On speaker change

Jens Allwood. Elisabeth Ahlsén
SCCIIL Interdisciplinary Center
University of Gothenburg

Abstract

There are many challenges for automatic recognition of content in multimodal face-to-face communication. In this paper, we focus on difficulties related to recognition of speaker change, i.e. who is the main speaker, when does another speaker get the floor and what happens in transitions between speakers? Phenomena such as overlaps (for giving feedback or other reasons) between speakers, laughter, short and very quiet utterances are studied in a discussion between four persons. The role of overlapping activities, for example serving and drinking coffee and eating cookies during a discussion, and the phenomenon of co-construction of content are discussed. The analysis of visual information, like eye gaze direction, gestures and actions, is also relevant for recognition of speaker changes. Taken together, these types of phenomena are very frequent and add to the difficulty in identifying speaker changes. In addition, the challenges connected with an analysis of the acoustic signal, suggest that "change of speaker" should perhaps better, from a multimodal point of view, be seen as "change of main communicator".

Thus, there are many challenges for automatic speaker identification as well as for automatic speech recognition. This paper provides a descriptive case study of an interaction with four participants, with the purpose of pointing to some of the relevant phenomena involved in speaker change. This could, hopefully, in turn, provide input to further attempts of capturing how such phenomena through improvement of tools for automatic speaker identification and speech recognition.

1. Introduction – Points of departure

Both humans and machines that interact with humans need to be able to discern transitions between different speakers, in order to identify who contributes what information when. Below, we will discuss this challenge in relation to an informal discussion between four participants.

1.1 Main communicator and types of contributions

The multimodal nature of human face-to-face communication makes simultaneous multidimensional collective information processing possible, involving several persons making contributions to a jointly constructed and shared content. However, continuous, totally overlapping communication from all contributors, focusing equally on production and interpretation, would be too difficult to process, given the limits of human attention span and capacity of processing. So humans have evolved procedures and mechanisms for change of main communicator and for sequential, to some extent simultaneous, creation of shared content by several contributors. Since all participants in this kind of face-to-face discussion, are "communicators" participating in a mutual multimodal flow of information and communication, we will use the expression "main communicator",

rather than “main speaker” for the person who is mainly holding the floor. The function of this terminological change is to recognize that other persons are also communicators and that contributions can occur in other modalities than speech.

Our analysis distinguishes three types of contributions:
- Utterance = unimodal or mainly vocal verbal contribution
- Gestural contribution = unimodal gestural contribution
- Multimodal contribution = vocal verbal + gestural contribution

In the case of “multimodal contributions”, we are dealing with “noticeable multimodality”, since it is very had to observe all the small body movements accompanying vocal verbal communication. (We are not using the term “turn”, since it is unclear what this term means, e.g. does it mean the right to speak, “holding the floor” or is a “turn” any utterance that is produced? The latter interpretation seems unlikely since it seems consistent with most interpretations of “turn” that utterances can be produced by one person while another has the “turn”. Our term “contribution” includes all utterances produced as well as gestural contributions, while “main communicator” denotes the person who mainly holds the floor. If there is competition between communicators, there is no clear main communicator.

1.2 Simultaneous multilateral flow of information

In face-to-face communication, all spoken contributions are really multimodal, cf. the McGurk effect (McGurk and MacDonald (1976)) which shows that facial movements of the mouth affect the perception of spoken words. There are, however, gestural expressions that are unimodal, since we can gesture without talking.

In face-to-face communication, all communicators are also potentially aware of each other and therefore potentially aware of what everyone else is contributing to communication. We write “potentially”, since attention is variable and people are not always aware of what is happening around them, even if they, in principle, could be. While A speaks, A sees B’s facial gestures, which B can produce while listening to A. Both A and B are therefore simultaneously producers and recipients in communication. However, limitations of attention, processing etc. have the effect that mostly one person at a time is the main communicator, while the other(s) are mainly recipients. What we investigate in this paper is how the transitions between contributions and main communicators are managed, and we address the questions: What are the procedures and mechanisms for transitions between main communicator and between contributions? How do we know when someone is letting go of the floor? How do we know how to fit our contribution into an on-going dialog? What does a virtual agent or robot need to know, to have the same ability? (Cf. Sacks, Schegloff and Jefferson 1974, Ahlsén 2012)?

1.3 Functional relevance and responsiveness of contributors

We start with the following theoretical assumptions. The first assumption is that for hearing communicators, speech is mostly the primary means of intentional “signaled” communication (cf. Allwood 2001) and gestures are auxiliary to speech. Sometimes contributions are balanced multimodally, often speech is dominant, sometimes

unimodal vocal and sometimes there is no speech and gesture takes over as the primary means (unimodal gestural). The second assumption is that contributions to communication are often or mostly made to try to achieve the purpose(s) of the activity for which the communication is instrumental. This, in turn, since human activities are interactive, coordinated, often collaborative and sometimes cooperative, has as a consequence that the contributions by the communicator should be (and mostly are) functionally adequate and relevant to contributions made by other communicators (especially the preceding main contribution) in the joint activity to which the contributions are made (Allwood, 1976, 2000) (see example 1). We distinguish at least three types of functional relevance or responsiveness. Contributions can be and mostly are functionally relevant or responsive in the following three ways:

1) To the preceding contribution
2) To the current exchange type
3) To the local sub-activity and global activity

Example 1.

A (Shop clerk): What do you like?
B (Customer): A bar of chocolate

In the example, B's contribution is relevant and adequate in relation to
- the preceding contribution, which is a question evoking an answer (see figure 1 below)
- the exchange type of question-answer
- the global activity purpose of shopping

Figure 1 shows the functional dependence between contributions

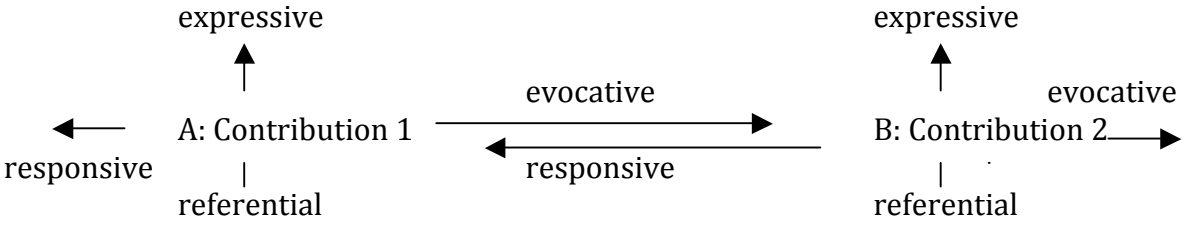

Figure 1. Functional dependence between contributions

The initial arrow pointing backwards from A's Contribution 1 indicates that the contribution has a responsive orientation towards the contribution(s) and context preceding this contribution. The final arrow, pointing forward from B's Contribution 2, indicates that the contribution has an evocative aspect with respect to the continuation of the interaction. Every contribution, thus, has both a responsive and an evocative orientation. In addition, the content of a contribution usually also has a referential orientation, concerning the topic talked about or referred to, as well as a more expressive orientation, concerning the affective/epistemic attitudes that are being expressed in the contribution.

A person can contribute information to interactive communication in three main ways. The first way is by making a contribution to the current main topic of communication. The second way is by giving feedback to contributions to the current main topic and the third way is through some other type of informative action, which is relevant for or parallel to the current main topic of the conversation.

1.4 Types of content

We have also distinguished five types of content (cf. Allwood 2002), that are common to vocal, gestural and multimodal contributions, namely

(i) Identity (information concerning the communicator's identity)
(ii) Physiological states, such as hunger, thirst, fatigue, alertness
(iii) Affective, epistemic and conative (will, intention) states
(iv) Communicative act function and factual information
(v) Communication management (CM) (Interactive CM, Own CM).

The types of content can influence each other, e.g. reinforce each other, and are therefore not mutually exclusive. All the types of content can in principle be expressed unimodally (vocal-verbally or gesturally) or multimodally, but factual information and communicative act function are mainly expressed by vocal verbal means, while physiological states as well as affective, epistemic and conative states are mainly expressed prosodically, gesturally or multimodally. Identity and communication management are expressed in all three ways (cf. Allwood, 2002, 2008).

2. Change of main communicator – humans and machines

2.1 Top-down and bottom-up

All communication utilizes a combination of top-down and bottom-up information. Top-down information influences communication through the assumptions and expectations participants have initially or build up as a result of the interaction. These assumptions and expectations concern all the requirements mentioned above: (i) evocative functions and response requirements of the preceding utterance, role and exchange type and global activity purpose. Bottom-up information is all information that is successively contributed to the interaction through the unimodal and multimodal contributions that are made.

In general, participants try to match bottom-up information with existing top-down assumptions and expectations, both in interpreting of what is being contributed and in producing new contributions. As an aid in this process, different types of feedback are used, both internally, in the communicator, and externally to other participants.

Let us now start our investigation of change of main communicator by considering the question: Why are contributions ended and how do we know they are finished? A possible answer is that contributions are finished when activity ceases and they have a "relevant completeness" or cannot for some (other) reason be continued. The requirements on "relevant completeness" can be described as follows: a relevantly complete contribution (i) has met the response requirements of the preceding

contribution's evocative function, (ii) has met the requirement s connected with its role in the exchange type, or (iii) is contributing an action relevant for the global activity purpose.

In example 1 above, B's contribution meets all of the above requirements and is, for this reason, a relevantly complete contribution that can be followed by a new contribution in a new exchange type relevant to the same local or global purpose (thus conforming to top-down expectations).

If we turn to machine based social signal processing, transitions between contributions are often important to recognize, e.g. in human-machine interaction or summarization of arguments in meetings. In both types, the machine must be able to recognize (and in the case of human-machine interaction also make) relevantly complete contributions and in both types of tasks and many others, a combination of top-down and bottom-up processes are needed. (Cf. Allwood and Ahlsén 2012, Salamin and Vinciarelli 2012, Vinciarelli 2009, Valente and Vinciarelli 2007.)

If a machine is to be able to participate smoothly in communication, it has to be able to interpret (i.e. discriminate, identify, recognize and understand) contributions. This means activation of top-down processes and behavior relevant to (i) the evocative function of the preceding utterance, (ii) the function in the exchange type, and (iii) the relation to the global activity purpose and the activity role of the communicator. For example, it should recognize when a contribution could end, when it actually ends and when it ought to end. The top-down activation should be combined with a bottom-up processing of gesture units, syntactic-semantic, phonetic – prosodic units and coordination of these three types of processing (Allwood 2000).

2.2 Overlaps

There are many difficulties for artificial recognition and production of naturalistic "face-to-face" dialog, for instance, in our data, as many as 20-40% of all transitions between contributions occur with overlaps, where more than one participant contributes simultaneously. "Overlaps" include all types of simultaneous communicative behavior from more than one person at the same time (vocal-verbal as well as body communication, contributions when having and not having the floor). Perhaps, the most common function of overlap is to give positive feedback with contributions, like *yes*, *m*, nods etc. Simultaneous contributions from several communicators are hard to analyze for humans and even harder for machines and make it difficult to decide when a change of speaker/contributor really takes place.

Taking multimodality as defined above into account, for two communicators, overlaps can, in fact, be of 16 possible types and for three communicators, overlaps can be of 32 types and so on, see table 1.

Table 1. Possible types of overlaps for two multimodal communicators

Communicator 1	Vocal verbal Communicator 2	Gestural Communicator 2	Multimodal Communicator 2	Other informative action Communicator 2
Vocal verbal				
Gestural				
Multimodal				
Other informative action				

2.3 Other difficulties

Other difficulties are related to the fact that face-to-face communication is full of quick, small contributions, which are sometimes of quite low intensity. Secondly, there is a large amount of individual and situational variation in the phonetic, gestural, syntactic and semantic properties of contributions and, thirdly, it is hard to know what features of interaction are relevant for communication. Therefore, a relevant question is how we can perceive, recognize and understand contributions and, more specifically, recognize when overlaps and small contributions are connected with a change of the main communicator and when they constitute contributions not connected with a change of the main communicator. An extra difficulty for artificial recognition and production of naturalistic "face-to-face" dialog is added if the analysis is limited to only auditory-acoustic properties, since in this case, the analysis neither has access to multimodal (mainly gestural) communication, nor does it have access to other non-acoustic informative action.

3. Analysis of a discussion between four girls

3.1 Empirical material

In the following, we will use a video-recorded discussion between four girls to exemplify some of the difficulties concerning recognition and production of transitions between the main contributors in human face-to-face communication.

The video-recording takes place in a studio, where four female students are seated in a sofa around a table, being filmed by one camera. They are given a topic for discussion by a researcher, who then leaves the room. Coffee and cookies are available on the table.

The 100 first contributions and the transitions between them were analyzed. The analysis was based on the video-recordings and on transcriptions using the GTS (Gothenburg Transcription System) (Allwood 1999, 2008) and MSO (Modified Standard Orthography) for spoken Swedish (Nivre 1999, 2004).

Figure 2. Four girls discussing (From left to right: M, A, C, K)

Example 2. Transcription of the opening of the discussion of the four girls
(B is the researcher)

B: då får ni dricka kaffe å ha de{t} / trevli{g}t < // >
B: then you can drink coffee and have a nice time < // >
<all look at B and laugh>
K: ja ska ni kolla på oss därigenom då < elle{r} >
K: yes will you look at us through there then < or >
<gesture: pointing towards the controlroom>
B: < ja vi ha{r} dratt för så vi ser inget >
B: <yes we have pulled the curtain so we don't see anything>
<event: paper rustle>
K: m < >
K. m < >
<laughing: several participants>
K: < hjälp >
K: <help>
< very quiet>, < event: opens the envelope>
A: < va står de{t} >
A: <what does it say>
<quiet>
C: hur kan vi på bästa sätt bevara den naturliga livsmiljön i < städerna >
C: how can we in the best way preserve the natural life environment in the <cities>
<chuckling>

3.2 Analysis

The transcriptions were then analyzed together with the video recordings, using the schema in table 2, below (left column).

First, each coded contribution is placed in an activity, a subactivity and an exchange type. Next there is a description of the communicative act function of the contribution.

The vocal verbal utterance is rendered in transcribed form, its syntactic-semantic properties are coded and prosody, gesture and other relevant actions are coded. This is followed by a coding of the responsive and the evocative functions of the contribution. Finally, the interactive effect, i.e. reactions of participants not having the floor, can be coded. The reaction of the participant who then becomes the next main contributor can be seen primarily in the responsive function of the next contribution. The coding categories refer back to the model in figure 1 above.

In table 2, the analysis schema is exemplified by an analysis of the initial contributions of the videorecorded interaction.

The table shows the coding of the three first contributions belonging to the subactivity of setting up the discussion (contributions 1 and 3 are utterances; contribution 2 consists of joint laughter by the four girls at the end of contribution 1).

The transcribed vocal verbal utterances are in Swedish, with an English translation of the utterance in italics. Some conventions of the transcription system are that: /, / /, / / / mark pauses of increasing length, { } encloses letters of the written word, which are not pronounced.
[ ] marks overlaps, which are also numbered.
We use < > to show (i) the position of comments and gestures, in the vocal verbal transcription line (it can surround words or phrases or just a space when there is no speech and only a gesture.) The line below the transcription of speech then contains the comment itself also enclosed in < >. When there are many comments, they can be numbered for clarity.
It is possible to add more coding in the schema, for example, prosody has not yet been included in the example below.

Table 2. Using the analysis schema to analyze the first utterances of the discussion in example 2. Main contributions are in boldface print.

CONTRIBUTION 1	
Activity	**University study of discussion of set topic**
Subactivity	**Start – setting up activity**
Exchange type	Request-acceptance (C1-C2-C3)
Contribution: communicative act function	**Contribution 1:** permission to have coffee, implying request to start interaction
Vocal verbal	**B: då får ni dricka kaffe å ha de{t} trevli{g}t** ***B: then you can drink coffee and have a nice time***
Syntactic-Semantic:	completed sentence
Prosody/ Gesture/ Action	B: Leaving room
Function RESPONSIVE (functional relevanc)	TO ACTIVITY AND ROLE AS EXPERIMENT LEADER
Function EVOCATIVE	EVOCATIVE: IMPLICIT REQUEST FOR ACTIVITY/ GIVES FLOOR/ DISPLAY OF LEAVING

	DIRECTED TO ALL – NO SPECIFIC NEXT SPEAKER SELECTED
Interactive effect (behavior by participants not having the floor)	Contribution 2: all look at B and laugh Function: responsive/acceptance of B's leaving, Evocative: B can leave
CONTRIBUTION 3:	
Activity	**Same as above**
Subactivity	**Same as above**
Exchange type	Request- acceptance, Question-answer, Acknowledgement of permission
Contribution: communicative act function	**Contribution 3: Acceptance of request+ question**
Vocal verbal	**K: ja ska ni kolla på oss därigenom då elle{r}** ***K: yes will you look at us through there then or***
Function RESPONSIVE.	FEEDBACK, CONTACT, PERCEPTION, UNDERSTANDING Acknowledgement of permission and implied acceptance of request
Function EVOCATIVE	ANSWER QUESTION GIVES FLOOR TO B (NEXT SPEAKER SELECTION)
Syntactic-Semantic:	Feedback "yes, - Interrogative "ska ni kolla på oss därigenom då", Feedback eliciting word "eller"
Prosody/Gesture/Action	K: Pointing towards the control room> K: EYE GAZE AT B
Interactive effect (behavior by participants not having the floor)	

If we look at the coding schema, we can see that many types of information contribute to the identification of a contribution, when it ends, how the speaker change is managed and who speaks or contributes next. The coded features are perhaps not all important in example 2, but can all be important in more problematic sequences than the one exemplified above.

4. Results of the analysis - Data related to the 100 first utterances

4.1 More on overlaps

In the first 100 contributions of the recorded discussion, we find overlaps between speakers in 29% of the contributions and of these 15% of the vocal verbal utterances were totally overlapped. Overlaps are, thus, frequent.

Example 3 contains three totally overlapped unimodal vocal verbal utterances made by K and one occurrence of a totally overlapped gesture (head nods) made by K, when M is the main contributor.

Example 3. Overlapped unimodal vocal verbal and gestural contributions.

M: [1 man kunde ha]1 gratis bussar kanske inne i stan liksom ja{g} tänker på dom hä{r} [2 parkerings]2 platser runt[3 om] < 4 å så bussar som gick i centrum >]4 då skulle de{t} ju: göra väldi{g}t mycke{t}
[1 you could have] free buses maybu in the city like I'm thinking of those [2 parking]2 places [3 around]3 [< 4 and then buses that ran in the center: >]4 then it would make a very much
K: [1 så ja{g} menar]1
[1 so I mean]1
K: [2 {j}aa]2
[2 yees]
K: [3 ja]
[3 yes]3
K: [4 < >]4
<three small head nods>

In overlap [1], M and K are competing for the role of becoming the main contributor, starting at the same time and since M wins the floor and continues, K's vocal verbal attempt is totally overlapped. Overlaps [2] and [3] are of a different type, where K contributes positive vocal verbal feedback, while M is talking and continues to be the main contributor. The same is true of the three small head nods from K in overlap [4], providing overlapped supportive positive gestural feedback, while M continues to speak.

Totally overlapped multimodal contributions are also quite common. Consider example 4 below, where C is the main contributor, talking and gesturing about big concrete buildings in cities resembling prisons, while M provides feedback in the form of vocal-verbal *m* and *mhm*, combined with simultaneous head nods. (There is also frequent overlapping laughter feedback during this contribution.)

Example 4. Overlapped multimodal contributions

C: … så synd å bygga hus på de{t} där viset [1 de{t} e stora]1 betong / [2 klumpar]2 bara så kommer man in i en jätteentre så här < >1 / / å så är de{t} korridorer på siderna å så e de{t} dörrar längs me{d} korridorerna < >2 så de{t} e precis som de{t} ba{ra} saknas galler för fönster < också >3
C … such a pity to build houses in that way [1 it is big] concrete / [2 lumps]2 only then you get into a giant entrance like this <1 >1 / / and then there are hallways on the sides There are doors along the halls <2 >2 so it is precisely like there are only bars across the windows missing <3 too >3
< gesture: shows a long hallway>1
< gesture: shows how the doors are lining the hallway>2
< laughter: several>3
M: [1 < m >]1
< head nod >
M: [2 < mhm >]2
<head nod >

4.2 Laughter and chuckle

Joint laughter episodes at the end of an utterance, often including more than two persons occurs in 11% of the utterances. Chuckles from a speaker simultaneously with her speech occurs in 5% of the utterances. (See example 2 above, the sections marked with the comment < laughing > and < chuckling > (in boldface).

4.3 Short contributions

One-word utterances make up 35% of all vocal verbal utterances in the recorded discussion and very quiet utterances (of one or more words) make up 7%. Thus, some of the one word utterances can be very quiet.

If the recognition is based only on the acoustic signal for such data, t his will cause difficulties for quite a large proportion of the utterances and speaker changes.

If we instead consider small contributions that are unimodally gestural among the first 100 contributions, we find 17 head nods (most of them repeated) and one headshake. The distribution of these head movements between the participants is very uneven. Of the four participants (M, A, C, K), one participant, M, produces 11 nods and 1 shake. A produces 3 nods, C 2 nods and K 1 nod. The uneven distribution can be attributed to the specific role of M as the participant who has recruited the other three participants as volunteers for the recording. This seems to make her the main or unmarked addressee for the others and as a consequence of this also the main producer of feedback to the others. This also means that in multiparty interactions, it would be wrong to assume that head nods always are spread uniformly among participants.

If we take a closer look at the three main phenomena that were mentioned above as causing difficulties in recognizing speaker changes as well as in identifying what is said when the speaker changes, i.e. overlaps, laughter/chuckle, and short contributions, we see that the three phenomena are all in general noticeably multimodal phenomena. In addition, speaker change as such involves multimodal interaction cues, like gaze direction, speech prosody etc. To what extent these phenomena are consciously noticed by the participants is not known, but, in any case, it seems clear that they react to cues in different modalities. Finding ways of identifying cues in different modalities and merging these cues is therefore a necessary, but difficult task in trying to achieve automatic recognition of speaker change.

5. When spoken interaction overlaps with another ongoing activity

An additional difficulty is that parallel to the main discussion, there is another on-going non-linguistic action, which might influence the interaction in specific ways, sometimes providing a "parallel" and sometimes a more integrated interaction pattern. An example of this is when coffee is served in the video recorded conversation.

In fact, in the analyzed interaction, two main types of simultaneous activity are going on simultaneously with the discussion. The first activity only takes place in the beginning of the interaction and consists in the handling of an envelope with an enclosed paper containing the assigned topic of discussion. This activity takes place in parallel with utterances 1-8, and, thus, overlaps with seven speaker transitions. What happens is that the envelope is opened by one of the participants (K) who then reads the set topic of the discussion, while the other participants produce reactions to this and the paper is then shown to everyone and especially one participant (A) studies it more closely, as the participants jointly try to specify and interpret the topic.

Figure 3. Simultaneous activities – Serving coffee and taking a closer look at the assigned topic of discussion

The second activity is the handling of the coffee and cookies, which continues in parallel throughout much of the discussion, but becomes especially salient, intervenes and replaces the discussion sometimes, see during utterances 12-19, 26-27 and 38-39, i.e. in total 11 speaker transitions.. When the participants engage in this activity, one of them serves coffee and they hand the plate with the cookies to each other. There are various comments regulating this activity, which are interspersed in the ongoing discussion of the assigned topic. An interesting issue concerns the nature of the relation between the discussion and the coffee drinking. For example, when one of the participants(C) repeatedly changes the focus to the coffee drinking, this could perhaps be seen as contributing in an less demanding way than by arguing in the discussion (see below, example 5).

Example 5. Serving coffee as an activity that is simultaneous and intervening with the discussion

K: [1den naturli{g}a livsmiljön]1 vadå den sköter < la sig själv >
[the natural life environment] what it takes <care of it self doesn't it>
<laughter: several>
M: de{t} e bara å e{h} [2plocka hit en massa (...)]2
it is just to e [bring here a lot of (...)]
C: < [2ska börja å dricka kaffe kanske]2>
<should start to drink coffee maybe >
<All: chuckle>, <event: C pushes cup towards K, K lifts thermos M pushes her cup towards K and moves A's cup to the side>
K: < upp med den lite grann >
< up with it a little >
<very quiet>
<M raises her cup>
M: < ops >
<ops>
<ingressive>
K: upp med koppen // < >
up with the cup // < >
<laughter: several>
M: <så>
<so>
<M raises her cup higher>
K: annars får du ju hälften i <// kan man ju >
[otherwise you know you'll get] half in < // can you you know>
<very quiet, <event: K pours coffee into M's cup, C bends to the side to watch>
K: ha /< / > va{d} tycker ni
yes // what do you think
M: [3vadå]3 nä den sköter sej [väl själv] de{t} e bara å stoppa dit en massa bilar å en massa höghus så [blir de{t} väl] naturli{g}t å0 bra
[3what3] no I guess it takes care of it |self] it is just to put a lot of cars there and a lot of high rise buildings so [I guess it will be] natural and good
A: [3vadå]3
[3what]3

In this case, C introduces serving coffee as an intervening topic, which is also focused on for a while in what is being said. C watches the coffee being poured very ostentatively, while A grabs the note with the assignment and studies it. K's and M's joking first comments on the topic frames the sequence, with M rephrasing and developing K's utterance. Except for the introduction of the coffee serving sequence, K keeps the initiative both in speaking and in serving coffee. This sequence contains overlaps and joint laughter, one very quiet comment and two completely overlapped utterances. It exemplifies how management of who is the main contributor to the interaction is maintained both in the discussion and in the serving of coffee.

6. Co-construction and the notion of main speaker or main contributor

Example 6, below, shows how contributions are interleaved in the discussion. This makes it difficult to identify one person as the main speaker or communicator at a specific time in the sequence. C and M are talking simultaneously, K adds a feedback word and C rephrases her own ending to complete M's sentence and then also finishes with an affirmative feedback word. Is C the main contributor here all the time? Are both C and M main contributors at the same time (overlapping and sharing in one contribution)? Is K the main contributor during her (non-overlapped) utterance of a feedback word? Or should rather the whole sequence be seen as a co-constructed contribution from C, M and K? Different types of analysis make it possible to choose different alternatives depending on the purpose of the analysis (summarization, argumentation or information giving). In general, a combination of bottom-up analysis with a top-down analysis (see section 2.1), considering the content and the expectations of the participants, can probably provide the best basis for a wise decision, concerning the extent to which it is necessary to distinguish the contributions of different participants.

Example 6. Co-construction

C: [1 att de{t}]1 inte finns nå{g}ra alkolister i frankrike å danmark för dom har stö+
    större alkolism än va{d} dom har i sverige < >
C: [1 that there]1 are no alcoholics in france and denmark because they have greater alcoholism than what they have in sweden < >
    <turns towards M>
M: [1 va{d} sa{de} du dom har de{t} dom har mycke{t} större]1
M: [1 what did you say they have it they have much greater]1
K: {j}aså
K: oh
C: än vi har i sverige ja:
C: than we have in sweden yes

7. Conclusions

This study has focused on factors that influence the dynamic recognition of the main contributor in communication and of changes of main contributor in multiparty face-to-face interactions.

We point to and illustrate the fact that automatic analysis of spoken interaction, in addition to speech reductions, involves several difficulties, that make it hard to identify main speakers/contributors, the transitions between them and what they actually say.

With respect to identifying the main communicator and shifts of main communicator, some of the problems are overlaps in transitions (in our example 30%) between contributions and many quick, small contributions, sometimes of very low intensity. There is also a large amount of individual and situational variation of the properties of contributions and it is difficult to identify what features of interaction are relevant for communication. As we have seen, in general, a combination of top-down and bottom-up processes is necessary for dealing with these tasks.

In summary, we want to stress three points:

1) If our goal is to capture all the clues for recognizing main communicator, shift of main communicator and the content in multiparty interaction, there are very many features of different types to register and analyze (cf. Table 2. above).
2) What is generally required is a combination of bottom-up and top-down strategies for a) recognition of main communicator and change of main communicator, b) processing the content of what is communicated as well as planning next contribution and c) production of relevant responses,
3) The notion and recognition of the main communicator has to be adapted to the fact that content and perhaps even contributions are often co-constructed by more than one contributor. This could be missed if individual contributions are not continuously related to other contributions

References

Allwood, J. (1976) *Linguistic Communication as Action and Cooperation,* GML 2, Univ of Gothenburg, Dept of Linguistics.

Allwood, J. (1999). "The Swedish Spoken Language Corpus at Göteborg University". In Proceedings of Fonetik 99, *GPTL 81*, Univ of Göteborg, Dept of Linguistics.

Allwood, J. (2000). An Activity Based Approach to Pragmatics". In Bunt, H., & Black, B. (Eds.) *Abduction, Belief and Context in Dialogue: Studies in Computatio-nal Pragmatics*. Amsterdam, Benjamins, pp. 47-80.

Allwood, J.(2002). Bodily Communication - Dimensions of Expression and Content. *Multimodality in Language and Speech Systems.* Björn Granström, David House and Inger Karlsson (Eds.). Dordrecht: Kluwer Academic Publishers, pp. 7-26.

Allwood, J. (2008). Multimodal Corpora. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics*. *An International Handbook.* Mouton de Gruyter, Berlin. 207-225.

McGurk H., & MacDonald J. (1976). "Hearing lips and seeing voices.". *Nature* 264 (5588): 746–8.

Allwood, J. ; Ahlsén, E. (2012). Multimodal Communication. *Handbook of Technical Communication. A. Mehler, R. Romary & D. Gibbon (eds). Mouton De Gruyter.*

Nivre, J. (1999). Modified Standard Orthography, Version 6 (MSO6). Univ of Gothenburg, Dept of Linguistics.

Nivre, J. (2004) Gothenburg Transcription Standard (GTS) V.6.4. Univ of Gothenburg, Dept of Linguistics

Sacks, H. Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50 (4), 696-735.

Salamin, H. & Vinciarelli, A. (2012).Automatic Role Recognition in Multiparty Conversations: an Approach Based on Turn Organization, Prosody and Conditional Random Fields. Accepted for publication by *IEEE Transactions on Multimedia,* to appear, 2012.

Valente, F & Vinciarelli, A. (2010). Improving Speech Processing trough Social Signals: Automatic Speaker Segmentation of Political Debates using Role based Turn-Taking Patterns. *Proceedings of the International Workshop on Social Signal Processing*, pp. 29-34, Florence, 2010.

Vinciarelli, A. (2009). Capturing Order in Social Interactions, *IEEE Signal Processing Magazine***,** Vol. 26, no. 5, pp. 133-137.

Vinciarelli, A. (2007).Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling
*IEEE Transactions on Multimedia***,** Vol. 9, no. 6, pp. 1215-1226.