Prosody is in the hands of the speaker

Bahia Guellai [1], *Alan Langus* [2], *Marina Nespor* [2]

[1] Laboratory of Ethology Cognition and Development, University Nanterre, France
[2] SISSA Language Cognition and Development Lab, Trieste, Italy

bahia.guellai@gmail.com

Abstract

It has been suggested that speech and hand gestures could form a single system of communication that facilitates the interaction between the speaker and the listener. What kind of information do gestures carry? In the present study, we tested the possibility that spontaneous gestures accompanying speech carry prosodic information. Results show that gestures provide prosodic information as adults are able to perceive the congruency between a low-pass filtered – thus unintelligible - speech stream and the gestures of the speaker. These results suggest that prosody is not a modality specific phenomenon and can be perceived in spontaneous gestures that accompany speech.

Index Terms: prosody, hand gestures, speech perception.

1. Introduction

Human language is a multimodal experience: it is perceived through both the ears and the eyes. Adults automatically integrate auditory and visual information as evidenced by the McGurk effect [1], and seeing someone talking improves performances on speech intelligibility tasks [2]. This visual information involved in speech is not limited to the lips and the mouth but includes also the movements of the head [3, 4]. Other regions of the body could also give information about speech. Indeed, when interacting with others, people usually also produce spontaneous gestures while talking. What is exactly the role of these gestures that accompany speech? A line of research evidenced that gestures accompanying speech ease the speaker's cognitive load and gesturing help solving diverse tasks in mathematics and spatial problems [5, 6]. Gestures are also believed to aid the conceptual planning of messages as well as facilitate lexical access [7, 8]. This suggests that gestures and speech go 'hand-in-hand' from the earliest stages of cognitive development. In this view, gestures should carry the same structure as spoken language. One way to test this possibility is to look at prosody, an essential aspect of language.

In the auditory modality, prosody is characterized by changes in duration, intensity and pitch [9]. Interestingly, some part of the grammatical structure of human language is automatically mapped onto prosodic structure during speech production [10]. An interesting issue is whether prosody is modality specific or not. Since it has been shown to characterize sign languages as well [11], prosody cannot be restricted to the oral modality. It is therefore possible that the grammatical structure of language is not only automatically mapped to the acoustic speech signal but also to the spontaneous gestures accompanying speech.

Adult listeners use prosodic cues for various tasks that range from segmenting speech, to constraining lexical access [12], to disambiguating sentences that have more than one meaning (e.g., [bad] [boys and girls] vs. [bad boys] [and girls]) [10]. If some elements of grammatical structure are automatically mapped also to the spontaneous gestures accompanying speech, we should ask whether listeners use these gestures while processing the speech signal.

Thus, while there is evidence suggesting a direct link between the prosody of the speech signal and the spontaneous gestures that accompany speech, it is unclear whether listeners can use these cues provided by gestures when perceiving speech audio-visually. In the present study, we investigate the role of gestures as prosodic cues in speech perception.

2. Method

2.1. Experiment 1

In this first experiment, we explored whether gestures carry prosodic information. We tested Italian-speaking participants in their ability to discriminate audio-visual presentations of lowpass filtered Italian utterances where the gestures either matched or mismatched the auditory stimuli. While low-pass filtering renders speech unintelligible, it preserves the prosody of the acoustic signal [13]. This guaranteed that only prosodic information was available to the listeners.

2.1.1. Participants

We recruited 20 native speakers of Italian (15 females, mean age 24 ± 5) from the subject pool of SISSA – International School of Advanced Studies (Trieste, Italy). Participants reported no auditory, vision, or language related problems. They received a monetary compensation.

2.1.2. Stimuli

We used sentences that contain the same sequence of words and that can be disambiguated using prosodic cues from one of two different levels of the prosodic hierarchy. The disambiguation could take place at the Intonational Phrase (IP) level – the higher of these two constituents, coextensive with intonational contours – signaled through final lengthening and pitch resetting [10]. For example, in Italian, *Quando Giacomo chiama suo fratello è sempre felice* is ambiguous because depending on the Intonational Phrase boundary *è sempre felice* (*is always happy*) could refer to either *Giacomo* or *suo fratello* (*his brother*): (1) [Quando Giacomo chiama]IP [suo fratello è felice]IP (*When Giacomo calls him his brother is always happy*); or (2) [Quando Giacomo chiama suo fratello]IP [è felice]IP (*When Giacomo calls his brother he is always happy)*. Alternatively, the disambiguation could take place at the Phonological Phrase (PP) level where phrase boundaries are signaled through final lengthening. The Phonological Phrase extends from the left edge of a phrase to the right edge of its head in head-complement languages (e.g. Italian and English); and from the left edge of a head to the left edge of its phrase in complement-head languages (e.g. Japanese and Turkish) [10]. An example of a phrase with two possible meanings is *mappe di città vecchie* that is ambiguous in Italian because depending on the location of the PP boundaries, the adjective *vecchie* (*old*) could refer to either *città* (towns) or

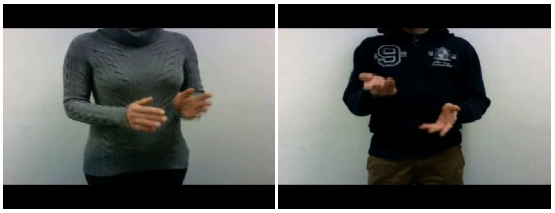
10.21437/SpeechProsody.2014-96

mappe (maps): (1) [mappe di città]PP [vecchie]PP (old *maps of towns*); or (2) [mappe]PP [di città vecchie]PP (*maps of old towns*). The presentation of the two types of sentences – those ambiguous at the IP level and those ambiguous at the PP level - was randomized across subjects. We video recorded two native speakers of Italian – a male and a female – uttering ten different ambiguous Italian sentences (see Table 1). The speakers were unaware of the purpose or the specifics of the experiments. The speakers were asked to convey to an Italian listener the different meanings of the sentences using spontaneous gestures. The videos of the speakers were framed so that only the top of their body, from their shoulders to their waist, was visible (see Figure 1). Thus the mouth – i.e. the verbal articulation of the sentences – was not visible. Two categories of videos were created from these recordings using the Sony Vegas 9.0 software. One category corresponded to the 'matched videos' in which the speakers' gestures and their speech matched and the second category corresponded to the 'mismatched videos' in which the gestures were associated with the speech sound of the same sequence of words, but with the alternative meaning. A total of 80 videos were created (each of the sentences was uttered twice). We ensured that, in the mismatched audio-visual presentations, gestures and speech were temporally aligned so that the beginning and the end of the gestures were aligned with the beginning and the end of the speech act. To remove the intelligibility of speech but to preserve prosodic information, the speech sounds were low-pass filtered using the Praat software with the Haan band filter (0-400 Hz). As a result it was impossible to detect from speech which of the two meanings of a sentence was intended. The resulting stimuli had the same loudness of 70 dB.

Table 1. *Example of a sentence with two different meanings depending on its prosody*.

Sentence	Meaning 1	Meaning 2
Quando Giacomo chiama suo fratello è sempre felice.	Giacomo è felice.	Suo fratello è felice.
When Giacomo calls his brother is always happy.	*Giacomo is happy*.	*His brother is happy*.

Figure 1: *Examples of the stimuli presented*.



2.1.3. Procedure

Participants were tested in a soundproof room and the stimuli were presented through headphones. They were instructed to watch the videos and answer – by pressing a key on a keyboard – whether what they saw matched or mismatched what they heard (i.e., [S] = yes or [N] = no). A final debriefing ensured that none of the participants understood the meaning of the sentences.

2.1.4. Results

The results show that participants correctly identified the videos in which hand gestures and speech matched (M=81.9, SD=11.03: t-test against chance with equal variance not assumed $t(19)$=12.93, p<.0001) and those in which they did not match (M=69.3, SD=10.17; $t(19)$=8.41, p<.0001). A repeated measures ANOVA with condition (Match, Mismatch) and type of prosodic contour (Intonational and Phonological Phrase) was performed on the mean percentage. The ANOVA only revealed a significant main effect for condition ($F(1,19)$=12.81, p=.002, ή² = 0.4), but neither for type of prosodic contour ($F(1,19)$=1.20, p=.287, ή² = 0.06) nor for an interaction of type and condition ($F(1,19)$=3.52, p=.076, ή² = 0.16). The results show that adult listeners detect the congruency between hand gestures and the acoustic speech signal even when only the prosodic cues are preserved in the acoustic signal. The spontaneous gestures that accompany speech must therefore be aligned with the speech signal, suggesting a tight link between the motor-programs responsible for producing both speech and the spontaneous gestures that accompany it. The results of Experiment 1 thus also show that adult listeners are sensitive to the temporal alignment of speech and the gestures that speakers spontaneously produce when they speak. We thus asked whether the prosodic cues that adult listeners use for understanding spoken language may automatically be mapped to gestures.

3. Discussion

Our findings show that when presented with acoustic stimuli that contain only prosodic information (i.e., low-pass filtered speech), participants are highly proficient in detecting whether speech sounds and gestures match. The prosodic information of spoken language must therefore be tightly connected to gestures in speech production that are exploited in speech perception. The syntactic structure and the meaning of utterances are therefore not necessary for the perceiver to align gestures and prosody. As opposed to the visual perception of speech in the speakers' face, where the movements of the mouth, the lips, but also the eyebrows [14] are unavoidable in the production of spoken language, the gestures that accompany speech belong to a different category that is avoidable in speech production. Our results suggest that prosody is a domain-specific phenomenon (i.e., characteristic of language) that extends from the auditory modality to the visual one in speech perception. This link between speech and gestures is congruent with neuropsychological evidence for a strong correlation between the severity of aphasia and the severity of impairment in gesturing [15]. While further studies are clearly needed to identify the specific aspects of spontaneous gestures that are coordinated with speech acts, our results demonstrate that part of speech perception includes the anticipation that bodily behaviors, such as gestures, be coordinated with speech acts.

4. Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 269502 (PASCAL), and the Fyssen Foundation.

5. References

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

[2] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

[3] Graf, P. H., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington D.C.

[4] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & VatikiotisBateson, E. (2004).Visual prosody and speech intelligibility. *Psychological Science, 15*, 133-137.

[5] Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General, 140*, 102-115.

[6] de Ruiter, J. P., Bangerter, A., & Dings, P. (2012), The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science, 4*, 232-248.

[7] Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*, 593-613.

[8] Pine, K. J, Bird, H., & Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. *Developmental Science, 10*, 747-754.

[9] Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken
language: A literature review. *Language and Speech, 40*, 141-201.

[10] Nespor, M., & Vogel, I. (2007). *Prosodic Phonology*. Berlin. Mouton De Gruyter.

[11] Nespor, M., & Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech. 42*, 143-176.

[12] Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language, 51*, 523-547.

[13] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago:
University of Chicago Press.

[14] Krahmer, E., & Swerts, M. (2004). More about brows: A cross-linguistic analysis-by-synthesis study, In: C. Pelachaud & Zs. Ruttkay (Eds.) *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers.

[15] Cocks , N., Dipper, L., Pritchard, M., & Morgan, G. (2013): The impact of impaired semantic knowledge on spontaneous iconic gesture production. *Aphasiology*.