Number of floors for a voice-only conference on packet networks – a conjecture

R.V. Prasad, H.S. Jamadagni and H.N. Shankar

Abstract: Voice conferencing is an essential block of any multimedia system used for collaborative work, as voice is shared by all participants. Floor control is mission-critical here and has been investigated by many to ensure fair resource sharing; yet fixing the number of floors has remained an open problem. A conferee (participant in a conference) can speak only after acquiring the floor. To allow impromptu speech, floor allocation must be made for many concurrent speakers. However, too many concurrent speakers degrade voice intelligibility. Therefore, setting an upper bound for the number of streams (floors) that may be mixed is *sine qua non* for quality conferencing. The problem of setting an upper bound on the number of floors to support concurrent multi-party audio sessions is addressed. A conjecture based on conversational and qualitative analysis is proposed. A pseudo-measure termed 'loudness number' used to manage the number of floors is briefly outlined. The implementation at a functional level on Windows© systems has yielded satisfactory performance.

1 Introduction

In today's shrinking world, collaborative work is centre stage in business, research and many maintenance activities. The main facilitators are computers and, with the advent of communication networks, the Internet. Such collaborative work is termed 'computer-supported co-operative work (CSCW)'. In this context audio-visual conferencing has several advantages [1]. Packet networks such as the Internet have their intrinsic advantages and limitations in transporting different types of traffic such as data, audio and video. We assume that the underlying network provides sufficient bandwidth for the application. The concern here, consequently, is not as specific to utilising the necessary bandwidth as it is to designing the application over the network that provides sufficient interactivity. We seek to build a CSCW application that supports audioconference mimicking acceptably closely a face-to-face dialogue involving several speakers. It may not be required to broadcast the laughter of every conferee in a moment of humour. In audioconferencing over the Internet, it is desirable for audio from 'selected' speakers to be broadcast/multicast to all conferees.

With improved network connectivity and bandwidth, exploring ways to support collaborative interaction between distributed participants has become a hot topic of research. Increasingly powerful systems for desktop conferencing, group authoring and distributed design have the potential to fundamentally change the way participants of modern society interact with each other, in both casual and formal business contexts.

IEE Proceedings online no. 20040282

doi:10.1049/ip-com:20040282

Paper first received 3rd April and in revised form 15th October 2003

R.V. Prasad and H.S. Jamadagni are with the Centre for Electronics Design and Technology, Indian Institute of Science, Sir C V Raman Avenue, Bangalore 560012, India

H.N. Shankar is with the Department of Telecommunication Engineering, PES Institute of Technology, Banashankari III Stage, Bangalore, and National Institute of Advanced Studies, Indian Institute of Science Campus, Bangalore, India

Recently, technology development has outpaced synthesis (even identification) of features for characterising and evaluating novel communication environments. Most efforts focus on bandwidth sharing, client–server designs, fair control etc. Resource-wise they adopt a more-the-merrier stance. It is assumed that more complex control mechanisms lead to better quality of interaction. These approaches are deaf to the functional utility of an audioconference. Doerry [2] criticises this as 'keeping *form* before *function*.' It is therefore imperative to investigate the usefulness of conversational analysis as an integral functional aspect of audioconferencing.

A conference facilitates real-time interactive data/voice/video transmission from 'active' members to all conferees. Service here is 'many-to-all', like a face-to-face physical conference, wherein at a given time, possibly more than one speaker may address all conferees.

2 Motivation

The ITU-T standard H.323 [3] mentions the selection of N concurrent speakers out of M participants but it does not set the value of N. This is left to the application developers [4, 5]. IETF's SIP [6] defines no standard architecture and control for conference service although it provides flexibility to its application developers.

There are detailed studies [7–9] on floor control for CSCW. Significantly, however, there has been no attempt to specify N. Perhaps this is due to the universal tacit assumption that only one person speaks at a time. An external stringent control deprives spontaneity and 'gags' the participants [10]. Then, even a fair and intelligent setup can be blamed for poor quality of support. Permitting all participants without any control increases interaction in networked gaming. Obviously, mixing many streams results in loss of spatialism. In a conference, speech intelligibility is compromised.

Thus, there is a pressing need to specify N. It must be more than one to enable impromptu speech. Indiscriminately large N raises issues such as bandwidth requirement

and computations at mixers and servers. In contrast to postponing the inevitable, this exercise will guide application designers [4, 5] to tread a well defined path. Moreover, fixing N paves the way for designing scalable distributed conference architectures [11].

3 Problem formulation

Let Ω be the set of all conferees and $M = |\Omega|$ (the cardinality of Ω) the number of conferees. A 'floor' is a virtual platform (as in any shared system) that a conferee must necessarily occupy to be permitted to transmit. At any given time, $S \subseteq \Omega$ is the maximum possible set of conferees provided with the permission or token to access the floor.

With small M, $S = \Omega$, or $|S| = M$, may be feasible when the service is ported on packet networks for CSCW. That is, every conferee has a token. However, as M becomes large, typically tens/hundreds, it is not pragmatic (in fact, not even necessary among well behaved conferees!) to have $|S| = M$. Hence the number of tokens $N = |S| < M$. ('Number of tokens' is synonymous with 'number of floors'.) This admits two scenarios. If S is static, i.e. the conferees in S are time-invariant, then tokens are not transferable among conferees. If S is dynamic, as when the conferees in S are time-variant, then not every conferee has a token, but tokens are transferable among conferees. This calls for floor control.

In this setting, we address a dyad of issues in this paper:

- first, specifying N, the number of floors;
- secondly, managing the N floors to ensure fair floor sharing when there is a conflict.

Before venturing into addressing this aspect of the problem, we must understand the underlying concerns that dictate the acceptability of a solution.

4 Solutions

In a voice-only conference, $|S|$ is the number of concurrent audio channels or floors. $N = 1$ yields best speech quality. Then any floor control severely constrains the conferees; the virtual conference falls short of getting acceptably close to the audio aspect of a face-to-face conference. Alternatively, if the conferees were to somehow adapt themselves to ensure that there is at most one speaker at any time, the conversation becomes doctored and unnatural.

We build up the strategy here onwards on the following findings of Sacks *et al.* [12] regarding behavioural aspects of conversations.

(i) overwhelmingly, one party talks at a time

(ii) occurrences of more than one person speaking at a time are common but brief

(iii) transitions (from one turn to another) with no gap and no overlap are common; together with transitions characterised by slight gap or slight overlap, they make up the majority of transitions (See Fig. 1, similar to [8])

(iv) turn order and size are not fixed

(v) talk can be continuous or discontinuous

(vi) repair mechanisms [13] exist for dealing with turn-taking errors; if two parties find themselves talking at the same time one of them will stop prematurely.

We use the above observations and many research findings by Schlegloff *et al.* [13] on conversational analysis gainfully in the remainder of the paper.

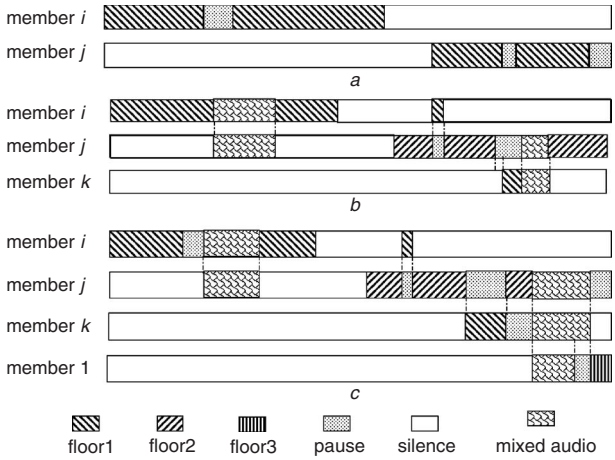

Fig. 1 *Examples of turn-taking by conferees*
a $N = 1$
b $N = 2$
c $N = 3$

Now we may state a simple proposition. We have used formal terms (like 'proposition') though they may sound a trifle contrived at times. The formalism is merely to assist the presentation of arguments as it is evolved based on the above discussions on conversational analysis (a social phenomenon), which is developed within the discipline of ethnomethodology.

Proposition 1: In a voice-only conference, $N = 1$ (*a*) is necessary; (*b*) is desirable; (*c*) is insufficient.

Proof:

Part (*a*): is trivially true.

Part (*b*): Desirability stems from goodness of speech quality as remarked above. Though utopian, it is indeed desirable that the conferees conduct themselves so that no two of them will speak concurrently (Fig. 1*a*).

Part (*c*): Investigations into conversational psychology [12, 14] and turn-taking repair mechanisms [13] have been reported. Providing for interruptions will render the conference closer to a natural face-to-face conference. Evidently, an interruption cannot be registered unless we provide for at least two simultaneous speech streams (Fig. 1*b*).

4.1 Mixing of audio streams

Clearly, mixing of audio streams is necessary for a conference. As a prelude to further debating the concerns for specifying N, it is necessary to take a look *en passant* at the mechanism of mixing.

With multiple active audio sources, the sound or pressure wave incident on the human ear is a sum of the individual pressure waves [15]. To quote Guyton and Hall [16], 'In the case of sound, the interpreted sensation changes approximately in proportion to the cube root of the actual sound intensity'. With multiple speakers in the same room, the signal captured by a microphone would be a linear combination of the signals. To get this effect with speakers in different locations and generating audio packets, the mixed stream should be the sum of the generated streams. If $X_i(j)$ is the jth linear sample of the ith audio stream, then the jth linear sample of the mixed stream is given by

$$S(j) = \sum_{i=1...N} W_i X_i(j) \qquad (1)$$

where W_i is the weight for the ith stream. Equation (1) forms a basis for a generic mixing algorithm [15]. As the amplitude of $S(j)$ cannot be increased beyond a certain level

(limited by the supply voltage to the sound card of the computer), invariably W_i are chosen to be in the open simplex, i.e. in (0,1) and adding up to unity. Thus clamping is precluded. Unbiased or fair mixing demands $W_i = W_j$ for all i, j.

For large values of N fair floor control is trivial; also the number of conferees awaiting floor access reduces. However, if too many conferees are concurrently accessing the floor, then it is possible that each weight is rendered so small as to result in deterioration of speech resolution. Thus, there is a strong case for specifying an upper bound for N. We must strike a balance between conflicting criteria by fixing the least possible value for N with a fairly good quality of performance. This is an issue to be probed in some depth.

When there are many active members in a conference there can be simultaneous speech streams. Usually, the human brain can neither comprehend nor even register more than one nontrivial concurrent speech stream. We do not delve into psycho-acoustic impact of mixing two or more streams but argue with an appraisal that the conference is 'well behaved' and fair. We do not intend to provide explicit 'grant floor' (GF) messages for the conferees [9] to speak. Explicit GF messaging precludes impromptu speech and disfavours natural interactions. With too many floors, GF messages may be dispensed with; however, intelligibility is compromised at the cost of liveliness. Hence we must find the smallest $N>1$ for the purpose. The number of floors thus fixed, together with loudness number (Section 5), should serve the purpose.

We require the following definitions for further discussion:

Definition 1: *Pause* is absence of a conferee's voice activity for duration of at most τ

Definition 2: *Silence* is absence of a conferee's voice activity for duration greater than τ

Definition 3: $T+$ is a *transition* of a conferee's state from silence to speech

Definition 4: $T-$ is a *transition* of a conferee's state from speech to silence

Lemma 1: $N=2$ is (*a*) necessary; (*b*) insufficient.

Proof:

Part (*a*): follows from parts (*a*) and (*c*) of proposition 1.

Part (*b*): A second token was made available to permit a conferee j to interrupt a speaker i, $i \neq j$. It is possible that both i and j are not silent thereafter. In a conference of well mannered participants this will not pose a problem as either i or j undergo $T-$. Else, the possibility of such prolonged or frequent occurrences cannot be simply ignored. Then, the conference would become impolite and messy in the absence of an intervention by a third conferee. Thus $N=2$ is insufficient.

Lemma 1 highlights the need for specifying $N>2$. In the process, the discussion on mixing must be borne in mind. It is but natural to ask at this stage whether any higher value of N will suffice. Before seeking an answer to this question, it is instructive to find out how the participants feel as the number of speech streams mixed exceeds one.

4.2 Qualitative study

Listeners in our study are English-speaking postgraduate students and faculty of CEDT, IISc, chosen at random. In this study a set of five male-and female-spoken sentences of approximate duration 10 s were taken from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. The transcript of one of the speech samples (referred to as a marked sample) was made available to the listeners. From these samples, four mixed samples were formed. The first mixed sample was a mix of the marked speech sample and one other sample speech from the set. The second mixed sample comprised a mix of the marked speech and two other speech samples from the set. Likewise the third and fourth mixed samples were mixes of the marked speech with three and four other samples, respectively, from the set. In all, a collection of eight mixed samples was synthesised for male and female voices.

First, the marked speech sample was played to the listeners. They were given the transcript of the marked speech. It was played on request repeatedly to ensure that the listeners were tuned in to the marked speech stream. This is necessary since in a conference the listeners can track a certain speaker's voice if the speaker has been active for some protracted period.

The listeners' task was to quantify the effort required to resolve the marked sample in the mixed streams. They graded the mixed speech from A (easily recognisable) to E (unrecognisable). The majority results for ten listeners are in Table 1. These results decimate the popular view that normal listeners can resolve speech despite voice stream mixing. With the male speaker's voice marked, when mixed with one other sample, the majority of listeners said that 'considerable effort' (grade C) went into latching on to the marked speaker's voice. Some reported that they could identify with 'none too great an effort'. Thus the overall response was between B and C. When mixed with two other voices, a majority said C; the others put it as D thus giving an overall response between C and D. When mixed with three and four other samples, the listeners unanimously reported inability even to identify the marked speech.

Table 1: Subjective quality of mixed speech streams

Speech sample	Marked speech file with X = no. of other speech files mixed			
	$X=1$	$X=2$	$X=3$	$X=4$
Male	C	C	E	E
Female	B	C	E	E

A: with no extra effort; B: with none too great an effort; C: with a considerable effort; D: very difficult; E: cannot recognise

Corresponding results with female voices (Table 1) indicates better intelligibility in some cases and no improvement in evaluation in the remaining cases. Improved resolution for female speech in the first case is perhaps due to the higher pitch frequency of the female voice relative to the male voice.

Thus for voice-only conferencing the duration of speech overlap reduces further. Moreover, there is a tendency for *repair* [13] in the event of three or more concurrent speakers. That is, in this mixed speech case our study bears very clear testimony to the hypothesis that the speech of any one of the participants is unclear!

4.3 The conjecture

This Section leads from the foregoing findings to the heart of this discussion. We have seen that $N=2$ is insufficient at times. The qualitative studies implied that more than three would render the mixed speech unintelligible. These implications are encapsulated in a conjecture here.

Conjecture: A maximum of three floors (i.e. $N=3$) are necessary and sufficient for a voice-only conference over packet networks.

Since our aim was to mimic a face-to-face blind conference as acceptably closely as possible, it is reasonably imperative to impose some etiquette in the rare event of the conference getting messy as above. We permit a third conferee to undergo $T+$ but restrict three floors to be for no longer than a duration of Γ. Here, Γ must exceed pause τ. Actually, the speakers tend to retract (self-control, see 'repairs' in [13]) as the playout becomes less intelligible. However, control may even be imposed. If the number of floors is forcibly reduced from three to two and it remains at two thereafter, when will the third floor be made available next? The requirement of a delay (lower-bounded by Γ) to allow the third floor after it has been disabled, has been recognised. Space constraints preclude this discussion in this paper. Which of the three conferees concurrently holding the floors will be forced to undergo $T-$? An answer to this will be based on 'loudness number' to be formulated in the following Section.

5 Loudness number (λ) for floor control

Setting N will not completely meet the requirements of hands-free conferencing without computer mediation. As already mentioned, one of the issues is: 'How are N floors allocated to $C(C>N)$ conferees competing for N floors?' This must be addressed in any conference [17]. In the context of video, for instance, it has been remarked [18] that '…..it was not obvious how to determine which sounds from the audience were appropriate to transmit'. So it is mandatory to resolve the conflict among speakers trying to access the limited number of floors. One way would be to rank packets from C conferees in a mixing interval by their energies, and choose the top N. This has been found to be inadequate at times because randomness in packet energies can lead to poor audio quality. For example, a noise burst in a listener's environment causes transient spikes in packet energy, and the packet is chosen amongst the N instead of a legitimate speaker's packet. This indicates the need for a quantifier different from the one based exclusively on current packet energies. It must have the following characteristics:

- A speaker currently holding the floor should not be cut off by a spike in the packet energy of another speaker. This suggests that a conferee's speech history should be given some weight. This is often referred to as 'persistence' or 'hangover'.
- A participant who wants to interrupt a speaker will have to (i) speak loudly and (ii) keep trying for a little while. In a face-to-face conference, body language can often indicate intent to interrupt. In a blind conference, however, a participant's intention to interrupt must be reflected unambiguously in the quantifier.
- A floor control mechanism empowered to cut off a speaker forcefully must be ensured.

We define, for each participant, a pseudo-metric called 'loudness number' λ which adapts slowly so that floor allocation is graceful. λ depends on the energies of the present and past packets. Among C conferees those with top N ranked λ have floor access.
Current activity L_1 (refer Fig. 2) of a conferee is computed within a recent past window W_{RP}

$$L_1 = \frac{1}{W_{RP}} \sum_{K=t_P}^{t_P-W_{RP}+1} X_K \tag{2}$$

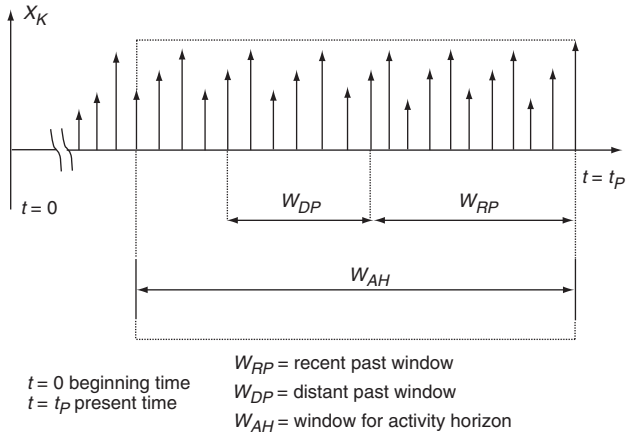

Fig. 2 *Windows for loudness number calculation*

where X_K is the r.m.s. of the samples in the packet.
Distant past activity L_2 of a speaker is over a distant past window W_{DP}

$$L_2 = \frac{1}{W_{DP}} \sum_{K=t_P-W_{RP}}^{t_P-W_{RP}-W_{DP}+1} X_K \tag{3}$$

Overall past activity L_3 of a speaker is spread over an activity horizon W_{AH}

$$L_3 = \frac{1}{W_{AH}} \sum_{K=t_P}^{t_P-W_{AH}+1} \theta I_{\{X_K>\theta\}} \tag{4}$$

where

$$I_{\{X_K>\theta\}} = 1 \text{ if } X_K > \theta$$
$$= 0, \text{ else}$$

The threshold θ is a constant and is the same for all conferees. We have set θ at 10–20% of maximum packet energy. Now the current loudness number λ_{tp} is given by

$$\lambda_{tp} = \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_3 \tag{5}$$

Here, $0<\alpha_1, \alpha_2, \alpha_3<1$ and $\alpha_1+\alpha_2+\alpha_3=1$. By appropriate choice of windows, α_1, α_2, α_3 and θ, λ can be tuned to smoothly provide or withdraw floor access.

5.1 Safety, liveliness and fairness

The parameter λ has some memory depending on the spread of the windows. After one conferee becomes silent, another can take the floor. Also, as there is more than one channel, interruption is enabled. A loud conferee is more likely to be heard because of elevated λ. This ensures fairness to all conferees. After all, even in a face-to-face conference, a more vocal speaker grabs special attention. All these desirable characteristics are embedded into λ. A discussion on how these parameters are selected and the dynamics of loudness number are beyond the scope of this paper.

6 Results and conclusions

We presented an argument to find an upper bound for the value of $N(=|S|)$. We tested our audioconferencing tool for $N=2,3,4$ on our test-bed [4, 5] with ten participants. We found the performance to be characterised by smooth turn-taking. Window sizes, α_1, α_2, α_3 and θ influence the complex dynamics of the system, and they help in fine-tuning the performance. After a limited survey of the perceptions of conferees on our test-bed and

with heuristics, we used $W_{RP} = 5$ s, $W_{DP} = 10$ s, $W_{AH} = 30$ s and $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.3, 0.3)$. $\tau = 660$ ms [19] and $\Gamma \approx 5\tau - 10\tau$ are typical.

We tested all the above proposals on our conferencing test bed [5, 20] and observed that the quality of conference was very close to a face-to-face conference for up to ten participants. Our preliminary studies, which are not very formal, lend credence to the values set for various parameters as above.

An important byproduct of setting $N = 3$ is the reduction of bandwidth in a distributed conference [5, 20] due to filtering of packets based on loudness number. This scheme may be extended to audio in a videoconference by including the pseudo-metric based on motion vectors. Voice activity detection (VAD) algorithms [21] can be used along with the present tool for enhanced performance.

This discussion does not consider limitations of the network support. Delay [22] introduced by the network may hamper smooth floor management. We know by experience that users gradually adapt to the effect of delays. Long-distance satellite calls are a case in point. Nonetheless delay merits attention for any real-time application.

6.1 Retrospection

A criticism of work such as this is the subjective and qualitative components of performance assessment. Comparison of perceived quality of service with existing conference solutions with controls is but one component. Allowing for multiple speech streams, to interrupt the speakers and limiting the number, thereby enhancing the quality of conference is the second aspect when we compare. Current literature assumes all streams to be mixed [10]. We have argued that it is not necessary. Specifying N, taken with loudness number, serves a larger purpose as in designing distributed architectures.

The label 'conjecture' for the main result here is not to convey a certain vagueness in the claim. It reflects (i) the use of hitherto unconventional and qualitative assessments of performance, and yet (ii) adopting conventional modes of analysis and rationale.

Some might suspect the forceful use of conversational analysis. Indeed, conversational analysis can be used exclusively to specify N. Conversational analysis in CSCW is used to tackle design issues for human interface with computers. There is no denying that the need to converse is basically sociological, technical or otherwise. Functional aspects are important and must be taken into account in designing the application.

While we do claim to have taken a couple of steps in the right direction, we make no pretext of having spoken the last word on the matter.

7 Acknowledgments

The authors thank the anonymous referees for their constructive criticisms that have helped them in refining this paper.

8 References

1 Silverman, L.R.: 'Coming of age: conferencing solutions cut corporate costs'. White Paper, Interactive Multimedia Collaborative Communications Alliance, http://www.imcca.org
2 Doerry, E.: 'Mosaic of creativity'. Proc. ACM SIGCHI'95, Denver, CO, USA, 7–11 May 1995, pp. 47–48
3 ITU-T Rec. H.323, 'Packet based multimedia communications systems',Vol. 2, 1998, http://www.itu.int/itudoc/itu-t/rec/h/h323.html
4 Prasad, R.V., Kuri, J., Jamadagni, H.S., Dagale, H., and Ravindranath, R.: 'Automatic addition and deletion of clients in VoIP conferencing'. Proc. 6th IEEE Symp. on Computers and Communications. Hammamet, Tunisia, July 2001, pp. 386–390
5 Prasad, R.V., Kuri, J., Jamadagni, H.S., Dagale, H., and Ravindranath, R.: 'Control protocol for VoIP audio conferencing support'. Proc. Int. Conf. on Advanced Communication Technology, Mu-Ju, South Korea, Feb 2001, pp. 419–424
6 Rosenberg, J., and Schulzrinne, H.: 'SIP: session initiation protocol', RFC 3261, IETF, Jun. 2002, ftp://ftp.isi.edu/in-notes/rfc3261.txt
7 González, A.J.: 'A distributed audio conferencing system'. MS Project, Department of Computer Science, Old Dominion University, Norfolk, VA 23529, 28 July 1997
8 Dommel, H.-P., and Garcia-luna-aceves, J.J.: 'Floor control for multimedia conferencing and collaboration', *Multimedia Syst.*, 1997, **5**, (1), pp. 23–28
9 Dommel H.-P., and Garcia-luna-aceves, J.J.: 'Network support for turn-taking in multimedia collaboration'. Proc. IS&T/SPIE Symp. on Electronic Imaging: Multimedia Computing and Networking, San Jose, CA, USA, Feb 1997
10 Radenkovic, M., and Greenhalgh, C.: 'Multi-party distributed audio service with TCP fairness'. Proc. ACM Int. Conf. on Multimedia, Juan-les-Pins, France, 2002, pp. 11–20
11 Prasad, R.V., Hurni, R., and Jamadagni, H.S.: 'A scalable distributed VoIP conferencing using SIP'. Proc. 8th IEEE Symp. on Computers and Communication, Antalaya, Turkey, 2003
12 Sacks, H., Schegloff, E.A., and Jefferson, G.: 'A simplest systematics for the organization of turn-taking for conversations', *Lang.*, 1974, **50**, (4), pp. 696–735
13 Schlegloff, E.A., Jefferson, G., and Sacks, H.: 'The preference for self-correction in the organization of repair in conversation', *Lang.*, 1977, **53**, (2), pp. 361–382
14 Isaacs, E., and Clark, H.H.: 'References in conversation between experts and novices', *J. Exp. Psychol., Gen.*, 1987, **116**, (1), pp. 26–37
15 González, A.J., and Abdel-Wahab, H.: 'Audio mixing for interactive multimedia communications'. Proc. JCIS'98, Research Triangle, NC, USA, Oct. 98, pp. 217–220
16 Guyton, A.C., and Hall, J.E.: 'Textbook of medical physiology' (W B Saunders Co., USA, 1996, 9th edn.)
17 Isaacs, E., and Tang, J.C.: 'What video can and cannot do for collaboration', *Multimedia Syst.*, 1994, **2**, pp. 63–73
18 Isaacs, E., Morris, T., and Rodriguez, K.: 'A forum for supporting interactive presentations to distributed audiences'. Proc. ACM. Conf. on Computer-Supported Cooperative Work (CSCW), Chapel Hill, NC, USA, 1994, pp. 405–416
19 Jaffe, J., and Feldstein, S.: 'Rhythms of dialogue' (Academic Press, New York, USA, 1970)
20 Prasad, R.V.: 'A new paradigm for audio conferencing on voice over IP (VoIP)'. PhD thesis, Indian Institute of Science
21 Prasad, R.V., Sangwan, A., Jamadagni, H.S., Chiranth, M.C., Sah, R., and Gaurav, V.: 'Comparison of voice activity detection algorithms for VoIP'. Proc. IEEE Symp. on Computer and Communications, Taormini-Giandini Naxos, Italy, July 2002, pp. 530–535
22 Ruhleder, K., and Jordan, B.: 'Co-constructing non-mutual realities: delay-generated trouble in distributed interaction', *Comput. Support. Coop. Work*, 2001, **10**, (1), pp. 113–138