chunk:block	0.002	0.001	873	1.631	.103
condition:chunking:block	0.003	0.001	873	2.636	.009

Table 2: WLI fixed-effect model summary.

	Est	SE	df	t	p
(Intercept)	1.126	0.052	13.3	21.527	<.001
condition (-rnd+str)	0.027	0.026	423	1.020	.308
block(-1+2)	0.046	0.026	423	1.729	.085
cond:block	0.058	0.026	423	2.196	.029

A t-test on the rating accuracy data did not detect a difference from the chance level (x = 44.75, 95% c.i. = 37.07–52.43, t = -1.457, df = 15, p = 0.166).

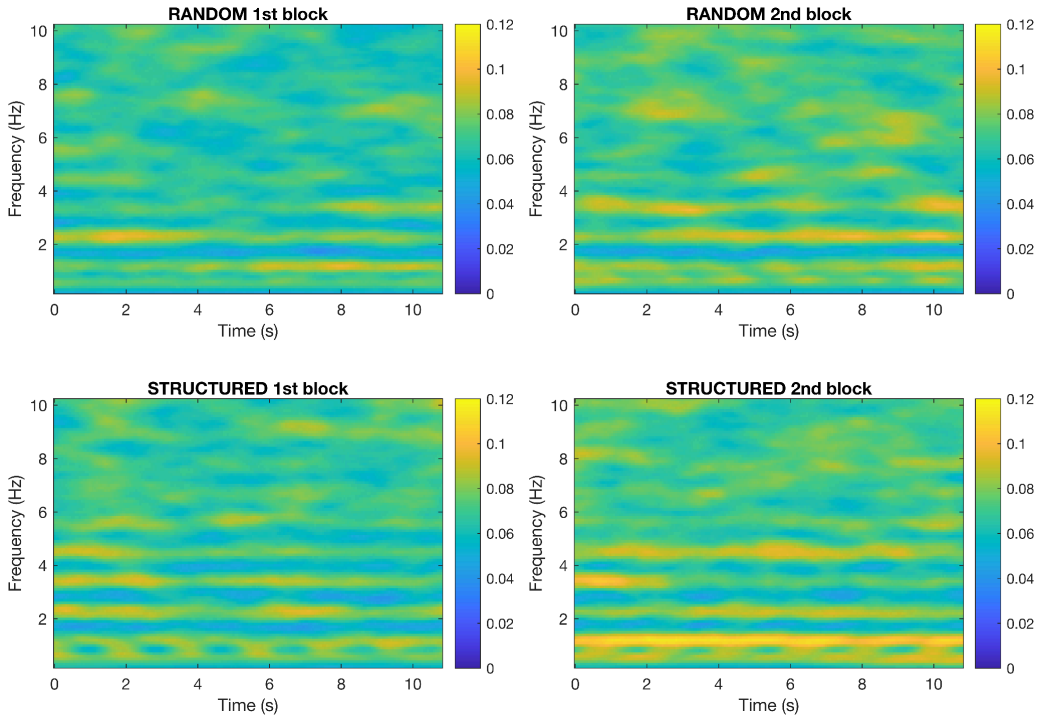

Figure 1: ITPC values in the first and the second exposure block of each condition, averaged across the 7 analysed channels (Pz, O1, O2, T3, T4, T5, T6) and across epochs.

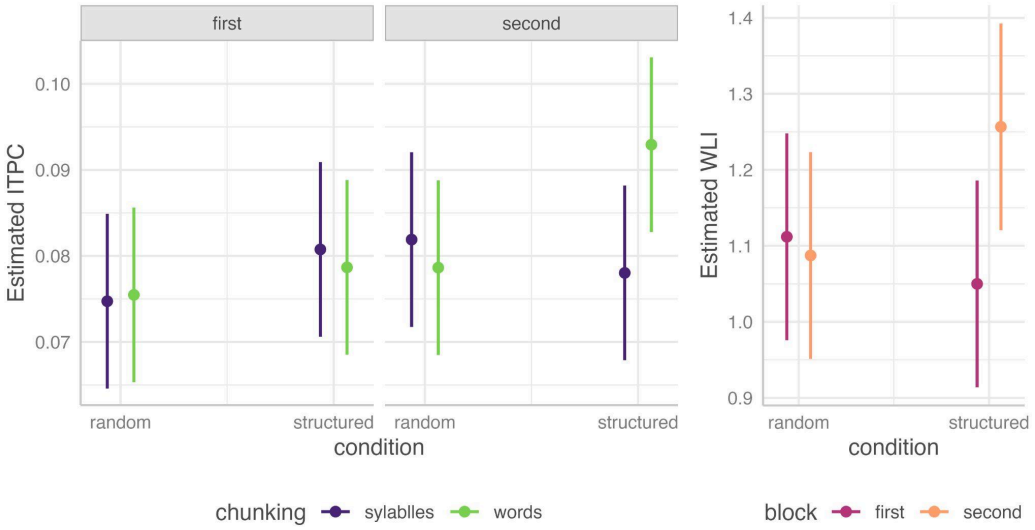

Figure 2: Left: Estimated ITPC per condition, exposure block, and chunking rate. Right: Estimated WLI per condition and exposure block (means and 95% confidence intervals).

Discussion

The present study investigated the online process of statistical learning (SL) in visual silent speech processing. Statistical learning is a mechanism that aids humans (and some nonhuman animals) uncover regularities and perceptually organise their environment. It has been shown to apply across modalities and domains, including spoken language (e.g. Aslin, et al., 1998; Batterink et al., 2015; Batterink & Paller, 2017; Choi et al., 2020; Fiser & Aslin, 2001; Handel & Buffardi, 1969; Lu & Vicario, 2014; Saffran et al., 1996a; Saffran et al, 1996b; Turk-Browne et al., 2009). In language development, it has been taken as one of the main mechanisms that children have available early and by which they figure out the word forms in their native language, the phonotactic patterning as well as sentence structures (Marcus et al., 1999; Saffran et al., 1996a). While statistical learning is a powerful mechanism allowing parsing the continuous auditory speech into words, it is unclear whether or to what extent it applies to speech without auditory cues, i.e. silent speech.

We tested whether adult brains exploit the statistical regularities in speech devoid of the audio signal, namely, in visual silent speech. We adapted the auditory speech SL design of Batterink and Paller (2017) for the visual modality and employed the EEG, which offers a real-time depiction of the gradual learning process that may take place during exposure to the novel artificial language. We hypothesised that statistical regularities in visual silent speech will elicit similar learning mechanisms reported for auditory speech and will enable parsing and segmentation into the underlying units. The present results showed that the ITPC at the word frequency increased significantly in the second block of the structured condition. Moreover, we found a greater WLI value in the structured stream compared to the random stream which indicates a higher sensitivity to the underlying trisyllabic structure in the structured condition. WLI significantly increased over the course of the first and second block in the structured condition which indicates that the sensitivity to the underlying structures increases with exposure as a result of the gradual process of statistical learning. This indicates that the statistics underlying the structured syllable stream have been detected and gradually acquired (at least at the preattentive level of processing), thus supporting our prediction of statistical learning in visual speech. Statistical learning appears to be a vital mechanism underlying speech perception regardless of its modality.

Despite the evidence of successful SL in the structured condition demonstrated by the ITPC and WLI values recorded with the EEG, the post-exposure behavioural word-recognition task did not reveal any effects of learning. The present results indicate that statistical learning occurs in artificial visual speech as demonstrated by the online monitoring of SL, despite there being no effect of SL shown in the post-exposure word-recognition task. This finding suggests that the brain is able to synchronise to speech and discover the underlying statistical patterns in it almost instantly within several minutes even in the absence of auditory cues. It thus remains to be shown what kind of exposure or how much input is needed for the adapted neural processing of word structures in novel (silent) speech to be reflected in behaviour.

The results support the idea that the effects of statistical learning might not be automatically reflected in post-exposure behavioural tasks perhaps due to cognitive limitations or inadequateness of the testing methods, at least for a certain type of input. Using neuroimaging methods to monitor statistical learning online might provide valuable new insights into the availability of learning mechanisms across input modalities and populations.

Although our results show significantly larger effects of learning in the structured condition compared to the random condition at the neural level, we cannot clearly pinpoint how and when the effects are reflected in behaviour. However, neuroimaging research shows that higher language proficiency leads to stronger neural phase-locking to the speech input, namely to the frequency of the salient linguistic units, such as words and intonational phrases, increasing the success rate of comprehension (Ding et al., 2016; Peelle & Davis, 2012). Successful neural tracking of speech is thus likely to be a prerequisite of effective language learning.

Since familiarity with the input immensely affects the process of learning, it is possible that one's primary or native modality might affect not only the detecting of the underlying structures but also their successful storing in the memory. Experiments investigating this effect in hearing-impaired and deaf adults are currently underway. Our preliminary results show that hearing-impaired participants perform better in the post-exposure word-recognition task, suggesting that familiarity with the input's modality might affect the success rate of storing the resulting statistical items or might decrease the time needed for a successful storing.

As noted in the Methods, during the recording of the stimuli, the speaker mispronounced several syllables in both streams. Due to its statistical nature, all mispronunciations had to be removed from the structured stream but some were kept in the random stream since frequent cuts would be more disrupting than the occasional mispronunciations. The unbalanced amount of mispronunciations between conditions might have affected the result. Recording of the individual syllables and their subsequent combination into a stream might solve this issue, but would yield unnaturally-looking visual speech. In the present study, rather than having a fully controlled but unnaturally looking concatenation of individual syllables, we opted for natural syllable streams albeit with cuts or occasional mispronunciations. For the future comparison across populations differing in hearing status and primary modality of communication, the potential additional modulating factor should not be problematic since the stimulus materials will be identical for all the tested groups.

Acknowledgments

This work was supported by the Czech Science Foundation, grant no. 21-09797S, by the Czech Academy of Sciences, project no. LQ300252401, and by the European Regional Development Fund, project Brain Dynamics, reg.no. CZ.02.01.01/00/22_008/0004643.

References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*(6), 952–964.

Abla, D., & Okanoya, K. (2009). Visual statistical learning of shape sequences: An ERP study. *Neuroscience Research, 64*(2), 185–190.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-Month-Old infants. *Psychological Science*, *9*(4), 321–324.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, *67*(1), 1–48.

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78.

Basu Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307.

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-Reading enables the brain to synthesize auditory features of unknown silent speech. *The Journal of Neuroscience*, *40*(5), 1053–1065.

Buiatti, M., Peña, M., & Dehaene‐Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage*, *44*(2), 509–519.

Calvert, G. A., Bullmore, E. T., Brammer, M., Campbell, R., Williams, S., McGuire, P., Woodruff, P., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596.

Capek, C. M., MacSweeney, M., Woll, B., Waters, D., McGuire, P., David, A. S., Brammer, M., & Campbell, R. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia*, *46*(5), 1233–1241.

Choi, D., Batterink, L. J., Black, A. K., Paller, K. A., & Werker, J. F. (2020). Preverbal infants discover statistical word patterns at similar rates as adults: evidence from neural entrainment. *Psychological Science*, *31*(9), 1161–1173.

Christiansen, M. H. (2019). Implicit Statistical Learning: a tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468–481.

Chládková, K., Podlipský, V. J., Nudga, N., & Šimáčková, Š. (2021). The McGurk effect in the time of pandemic: Age-dependent adaptation to an environmental loss of visual speech cues. *Psychonomic Bulletin & Review*, *28*(3), 992–1002.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1980). Segmenting speech into words. *Journal of the Acoustical Society of America*, *67*(4), 1323–1332.

Crosse, M. J., ElShafei, H. A., Foxe, J. J., & Lalor, E. C. (2015). Investigating the temporal dynamics of auditory cortical activation to silent lipreading. *7th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 308-311). Montpellier, France.

Dal Ben, R., De Hollanda Souza, D., & Hay, J. F. (2021). When statistics collide: The use of transitional and phonotactic probability cues to word boundaries. *Memory & Cognition*, *49*(7), 1300–1310.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 134*(1), 9-21.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164.

Fiser, J., & Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. *Psychological Science*, *12*(6), 499–504.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12).

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading Fluent Speech from Talking Faces: Typical Brain Networks and Individual Differences. *Journal of Cognitive Neuroscience*, *17*(6), 939–953.

Handel, S., & Buffardi, L. C. (1969). Using Several Modalities to Perceive one Temporal Pattern. *Quarterly Journal of Experimental Psychology*, *21*(3), 256–266.

Jost, E., Conway, C. M., Purdy, J., Walk, A. M., & Hendricks, M. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. *Brain Research*, *1597*, 95–107.

Kabdebon, C., Peña, M., Buiatti, M., & Dehaene‐Lambertz, G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and Language*, *148*, 25–36.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software*, *82*(13), 1–26.

Lü, K., & Vicario, D. S. (2014). Statistical learning of recurring sound patterns encodes auditory objects in songbird forebrain. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(40), 14553–14558.

Lüdecke, D. (2018). "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." *Journal of Open Source Software*, *3*(26), 772.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., Williams, S., Woll, B., & Brammer, M. (2000). Silent speechreading in the absence of scanner noise. *NeuroReport*, *11*(8), 1729–1733.

MacSweeney, M., Calvert, G. A., Campbell, R., McGuire, P., David, A. S., Williams, S., Woll, B., & Brammer, M. (2002). Speechreading circuits in people born deaf. *Neuropsychologia*, *40*(7), 801–807.

Marcus, G., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by Seven-Month-Old infants. *Science, 283*(5398), 77–80.

McGurk, H., & Macdonald, J. B. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Muthukumaraswamy, S., Johnson, B. W., Gaetz, W., & Cheyne, D. (2006). Neural processing of observed oro-facial movements reflects multiple action encoding strategies in the human brain. *Brain Research*, *1071*(1), 105–112.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences*, *23*(11), 913–926.

Park, H., Kayser, C., Thut, G., & Groß, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*.

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology, 3*: 320.

Podlipský, V.J., Chládková, K., Paillereau, N., Šimáčková, Š. Native variety influence on speech segmentation in a novel language. In Carlet, A. et al (Eds.) *Book of Abstracts. New Sounds 2022.* Barcelona, p. 133.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-Month-Old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

The MathWorks Inc. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc.

Tune & Obleser (2022): Chapter on Neural Oscillations in Speech Perception. Retrieved from osf.io/6b7eg

Turk‐Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural Evidence of Statistical Learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4.