Auditory-Visual Speech Segmentation in Infants

S. H. Jessica Tan, Denis Burnham

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

j.tan@westernsydney.edu.au, denis.burnham@westernsydney.edu.au

Abstract

Speech segmentation, breaking the heard speech stream into words, is necessary for language acquisition. Visual prosody, like acoustic prosody, aids speech segmentation in adults [1], [2]. By contrast, surprisingly little is known about how visual speech information influences speech segmentation in infants despite the important role that speech segmentation plays in language development and past research demonstrating that young infants can segment auditory-only speech. Further, studies on infants' gaze behavior to the eye and mouth regions of the speaker's face have found that infants perceive the mouth region as an important conveyor of articulatory information [3]. Such evidence suggests two hypotheses: (i) that infants should benefit from visual speech information in word segmentation, and (ii) any visual speech benefit should be related to greater gaze directed to the speaker's mouth than the eyes. This study investigated whether (1) 7.5-month-old infants' speech segmentation differed between auditory-only and auditory-visual conditions, and (2) gaze behavior modulated segmentation performance. Preliminary analyses reveal better segmentation performance in the auditory-visual condition that may be accounted for by greater attention on the speaker's mouth.

Index Terms: speech segmentation, visual speech information, auditory-visual speech perception

1. Introduction

The ability to segment continuous speech is an essential skill for language acquisition. To acquire a vocabulary, young infants must first be able to identify where one word ends and another begins. As the temporal patterns of mouth movements are similar to the acoustic timescale of syllables, it is likely that visual information from the speaker's face, such as mouth movements, assists with segmentation by providing useful information regarding the start and end points of syllables [4]. Studies with adults have demonstrated better performance of artificial speech segmentation when speech was paired with synchronous videos of a speaker's talking face compared to when the artificial language stream was presented only in auditory modality, suggesting that visual speech information aid adults' segmentation of a newly learnt artificial language [1], [2]. These findings with adults raise the question of whether visual speech information has the same enhancing effect on infants' speech segmentation.

The majority of infant speech segmentation studies have been conducted with auditory-only speech [5], [6], [7], and [8] even though studies on infants' gaze behavior [3] have established that infants perceive the mouth as an important conveyor of articulatory information. To date, only one study has examined whether the presentation of a speaker's talking face augments infants' speech segmentation [9]. That study found that 7.5-month-old infants segmented words from passages that were blended with a distractor voice when the passages were paired with the speaker's talking face, suggesting that visual information from a speaker's talking face enhances infants' segmentation of a fluent speech stream in the presence of background noise.

In addition, it has been found that infants as young as six months are already seeking linguistic information from the mouth; 6- to 12-month-olds fixate more on the mouth when the speaker is talking than when the speaker is only smiling [10]. Given this evidence, it stands to reason that visual speech information from a speaker's face may facilitate infants' speech segmentation. Whether this is indeed the case remains unclear.

The overall aim of the current study is to investigate whether the addition of visual speech information from a speaker's talking face augments 7.5-month-olds' segmentation of fluent speech. A secondary aim was to examine if individual differences in fixation durations to the eye and mouth regions modulate the enhancement derived from visual speech information. It was expected that (1) speech segmentation performance will be better when the auditory recordings are paired with the synchronous video the speaker's talking face, and (2) greater gaze directed to the speaker's mouth than the eyes will facilitate speech segmentation performance.

2. Method

2.1. Participants

Participants were 16 7.5-month-old monolingual Australian-English learners. These infants were born full-term, with no vision and hearing deficits, were not at risk for any language or cognitive delay and had no history of ear infections. Data from two infants were excluded because they had less than 40% weighted gaze samples. Thus, the final data set consists of 14 infants with 6 infants in an auditory-only (AO) condition (mean age = 7.48 months, *SD* = 0.10) and 8 infants in the auditory-visual (AV) condition (mean age = 7.37 months, *SD* = 0.04).

2.2. Stimuli

Stimulus materials consisted of four different passages used in a previous study [5]. These passages centered around the target words 'cup', 'dog', 'bike' and 'feet'. A female native Australian English speaker was recorded saying these passages in infant-directed speech. The recordings were auditory-visual and included the speaker's head, face and neck. Additionally, for use in test trials, recordings were made of each of the four target words repeated eight times in succession with varying intonation.

For the AO condition, auditory recordings were extracted from the video recordings. The auditory recordings were then

10.21437/AVSP.2019-9

paired with a still image of the speaker's resting (non-articulating) face. For the AV condition, the video and audio recordings of the speaker's talking face were used.

2.3. Apparatus

Video recordings were presented via a 17-inch DELL LCD monitor while auditory recordings were played via two loudspeakers (Edirol MA-15 Digital Stereo Micro Monitors) placed at the left and right sides of the monitor. A Tobii X120 eye tracker was fitted at the bottom of the screen to record infants' gaze patterns throughout the entire session.

2.4. Procedure

A visual preference procedure was employed using a single central screen [11]. Infants sat on their parent's lap approximately 60cm from the centre of the screen. The experimenter remained in an adjacent control room throughout the entire experiment.

During a familiarisation phase, half of the infants were presented with passages containing 'cup' and 'dog' while the other half were presented with passages containing 'bike' and 'feet'. The two passages were presented twice on alternate trials. During the test phase, all four words (*cup*, *dog*, *bike*, *feet*) were presented to the infants, two of them being targets (words that appeared in the familiarisation passages) and two being non-targets, depending on the familiarisation condition. The test phase consisted of two blocks of test trials in which all four lists of words were presented once per block. A refamiliarisation phase followed the test phase. In this phase, the two passages were presented once again. After the refamiliarisation phase a single block of test trials was presented. To capture infants' attention, throughout the entire session an attention-getter animation was played in between each trial and the experimenter initiated the next trial only after the infant fixated on the screen for at least 2s. Once a trial was initiated, the trial was played through to completion regardless of infant gaze behaviour.

Infants in the AO condition were presented with familiarisation and test trials that consisted of a still image of the speaker's resting face and the auditory recordings. Infants in the AV condition were presented with familiarisation and test trials in auditory-visual modality (both video and auditory recordings).

2.5. Eye-Tracking

Dynamic areas of interest (AOIs) in the eye, mouth and face regions were defined and fixation durations to these regions were collected (Figure 1). Custom MATLAB (R2018b, Mathworks) scripts were used to extract raw looking times to these regions which were then computed as the proportion of total looking times (PTL).

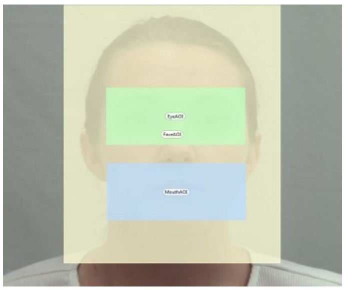

Figure 1: *Example of AOIs in a video frame.*

The following measures of PTL were computed:

Attention on each trial:

$$\frac{fixation\ duration}{trial\ duration} \times 100\% \qquad (1)$$

It was expected that infants in the AV condition will pay greater attention during familiarization and test phases than infants in the AO condition.

Mouth preference:

$$\frac{mouth}{mouth+eyes} \times 100\% \qquad (2)$$

It was expected that infants in the AV condition will show greater mouth preference than infants in the AO condition.

3. Results

3.1. Preliminary Analyses

Independent-samples t-tests were conducted to examine if there was any difference in attention between conditions during the familiarisation and test phases. Infants in the AV condition had greater attention than infants in the AO condition during the familiarisation phases for Blocks 1 and 2 (Block 1: $t(12) = 2.46$, $p = 0.03$; Block 2: $t(12) = 2.70$, $p = 0.02$).

Table 1: *Means and standard deviations (in parentheses) of attention during familiarization and test phases.*

Phase	**Block 1**		**Block 2**	
	.AO	**AV**	**AO**	**AV**
Familiarisation	0.65 (0.16)	0.85 (0.15)	0.34 (0.12)	0.60 (0.21)
Test	0.58 (0.06)	0.72 (0.18)	0.55 (0.13)	0.53 (0.15)

3.2. Speech Segmentation Performance

To examine speech segmentation performance in each condition, a 2 (Condition: AO vs. AV) x 2 (Trial Type: Target vs. Non-Target) x 2 (Block: 1 vs. 2) mixed-measures ANOVA was conducted.

The mixed-measures ANOVA revealed a significant main effect of Block, $F(1,12) = 7.94$, $p = 0.016$, with greater attention in Block 1 ($M = 0.65$, $SE = 0.04$) than in Block 2 ($M = 0.54$, $SE = 0.04$). The main effect of Trial Type and the Condition x Trial Type x Block interaction approached significance: $F(1,12) = 4.39$, $p = 0.058$, and $F(1,12) = 4.45$, $p = 0.056$, respectively.

The non-significant Condition x Trial Type interaction ($F(1,12) = 3.11$, $p = 0.10$) suggests that infants' looking times to target and non-target trials did not differ between conditions. Therefore, further post-hoc comparisons were conducted by means of paired-samples t-tests to examine speech segmentation performance in each condition separately. Means and standard deviations of attention to target and non-target trials for each block are reported in Table 2.

3.2.1. Auditory-only condition

Paired-samples t-tests revealed that infants in the AO condition spent a similar proportion of time looking at target and non-target trials in Block 1, $t(5) = 2.05$, $p = 0.096$, and Block 2, $t(5) = -0.97$, $p = 0.38$.

3.2.2. *Auditory-visual condition*

In Block 1, a paired-samples t-test showed that infants in the AV condition spent a similar proportion of time looking at target and non-target trials, $t(7) = 0.99$, $p = 0.35$. In Block 2, a infants spent a significantly larger proportion of time looking at target compared to non-target trials, $t(7) = 2.55$, $p = 0.038$).

Table 2: *Means and standard deviations (in parentheses) of attention during target and non-target trials.*

Trial Type	**Block 1**		**Block 2**	
	.AO	**AV**	**AO**	**AV**
Target	0.62	0.74	0.52	0.60
	(0.07)	(0.06)	(0.07)	(0.06)
Non-Target	0.54	0.70	0.57	0.46
	(0.05)	(0.05)	(0.07)	(0.06)

3.3. Infants' Gaze Patterns in the AV Condition

To investigate infants' gaze patterns, a 2 (Condition: AO vs. AV) x 2 (Trial Type: Target vs. Non-Target) x 2 (Block: 1 vs. 2) mixed-measures ANOVA was conducted using the derived PTL that quantified mouth preference as the dependent variable (DV). Only the main effect of Condition was significant, $F(1,11) = 8.09$, $p = 0.016$, with infants in the AV condition showing a stronger mouth preference than infants in the AO condition across trials and blocks (Table 3).

Table 3: *Means and standard deviations (in parentheses) of mouth preference during target and non-target trials.*

Trial Type	**Block 1**		**Block 2**	
	.AO	**AV**	**AO**	**AV**
Target	0.19	0.53	0.20	0.66
	(0.16)	(0.28)	(0.34)	(0.23)
Non-Target	0.26	0.60	0.19	0.68
	(0.41)	(0.26)	(0.28)	(0.19)

One possible reason behind the stronger mouth preference of infants in the AV condition is that they pay more attention to the dynamic and moving face in general. To examine this possibility, a face preference was quantified by the following equation:

$$\frac{face}{fixation\ duration} \times 100\% \qquad (3)$$

A 2 (Condition: AO vs. AV) x 2 (Trial Type: Target vs. Non-Target) x 2 (Block: 1 vs. 2) mixed-measures ANOVA was conducted with face preference as the DV (see Table 4 for means and standard deviations). Only the main effect of Trial Type was significant, $F(1,12) = 8.57$, $p = 0.013$. Infants in both conditions spent a greater proportion of time looking at the face in target than in non-target trials. The main effect of Condition approached significance, $F(1,12) = 4.34$, $p = 0.059$.

Table 4: *Means and standard deviations (in parentheses) of face preference during target and non-target trials.*

Trial Type	**Block 1**		**Block 2**	
	.AO	**AV**	**AO**	**AV**
Target	0.87	0.96	0.74	0.91
	(0.12)	(0.40)	(0.20)	(0.09)
Non-Target	0.89	0.94	0.71	0.88
	(0.08)	(0.08)	(0.28)	(0.12)

3.4. Did looking behavior in the familiarisation phase predict segmentation performance?

Linear regression analyses were conducted with the PTLs derived from equations (1) to (3) as predictor variables to examine if individual differences in infants' looking times predicted differences in proportion looking times to target trials and non-target trials. None of the predictor variables was significant for either trial block or condition (all *p*s > 0.15).

4. Discussion

This study investigated whether visual speech information from a speaker's face aids infants' speech segmentation and whether gaze patterns modulated segmentation performance.

Results suggest that infants in the AO condition did not segment target words from the passages in blocks 1 and 2. By comparison, infants in the AV condition spent a greater proportion of looking times to target than non-target trials in block 2, suggesting that they successfully segmented target words from the passages in block 2 but not in block 1. Further investigations of infants' looking behavior revealed that infants in the AV condition had a stronger mouth preference than infants in the AO condition across trial types. It is possible that infants in the AV condition are attracted to the speaker's moving face. This warrants further research as the difference between both groups of infants in proportion of time spent looking at the face approached significance. Future work needs to be done to examine if both groups of infants pay attention to different regions of the face. For instance, infants in the AO condition may focus more on the eye region whereas infants in the AV condition may focus more on the mouth region as there is more linguistic information from a talking mouth.

Finally, non-significant findings from regression analyses suggest that infants do not actively rely on a single region of the speaker's face to segment speech. This too requires further investigation.

5. Conclusions and Implications

This study provides preliminary findings on the influence of visual speech information on infants' speech segmentation. A larger sample size is necessary to draw stronger conclusions. Understanding how visual speech information may facilitate infants' segmentation of continuous speech is particularly important given that speech is a multimodal phenomenon. Further, there are potential implications for infants who do not have clear access to auditory speech, e.g., infants with hearing loss. As previous research has established that children and adults with hearing loss perform better on speech perception tasks when visual speech information is provided, e.g., [12], findings from the current study have ramifications for infants with hearing loss especially since this group of infants may already be relying more on visual speech cues for linguistic information.

6. Acknowledgements

This research was supported by a doctoral scholarship to the first author funded by the MARCS Institute at Western Sydney University and the HEARing Cooperative Research Centre (CRC), and by HEARing CRC funding to the second author. We gratefully appreciate the assistance of Caitlin Hooper for the recording of stimuli, the MARCS technical team, and the parents and their children who participated in the study.

7. References

[1] A. D. Mitchel and D. J. Weiss, “What’s in a face? Visual contributions to speech segmentation,” *Language and Cognitive Processes*, vol. 25, no. 4, pp. 456–482, 2010.

[2] A. D. Mitchel and D. J. Weiss, “Visual speech segmentation: Using facial cues to locate word boundaries in continuous speech,” *Language, Cognition and Neuroscience,* vol. 29, no. 7, pp. 771–780, 2014.

[3] C. Kubicek, J. Gervain, A. H. de Boisferon, O. Pascalis, H. Loevenbruck and G. Schwarzer, “The influence of infant-directed speech on 12-month-olds’ intersensory perception of fluent speech,” *Infant Behaviour and Development*, vol. 37, no. 4, pp. 644–651, 2014.

[4] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, “The natural statistics of audiovisual speech,” *PLoS computational biology*, vol. 5, no. 7, pp. e1000436, 2009.

[5] P. W. Jusczyk and R. N. Aslin, “Infants’ detection of the sound patterns of words in fluent speech,” *Cognitive Psychology,* vol. 29, no.1, pp. 1–23, 1995.

[6] E. K. Johnson and P. W. Jusczyk, “Word segmentation by 8-month-olds: When speech cues count more than statistics,” *Journal of Memory and Language,* vol. 44, no. 4, pp. 548–567, 2001.

[7] S. L. Mattys, P. W. Jusczyk, P. A. Luce and J. L. Morgan, “Phonotactic and prosodic effects on word segmentation in infants,” *Cognitive Psychology*, vol. 38, no. 4, pp. 465–494, 1999.

[8] P. W. Jusczyk, E. A. Hohne and A. Bauman, “Infants’ sensitivity to allophonic cues for word segmentation,” *Perception and Psychophysics,* vol. 61, no. 8, pp. 1465–1476, 1999.

[9] G. Hollich, R. S. Newman, and P. W. Jusczyk, “Infants’ use of synchronized visual information to separate streams of speech,” *Child Development*, vol. 76, no. 3, pp. 598–613, 2005.

[10] E. J. Tenenbaum, R. J. Shah, D. M. Sobel, B. F. Malle and J. L. Morgan, “Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study,” *Infancy*, vol. 18, no.4, pp. 534–553, 2013.

[11] E. D. Thiessen, “Effects of visual information on adults’ and infants’ auditory statistical learning,” *Cognitive Science*, vol. 34, no. 6, pp. 1093–1106, 2010.

[12] R. Taitelbaum-Swead and L. Fostick, “Audio-visual speech perception in noise: Implanted children and young adults versus normal hearing peers,” *International Journal of Pediatric Otorhinolaryngology,* vol. 92, pp. 146–150, 2017.