Turn-taking patterns in human discourse and their impact on group communication service design

Kostas KATRINIS*, Gisli HJÁLMTÝSSON**, Bernhard PLATTNER*

Abstract

Recent studies demonstrated the benefit of integrating speaker prediction features into the design of group-communication services supporting multiparty online discourse. This paper aims at delivering a more elaborate analysis of speaker prediction by analyzing a larger volume of data. Moreover, it tests the existence of speakers dominating speaking time. Towards this end, we analyze tens of hours of recorded meeting and lecture sessions. Our principal results for meeting-like interaction manifest that the next speaker is one of the last four speakers with over 90% probability. This is seen consistently across our data with little variance (standard deviation of 8.71%) independent of the total number of potential speakers. Furthermore, lecture time is in most cases significantly dominated by the tutor. In meetings, although a single dominating speaker is always evident, domination exhibited high variability. Generally, our findings strengthen and further motivate the act of incorporating user-behavior awareness into group communication service design.

Key words: Group dynamics, Speech, Speaker, Linguistic analysis, Human communication, Oral communication, Teleconference.

CONFIGURATIONS DE PRISES DE PAROLE DANS LE DISCOURS HUMAIN ET LEUR IMPACT SUR LA CONCEPTION DES SERVICES DE COMMUNICATION DE GROUPE

Résumé

Des études récentes montrent l'intérêt d'intégrer des caractéristiques relatives à la prédiction du locuteur dans la conception des services de communication de groupe qui admettent une conversation en ligne entre plusieurs personnes. L'article vise à obtenir une analyse plus élaborée de la prédiction du locuteur par l'analyse d'une plus grande quantité de données. Il teste en outre l'existence de locuteurs qui dominent le temps de parole. On analyse dans ce but des dizaines d'heures d'enregistrement de réunions et de conférences. Les principaux résultats pour des interactions de type réunion indiquent que le locuteur suivant est l'un des quatre derniers locuteurs avec une probabilité de plus de 90 %. Cela est vérifié de

* Communication Systems Group, Swiss Federal Institute of Technology (ETH Zurich) – Gloriastrasse 35, 8092 Zürich, Switzerland; {katrinis, plattner}@tik.ee.ethz.ch
** Network Services Laboratory, Reykjavik University – Ofanleiti 2, 103 Reykjavík, Iceland; gisli@ru.is

façon constante pour l'ensemble des données avec une faible variance (écart-type de 8,71 %) quel que soit le nombre de locuteurs potentiels. Dans les conférences, le temps de parole est dominé dans la plupart des cas par le conférencier. Dans les réunions, bien qu'un seul locuteur dominant apparaisse toujours, la domination se révèle plus variable. De façon générale, les résultats renforcent et motivent l'incorporation du comportement des utilisateurs dans la conception des services de communication de groupe.

Mots clés: Dynamique groupe, Parole, Locuteur, Analyse linguistique, Communication humaine, Communication orale, Téléconférence.

Contents

I. INTRODUCTION

In various fields of computer science, patterns in resource usage or user behavior are often exploited for further optimization of algorithms and services. For instance, the execution of computer programs is associated with high locality in terms of memory references. Page replacement algorithms in paged operating systems take advantage of this phenomenon to reduce the probability of page fetches from secondary storage to main memory, thus achieving faster code execution (e.g. Least Recently Used (LRU) replacement strategy [1]). Particular to computer networks, locality has been used in large-scale distributed simulations and multiplayer online games [2] as a means of limiting the amount of data received by a node participating in the session. In short, each node interacts at full-rate only with the nodes residing in its declared vicinity (termed "area of interest"), while exchanging only coarse-grained control information with the rest of the session's nodes.

Out of the large set of potential multiparty applications, herein the focus is on "teleconferencing-like" applications enabling online discourse. For instance, we consider Voice-over-IP (VoIP) multiparty meetings, synchronous distance learning sessions and online workshops. In our previous work on multicast tree management, we have demonstrated the benefit of tree caching in multi-source conferencing sessions over Source-Specific Multicast (SSM) [3]. One alternative for realizing multi-source SSM sessions mandates creating a tree rooted at a group member as soon as the latter starts speaking (on demand). We showed that there is value in maintaining the tree rooted at a recently spoken conferee alive, as the probability of the same conferee speaking again in the near future is considerably high (*temporal locality*). This observation motivated us to extend our analysis to a larger set of interaction traces in an attempt to extract patterns that could be generally of interest to the designer of a network service supporting online multiparty applications.

In this paper, we analyze fifty-two hours of recorded discourse from thirty-nine interactive multiparty events with physically collocated participants (face-to-face). The recorded events comprised both of lectures and meetings recorded in a non-televised setting. We ana-

lyze the sequence and duration of turns taken in each event and test the following two hypotheses:

- H1: The existence of *speakers dominating speaking time* and
- H2: The ability of *predicting with high probability the future speaker* from the short term interaction history.

Our principal finding suggests that in meeting-like interaction, predicting the next speaker from a small constant number of previous speakers – independent of the group size – can be accomplished with considerably high probability (over 90% on average). Applying the same to lecture-like discourse proved irrelevant, solely due to low interactivity caused by the lecturer's domination in terms of speaking time. Last, our analysis confirms the hypothesis of a single dominating speaker across all meeting traces processed, however with varying significance.

The primary contribution of this work is showing that there are indeed specific patterns in human discourse that can be exploited to design more efficient group communication services. Although designs that take such patterns into consideration have been already introduced (partly by the authors, see Section II), a strict and large-scale statistical study of these patterns has so far been missing. As such, we see our work valuable in validating and optimizing existing designs and in possibly inspiring new approaches taking interaction behavior into consideration. In general, our work validates and strengthens the impact of incorporating user-behavior awareness to the design of group communication services.

The structure of this paper is as follows. The next section outlines previous work on the topic and discusses how this interrelates to the content presented herein. Section III presents the formal framework of the theory put under test and elaborates in the practical applicability of our primary hypotheses. The next two sections describe the method and scope of the performed data analysis, whose results are presented and discussed in Section VI. We conclude in Section VII with discussion and future work.

II. RELATED WORK

By definition, conversation is a sequence of possibly overlapping speaking periods – called *turns* – and pauses. In each turn, one person talks, while another (or more) listens. Hence there is a subconscious assignment of speaker and auditor roles to the conversation parties. This assignment changes frequently as the conversation evolves, with or without the current speaker's consent (interruption) and either explicitly (e.g. by forcing a turn transition with addressing the word to a listener) or implicitly (e.g. eye gaze, gestures, content). Analyzing and modelling turn-taking has been an ongoing effort for more than thirty years. Discourse analysts [4] noted that humans innately delineate the start/end of turns and have focused on the factors driving these systems. Additionally, they investigated the various types of transitions between adjacent turns and elaborated on the human factors motivating them. Generally, turn-taking constitutes a multidisciplinary topic that has been studied by various fields of science, like linguistics [4], psychology [5] and sociology [6] to name a few. Although we adhered to the established terminology and borrowed various definitions from these studies, this work approaches turn-taking from a completely different perspective. More precisely, we perceive turn-taking as a sequence of three types of finite duration events: turns, overlaps

and pauses. Using this simple model, we test the existence of specific patterns, leaving out the process of inquiring the aspects of human behavior that may cause the occurrence of such patterns.

Particular to studying turn-taking habits as a means of improving online human interaction, McKinlay et al. [7] evaluated the impact of various turn-taking protocols on the performance of a small group of humans collaborating over a networked CSCW (Computer Supported Collaborative Work) platform. Although the authors mentioned the frequent occurrence of temporal locality in turn-taking (referred to as "adjacency pairs"), they neither specified the occurrence probability of this phenomenon, nor did they test the existence of locality in turn subsequences of size larger than two (therefore the term "pair").

To our knowledge, [8] was the first work that combined turn-taking analysis with the reflection of its findings to network design, resulting in the ALNAC system. ALNAC [8] (Application-layer Network Audio Conferencing) is a special-purpose Application-Layer Multicast (ALM) routing protocol targeted at audio-conferencing. The novelty of its design rests on the observation that minimizing latency over an ALM infrastructure from the active speaker to every other group member is not the most effective choice in terms of perceived quality. In fact, this can result in a number of participants experiencing unacceptably high (for real-time interaction) delays. On the contrary, ALNAC minimizes latency from the current speaker to a small (constant) set of participants, who are most likely to interact with the current speaker. It is exactly these participants, who need interactive latencies in order to react to conversational cues. The rest of the group members, participating as passive listeners at that particular time of the session, can tolerate higher latencies using normal (non ALNAC-optimized) overlay routing. ALNAC builds the set of future speakers from the last five spoken participants. This decision resides on results obtained after analyzing four multiparty conversations (two audio-conferencing sessions and two public-meeting traces) and which essentially state that the accuracy of picking the future speaker out of the last five spoken participants is on average *92%*. The connection of the present work to ALNAC is twofold. While [8] offered a small scale proof of locality in turn-taking, herein we provide for a more robust study of the phenomenon by analyzing a much larger volume of discourse sessions. Still, ALNAC constitutes a straightforward example of how the results introduced in this paper can be put to good use in order to design more efficient services in support of online discourse.

In our previous work [9], we focused on the dynamic allocation of source-rooted trees in multiparty conferences over SSM. One approach [10] for implementing sessions with multiple sources over SSM is to build a distribution tree rooted at a session participant on demand, i.e. at the time this specific participant starts speaking. This reactive approach suffers from two critical shortcomings: first, it often leads to considerably long communication outages due to delay in the creation of the on demand tree. Second, it increases the cost of the service in terms of router processing resources, due to intensive tree creation/tear-down activity. As a remedy, we proposed caching a constant number of least recently used trees, building up on the observation that a recently spoken conferee is highly probable to speak again in the near future. Through simulation of two real-life meetings, we showed that the caching approach manages to alleviate both shortcomings of the elementary on demand approach to a great extent. The results presented herein form both a complement to and an extension of [9]: a) they strengthen the applicability of the tree caching approach by proving the existence of temporal locality in a much larger dataset of interaction traces and b) they show that application-aware tree management can potentially benefit from other interaction patterns beyond locality.

III. THEORY AND ANALYSIS

In this section, we present the hypotheses put under test by our study. Additionally, we discuss for each pattern how it can be exploited by two specific network services – namely Source-Specific Multicast and Application-Layer Multicast – to provide for more efficient online discourse.

III.1. Dominance

Often, a small number of speakers tend to monopolize the word during multiparty conversations. Many factors may lead to this pattern, like for example the discussion topic (e.g. when the topic mostly concerns only a fraction of the speakers, who tend to express themselves more frequently), the status of a speaker (e.g. the CEO in a company meeting is more probable to monopolize the podium) and/or simply due to human nature for some humans are more extroverted than others. Regardless of the reason, we are interested in finding out the frequency of monopolization effects and the extent of monopolization. The latter expressed both in the number of speakers monopolizing a single session and in terms of speaking duration of each monopolizing speaker.

We use the term "*dominance*" to refer to the phenomenon of a speaker exceeding his fair speaking ratio and define the "*dominance factor*" metric to quantify the degree of monopolization. More formally, let T_{event} be the entire duration of a multiparty session and T_i stand for the cumulative speaking time of speaker i throughout the session. Let also *fst* (measured in seconds) stand for the fair[1] speaking time of a conversation, i.e. the nominal total speaking time of any speaker, if the session's total speaking time were equally allocated to all participants. Assuming that S is the set of participants spoken over the entire session, *fst* is given by the term:

$$fst = \frac{T_{event}}{|S|} \quad (1)$$

We define the dominance factor d_i for speaker i as the ratio:

$$d_i = \frac{T_i}{fst}, \forall i: T_i > fst \quad (2)$$

Essentially, the dominance factor metric captures the degree of significance, by which a particular speaker exceeds his theoretical speaking time, if fair sharing of speaking time were employed. To name an example, consider a two hours meeting with 10 participants. The fair speaking time common to all participants is 0.2 hours in this case. Assuming that speaker 3 talks for 36 minutes in total, his dominance factor is then $d_3 = \frac{36}{12} = 3$. Equivalently, speaker

1. While acknowledging that uneven speaking time may be natural and proper, we use the term "fair" to refer to equal speaking time among the participants.

3 spoke three times more than he would talk, were speaking time allocated in a fair manner. Note in Equation 2 that we constrained the dominance factor definition only to speakers, whose participation level exceeds their fair speaking ratio. Thus, it is straightforward that $d_i > 1$ will always hold.

Network services aware of speaker dominance are capable of improving service quality and/or cost by differentiating the service offered to dominating speakers as opposed to less talkative participants. Specifically, in the case of multiparty conferencing over SSM, each dominating speaker can be assigned a static tree – i.e. a tree that is kept alive for the entire session lifetime – whereas the rest of the participants can be served with on demand created trees. Since creating on demand trees can potentially cause interruptions in data delivery [9], assigning static trees to dominating speakers improves quality of service. Moreover, reducing the cumulative number of tree setup events that indeed alter router state economizes on router processing resources and therefore reduces cost.

In many applications, the potential dominating speaker(s) may be known in advance. This is for example particularly true for lecture sessions, where the tutor will with very high probability (almost deterministically) dominate the speaking time. This is indeed verified by our analysis results presented later in this paper. However, specifying the potential dominating speaker(s) in advance of the session is not trivial for other types of online discourse, collaborative-work meetings being a typical example. For these cases, devising (learning) algorithms that are capable of predicting dominating speakers as the session advances constitutes an interesting research topic. This is however out of the scope of the present paper.

III.2. Speaker Prediction

In the following analysis, we make use of the sorted set $U = (u_1, u_2, \ldots, u_n)$ of global turns, n being the total number of turns taken throughout the session. Additionally, given the set of identities of all spoken participants S, we define the function $sp: U \rightarrow S$ as the operator that matches a turn to the identity of its speaker.

Here, we are primarily interested in testing the existence of temporal locality in turn taking. By the term "temporal locality" we refer to the probability of previous speakers appearing as speakers in the near future. If this pattern occurs with high probability, we can then predict the future speaker with high accuracy from the short-term interaction history. More formally, let u_i be the next turn, taken by participant j, i.e. $j = sp(u_i)$. We define the past speakers window PSW_i of size w ($w \leq |S|$) at turn i as the sorted set of identities of the last k distinct speakers prior to turn i, where k is given by:

$$k = \begin{cases} i-1, \text{ if } i \leq w \\ w, \text{ if } i > w \end{cases} \tag{3}$$

We test the existence of locality by calculating the probability P of the identity of the next speaker matching one of the identities contained in the past speakers window or equivalently:

(4) $$P(\{j\} \cap PSW, \neq \varnothing)$$

Particular to the problem of on demand tree management in multiparty SSM sessions (see Section II), the above probability equals to the hit ratio of an LRU (Least Recently Used) tree cache of size w. To enhance intuition, we give an example of an audio conference among a set S = {1,2,3,4} of four speakers. The SSM middleware at each of the four conferees maintains a cache with the w least recently spoken participants (in this example we set $w = 2$). For each cache entry, a tree rooted at the respective SSM source is maintained. Every new turn taken triggers an update of the cache: if a tree rooted at the new speaker is not cached, the SSM middleware builds a new tree towards the new speaker. Additionally, if the cache is full, the least recently used tree in the cache is torn down and replaced by the new entry. In case the cache is not full at the start of the new turn, the newly created tree is just added to the cache and the usage flags of all cache entries are updated. Let the global turn sequence be given by the set $U = (2, 3, 4, 3, 4, 1, 2)$. The first three turns cause a cache miss, leading thus to creating new trees rooted at the respective speakers. However, at the fourth and fifth turn, every conferee has already a cached tree to speakers 3 and 4 respectively (cache hit) and thus the overhead of tree creation is for these two turns avoided. Finally, the last two turns lead to a cache miss. Overall, the probability P of the next speaker being among the last 2 speakers is $P = 28.57\%$ in this example.

Note in the last example that if we increased the size of the last speakers window to $w = 3$, the overall probability would increase to $P = 42.85\%$. In general, there is a trade-off between the accuracy of next speaker prediction and the cost of taking locality into consideration. For instance, in the case of tree caching in SSM, increasing the size of the cache (and thus the probability of a correct prediction) increases the cost due to the additional amount of router state required to keep the additional cached trees alive. In application-layer multicast, the out-degree of each overlay node in the distribution tree is normally bounded by a maximum number threshold d_{out}. If ALNAC is used on top of an overlay routing protocol, the current speaker j streams to the w potential next speakers per unicast and to the rest $d_{out} - w$ children using overlay routing. The w overlay nodes, that j would serve, if ALNAC were not used, are delegated by j to his d_{out}-w children. Here, increasing the size w of the last speaker's window causes the current speaker to delegate more overlay neighbors to his children, thus increasing the deviation from "routing optimality" as mandated by the underlying overlay routing protocol.

IV. EXPERIMENTAL SETUP

IV.1. Input Dataset

For the purpose of our study we analyzed various multiparty sessions contained in the "MIchigan Corpus of Academic Spoken English" (MICASE [11]). We first classified online discourse to two major categories of interest: conversational meetings, where all meeting partners can potentially equally contribute as the meeting evolves, and classroom-like ses-

sions (e.g. courses, tutorials, talks), where implicit role assignment (e.g. tutor/student, main speaker/audience) causes a small fraction of participants to dominate in terms of speaking time. This classification corresponds fully with our intuition on types of discourse prominent in the Internet today and – as it will be made clear in the results section – is necessary, since we expect to find different patterns of interaction in the two classes of multiparty discourse.

Table I summarizes the nineteen events making up our input dataset pertaining to meeting-like interaction. All but one event were held within the academic community (higher education or research). Participation level in sixteen of the meetings ranged from 3 to 11 participants (6.67 on average), whereas the rest three meetings were more highly populated (21, 34 and 84 respectively). Note that across all meetings, no participant remained silent and therefore the number of participants equals the number of speakers in our dataset. Last, the mean duration of a meeting session was 1.54 hours, ranging between 0.68 and 3.12 hours. In total, we analyzed 29.25 hours of meeting time. Accordingly, Table II lists the titles of the MICASE lectures we processed, together with participation and duration information (22.81 hours lecture time in total). To allow for generalization of results, we included both undergraduate and graduate sessions, with population ranging from 17 to 400 participants.

TABLE I. – Group size, duration and short description of analyzed MICASE meetings.

Taille du groupe, durée et courte description des réunions MICASE analysées.

Meeting-ID	Description	Group Size	Duration (sec)
Meeting-1	Artificial Intelligence Research Group Meeting	9	5640
Meeting-2	Immunology Lab Meeting	8	3600
Meeting-3	Natural Resources Research Group Meeting	6	4980
Meeting-4	Physics Research Group Meeting	9	2460
Meeting-5	Forum for International Educators Meeting	21	6120
Meeting-6	Student Government Meeting	34	3960
Meeting-7	Media Union Service Encounters	84	11220
Meeting-8	Economics Office Hours	11	5520
Meeting-9	Art History Office Hours	5	3960
Meeting-10	Computer Science Office Hours	11	6960
Meeting-11	Intro Biology Study Group	5	6180
Meeting-12	Biochemistry Study Group	5	6540
Meeting-13	Chemical Engineering Group Project Meeting	4	4620
Meeting-14	Organic Chemistry Study Group	8	6060
Meeting-15	Math Study Group	3	7920
Meeting-16	American Family Group Project Meeting	6	5100
Meeting-17	Objectivism Student Group	6	7500
Meeting-18	Undergrad. Social Science Study Group	4	3840
Meeting-19	Honors Advising	4	3120
			Total:105300

IV.2. Data Processing

Each MICASE session is transcribed using custom semantic and marked up in SGML (Standard Generalized Markup Language) format. Of the various types of annotated events, three are of interest to our analysis:

- Speaker turns, containing the full text of the turn and marked with the identity of the speaker.
- Overlapping utterances by two or more participants.
- Pauses in speech, either within an ongoing turn or between two adjacent turns, and their respective duration in seconds.

TABLE II. – Group size, speaker number, duration and short description for each of the twenty analyzed MICASE lectures.

Taille du groupe, nombre d'intervenants, durée et courte description des vingt conférences MICASE analysées.

Lecture-ID	Description	Group Size	# Speakers	Duration (sec)
Lecture-1	Perspectives on the Holocaust Lecture	40	11	6000
Lecture-2	Principles in Sociology Lecture	50	20	4920
Lecture-3	Fantasy in Literature Lecture	150	8	4980
Lecture-4	Golden Apple Award Statistics Lecture	100	5	2700
Lecture-5	Drugs of Abuse Lecture	160	6	4080
Lecture-6	History of the American Family Lecture	100	9	4860
Lecture-7	Archeology of Modern American Life Lecture	25	18	4380
Lecture-8	Biology of Birds Lecture	17	9	5040
Lecture-9	Ethics Issues in Journalism Lecture	26	26	4980
Lecture-10	Intro Programming Lecture	17	6	3000
Lecture-11	Intro Anthropology Lecture	400	2	4440
Lecture-12	Medical Anthropology Lecture	40	9	4140
Lecture-13	Twentieth Century Arts	100	4	2460
Lecture-14	Intro Engineering Lecture	200	7	3120
Lecture-15	Renaissance to Modern Art History Lecture	150	2	3000
Lecture-16	Behavior Theory Management Lecture	60	54	4800
Lecture-17	Intro to Evolution Lecture	65	3	5880
Lecture-18	Media Impact in Communication Lecture	150	13	4320
Lecture-19	Literature and Social Change Lecture	45	4	5040
Lecture-20	Race and Human Evolution Lecture	103	8	4620
				Total Time: 82140

Previous studies on human conversation have approached turn taking from a linguistic point of view, therefore usually defining a speaker's turn as the stretch of speech by a speaker that consists of one or more utterances. Various criteria (prosody, semantics) are used in these studies to decide on the start and the ending of a turn. For our engineering purposes, we

adopted the definition proposed by Weilhammer [12]: the start of a turn is positioned either at the first word of the conversation or the first word interrupting the silence that follows the previous turn. Additionally, two successive turns by one speaker are always interrupted by an utterance of an interlocutor. The transcription methodology followed in the MICASE Corpus samples conforms to this definition. The only deviation is that MICASE transcriptions contain successive turns taken by the same speaker. Therefore, we applied the preprocessing step of merging adjacent turns taken by the same speaker into a single turn. Also note that overlapping utterances account in our analysis as distinct turns in the global turn sequence.

At various points of the present work, we are interested in the duration of a turn, apart from the turn taking sequence itself. As the MICASE transcriptions do not provide for the duration of each turn, except for the duration of pauses and the entire session duration, we devised a custom technique to calculate turn duration. In particular, we first count the total number of letters comprising the entire session's speech and the effective speaking time, the latter given by subtracting the total pause time from the session duration. Subsequently, we divide the total letter count by the effective speaking time, resulting to the time spent on pronouncing a single letter (termed *lettertime*)[2]. The duration of a turn can then be easily computed by counting the number of letters in the turn and multiplying it with the *lettertime*. Note, that in the process of letter counting, we incorporate spaces between words as well, for breaks do also add to the total speaking time. Timed proof reading of random turn samples confirmed that our automatic technique gives a good approximation to actual turn duration.

V. PATTERN SIMILARITY BETWEEN ONLINE AND REAL SESSIONS

Due to the lack of online traces, we conducted our analysis using traces captured during face-to-face multiparty sessions. In face-to-face sessions, visual cues and other bodily gestures aid communication, yet in online interaction these additional communication channels are not provided. Normally, this deficiency of online interaction should dramatically increase the number of overlaps and backchannel interactions. However, users becoming increasingly familiar with the medium realize that the single communication channel (or the couple of channels in case of synchronized audio/video) is exclusively important for communication and therefore tend to adhere to a gentle social interaction protocol (i.e. try to minimize overlapping turns or remain silent until the current speaker concludes). This is partially confirmed for two-speaker interactions in [13], where overlaps or short interrupting utterances are found by only 13% higher in telephone conversations compared to the face-to-face analog. With the advent of video conferencing and the ability to have real-time visualization of more participants beyond the current speaker, we expect to see an even closer match between online and real-life interaction patterns.

Particular to the patterns we seek for in this paper, we perceive both of them – locality and dominance – as features inherent to human communication per se, independent of the communication medium used. This lies in the fundamental way that people communicate with each other, no matter whether online or face-to-face. In fact, people do predict the next

2. This derivation holds under the assumption that speakers have uniform speaking speed. Although we acknowledge that this is not generally true, we don't expect this to heavily bias the results.

speaker, even if they just do not conceptualize it. If this were not the case, conducting a sensible conversation would not be possible at all. Consider the counter-example of user A asking user B a question, user C responding to A's question and then user D commenting on C's answer. If such patterns of interaction appeared frequently, reaching a point of understanding – essentially the ultimate goal of human conversation – would be impossible. Summarizing, if locality is frequently found in real-life interaction traces, the same will be true in online traces as well. The same argumentation applies to speakers dominating conversation time: the phenomenon is rather caused by human factors (e.g. due to some humans being more communicative or possessing more developed leadership skills than others) and/or context (e.g. when part of the speakers is specialized on the discussion topic and therefore dominates) and not influenced by the communication medium in use. In fact, it has been shown [14] that in business meetings the domination of the highest rank participant is magnified in online conversations as compared to the face-to-face paradigm.

The above argumentation holds under two reasonable assumptions, namely a) that all communication parties perceive acceptable audio quality from any potential speaker and b) that all speakers are well accustomed with the communication medium (e.g. VoIP or videoconferencing tool used). The above two assumptions guarantee that communication parties will not be impeded to behave naturally due to problems inherent to or otherwise caused by the communication medium.

VI. RESULTS

As manifested by the analysis outcome presented below, the two types of multiparty discourse – meetings and lectures – exhibited different interaction patterns. We present first the results stemming from the analysis of meetings and then proceed to lecture results.

VI.1. Meetings

VI.1.1. Dominance

We first tested the hypothesis of dominating speakers. For each meeting, we calculated the dominance factor of every speaker exceeding his fair speaking ratio (as given by Equation 2) and counted the number of speakers exceeding a specific dominance factor threshold D_{thresh}. We repeated the same procedure for various threshold values, ranging from $D_{thresh} = 1$ to $D_{thresh} = 6$ and using a step of 0.01. For each threshold value, we computed the mean, standard deviation, minimum and maximum number of speakers exceeding it over all meetings. Figure 1 plots the rates of the four statistical indices against threshold value. Unfortunately, the plot does only convey information about the high variability of the minimum and maximum number of speakers exceeding a given dominance factor across all meetings. We avoid drawing any conclusions using the average index due to the relatively high standard deviation of samples from the mean. In fact, the magnitude of standard deviation

remained always comparable to the mean, motivating us to further explore the cause of increased variance.

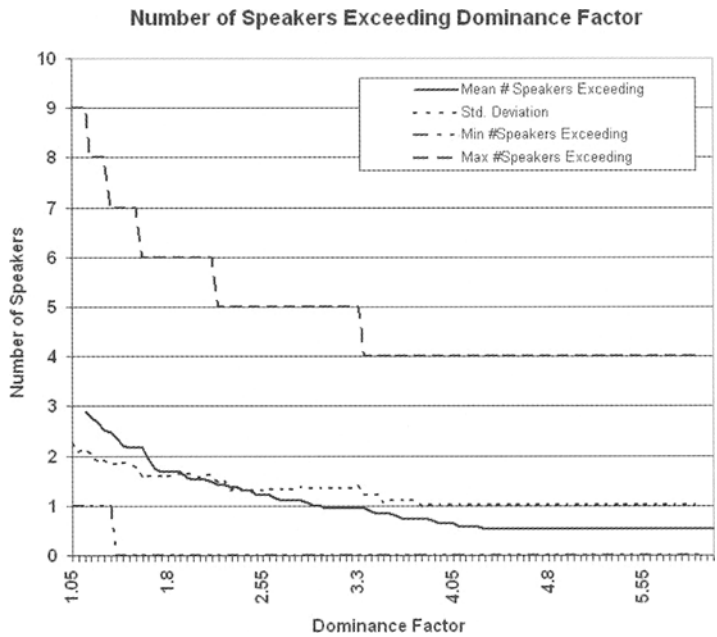

FIG 1. – Mean number of speakers exceeding a given dominance factor, averaged over all meetings and for various factor values. The plot shows the high variability, as manifested by the relatively large standard deviation.

Nombre moyen d'orateurs dépassant un facteur de dominance donné, calculé sur toutes les réunions et pour plusieurs facteurs. Le tracé montre une grande variabilité illustrée par un écart-type important.

Towards this, we first calculated for each meeting the cumulative speaking ratio $\frac{T_i}{T_{event}}$ for the two most talkative speakers – termed 1st and 2nd dominating speakers hereafter – and plotted it against the fair speaking ratio of the event, the latter defined as $\frac{1}{S}$. The graph is depicted in Figure 2. We also show in the figure the reference "fair ratio" line, where all pairs would lie, if speaking time were equally allocated. The scattering of < fair speaking ratio, cumulative speaking ratio> pairs manifests that in the majority of meeting events, the 1st dominating speaker exceeded to a great extent his fair speaking ratio. However, four of the events did not follow this pattern. In fact, half of these "non-conforming" pairs (marked with a circle in Figure 2) correspond to the two highest fair speaking ratios of the plot or equivalently to two of the meetings with the lowest number of speakers. A possible explanation for this is that in very small meetings (in terms of speaker number), speakers are more easily

prompted to speak. Also, any potential implicit denial to communicate becomes much more apparent in very small meetings. This motivates all parties to speak more frequently and thus leads to speaking time being allocated closer to equally to all speakers. Particular to the second most talkative speaker, the deviation from fair speaking ratio was not noticeably high, as shown in the scatter plot, and therefore the hypothesis of having two speakers significantly dominating is defeated.

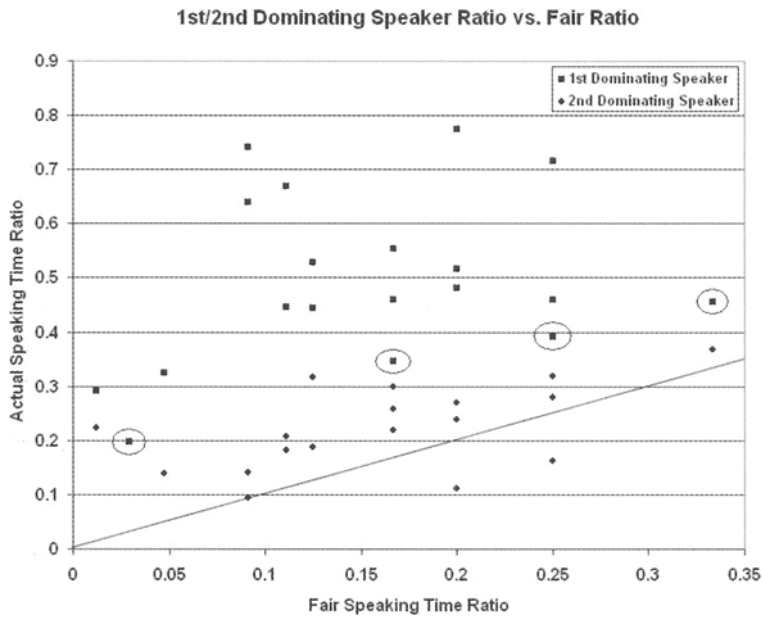

FIG 2. – Scatter plot comparing the fair speaking ratio to the actual speaking ratio of the two most talkative speakers and for each meeting.

Nuage de points comparant le rapport de parole juste au rapport effectif pour les deux orateurs les plus bavards de chaque réunion.

We quantify the degree of the 1st dominating speaker's dominance in the histogram presented in Figure 3. The dominance factor was at least 3 in 57.89% of the meetings, whereas only 15.79% had the 1st dominating speaker talk less than twice its fair speaking time. Recognizing that the dominance factor is a measure dependent on total speaker number, we also plot the mass of the dominance compared to total meeting duration. For this, we cluster the cumulative speaking ratios of the 1st dominating speakers using a bin-interval of 10% and illustrate the clusters' frequency over all meetings in Figure 4. The cumulative curve in this histogram conveys that in 73.68% of the meetings, the dominance of the most talkative speaker spanned at least 40% of the total session duration. Given the fact that all but one meetings comprised of four or more participants (i.e. fair ratio $\leq$ 25%), we conclude that dominance was clearly noticeable in almost 75 % of the cases.

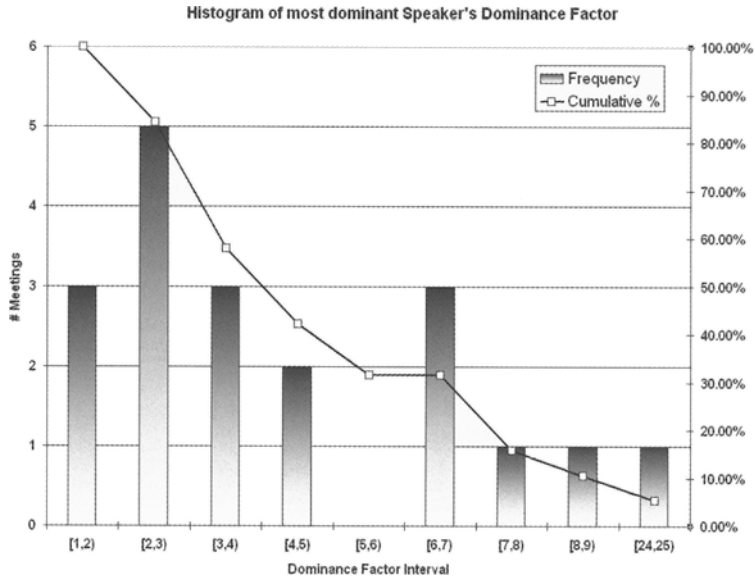

FIG 3. – Frequency of dominance factor for the 1st dominating speaker, clustered in one unit intervals over all meetings.

Fréquence du facteur de dominance de l'orateur le plus dominant avec regroupement par intervalle d'une unité.

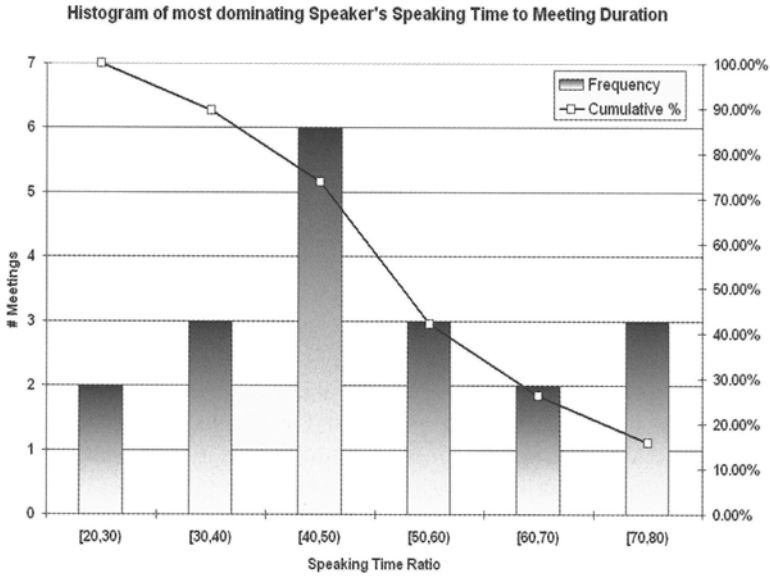

FIG 4. – Distribution of the ratio of the 1st dominating speaker's total speaking time to meeting duration over all meetings. The ratios are clustered in 10%-bins.

Distribution des rapports du temps de parole de l'orateur le plus dominant au temps total de la réunion, regroupés par blocs de multiples de 10 %.

VI.1.2. Speaker Prediction

We further tested the MICASE meeting traces against the hypothesis of correct future speaker prediction (with acceptably high success probability). As discussed in Section III.2, we hypothesized the existence of temporal locality in turn-taking and therefore specified for each turn the probability of the turn's source matching one of the speakers in the near past (see Equation 4). We experimented with three separate previous speakers' window sizes w, namely $w=2$, $w=3$ and $w=4$ respectively. We excluded *Meeting-15* from the entire analysis, for its low participation level (three speakers) would obscure the analysis with positive results not caused by the prediction method's efficiency. For the same reason, we excluded meetings with four and five speakers, when testing locality with window sizes of $w = 3$ and $w = 4$.

Table III shows the mean probability of successful prediction for all three window sizes. Clearly, guessing the next speaker correctly out of the last two speakers failed in almost $\frac{1}{3}$ of the times, exhibiting also high variability. Adding one more speaker to the prediction possibilities ($w = 3$) improved the mean success probability by 14.08%. As it can be seen in the Table, the magnitude of the standard deviation for $w = 3$ turned the benefit of using a larger window to marginal compared to the mean success probability of $w = 2$. For this reason, we also show for each window size the highest of the 10 %, 15 % and 25% worst success probabilities over all meetings (percentiles). For $w = 3$, the 25 %-percentile conveys that correct prediction failed in $\frac{1}{4}$ of the cases with probability of up to fairly 22%. Therefore, we decided to further increase the previous speaker's window to $w = 4$. Although this brought almost half the benefit of the last window expansion, it yielded average successful prediction in more than 91% of the times. Also, it reduced standard deviation to less than 10% of the mean. Overall, 90% of the meetings had a failure probability of fairly less than 15%.

We further depict in Figure 5 the cumulative distribution of success probabilities over all meetings. For a given prediction probability p on the horizontal axis of the graph, the respective value on the vertical axis corresponds to the fraction of meetings with successful prediction probability less or equal than p. As an overview, the figure illustrates the increase of benefit as the window size increases. Additionally, the graph is useful in studying the worst case behavior of the best performing window size ($w = 4$ depicted by the solid curve). We observe that more than almost 91% of the meetings yielded a success probability of over 80%. In fact, the two worst correct prediction probabilities for $w = 4$ were 67.91% and 84.53 %. This manifests that prediction succeeded with high probability in almost all of the meetings. More importantly, the latter occurred independently to the total number of potential future speakers (group size). For instance, this was the case for Meeting-7 (84 speakers) and Meeting-5 (21 speakers) with success probabilities of 94.93% and 84.53% respectively. Also, it is worth mentioning that of the three meetings with the least number of total speakers (six), two were not among the first five meetings that scored the best prediction ratios. Notice in Figure 5 that for correct prediction probabilities over 94%, the curve for $w=2$ surpasses the curve for $w=3$, against the intuition that narrowing the previous speakers' window should normally yield worse or in the best case equal prediction results. This discrepancy lies on the fact that we excluded all meetings with four or five speakers from the analysis for $w \geq 3$ and at the same time in part of the excluded meetings the success probability was over 94% for window size $w = 2$.

TABLE III. – Average probability of successful next user prediction, variability and percentiles over all meeting events and for three distinct history sizes (2, 3 and 4).

Probabilité moyenne de prédiction avec succès du prochain orateur, calculée sur toutes les réunions pour trois tailles d'historique (2, 3 et 4).

	Mean Success Probability	**Std. Dev.**	**10% Percentile**	**15% Percentile**	**25% Percentile**
History Size $w=2$	70.45%	17.07%	50.80%	53.28%	57.95%
History Size $w=3$	84.53%	11.39%	75.12%	76.80%	78.36%
History Size $w=4$	91.50%	8.71%	84.86%	86.64%	88.68%

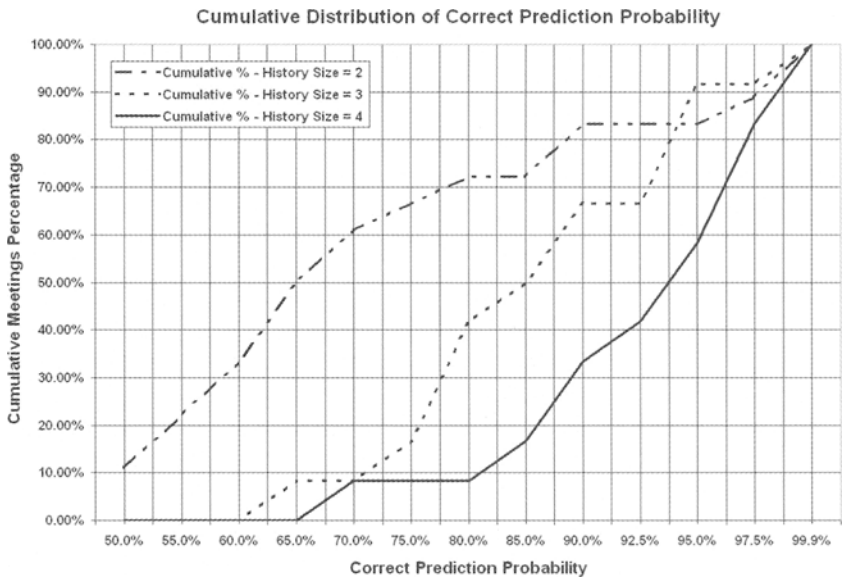

FIG 5. – Cumulative percentage of meetings against correct prediction probability for three interaction history sizes. For a given prediction probability p (x-axis), the respective value on the y-axis shows the fraction of meetings with successful prediction probability less or equal than p.

Pourcentage cumulé de réunions en fonction de la probabilité correcte de prédiction. En abscisse, la probabilité de prédiction, en ordonnée le taux de réunions avec une probabilité de prédiction réussie inférieure ou égale à p.

VI.2. Lectures

Unlike meeting-like discourse, in educational lectures only a small fraction of the participants speaks throughout the session, while the rest attends passively. The histogram in Figure 6 quantifies this phenomenon. Not unexpectedly, the ratio of speakers to group size is

below 10% in 60% of the analyzed lecture sessions, whereas only 15% of the lectures triggered the active participation of more than 50% of the attendees.

Likewise, the domination of the lecturer in terms of speaking time is also an expected characteristic of lecture discourse. Still, it is unclear for those cases, where the active participation of students is high, whether particular students dominate the total student speaking time. Figure 7 illustrates the ratio of speaking time to lecture duration of: a) the lecturer, b) the first and second most talkative students and c) the total student contribution to speaking time. The first and most obvious observation is the clear domination of the lecturer (with one exception in lecture 16, where the tutor's speaking time was under 60%). As for the existence of a student dominating total student time, this is hardly evident in the results. Even in the few cases, where the ratio of the most talkative student was a large fraction of the total student time, this does not justify for an application or service to differentiate this particular student from the rest of the students. For, his absolute speaking time is still insignificant compared to total session duration. This is particularly true in 16 out of 20 lectures (or 80%) of our dataset.

Due to low degree of interactivity – as indicated by the dominance results presented above – the amount of lecture time exhibiting temporal locality was very small compared to total event duration. Thus, applying speaker prediction to services supporting lecture discourse would be overkill, for such an optimization would be infrequently used. For this reason, we did not further study speaker prediction in the case of lecture interaction.

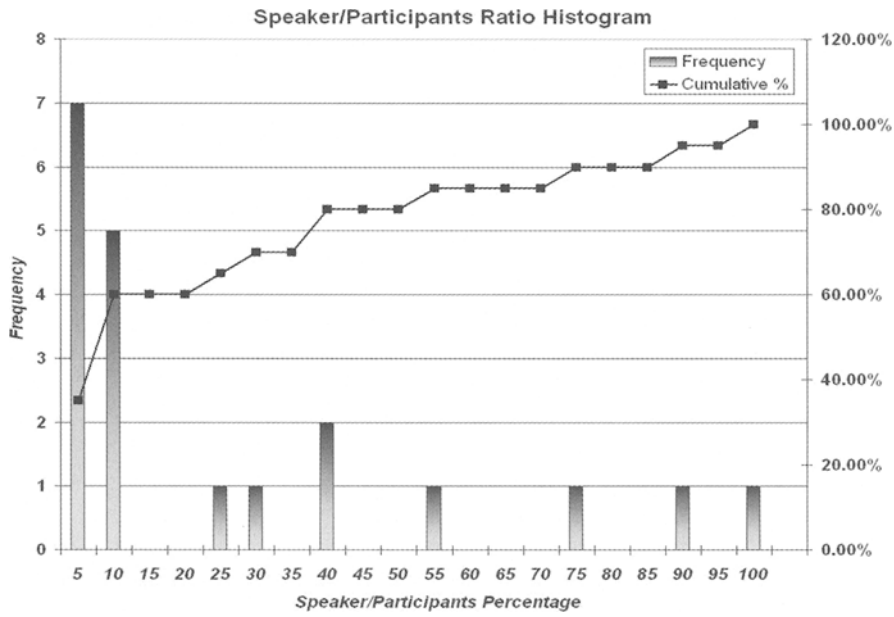

FIG 6. – Fraction of lecture participants, who spoke at least once, to total number of participants. The frequencies show clearly that only a small fraction of participants talked.

Taux de participants ayant pris au moins une fois la parole. On voit nettement que seule une petite partie des participants s'est exprimée.

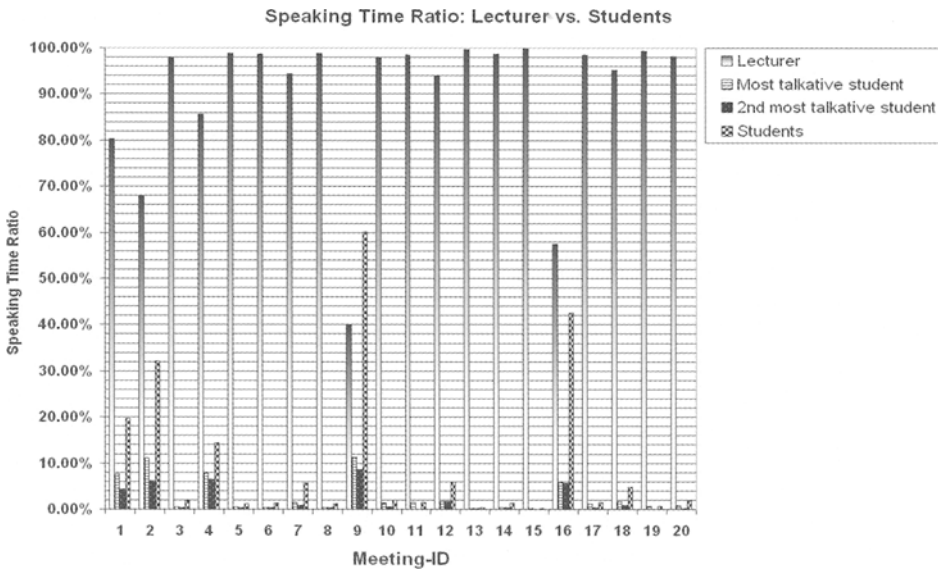

FIG 7. – In only 3 out of the 20 lecture sessions was the speaking time of students comparable to the lecturer's speaking time.

Le temps de parole des étudiants n'est comparable au temps de parole des conférenciers que dans seulement 3 cas sur 20.

VII. CONCLUSION

The results of this study confirmed the claim that natural multiparty human interaction (i.e. interaction without explicit role assignment to speakers) exhibits high temporal locality. Even more, the analysis outcome showed that for varying group sizes the speaker taking a new turn is with high probability one of the last four spoken participants. This finding strengthens the validity of previous designs that employed future speaker prediction as a means of optimizing group communication. In general, it motivates future services to bias design optimization towards locality in turn-taking.

Furthermore, our results indicated that speaking time is not equally allocated to all discourse parties, but instead speakers that monopolize the word do exist. This was clearly evident in the results of lectures' analysis, where a single dominating speaker (the lecturer) outweighed by far the students' activity. The same was true for meeting-like interaction, however here the speaking time mismatch between the most talkative speaker and the rest of the speakers was not as significant as in the case of lectures. As such, it remains unclear, whether the benefit of differentiating the group communication service offered to meeting participants according to speaking time would sufficiently exceed the cost of realizing this differentiation. On other hand, this is definitely true for lecture sessions.

Note that the scope of our results is strictly limited to the domain, in which we studied human interaction. Herein, we exclusively analyzed discourse in the academic community,

where speakers have a certain profile and act in a specific environment. Generalizing the validity of our findings in other fields of social life – e.g. business meetings or political debates – requires further analysis of related interaction traces and surely constitutes an interesting extension to the contributions of this paper. Moreover, former studies [12] have shown that interaction behavior may differ depending on language spoken and ethnical characteristics. Hence, the interpretation of our results applies primarily to Native American English speakers, as manifested by the profile of all speakers in our dataset.

In general, we see our work strengthening the argument that there is value in incorporating user-behavior awareness into group communication service design. For services that are realized in the network layer (e.g. Source-Specific Multicast), this implies shifting part of the functionality to the application layer, where the exploitation of user-behavior characteristics becomes possible. In our future work, we plan to evaluate the efficiency of further previous window management (or cache management) strategies complementary to least recently spoken participants. Additionally, extending the scope of the present analysis to other domains – primarily to business discourse – forms an interesting topic of the related research agenda.

Acknowledgements

The authors would like to kindly thank Thomas D_bendorfer, Karoly Farkas and Placi Flury for providing valuable feedback on the entire paper. Also, we are grateful to Nick Blundell at Lancaster University for sharing his thoughts on pattern similarity between face-to-face and online discourse.

Manuscrit reçu le 30 juin 2005
Accepté le 9 mars 2006

REFERENCES

[1] SILBERSCHATZ (A.), GALVIN (P.), GAGNE (G.), Applied Operating System Concepts, *Wiley*, 1999.

[2] PARMENTELAT (T.), BARZA (L.), TURLETTI (T.), DABBOUS (W.), A Scalable SSM-based Multicast Communication Layer for Multimedia Networked Virtual Environments", *Technical Report RR-5389*, INRIA, Sophia Antipolis, November 2004.

[3] HOLBROOK (H.), CAIN (B.), Source-specific Multicast for IP, Internet Draft (work in progress), *Internet Engineering Task Force*, 2003.

[4] SACKS (H.), SCHEGLOFF (E.), JEFFERSON (G.), Simplest Systematics for the Organization of Turn-taking for Conversation, *Language*, **50**, n° 4, pp. 696-735, 1974.

[5] POTTER (J.), Discursive Psychology: Between Method and Paradigm, *Discourse Society*, **14**, n° 6, pp. 783-794, 2003.

[6] HAMMERSLEY (M.), "Conversation Analysis and Discourse Analysis: Methods or Paradigms?", *Discourse Society*, **14**, n° 6, pp. 751-781, 2003.

[7] MCKINLAY (A.), ARNOTT (J.), A Study Of Turn-Taking In A Computer-Supported Group Task, Proceedings of the *Eighth Conference of the British Computer Society, Human Computer Interaction Specialist Group, People and Computers VIII (HCI '93)*, Loughborough, UK, 1993.

[8] BLUNDELL (N.), MATHY (L.), Minimizing Perceived Latency in Audio-Conferencing Systems over Application-Level Multicast, *Second International Workshop on Multimedia Interactive Protocols and Systems (MIPS 2004)*, Grenoble, France, November 16-19, 2004, pp. 1 – 12.

[9] KATRINIS (K.), BRYNJÚLFSSON (B.), HJÁLMTÝSSON (G.), PLATTNER (B.), Dynamic Adaptation of Source Specific Distribution Trees for Multiparty Teleconferencing, to appear in *CONEXT 2005*, Toulouse, France, October 24-27, 2005.

[10] HOLBROOK (H.), CHERITON (D.), IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications, *ACM SIGCOMM Computer Communication Review*, **29**, n° 4, pp. 65-78, Oct. 1999.

[11] "MICASE: Michigan Corpus of Academic Spoken English", http://www.hti.umich.edu/m/micase/ [online source], 2002.

[12] WEILHAMMER (K.), RABOLD (S.), Durational aspects in Turn Taking, Proceedings of the *International Conference of Phonetic Sciences*, Barcelona, Spain, 2003.

[13] TEN BOSCH (L.), OOSTDIJK (N.), DE RUITER (J. P.), Durational Aspects of Turn-taking in Spontaneous Face-to-Face and Telephone Dialogues, Proceedings of *Conference of Text, Speech and Dialogue*, Brno, September, 2004.

[14] FRANCE (E.), ANDERSON (A.), GARDNER (M), The Impact of Status and Audio Conferencing Technology on Business Meetings, *International Journal of Human-Computer Studies*, **54**, n° 6, pp. 857 – 876, 2001.