SPATIALIZED AUDIO CONFERENCES
IMS Integration and Traffic Modelling

Christopher J. Reynolds, Martin J. Reed
University of Essex, Colchester, Essex, UK

Peter J. Hughes
Broadband Applications Research center, BT Group, Adastral Park, Ipswich, UK

Keywords: Spatial audio, audio conferencing.

Abstract: Existing monophonic multiparty VoIP conferencing applications are currently limited to supporting a single conversation floor, with limited numbers of simultaneous speakers. We discuss the additional requirements and benefits of delivering a spatially enhanced audio application via Head Related Transfer Function (HRTF) filtering, which may support many conversation floors. Several network delivery architectures are presented, including integration to the Next Generation Network (NGN) IP Multimedia Subsystem (IMS). The delivery architectures are compared using traffic models, and implications for the scope of such an application are discussed.

1 INTRODUCTION

Multiparty VoIP conferencing is among a range of advanced voice services to be offered by next generation networks (NGN's), with the IP multimedia subsystem (IMS) providing native support both for media delivery, and session management via Session Initiation Protocol (SIP). We explore a proposal for the delivery of a new headphone based VoIP multiparty conferencing application, with Head Related Transfer Function (HRTF)(Cheng and Wakefield, 2001) filtering used to provide a spatially enhanced audio environment. Existing *monophonic conferencing* systems impose severe limits upon the participant's ability to naturally converse, particularly for large groups. *Spatialized audio conferencing* allows for a much more natural audio environment, and extends support for larger groups by allowing overlapping speech to be distinguished as separate perceptual streams (Bregman, 1994). Whilst the theoretical limit to the number of participants within a monophonic conference may be large, users are typically limited to interacting via a single conversation floor in which they align their turns of speech, with a suggestion that the maximum number of simultaneous speakers be set at 3 (Venkatesha et al., 2003). As such, support for multiple conversation floors is restricted and indeed not expected to occur, as many overlapping speakers presented monaurally are difficult to distinguish. However the addition of spatial cues can extend support for multiple conversation floors. The ability to focus upon a particular talker in the presence of other conversations is greatly enhanced when the sources are spatially separated, a phenomenon well known as the *cocktail party effect* (Cherry, 1953). It is known that presenting multiple audio sources from different spatial locations aids the perceptual organization of sound streams (Bregman, 1994), and can enhance memory, comprehension and intelligibility (Baldis, 2001). Conferences with spatial cues more closely resemble face to face meetings and conversations, and represent a significant advance over existing monophonic conferencing applications.

The audio mixing and filtering process may be performed locally at a users terminal, or centrally via a dedicated server, and whilst methods for mixing audio using these models have been discussed (Singh et al., 2001), spatialized audio conferences have not yet been covered. We discuss the additional requirements for delivering such an application, the relative merits of adding spatial cues, and how such an application may be integrated within the NGN/IMS model. We cover the limits imposed by both the psychoacoustic properties of such an audio environment, and

J. Reynolds C., J. Reed M. and J. Hughes P. (2007).
SPATIALIZED AUDIO CONFERENCES - IMS Integration and Traffic Modelling.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 243-248
DOI: 10.5220/0002134002430248

the network delivery architecture. Three network delivery architectures are then considered, *Centralized*, *Unicast Full Mesh*, and a brief outline of a *Hybrid* system. A traffic model is constructed with reference to NGN core/access partitioning, and comparisons of resulting traffic are made for each architecture.

2 CONFERENCE MODELS AND SPATIAL AUDIO

Spatialized or 3D audio for virtual multiparty conferencing has been implemented by Kilgore et al (Kilgore et al., 2003), with simple manipulation of Inter Aural Time Differences (ITD) and the Inter Aural Intensity Differences (IID) in accordance with duplex theory (Cheng and Wakefield, 2001). HRTF based systems are known to produce effective spatial reproduction (Crispien and Ehrenberg, 1995) (Evans et al., 2000) and have been integrated into a conferencing application under our development.

Using HRTF based spatial audio, a participant's mono voice stream may be convolved with a HRTF to give a binaural audio stream that has temporal and spectral effects that mimic a sound source from a given point in space. Convolving each participant with a different HRTF (relating to a different azimuth and/or elevation), and then mixing the output for all participants produces an audio space in which each different speaker's utterance will appear to emanate from a different spatial location. As mentioned previously, this has many benefits for communication and more importantly allows multiple conversation floors to emerge through the process of *schisming* (Egbert, 1997), in which a large conversation floor involving many participants may fragment into several smaller floors. Users make use of the cocktail party effect to ignore other conversations within the audio space, and to align their speech turns to a conversation floor of their choosing. As a result many floors may exist within the space/conference. The floor control mechanism of limiting and choosing the number of simultaneous speakers is no longer required, as many participants may speak simultaneously without masking each other. Limits to the number of conferees are discussed later in relation to the delivery architecture.

Where the mixing and HRTF filtering is performed has direct implications for both the scope of such an audio space, and the resulting network traffic. The next section introduces the possible architectures, with a brief discussion on NGN partitioning.

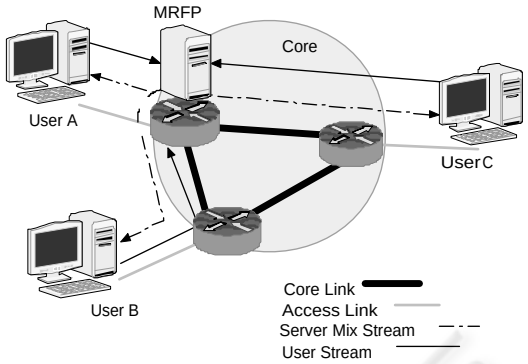

Figure 1: Core/Access Network Division.

2.1 NGN and IMS: Centralized Conferencing

The NGN architecture provides logical division between service functions and the underlying transport technologies. The transport functions are further divided into access and core network functions, which perform a range of quality of service mechanisms including packet filtering, marking, shaping, buffer management, scheduling and queuing (Knightson et al., 2005). The core transport network and its associated control functions provide a platform to deliver traffic for services such as the IMS, and may be logically separated by technology, ownership or administrative boundaries. An IMS may be located within a core network partition, and can provide support for media services such as audio conferencing. An Application Server (AS) within the IMS can be used for conference control, with SIP based session control through call session control functions (CSCF). In the NGN/IMS model, ASs have control over audio mixing and filtering through the media resource function controller (MRFC) that directly controls the media resource function processor (MRFP) which is responsible for audio processing. The AS and MRFP may be physically separate, and thus it is the MRFP location that is critical as the audio traffic dominates the signalling traffic.

2.1.1 Mixing

The MRFP allows for a centralized audio conferencing model, under the control of an application server. An outline for server based audio mixing for monophonic conferencing is described in (Singh et al., 2001), including a discussion of the decoding, jitter buffering and mixing procedure, as well as some performance statistics. Figure 2 shows the additional filtering process within the MRFP required to pro-

vide a spatially enhanced audio scene for a set of 3 participants. The controllers may act upon SIP messages as users leave or join the conference, to signal the MRFP to select the relevant HRTF and to determine the mixing configuration. The MRFP may convolve each stream with a different HRTF according to some pre-defined spatial arrangement (some preliminary suggestions have been made (Brungart and Simpson, 2003)), and deliver a mix back to each participant. As such, each participant will receive a custom mix consisting of all the other participants' binaural streams, with their own stream missing. As suggested by (Singh et al., 2001) local removal of a participant's stream from a mix may be difficult, hence a single (possibly multicast) stream for all participants is not possible as the participant would hear their own voice. As participants leave or join the conference, the spatial arrangement may be altered by the server by applying different HRTF's to each stream. For example a mix of x streams may be spatially arranged in the frontal hemisphere with equal θ degrees of separation. Should a participant leave, the angular separation θ may become $180/(x-1)$.

As mixing is performed by the MRFP, clients that are not capable of mixing and spatializing audio such as smart phones and PDA's, may still participate in conferences. Each mix may then be encoded with an arbitrary waveform codec and distributed to the appropriate participant.

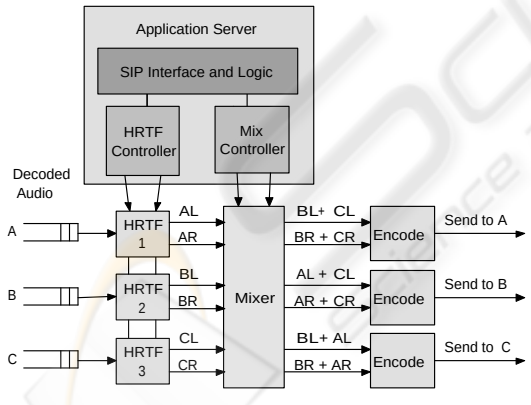

Figure 2: Convolution and Mixing Arrangement.

2.1.2 Centralized Conference Size Limits

Whilst the audio processing is handled by a server, thus reducing the processing load on the clients, in practice a limit to the number of conference participants may still be imposed. A limit may occur due to the degree of localization performance that allows perceptual voice streams to be spatially distinguished from each other, as the mix returned by the MRFP presents each talker at a fixed spatial location. Also spatialization effects such as reversals (Begault and Wenzel, 1993) may limit the arrangement of audio sources to the frontal hemisphere only (a back to front reversal may lead to the perceptual overlap of two speakers if another source exists in a position mirrored in the interaural axis). Limits may also be imposed by the MRFP processing capabilities. However, network bandwidth limits should not be restrictive as participants only ever send one audio upstream, and receive the single audio mix from the server downstream. A comparison of traffic estimates is made in section 3. Next we consider an alternative delivery model.

2.2 Unicast Full Mesh

With the unicast full mesh model, participants send a copy of their own voice stream to every other conference participant. HRTF convolution and mixing is then performed locally at the participant's terminal. This method would be restricted to terminals with the capabilities to filter and mix an allocated number of streams. Since the spatial cues are added to each stream locally, users have full control over their own audio space. The filters may be adjusted to allow each user to fully customize where they hear other members of the conference. For example a user may choose to group a number of voice streams with whom they are not conversing to a similar azimuth (effectively merging the multiple streams to one perceptual stream), or make adjustments to the volume of each speaker.

2.2.1 Unicast Full Mesh Conference Size Limits

In the unicast full mesh model, limits to the number of conferees are imposed by the terminal resources and access network technologies, rather than the perceptual spatial arrangement. This may be restrictive for asymmetric technologies where upstream bandwidth is limited.

2.3 Hybrid

With a centralized model, the spatial locations and the mix for each user are fixed, as any changes a participant makes to the HRTF set would be common for the group. However, a hybrid model may be implemented to give users control over their mix. We propose that the application server may respond to SIP INFO messages sent from participants, and adjust the users mix upon request via the Mix Controller shown

in Figure 2. This would allow each user to control the volume at which they hear other participants, perhaps at a reduced level for conversation floors they are not involved in. Traffic modelling for this architecture may be considered equal to the centralized model, on the assumption that SIP INFO signalling traffic may be ignored.

3 TRAFFIC MODELLING

This section describes a comparison of network traffic generated for two suggested conference delivery models, centralized and unicast full mesh, across a network logically divided into core and access partitions. Figure 1 shows an example of the network partitioning with an MRFP (placed at a single node) and 3 users. Four conference group sizes were investigated with varying degrees of distribution across the network. Each group consisted of N users, and only one group was modelled at a time. A random core network was generated using the BRITE topology generator with the AS Waxman configuration. The network consisted of 130 nodes, based upon a hypothetical national sized NGN core, with node degree 3. Edge bandwidths were set to infinite on the assumption of an unconstrained capacity model. Edges were also assumed to be of unitary cost, *i.e.* hop count considered more dominant in cost than distance, though for long distance topologies this may require revision. A fixed low bit rate voice codec rate r was set to 16kb/s for user voice streams, whilst the MRFP return rate s was set to 128kb/s, based upon an MPEG II layer 3 waveform codec to preserve stereo reproduction. The MRFP was positioned such that in each scenario, the sum of all paths between users and the MRFP was at a minimum.

3.1 Access Network

The access network was modelled as a single link from each user to core as shown in Figure 1. Since access traffic is independent of how distributed the users within the group are, the traffic is trivial to calculate but for completeness is shown below. Unicast access upstream traffic U_u may be defined as:

$$U_u = rN(N-1) \tag{1}$$

Unicast downstream access traffic D_u is defined as:

$$D_u = rN(N-1) \tag{2}$$

Access upstream traffic with application servers may be defined by (3), whilst downstream traffic is defined by (4)

$$U_c = Nr \tag{3}$$

$$D_c = Ns \tag{4}$$

Figure 4 shows a comparison of downstream access traffic between the two models. Clearly at group size 9 the traffic for both models is equal, whilst traffic is reduced using a centralized model when group sizes grow beyond this. A significantly greater saving is made using a centralized model when considering upstream traffic, as illustrated in Figure 3, as less traffic is generated for all group sizes. This has significant implications for asymmetric access technologies.

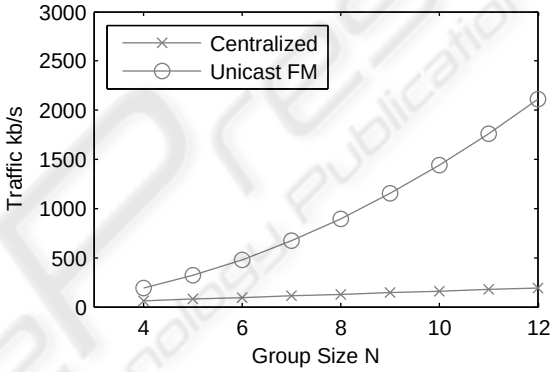

Figure 3: Access Upstream Traffic.

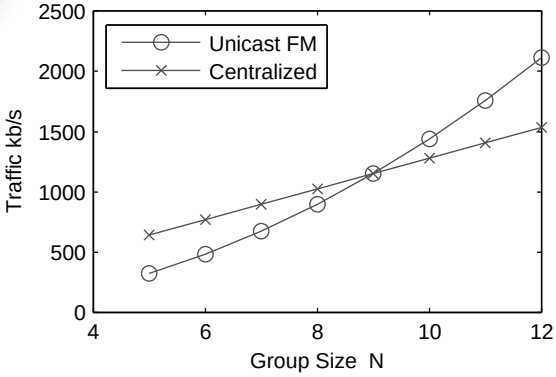

Figure 4: Access Downstream Traffic.

3.2 Core Network

Traffic generated in the core is dependent on how distributed the group of users are. To measure this distribution a value of mean hop count (MHC) was used,

and may be defined as follows. Let A be the set of N users where each user is connected to one node V in the core network $G(V,E)$. We define the mean hop count L between all users in A as follows:

$$L = 1/N \sum_{\{s,t:s,t\in A, s\neq t\}} P(s,t) \quad (5)$$

where $P(s,t)$ is the length of the shortest path in G between users s and t measured using unit length for each edge E. Note that this does not include any access cost.
Core traffic for unicast T may be summed as:

$$T = r \sum_{\{s,t:s,t\in A, s\neq t\}} P(s,t) \quad (6)$$

Unicast traffic and mean hop count are trivially related. However it is possible to distribute a group with a fixed MHC (and hence fixed unicast traffic), and calculate the traffic saving with an MRFP.

For each group size, 50 scenarios were simulated each with the same MHC, and their results averaged. The traffic generated with the centralized model was then deducted from the Unicast model to generate a value of *traffic saving*. Figure 5 shows the values for traffic saving for each group size, with varying distributions measured by MHC. For a group size of 6 no saving is made by using the centralized model, as indicated by the negative saving values across all distributions. For group size 8, traffic savings increase slightly as the group spreads out, though savings are always made for this group size. Significant savings are made for group sizes of 10 and 12 showing an increased traffic saving as the group becomes more distributed.

Thus for larger, more distributed groups, large savings are made when using the centralized model. For smaller, more concentrated distributions, audio filtering and mixing should not be done centrally, rather locally at the participant's terminal. When considering the case for supporting terminals with no filtering and mixing capabilities, a centralized model is the only possible solution, though at a cost of increased network traffic for small groups.

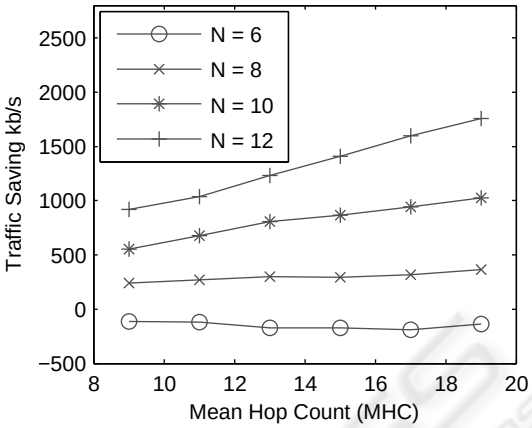

Figure 5: Core Traffic Saving.

4 CONCLUSIONS AND FUTURE WORK

We have presented three models for the delivery of a collaborative spatially enhanced audio space, and outlined the necessary modifications to existing conferencing mixing architectures to support such an environment. This included a discussion of changes related to floor control, in order to support many simultaneous conversation floors within a space/conference.

The centralized model fits naturally within an IMS infrastructure located within an NGN core network partition, in order to reduce core traffic for larger group sizes, demonstrated by a traffic modelling investigation for a randomly generated core network. Access traffic has also been shown to drop for a centralized model, in particular upstream where access network bandwidth for asymmetric technologies may be restricted. The advantages of convolution and mixing at the users terminal have also been discussed, a process which allows a fully customizable audio environment for the user, and potentially larger conference sizes. Future work in this area needs to address the modelling of multiple groups and optimal server locations, some of which has been discussed by (Venkatesha et al., 2005), as well as investigation into the psychoacoustic limits for conferences with fixed spatial locations. The limit to the maximum number of simultaneous conversation floors, and hence simultaneous speakers needs to be found. This may require an analysis of users conversing within a spatially enhanced environment, in order to determine how difficult they find it to communicate. However, early experiments point to spatialized audio conferencing as a highly attractive technology.

REFERENCES

Baldis, J. (2001). Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 166–173. ACM Press.

Begault, D. and Wenzel, E. (1993). Headphone localization of speech. *Human Factors*, 35:361–376.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.

Brungart, D. and Simpson, B. (2003). Optimizing the spatial configuration of a seven-talker speech display. In *Proceedings of the 2003 International Conference on Auditory Display.*

Cheng, C. and Wakefield, G. H. (2001). Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *J. of the AES*, 49:231–249.

Cherry, E. C. (1953). Some experiments on the recognition of speech, withone and with two ears. *J. Acoust. Soc. Am*, 25(5):975–979.

Crispien, K. and Ehrenberg, T. (1995). Evaluation of the cocktail party effect for multiple speech stimuli within a spatial audio display. *J. of the Aud. Eng. Soc*, 43(11):932–941.

Egbert, M. (1997). Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language and Social Interaction*, 30(1):1–51.

Evans, M., Tew, A., and Angus, J. (2000). Perceived performance of loudspeaker-spatialized speech for teleconferencing. *J. of the Aud. Eng. Soc*, 48(9):771–785.

Kilgore, R., Chignell, M., and Smith, P. (2003). Spatialized audioconferencing: what are the benefits? In *Conference of the Centre for Advanced Studies on Collaborative research*, pages 135–144.

Knightson, K., Morita, N., and Towle, T. (2005). NGN architecture: generic principles, functional architecture, and implementation. *IEEE Communications Magazine*, 43(10):49–56.

Singh, K., Nair, G., and H, S. (2001). Centralized conferencing using SIP. In *Proceedings of the 2nd IP-Telephony Workshop (IPTel).*

Venkatesha, P., Jamadagni, H., and Shankar, H. (2003). On the problem of specifying the number of floors for a voice-only conference on packet networks. In *ITRE2003: International Conference on Information Technology: Research and Education*.

Venkatesha, P., Shankar, H., Jamadagni, H., and Vijay, S. (2005). Server allocation algorithms for VoIP conference. In *Proceedings of the First International Conference on Distributed Frameworks for Multimedia Applications.*