What you see is what you hear: How visual prosody affects artificial language learning in adults and children

Jaspal Brar, Michael D. Tyler, and Elizabeth K. Johnson

Citation: Proc. Mtgs. Acoust. **19**, 060068 (2013); doi: 10.1121/1.4800523

View online: https://doi.org/10.1121/1.4800523

View Table of Contents: https://asa.scitation.org/toc/pma/19/1

Published by the Acoustical Society of America

Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication

Session 2aSC: Linking Perception and Production (Poster Session)

2aSC24. What you see is what you hear: How visual prosody affects artificial language learning in adults and children

Jaspal Brar*, Michael D. Tyler and Elizabeth K. Johnson

***Corresponding author's address: Psychology, University of Toronto, 3359 Mississauga Road N, Mississauga, L5L 1C6, Ontario, Canada, pauly.brar@mail.utoronto.ca**

Speech perception is a multimodal phenomenon, with what we see impacting what we hear. In this study, we examine how visual information impacts English listeners' segmentation of words from an artificial language containing no cues to word boundaries other than the transitional probabilities (TPs) between syllables. Participants (N=60) were assigned to one of three conditions: Still (still image), Trochaic (image loomed toward the listener at syllable onsets), or Iambic (image loomed toward the listener at syllable offsets). Participants also heard either an easy or difficult variant of the language. Importantly, both languages lacked auditory prosody. Overall performance in a 2AFC test was better in the easy (67%) than difficult language (57%). In addition, across languages, listeners performed best in the Trochaic Condition (67%) and worst in the Iambic Condition (56%). Performance in the Still Condition fell in between (61%). We know English listeners perceive strong syllables as word onsets. Thus, participants likely found the Trochaic Condition easiest because the moving image led them to perceive temporally co-occurring syllables as strong. We are currently testing 6-year-olds (N=25) with these materials. Thus far, children's performance collapsed across conditions is similar to adults (60%). However, visual information may impact children's performance less.

Published by the Acoustical Society of America through the American Institute of Physics

Received 25 Jan 2013; published 2 Jun 2013

INTRODUCTION

Artificial language studies have shown that native language experience shapes the way listeners segment words from speech (e.g. Tyler & Cutler, 2009). For example, most content words in English carry word initial stress (e.g. Cutler & Carter, 1987). Accordingly, if you present English learners with an artificial language that contains stress aligned with the onsets of word boundaries, they will find words more accurately than if they are presented with stress in word final position (e.g. Johnson & Seidl, 2009).

At the same time, speech perception is a multimodal phenomenon, with what we see influencing what we hear. As such, visual information has been shown to aid in the segmentation of an artificial language. For example, transitional probabilities between syllables (henceforth TPs) are learned more readily when routinely accompanied by a visual referent (Thiessen, 2010; Yurovsky, Yu, & Smith, in press). And the presence of any visual cue marking the onset of either the beginning or end of words in an artificial language makes it easier to identify words in the speech stream (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010).

In the current study, we examine how visual information impacts English listeners' segmentation of words from an artificial language containing no auditory cues to word boundaries other than transitional probabilities between syllables. More specifically, we ask whether a visual movement accompanying particular syllables in a speech stream can lead listeners to perceive those syllables as more prominent than others, and thus make the language either easier or more difficult to learn depending on the alignment of the visual cue with word boundaries. We also manipulate language difficulty to investigate how heavily the visual information influences segmentation abilities in easy versus difficult artificial language learning tasks. We propose that although all participants will track transitional probabilities between syllables, adults may do so more accurately than children, at least with the more difficult language. Moreover, if syllables aligned with visual movement are perceived as stressed, then participants will perform best when visual stress occurs with syllable onsets, consistent with the trochaic stress patterns of English.

METHOD

Participants

Forty-three native English speaking adults were tested in this study (mean age = 22 years; range = 17 years – 57 years; 16 males and 27 females). Consent was obtained from all participants. Course credit or monetary compensation was provided for their participation. Twenty-five English speaking 6 year olds were also tested (mean age: 6 years, 10 months; range = 5 years, 11 months – 7 years, 6 months; 11 males and 14 females). Parental and participant consent was obtained. Participants received a Junior Scientist Certificate along with a toy for their participation.

Materials

Twelve CV syllables were recorded in isolation by a young female from Ontario. The syllables were edited in PRAAT to ensure their relative uniformity in pitch, amplitude, and duration. Four artificial languages (two easy languages and two difficult languages), each containing all bisyllabic words, were constructed from the twelve CV syllables. Language difficulty was defined using within word and between syllable transitional probabilities (TP's). The easy language contained 6 words with a within word TP of 1.0 and a between syllable TP of 0.17. The difficult language contained 12 words with a within word TP of 0.5 and a between syllable TP of 0.17. The easy and difficult languages were counter-balanced so that the words of one language formed the partwords of the other language. That is, a syllable that occurred word-initially in one language occurred word-finally in another language. The familiarization streams were created by stringing together the words of each language in a randomized order. The speech streams contained no pauses or any other auditory cues to word boundaries besides TPs between syllables. However, each of the four languages was paired with both of two types of visual prosody: word initial stress (henceforth trochaic) or word final stress (henceforth iambic). For the trochaic condition, a character on the screen rapidly loomed towards the participant in synchrony with the occurrence of the first syllable of every word in the language (no movement occurred to accompany the second syllable of the word). For the iambic condition, a character on the screen loomed toward the participant in an identical manner in synchrony with the last syllable of each word (no movement occurred in synchrony with the occurrence of the first syllable in every word). Past studies have demonstrated that when speakers describe the movement of a visual stimulus, the prosody of their utterances often conveys information about the type of movement they are observing. For example, when people see a dot rise on a screen, they are more likely to use a rising pitch to describe its movement than when they see that same dot sink to the bottom of the screen (e.g. Shintel, Nusbaum, & Okrent, 2006). And it is well known that what we hear is influenced by what we see (e.g. McGurk & MacDonald, 1976). Thus, we reasoned that seeing a rapidly approaching visual stimulus accompanying an auditory syllable would lead that syllable to be perceived as more perceptually pronounced. That is, the languages accompanied by trochaic visual prosody would be perceived as having word-initial stress whereas the languages accompanied by iambic visual prosody would be perceived as having word-final stress. Since our participants spoke English, a language characterized by word initial stress (Cutler & Carter, 1987), we predicted that the languages accompanied by trochaic visual prosody would be easier to learn than the languages accompanied by iambic visual prosody.

Procedure

The experiment consisted of two phases: familiarization and test. During the familiarization phase, participants watched 3 four-minute videos while listening to an artificial language. In each video, participants saw a character who loomed toward the listener in synchrony with either the first (trochaic condition) or last syllable (iambic condition) of every word in the language (see Figure 1). To avoid fatigue and encourage participants to pay close attention to the character in the video, participants were given brief breaks in between each video and asked to draw a picture of the character they had been viewing for the last four minutes. Participants did not receive any information about the length or the structure of words during the familiarization phase. They were simply told to watch each video and listen to the language. After completing the familiarization phase, participants immediately completed a 2 Alternative Forced Choice Test (2AFC). Participants heard a sequence of two words, and were asked to determine which word belonged to the language they had been familiarized to. On each trial participants were presented with one item that was a word in the language that they were familiarized with and one item that was a partword (last syllable of one word plus the first syllable of another word). The word item was presented equally often in first and second position. Adult participants circled their responses on paper, while children provided verbal responses which were recorded by the experimenter.

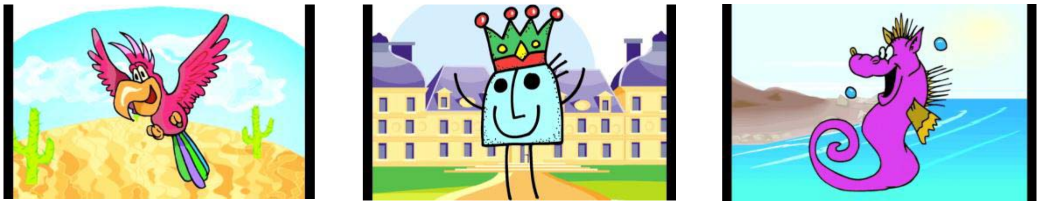

Figure 1. Screen shots of animated characters presented during the familiarization phase, in sequential order from left to right. Each character zoomed toward the participant in synchrony with either the initial (trochaic condition) or final (iambic condition) syllable of each word in the familiarization speech stream.

Design

All participants were randomly assigned to one of two visual stress conditions (trochaic or iambic) and one of two language types (easy or difficult). Thus, this experiment included four different familiarization conditions: trochaic easy, iambic easy, trochaic hard, and iambic hard. Test items were the same for participants in the four different conditions. Adults were presented with two blocks of trials, for a total of 24 test trials. Due to limitations in their attention spans, children were presented with only one block of 12 test trials. In the current report, we focus our analysis on only the first 12 trials to facilitate comparison between adults and children.

RESULTS

Percent correct scores were calculated for all participants, with chance performance equal to 50%. Collapsing across conditions, children's overall performance (60%) was no different than adults' (62%), *t*(66) = .43, p = .67. Although contrary to our predictions, this could be seen as fitting well with what is in the literature (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). Next, we compared children's and adult's performance on the easy versus difficult language (see Figure 2). Both children and adults performed significantly better with the easy language than the difficult language (adults: *t*(41)=3.22, p=.003; children: *t*(23)=2.23, p=.04). When we collapsed across age group, participants performed above chance in the easy language [*t*(35)=6.4, p <.0001], but only marginally above chance in the difficult language [*t*(31)=1.4, p=.08].

Next, we turned our attention to the influence of visual prosody on participants' performance. Here, we observed dramatic differences in adult's and children's performance and in trochaic and iambic stress conditions (see Figure 2). Adult performance in a 2AFC test was better in the trochaic (67%) versus iambic (56%) conditions *t*(41) = 2.09, p > 0.04). Children did not show a significant difference in their performance between trochaic (60%) and iambic (60%) visual stress conditions, *t*(23)=.004, p=.99.

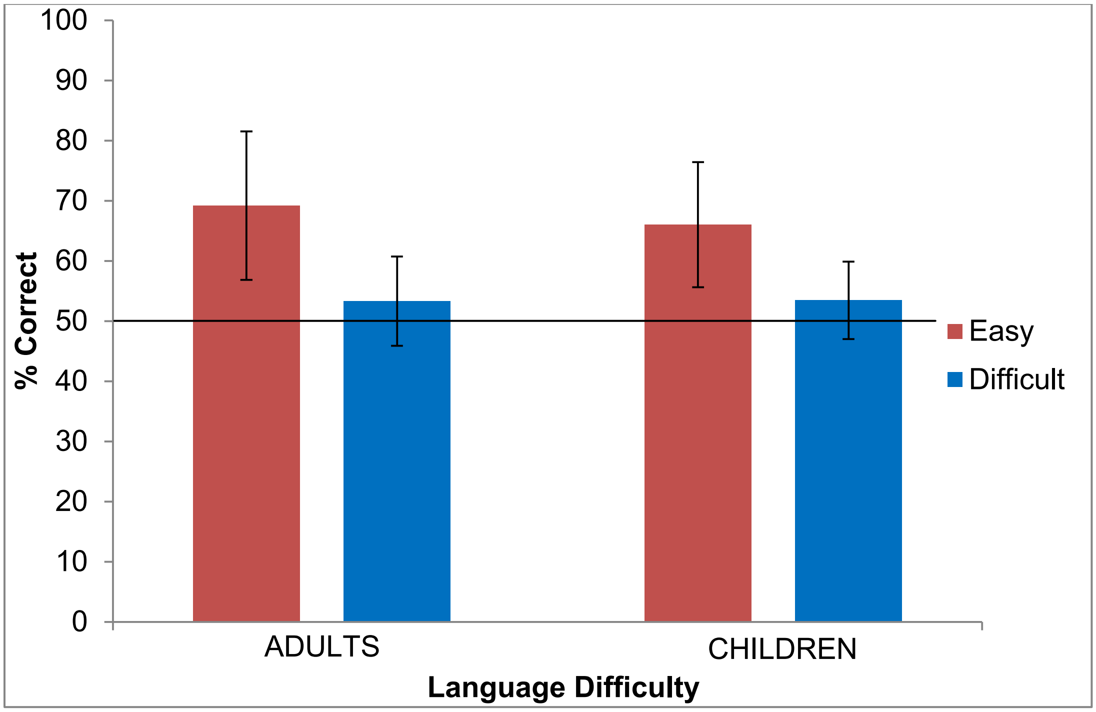

FIGURE 2. Mean performance of adults and children in a 2AFC test broken down by language difficulty. Chance performance is shown at 50%. Error bars indicate Standard Deviation.

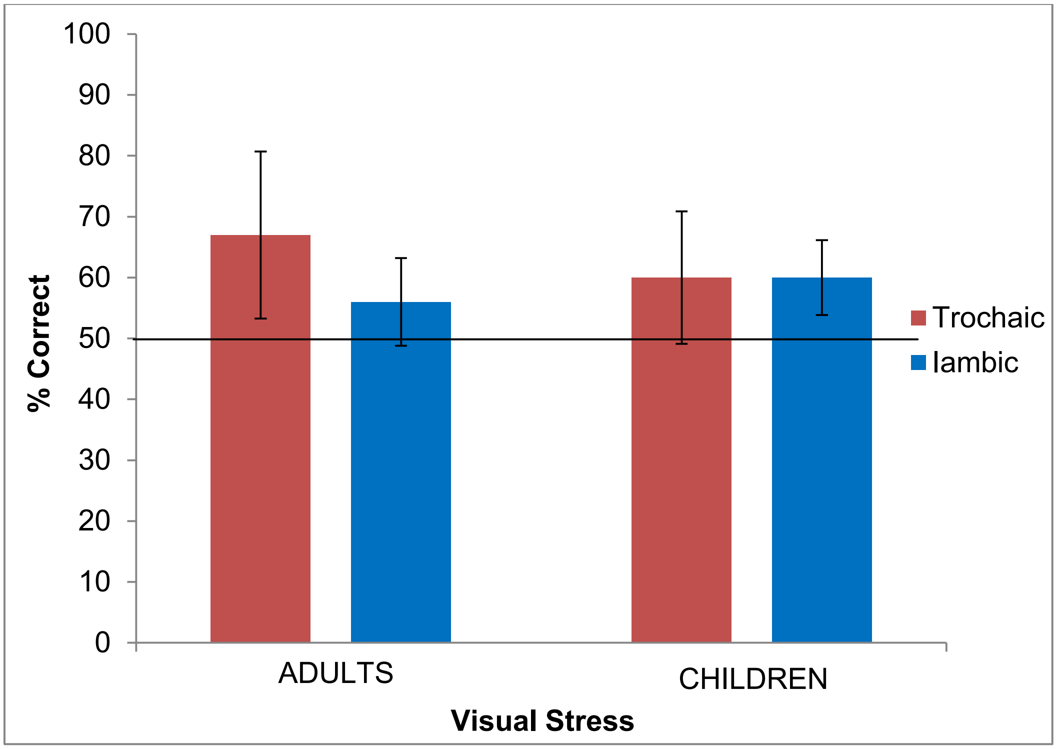

FIGURE 2. Mean performance of adults and children in a 2AFC test broken down by visual stress. Chance performance is shown at 50%. Error bars indicate Standard Deviation.

DISCUSSION

In the current study we tested two hypotheses. First, adults should be better at tracking transitional probabilities between syllables than children. Second, all English speakers should segment an artificial language better when provided with a trochaic rather than iambic visual stress cue. Surprisingly, our results failed to fully support either of these predictions. First, adults and children learned both the easy and difficult artificial languages equally well. Second, only adults' segmentation performance benefited from the presence of trochaic visual stress. Children performed identically regardless of where the visual stress was placed. Thus, our second hypothesis was only partially supported. Below, we consider both of these interesting findings in turn.

The finding that children track TPs between syllables as well as adults fits well with the notion that children depend heavily on statistical learning mechanisms to acquire language (Saffran, Werker, & Werner, 2006). However, it is worth noting that in this particular study, it is not the case that children performed better than expected so much as it is the case that adults performed worse than expected. The difficult language in this study only contained 12 words and the TPs defining word boundaries, although not as clear as those in our easy language, were surely far clearer than those defining word boundaries in natural language. Therefore, we expected adults to readily learn both the easy and difficult languages. We expected children to potentially have trouble learning the difficult language. Instead, we found that both adults and children struggled to learn the difficult language. This finding could be interpreted as evidence that listeners may not track TPs in speech as skilfully as past studies have suggested (Johnson & Tyler, 2010).

We found that adult English speakers perform better when visual stress is aligned with syllable onsets rather than syllable offsets. Since the artificial languages in this study lacked auditory prosody, we conclude that the moving image led adults to perceive temporally co-occurring syllables as stressed. A good way to test this hypothesis further might be to repeat the study with native speakers of another language that does not have a trochaic stress pattern, such as French. The fact that children's performance appeared to be unrelated to the type of visual prosody they received in this study is difficult to explain and deserves further investigation. Perhaps the results would have differed had we used a more ecologically valid cue to visual stress, such as a cartoon voice that opened very widely on either initial or final syllables.

ACKNOWLEDGEMENTS

This research has been supported by NSERC and SHHRC grants awarded to EKJ and an NSERC USRA awarded to JB. We thank Dena Krieger, Caterina Minaudo, Juwairia Sohail, and Joshua Romano for assistance in testing participants.

REFERENCES

Cunillera, T., Càmara, E., Laine, M., and Rodríguez-Fornells, A. **(2010)**. "Speech segmentation is facilitated by visual cues," The Quarterly Journal of Experimental Psychology, **63**, 260-274.

Cutler, A., and Carter, D. M. (**1987**). "The predominance of strong initial syllables in the English vocabulary," Computer Speech and Language, **2**, 133-142.

Johnson, E.K., and Seidl, A. **(2009)**. "At 11 months, prosody still outranks statistics," Developmental Science, **12,** 131-141.

Johnson, E.K., and Tyler, M. (**2010**). "Testing the limits of statistical learning for word segmentation," Developmental Science, **13**, 339-345.

McGurk, H., and MacDonald, J. **(1976)**. "Hearing lips and seeing voices," Nature, **264**, 746-748.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (**1997**). "Incidental language learning: Listening (and learning) out of the corner of your ear," Psychological Science, **8**, 101-105.

Saffran, J.R., Werker, J., and Werner, L. (**2006**). "The infant's auditory world: Hearing, speech, and the beginnings of language," in *Handbook of Child Development,* edited by R. Siegler and D. Kuhn (Wiley, New York, 2006), pp. 58-108.

Shintel, H., Nusbaum, H. C., and Okrent, A. (**2006**). "Analog acoustic expression in speech communication," J. of Memory and Language, **55**, 167-177.

Thiessen, E. D. **(2010)**. "Effects of visual information on adults' and infants' auditory statistical learning," Cognitive Science, **34**, 1092-1106.

Tyler, M., and Cutler, A. **(2009)**. "Cross-language differences in cue use for speech segmentation," J. of the Acoustical Society of America, **126**, 367-376.

Yurovsky, D., Yu, C., and Smith, L. B. (in press). "Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech," Frontiers in Language Sciences.