Supporting Collaborative Audio in the Internet

Milena Radenkovic, Chris Greenhalgh
School of Computer Science and IT
University of Nottingham

Abstract: - Distributed Partial Mixing (DPM) is an approach to creating a distributed audio service that supports optimisation of bandwidth utilization across multiple related audio streams (e.g. from concurrently active audio sources) while maintaining fairness to TCP traffic in best effort networks. This paper extends our DPM prototype to include the application control of selector and database components of DPM as a necessary complement to purely network-driven DPM presented till now. In particular we introduce awareness driven and Interest Management driven DPM as one way of that smoothing out the effects of the network adaptation on the end user audio quality in Collaborative Environments.

*Key words :*Real-time audio, Mixing, Distributed Partial Mixing, Spatial Model, Locales, Collaborative Virtual Environments (CVEs)

1. Introduction

Over the past several years the Internet has been experiencing the emergence of large-scale collaborative networked applications that aim to support real world scenarios and the associated patterns of audio activity. For example, collaborative virtual environment technology has been used for applications that support large on-line communities and highly interactive social events such as multi-player games and inhabited television [4]. Experiments have shown that these applications involve potentially significant numbers of simultaneous speakers [3]. The deployment of such applications is expected to increase and pattern of simultaneous speaking to continue.

The current Internet and current approaches are not well suited to support such applications. Today's Internet operates on a best-effort basis and cannot guarantee an upper bound on end-to-end delay nor lower bound on available bandwidth. These applications can easily cause network congestion, and hence packet loss and increased delays that can significantly the quality of delivered audio. As such applications become more widespread, large number of audio streams may form a considerable portion of the Internet load.

We proposed DPM as an efficient technique that dynamically adopts both to varying number of active audio streams in collaborative environments and to congestion in the network. The single most distinctive aspect of the DPM design is the mixing of a subset of incoming streams. The choice of streams to be mixed – which and how many – determines the effectiveness of distributed partial mixing.

Our previous work was primarily focused on network-driven DPM. In contrast, this paper concentrates on application-driven DPM in order to maximise user satisfaction within prevailing network resources.

This paper is organised as follows. Section 2 introduces some aspects of collaborative audio and the underlying networks used for communication that are of direct relevance to the work presented in this paper. Section 3 gives overview of Distributed Partial Mixing as the basis for building any audio service that efficiently supports applications that allow simultaneous speaking as a natural pattern of audio activity among participants, and briefly shows results of network-driven DPM. Section 4 proposes awareness and interest management techniques as a flexible framework for controlling the selector and database components of the DPM system. Section 5 concludes and outlines directions of our future work.

2. Collaborative Audio

Audio is arguably the most important component of a real-time collaborative application and perhaps the most complex one. It often carries the core content of our interaction and serves as

the baseline for successful collaboration. For example, recent studies have noted how participants fall back on audio to resolve difficulties with other aspects of collaborative applications such as negotiating a shared perspective [7].

The design of audio for collaborative interfaces is complicated by: constraints imposed by the environments in which the audio service must operate, potentially large numbers of simultaneous participants, as well as a wide range of interpersonal communication models.

The underlying network requirements for any collaborative audio are very strict (stricter than for non-interactive audio). Collaborative audio is sensitive to network packet loss, latency and jitter, and to gaps in the audio caused by the lack of the real-time support on general purpose hosts. These factors affect intelligibility of audio and speech.

Different models of communication among participants will have different bandwidth requirements and architectural needs. After a brief overview of the traditional applications and their needs, a newly emerging style of applications with different needs is introduced.

The most common scenarios for the traditional use of networked audio include public presentations, on-line lectures and small-scale interactive conferencing.

In public demonstration applications only a small number of participants produce audio while the majority of the attendees can only receive. In an on-line lecture set-up, the subset of participants allowed to send audio data can change during the evolution of the application. In this scheme the active participants are not predetermined but there is a control over who will speak at any time. In interactive small scale conferencing (such as some telephony and more traditional CSCW applications) more than one (potentially all) participant(s) can transmit audio at the same time. They do indeed support several simultaneous speakers as there is no control over who will speak at any time. But they focus mostly on small groups and a slow pace of interaction.

The distribution of audio streams in the network for such applications is either one-to-many or few-to-many (M-to-N where usually M<<N). Various techniques and protocols have been proposed (and used) for reducing the bandwidth required for such broadcasts, most notably network multicasting [9], but also RTP[12].

However, newly emerging large scale networked applications such as collaborative virtual environments (CVEs) can support large on-line communities and highly interactive social events such as multi-player games and inhabited television [9]. These applications have a fast pace of interaction, and have M-to-N distribution of audio streams where M is almost equal to N. The work of Communications Research Group (CRG) from Nottingham University on inhabited television has focused on enabling public participation in on-line TV shows within shared virtual worlds [4]. Figure 1 shows a screenshot of a CVE world with multiple mutually aware participants in which we run our experiments. As part of a public experiment in inhabited TV called Out of This World (OOTW), patterns of user activity, including audio activity, were studied by statistically analysing system logs [3]. The results showed that overlapping audio transmissions from several participants were common for this event. Indeed, during a 45 minute show, there were several minutes when all 10 of the participants were generating audio traffic at the same time. Periods of high activity included teams shouting instructions to one another during action games and also shouting out answers to questions (OOTW was a gameshow).

Similar analyses of patterns of audio activity in other CVE applications and platforms, for example virtual teleconferencing in the DIVE system [2], have also revealed significant periods when several participants are simultaneously generating audio traffic. Indeed, previous experiences suggest that, even for relatively focused applications such as teleconferencing, audio activity is best approximated by a model of people transmitting audio at random, rather than deliberately avoiding overlapping speech. For more socially dynamic kinds of events such as OOTW, audio activity appears to be even more strongly positively correlated among participants.

Fig 1 Virtual World with multiple mutually aware participants

Not all of this activity is necessarily verbal; it may also include non-verbal utterances and background noise. However non-verbal utterances form a significant part of human communication. In addition, studies of social interaction in CVEs have noted how communication in a virtual world can be influenced by events in participants' local physical environments [1]. Obtaining awareness of these events through overheard background audio might help participants account for actions in the virtual world. Media spaces are one class of collaborative application that employ open audio connections to provide this kind of awareness. For example, when discussing patterns of usage for the Thunderwire audio media-space, Hindus *et al*. note that it might be better to consider the number of live microphones, rather than the number of active participants [8].

Therefore, overlapping audio, including talk and other noise, is likely to be the norm for future CSCW applications. This paper anticipates that as audio-enabled applications become more widespread, and are used to support larger virtual communities, so designers of audio services will see increasing numbers of simultaneous speakers.

3. Distributed Partial Mixing

[10], [11] proposed a novel technique called Distributed Partial Mixing (DPM) as the basis for this service. DPM can dynamically adapt both to varying numbers of active audio streams in the collaborative network application and to congestion in the network. Each distributed partial mixing component adaptively mixes subsets of its input audio streams into one or more mixed streams, which it can forward to the other components along with any unmixed streams. DPM minimises the amount of mixing performed so that end user recipients receive as many separate audio streams as possible within prevailing network resource constraints. Delivering large numbers of independent streams to the end user is important in order to allow maximum flexibility of audio control to the end users. Results of a series of experiments over a single network link demonstrate the effectiveness of congestion control performed by distributed partial mixing. Distributed partial mixing manages the tradeoff between preserving stable audio quality, being responsive to congestion and achieving fairness towards TCP traffic. This is shown in Fig. 2.

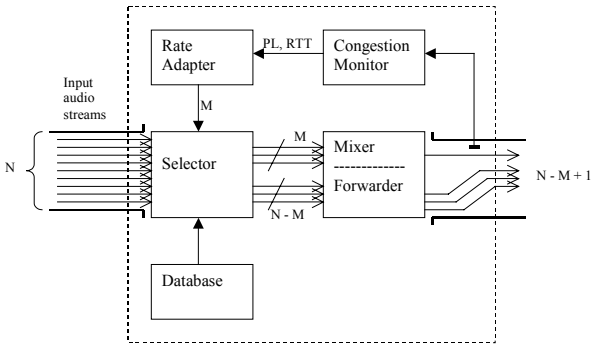


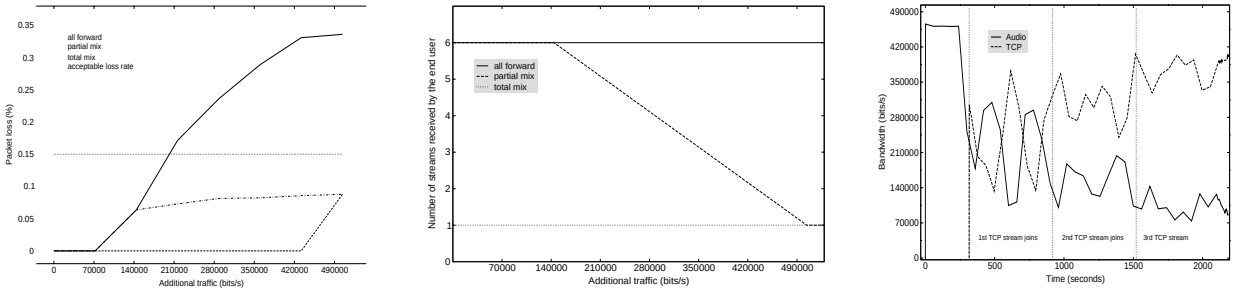
Figure 2. Functional Architecture of DPM

The topology of DPM components and clients can be static or dynamic. This paper assumes that the application configures all the DPM servers and clients statically in the beginning of the session.

[11] also proposed a number of application criteria that could be used by CSCW applications and that need to be considered when deciding which audio streams to mix. These criteria have to be taken into consideration by the DPM selector and they included: roles of speakers, roles of listeners, natural groupings of sources, patterns of activity and voice characteristic. We propose one extension to this set of criteria that includes selection of audio streams according to he predetermined criteria For example, the allocation or reservation of bandwidth may be controlled by different charging tariffs associated with the quality of

service required. In this way a user may specify a quality of service requirement of say 3 x 64kb/s audio channels in which case selected audio streams will be mixed in the network when more than three separate audio streams are to be transmitted by the network. In this case available bandwidth may be considered as allocated or reserved bandwidth for use according to user specified quality of service requirements.

DPM prototype was implemented in MASSIVE-3 CVE platform [5]. Network-driven DPM prototype was evaluated extensively in small scale demonstration system and its effectiveness evaluated in term of minimising packet loss (Fig 3a) maximising levels of specialisation (Fig 3b), being adaptive and TCP fair (Fig 3c) as well as supporting heterogeneous networks and supporting efficient distribution of streams. A wide rage of experiments was carried out over a single network link to show its steady state and dynamic behaviour of the built system against both non-adaptive traffic and adaptive traffic. To determine effectiveness of DPM, it was compared to multicast and full mixing approaches (Fig 3a and 3b). For more detailed discussion of the results see [11].



3a Packet Loss rates 3b Levels of Spatilaisaiton 3c TCP Fairness

Figure 3. Steady-state (3a, 3b)and dynamic characteristics (3c) of DPM

4. Application Control of DPM

We first introduce the idea of awareness driven DPM and interest-management driven DPM, and then move to describing the DPM framework that incorporates that. This refers to how the selector and database decide which streams to mix and which to forward given that the comparator has limited the total number of streams to be transmitted based on congestion in the network.

4.1 Introducing Spatial Awareness Driven DPM

Awareness driven DPM involves the application of the previously defined spatial model of interaction [6] to the medium of audio. The inhabitants of a CVE move about a shared virtual world. These movements are used to calculate different levels of mutual awareness through the spatial model mechanisms of aura, focus, nimbus and third party objects. The DPM system uses the awareness level associated with each audio source to decide whether to mix that audio stream. Users' spatial movements therefore provide a high level, medium independent way of dynamically managing the QoS of multiple audio streams which can then be played within the virtual world.

The spatial model of interaction is a computational framework that allows a CVE to estimate the relevance of different elements of a virtual world to each user. This is expressed as the user's "awareness" of these elements. The spatial model uses three concepts to negotiate awareness:

- ***nimbus:*** controlled by the originator of information to express social behaviours such as interrupting, shouting and whispering, as well as security restrictions and directionality.
- ***focus:*** controlled by the (potential) recipient of information to express allocation of attention, preference, and interest.
- ***third party objects:*** the context in which the communication occurs, which may include boundaries (e.g., walls and windows) and modifiers (e.g., a speaker's podium).

Focus and nimbus are defined as spatial fields whose values may vary with distance from the user. Focus, nimbus, and the effects of third party objects can be defined independently for each potential medium of interaction. Considering two participants A and B, A's awareness of B in medium M is a function of A's focus on B in M and B's nimbus on A in M, modified by any relevant third party objects. Although the role of focus is relatively intuitive – supporting users in selecting information of interest, the role of nimbus requires more consideration. Essentially, nimbus allows a user to project their information at others to a greater or lesser extent, thereby engaging in activities such as whispering, shouting and interrupting.

4.2 Introducing Interest Management Driven DPM

In addition to having only one unit of application controlled DPM based only on spatially defined region (associated with a third party object), we drive DPM with a more flexible approach which is not limited to spatial organisation, and which does not have excessive overheads (change with every movement in the virtual world). These include having and integrating several elements of the state of the art in interest management including functional and organisational distinctions, and abstractions in DPM.

In MASSIVE-3 we have adopted (and extended) the concept of "Locales", introduced in the SPLINE system as an effective means of subdividing a virtual world according to spatial characteristics. In MASSIVE-3 we further sub-divide a locale into one or more "Aspects". These are the fundamental unit of interest management in MASSIVE-3, and each aspect is realised as one environment database. Following [6] we add the notion of organisational scopes i.e. these will normally be defined by the application to reflect organisational associations between objects. For example, a game may define an organisational scope for each team, while a military simulator may define an organisational structure to mirror the military hierarchy. The system could also use organisational scopes to split overcrowded aspects into sub-parts. Each Aspect can have different fidelities e.g. an observer can then choose the aspect with the fidelity (fully mixed, partially mixed or fully forwarded audio streams) that suits their requirements and available resources. We call reduced-fidelity aspects "abstractions", as in [13] (c.f. "aggregates" in [6]). Abstractions may display audio at a lower quality, for example mix all the audio streams belonging to a group. Alternatively, abstractions may contain composite representations of multiple objects.

The main "work" of interest management in DPM is to determine what audio streams should be received and transmitted to and from the aspects. In MASSIVE-3 we avoided a single specific policy, adopting instead a framework into which application-defined policies can be dynamically inserted, as required. Currently MASSIVE-3 includes standard policies to select aspects based on topological distance, Euclidean distance, awareness (each inter-locale link includes an awareness value describing the level of awareness across the link), and benefit/cost (which are statically configured at present, but could be derived from actual activity). Different policies can be specified for different functional and/or organisational types, and for different fidelities. Multiple policies can then be combined using logical conjunctions (e.g. policy A OR policy B). The awareness of the participants within the same aspect is determined via the Spatial Model presented in the previous section.

The general overview of the Spatial Awareness and Interest Management driven DPM is shown in Fig 4.

4.3 Incorporating application criteria in DPM packet

In order to clarify DPM protocol we describe DPM audio packet format in more detail. Each DPM packet includes audio data, feedback and control information. The packet (as used in our implementation of DPM within MASSIVE) is shown in Figure 5. Feedback information is piggybacked in the header of a returning DPM packet, and includes sequence number of the packet that is being acknowledged (unsigned long ACKseq_num), the difference between the time when it was received and time when it is being sent back to its sender (ackRTT). Packet loss rate estimated in the receiver is also sent

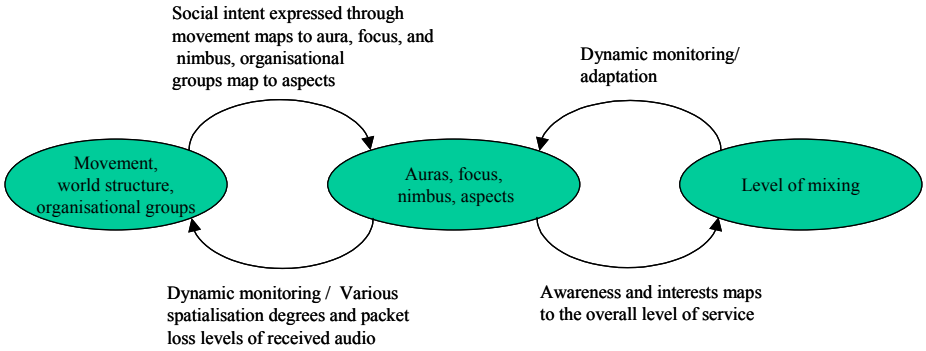


Figure 4. Overview of Awareness and Interest Management driven DPM

back to the sender (even though the sender estimates packet loss event rates as well). Various control information such as speaker's identification number (e.g. num_speaker_ids identify speaker's role) and conference ID number (which identifies the scope of speaker's interaction with the other users) are also piggybacked in the returning DPM packet header. The field (unsigned short pad) is used to express the preferences of the current speaker as

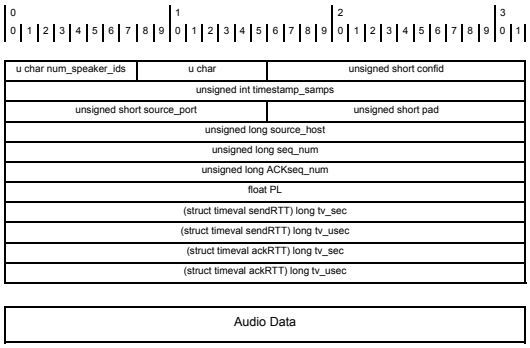


Figure 5. DPM packet format

well as the ID of the particular receiver and/or target multicast group. The selector in each DPM server processes these fields in order to choose which streams are to be mixed in order to maximise the end user experience. Each DPM packet audio header also included sequence number of the packet being sent (seq_num) and the times when it has been sent (sendRTT). Flag field (Flags) is set to different values depending on whether it marks the start of a new talkspurt, or audio format bits or packet type bits.

4.4. Incorporating Application Criteria within Selector and Database

In our extended prototype, the application is responsible for providing the following information to each DPM:

- A spatial awareness matrix that represents the awareness between all the users relevant for that DPM. This matrix changes dynamically with each user movement within the virtual world. This can be "too fine" granularity for the majority of applications and it can result in large overheads. DPMs uses this matrix only if the application involves tasks where spatialiasation is of highest importance for the success of the task, and when the network is estimated as stable (LANs or private networks), or if the users subscribed to use DPM at high tariffs.
- An aspect matrix that represents relationship between aspects. This policy is less dynamic (e.g. users' interests do not change as rapidly as their movements) has small overheads. This is the prototype-DPM's default policy that it uses it for all networks and applications.
- A linked list of neighbouring DPMs and clients connected to these DPMs that represents the topological structure of the DPM system relevant to a particular DPM.
- A list of interest profiles for each client relevant for the DPM. This is based on the static parameters such as roles of user and their tariffs.

In this way each DPM, on receipt for a DPM audio packet is able to perform the following:

- extract user source ID/role and user destination ID/role from the packet,
- check their preferences against their static interest profiles in the database,
- check their aspects against the aspect matrix
- check their dynamic mutual awareness against the dynamic awareness matrix in the database,
- decide which streams to mix/forward and send it to the next DPM as specified in the list of DPM/client topology by the application.

With this controlling scheme, conflicts (such as sender requiring their stream to be mixed and receiver requiring not to have that sender mixed and vice versa) are resolved by the virtual world computations of mutual awareness and aspect transformations.

5. Conclusions and future work

In this paper we argued that DPM is necessary both for the health of Internet and the quality of delivered audio and stability of the applications. Our previous work has demonstrated the effectiveness of the network-driven DPM over wide range of link conditions.

This paper moved forward and suggested ways of integrating application preferences into DPM. It discussed in detail how application and user criteria can be integrated in the DPM packet, selector and database of each DPM. It described how spatial model of interaction could be used as the main controlling mechanism for the selector and database components of a fine-grained DPM. The paper also proposed the use of aspects to control coarse-grained DPM mixing policies and decrease control overheads. Further evaluation of the application-driven DPM was out of the scope of this paper and is the matter of our future work.

References

1.Bowers, J., Pycock, J. & O'Brien, J., Talk and Embodiment in Collaborative Virtual Environments, in Proc. of CHI'96, 1996.

2.Frécon, E., Greenhalgh, C. & Stenius, M., The DiveBone:AnApplication-Level Communication Infrastructure for Internet-Based CVEs, in Proc. of VRST'99 (London, 1999), 58-65.

3.Greenhalgh, C., Benford, S. & Craven, M., Patterns of Network and User Activity in an Inhabited Television Event, in Proc. of VRST'99 (London, Dec 20-22 1999), ACM, 58-65.

4.Greenhalgh, C., Benford, S., Taylor, I., Bowers, J., Walker, G. & Wyver, J., Creating a Live Broadcast from a Virtual Environment, in Proc. of SIGGRAPH'99 (1999), 375-384.

5. Geenhalgh C., Purbrick J., & Snowdon D., Inside MASSIVE-3: Flexible Support for Data Consistency and World Structuring, in Proc. of CVE'2000, San Francisco, ACM, 119-127.

6.Greenhalgh, C. M. and Benford, S. D., Supporting Rich And Dynamic Communication In Large Scale Collaborative Virtual Environments, Presence: Teleoperators and Virtual Environments, Vol. 8, No. 1, February 1999, pp. 14-35, MIT Press

7. Hindmarsh, J., Fraser, M., Heath, C., Benford, S. & Greenhalgh, C., Fragmented interaction: establishing mutual orientation in virtual environments, in Proc. of CSCW'98 (Seattle, WA, USA, Nov. 1998), ACM Press, 217-226.

8. Hindus, D., Ackerman, M., Mainwaring, S. & Starr, S., Thunderwire: A Field Study of an Audio-Only Media Space, in Proc. of CSCW'96 (USA), ACM Press, 238-247

9.Macedonia, M., Brutzman, D., Mbone Provides Audio and Vision Across the Internet, IEEE Computer, pp30-36, April 1994.

10.Radenkovic M, Greenhalgh C., Patent published: Audio Data Processing, International Publication Number: WO 02/17579 A1, 28 February 2002

11.Radenkovic M, et al "Scaleable and Adaptable Audio Service for Supporting Collaborative Work and Entrainment over the Internet", in the Proc. of International Conference on Advances in Infrastructures for e-Business, e-Education, e-Science and e-Medicine in Internet (SSGRR 2002), L'Aquila, Italy, January, 2002.

12.Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V., RTP: A Transport Protocol for Real-Time Applications, IETF RFC 1889, January 1996.

13.Singhal and D.R. Cheriton, "Using Projection Aggregations to Support Scalability in Distributed Simulation", in Proc. of ICDCS' 96.