

Perception of ‘Speech-and-Gesture’ Integration

Gianluca Giorgolo¹, Frans A.J. Verstraten²

¹UiL-OTS, Universiteit Utrecht, The Netherlands

²Helmholtz Instituut, Experimental Psychology Division, Universiteit Utrecht, The Netherlands

gianluca.giorgolo@let.uu.nl, f.a.j.verstraten@uu.nl

Abstract

This paper describes two experiments conducted to identify the role of synchronization in the perception of ‘speech and gesture’ communication and to isolate the parameters that determine the perception of temporal alignment. The results of the first experiment show that the *synchronization* between audio and visual signals determines the felicitousness of a multimodal utterance. With the second experiment we were able to determine that *prosodic alignment* is a parameter that our subjects used to judge the ‘well-formedness’ of speech and gesture input.

Index Terms: speech and gesture integration, audiovisual perception

1. Introduction

Synchronization is an important concept in human communication. It determines the felicitousness of many types of human communication: consider for example the unpleasant effect of an instrument playing out of *tempo* in an orchestra. Audiovisual perception relies to a large extent on temporal alignment, so that it can be easily manipulated by exploiting synchronization: more than thirty years ago McGurk and MacDonald ([1]) showed the strong influence of visual stimuli on auditory perception¹. We are sensitive to synchrony from very early stages in our life: Rosenblum et al. ([3]) describe a set of experiments that lead them to obtain a ‘McGurk-like effect’ in 5-month-olds and Hollich and colleagues ([4]) showed that in general 7.5-month-olds use synchronized audio-visual cues when filtering target words in noisy environments. The intuition that temporal alignment plays an important role in human communication has been exploited also in gestural studies. McNeill ([5, 6]) proposes three *synchronization patterns* to describe the relation between the verbal and the gestural channels:

1. Temporal/prosodic synchronization: the meaningful part of a gesture (the so called *stroke*) co-occurs with (or slightly precedes) the prosodically most prominent segment of the accompanying speech.
2. Semantic synchronization: the meaning expressed by gesture is compatible with the one expressed by the connected speech fragment.
3. Pragmatic synchronization: gesture and speech work together to achieve the goals of communication.

These patterns have been the base for much of the recent research on gesture, which, on the other hand, has been mainly concentrated on gesture production related issues. Only recently, researchers have started looking into the perception of

¹However, notice that Munhall et al. ([2]) proved that the temporal constraints for the ‘McGurk effect’ are quite loose.

‘speech and gesture’ communication (see for example [7, 8, 9]). For example they try to identify the processes underlying the integration and the comprehension of a message distributed among different channels.² The current experiments are a contribution to this enterprise. Our experiments are related to those of Özyürek et al. and Willems et al. [7, 8]. While their main interest lays in investigating the integration of the semantic content of speech and gesture on the receiving side, we tried to analyze the role of ‘low-level’ parameters -more specifically prosody- in speech-gesture integration.

We designed two behavioral experiments based on a ‘violation’ paradigm. The idea is to exploit those parameters which we assume are used by the subjects during the cognitive task. By changing the values of these parameters from a maximally acceptable value to highly deviant ones, we expect a corresponding change in the response of the subjects. In our case the parameter under investigation is temporal alignment between gesture and speech, so the stimuli will comprise a hierarchy of differently aligned audio-visual clips. The responses of the subjects should change accordingly.

We can schematically summarize the rationale behind the experiments with the following (abductive) reasoning:

- We know that listeners/viewers are sensitive to synchronization/temporal alignment of audio and visual stimuli ([12, 1]).
- We know that listeners/viewers are continuously exposed to multimodal input and use the two channels for their communicative goals (both producing multimodal utterances and perceiving them).
- We know that prosody patterns quite regularly with gestures (at least at level of pitch accents and strokes, and, to a less extent, at level of gesture phrases and intermediate phrases, see [13]).
- *Assuming* that one of the parameters for well-formedness in the case of multimodal utterances is the temporal alignment between prosodic ‘movement’ and manual activity.
- Thus, we can expect that listeners/viewers will be sensitive to not properly prosodically aligned multimodal utterances and will rate them as ill-formed.

In other words, our hypothesis is that speech and gesture integration relies on prosodic and kinematic cues, similarly to the cross-modal integration described in the literature.

The paper is structured as follows: section 2 describes the experiments in detail, justifying some choices made to adapt the

²The fact that gestures convey a meaning is a disputed topic, while much of the research seems to suggest that this is the case, the consensus in the field is not unanimous. See [10, 11] for a discussion.

general violation paradigm to the specific case under discussion, section 3 summarizes the results and the statistical analyses performed on them and finally in section 4 we propose our interpretation of the results.

2. Experiments

We describe two experiments that address the question regarding the role of prosody in the perception and integration of speech and gesture.

The first of the two experiments was designed to test a precondition to run the second experiment. As said, listeners should be sensitive to the temporal alignment of the audio and visual channels. However, this is a known fact for facial stimuli. Our experiments look for the effect of gestures. We try to find evidence, showing that the same type of sensitivity applies also for the case of manual gesture. The first experiment tries to verify the assumption that temporal alignment matters also for speech and gesture stimuli. Participants were presented with audio-visual stimuli displaying different alignments between the audio and the video channels and were asked to judge the naturalness of the alignment.

The second experiment is aimed at answering the main question: is prosody relevant for speech and gesture integration? In this case, we isolated the contribution of prosody from any other parameter that may take part in the integration process (mainly semantic cues), and asked our subjects to give a judgement in a setting similar to the first experiment.

2.1. Materials

2.1.1. Data collection

The stimuli used in both experiments were extracted from audio-visual recordings of dyadic conversations. The content of the conversation was elicited by showing one of the two participants (the ‘narrator’) a short cartoon and the task was to report the plot to the other participant (the ‘listener’). All the participants were native Dutch speakers and each pair was formed by people well acquainted with each other. Only the ‘narrator’ was video recorded, using an almost frontal perspective (see figure 1). The audio was digitally recorded with a general purpose microphone in a sound proof environment. The equipment used for the video recording was a consumer level miniDV camcorder (25 frames per second). In total 7 pairs (2 female and 5 male ‘narrators’, average age 31.14) were recorded for a total of ~ 37 minutes of conversation.

2.1.2. Data preparation

27 utterances were manually selected from the recordings. 21 utterances were selected as containing gestures (*target clips*), while 6 were selected as control clips for the absence of hands movements (*control clips*). The criteria used for the selection were:

- Complete linguistic utterance formed by one or more sentences.
- *Target clips* were selected for the presence of one, maximum two iconic gestures, surrounded by conversation fragments during which the hands of the speaker were more or less static. Similarly, *control clips* were selected for the absence of manual movements before, during, and after the utterance.

The resulting clips had different lengths, ranging from 1625ms to 10025ms, with an average length of ~ 5 seconds, for a total of

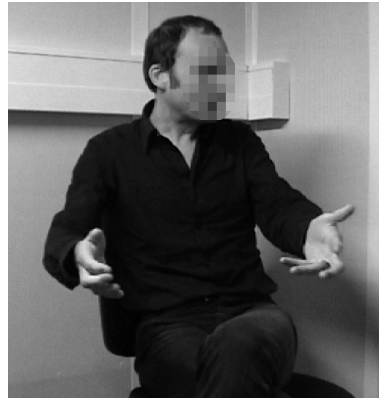


Figure 1: Sample frame from the original recordings. The pixelization is used here to protect the subject’s identity.



Figure 2: Sample frame from the original recordings. The pixelization is used here only to protect the subject’s identity.

140 seconds. Each clip was then clipped so that only the torso of the ‘narrator’ was visible (see figure 2).

Finally, 9 different stimuli clips were generated from each clip by applying a delay to the video channel (in total we had 243 stimuli):

- 9 different delays were employed: -1000ms, -750ms, -500ms, -250ms, 0ms, 250ms, 500ms, 750ms, 1000ms³. The choice of a delay step of 250ms is primarily dictated by the average word length in our data (259ms), a parameter that can be used for a rough but simple sentence segmentation strategy. Notice that in the *target* clips the gesture was still visible in all delayed versions.
- The audio of the original 27 clips was kept constant while the video was shifted to the left of the time axis for negative delays (i.e. video presented before the original time position) or to the right for positive delays (i.e. images presented later than their original position).

The clips were further elaborated on for the second experiment by replacing the original audio with a synthesized version. The synthesized version was meant to isolate the prosodic components of the recorded utterance. This was achieved by automatically extracting the pitch⁴ and the amplitude contour⁵. The results were then manually checked and corrected. The contours obtained with this procedure were then used to synthesize a ‘humming’ version of the utterance. This was done to obtain a copy of the original stimulus, from which all the linguistic information (with the exclusion of suprasegmental features) had been filtered out.

2.2. Experiment A: design

The setting of the experiment was straightforward: each subject was presented with one of the 243 clips and was asked to judge if the clip was synchronized or not. All 243 clips were rated this way. The subjects were told that the alignment of the audio and the video channel had been randomly changed and their task was to determine whether a clip appeared synchronized. They were naive as to the purpose of the experiment and there was no reference to the notion of gesture in the instructions. The sequence of stimuli was randomized to avoid order effects. Each subject completed the experiment in ~ 1 hour with a possible break after about 30 minutes.

We tested 19 subjects (2 males, 17 females, mean age 22.26) for experiment A. The subjects were recruited among bachelor students and were paid €7 for participating. We collected in total 4617 judgments.

2.3. Experiment B: design

In the second experiment the subjects were asked to compare two clips obtained from the same original conversation fragment and determine which one of the two looked more natural. Again, the subjects were naive as to the goal of the experiment, and the notion of gesture was not introduced in the experiment instructions. The delayed clips were initially grouped according to the original recording fragment, the different versions were

³We will use the word *delay* to indicate any measurement of alignment between two channels, using positive values for what is usually meant with *delay* and negative ones in the opposite case.

⁴For the pitch contour only the fragments with glottal pulses were considered, defaulting the value of the extracted contour to 0 in the remaining fragments. This was done in order to maintain the rhythm of the original utterance.

⁵We used the free tool Praat ([14]).

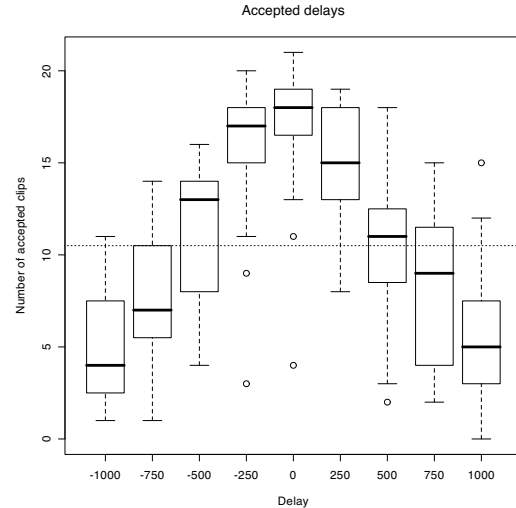


Figure 3: Mean acceptance over the different delays for target clips in Experiment A. The dotted line represents the level of a random response.

then shuffled and finally the order of presentation of the possible pairs was randomized. In other words, the subjects had to select a clip from a pair. Then, after comparing a random number of pairs extracted from other recorded fragments, they were presented again with the previously selected clip. At this point, they had to compare it with a new different clip, obtained from the same original fragment. This in order to avoid stimuli sequence effects. In total, 216 pairs were presented, for a total of 432 clips. On average it took 1 hour to complete the experiment.

We tested 48 subjects (7 males, 41 females, mean age 21.27). Again, the subjects were recruited among bachelor students and were paid €7. We collected a total of 1296 judgments.

3. Results

3.1. Experiment A

The results of the first experiment are plotted in figure 3. The nine delays are plotted on the horizontal axis, while the vertical axis represents the distribution of the average number of accepted clips for each delay, where accepted means that the subject judged the clip synchronized. The scale on the vertical axis goes from 0 (no clip accepted) to 21 (all clips accepted). The shape of the graph clearly shows that there is a significant effect of the specific delay on the number of clips considered as synchronized. A one way analysis of variance (ANOVA in the remainder) confirms it [$F(8, 162) = 23.02; MSE = 15.01; p < 2.2e - 16$]⁶. Figure 4 shows the rate of acceptance for the control clips. In this case we did not find a significant effect of delay on preference ($[F(8, 162) = 1.06; MSE = 2.40; p = 0.3984]$).

We also performed a Tukey HSD pairwise comparison among the different delays (a short summary of the results is

⁶In this and the following analyses we consider a significance level of 0.001. Such a low significance level is justified by the high number of observations and it is further supported by a pairwise comparison among the delays, which shows no significant difference.

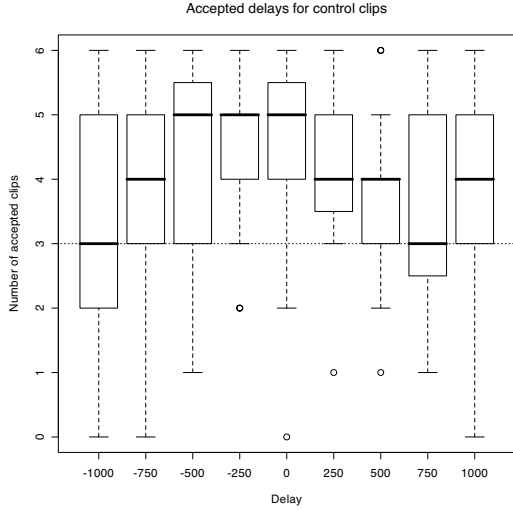


Figure 4: Mean acceptance over delays for control clips in Experiment A. The dotted line represents the shape of a completely random response.

presented in table 1). This analysis shows a symmetric pattern in the response: preferences for inverse delays (e.g. $-250ms$ and $250ms$) are not significantly different. At the same time we can see that a difference between of 250 milliseconds is not significant for the response recorded. On the other hand, differences $\geq 500ms$ are highly significant.

Compared delays	p-value
-1000ms : -750ms	0.4812351
-1000ms : -500ms	0.0001147
-1000ms : -250ms	~ 0
-1000ms : 0ms	~ 0
-750ms : -500ms	0.1484447
-750ms : -250ms	0.0000001
-750ms : 0ms	~ 0
-500ms : -250ms	0.0161005
-500ms : 0ms	0.0009160
-250ms : 0ms	0.9968620

Table 1: Results of a Tukey HSD test comparing the different delays in the first experiment (We present only the comparisons for negative delays as the positive ones present a specular pattern of significance, i.e. comparisons between opposite delays (e.g. $-250ms$ and $250ms$) are non-significant and the remaining ones can be extracted from the table by dropping the minus sign.)

3.2. Experiment B

Figure 5 shows the results for the second experiment. As was true for Experiment A, the preferences peak is centered around the null delay. However, in this case the pattern of preference is much more concentrated around the center, while all the other delays present a similar degree of acceptance below the chance level. Performing an ANOVA to test the effect of the delay on the preference shows again a highly significant effect

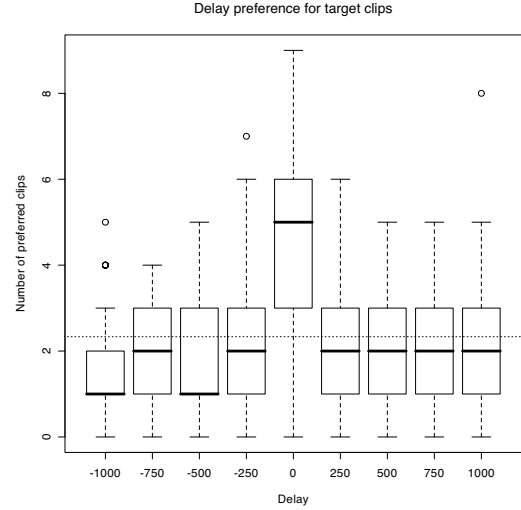


Figure 5: Mean preference over delays for target clips in Experiment B. The dotted line represents the shape of a completely random response.

($[F(8, 423) = 18.21; MSE = 2.16; p < 2.2e - 16]$). However, the result of a Tukey pairwise comparison between the delays is more interesting: the only significant difference can be observed between the null delay and other delays ($p\text{-value} \approx 0$).

As was the case in Experiment A, the results for control clips support the hypothesis that the participants correlated prosodic cues with hand movements: the result of an ANOVA analysis ($[F(8, 423) = 2.59, MSE = 0.645; p = 0.009133]$) shows no significant effect of the delay on preferences (see fig. 6).

The design of the second experiment allowed us to check also for the presence of a learning effect during the test. We considered the number of correct judgments for each trial. A judgment was considered ‘correct’ if the absolute value of the selected delay was smaller or equal to the absolute value of the discarded delay. It is clear from the graph (see figure 7) that there was no learning effect during the experiment.

4. Discussion

While it is a widely accepted fact that speakers tend to produce speech and manual gestures following some pattern of synchronization, the observable parameters determining this intuition are still unexplored. This is a direct consequence of the difficulty of studying large amounts of multimodal data directly, given the lack of automatic techniques for annotating it and the extreme slowness of manual annotation. In this study, we tried to shed some light on this topic by taking a different perspective. Listeners/viewers are sensitive to synchronization/temporal integration and are continuously exposed to multimodal input in everyday life. Therefore, it seems reasonable to hypothesize that they are sensitive to ill-formed multimodal input if the ill-formedness is caused by asynchrony. With the two experiments described here, we attempted to show that listeners/viewers are sensitive to temporal integration in the case of manual gestures and that a parameter they may rely on (among others) is the

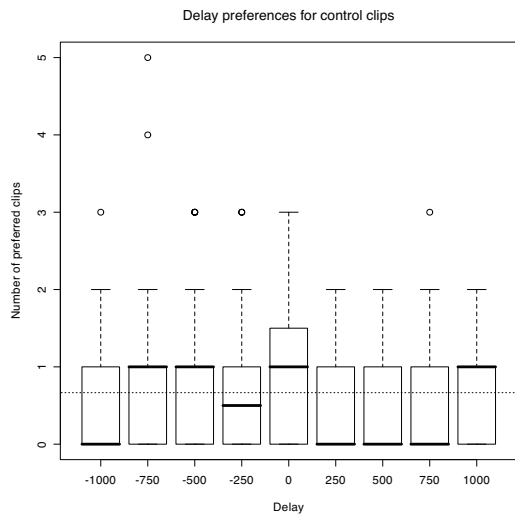


Figure 6: Mean preference over delays for control clips in Experiment B. The dotted line represents the shape of a random response.

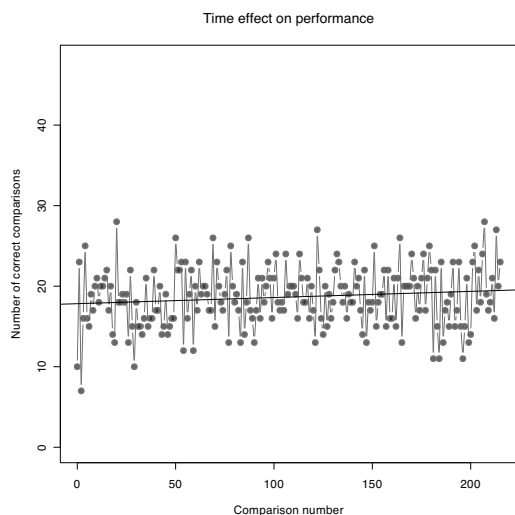


Figure 7: Effect of time on performance in Experiment B. The series of stimuli is represented on the x-axis, while the level of correct answers per stimulus is represented on the y-axis.

prosodic contour of verbal utterances.

Experiment A was designed to actually test the ability of listeners/viewers to temporally integrate manual gestures and verbal utterances. We found that variations in the alignment between the verbal and the gestural channel produce a noticeable change in the judgments of acceptability of our participants. The fact that this effect is not present in the control clips assures us that participants used manual gestures as a parameter to form their judgments. This result is not surprising, and can be well explained as an appropriate perception of patterns that are present in the surface expression of multimodal utterances ([15, 5, 13]). Not surprisingly, the delays that present a preference above the chance level are concentrated around the 0ms delay and span in total 500ms. In general we observed a significant change in the level of preferences only when the difference between two alignments was $\geq 500ms$.

This result is particularly interesting if we compare it with the picture that emerges from Experiment B. First of all, the distributions of preferences in the two experiments have a rather different shape. Even if the mean values and the standard deviations are quite similar (mean=5.65ms, standard deviation=533.815ms in the distribution represented in figure 3, mean=44.39ms, standard deviation=589.612ms in the case of the distribution of figure 5), the results of Experiment B present a much more accentuated peak, and as we saw, only the null delay reaches a level above the chance level.

One possible reason for the difference between the responses in the two experiments can be identified with the different design of the two experiments. Participants of the first experiment could accept more than one delay for each utterance, while in the second experiment they were forced to select only one delay for each original fragment. However, we could interpret this particular difference between the results as suggesting the existence of a sort of ranking among the parameters that determines the felicity of a multimodal utterance. If we accept that the synchronization patterns described in the literature ([5]) are in some way reflected in the perceptual process, then we could explain the difference between the two responses as a higher ranking of semantic synchronicity with respect to prosodic/kinematic alignment. In other words, the 'larger' peak we observe in the case of Experiment A could be caused by a semantically based judgement overriding a prosodically based one: the target gesture is in the vicinity of its semantic correlate (the 500ms 'band' spans over only two words) and thus, even if the gesture is not perfectly aligned from the prosodic point of view, the combination is accepted.

5. Conclusion

The primary goal of this study was to investigate the existence and the nature of a notion of 'well-formedness' in the interaction between speech and manual gestures. We also tried to isolate a parameter that influences the perception of this 'well-formedness'. Our results strongly suggest that prosody may play such a role but they also showed that this particular parameter is not applied in isolation and that its contribution to the final judgement can be masked by other intervening factors (i.e. semantic accommodation).

We believe that this new line of research can be fruitful not only to gain a better understanding of the perception of multimodal input but also to obtain a deeper knowledge of the nature of cross-modal communication, given the difficulties of studying the data directly.

While it would be of primary importance to replicate our

results, we are at the moment planning a similar set of experiments in order to identify the correlate of prosody in the gestural channel.

6. Acknowledgments

Gianluca Giorgolo wishes to thank Michael Moortgat, Bert Le Bruyn and Sander van der Harst for the fruitful discussions before, during and after the experiments. Frans Verstraten was supported by the Netherlands Organisation for Scientific Research (NWO-pionier).

7. References

- [1] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264, 1976.
- [2] Kevin G. Munhall, P. Gribble, L. Sacco, and M. Ward. Temporal constraints on the mcgurk effect. *Perception & Psychophysics*, 58(3):351–362, 1996.
- [3] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson. The mcgurk effect in infants. *Perception & Psychophysics*, 59(3):347–357, 1997.
- [4] George Hollich, Rochelle S. Newman, and Peter W. Jusczyk. Infants’ use of synchronized visual information to separate streams of speech. *Child Development*, 76(3):598–613, 2005.
- [5] David McNeill. *Hand and Mind*. University of Chicago Press, 1992.
- [6] David McNeill. *Gesture and Thought*. University of Chicago Press, November 2005.
- [7] Asli Özyürek, Roel M. Willems, Sotaro Kita, and Peter Hagoort. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, 19(4):605–616, 2007.
- [8] Roel M. Willems, Asli Özyürek, and Peter Hagoort. When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, December 2006.
- [9] S. D. Kelly, C. Kravitz, and M. Hopkins. Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89(1):253–260, 2004.
- [10] Adam Kendon. Do gestures communicate? a review. *Research on Language and Social Interaction*, 27(3):175–200, 1994.
- [11] Geoffrey Beattie and Heather Shovelton. Do iconic gestures really contribute anything to the semantic information conveyed by speech? an experimental investigation. *Semiotica*, 123:201–212, 1999.
- [12] E. Krahmer and M. Swerts. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414, 2007.
- [13] Daniel P. Loehr. *Gesture and Intonation*. PhD thesis, Faculty of the Graduate School of Arts and Sciences of Georgetown University, March 2004.
- [14] Paul Boersma and David Weenink. Praat: doing phonetics by computer (Version 5.0.24) [Computer program]. Retrieved September, 2007, from <http://www.praat.org/>
- [15] Adam Kendon. Some relationships between body motion and speech: An analysis of an example. In A. Siegman and B. Pope, editors, *Studies in Dyadic Communication*. Pergamon Press, New York, 1972.