Fixing Number of Floors for Virtual Voice-Only Conference – an Empirical Study

R. Venkatesha Prasad*, H. N. Shankar†, Przemysław Pawełczak* and H. S. Jamadagni‡
*Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, Mekelweg 4, 2600 GA Delft, The Netherlands
Email: {vprasad, p.pawelczak}@ewi.tudelft.nl
‡Center for Electronics Design and Technology, Indian Institute of Science, Bangalore - 560012, India
†Department of Telecom Engineering, PES Institute of Technology, Bangalore - 560085, India

***Abstract*— For efficient Computer Supported Cooperative Work (CSCW) audio conferencing is an essential component where video and text are add-ons. The specifications for enabling CSCW over Internet are incomplete if they are blind to actual conduct of participants. Indeed, a blind conference mimics quite closely a virtual voice-only conference. In this paper, we analyze the results of sessions of face-to-face blind conversations and gain penetrating insights. In particular, we focus on the impact of users' behavior on the design of a scalable architecture for virtual voice-only conferencing over VoIP and arrive at a meaningful number of *floors* for such conferences. We also present the features and the requirements for the proposed service.**

***Index Terms*— Computer Supported Cooperative Work, Number of Floors, Virtual Audio Conferencing Architecture and Design, Human Computer Interface.**

I. INTRODUCTION

As the world is becoming more and more networked, the traditional circuit switched services are increasingly being ported onto the Internet. The realtime audio is an important basic medium that draws attention for its low cost and effective connectivity across the globe (see for example Skype [2] or Yahoo Voice Chat [3]). While there is so much activity on the Internet catering to communications between individuals, applications for collaborative work is progressing steadily. We have seen over the years some important applications that marked a niche in collaborative games and distributed virtual environments. They are are Timewarp [5], MASSIVE [6], DIVE [7], JASMINE [37], and many groupware applications such as GroupWeb and GroupKit [8]. Presently, Computer Mediated Collaborative (CMC) platforms are the most sought after applications for many corporate members and individuals, who are working in different time zones. It not only allows integration of many data types but also saves load on the exchequer [9]. For all these developments Internet has become a ubiquitous facilitator.

Audio conferencing is a bare essential block of today's collaborative work – video and text are add-ons. Hitherto this service was available only on PSTN using conference bridges. Providing a similar service on the Internet has been an area of intensive research and development in recent times. Some standardization in this direction has been done by IETF's SIP [18] and, ITU-T's H.323 [19]. Many early applications that addressed these problems are RAT [20], VAT [21].

We broadly divide many issues concerning the service into two groups *technical* and *functional*. Technical issues – mainly related to the underlying network – are bandwidth, delay, delay jitter, loss concealment [22], [23], [24], [25], [26], conference control [27] and multicasting support [28]. We can see that significantly enhanced bandwidth, powerful systems for desktop conferencing, group authoring, and distributed design for audio conferencing driving the fundamental changes in casual and formal business interactions among participants of modern society.

Functional issues pertain to operational and maintenance aspects of a CMC application related to interactions between users and applications. It concerns with behavioral aspects and other important issues such as ease of use, comfort levels of users during usage, the effects of delay and/or jitter, etc. [29], [30], [31], [32]. Though there are many facets for the functional issues being addressed by the multimedia community, evaluation of techniques for novel communication environments has by far fallen short of aggressive technological advancements. Usually it is *more-the-merrier* attitude towards bandwidth, facilities and complex control mechanisms. However this has aptly been criticized as "keeping form the before function" [4].

Therefore the concern here is to build a Computer Supported Cooperative Work (CSCW) application that supports audio conference mimicking acceptably closely a face-to-face real-life conference by effective utilization of available bandwidth. The specifications for this problem are definitely incomplete if they are blind to actual conduct of cultured[1] participants. In this paper we study empirically one of the aspect of users' behaviour in a conference. We do not intend to redo the study on conversational analysis [35], but we only reinforce, further by our experimentation, the advantage of such a study on the architecture for virtual audio conferencing on Internet. We leverage the results of the analysis of our experiments of a blind conference to gain penetrating insights into building a virtual audio conferencing tool over VoIP. We demonstrate the significance of users' behaviour in designing a scalable architecture.

[1] In the sense that participants are striving to make a sense out of their interactions during a conference.

Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM'05)
0-7695-2489-3/05 $20.00 © 2005 IEEE

First we try to find some of the requirements of a virtual voice conferencing over Internet. Then with the help of a study of many blind conferences, we try to infer an important facet of such conferences. We briefly study how it can be effectively used in building a tool. The core of the paper is on our experiments and its influence on architecture, We feel this is one of the examples on how technical and functional aspects could be bridged to design a meaningful CSCW application.

The rest of this paper is organized as follows. In Section II we list important requirements and studies related to CSCW. In Section III we motivate our work. In Section IV we present the results of our experiment along with statistical analysis. In Section V we discuss impact of chosen number of floors for CSCW application and propose our architecture for virtual voice-only conferencing. Finally in Section VI we discuss our results and conclude the paper.

II. Requirements and Related Studies

A virtual audio conferencing over Internet throws many challenging problems. Many of them have been addressed already by various studies [16], [17], [22], [23]. However, we feel that there are many issues and requirements that have been overlooked, which can make a virtual conferencing more realistic and useful. Here we list a few of them.

- In an audio conferencing over Internet, it is desirable that impromptu speech is enabled so that participants rather than waiting in the queue for their turn to address the gathering, can interact without inhibitions. Enabling immediate feedback by intercepting the current speaker and argumentative conversation with the person who is addressing presently are essential ingredients of a successful CSCW application.
- With more than one concurrent speaker, mixing of audio streams prior to play out is essential. Thus a CSCW tool that caters audio conferencing should take care of mixing of different streams without losing the spatialism (the ability to discern who is speaking and what is spoken). Many applications avoid this by having a control protocol in place which allows participants to take turns one-by-one with a scheduling algorithm as in JASMINE [37], and Interactive Remote Instruction (IRI) [11] but losing out on spontaneity. Therefore we don't want to trek on the same path.
- With multiple speakers allowed to address the conference, it is but imperative that streams from only a small set of selected speakers be broadcast/multicast to all participants as in a many-to-all fashion. Mixing arbitrarily many streams makes the resultant speech unintelligible [14].
- Giving users the ability to determine the weights for individual streams before mixing at their terminals results in increasing quality of mixed speech stream since, they can boost the volume of speaker(s) whom they want hear loudly.
- A simple and dynamic rule or metric for selection of speakers to address in case of contention if the total number of streams selected for mixing are limited.

With these broad requirements we shall have a look at some of the existing applications, tools and studies. Since we use the term *floor* frequently, we define it here as a virtual token that is necessarily be acquired by a participant to be able to address everyone in a conference.

Floor allocation in IRI [11] is unnatural as it is on a first-come first-serve basis and with holding time depending on how many are already waiting in the queue. It asks users to first request a floor and wait for their turns, thus losing out on interactivity. JASMINE [37], has a tight floor control with only one participant speaking at an instance and is selected on first-come first-served basis. MASSIVE and DIVE [6] implicitly assume that arbitrarily several participants with their *aura* (a virtual boundary around each participant in a virtual 3D space) colliding is able to interact with each other. They mix streams from all participants who are in the vicinity. Redenkovic *et al.* [16] and Rangan *et al.* [17], on the other hand, allow everyone to speak in a conference using mixing architectures. Redenkovic proposes to judiciously mix a subset of streams and keep some streams as it is depending on the available bandwidth using distributed partial mixing. The partial mixing proposed in [16] primarily selects all M participants to be transported separately, if sufficient bandwidth is available. When bandwidth is limited a subset of M streams are mixed. Thus, this technique uses unnecessarily high bandwidth and when streams are mixed it loses spatialism. Rangan [17] proposes to have a set of hierarchical mixing entities arranged in the form of a tree that mix each stream they get from their children and pass it to their parents. At every stage single stream is formed. Thus at the root of the tree only one stream would be formed which is broadcasted to all the nodes in the tree. The idea here is to allow everyone who tries to speak to be heard by one and all. Though their technique improves interactivity and reduces bandwidth it simultaneously reduces spatialism of the mixed speech and in turn makes mixed audio unintelligible[2].

III. Motivation

The number of floors N (i.e. allowing N concurrent speakers out of M participants) is an important parameter in the design of applications since it influences many aspects of collaborative tools. It governs how a floor control need to be designed and it also influences the design of applications if there are more than one floor.

Standards for such collaborative application development mention general guidelines for architecture designs, protocol, etc. but do not explicate on the value of N to be chosen. IETF's SIP [18] or ITU-T's H.323 [19] does not specify N or even a bound thereon to N. While H.323 leaves it to collaborative tool designers, SIP doesn't even admit an instance where it may be useful to find it. Architectures designed independent of fixing and managing N necessarily serve little point or purpose. Moreover, to mimic closely a face-to-face voice conference

[2]Digitally mixing of even two streams reduces intelligibility of mixed speech and it becomes almost noise like when more streams are mixed [14].

allowing multiple simultaneous speakers and a user controlled mix of the selected voice streams are a must. Any system with tight control gags the participants due to loss of spontaneity [16], which has been identified as *process loss phenomenon* in verbal brainstorming [36]. Furthermore the ideas can be lost in group conversation as a result of single voice controlling the floor, i.e. *productivity loss* [38]. Therefore interactivity and user-friendliness are at a premium for any CSCW tool.

Clearly we need to understand human interactions in a conference before we set forth towards designing a conferencing tool. It is clear from the above discussion that N can not be too large which will then reduces intelligibility for example complete mixing [17]. Moreover it should not be as low as 1 that calls for a floor control. Therefore it is important to find an optimal number of floors, which we call N_{max}, that can provide interactivity and spatialism, as well as keeping it to a minimum compared to complete mixing architectures [17], [16]. In this paper we present the results of an experiment to imply how to find this value, and briefly its impact on the design of application architectures taking off from previous studies [14].

IV. EXPERIMENTAL ESTIMATION OF N_{max}

Turn-taking is one of the most important aspect in any social interaction with which humans interact with others meaningfully. Using turn-taking mechanisms they organize to take turns to speak to others in group conversations this has been thoroughly studied by Sacks *et al.* [35]. In the present context of the virtual voice-only conferencing turn-taking is nothing but accessing the floor, i.e. a participant getting the rights to address all the participants. As we have discussed in earlier sections we do not intend to *explicitly* control the user behaviour by notifying who should address at a given instance of time. We strive only to *facilitate* instinctive conversation with a natural turn-taking in a CSCW tool.

Some work in this regard from ethnomethodological perspective has been done earlier [13], [14]. We take a completely different approach here when we try to fix N_{max} by analyzing a few conferences that lacks visual aid. We conducted a simple experiment, which is to some extent equivalent to voice-only (blind) conference. As we can see later, these experiments reported here strongly influence the selection of N_{max}. We observe how participants behave in a conference, which to a certain extent mimicked a CSCW environment. In principle we specifically wanted to simulate conferences that inherently encouraged many concurrent speakers contributing to more interruptions and more number of turn-taking without any inhibition. However, during meetings in a corporate world, floor occupancy is usually minimal and is decided by many aspects such as hierarchy, leader, perceived authority of a person over a topic, etc. Therefore, we organized the experiment that emulated extemporized conferences – wherein each participant can actively take part in an ad-libbed way – rather than well planned meetings. We now furnish relevant details of some aspects of our experiments briefly, before presenting the data for analysis.

A. Experiment setup

We took the help of ten persons, four women and six men from a similar age, group and social status, who volunteered to take part in a set of discussion sessions. Volunteers knew each other well. This aspect influenced them to express themselves freely which increased the possibility of higher overlapped speech. Each discussion was recorded on different days and involved not less than six participants. Moreover, no two experiments involved the same subset of participants. The volunteers were seated around in a conference hall. They were not informed *a priori* how the experiment will be performed, to make sure that knowledge of the aim of experiments will not influence the outcome. After about 10 minutes of grounding time [40] they were asked to continue their talk but with an opaque eye shields counteracting eye contacts with other participants. Participants have chosen various topics of discussions – varying from cultural (i.e. "Current trends in movies"), through social (i.e. "Career options: entrepreneur or employee"), political (i.e. "Modern terrorism") towards philosophical ones (i.e. "Mind, meditation and happiness"). Not all of these subjects were given to participants. After some time of starting discussions they broadly got evolved into a major topic of discussion. Some of them becoming a hotly debated discussions and one of them was steered by a participant who assumed leadership. The end to end duration of conducted experiments varied between 15 and 60 minutes.

All the volunteers were unaware of the fact that their conversations were recorded. This was to minimize the influence of that knowledge on the conduct of participants during the discussions. However, we informed them at the end of the experiments about the recording and how it will be used further. The person who was recording the discussion participated in it only in the beginning of the experiments. However, he withdrew slowly so that his influence did not affect the course of the discussion. Person recording was also silent throughout the rest of the discussions. Recording was done after participants have lost eye contacts and the person who was recording became silent.

B. Data processing and statistical analysis

We have recorded each of the discussions in a WAVE format. From each of ten recordings we cropped fragments of around five minutes length, excluding beginning, when grouding was performed, and the end when participants were leaving the room. Recordings were later post-processed with Goldwave package [12]. Length of overlapped speech were fixed by repeatedly listening to a smaller segments and correlating it with the visual aid in Goldwave. We identify a set of overlapped and non-overlapped speech samples $D = \{D_0^i, D_1^i, D_2^i, D_3^i, D_4^i\}$, where for recording i:

1) D_0^i – silence duration,
2) D_1^i – total duration when one participant was speaking,
3) D_2^i – total duration of two concurrent participants speaking,
4) D_3^i – total duration of three participants speaking concurrently,

TABLE I

MEASURED DURATIONS OF EACH ELEMENT OF SET D FOR $i = 10$ RECORDINGS. DATA GIVEN IN SECONDS.

i	D^i_0	D^i_1	D^i_2	D^i_3	D^i_4	D^i
1	85.786	181.087	12.254	2.291	1.264	282.682
2	40.156	212.721	12.984	2.159	0.000	268.020
3	78.785	191.521	17.251	1.325	2.020	290.902
4	33.514	233.828	22.632	1.951	0.000	291.925
5	76.740	145.340	19.700	1.870	0.000	243.650
6	104.450	150.450	23.560	1.950	6.910	287.320
7	56.890	234.870	15.890	1.784	1.820	311.254
8	68.870	196.380	15.670	2.490	0.000	283.410
9	73.040	218.630	17.760	2.270	3.050	314.750
10	52.870	220.540	21.840	3.460	4.500	303.210

TABLE II

ESTIMATES OF MEAN μ_j AND VARIANCE σ_j OF SAMPLES' LENGTH IN SET D_j . DATA GIVEN IN SECONDS.

	μ'_j	σ'_j
D_0	66.9660	21.4580
D_1	199.9420	41.9875
D_2	17.9830	4.2429
D_3	2.1620	0.6303
D_4	2.0230	2.3544

5) D^i_4 – total duration of when more than three participants speaking (and/or laughing).

The accuracy of measured durations were in the order of milliseconds. Results of the post processing are presented in Table I. We have to emphasize that D^i – total duration of recording i – was marginally different for each i since we took a complete discussion segment. Mean value of recording durations D^i (see Table I) is $\overline{D^i} = 287.71$ seconds.

Next we have performed a set of statistical hypothesis tests to determine whether the length of gathered samples D_j, $j = 0 \ldots 4$ for all i recordings follow the normal distribution. This is to statistically make some decisions. Since we do not know exactly the first two moments of the underlying distribution we performed Lilliefors test [33]. It is to test null hypothesis H_0 that each element of set D_j is normally distributed. Because duration of each recording was marginally different we first have to normalize samples $D^i_j, \forall i, j$ in the Table I by $D^i_j \frac{D^i}{\overline{D^i}}$. We have also performed sample standardization according to $\widetilde{D}^j_i = \frac{D^j_i - \mu_j}{\sigma_j}$ where μ_j and σ_j are mean and standard deviation of D_j, respectively. $\widetilde{D}^j_i$ are used in the Lilliefors hypothesis test. We found that at we cannot reject the aforementioned null hypothesis H_0 at confidence levels 95%, 97.5% and 99.5%. Now armed with this result we assume that all samples are normally distributed and we perform estimation of first two moments, μ'_j and σ'_j, of distribution of non-standardized samples D_j, according to [34]. Results of the estimations are presented in Table II that shows the normalized mean and variance of duration of speech for each D_j.

C. Analysis

After data processing and statistical analysis we outline our observations here.

- Silence has occupied no less than 11.48% with a mean of 23.28% of total discussion time since participants took time to co-ordinate themselves, which can increase with a CSCW tool possibly to allow for catching up with network and computational delays. It must be emphasized, however, that such result mainly depends on participants.
- Predominantly, (minimum of 52.36% and mean of 69.49%) only one participant was speaking. It implies that the turn-taking were effective in these experiments and any intelligible conference will have to have a repair mechanisms so that the turn-taking effectively allows one person to address.
- Two simultaneous streams amount to nearly the rest of the time (maximum of 8.09% and mean of 6.25%).
- A third concurrent speaker comes into picture for no more than 1.14% of the conversation. Moreover this phenomenon was observed when two participants got into fierce argument or when participants tried to get the attention of others during heated debate.
- More than three simultaneous streams are effectively nonexistent. This is pivotal in choosing the number of floors N_{max}.
- One of the recording was of the nature of briefing by a leader. It comprised no significant change in the speech amplitude. This was possibly due to the fact that interruptions were few and far apart. Apparently, auto restraining had played its part.
- Some recordings were of debates in nature. Speakers occupied floor for shorter durations. It contained several interruptions by one participant, as also significantly many interruptions by two participants. Voice modulation was more pronounced and it was used for interruption with telling effect. Silences were infrequent but for relatively longer durations. Bursts after silence were more the rule than exception.
- Speech intelligibility deteriorated very sharply with the third participant getting into addressing mode in any conversation. Naturally, this has led to one or more of them retracting quickly.

D. Fixing N_{max}

It is clearly seen from Table II that mostly one or two participants were talking during these experiments even without explicit floor control. Three and more participants occupy statistically less than 4 seconds of duration. Thus we may conjecture that $N_{max} = 2$ is sufficient. However, during a voice-only conference CSCW setup and when participants are geographically apart, different delays between them result in more collisions. To allow the feedback of these collisions that are detrimental to facilitate turn-taking we increase the threshold of N_{max} from two to three. Hence we conjecture that in a blind (voice-only) conference three simultaneous speakers are sufficient for interactivity. Recently, we had arrived at this result in a slightly different manner applying conversational analysis rigorously for all its intents and purposes. Thus our work presented here validates the result in [13].

After fixing N_{max}, the next step is to choose speakers out of $M(\gg N_{max})$ speakers possibly vying for floors. This needs to be done dynamically, smoothly and in realtime without participants being able to notice it. In [14] this issue was settled by employing a pseudo-metric termed *Loudness Number* (LN). It quantified the following attributes of speakers:

1) current loudness of a speech,
2) loudness of a speech in the recent past duration,
3) level of activity in a reasonably past duration.

Therefore LN has been defined as a function of the current energy of the audio stream, energy in the past and level of activity for some duration. A persistent participant would also get a chance even the loudness is lower. It is identified that there are some participants who shy away from competing for a floor [39]. We can always provide an option for participants by reserving one floor out of three, which may be allotted using techniques proposed in [11] and [37].

V. Impact of $N_{max} = 3$ on the Design of Architectures for CSCW

As we have stated earlier Rangan *et al.* [17] have proposed a tree mixing architecture wherein each node is a mixer and leaves are the terminals of participants. Each node mixes speech from each leaf nodes or from child mixers and passes it on to the root. Using $N_{max} = 3$ in their architecture, each mixer can select N_{max} streams from its child nodes and pass them to its parent without mixing. With this sort of filtering at each level, one can avoid mixing so many streams and losing the intelligibility of the mixed speech. Further the spatialism can be held intact. In the case of *Distributed Partial Mixing* (DPM) proposed by Redenkovic *et al.* a similar method can be employed. That every DPM would not mix but it filters the speech and sends it to the next stage. One major difference we propose here compared to DPM is that the number of streams ultimately mixed is going to be constant. This avoids continuous change in the volume level of speakers in the mixed speech. To avoid bandwidth overshooting one can filter some streams if their LN is very low for a considerable amount of time. Therefore in both these approaches only N_{max} streams are transmitted back to all participants' terminals, otherwise which would have been M or single but completely mixed stream.

A. Proposed architecture

In the above two architectures the total delay between the speaker and the listener terminals depends on the number of nodes or DPMs connected serially. This determines the interactivity of the conversation. To avoid more delay, to provide customized mixing and to allow spatialism we present a new architecture with *Conference Servers (CS)* – see Figure 1. CSs are equivalent to mixers of the above architectures. However, CSs do not mix streams – they select N_{max} out of all the clients assigned. Thereafter CSs exchanges these N_{max} streams amongst themselves to select the N_{max} streams to be returned to clients in their respective domains. CSs may communicate with each other on unicast/multicast depending on the network. We can think of CSs forming an application level multicasting nodes (an overlay network). Since speech streams cross only two CSs, the delay is limited to transit time and queuing time at two nodes. The decision to select speakers is based on LN [14] at every packet instant (at most every 40 miliseconds). Thus it is dynamic enough to invoke the speaker change on the fly. Mixing is done at the user terminals so that users can set weights for mixing these N_{max} streams according to their preferences. It can be noted here that the number of streams exchanged here would always be constant and user interface designs are going to be simpler in this case compared to [16]. We can also harness the property of LN and fewer change in speakers – when the participants are managing themselves – to reduce the number of streams exchanged across CSs [15].

B. Informal Evaluation

We have realized a prototype of this new architecture with $N = 3$ on a campus-wide network. In the context of network delay, since perception of a second speaker is not immediate, it is required to investigate into the sufficiency of two conference servers in the path of an audio stream. However we have some observations from our implementation.

- $N_{max} = 3$ obviates explicit floor control and hence paves the way for a speaker to gracefully relinquish the floor or acquire the floor.
- $N_{max} = 3$ implies enabling impromptu speech to offer feedback or conveying the intention to interrupt.
- As mixing is done at each participants' terminal customized mix can be achieved so that spatialism of the mixed speech is maintained.
- Participants felt overwhelmingly that this method is more interactive than applications where they had to request for a floor before they could speak.

When at each node only N_{max} streams are selected scalability is achieved. Nevertheless it is less than what can be accomplished from other mixing architectures [17], [16] but at the cost of losing the interactivity of the conversation.

VI. Discussion and Conclusions

A. Discussion – Are we treading a 'soft' path?

Though there has been much discussion within HCI community about using inferences based on ethnomethodology, some question the much-hyped use of it. Graham Button [1] opines, "... ethnomethodologically affiliated studies have produced a strong critique of the design of technology at work for ... technology, at best, often fails to support the work it is designed for, or at worst, does not allow people to actually engage in their work, because the technology is not aligned to the practices through which they organize their actions, interactions and work". Heath et al. [10] observe, "There are relatively few examples of successful applications in real world ... the lack of success of CSCW systems derives ... more from their insensitivity to ... real work environments". Thus it is imperative that we use some of the real world situations and

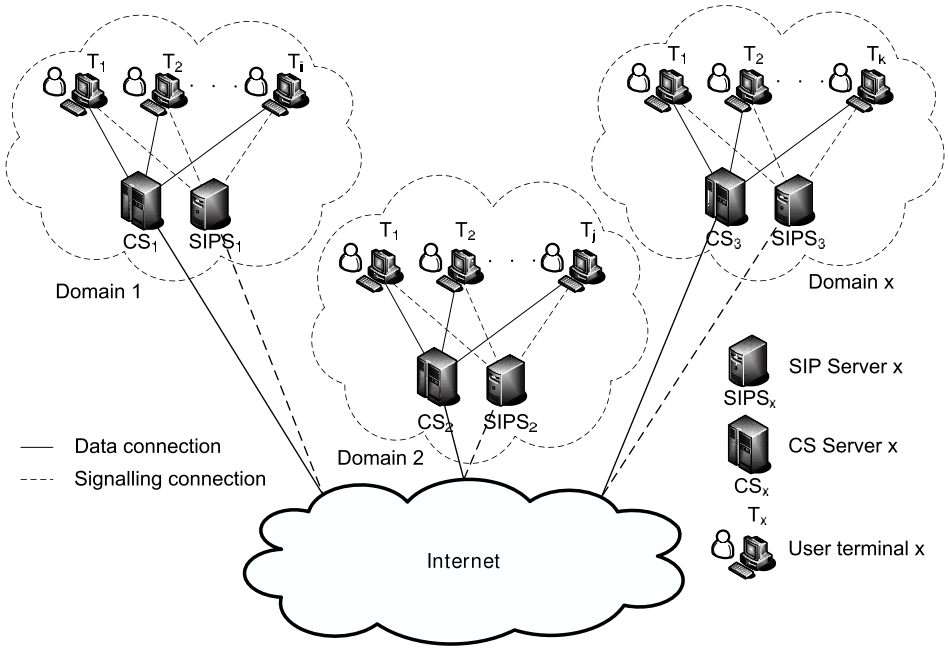

Fig. 1. Proposed architecture for CSCW over Internet.

participants' behaviour before we start making an effort at building a CSCW application.

We argue as follows. In HCI studies it is mandatory to have insights purely from the users' and sociological perspective. There is no denying that the need to converse is basically social – technical or otherwise. It is worthwhile to have an in-depth knowledge that helps in design of CSCW application. Unlike in a face-to-face conference, in a blind conference (with no visual clues) it hardly matters whether it is three or more participants. Functional aspects are important and must be taken into account before designing the application. The very approach here is therefore novel compared to the previous studies such as [16], [17], [37].

Being on the network, introduction of end-to-end delay would pose some problems [41]. We have discounted the effect of delay here and presumed that the applications would strive to minimize end-to-end delay. Nonetheless delay merits attention for any realtime application. Further, we make no pretext that the data presented here and the methodology of the experiments followed here can provide a complete evidence for the fact that the three floors are sufficient. The methodology of experiments and analysis presented here pertaining to ad hoc dicussions/meetings, to some extent it doesn't represent the well defined conferences. However, by considering informal conferences for our study we, in a way, have studied conferences that are difficult to conduct, where the turn-takings are higher than in formal well defined conferences or meetings.

B. Conclusions

Audio conferences are an important component of almost any CSCW system. Voice-only conferences are easy to support due to their reduced bandwidth requirements. We have studied blind conferences to find some insights into the behaviour of participants when they lack visual clues. We confirmed statistically that predominantly, no more than two participants speak for most of the time, even in a face-to-face conversation. It would therefore be unnecessary to provide for more than three floors in a virtual audio conference over VoIP. With three floors as the ceiling, the impact on the choice of the supporting application architecture are improved scalability and simplicity of the architecture which is eminently viable. However, there are some questions that need to be answered: If video clues are involved, is explicit floor control required? What then is an appropriate N_{max}?

More fundamentally, intrinsic analysis of blind conferences without spatialism, vibrational cues and auditory and with geographical apart is planned for the next phase. Moreover, more number of trials with heterogeneous participants is ongoing. We believe more work in this direction is fruitful for successful CSCW tools.

References

[1] G. Button and P. Dourish, "Technomethodology: Paradoxes and Possibilities," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 19-26, Vancouwer, British Columbia, Canada, 13-18 April 1996.

[2] http://www.skype.com.

[3] http://vc.yahoo.com.

[4] E. Doerry, "Evaluating Distributed Environments Based on Communicative Efficacy," *Proceedings of Conference on Human Factors in Computing Systems*, pp. 47-48, Denver, Colorado, USA, 7-11 May 1995.

[5] W. K. Edwards and E. D. Mynatt, "Timewarp: techniques for autonomous collaboration," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 218-255, Atlanta, Georgia, USA, 22-27 March 1997.

[6] C. Greenhalgh and S. Benford, "MASSIVE: A Collaborative Virtual Environment for Teleconferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 2, pp. 239-261, September 1995.

[7] E. Frcon, C. Greenhalgh and M. Stenius, "The DiveBone an application-level network architecture for Internet-based CVEs,", Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp.58-65, London, United Kingdom, 20-22 December 1999.

[8] http://grouplab.cpsc.ucalgary.ca/projects.

[9] http://www.imcca.org/wpcomingofage.asp.

[10] C. Heath, M. Jirotka, R. Luff and J. Hindmarch, "Unpacking Collaboration: The Interactional Organisation of Trading in a City Dealing Room," *Journal of Computer Supported Cooperative Work (CSCW)*, Vol. 3, pp. 147-165, Springer Science, June 1994.

[11] K. Maly, C. M. Overstreet, A. Gonzlez, M. Ireland and N. Karunaratne, "Experiences with Structured Recording and Replay in Interactive Remote Instruction," *Second International Conference on New Learning Technologies*, University of Bern, Bern, Switzerland, 30-31 August 1999. Available at http://www.cs.odu.edu/~iri-h/publications/conference_at_Berne.pdf.

[12] http://www.goldwave.com.

[13] R. V. Prasad, "A New Paradigm for Audio Conferencing on Voice over IP (VoIP)," PhD Thesis, Center for Electronics Design and Technology, Indian Institute of Science, Bangalore, India, 2003.

[14] R.V. Prasad, H.S. Jamadagni and H. N. Shankar, "Number of Floors for a Voice-Only Conference on Packet Networks – A Conjecture," *IEE Communications Proceedings*, Vol. 151, No. 3, pp. 287- 291, 25 June 2004.

[15] R.V. Prasad, R. Hurni, H. S. Jamadagni and H. N. Shankar, "Deployment Issues of a VoIP Conferencing System in a Virtual Conferencing Environment," *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pp. 150-159, Osaka, Japan, 1-3 October 2003.

[16] M. Redenkovic, C. Greenhalgh and S. Benford, "Deployment Issues for Multi-User Audio Support in CVEs," *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pp. 179-185, Hong Kong, China, 11-13 November 2002.

[17] P.V. Rangan, H. M. Vin and S. Ramanathan, "Communication Architectures and Algorithms for Media Mixing in Multimedia Conferences," *IEEE/ACM Transactions on Networking*, Vol. 1, No. 1, pp. 20-30, February 1993.

[18] J. Rosenberg , H. Schulzrinne, G. Camarillo , A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, "SIP: Session Initiation Protocol," *Request For Comments 3261*, Internet Engineering Task Force, June 2002.

[19] ITU-T Recomendation, "H.323: Packet based Multimedia Communications Systems". Available at http://www.itu.int/itudoc/itu-t/rec/h/.

[20] http://www-mice.cs.ucl.ac.uk/multimedia/software/rat.

[21] http://www-nrg.ee.lbl.gov/vat.

[22] J.-C. Bolot and A. Vega-Garca, "Control Mechanisms for Packet Audio in the Internet", *Proceedings of the IEEE INFOCOM'96*, vol. 1, pp. 232-239, San Francisco, CA, USA, 24-28 March 1996.

[23] S. B. Moon, J. Kurose and D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms" *Multimedia Systems*, vol. 6, no. 1, pp. 17-28, Springer Science, February 1998.

[24] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, May 1991.

[25] R. Ramjee, J. Kurose, D. Towsley and H. Schulzrinne, "Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks," *Proceedings of the IEEE INFOCOM'94*, vol. 2, pp. 680-688, Toronto, Canada, 12-16 June 1994.

[26] Y. J. Liang, N. Farber and B. Girod, "Adaptive Playout Sceduling and Loss Concealment for Voice Communication over IP Networks," *IEEE Transactions on Multimedia*, vol. 5, no. 4, December 2003.

[27] O. Levin, R. Even and P. Hagendorf, "Centralized Conference Data Model," *Internet Draft draft-levin-xcon-cccp-02*, Internet Engineering Task Force, 20 February 2005.

[28] D. Thaler, M. Handley and D. Estrin, "The Internet Multicast Address Allocation Architecture", *Request For Comments 2908*, Internet Engineering Task Force, September 2000.

[29] T. Henderson, "Latency and User Behaviour on a Multiplayer Game Server," *Proceedings of the Third International COST264 Workshop on Networked Group Communication*, pp. 1-13, London, UK, 7-9 November 2001.

[30] P. T. Brady, "Effects of Transmission Delay on Conversational Behaviour on Echo-Free Telephone Circuits", *Bell System Technical Journal*, pp. 115-134, January 1971.

[31] M. Muhlhauser, "Content Development for the Internet as a Mass Medium," *Proceedings of International Workshop on Multimedia Software Engineering*, pp. 2-9, Kyoto, Japan, 20-21 April 1998.

[32] J. Encarnacao and J. Foley, (editors), "Multimedia,", Springer-Verlag, Berlin, Germany, 1994.

[33] H. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", *Journal of the American Statistical Association*, vol. 62, pp. 399-402, 1967.

[34] M. Evans, N. Hastings, and B. Peacock, "Statistical Distributions," John Wiley & Sons, Second Edition, pp. 170, 1993.

[35] H. Sacks, E.A. Schegloff and G. Jefferson, "A Simplest Systematics for the Organization of Turn-taking for Conversations", *Language*, vol. 50, pp. 696-735, December 1974.

[36] D. Shaw, "Evaluating Group Support Systems: Improving Brainstorming Research Methodology," Report RP0203, Aston Business School, Aston University, March 2002.

[37] S. Shirmohammadi, A. El-Saddik, N. D. Georganas and R. Steinmetz, "JASMINE: A Java Tool for Multimedia Collaboration on the Internet," *Journal of Multimedia Tools and Application*, vol. 19, pp. 5-28, Springer Science, January 2003.

[38] Stroebe, W and, Diehl, M. "Productivity Loss in Idea Generating Groups: Tracking Down the Blocking Effect", *Journal of Personality and Social Psychology*, vol. 61, no. 3, pp. 392-403, September 1991.

[39] E. Doerry, (personal communication).

[40] S. B. Brennan, "The Grounding Problem in Conversations With and Through Computers," *Social and Cognitive Psychological Approaches to Interpersonal Communication*, pp. 201-225, Lawrence Erlbaum Associates, 1998.

[41] Karen Ruhleder and Brigitte Jordan, "Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction", *Computer Supported Cooperative Work*, vol. 10, no. 1, pp. 113-138, Springer Science, March 2001.