A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualisation

Angelika Hönemann, Hansjörg Mixdorff, Sascha Fagel

Abstract

This study investigates the relationship between prosodic speech features, such as syllable prominences and phrase boundaries, and visual cues, such as head and facial movements. The insights gained from the study provide the basis for a predictive model that generates prosodic visual cues from speech signals. Such predictive models have many interesting applications, for example, the control of non-verbal movements for the visualisation of voice messages by an avatar. Our dataset consists of synchronously recorded audio and video signals, as well as motion capture data from seven speakers. The 3D data were recorded by means of an optical method, using the Qualisys motion capture system. The acoustic data was segmented at the syllable level and the prominent syllables, phrase types and phrase boundaries were annotated using Praat. We annotated the visible motions manually in the digital video sequences using Anvil. This audiovisual corpus was subjected to a preliminary statistical analysis which included the eye, eyebrow, lip and head movements in relation to prominent syllables, phrases and phrase boundaries. Our results show that for each speaker 20–30% of events in each motion class are aligned with prominent syllables in phrase-initial or -medial position and that the speakers moved most often at the end of an intonation phrase. Movements, however, differed in strength and frequency between speakers.

1. Introduction

Speech is a powerful means of expression that provides a wide variety of ways to convey information for mutual understanding. This information is generated, transmitted and received on acoustic and visual channels. Speech is therefore a multi-modal process. We can see it every day when people talk. We can observe movements such as the raising of the eyebrows, the nod of the head or a smile.

Studies confirm that not only articulatory speech processes produce visible movements, but also prosodic features show up visually. It has been shown that there is a strong correlation between facial and head movements and the prosodic structure of a text (Graf et al., 2002).

Facial and head movements are associated with speech and emphasise what has been said and lead to a better understanding by the listener. It has been shown that the influence of visual cues is significant for the perception of prominences. When eyebrow and head movements are associated with certain words these are perceived as emphasised by the listener, as opposed to when no movements are observed (Krahmer & Swerts, 2007).

Our main interest is to develop a predictive model for the control of non-verbal movements which could be used for speech visualisation by an avatar. The persuasiveness of an avatar depends on the interaction between the acoustic and visual cues, and should therefore be realistic and natural. Speech visualisation via avatar requires a high agreement of these two modalities; otherwise they will appear less convincing or even unintelligible. On that account we need a better understanding of the correlation between these audiovisual features in human-human communication.

To this end we performed a preliminary experiment providing a natural environment in which the test subjects felt relaxed. We collected a dataset in the style of a free narrative recounting the vacation of the test subjects. The individual stories offer a wide prosodic range: for example, they contain sentences of different lengths, breaks, hesitations, etc. Also the speakers displayed a large variety of movements.

While the test subjects talked we synchronously captured the facial and head movements with three Qualisys infra-red cameras (www.qualisys.com). We also recorded a digital video on which we based our first perceptual evaluation. Motion events investigated included the eyes, eyebrows, lips and rigid head motion. Our rationale is that any relevant motion should be visually detectable and only in that case the underlying motion capture data should be subjected to a closer analysis. In the scope of this paper we therefore concentrate on the video only.

This paper is structured as follows: Section 2 describes in detail the annotation of the acoustic and visual data and the results which it has yielded. Here you can find the data used for the statistical analysis, for example, the numbers of syllables, prominences, phrases and motion events of each speaker, the average duration and standard deviation. Section 3 contains the results of the statistical analysis which includes the distribution of motion events of each speaker in relation to three target syllable types and the percentage share of the total number of each motion at the end of the phrase is shown. Section 4 offers the conclusion of our paper.

2. Audiovisual data corpus

Our audiovisual corpus consists of narratives with a length of about one minute spoken by three male and four female German native speakers. Due to the free narrative, the speakers behaved in a natural way, so that an investigation of natural facial expressions is possible. On the down side, materials produced are unrestricted and therefore direct utterance comparisons impossible. The stories offer wide prosodic variety.

2.1. Acoustic data

The acoustic data were segmented at the syllable level and the target syllables, phrase and phrase boundaries were annotated using Praat (Boersma & Weenink, 2012). As target syllables we labelled (A) accented syllables in phrase-initial and medial positions, (B) unaccented syllables phrase-finally, (A/B) accented syllables phrase-finally.

Table 1: Total number of syllables, total number of target syllables and the percentage share of the target syllable.

Feature / Speaker		*sp01*	*sp02*	*sp03*	*sp04*	*sp05*	*sp06*	*sp07*
Syllables	*Total*	*282*	*320*	*232*	*176*	*360*	*242*	*308*
Target Syllables Percentage Ratio (%)	*Total*	114(40.4)	116(36.2)	91 (39.2)	66 (37.5)	133(36.9)	87 (36.0)	124(40.3)
	A	59 (20.9)	76 (23.8)	53 (22.8)	36 (20.5)	80 (22.2)	54 (22.3)	74 (24.0)
	B	28 (9.9)	26 (8.1)	24 (10.3)	11 (6.2)	27 (7.5)	18 (7.4)	30 (9.7)
	A/B	27 (9.6)	14 (4,4)	14 (6.0)	19 (10.8)	26 (7.2)	15 (6.2)	20 (6.5)

Table 1 shows the total number of syllables, the total number of target syllables and the percentage ratio of target syllables to the total number of syllables for each speaker. The average duration of the syllables is 214 ms and the standard deviation 171 ms. The average ratio of syllables labelled as prominent is 38.1% thereof A accented syllables 22.5%, B unaccented syllables 7.0% and A/B accented syllables 8.5%.

In addition we segmented the narratives into phrases which were labelled as follows: C non-terminal phrase which is characterised by a rise in intonation at the end of the phrase, D declarative phrase which is identified by a phrase-final fall in intonation, I interrogative phrase, i.e. for questions. Whether the intonation at the end of an interrogative phrase is rising or falling depends on the kind of question, i.e. for yes/no questions the intonation indicates an increase. Due to the spontaneous production the speakers showed hesitations labelled with H. BR indicates breaks longer than 150 ms. Typically, the speakers took a break to inhale or pre-plan the discourse.

The phrase boundaries were annotated following the GToBI conventions (Grice et al., 2005). The break index describes the degree of perceived separation between the final syllable and the silence at the end of an utterance, i.e. the subjective strength of the boundary.

We used three break indices (BI): 2 for breaks without tonal marking, hesitations and when the final syllable of a phrase was delayed, 3 for an intermediate

phrase, and 4 for an intonation phrase. Table 2 shows the distribution of the different types of phrases and phrase boundaries.

Table 2: Total number of different types of phrases and phrase boundaries.

Phrases / Speaker	*sp01*	*sp02*	*sp03*	*sp04*	*sp05*	*sp06*	*sp07*
C non-terminal	34	31	25	20	34	17	46
D declarative phrase	21	9	13	10	17	16	4
I interrogative phrase	0	0	0	0	2	0	0
H Phrase	4	6	3	9	4	3	15
BR Phrase	17	18	13	8	21	12	19
Phrase-Boundaries							
2 without tonal breaks	30	32	27	18	35	21	46
3 intermediate phrase	19	9	8	17	16	7	18
4 intonation phrase	27	23	19	12	27	20	20

As can be seen the C phrases, the non-terminal phrases, have the greatest share of the distribution. We suppose that the reason could be that the speakers call for the listener's attention. A speaker signals by a rise of intonation at the end of the phrase that he wants to continue speaking. He does not want to lose the listener's attention. Conversely, when the intonation falls, the probability is high that the speaker has finished or is about to finish his turn.

The average duration and the standard deviation (s.d.) of the different phrase types are as follows: C non-terminal phrases: 1056 ms (s.d. 571 ms), D declarative phrases: 1188 ms (s.d. 580 ms), I interrogative phrases: 909 ms (s.d. 493 ms), BR Phrases: 625 ms (s.d. 306 ms) and the H Phrase: 466 ms (s.d. 263 ms).

2.2. Visual data

We annotated the motions which we perceived on the basis of the digital video sequences. We used the Anvil annotation tool which offers the possibility to define our own coding scheme (Kipp, 2001).

Figure 1 exhibits the Anvil Annotation Tool and the annotation of the motion events of sp06. It shows the digital video frame on the left side and the annotation window on the right side. The annotation window includes from top to bottom the oscillogram of the audio, the F0 curve (dotted line) and intensity curve (continuous line), the segmented words and syllables, the topic- and phrase structure and the facial expressions conveyed by the head, eyes, eyebrows and lips.

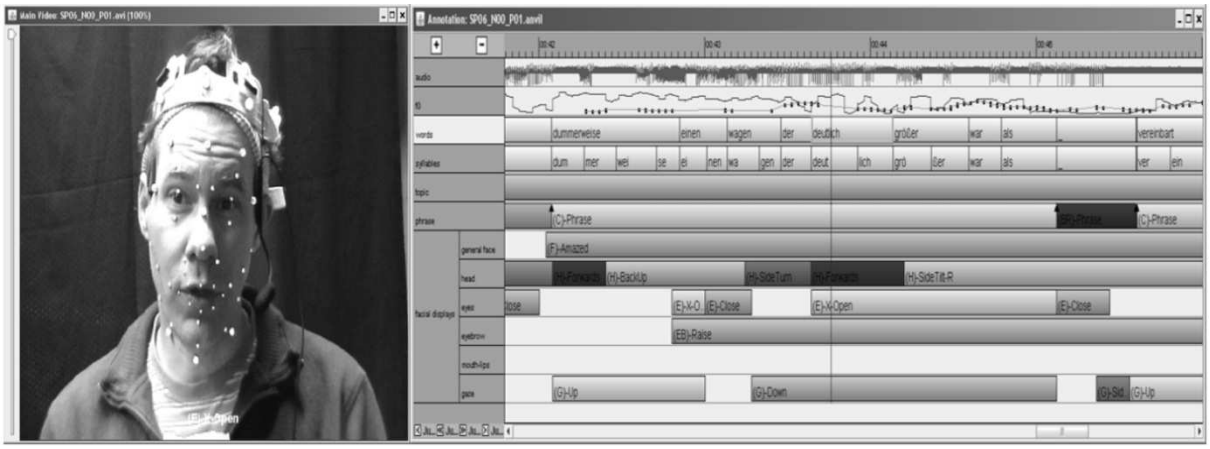

Figure 1: Anvil Annotation Tool: Annotation of sp06.

We used the MUMIN coding scheme but only in parts because it was originally developed for coding dialogues (Allwood et al., 2004). There were many codes which were not suitable and we also needed additional codes. Hence we decided to expand the MUMIN coding scheme with our own classifiers.

We defined, for example, for rotational movements of the head (H)-Down as the head motion down and back to the neutral position, (H)-BackUp as the head motion up and back to the neutral position, (H)-Nod as the repeat up and down of the head motion, and (H)-SideTurn-R as the repeat of a side turn, i.e. a shake. For translational head movements we defined (H)-Backwards and (H)-Forwards as the head motions backwards and forwards and back to the neutral position, and (H)-SideTilt as a short tilt motion to the side. There was a special kind of movement – a mix of head movements which we called (H)-Waggle.

Table 3: Overview of the perceived movements of head, eyes, eyebrow and lips.

Head	*Eye*	*Eyebrow*	*Lips*
(H)-Down	(E)-Close	(EB)-Rise	(L)-Down
(H)-BackUp	(E)-X-Open	(EB)-Frown	(L)-Up
(H)-Nod	(E)-Flutter		(L)-Protruded
(H)-SideTurn			(L)-Open-M
(H)-SideTurn-R			
(H)-Backwards			
(H)-Forewards			
(H)-SideTilt			
(H)-SideTilt-R			
(H)-Waggle			

The movements of the eyebrows were classified as (EB)-Raise and (EB)-Frown, and those of the eyes as (E)-Close, (E)-X-Open, (E)-Flutter which means a fast eyelid motion but not a closure of the eyes. For the lips we defined for example (L)-Up, (L)-Down for the up and down movements of the corner of the mouth, and (L)-Open-M. The (L)-Open-M classifier we used only at a speech break. It is interesting to see what the speaker does during a speech break because it is not clear why the speaker makes a break. We suspected that there are other reasons besides the inhale. Table 3 gives an overview of the all perceived movements of the head, eyes, eyebrows and lips.

The result shows significant differences between the speakers' motions. There are many possible reasons for this such as the temperament or the origin of a speaker. Table 4 shows the number and average duration of each motion event of each speaker.

Table 4: Number and average duration (ms) of each motion class of each speaker.

Speaker/ Motion Events	*Head*		*Eyes*		*Eyebrows*		*Lips*	
	No.	*Average duration ms*	*No.*	*Average duration ms*	*No.*	*Average duration ms*	*No.*	*Average duration ms*
Sp01	36	714	29	396	-	-	16	1595
Sp02	52	624	32	321	1	224	8	795
Sp03	33	1145	53	461	8	945	10	892
Sp04	30	848	27	404	1	320	8	2730
Sp05	34	915	46	307	-	-	15	765
Sp06	43	971	41	339	4	148	10	724
Sp07	65	788	25	398	3	280	18	673

The frequency of each motion class describes how often one motion event occurs per second. It shows the activity of each speaker. The basis is the total duration of all motion events, i.e. the head, eye, eyebrow and lip motions of each speaker. The frequencies are as follow: sp01: 1.3 Hz, sp02: 1.8 Hz, sp03: 1.3 Hz, sp04: 1.1 Hz, sp05: 1.6 Hz, sp06: 1.4 Hz, sp07: 1.5 Hz. On average over all speakers there are approximately 1.5 motion events per second. In detail the average frequency of all speakers for the head motion is 1.2 Hz, for the eye motion 2.7 Hz, for the eyebrow motion 1.1 Hz and for the lip motion 1.1 Hz.

3. Results

Our first statistical analysis is the investigation of our speakers' movements in relation to prominent syllables, phrase and phrase boundaries.

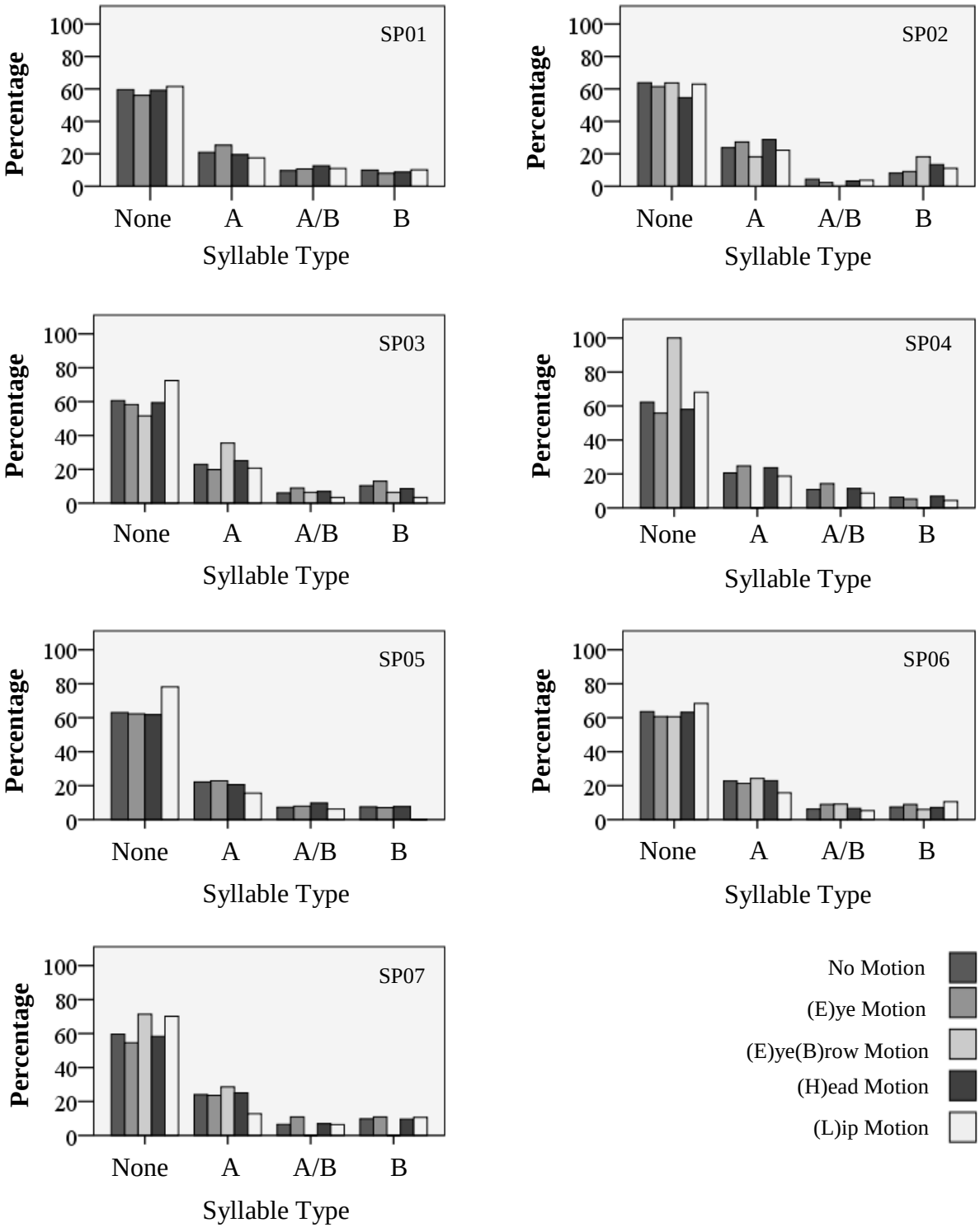

Figure 2: Distribution of motion events for each speaker in relation to the target syllable types A accented syllables, B unaccented syllables and A/B accented syllables phrase-finally, and non-target syllables.

3.1. Prominent syllables

The seven graphs in Figure 2 show the distribution of motion events of each speaker in relation to the target syllable types, the A, B and A/B syllables, and non-target syllables.

The speakers' movements were greater at prominent syllables in phrase-initial or medial position than phrase-final: 20% to 30% of the motions that the seven speakers made align with A syllables. At the A/B and B syllables all speakers show fewer movements, less than 15%.

Our results indicate that the speakers have different preferences of movements to mark the target syllables. For example, sp01 prefers head motions such as (H)-SideTilt-R and (H)-Down in contrast to sp05 and sp07 who did not move or moved less in that way. They show movements such as (H)-Nod or (H)-BackUp. Sp02, sp04 and sp05 obviously prefer (H)-forwards. We perceived many (E)-close motions but the reason is the blink to wet the eyes. We did not differentiate if the eyes were closed because of the eye reflex or for another reason.

The graphs in Figure 2 also show that more than half of the motion events occur at syllables which were not classified as target syllables. There are of course other reasons for facial and head movements beside prominences. Our assumption is that the content of the spoken text could be a main point. This should be further examined.

3.2. Phrase and phrase boundary

Table 5: Average percentage share of each motion event of all speakers at the end of the phrase.

Phrase-Boundaries	*Head %*	*Eyes %*	*Eyebrows %*	*Lips %*
C-2	2.6	2.5	0.0	0.3
C-3	4.2	4.0	2.4	5.9
C-4	5.4	5.3	8.3	4.5
D-2	1.0	1.7	0.0	1.1
D-3	0.8	1.1	0.0	1.1
D-4	2.8	3.4	2.4	2.3
I -4	0.2	0.0	0.0	0.0
H -2	1.7	1.6	1.2	4.2
BR -2	5.4	7.5	7.1	19.5

Table 5 shows the percentage distribution of each motion class of all speakers at the end of the phrase. In the left column there are the labels of phrases and phrase boundaries. For example, C-2 means a non-terminal phrase without a clear tonal marker. D-3 is a declarative phrase as an intermediate phrase, H-2 a hesitation and BR-2 a pause both of which do not have a clear tonal marker.

It is obvious that the non-terminal phrases with a BI of 4, that is, the C-4, have the most head, eye and eyebrow movements. The declarative phrases have fewer movements than the non-terminal phrases, although still the highest share in the BI 4 boundary. Obviously the speakers moved most often at the end of an intonation phrase. This indicates that the speakers' movements accompany the rising of intonation.

There are also many lip motions at a speech break. We annotated the tag (L)-Open-M at a break if the speakers opened their mouth. In addition to inhaling we assume that the speakers used the break to prepare for the discourse.

4. Conclusion

This paper is a preliminary investigation of the relationship of prosodic features, that is, prominent syllables, phrase and phrase boundaries with facial and head motions, especially rigid head movements, eye, eyebrows and lip movements.

Our results show that 20–30% of the speakers' movements occur at accented syllables which are at a phrase-initial or medial position. However, there is also a great ratio of about 60% of motion events which are not assigned to the target syllables. We assume that the content of the spoken text plays an important role. We also need to expand the analysis to the syllables before or after a prominent syllable.

We also found that the test subjects moved at an intonation phrase more often than at an intermediate phrase, hesitation or during a pause. This suggests that there is a strong relationship between the rising of the intonation and facial and head movements.

We pointed out that the movements of the speaker differ with respect to motion and frequency. How a speaker moves obviously depends on his individual preferences. Probable reasons for the difference in strength of motion are the temperament and emotional state of a speaker.

A more detailed future analysis will concern rises and falls of the fundamental frequency (F0). We need a more accurate description of the target syllables such as the GToBI transcription scheme for pitch accents. Another important parameter is intensity which we did not consider for this investigation.

Acknowledgements

The first author is funded by the European Social Fund (ESF) and supported by the Berlin Senate for Economics, Technology and Research. We thank all our test subjects and a special thanks go to Adam Cleeve for his patience and support.

References

Allwood, J., Cerrato, L., Dybkjær, L., Jokinen, K., Navarretta, C., & Paggio, P. (2004). *The MUMIN multimodal coding scheme*. Technical Report retrieved from <www.cst.dk/mumin/>. CST, University of Copenhagen, Denmark.

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer* (Version 5.3) [Computer program]. Retrieved February 21, 2012, from <www.praat.org/>.

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 02)* (pp. 396–401).Washington, DC., USA.

Grice, M., Baumann, S., & Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 55–83). Oxford: Oxford University Press.

Kipp, M. (2001). Anvil - A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European conference on speech communication and technology (Eurospeech)* (pp.1367–1370). Aalborg, Denmark.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414.

Next contribution