Are gesture and prosodic prominences always coordinated? Evidence from perception and production

Núria Esteve-Gibert[1], *Ferran Pons*[2], *Laura Bosch*[2], *Pilar Prieto*[3,1]

[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[2] Department of Basic Psychology, University of Barcelona, Barcelona, Spain
[3] ICREA – Institució Catalana de Recerca i Estudis Avançats

nuria.esteve@upf.edu, ferran.pons@ub.edu, laurabosch@ub.edu, pilar.prieto@upf.edu

Abstract

This study explores the temporal coordination between gesture and speech by addressing two main questions: (1) Are speakers sensitive to the misalignment between gesture prominence and prosodic prominence? (2) Is this sensitivity modulated by the semantic information conveyed by gesture and speech modalities in production? Experiment 1 tested question (1) and Experiment 2 tested question (2). Results from Experiment 1 revealed that the combinations in which prominences were misaligned were less acceptable than combinations with aligned prominences, and that the metrical pattern of the target word had an effect on the speakers' sensitivity: unsynchronized trochees (with the gesture prominence at the post-tonic syllable) were frequently accepted, while unsynchronized iambs (with the gesture prominence at the pre-tonic syllable) were rejected. Results from Experiment 2 revealed that when the pointing gesture adds information to speech, i.e. it is supplementary to speech, the prominences are frequently misaligned (with gesture occurring after the speech), as if two different speech acts were produced. These findings suggest that the semantic content of gesture-speech combinations might influence the speakers' sensitivity of the misalignment between prosodic and gesture prominences.

Index Terms: gesture-speech synchronization, audiovisual prosody, multimodal prominence

1. Introduction

There is ample evidence in the literature that humans coordinate gesture movements with speech, suggesting that both modalities are in fact part of an integrated system [1-3]. This coordination is evidenced from both semantic and temporal points of view. What speakers express with their hands is semantically related with what they express with their speech (what could be called 'semantic coordination'). Also, gesture and speech timings are coordinated, since the most prominent part of the gesture co-occurs with the most prominent part of speech ('temporal coordination') [3].

1.1. Temporal and semantic coordination

Studies investigating the temporal coordination of gesture and speech have found convincing evidence that gesture and speech co-occur in time in the sense that the point of maximal expression of a gesture (hereafter 'gesture prominence') coincides with the moment of maximal prosodic prominence in speech [4]. In order to define gestural prominence, most studies use either the stroke of the gesture (the interval involving the greatest physical effort in the gesture) or the apex of the gesture (the point in time in which the gesture reaches its maximal extension). As for the prominent feature of speech, a growing body of research has found that the speech landmark with which the gesture prominence aligns is the lexical stress [5, 6] and even the pitch peak within the stressed syllable when it is uttered in a contrastive focus situation [7, 8].

However, it has also been proposed that the temporal synchronization between gesture and speech may depend on their semantic coordination: when the meanings expressed by the co-speech gesture and by the accompanying lexical affiliate are complementary, the onset of the gesture stroke is closely aligned with its lexical affiliate; but when the two modalities express supplementary semantic features, stroke onset and lexical affiliate are not so closely aligned [9]. But more evidence is needed to corroborate this hypothesis.

1.2. Perception of temporal asynchrony

But how important is this tight temporal coordination? As interlocutors, do we expect the gesture apex to co-occur with the lexical stress? Do we perceive misalignments in their temporal coordination?

Most of the studies examining the perception of audio-visual asynchrony have focused on the human ability to perceive unsynchronized audiovisual events in articulatory gestures of a person producing syllables or a list of words. They found that adults can detect an audiovisual asynchrony of around 200 ms when the visual attributes of an audiovisual event precede the auditory attributes, and around 100 ms when the auditory attributes precede the visual attributes [10]. However, the articulatory synchronization patterns tested in these experiments did not answer the question of whether the temporal coordination of prominences found for co-speech gestures is relevant in perception.

Few studies have examined the effects of the gesture-prosodic misalignment in the perception of the lexical stress [11-13]. Results seem contradictory, some finding a clear influence [12, 13] and some not [11]. From these, only in [11] the authors analyzed pointing gestures and they did not find a clear influence and the authors suggest that their results might be influenced by some methodological problems with the procedure. Thus, the influence of the timing of the gesture prominence with respect to the speech prominence needs to be further analyzed.

1.3. Aim of the study

The aim of the present study was two-fold: first, to investigate speakers' ability to perceive a temporal asynchrony between gesture and speech prominences (Experiment 1); second, to investigate whether this perceptual ability is related to how

speakers align gesture and speech when the semantic information expressed by gesture supplements what is expressed in speech (Experiment 2).

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty-two adult Catalan-speakers took part in an online acceptability judgment task. They were unaware of the purpose of the study and participated voluntarily.

2.1.2. Materials

An online survey was prepared using the SurveyGizmo application. Participants watched a series of video clips each showing a woman producing a disyllabic word accompanied by a deictic pointing gesture. The woman appeared sideways in the right part of the screen and pointed to the left part of the screen. In order to prevent participants from looking at her lip movements, the woman covered her mouth with the hand not used for pointing (see Figure 1, left panel).

Sixteen disyllabic words were used, half of them iambs (with stress on the second syllable) and the other half trochees (with stress on the first syllable). They were all common words, such as "miRALL" ('mirror'), "ioGURT" ('yogurt'), "AIgua" ('water'), or "COtxe" ('car').[1] Words were pronounced in an exaggerated manner so that syllable duration values were longer and pitch range values were higher than in spontaneous speech (see Figure 1, right panel).

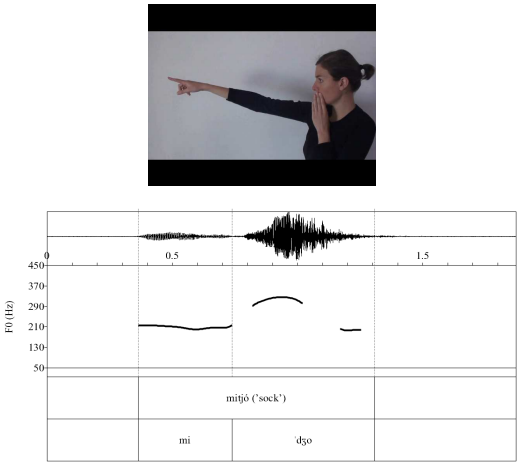

Figure 1: *Top panel, visual stimulus presented in the survey: a video clip showing a woman pointing while saying a word (frame showing the apex of the gesture). Bottom panel, waveform and pitch contour of the word produced while pointing, with the F0 peak coinciding with the gesture apex.*

Of the total number of video clips that participants observed (N = 32), half were synchronized (gesture apex coinciding with lexical stress) and half were unsynchronized (gesture apex not coinciding with lexical stress). Using Adobe Premiere Pro, all clips were constructed with the same pointing gesture and then the various audio inputs were juxtaposed on it, either synchronized or not. To create the synchronized stimuli, we combined the audio track of the different target words with the video track of the pointing movement so that the apex of the gesture movement coincided with the pitch peak of the target word (see the frame in Figure 1). To create the unsynchronized stimuli, we combined the audio track of each target word with the video track of the pointing movement in such a fashion that the apex of the gesture movement occurred in the middle of the unaccented syllable. Synchronized and unsynchronized stimuli were randomly mixed during the survey.

2.1.3. Procedure

Participants were asked to rate the acceptability of the video clips containing either synchronized or unsynchronized gesture-speech combinations on a 5-point Likert scale (1 = totally unnatural; 2 = quite unnatural; 3 = slightly unnatural; 4 = quite natural; 5 = totally natural).

Before the survey, participants were asked to imagine that the person in the videos was pointing at an object while naming it because she wanted to show them where the object was. Also, they were told that they had to base their acceptability judgments on the degree of coordination between gesture and speech that they perceived. The duration of the experiment was approximately 6 minutes.

2.2. Results

The total number of ratings obtained were 736 (23 participants x 32 clips), but 20 clips were found to have been left unrated by one or the other participant, so the total number of ratings analyzed was 716 (179 ratings for each of the four stimulus types, i.e. synchronized trochee, synchronized iamb, unsynchronized trochee, and unsynchronized iamb). An ANOVA analysis was carried out with acceptability rate as the dependent variable and stimulus type as the independent variable (four levels: synchronized trochee, synchronized iamb, unsynchronized trochee, unsynchronized iamb). The statistical analysis revealed that stimulus type significantly affected the acceptability rate ($F(3,715)=73.778$) = $p< .001$). Bonferroni post-hoc comparisons showed that, as expected, ratings for synchronized and unsynchronized trochees were significantly different ($p < .01$), and ratings for synchronized and unsynchronized iambs were also significantly different ($p < .001$), while synchronized trochees and synchronized iambs were rated similarly ($p> .05$). Surprisingly, ratings for unsynchronized trochees were also significantly different from ratings for unsynchronized iambs ($p< .001$). As Figure 2 shows, the mean acceptability rating for synchronized stimuli was very close to '4 = quite natural' (M = 3.79, *SD* = 0.983 for trochees; M = 3.89, *SD* = 0.963 for iambs). Unsynchronized iambs were rated very close to '2 = quite unnatural' (M = 2.36, *SD* = 1.331). However, participants judged unsynchronized trochees between '3 = slightly unnatural' and '4 = quite natural' (M = 3.39, *SD* = 1.050), thus more acceptable than unsynchronized iambs.

The results from Experiment 1 indicate that speakers detect the asynchrony between gesture and speech prominence, but it is more acceptable to them when the gesture apex occurs during an unaccented syllable in word-final (and also phrase-final in our stimuli) position (trochees) than during an unaccented syllable in word-initial position (iambs). Experiment 2 aimed at investigating the reason why misaligned trochees are more accepted than misaligned iambs.

[1]Capital letters indicate the accented syllable.

We hypothesized that when the pointing gesture conveys supplementary information to speech, speakers may misalign both modalities such that the gesture prominence can occur in post-tonic position but not in a pre-tonic one.

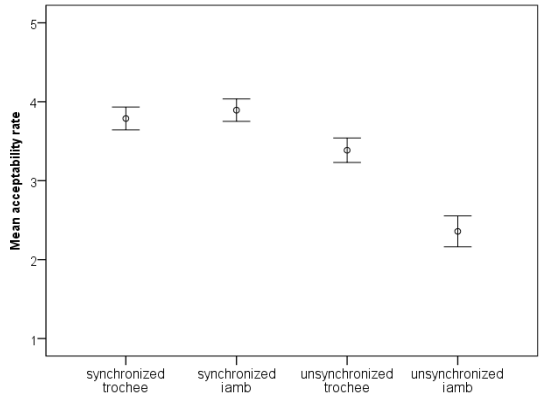

Figure 2: *Error bars of the mean acceptability rating as a function of stimulus type in Experiment 1.*

3. Experiment 2

In the second experiment we explored whether when the gesture is supplementary to speech speakers produce misaligned trochees (with gesture apexes in post-tonic position, i.e., phrase-final positions) but not misaligned iambs (with the apex in pre-tonic position).

3.1. Methods

3.1.1. Participants

Six Catalan-speakers participated in a pointing task. They were unaware of the purpose of the study and participated voluntarily.

3.1.2. Materials

In this pointing task, participants were asked to teach the experimenter the name of eight strange objects that were lined up in a row on a table (see Figure 3). The names of these objects were disyllabic nonsense words, half trochees (CVcv) and the other half iambs (cvCV), but all consisting of combinations of the same vowels and consonants, e.g. 'DUbi', 'duBÍ', 'BIdu' 'biDÚ'. Nonsense words were used to give meaning to the act of teaching and they were similar to make the game more challenging for the participants. Crucially, the participants had to name the object in the context of the sentence "*Agafa el* [target name]" ('Pick up the [target name]') and they were instructed not to produce any other kind of speech. Since the experimenter did not know which name referred to which object, participants were offered the possibility of using gestural strategies to indicate which object they were referring to.

3.1.3. Procedure

During the task, participants were recorded using a Panasonic HD AVCCAM recording at 25 frames per second. The sound was recorded through a small microphone that was placed somewhere on their clothing and as close as possible to their mouth.

Figure 3: *Setting of Experiment 2.*

At the beginning of the experiment, participants were given a legend in which the objects were labeled with their names. Participants were instructed to keep it hidden on their lap during the experiment and were then told that they were going to play a game in which they had to teach the experimenter the name of each object. In this teaching phase, the participant indicated the name of an object and its location to the interlocutor, then the interlocutor picked up that object, held it for a couple of seconds, and then put it back on the table. The participant then moved on to the next object. The task continued until the participant thought that the interlocutor would now be able to remember all the objects' names and locations. At that point the task ended and the interlocutor attempted to name all the objects.

3.1.4. Coding

All gesture-speech combinations that appeared in the video recordings were annotated using ELAN software in terms of the temporal features of both speech and pointing gestures. For speech, we annotated the temporal limits of the target name within the sentence, the metrical pattern of the name (either trochaic or iambic), and the temporal limits of the accented syllable within it. For pointing gestures, we annotated the preparation, stroke, and retraction phases of the gesture, and the location of the gesture apex [3].

3.2. Results and discussion

To examine whether participants produced unsynchronized trochees but not unsynchronized iambs, we calculated the location of the apex with respect to the end of the accented syllable as a function of the two metrical patterns. In total, 147 instances of items were analyzed, 73 with trochaic words and 74 with iambic words. Figure 4 and 5 illustrate the position of all the gesture apexes with respect to the accented syllable in trochaic and iambic words, separated by participant. In both figures, the solid horizontal line indicates the end of the accented syllable and the dotted line indicates the beginning of the accented syllable. Thus, circles occurring below the dotted line are cases in which the gesture apex occurs in the pre-tonic position and circles occurring above the solid horizontal line are cases in which the gesture apex occurs in the post-tonic (phrase-final) position.

Despite the high variability within and across participants, some patterns can be observed: (1) apexes occurring during the pre-tonic material are extremely scarce (3 cases in trochees and 4 cases in iambs, i.e. 4% and 5.4% respectively), and crucially all of them contain a pause between the pointing gesture and the upcoming speech; (2) in around one third of all instances, gesture apexes occur within the accented syllable (19 cases in trochees and 27 cases in iambs, i.e. 26.1% and 36.6% respectively); and (3) more than half of the participants produced the gesture apexes in phrase-final position,

irrespective of the metrical pattern (51 cases in trochees and 43 cases in iambs, i.e. 69.9% and 58% respectively).

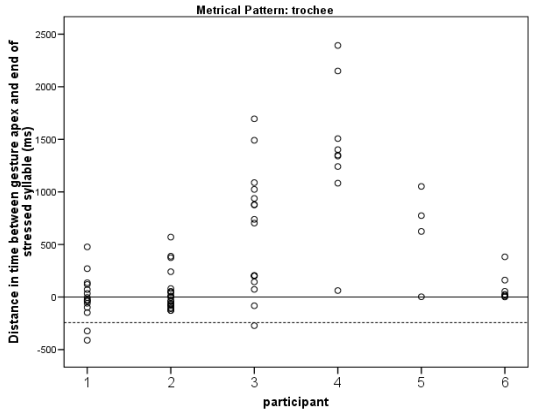

Figure 4: *Dispersion graph of the distance between gesture apex and end of the stressed syllable (in milliseconds) in trochaic words as a function of each participant.*

Chi-square tests indicated that the proportion of gesture apex occurring at a pre-tonic, tonic, or post-tonic position did not change across the two metrical patterns ($\chi^2(2) = 2.597$, $p > .05$). They also showed that the proportion of apexes at a pre-tonic position differed significantly from the proportion of apexes at tonic ($\chi^2(1) = 27.769$, $p < .001$) and post-tonic positions ($\chi^2(1) = 74.941$, $p < .001$), and a significant difference was also seen when comparing the proportion of apexes occurring at the tonic and post-tonic positions ($\chi^2(1) = 17.273$, $p < .001$).

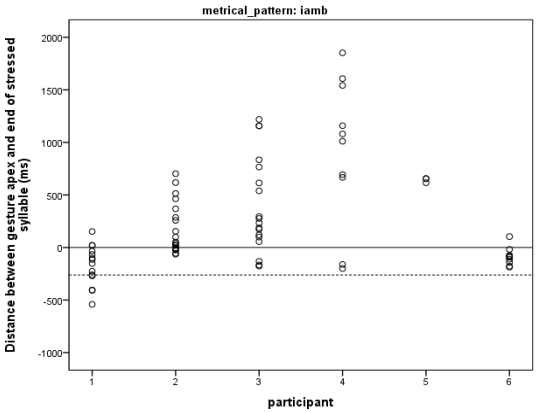

Figure 5: *Dispersion graph of the distance between gesture apex and end of the stressed syllable (in milliseconds) in iambic words as a function of each participant.*

We observed three strategies regarding the use of gesture and speech. The most frequent strategy was to utter a sentence and follow it by a pointing gesture, e.g. "Take the" [speech] + "object's name" [speech] + *it is this one* [gesture]. The second most frequent strategy was to utter unsynchronized pointing plus speech combinations, e.g. "Take the" [speech] + "object's name/ *it is this one* [gesture-speech combination]. And finally, there were few instances where the gesture apexes occurred during pre-tonic material, and these were produced with a pause between the pointing and the following word, e.g. "Take" [speech] + *this one* [gesture] + "which is called object's name" [speech]". These results show that pointing gestures can be produced before or after the target words, i.e. they are positioned at the edges of prosodic phrase boundaries, provided that they are perceived as separate speech acts carrying different semantic information.

4. Discussion

The purpose of this study was to investigate whether speakers detect temporal asynchrony between gesture and speech prominences (Experiment 1) and whether this perceptual ability is related to how they actually align gesture and speech in natural interactions (Experiment 2).

The results of Experiment 1 indicated that speakers do indeed detect asynchrony between gesture and speech prominences. However, surprisingly, unsynchronized trochees were perceived as more natural than unsynchronized iambs. More research is needed to investigate whether this effect is also found in trisyllabic words in which the misalignment of prominences can lead to an apex occurring at the pre-tonic or at the post-tonic position. This unexpected finding was further explored through a production experiment which elicited pointing gestures with the goal of teaching the name of the object and at the same time indicating its location. Our hypothesis was that speakers would rate unsynchronized trochees as fairly natural because in natural interactions speakers frequently align gesture prominences with phrase-final positions, especially when the semantic information conveyed by gesture is supplementary to the one conveyed in speech. Results of the production experiment (Experiment 2) confirmed this hypothesis: speakers produced practically no apexes during the pre-tonic material while apexes aligned during the post-tonic material were fairly frequent.

In our production study participants signaled the object they were referring to through a pointing gesture that frequently occurred after the object naming. It seems that speakers were actually saying "Pick up the object" using speech strategies + "that is there" using a pointing strategy. Thus, the gesture supplemented the meaning of speech and this affected the temporal coordination of the two modalities. This is not the first study showing evidence for the interrelation between semantic and temporal synchrony [9]. In [9] the authors found that gesture and speech timings were better aligned in complementary gesture-speech combinations than in supplementary gesture-speech combinations.

In sum, our results suggest that speakers perceive the alignment of gestural prominences by taking into account the temporal coordination of these gestures to prosodic heads (i.e. stressed syllables) or prosodic edges (i.e. phrase boundaries), and also by taking into account the semantic coordination of those gestures. Although further research is needed, this study has attempted to contribute to gain a better understanding of the temporal coordination between gesture and speech.

5. Acknowledgments

We thank Alfonso Igualada, Rafèu Sichel, and Santiago González for help with running the studies, and also the participants in the experiments. This research has been funded by grants FFI2012-31995, PSI-2011-25376, 2009SGR-701, and by the RECERCAIXA 2012 grant "Els precursors del llenguatge: una guia TIC per a pares i educadors".

6. References

[1] Birdwhistell, R. L. “Introduction to kinesics: An annotated system for analysis of body motion and gesture”. Washington, DC: Department of State, Foreign Service Institute, 1952.

[2] Kendon, A. “Gesticulation and speech: Two aspects of the process of utterance”. In M. R. Key (Ed.), The relationship of verbal and nonverbal communication (pp. 207–227). The Hague, the Netherlands: Mouton, 1980.

[3] McNeill, D. “Hand and Mind: What Gestures Reveal About Thought”. The Chicago University Press, Chicago, 1992.

[4] Wagner, P., Malisz, Z., and Kopp, S. “Gesture and speech in interaction: An overview”. Sp. Comm 57:209-232, 2014.

[5] Loehr, D. “Aspects of rhythm in gesture and speech”. Gesture 7: 179-214, 2007.

[6] Rusiewicz, H., Shaiman, S., Iverson, J., Szuminsky, N., Smith, A., and van Lieshout, P. “Effects of Prosody and Position on the Timing of Deictic Gestures”. J. Speech Lang. Hear. Res. 56(2):458-470, 2013.

[7] De Ruiter, J. P. “Gesture and speech production”. Doctoral dissertation. Katholieke Universiteit, Nijmegen, 1998.

[8] Esteve-Gibert, N. and Prieto, P. “Prosodic structure shapes the temporal realization of intonation and manual gesture movements”. J. Speech Lang. Hear. Res. 56(3): 850-864, 2013.

[9] Bergmann, K., Aksu, V., and Kopp, S. “The Relation of Speech and Gestures: Temporal Synchrony Follows Semantic Synchrony” in Proceedings of the 2nd Workshop on Gesture and Speech in Interaction. Bielefeld, Germany, 2011.

[10] Vatakis, A., and Spence, C. “Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli”. In M. J. Naumer & J. Kaiser (Eds.), Multisensory Object Perception in the Primate Brain. New York, Springer, 2010.

[11] Jesse, A., and Mitterer, H. “Pointing gestures do not influence the perception of lexical stress” in Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011): 2445-2448, 2011.

[12] Treffner, P., Peter, M., and Kleidon, M. “Gestures and phases: the dynamics of speech-hand communication”. Ecol. Psychol. 20:32-64, 2008.

[13] Leonard, T. and Cummins, F. “The temporal relation between beat gestures and speech”. Lang. Cognitive Proc. 20(1):32-64, 2008.