Pointing Gestures do not Influence the Perception of Lexical Stress

Alexandra Jesse[1,2], *Holger Mitterer*[2]

[1] Department of Psychology, University of Massachusetts, Amherst, U.S.A.
[2] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
ajesse@psych.umass.edu, Holger.Mitterer@mpi.nl

Abstract

We investigated whether seeing a pointing gesture influences the perceived lexical stress. A pitch contour continuum between the Dutch words "CAnon" ('canon') and "kaNON" ('cannon') was presented along with a pointing gesture during the first or the second syllable. Pointing gestures following natural recordings but not Gaussian functions influenced stress perception (Experiment 1 and 2), especially when auditory context preceded (Experiment 2). This was not replicated in Experiment 3. Natural pointing gestures failed to affect the categorization of a pitch peak timing continuum (Experiment 4). There is thus no convincing evidence that seeing a pointing gesture influences lexical stress perception.
Index Terms: speech perception, lexical stress, gestures

1. Introduction

The perception of spoken language is often an audiovisual phenomenon: Humans identify what a speaker says based on listening to speech and based on observing the speaker's facial and manual gestures. Seeing a speaker's talking face, for example, aids the recognition of spoken sound segments and hence the recognition of words [e.g., 1]. Seeing a speaker also provides prosodic information [e.g., 2-8]. On the word level, seeing a talking face can provide sufficient information to recognize the relative emphasis of syllables within a word, that is, to recognize a word's lexical stress pattern [2,3]. For example, English and Swedish minimal stress word pairs (such as '(to) preSENT' and '(a) PREsent'; capital letters indicate primary lexical stress) can be identified above chance in visual-only presentations of a speaker [2,3]. For English, mainly articulatory correlates, especially chin opening, contribute to the perception of lexical stress [2]. Here, we examine whether seeing a manual pointing gesture can influence the perception of lexical stress.

Lexical stress information is important for recognizing spoken words. Considering lexical stress reduces, for example, the number of embedded words from an average of 0.94 words to 0.59 words in English and from 1.52 words to 0.74 words in Dutch [9,10]. The use of lexical stress is therefore more important for Dutch than for English word recognition. The relative acoustic implementation of stress cues is also language-specific. Lexical stress in English is mostly cued by vowel reduction, a segmental cue. But in Dutch, lexical stress is often only cued suprasegmentally. That is, stressed syllables are louder, longer, and have higher pitch than unstressed syllables. In listening to speech, Dutch listeners are more sensitive to suprasegmental cues of lexical stress than English listeners [11]. Dutch listeners can reliably identify whether or not a syllable excised from a suprasegmental minimal stress pairs, such as "CAnon" ('canon') and "kaNON" ('cannon'), is stressed [12]. English listeners have difficulties perceiving suprasegmental cues for lexical stress [13]. Dutch listeners use suprasegmental lexical stress information already efficiently during word recognition before segmental information disambiguates the target word from its competitors [14]. The perception of lexical stress is thus important in understanding spoken words, especially in Dutch.

Here, we tested whether seeing a pointing gesture influences the perception of lexical stress location. A pointing gesture to an intended referent tends to be synchronized to the production of the referent's label. More precisely, the apex of the pointing gesture is, for example, synchronized with the onset of the demonstrative in the utterance "this/that lamp" [15]. Speakers align the apex of a pointing gesture with the maximal point of jaw opening in the stressed syllable of bisyllabic nonsense words, such as "PApa" vs. "paPA" [16]. This suggests that the apex of the pointing gesture is aligned with the emphasis in the speech signal. Experiment 1 through 4 tested whether listeners are sensitive to this cross-modal temporal alignment and hence whether the timing of a pointing gesture influences auditory stress perception.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Ten native Dutch speakers from the Max Planck Institute's subject pool were paid for their participation.

2.1.2. Materials

One token each of the Dutch words "canon" (/'ka-nɔn/, 'canon') and "kanon" (/ka-'nɔn/, 'cannon') were synthesized using MBROLA with a female Dutch voice (nl3). Segment durations were based on those of utterances of these words found in a Dutch speech corpus [17]. For the synthesis, segment durations were averaged for each syllable across tokens. A pitch-contour continuum was generated by interpolating between the pitch contours for each of the five sound segments separately. Pitch contours were mixed in different proportions and seven mixed versions, ranging from 20% to 80% mixtures, were used in the experiment.

A black-and-white drawing of a hand with a pointing gesture (see Figure 1) was animated in Matlab. All animations were created as avi files (size 720 x 576, 25 fps). Animations showed the hand moving down straight towards a question mark positioned in the middle bottom of a white background. The hand then moved up again. The speed of the movement followed a Gaussian function and lasted 400 ms (i.e., 10 frames). The apex of the movement was aligned to the maximal acceleration of the amplitude contour of the first or the second syllable (i.e., the p-centre, [18]). Two catch trial animations were created where the question mark changed into the word "stop" when the manual movement reached its apex. A fade in and fade out of five frames was added to all four animations in Adobe Premiere. The two target video tracks were combined with all audio files of the continuum. The two catch trial video tracks were only combined with auditory endpoint versions. Final files were converted to mpg.

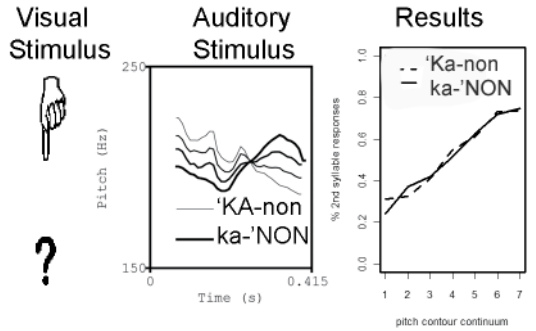

Figure 1: *Stimuli and results from Experiment 1. Pitch contours (middle panel) are shown with thicker lines, the more similar they become to second-syllable stress. Only contours for steps 1, 3, 5, and 7 are shown throughout the paper.*

2.1.3. Procedure

Listeners were instructed to indicate by button press whether they heard "canon" or "kanon". Buttons were labeled "lied" ('song') for "canon" and "wapon" ('weapon') for "kanon". Labels were shown on the bottom of the screen; "lied" was always on the left side. These response labels appeared 500 ms before a video was shown. Figure 1 shows a screenshot of a trial. Responses were collected 300 ms after video onset for up to 3700 ms. Participants failing to respond in time were given feedback in form of a displayed stopwatch. Participants were instructed not to respond on catch trials where the question mark changes into the word "stop". The experiment consisted of ten repetitions of all stimuli. This resulted in a total of 140 trials, including sixteen catch trials.

2.2. Results and discussion

The mean response rate on catch trials was 50%. The stop signal was given quite late, which explains this high error rate. Catch trials were excluded from any analyses. Figure 1 shows the categorization data. Results were analyzed in the R statistical program with mixed effect models [19]. A binomial logit linking function was included to deal with the categorical nature of the response variable. Best-fitting models were established through systematic step-wise model comparisons using likelihood ratio tests. Continuum step was assessed as a numerical factor (steps 1 to 7, centered at zero for the analysis). Pointing was assessed as a categorical factor (1st syllable, 2nd syllable (intercept condition)). If the estimated adjustment for a factor differs significantly from zero, then the variable has an effect on performance. For numerical factors, the adjustment would be applied to the slope; for categorical factors, to the intercept. For categorical fixed factors, one condition is mapped onto the intercept of the model. All best-fitting models included subject as a random factor, allowing for subject-specific adjustments to the regression weights.

Listeners were sensitive to the auditory manipulation (β=.37, p<.0001) and gave more kaNON responses the more the pitch contour continuum was indicating lexical stress on the second syllable. The timing of the pointing hand had no effect on categorization ($X^2(1)$=0.03, p=.86).

3. Experiment 2

Experiment 1 failed to show an effect of pointing on the perception of lexical stress location. Animations of the pointing hand were artificially created by assigning the speed of the hand to follow a Gaussian function. Their apex was aligned to the point of maximal acceleration in the speech amplitude contour. Natural pointing gestures may, however, have a different timing [16]. Animations in Experiment 2 were therefore based on natural recordings of a speaker pointing either to emphasize stress on the first or second syllable. Additionally, the pointing gesture was shown against a textured background to provide observers with a more contrastive frame of reference to follow the hand movement.

We also added a context condition with "een" ('a') as a precursor. Pointing movements emphasizing the first syllable started before word onset. A preceding context provides speaking rate information and may therefore help estimate the arrival time of the apex relative to the spoken word. Preceding context also enables the listener to interpret the pitch of the first syllable relative to the context.

3.1. Methods

3.1.1. Participants

Twenty-eight new subjects from the same population as in Experiment 1 were paid for their participation. Twelve subjects participated in the no-context condition; sixteen in the context condition.

3.1.2. Materials and procedure

A female native Dutch speaker was recorded pointing sideways with her dominant right hand while saying "canon" or "kanon". No instructions in regard to the timing of the gesture were given. The speaker was wearing a CyberGlove. The x and y location of the tip of the forefinger was tracked (100 Hz) with two Ascension Flock of Bird location trackers.

One token of each word was selected. The pointing movement rotated by 90 degrees and smoothed over the coordinates of three neighboring time samples. The timing of the movement was rescaled for the duration of the continuum stimuli. Animations showed a pointing hand with a black contour and a light skin-colored filling. The background was grey and textured. The hand pointed to a black dot shown in the middle bottom of the screen. Figure 2 shows a screenshot from a trial.

The auditory continuum was based on two endpoint tokens spoken by the same speaker who provided the movement data. These tokens were re-synthesized with segment durations that corresponded to the mean durations of ten CAnon and kaNON tokens. The resulting syllable durations were hence ambiguous in regard to stress. The first syllable was 249 ms long; the second syllable lasted for 493 ms. These tokens were longer than in Experiment 1 (742 ms vs. 420 ms). The precursor "een" lasted 420 ms. A stress continuum was created as done in Experiment 1. Five intermediate steps and the two endpoints were presented. The testing procedure was the same as for Experiment 1, but an audio-only condition was included. The auditory-only condition was always presented first; the two visual conditions were then presented intermixed. Each stimulus was presented ten times.

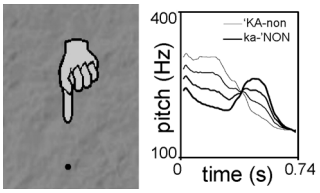

Figure 2: *Visual and auditory stimuli (left and right, respectively) used in Experiments 2 and 3.*

3.2. Results and discussion

The mean response rate on catch trials was 52%. Figure 3 shows the categorization data for both context conditions. Analyses were conducted separately for each modality condition. On auditory-only trials, the continuum manipulation affected responses. (β=.52, p<.0001). Context had no effect (X^2(1)=2.5, p=.11). On audiovisual trials, listeners were sensitive to the auditory continuum manipulation (β=.49, p<.0001), but less so without the precursor (β=-.06, p<.0001). As predicted, listeners gave fewer kaNON responses when pointing emphasized the first syllable (β=-.71, p=.01). This effect was smaller with more kaNON-like continuum steps (β=.08, p=.015). The context manipulation had a marginally significant main effect (β=.45, p=.11), but modulated the pointing effect (β=.36, p=.03). This suggests that the pointing effect was stronger when a preceding context was given.

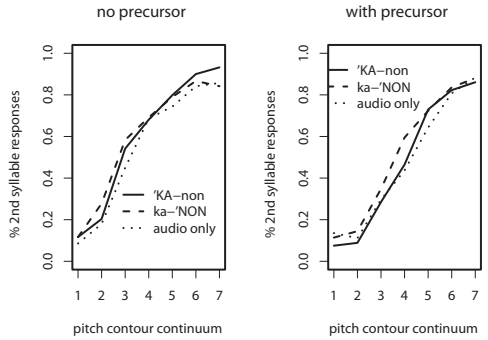

Figure 3: *Results from Experiment 2.*

4. Experiment 3

Experiment 2 showed that a natural pointing gesture seems to influence the perception of the location of lexical stress. The effect was stronger when a preceding context was provided. In Experiment 3, we assessed whether seeing a pointing gesture influences only categorization or also discrimination. The latter result would indicate an influence on early perceptual stages (cf. [19], for a similar rationale).

4.1. Methods

4.1.1. Participants

Forty-four new subjects from the same population as in the previous experiments were paid for their participation.

4.1.2. Materials and procedure

The same materials as in the context condition in Experiment 2 were used. For the same-different discrimination tasks, videos were paired so that the visual gestures would enhance or diminish the auditory difference. Three different versions of the discrimination task were used that differed in step size (2 vs. 3 steps) and the direction of the different pairs (from 1st to 2nd syllable stress or vice versa).

Participants first completed one version the discrimination and then the categorization task. This order prevented influences of labeling on discrimination performance (cf. [20]). The procedure of the categorization task was the same as in Experiment 2.

4.2. Results and discussion

The mean response rate on catch trials was 18% in the categorization task. Figure 4 shows the categorization data. Results were analyzed as done for the previous experiments. Listeners were sensitive to the continuum manipulation (β=.92, p<.0001), but pointing did not affect categorizations (X^2(1)=0.22, p=.64). Given the absence on categorization, it is not surprising that performance in the discrimination tasks was not influenced by the visual display. The percentage correct on different trials was 87% with cooperating visual cues and 86% with conflicting cues.

Experiment 3 with a larger sample thus failed to replicate the effects of pointing on stress categorization found in Experiment 2. Such a negative association between effect and sample size is often taken as an indication for an underlying null-effect (cf. [21]).

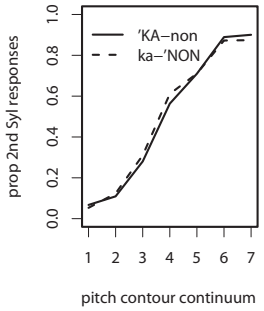

Figure 4: *Results from the categorization task in Experiment 3.*

5. Experiment 4

In all previous experiments, lexical stress location was manipulated as change in the height of the pitch contour. Tokens with stress on the first syllable had a pitch peak in the first syllable, while having a lowered pitch in the second syllable. The opposite was the case for tokens with stress on the second syllable. We assessed therefore whether pointing affects the perceived pitch height. Pointing may instead affect the perceived timing of the pitch peak. Experiment 4 tested whether pointing can change the perceived pitch peak location and hence affect lexical stress perception. A continuum in pitch location was combined with the pointing animations from Experiment 3. Only categorization was tested.

5.1. Methods

5.1.1. Participants

Twenty new subjects from the same population as in the previous experiments were paid for their participation.

5.1.2. Materials and procedure

The same visual stimuli as in the context condition in Experiment 2 were used. The auditory stimuli were generated by adding a peak to a basic declinating contour (see Figure 5). The procedure was the same as for the categorization task in the previous experiments.

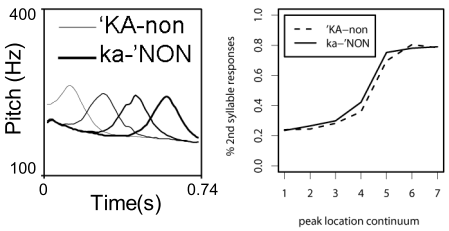

Figure 5: *Stimuli and results from Experiment 4.*

5.2. Results and discussion

The mean response rate on catch trials was 46%. Figure 5 shows the categorization data. Results were analyzed as done for the previous experiments. Listeners were sensitive to the continuum manipulation (β=.53, p<.0001), but pointing did not affect categorizations (X^2(1)=2.04, p=.15).

6. Discussion

Pointing gestures are synchronized with the emphasis in the speech stream in speech production tasks [16]. The present study investigated whether seeing an accompanying pointing gesture can influence the perceived location of (suprasegmental) lexical stress. A pitch contour continuum between the two bisyllabic words of a Dutch minimal stress pair was created and combined with a pointing gesture with an apex either during the first or the second syllable, respectively. For Experiments 1 through 3, this continuum was created by interpolating the height of the pitch peaks on both syllables of the words. In Experiment 1, pointing gestures animated with speed following a Gaussian function did not influence stress perception. Pointing gestures based on a natural recording of a speaker seemed to influence stress perception in Experiment 2. Their influence was stronger when an auditory preceding context was provided. In Experiment 3, however, we failed to replicate these results. In Experiment 4, the auditory stress continuum was created by systematically shifting the timing of the pitch peak from the first to the second syllable. No effect of the timing of the pointing gesture on stress perception was found. Overall, the results show no convincing evidence that seeing a pointing gesture influences lexical stress perception.

One possibility why the present experiment failed to find an effect of pointing on stress perception could be that a disembodied hand was shown. Listeners may not be influenced in their perception of an auditory event by observing a visual event if both are not perceived as being part of the same multisensory event. It could also be the case that the pointing gesture influences the perceived duration of a syllable rather than the perceived pitch. Syllable durations were here set to a neutral value. If seeing a pointing hand makes a syllable sound longer, then this stress information should have, however, nevertheless shifted the overall perception of stress. A third possibility is that manual gestures, or maybe in particular pointing gestures, are not able to influence lexical stress perception. Pointing gestures were temporally aligned with lexical stressed syllables [16] in a task with nonwords, but it is not clear that the same strategy is used in spontaneous dialogue. Furthermore, manual beat gestures, just like eyebrow and head movement, for example, affect the perception of sentence-level emphasis [7,23]. It could be the case, that manual gestures only provide information that a particular word is emphasized, but not more fine-grained information about relative emphasis placed on the syllables within a word. Further research should directly compare the role of manual gestures on word-level and sentence-level emphasis and the temporal alignment in more natural situations than in the previous studies.

7. Acknowledgements

This research was supported in part by an Innovational Research Incentive Scheme Veni grant from the Netherlands Organization for Scientific Research (NWO) awarded to first author. The authors thank Lies Cuijpers for her help with the experiments.

8. References

[1] Sumby, W.H., and Pollack, I., "Visual contribution to speech intelligibility in noise", JASA, 26:212-215, 1954.

[2] Scarborough, R., Keating, P., Mattys, S. L., Cho, T., and Alwan, A. "Optical phonetics and visual perception of lexical and phrasal stress in English", Lang. Speech, 52:135-175, 2009.

[3] Risberg, A., and Lubker, J. "Prosody and speech-reading", Speech Transm. Lab. Quart. Progress Status Rep., 4:1-16, 1978.

[4] Bernstein, L.E., Eberhardt, S.P., and Demorest, M.E. "Single-channel vibrotactile supplements to visual perception of intonation and stress", JASA, 85:397-405, 1989.

[5] Dohen, M., Loevenbruck, H., Cathiard, M.-A., and Schwartz, J.-L. "Visual perception of contrastive focus in reiterant French speech", Speech Comm., 44:155-172, 2004.

[6] Swerts, M., and Krahmer, E. "Cognitive Processing of Audiovisual Cues to Prominence", Proc. AVSP, 29-30, 2005.

[7] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. "About the relationship between eyebrow movements and F0 variations", Proc. Interna. Conf. Speech Lang. Processing, 2175-2178, 1996.

[8] Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., and Vatikiotis-Bateson, E., "Visual prosody and speech intelligibility: Head movement improves auditory speech perception", Psych. Sci., 15:133-137, 2004.

[9] Cutler, A., Norris, D., and Sebastián-Gallés, N., "Phonemic repertoire and similarity within the vocabulary", Proc. 8th Interna. Conf. Speech Lang. Processing, 65-68, 2004.

[10] Cutler, A., Pasveer, D., "Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. Proc. of 3rd Conf. on Speech Pros., 237-240, 2006.

[11] Cooper, N., Cutler, A., and Wales, R. "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners", Lang. Speech, 45:207-228, 2002.

[12] Cutler, A., and Van Donselaar, W., "Voornaam is not a homophone: Lexical prosody and lexical access in Dutch", Lang. Speech, 44:171-195, 2001.

[13] Fear, B. D., Cutler, A., and Butterfield, S. "The strong/weak syllable distinction in English", JASA, 97:1893-1904, 1995.

[14] Reinisch, E., Jesse, A., & McQueen, J. M. "Early use of phonetic information in spoken word recognition: Lexical stress drives eye-movements immediately", Quart. J. Exp. Psych., 63: 772-783, 2010.

[15] Levelt, W. J. M., Richardson, G., and La Heij, W., "Pointing and voicing in deictic expressions", J. Mem. Lang., 24:133–164, 1985.

[16] Rochet-Capellan, A., Laboissiere, R., Galvan, A., and Schwartz, J.-L., "The speech focus position effect on jaw-finger coordination in a pointing task", J. Speech, Lang. and Hearing Res., 51:1507-1521, 2008.

[17] Oostdijk, N., "The Spoken Dutch Corpus Project", ELRA Newsletter, 5:4-8, 2000.

[18] Scott, S.K., "The point of P-centres", Psych. Res., 61:4-11, 1998.

[19] R Development Core Team, "R: A language and environment for statistical computing", Vienna: R Foundation for Statistical Computing, 2009.

[20] Mitterer, H., Csépe, V., and Blomert, L., "The role of perceptual integration in the recognition of assimilated word forms", Quart. J. Exp. Psych., 59(8): 1395-1424, 2006.

[21] Gerrits, E. and Schouten, M.E.H., "Categorical perception depends on the discrimination task", Perc. Psychoph., 66:363-376, 2004.

[22] Egger, M., Smith, G. D., Schneider, M., and Minder, C., "Bias in meta-analysis detected by a simple, graphical test", BMJ, 315(7109):629-634, 1997.

[23] Krahmer, E., and Swerts, M. "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception", J. Mem. Lang., 57:396-414, 2006.