Pathways of Alignment between Gesture and Speech: Assessing Information Transmission in Multimodal Ensembles

Alexander Mehler and Andy Lücking

Text Technology Lab, Computer Science, Goethe University Frankfurt am Main
{Mehler, Luecking}@em.uni-frankfurt.de

Abstract. We present an empirical account of multimodal ensembles based on Hjelmslev's notion of selection. This is done to get measurable evidence for the existence of speech-and-gesture ensembles. Utilizing information theory, we show that there is an information transmission that makes a gestures' representation technique predictable when merely knowing the part of speech of their lexical affiliate. Thus, there is evidence for a one-way coupling – going from words to gestures – that leads to speech-and-gesture alignment and underlies the constitution of multimodal ensembles.

1 Introduction

We present a information-theoretic measurement procedure of speech-and-gesture alignment. Alignment of the representations of interlocutors is a well-observed phenomenon that has led to the formulation of the *Interactive Alignment Model* (IAM) of [1]. According to this model, interlocutors tend to align (e.g., make similar or functionally complementary) their linguistic representations as their conversation unfolds. Alignment affects the phonetic, the morphological, the lexical, the syntactic, and the semantic level as well as the level of situation models (see [2–5] for examples of respective studies). Recent studies also shed light on the alignment on non-linguistic levels as, for example, on the level of gestures and *Multimodal Ensembles* (MME) (cf. [6, 7]). Although there exist already some approaches to measuring alignment on the lexical and syntactic level [8, 9], we still lack a measurement procedure of multimodal alignment that includes, for example, gesture and speech.

In Section 2 we present such procedure and give an empirical account of speech-and-gesture alignment. The analysis is carried out on a subset of the *Bielefeld Speech and Gesture Alignment Corpus (SaGA)* [10], that has been extended for an explicit annotation of affiliation (see Section 2.1). Finally, Section 3 gives a brief conclusion with an outlook to an encompassing graph model of multimodal alignment in multilogs.

2 Fingerprints of Bimodal Ensembles

In [7] we presented a quantitative account of the interaction of gesture and speech. Basically, we observed only a latent tendency towards the formation of such ensembles. Since in this study no annotations of speech-and-gesture ensembles were available, it provided an *indirect* measurement using the Hartley information [11] by exploring the extension of the set of gestures and of the set of parts of speech that were in use in the corresponding dialogs. That is, [7] did not directly explore the relation R of gesture and speech units as a model of their interaction. As a consequence, the latter study hinted at a latent effect of the formation of MMEs that may also be understood to support neglecting the existence of such ensembles.

In contrast to this study, we explore here a small corpus of seven dialogs in which we manually annotated every gesture together with its links to the speech layer. This gives us a first corpus for exploring the interaction of gesture and speech in terms of information theory [12]. Section 2.1 describes this corpus and its annotation. Section 2.2 uses this corpus to give a quantitative account of the relation R. In contrast to [7], we additionally explore the probabilities of speech and gesture linkage and therefore compute Shannon information instead of Hartley information.

2.1 The Corpus and its Annotation

As an empirical basis for the exploration of MMEs, we use the so called SaGA corpus [10]. The SaGA corpus is built around 25 directions dialogs in which interlocutors combine giving directions and describing sights. The topic of the directions is a route through a Virtual Reality (VR) town model. After having finished a bus ride through the VR town, the bus rider, henceforth called *router*, describes the route and the wayside landmarks to a naïve addressee (*follower*). Each dialog has been recorded on audio and video tapes. Recordings have been annotated for gestures and transcribed for speech. In sum, there are approximately 5,000 iconic and/or deictic gestures and nearly 40,000 words. For more details on the set-up as well as on data annotation and evaluation see [10]. In what follows, we describe the annotation data relevant to our present study.

Gestures are annotated by means of ELAN[1] on three levels, namely for gesture phrases, gesture phases, and gesture practices. On the phrase level, gestures have been identified and, following [13], classified as being *iconic* (depictive gestures), *deictic* (pointing gestures), or both. The gestures have been segmented into *preparation*, *stroke* and *retraction* phases [14] on the gesture phase level. Additionally, holds are recognized. The stroke phase is the core phase which carries semantic information.

According to the gesture practice annotation, each gesture (i.e. stroke) is assigned a *representation technique* [15]. For SaGA annotation, we compiled the following set of representation techniques out of the proposals found in the literature [15, 6, 16]:

[1] http://tla.mpi.nl/tools/tla-tools/elan.

gesture layer X :	g_{i_1}	g_{i_2}	g_{i_3}	...	g_{i_t}
linguistic layer Y :	l_{j_1}	l_{j_2}	l_{j_3}	...	l_{j_t}
time T :	1	2	3	...	t

Fig. 1. The unfolding of gesture and speech mapped by two stochastic processes.

- *Indexing*: pointing to a location of the gesture space;
- *Placing*: as if an object is put in a certain location of the gesture space;
- *Shaping*: as if an object's shape is sculptured;
- *Drawing*: as if an object's outline is traced;
- *Modeling*: the hand(s) is (are) a proxy for an object;
- *Sizing*: as if hands or fingers indicate a certain distance or size;
- *Counting*: enumerating things with the fingers;
- *Hedging*: expressing uncertainty (for instance, by hand wiggling).
- *Unclear*: if no representation technique can be assigned.

Since single gestures might perform more than just one representation technique, it is also allowed to combine the respective annotation predicates.

In addition to the above-mentioned annotations, subsets of seven dialogs, namely dialogs $\mathcal{D}_2, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_{24}$, and $\mathcal{D}_{25}$, have also been annotated for *affiliation* and *parts of speech* (POS). A co-verbal gesture is related to one or more words from its accompanying speech – the gesture's *affiliate*. Affiliation is made explicit by using an in-house tool that allows one to combine annotation elements (viz., words and gestures) from different layers into newly created cross-layer elements. These new elements are called MMEs (*Multimodal Ensembles*). In those cases where a gesture is affiliated with more than one word, the syntactic type of this complex affiliate is specified. In addition to the conventional phrase types *S*, *VP*, *NP*, *NOM*, *AP* (adjective phrase), *ADVP* (adverbial phrase), and *PP*, we also recognize discontinuous verb phrases (*VP-disc*) and constructions made up of prepositions and adverbs (*P+ADV-CX*), as needed for frequently-occurring "[drive] towards [...] to" directions. If the affiliation to multiple words is due to a self-correction or repair, the annotation value *REPAIR* is used. Altogether 1196 MMEs have been annotated, out of which 225 have a complex affiliate (that means that 971 MMEs contain only lexical affiliates).

In a follow-up annotation, it will be specified on top of the affiliation annotation which speech-gesture pairs re-occur in the same or in a derived (e.g., simplified) form. An uptake of speech-gesture pairs is presumably the strongest directly-observable evidence for their status as ensembles. In this way, we will be able to study the dynamics of speech and gesture chaining within direction-giving dialogs.

2.2 On the Informational Uncertainty of Speech-Gesture Selection

In order to shed light on the existence of bimodal speech-and-gesture ensembles with the help of the corpus data of the latter section, we perform a quantitative analysis of the interaction of their constituents based on information theory

Table 1. Speaker-dependent information transmission of *gesture types* and *phrase types*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	21	12	10	6	.941	.861	.384	.145	.593	.716
$\mathcal{D}_4$	45	49	22	10	.821	.828	.555	.472	.672	.356
$\mathcal{D}_5$	49	33	14	10	.787	.853	.494	.516	.661	.337
$\mathcal{D}_6$	39	14	11	6	.879	.778	.468	.229	.601	.550
$\mathcal{D}_7$	20	12	6	6	.673	.897	.367	.591	.632	.306
$\mathcal{D}_{24}$	30	26	16	8	.903	.838	.506	.309	.648	.529
$\mathcal{D}_{25}$	21	14	8	5	.920	.957	.482	.391	.689	.566

Table 2. Speaker-dependent information transmission of *gesture types* and *parts of speech*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	34	39	15	11	.894	.805	.549	.416	.669	.390
$\mathcal{D}_4$	51	119	34	13	.769	.764	.634	.579	.689	.186
$\mathcal{D}_5$	55	119	37	13	.646	.707	.520	.529	.597	.178
$\mathcal{D}_6$	49	84	31	11	.756	.769	.579	.516	.657	.253
$\mathcal{D}_7$	30	43	12	10	.645	.849	.488	.679	.662	.169
$\mathcal{D}_{24}$	36	82	22	12	.803	.801	.632	.589	.708	.212
$\mathcal{D}_{25}$	31	45	16	9	.852	.772	.629	.490	.692	.282

[12]. This is done by means of Hjelmslev's notion of *selection* [17]. Hjelmslev calls an item c a constant of what he calls a function between two items c and y if the presence of c is a necessary condition for the presence of y in this function. Conversely, an item v is called a *variable* if it is not a necessary condition for the presence of its counterpart in the function. Henceforth, we speak of (syntagmatic) co-occurrences instead of functions in order to prevent any confusion with the mathematical notion of a function. In this sense, a constant c is said to be the *selecting* and a variable v the *selected* constituent of a co-occurrence. Starting from the binarism of constants and variables, Hjelmslev distinguishes three syntagmatic relations:[2] selections $c \leftarrow v$ hold between a constant and a variable, solidarities $c_m \leftrightarrow c_n$ link two constants and combinations $v_i \updownarrow v_j$ connect two variables. Based on sets of such co-occurrences, Hjelmslev distinguishes items that always enter into selections (as constants, as variables, or as both constants and variables), into solidarities or into combinations. Obviously, the underlying logic of these notions is too coarse-grained to characterize the interaction of gesture and speech appropriately. A gesture may enter, for example, into different ensembles so that the corresponding class of solidarities is probably always empty: there is no one-to-one mapping between gesture and speech units. Nevertheless, this approach gives rise to a generalization in the framework of information theory by means of which we can distinguish between selections and solidarities in a more flexible way that accounts for probabilities of co-occurrences.

[2] For their paradigmatic counterparts see [17].

Table 3. Speaker-dependent information transmission of *gesture types* and *parts of speech and phrase types.*

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	40	51	15	17	.898	.815	.513	.447	.667	.384
$\mathcal{D}_4$	61	168	34	23	.764	.771	.608	.595	.685	.176
$\mathcal{D}_5$	65	152	37	23	.643	.691	.503	.529	.590	.161
$\mathcal{D}_6$	55	98	31	17	.756	.727	.563	.492	.637	.234
$\mathcal{D}_7$	36	55	12	16	.631	.834	.457	.677	.655	.174
$\mathcal{D}_{24}$	44	108	22	20	.812	.792	.606	.579	.697	.213
$\mathcal{D}_{25}$	36	59	16	14	.842	.789	.591	.525	.687	.264

Table 4. Speaker-dependent information transmission of *gesture types* and *parts of speech or phrase types.*

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	40	37	15	15	.873	.835	.426	.389	.631	.447
$\mathcal{D}_4$	61	109	34	20	.763	.763	.538	.498	.641	.265
$\mathcal{D}_5$	65	130	37	20	.639	.674	.477	.478	.566	.196
$\mathcal{D}_6$	55	88	31	17	.745	.739	.520	.466	.619	.272
$\mathcal{D}_7$	36	47	12	15	.669	.809	.478	.634	.650	.191
$\mathcal{D}_{24}$	44	73	22	20	.765	.778	.493	.497	.633	.280
$\mathcal{D}_{25}$	36	43	16	11	.852	.812	.538	.449	.665	.363

Look at Figure 1, which shows two random variables that emit gesture (X) and speech units (Y) in the course of time (T) of a dialog. Now, suppose that gestures only occur as constituents of perfectly distinguishable ensembles in which they interact with items of a certain layer of linguistic resolution (e.g., of lexical or phrasal units). In line with this extremum we can think of a dialog lexicon that consists of ensembles as a sort of bimodal *solidarities* where the gestural constituent always *selects* its linguistic counterpart and vice versa. Conversely, we may think of an extremum where any gesture co-occurs with any item of the linguistic layer under consideration and vice versa. Obviously, this corresponds to a combination of gesture and speech in terms of Hjelmslev. As said before it is unlikely that these extrema are observable in reality. Rather, we expect intermediary states according to various mixtures of selecting and selected items. However, if we assume that bimodal ensembles are generated in the course of a dialog then this should be reflected by a considerable amount of *information transmission* between the gesture layer and its linguistic counterpart. In order to measure this sort of interaction, we compute several functionals based on Shannon's entropy [18]. More specifically, for a given dialog $\mathcal{D}$ we span the bipartite graph

$$G(\mathcal{D}) = (X \cup Y, E),\ E \subseteq [X \cup Y]^2$$

that is partitioned into the set of gesture types X and the set of linguistic types Y.[3] To derive probability distributions from $G(\mathcal{D})$, we define the relation $R = X' \times Y'$ together with a joint probability distribution P over R such that for any $r = (x, y) \in R$, $P(r)$ is the probability of observing co-occurrences of x

[3] Note that since $G(\mathcal{D})$ is a bipartite graph, there is no monomodal edge in E.

Table 5. Speaker-dependent information transmission of *gestures* and *words*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	302	69	15	50	.894	.961	.152	.448	.630	.742
$\mathcal{D}_4$	361	261	34	130	.769	.889	.274	.530	.630	.496
$\mathcal{D}_5$	525	356	37	187	.646	.923	.227	.633	.639	.419
$\mathcal{D}_6$	430	171	31	96	.756	.912	.210	.501	.611	.546
$\mathcal{D}_7$	317	113	12	80	.645	.930	.175	.663	.657	.470
$\mathcal{D}_{24}$	307	150	22	92	.803	.924	.214	.521	.636	.589
$\mathcal{D}_{25}$	241	91	16	69	.852	.941	.133	.470	.621	.719

and y in $\mathcal{D}$ based on our annotations (see Section 2.1). $X' \cup Y' \subseteq X \cup Y$ is the subset of non-isolated vertices in $G(\mathcal{D})$. That is, we ignore any gesture or speech unit that is isolated in $G(\mathcal{D})$. One reason is that there are likely many more word types that are not affiliated with any gesture, than gesture types that are not affiliated with any word. Based on these preliminaries we can compute several entropies together with the corresponding information transmission [11]:

- the normalized joint entropy $\hat{S}(X' \times Y')$ of the joint probability distribution over $X' \times Y'$:

$$\hat{S}(X' \times Y') = \frac{S(X' \times Y')}{\log_2(|X'| \times |Y'|)} = \frac{\sum\limits_{(x,y) \in X' \times Y'} p(x,y) \log_2 p(x,y)}{\log_2(|X'| \times |Y'|)} \in [0,1] \quad (1)$$

- the normalized simple entropies $\hat{S}(X')$, $\hat{S}(Y')$ based on the marginal probability distributions derived from the joint probability distribution over R;
- the normalized conditional entropies $\hat{S}(X'|Y')$ and $\hat{S}(Y'|X')$;
- and the normalized information transmission

$$\hat{T}_S(X', Y') = \frac{S(X') + S(Y') - S(X' \times Y')}{\min\{\log_2 |X'|, \log_2 |Y'|\}} \in [0,1] \quad (2)$$

Note that we normalize all entropies to the unit interval as described in [11] to provide comparability between dialogs. Note also that we compute these functionals for different linguistic layers Y'. More specifically, we compute the information transmission between gestures on the one hand and parts of speech, phrase types and word types on the other. Further, we consider two variants of combining parts of speech and phrase types: in one case, we consider edges between gestures on the one hand and parts of speech and phrase types on the other, whenever they are annotated. In the other case we ignore links to parts of speech if the gesture is affiliated with a phrase. The reason for distinguishing these cases is that we often observed gestures that are affiliated with items on the level of preterminals, while only few gestures are affiliated with complex phrases. Finally, we distinguish between interlocutors (in the role of the *router*, i.e., direction-giver (see Section 2.1), and the *follower*) such that entropies are computed intra- as well as interpersonally.

According to our hypothesis we expect the existence of bimodal ensembles. This hypothesis is connected with three expectations regarding the relation of entropy measurement and Hjelmslevian functions as considered here:

– Firstly, an extensive set of solidarities should result in high rates of information transmission while an extensive set of combinations should be reflected by low transmission rates.
– Secondly, a tendency towards a selection of gestures by linguistic units should be reflected by a low conditional entropy $\hat{S}(X'|Y')$.
– Thirdly, a tendency towards a selection of linguistic units by gestures should be reflected by a low conditional entropy $\hat{S}(Y'|X')$.

Thus, entropies that are in support of our hypothesis are low values of both conditional entropies that coincide with a high transmission rate.[4] Tables 1–10 report the corresponding results of our measurements.

2.3 Discussion

Looking at Table 1, we observe a remarkably high rate of information transmission in the case of $\mathcal{D}_2$. Even in the case of $\mathcal{D}_6$, $\mathcal{D}_{24}$ and $\mathcal{D}_{25}$ we get a transmission rate slightly above the transmission equilibrium 0.5 (last column). Looking at Table 2 and 3, we observe much lower transmission rates. Obviously, there is a higher degree of solidarity between gestures and phrases than between gestures and parts of speech. In any event, in all these tables we additionally observe conditional entropies around the entropic equilibrium of 0.5. Note that a strong tendency to selection would be reflected by a corresponding conditional entropy near to zero. Thus, we can conclude that although there is a latent tendency to mutual selection (in the sense of a Hjelmslevian solidarity) between gesture and phrase types, the constituent selections are only weakly developed. The same is true for Table 4, although there we observe an increase of the transmission rate as expected by the way these graphs have been spanned. Highest rates of transmission are observable for the affiliation of gestures and words (see Table 5). The maximum rate of transmission is reached by $\mathcal{D}_2$. Further, if we look at the conditional entropies in Table 5, we see a strong tendency to the selection of gestures by words (column $\hat{S}(X'|Y')$), while the opposite entropy $\hat{S}(Y'|X')$ indicates a much weaker tendency to the selection of words by gestures. Obviously, knowing a word that affiliates with a gesture, tells you much about this gesture. That is, words *select* gestures in the sense of Hjelmslev. Since this finding goes together with a latent selection of words by gestures, the final rate of information transmission is the highest among all cases of gesture and speech affiliation studied here. Obviously, this observation is in line with our expectation about the existence of bimodal speech-and-gesture ensembles. In any event, we can now quantify this expectation.

If we finally look at the same range of speech-and-gesture ensembles, but now in a speaker-*independent* manner (see Table 6–10), we observe in all five modes in almost all cases a decreasing rate of information transmission together with an increasing conditional entropy $\hat{S}(Y'|X')$ that reflects a decreasing tendency to the selection of speech units by gestures. This hints at the formation of

[4] Note that we do not provide a significance test. A corresponding random model of the affiliation of gesture and speech units will be the task of future work.

Table 6. Speaker-independent information transmission of *gesture types* and *phrase types*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	18	11	8	6	.943	.861	.383	.212	.604	.649
$\mathcal{D}_4$	34	43	16	10	.799	.828	.553	.533	.678	.295
$\mathcal{D}_5$	35	29	10	10	.797	.853	.500	.555	.676	.297
$\mathcal{D}_6$	30	14	11	6	.879	.778	.468	.229	.601	.550
$\mathcal{D}_7$	20	12	6	6	.673	.897	.367	.591	.632	.306
$\mathcal{D}_{24}$	23	23	12	8	.894	.838	.499	.366	.653	.472
$\mathcal{D}_{25}$	18	14	7	5	.943	.957	.515	.440	.715	.516

Table 7. Speaker-independent information transmission of *gesture types* and *parts of speech*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_{24}$	29	71	15	12	.786	.801	.629	.631	.712	.171
$\mathcal{D}_{25}$	28	40	13	9	.815	.772	.602	.524	.681	.248
$\mathcal{D}_2$	31	34	12	11	.860	.805	.521	.454	.661	.351
$\mathcal{D}_4$	40	98	23	13	.747	.764	.624	.614	.687	.151
$\mathcal{D}_5$	41	93	24	13	.637	.707	.511	.551	.598	.156
$\mathcal{D}_6$	40	71	23	11	.747	.769	.580	.552	.662	.218
$\mathcal{D}_7$	30	43	12	10	.645	.849	.488	.679	.662	.169

speech-and-gesture ensembles that are more developed in the case of intrapersonal alignment than in the case of interpersonal alignment.

Based on our measurements presented so far we conclude that there is a latent tendency to the formation of speech-and gesture ensembles that can be best observed on the level of words: there is a stronger tendency to the selection of gestures by words than the other way round. Moreover, this tendency goes beyond what has been reported in [7] so that we conclude that entropy measures are more appropriate for this sort of measurement.

3 Conclusion

We presented an empirical procedure for assessing the influence of multimodal ensembles in dialog that utilizes information theory to compute the information transmission between their constituents of different mode. More specifically, we used the Hjelmslevian notion of selection to show that higher rates of transmission concern the mutual selection of words and gestures as part of bimodal ensembles. We found that there is a detectable degree of informativity between the verbal and the gestural modality, *even if we only consider parts of speech and representation techniques*. Note that the tranmission relation is directed: it runs from words to gestures (and not *vice versa*). This finding is in line with the *primacy of language* in multimodal communication [19].

In order to map the complex dynamics of multimodal conversations in an encompassing model, the graph model of alignment in dialogs [9] and in multilogs [20] has to be extended in order to additionally map gestural alignment as a further reference point of semiotic alignment. This gives rise to a more flexible

Table 8. Speaker-independent information transmission of *gesture types* and *parts of speech and phrase types.*

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	37	45	12	17	.866	.815	.489	.484	.663	.377
$\mathcal{D}_4$	50	141	23	23	.744	.771	.600	.628	.686	.144
$\mathcal{D}_5$	51	122	24	23	.635	.691	.495	.549	.592	.141
$\mathcal{D}_6$	46	85	23	17	.75	.727	.566	.523	.642	.204
$\mathcal{D}_7$	36	55	12	16	.631	.834	.457	.677	.655	.174
$\mathcal{D}_{24}$	37	94	15	20	.799	.792	.604	.616	.703	.195
$\mathcal{D}_{25}$	33	54	13	14	.811	.789	.572	.557	.682	.239

Table 9. Speaker-independent information transmission of *gesture types* and *parts of speech or phrase types.*

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	37	33	12	15	.837	.835	.392	.427	.623	.445
$\mathcal{D}_4$	50	93	23	20	.726	.763	.509	.535	.633	.228
$\mathcal{D}_5$	51	104	24	20	.63	.674	.465	.498	.566	.175
$\mathcal{D}_6$	46	76	23	17	.731	.739	.517	.502	.622	.236
$\mathcal{D}_7$	36	47	12	15	.669	.809	.478	.634	.650	.191
$\mathcal{D}_{24}$	37	65	15	20	.736	.778	.475	.542	.634	.261
$\mathcal{D}_{25}$	33	39	13	11	.801	.812	.499	.488	.650	.323

graph model of multimodal ensembles that maps interpersonal ensembles, the variability of their manifestations and the laws of their clustering in the course of conversations.

Acknowledgement

Support by the LOEWE Priority Program *Digital Humanities* at Frankfurt University and by the Special Research Center *Alignment in Communication* at Bielefeld University is gratefully acknowledged.

References

1. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences **27** (2004) 169–226
2. Giles, H., Powesland, P.F.: Speech Styles and Social Evaluation. Academic Press, London ; New York (1975)
3. Branigan, H.P., Pickering, M.J., Stewart, A.J., McLean, J.F.: Syntactic priming in spoken production: Linguistic and temporal interference. Memory & Cognition **28**(8) (2000) 1297–1302
4. Garrod, S., Anderson, A.: Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. Cognition **27**(2) (1987) 181–218
5. Watson, M.E., Pickering, M.J., Branigan, H.P.: An empirical investigation into spatial reference frame taxonomy using dialogue. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society. (2006) 2353–2358
6. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, MA (2004)

Table 10. Speaker-independent information transmission of *gestures* and *words*.

$G(\mathcal{D})$	$\lvert X \cup Y\rvert$	$\lvert E\rvert$	$\lvert X'\rvert$	$\lvert Y'\rvert$	$\hat{S}(X')$	$\hat{S}(Y')$	$\hat{S}(X'\lvert Y')$	$\hat{S}(Y'\lvert X')$	$\hat{S}(X' \times Y')$	$\hat{T}_S(X', Y')$
$\mathcal{D}_2$	299	67	12	50	.860	.961	.152	.512	.647	.707
$\mathcal{D}_4$	350	242	23	130	.747	.889	.258	.574	.642	.489
$\mathcal{D}_5$	511	330	24	187	.637	.923	.227	.674	.660	.409
$\mathcal{D}_6$	421	163	23	96	.747	.912	.207	.541	.625	.540
$\mathcal{D}_7$	317	113	12	80	.645	.930	.175	.663	.657	.470
$\mathcal{D}_{24}$	300	139	15	92	.786	.924	.189	.566	.649	.597
$\mathcal{D}_{25}$	238	87	13	69	.815	.941	.117	.518	.630	.697

7. Lücking, A., Mehler, A., Menke, P.: Taking fingerprints of speech-and-gesture ensembles: Approaching empirical evidence of intrapersonal alignment in multimodal communication. In: LONDIAL 2008: Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL), King's College London (2008) 157–164
8. Reitter, D., Moore, J.D., Keller, F.: Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society. CogSci'06 (2006) 685–690
9. Mehler, A., Lücking, A., Weiß, P.: A network model of interpersonal alignment. Entropy **12**(6) (2010) 1440–1483
10. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Malta, 7th International Conference for Language Resources and Evaluation (LREC 2010) (2010) 92–98
11. Klir, G.J.: Foundatins of Generalized Information Theory. Wiley, Hoboken (2006)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, Hoboken (2006)
13. McNeill, D.: Hand and Mind—What Gestures Reveal about Thought. Chicago University Press, Chicago (1992)
14. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In Key, M.R., ed.: The Relationship of Verbal and Nonverbal Communication. Volume 25 of Contributions to the Sociology of Language. Mouton Publishers, The Hague (1980) 207–227
15. Müller, C.: Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich. Volume 1 of Körper – Kultur – Kommunikation. Berlin Verlag, Berlin (1998)
16. Streeck, J.: Gesture Craft: The Manu-Facture of Meaning. Volume 2 of Gesture Studies. John Benjamins (2009)
17. Hjelmslev, L.: Prolegomena to a Theory of Language. University of Wisconsin Press, Madison (1969)
18. Shannon, C.E.: A mathematical theory of communication. The Bell Systems Technical Journal **27**(3) (1948) 379–423
19. de Ruiter, J.P.: On the primacy of language in multimodal communication. In: Workshop Proceedings on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces, Lisbon (2004) 38–41
20. Mehler, A., Lücking, A.: A graph model of alignment in multilog. In: Proceedings of IEEE Africon 2011. IEEE Africon, Zambia, IEEE (2011)