On the Problem of Specifying the Number of Floors for a Voice-Only Conference on Packet Networks

R. Venkatesha Prasad, H. S. Jamadagni
CEDT, Indian Institute of Science,
Bangalore, INDIA
{Vprasad, hsjam}@cedt.iisc.ernet.in

H. N. Shankar
P. E. S. Institute of Technology
& National Institute of Advanced Studies,
Bangalore, INDIA

Abstract **--- Voice Conferencing is an essential block of any multimedia system used for collaborative work. In a collaborative environment *Floor Control* is an important issue that is dealt by many; yet fixing the number of *floors* is an open problem. In an audio conference, mixing streams from too many concurrent speakers degrades the voice quality. Therefore setting an upper bound for the number of streams (*floors*) that may be mixed is *sine qua non* for providing quality conferencing service. In this paper we address the problem of setting the upper bound on number of floors for a system meant to support concurrent multi-party audio sessions on top of IP multicasting. A measure called "Loudness Number" that is used to manage the number of floors is briefly outlined to the extent of making this paper self-contained. Our implementation at a functional level on a campus-wide network of Windows© systems has yielded satisfactory performance.**

Index Terms – VoIP, Conference, Floors, Loudness Number

I. INTRODUCTION

In today's shrinking world, collaborative work is occupying a center stage in business, research and many maintenance activities. The main facilitators are computers and, with the advent of communication networks, Internet. Such collaborative work is termed Computer Supported Cooperative Work (CSCW). In this context audio-visual conferencing has several advantages [13]. Packet networks such as Internet have their intrinsic advantages and limitations in transporting different types of traffic like 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 *utilizing* the necessary bandwidth as it is to *designing* the application over the network that provides sufficient bandwidth. We seek to build a CSCW application that supports audio conference that mimics *acceptably closely* a face-to-face audio conference. It may not be required to broadcast the laughter from every conferee in the event of a joke. In audio conferencing over Internet, it is desirable that audio from "selected" speakers be broadcast/multicast to all conferees (participants of a conference).

A conference facilitates real-time interactive data/voice/video transmission from "active" members to all conferees. Service here may be classified as:

- One-To-All (OTA), like for instance, in a radio broadcast;
- Many-To-All (MTA), for example, like in a concert, wherein possibly more than one speaker is permitted, at a time, to address all conferees;
- Some-To-All (STA), wherein a select few conferees are privileged to address; i.e., the rest are receive-/listener-only-parties; and
- All-To-All (ATA), where the maximum number of concurrent speakers equals the number of conferees.

In STA and MTA providing for different speakers to speak is mutually exclusively in time [1]. Clearly, the audio conference studied here falls into MTA.

In this paper we venture into Conversational Analysis [15, 16] which is primarily a domain of social scientists. The argument here is based on the findings of Conversational Analysts. However, we present our results in a formal framework. They may seem anecdotal at times. Hence we term our final result here a "conjecture".

II. PROBLEM FORMULATION

Let Ω be the set of all conferees. Then $M = |\Omega|$ (the cardinality of Ω) is the number of conferees. We define *Floor* as a virtual platform (as in any shared system) which a conferee must necessarily occupy to have the permission to transmit. $S \subseteq \Omega$ is the set of conferees provided with the permission or token to access the floor.

With M small, $S = \Omega$, or $|S| = M$, may be feasible when the service is ported on packet networks for a 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, $N = |S| < M$. (The "number of

0-7803-7724-9/03/$17.00 © 2003 IEEE

tokens" is also referred to as the "number of floors".) This admits two scenarios. If S is static, which is the case when the conferees in S are time-invariant, the implications are that not every conferee has a token and tokens are not transferable among conferees. On the other hand, if S is dynamic, which is the case when the conferees in S are time-variant, the implications are that not every conferee has a token, but the tokens are transferable among conferees. This calls for floor control.

We address a dyad of issues in this paper:

- First, specifying N, the number of floors.
- The second is managing the N floors to ensure "fair" floor sharing when there is a conflict.

The ITU-T standard H.323 [11] mentions about selection of N out of M members but it does not set the value of N. The onus is on the application developers [17, 18]. SIP [22] does not propose any standard approach for a conferencing service and assumes that User Agents would implement it with SDP.

Though there are many detailed studies [5, 3] on floor control for a CSCW there has been no attempt to specify N. Our attempt here is to specify N for a voice-only conference. Fixing N would in fact help in defining conference architectures, floor control, etc,.

Before we venture into addressing this aspect of the problem, there is a need to understand the underlying concerns that dictate the acceptability of a solution.

III. THE SOLUTIONS

In a *voice-only conference,* $|S|$ is the number of audio channels referred to as *floors* in the context of audio conference. If $N = 1$, best speech quality would be achieved. Then any floor control will make the conferees too constrained and, in this sense, the conference itself will fail to 'mimic acceptably closely the audio aspect of a face-to-face conference'. Alternatively, if the conferees somehow adapt to ensure that there is no more than one speaker at any time, the conversation may be a little unnatural.

We shall build up the strategy from here onwards on the findings of Sacks et al. [15] regarding properties of conversations. Their findings are: (i) Overwhelmingly, one party talks at a time; (ii) Occurrences of more than one person speaking at a time is common but brief; (iii) Transitions (from one turn to a next) with no gap and no overlap are common. Together with transition characterized by slight gap or slight overlap, they make up the majority of transitions (See Fig.1, similar to [5]); (iv) Turn order and size are not fixed; (v) Talk can be continuous or discontinuous; and (vi) Repair mechanisms [16] 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 [16] on conversational analysis, gainfully, in the sequel. Now we may state a simple proposition. We used formal terms (like proposition) for expressing the solution though it looks *contrived* at times. The formalism is ***only*** to assist the presentation of arguments towards a solution as it is evolved based on the above discussions on conversational analysis (a social phenomenon), which is developed within the discipline called *Ethnomethodology.*

Proposition: *In a voice-only conference, N=1 (a) is necessary; (b) is desirable; and (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 [7, 15] and turn taking repair mechanisms [16] 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)). ■

A. Mixing of Audio Streams

Clearly, mixing of audio streams is necessary for a conference. As a prelude to further debating the concerns that dictate the specification of the number of audio channels, N, it is necessary to take a cursory look 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 [2]. Gyton and Hall [10] say: '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 $X_i(j)$ be the j^{th} linear sample of the i^{th} audio stream, then the j^{th} linear sample of the mixed stream is given by

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

where W_i is the weight for the i^{th} stream. Eqn. (1) forms a basis for a generic mixing algorithm [2]. As the amplitude of

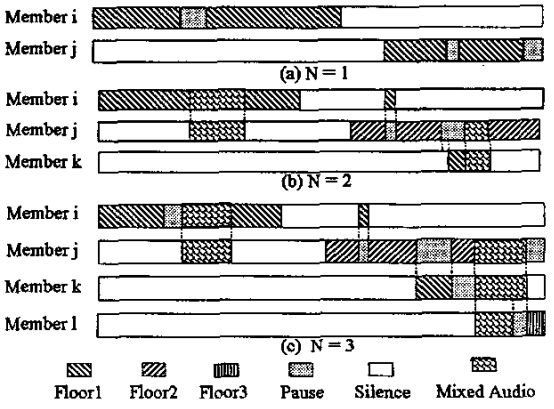

Figure. 1. Examples of Turn Taking by Conferees

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 interval (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 *N* fair floor control is trivial; also the number of conferees waiting for floor 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 requires probing in some depth.

Let us take a look at some parameters of performance. When there are many active members in a conference there can be simultaneous speech streams. Generally, human brain cannot register and comprehend more than one audio signal at a time. 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 make explicit "Grant Floor" (GF) messages to conferees [6] mandatory before they may speak. Explicit GF messaging hinders liveliness as it precludes impromptu speech avoiding natural interactions. A large number of floors hampers speech quality even though it allows more liveliness in a conference, since there is no need for GF messages. We shall find the minimum number of floors for the purpose. The number of floors thus fixed along with Loudness Number (see Section 4) should serve the purpose.

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

Proof: Part (a) follows from Part (c) of Proposition.

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. This will not pose a problem in a conference of well-mannered participants as either *i* or *j* can undergo *T–*. Yet, such prolonged and infrequent occurrences cannot be ruled out. Then, the conference would become impolite and messy in the absence of an intervention by a third conferee. Thus *N*=2 is insufficient. ■

We have to find a larger *N* that overcomes this infirmity. It is but natural to ask at this stage whether *any* higher *N* will suffice. 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 for a third conferee to undergo *T+* but restrict three floors for no longer than a duration Γ. Here, Γ must exceed pause, τ. Actually, the speakers stop speaking as soon as the speech becomes less intelligible (*vide* 'Repairs' in [16]).

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 Section 4.

If *N* is forced from 3 to 2, and it remains 2 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 recognized. Space constraints preclude this discussion in this paper.

Now we state the result of our arguments in the form of a conjecture (an in depth account of which is not furnished here due to space constraints, but can be easily shown as it follows from the above lemma) under appropriate floor control [20].

Conjecture: ***Three floors (i.e., N=3) are sufficient for a voice-only conference over packet networks.***

IV. LOUDNESS NUMBER (λ) FOR FLOOR CONTROL

Setting *N* will not completely meet the requirements of a 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 [8]. In the context of video, for instance, it has been remarked [9]: '... 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 measure different from the one based exclusively on current packet energies. The measure should have the following characteristics:

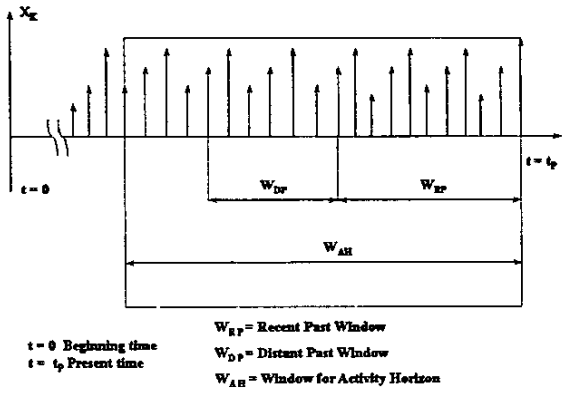

Figure 2. Windows for Loudness Number Calculation

- A speaker (floor occupant) should not be cut off by a spike in the packet energy of another speaker. This implies that a speaker'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. But in a blind conference under discussion, a participant's intention to interrupt can be conveyed effectively through the loudness metric.

A floor control mechanism empowered to cut off a speaker forcefully must be ensured.

We define a new metric called *Loudness Number*, which adapts smoothly so that the floor allocation is graceful. Therefore, for each participant, we define Loudness Number, λ, as a function of the amplitude of the present and past audio stream. When C participants compete for N floors, winners are decided by the higher magnitude of their Loudness Number.

Current activity L_1 (refer Fig. 2) of a conferee is computed by the moving average of packet amplitude (X_K equal to the r.m.s. of the samples in the packet) within a *Recent Past Window*, W_{RP}.

$$L_1 = \frac{1}{W_{RP}} \sum_{K=t_p}^{t_p - W_{RP}+1} X_K \qquad (2)$$

Distant *Past activity* L_2 of a speaker is the moving average of X_K within a *Distant Past Window* W_{DP}.

$$L_2 = \frac{1}{W_{DP}} \sum_{K=t_p - W_{RP}}^{t_p - W_{RP} - W_{DP}+1} X_K \qquad (3)$$

Overall past activity L_3 of a speaker is found over the *Activity Horizon*, W_{AH}.

$$L_3 = \frac{1}{W_{AH}} \sum_{K=t_p}^{t_p - W_{AH}+1} \theta * I_{\{X_K > \theta\}} \qquad (4)$$

where $I_{\{X_K > \theta\}} = 1$ if $X_K > \theta$

$= 0$, otherwise

The threshold θ is a constant and is same for all conferees. We have set θ at 10-20% of maximum packet amplitude. Now the current Loudness Number λ_{tp} is given by

$$\lambda_{tp} = \alpha_1 * L_1 + \alpha_2 * L_2 + \alpha_3 * L_3 \qquad (5)$$ with

$0 < \alpha_1, \alpha_2, \alpha_3 < 1$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

By appropriate choice of window lengths, α_1, α_2, α_3 and θ, λ can be tuned to smoothly provide or withdraw floor access.

A. *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 the Loudness Number. The discussion on how these parameters are selected and the dynamics of loudness number are beyond the scope of this paper.

V. RESULTS AND CONCLUSIONS

We presented an argument to find an upper bound for the value of N (= |S|). We tested our audio conferencing tool for N = {2,3,4} on our test-bed [17, 18]. We found the performance to evolve to satisfaction characterized 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 about the perceptions of conferees on our Test-bed and with heuristics, we used W_{RP} = 5 s, W_{DP} = 10s, W_{AH} = 30s and (α_1, α_2, α_3) = (0.4, 0.3, 0.3). τ = 660ms[21], and $\Gamma \approx 5\tau$ to 10τ are typical.

We tested all the above proposals on our Conferencing *Test-bed* [18, 20] and observed that the quality of conference is very close to a face-to-face conference for ten participants. Our preliminary and not too formal studies lend credence to the values set for various parameters as above.

A qualitative analysis [20] of mixed speech interestingly reveals that 'it is difficult to identify at least one known speech stream when three or more speech streams are mixed'; this supports our arguments in section 3.

An important byproduct of setting N = 3 is reduction of bandwidth in a distributed conference [18, 20] due to filtering of packets based on Loudness Number. This scheme may be extended to video in a video conference by including the metric based on motion vectors. Voice Activity Detection (VAD) algorithms [19] can be used along with the present tool for further performance enhancement.

This discussion does not consider limitations of network support. Effects of delay [12] introduced by the network may hamper smooth floor management in a conference. 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. Interesting issues for future work include *Data Sonification* [14], *Earcon*, and low bandwidth graphics [4] so as to facilitate more effective use of *floors* in the same environment.

A. *In Retrospection*

Studies such as this are often criticized for performance evaluation being subjective, i.e., not quantified. Comparison of perceived quality with existing conference solutions is just one component. Allowing for multiple speech streams to interrupt the speakers, thereby enhancing quality of conference is the second aspect that must be compared. Redenkovic's work [23] appears to be the only report providing for impromptu speech. We believe that our approach of proposing the bound is a concrete and pragmatic step and that it would carry their work further.

VI. REFERENCES

[1] Agustín J González, "A Distributed audio conferencing system", *MS Project Department Of Computer Science Old Dominion University*, Norfolk, VA 23529, July 28, 1997.

[2] Agustín J González and Hussein Abdel-Wahab, "Audio mixing for interactive multimedia communications," *JCIS'98*, Research Triangle, NC, pp. 217-220, Oct. 98.

[3] Agustín J González, "A Semantic-based middleware for multimedia collaborative applications", *PhD Thesis, Computer Science,* Old Dominion University, May 2000.

[4] Alex R Colburn, Micheal F Cohen, Steven M. Drucker, Scott Lee Tiernan and Anoop Gupta, "Graphical enhancements for voice only conference calls", *Technical Report, MSR-TR-2001-95*, Microsoft Research, Redmond, WA98052, Oct 2001.

[5] Dommel H P, and Garcia-Luna-Aceves J J, "Floor control for multimedia conferencing and collaboration", *Multimedia Systems Journal (ACM/Springer)*, Vol.5, No.1, pp. 23-28, 1997.

[6] Dommel H P, and Garcia-Luna-Aceves J J, "Network support for turn-taking in multimedia collaboration", *Proc. IS&T/SPIE Symposium on Electronic Imaging: Multimedia Computing and Networking,* San Jose, CA, Feb 1997.

[7] Ellen Isaacs and Herbert H. Clark, "References in conversation between experts and novices" *Journal of Experimental Psychology: General*, Vol. 116, No 1, pp. 26-37, 1987.

[8] Ellen Isaacs and John C. Tang, "What video can and cannot do for collaboration", *Springer-Verlag journal, Multimedia Systems*, vol. 2, pp. 63-73, 1994.

[9] Ellen Isaacs, Trevor Morris, and Thomas K. Rodriguez, "A Forum for supporting interactive presentations to distributed audiences", *ACM Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW)*, Chapel Hill, NC, pp. 405-416, 1994.

[10] Gyton A C and Hall J E, "Text book of medical physiology", W B Sanders co., USA 9e.

[11] ITU-T Rec. H.323, "Packet based multimedia communications systems", vol. 2, 1998, http://www.itu.int/itudoc/itu-t/rec/h/h323.html.

[12] Karen Ruhleder, Brigitte Jordan, "Co-Constructing non-mutual realities: delay-generated trouble in distributed interaction", *Computer Supported Cooperative Work*, Vol.10, No.1, pp. 113-138, 2001.

[13] Lisa R. Silverman, "Coming of age: conferencing solutions cut corporate costs" *White Paper, Interactive Multimedia Collaborative Communications Alliance*, http://www.imcca.org/

[14] Madhyastha T M, and Reed D A, "Data sonification: do you see what i hear?", *IEEE Software*, March 1995.

[15] Sacks H, Schegloff E A, and Jefferson G "A Simplest systematics for the organization of turn-taking for conversations", *Language*, vol. 50, No. 4, pp. 696-735, 1974.

[16] Schlegloff E A, and Jefferson G, Sacks H, "The preference for self-correction in the organization of repair in conversation", *Language*, vol. 53, No. 2, pp. 361-382, 1977.

[17] Venkatesha Prasad R, Joy Kuri, H S Jamadagni, Haresh Dagale, Ravi R Ravindranath, "Automatic addition and deletion of clients in voip conferencing" *6th IEEE Symposium on Computers and Communications,* Hammamet, Tunisia, pp. 386-390, July 2001.

[18] Venkatesha Prasad R, Joy Kuri, H S Jamadagni, Haresh Dagale, Ravi R Ravindranath, "Control protocol for voip audio conferencing support", *International Conference on Advanced Communication Technology (ICACT'01)*, Mu-Ju, South Korea, pp. 419-424, Feb 2001.

[19] Venkatesha Prasad R, Abhijeet Sangwan, H.S. Jamadagni, Chiranth, Rahul Shah, Vishal Gaurav, "Comparison of voice activity detection algorithms for VoIP", *IEEE Symposium on Computer and Communications*, Italy, pp. 530-535, July 2002.

[20] Venkatesha Prasad R, " A New Methodology for Audio Conferencing on Voice over IP (VoIP), *PhD Thesis*, Manuscript.

[21] Jaffe J and Feldstein S, "*Rhythms of Dialogue*" New York, Academic Press, 1970.

[22] Rosenberg R, Schulzrinne H, et al., "SIP: session initiation protocol", *RFC 3261, IETF*, Jun. 2002, ftp://ftp.isi.edu/in-notes/rfc3261.txt

[23] Redenkovic M, Greenhalgh C, "Multi-party distributed audio service with tcp fairness" *ACM MM*, pp.11-20, Dec. 2002.