Journal of Speech Sciences (9): 01-06. 2020.
Available at: http://revistas.iel.unicamp.br/joss

EDITORIAL: MULTIMODALITY, SEGMENTATION AND PROMINENCE IN SPEECH

MELLO, Heliana[1,2]
FERRARI, Lúcia[1]
ROCHA, Bruno[1]

[1]Universidade Federal de Minas Gerais
[2] CNPq

1 Widening the lens: joining multimodality, segmentation and prominence in speech

Speech and gestures meet at their departure point which is actionality. The same departing point keeps the two channels connected through their execution in the creation of meaning and interactivity. Both speech and gestures require segmentation in order to be studied and understood scientifically, as knowing what the units of analysis are is crucial to the scientific endeavor. Prominence is both a characteristic carried by prosody (be it defined functionally, physically or cognitively), as well as by several gestural acts, such as widening of the eyes, increased speed in hand motion, head tilting, among others. This link permits our joining multimodality, segmentation and prominence in speech as a topic for a scientific journal. As our knowledge about spoken language grows, thanks to empirically and experimentally based studies, the necessity for the never ending refining of methodologies is called into action, as well as the broadening of their boundaries. The understanding that gestuality actively interacts and partakes in communication is not a novel perception, as gesture forms a single system with speech and is an integral part of the communicative act (Kendon 1980; McNeil, 1992). However, the accurate pairing of how this interaction occurs is still not fully understood. Are gestures and speech additive, parallel, complementary? How are they linked in terms of the cognitive-neurological and motor routines involved?

This issue of JoSS emerged from the topics proposed at the X LABLITA and XI LEEL International Workshop: Prosody and Gesture, an initiative of the Empirical and Experimental Language Research Lab (LEEL) that took place in 2019, at the Federal University of Minas Gerais (UFMG). The LEEL lab focuses a large part of its team efforts into the compilation and study of spoken corpora. By organizing this event, the lab team aimed at bringing the correlations between gesture and prosody to the front, as well as supporting discussions about multimodal corpora compilation.

2 The papers in this volume

In this volume, contributions are made as to how gesture, segmentation and prominence occur in multimodal communication, taking a range of topics of study, as well as theoretical viewpoints. The volume is comprised of six articles which are sequenced according to the following rationale: the departing point is a theoretical-empirical proposal for the pairing of gesture and speech, taking into account prosodic segmentation as a defining factor for utterance recognition, gestural phrase segmentation and the correlation between information units and gestural phrases. Subsequently, studies that look into the analysis of multimodal communication, covering either a set of gestures or specific ones and their relationship to verbal

expression are featured. These studies cover the gestuality in L1 and L2, gestuality in intercultural communication, visual and auditory cues in the expression of specific illocutionary acts, and the correlation of lexical tone with auditory cues along with head and face movements. To close this JoSS issue, a study on the methodological steps for the automatic segmentation of speech intro prosodic units is presented.

The first paper in this JoSS issue, by Cantalini and Moneglia, entitled *The annotation of gesture and gesture / prosody synchronization in multimodal speech corpora* focuses on the functional and structural correlation between gestures and prosody. The authors empirically investigate the synchronization of gesture and prosody in spontaneous spoken Italian. Gestures were segmented and annotated following the LASG model (Bressem et al. 2013), while the segmentation and annotation of spoken Italian followed the L-AcT model according to Cresti (2000) and Moneglia and Raso (2014). The authors found that there was synchronization of verbal and gestural moves in 90% of the examined data, stressing the fact that gestural arcs coincide with prosodic boundaries. A very relevant finding pointed out by Cantalini and Moneglia is the fact that Gesture Phrases containing the Stroke (or expressive gestural phase) never cross terminal prosodic boundaries. This means that the scope of the utterance dictates the domain for the interaction between gesture and verbal activity. Another interesting result brought about by the authors is that Strokes coincide with all types of textual information units (these carry the core linguistic material in an utterance), but very rarely do they pair with dialogic information units, whose main function is to manage the flow of interactive communication, keeping cohesion among speakers.

Melussi and Capussotti in their paper entitled "*The egg and Jerry": narration and gesture in L1 and L2 by Italian schoolchildren* investigate the relation between speech and gestures in Italian monolingual schoolchildren, while speaking Italian (L1) and English (L2). The authors analyzed their production in order to understand whether factors such as language, style (monologue vs. dialogue) and sex influence the quantity, type and function of gestures. The subjects were 15 children (7 boys and 8 girls) aged around 9 years old, who have been studying English as L2 for approximately 3 years. Subjects were audio and video recorded while performing two different tasks: the first one consisted of describing a muted *Tom & Jerry* cartoon which they had previously watched. The second task was a structured interview conducted by one of the researchers. Both tasks were conducted first in L1 and, one week later, in L2. The gestures were annotated in ELAN according to a protocol developed by Capussotti (2019: 50-53), which distinguishes different types of hand gestures (McNeill, 1992), facial expressions (Ekman *et al*., 2002) and gestures produced with other parts of the body (McNeill, 1992). The data revealed that gestures vary quantitatively and qualitatively in L1 and L2 narrations: in L1, gestures were more frequent and had mainly the function of organizing the discourse, by complementing the speech and providing extra information to it. In L2, on the other hand, there were mostly iconic gestures, produced with lower amplitude, as part of problem-solving strategies. Furthermore, the authors have noticed differences between sexes: both in L1 and L2, female particpants gesticulated less than male participants, and their gestures had a smaller amplitude. However, they produced more articulated narrations in terms of lexical variety and syntactic structures than male participants.

The article by Schröder entitled *Between cultures: verbal, prosodic and gestural conceptualizations of interculturality in talk-in-interaction* presents an innovative proposal of a conversation and interactional analysis that takes into account cognitive and cultural aspects of real interactions in a multimodal perspective. The author discusses how metaphor and metonymy in gestures contribute for grounding and contextualizing the organization of face to face communication, highlighting elements and aspects of linguistic expressions. After

presenting the work of the research group Intercultural Communication in Multimodal Interactions, she discusses in detail two sequences of interactions extracted from the ICMI corpus. Videotape of the interactions are analysed through the software program EXMARaLDA (Schmidt and Wörner, 2009) and the transcription utilizes GAT 2 system convention (based on the transcription system of CA of Jefferson, 2004) focussing on the integration of the co-ocurrent cues of syntactic, pragmatic and prosodic levels. Schröder shows how the co-construction of intercultural experience can be distinguished by different aspects such as verbal and visual cues, signalled by gazes, pitch jumps and contours, rhythmic and intonational stylization, accents, lengthening and volume, as well as gestures, body movements, postures and enactments.

In the article by Miranda *et al.*, *Visual and auditory cues of assertions and questions in Brazilian Portuguese and Mexican Spanish: a comparative study*, the authors, based on previous literature, start from the premise that both production and perception of speech are multimodal. They focus on a visual-prosodic investigation of facial expressions comparing the production of Brazilian Portuguese (BP) both assertions and echo questions and Mexican Spanish (MS) assertions and yes-no questions. Analyses were carried on in forty BP and thirty MS set of data, focussing on facial gestures of speakers while uttering the speech acts. Prosodic investigation was made with a phonological notation of the nuclear region of assertive and interrogative intonational contours following the Autossegmental-Metric model (Pierrehumbert, 1980): Portuguese_ToBI (Frota et al., 2015) was used for analyzing BP data and Sp_ToBI (Prieto and Roseano, 2018) for MS. The facial movements in both languages were described using the FACS manual developed by Ekman et al. (2002). The authors describe the statistical tests employed to verify the variance of data and guarantee the significance of each dependent variable and check the differences among them. Their findings suggest that prosodic cues encode specific speech acts in each language, while the same facial gestures, as eyebrow lowering, lid tightening and nose wrinkle, are used in questions in both languages.

The article *A method for lexical tone classification in audio-visual speech* by Menezes *et al.* discusses multimodality in tone languages and proposes, using statistical comparisons of acoustic signal and face movement components, a method for assessing tone classification through acoustic parameters and the impact of face and head movements on speech. The authors tested the procedure in a data collection of experimental speech production of a female native speaker of Cantonese that uttered a set of syllables and words covering the 6 lexical tones of the Cantonese language. An NDI Optotrak device along with a high-quality microphone captured the acoustic signal and the visual component by 33 active tracking markers placed on the speaker's face, head and neck. Data was then organized in four signal groups (F0, Face, Head, Face + Head) and parametrized with a K-Fold Cross Validation (KFCV) to train an LDA (Linear Discriminant Analysis: James *et al.*, 2013) classifier. Those steps permitted that each signal group predicted the correct lexical tones. As expected, the F0 signal presented the highest accuracy but it is interesting that all signal groups presented considerable reliability to lexical tone classification. The meticulous description of this method allows its replication in future studies, whether multimodal or not.

In the article *Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech*, Raso *et al*. investigate the phonetic-acoustic parameters responsible for the production/recognition of prosodic boundaries. The primary aim of the paper is to build models that can automatically identify two levels of boundaries (terminal and non-terminal) in male monological spontaneous speech of Brazilian Portuguese. Furthermore, the paper discusses the parameters that compose the models in order to make assumptions regarding the physical cues that guide the human perception of prosodic breaks. The models were

developed to reproduce the results obtained by two groups of trained annotators while segmenting two samples of spontaneous speech into intonational units. The positions in which at least 50% of the annotators indicated a break of the same type were considered by the script to be a prosodic break of that type. For each position, the script extracted a set of 111 acoustic parameters comprising measures of speech rate and rhythm, standardized segment duration, fundamental frequency, intensity and silent pause. Different models were developed to analyze terminal and non-terminal breaks. By analyzing each model individually, the authors found that almost every model takes into consideration pause-related parameters. Additionally, the authors have noticed that f0-related parameters are more relevant to the models aimed to identify terminal breaks, while the duration-related measurements are crucial for every non-terminal break model. However, the models cannot be summarized by those few observations: actually, they are formed by a large set of parameters and by the weights assigned to each one of them. In the final part of the paper, the authors discuss various methodological aspects that emerged during the research and, based on them, point out some strategies for the improvement of the models.

3 Connecting speech and gesture at multiple levels

In order to study the correlation of the acoustic signal in speech and gesture in multimodal communication, authors resort to segmentation and the study of prominence, following a host of proposals and theories. Speech and gesture have been linked in production, departing from a common source in working memory and are connected through different analytical levels. The synchronization of gesture and speech is summarized through three models that can be seen in Figure 1, below:

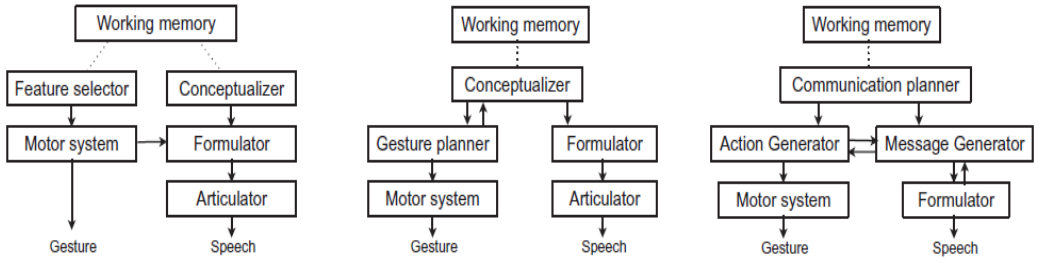

Figure 1: Three different models that can account for interaction between speech and gesture production From left to right: Krauss and Hadar (1999), de Ruiter (2000) and Kita and Özyürek (2003). Source: Wagner *et al*. (2014:2018)

These models, however, do not fully explain the minutiae of the workings of the pairing speech/gesture. One of the many still open questions that remain regards what kind of elements are correlated in the pairing – are they specific lexical items or prominent syllables that correlate temporally to particular gesture types? What happens in quick sequences of gestures, compressed into a single phrase? What are the processing effects that are associated individually to speech and gesture?

Yet another front of investigation that requires attention and research deals with of prominence in prosody and gestuality. The later have been called audiovisual prosody and are believed to closely interact with speech (Wagner *et al*, 2014:220), facilitating comprehension or serving the purpose of focus, and multimodal prominence. The findings, although usually

spotlighting head and hand movements, can be performed by other body parts as well. Head movements, however, have been pointed out as being very prominent in taking part in the expression of emotions, attitudes, engagement, besides expressing evaluations of the ongoing discourse.

Last, but not least, a very relevant investigation front related to multimodal communication deals with the technical elements related to data collection, storage, annotation, computational tools and the applications derived thereof. All the papers in this JoSS issue discuss extensively the methodological decisions taken and demonstrate that there are many possible annotation schemas and software available to be adopted, depending on the issues to be handled and a researcher's own preference. The size of datasets examined still faces restrictions dictated by the amount of time and resources involved in the treatment of multimodal data. The perspectives for the future, however, are bright, as more interest from the community and the rekindled effervescence of empirically based studies seems to be on the increase at present.

It is our hope that the contribution brought about by this JoSS issue will help support research initiatives into the nature of multimodal communication and its contributing elements.

The authors of this volume are grateful to Fapemig for financing the research.

REFERENCES

1. Bressem J, Ladewig SH., Müller C. Linguistic Annotation System for Gestures (LASG). In Müller C, Cienki A, Fricke E. Ladewig SH., McNeill D, Teßendorf S. (eds.), *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38)* Vol. 1, Berlin: De Gruyter Mouton, 2013, 1098–1125.
2. Capussotti G. *La multimodalità in italiano e inglese LS: relazione tra gesto e discorso nei bambini di una scuola primaria.* Unpublished Master's Degree Thesis, University of Pavia, Italy, 2019.
3. Cresti E. *Corpus di italiano parlato.* Firenze: Accademia della Crusca, 2000.
4. de Ruiter J P. The production of gesture and speech. In McNeill, D (ed.) *Language and gesture*, Cambridge: Cambridge University Press, 2000, 284-311.
5. Ekman P, Friesen WV, Hager JC. *The Facial Action Coding System*. Salt Lake City: Research Nexus, 2002.
6. Frota S, Oliveira P, Cruz M, Vigário M. *P-ToBI: tools for the transcription of Portuguese prosody*. Lisboa: Laboratório de Fonética, CLUL/FLUL, 2015b. ISBN: 978-989-95713-9-6. [http://labfon.letras.ulisboa.pt/InAPoP/P-ToBI/]
7. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York: Springer, 2013.
8. Jefferson G. Glossary of transcript symbols with an introduction. In: Lerner G (ed.). Conversation analysis: Studies from the first generation. Amsterdam, Philadelphia: John Benjamins, 2004, 13–31.
9. Kendon A. Gesticulation and speech: two aspects of the process of utterance. *Relationship of verbal and nonverbal communication* (ed. & Key MR), pp. 207–228. The Hague, The Netherlands: Mouton, 1980.
10. Kita S, Özyürek A. What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language* 48(1), 16-32.
11. Krauss R M, Hadar, U. The role of speech-related arm/hand gestures in word retrieval. In: Campbell R, Messing L. (eds.) *Gesture, Speech, and Sign*. Oxford: Oxford University Press, 1999, 93–116.
12. McNeill D. *Hand and mind: what gestures reveal about thought*. Chicago, IL: University of Chicago Press, 1992.
13. Moneglia M, Raso T. Notes on the Language into Act Theory. In Raso T, Mello H. (eds), *Spoken corpora and linguistics studies.* Amsterdam: Benjamins, 2014, 468–494.
14. Pierrehumbert J. *The phonology and phonetics of English intonation*. Bloomington: Indiana University Linguistics Club. PhD thesis, MIT, 1980. [Published 1987 by IULC edition, Bloomington, IN.].

15. Prieto P, Roseano P. Prosody: Stress, Rhythm, and Intonation. In: Geeslin KL. (ed.) *The Cambridge Handbook of Spanish Linguistics*. Cambridge: Cambridge University Press, pp. 211-236, 2018.
16. Schmidt T, Wörner K. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19, 2009, 565–582.
17. Wagner P, Malisz Z, Kopp S. Gesture and speech in interaction: an overview. *Speech Communication* 57, 2014, 209-232.