Enhanced Reality Audio in Interactive Networked Environments

Vicky Hardman and Marcus Iken

Abstract

The human auditory system has a two-fold function, to act as a receiver for spoken language, and also as an interpreter of the omni-directional environment that the listener is in (to direct the eyes). The acceptance of a network audio tool depends on both its ability to provide communication between remote participants, and its ability to immerse the listener in a natural auditory environment. The first part of this paper describes the multimedia conferencing piloting activities at UCL, that high-lighted the requirement for improved and more natural audio. The Robust-Audio tool was initially developed to address problems with speech intelligibility, that stem from the use of low-cost shared networks, and general purpose hardware. Recent developments in RAT have addressed the second function of the auditory system and provide 3D spatialisation supported by higher bandwidth speech,

1. Introduction

There are many applications for group-based virtual reality systems, such as visualisation, enhanced reality multimedia conferencing, and entertainment systems. These applications use sound localisation to enhance the presence in the artificial world. Current group-based virtual reality systems use high quality, high bandwidth audio, with specialist hardware to provide 3D spatial audio. If this technology is to mature quickly, then the cost (network bandwidth and computing power) must be reduced, and the systems tested by many simultaneous users on a global basis.

In contrast, the approach presented in this paper is based on the Mbone (the multicast backbone overlay on top of the Internet), which is a multi-way global network that is used to carry multimedia traffic. A variety of multimedia conferencing tools have been developed by various research projects to assess the usability of solutions to the low-cost requirements. Evaluation results from application trials are used to focus new research. Our experience shows that audio is the most important media, in terms of effective communication, and the one with the biggest problems. The audio problems identified so far include, sensitivity to network packet loss, gaps in the audio caused by the lack of real-time support on general purpose hosts, poor hands-free operation, restricted speech intelligibility, and lack of distance cues for speakers.

The Robust-Audio Tool (RAT) is a multi-way multimedia conferencing audio tool that has recently been developed in the Computer Science Department, UCL. RAT offers benefits over most other available audio tools, since it provides packet loss protection, and a mechanism that copes with the lack of real-time support found on most general purpose computing facilities. RAT also offers improved hands-free operation.

Recent work at UCL has begun to address problems such as restricted intelligibility, and the lack of distance cues. The ability to perform real-time 3D audio spatialisation has been added into RAT without any extra specialist hardware. In conjunction, a wideband speech coding algorithm, suitable for use over packet networks, is also being developed for RAT, which will be used to minimise the network bandwidth required by 3D audio. The latest developments begin to address some of the remaining audio problems in multimedia conferences, and produce an audio tool which is suitable for use in virtual reality environments. The recent ability of general purpose computing facilities to support this level of computation, and the emergence of the Mbone as a useable multi-way multimedia network means that virtual reality systems can begin to be realised with low-cost hardware, and evaluated on a large scale.

This work stems from multimedia conferencing piloting applications, such as project MICE[1] (Multimedia Integrated Conferencing for Europe, and its follow-on Project MERCI - ESPRIT), project ReLaTe[2] (Remote Language Teaching over SuperJANET - BT/JISC, and its follow-on ReLaTe/2 - JISC), and more recently the RAT project[3] (EPSRC). Virtual reality work at UCL includes projects such as COVEN[4], and DEVRL[5]. We envisage the possible integration of RAT into Coven.

Some background is given in the first section, which discusses the applications projects in the department. The second section discusses the evaluation results relating to audio, and identifies possible improvements. The third section of the paper briefly discusses the RAT architecture, and highlights some of the recent developments, which improve the quality of the audio over the network, and on the host platform. The fourth section concentrates on the implementation of sound localisation in RAT.

2. Networked Virtual and Enhanced Reality Applications

Networked virtual and enhanced reality applications all have one thing in common; the requirement to communicate over a network in a manner similar to that used in the face-to-face situation.

The applications include the following:

- enhanced reality multimedia conferencing [6][7][8]
- information visualisation (VR-VIBE[9])

Multimedia conferencing is the technology that under-pins many of the new virtual and enhanced reality applications, and it comprises audio, video and shared workspace. Of the three media, audio is still the most important means of communication.

2.1 Multimedia Conferencing

The emergence of a global low-cost shared network - the Mbone - has meant that multimedia conferencing applications are beginning to migrate from the research environment into the commercial world. A strategy of restricting costs enables a critical mass of users to be established, which allows the technology to mature through design iterations and improvements. Low cost multimedia conferencing solutions can only be achieved by using human perception techniques to focus precious computing resources, for example; image rendering in virtual reality worlds is often of much lower resolution when an object is far away, some audio tools in multimedia conferencing only mix the maximum number of sources that a human can discern.

2.2 Low-cost Network and Platforms

The Mbone provides multi-way communication by using a technique known as multicast [10]. Multicasting enables the transmission of multi-way real-time audio and video communication to interested participants, without having to use either complicated set-up procedures (circuits), or consume large amounts of bandwidth (broadcast). The network nodes contrive to transmit only to interested receivers (figure 1).

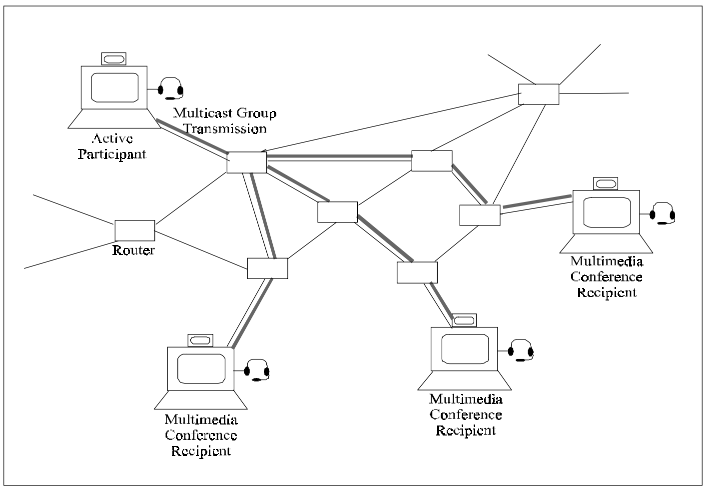


Figure 1: Multicast Transmission over the Mbone

The Mbone is a shared heterogenous network, where information is transmitted in packets. Communication is usually on a ‘best-effort’ basis, which means that there are no quality of service guarantees (the ability to perform resource reservation is currently being incrementally deployed[11]). The heterogenous nature of the Internet and Mbone has meant rapid global deployment, and cheap network access charges.

The availability of general purpose hosts, such as UNIX-based workstations, and more recently multimedia personal computers, is another factor in the low-cost strategy that promotes a critical mass of users. This is especially true if extra equipment is kept to a minimum. Multimedia conferencing components are consequently implemented as separate software processes on the multi-tasking operating systems, and the implication is that the host must be able to perform real-time compression and delivery (rendering) of the media in software.

2.3 Multimedia Conferencing Components

Multimedia conferencing over the Mbone provides multiway audio, video and shared text facilities:

- **Audio tools**, such as vat [12], and RAT [13], enable every participant to hear everybody else in the conference simultaneously. The system interface usually comprises a window showing a list of the participants in the audio conference, gain and volume controls, power meters to indicate speech activity etc. (Figure 2). In order to talk, the pointer has to be placed in the window, and the left mouse button pressed. This mechanism is known as 'push-to-talk'.

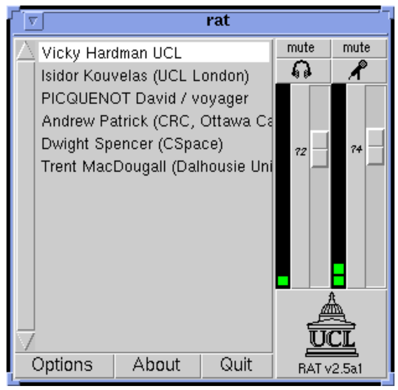


Figure 2: Audio Tool Window

- **Video tools**, such as vic [14], enable every participant to see video from every other participant (depending on the available bandwidth). Each video image is displayed in a separate window, and these can be arbitrarily positioned on the screen.
- **Shared text and whiteboard tools**, such as wb [15], and nt [16] enable each participant to see and edit the same text (wb and nt) and images (wb). Text and drawings can be entered via the mouse and keyboard, and each participant draws in a separate colour. Previously stored postscript files can also be imported.

2.4 Business Meeting and Distance Learning Applications

Real applications are currently being piloted over the Mbone, such as business meetings [1], and distance learning trials [2][1], using audio, video and shared text tools. The ReLaTe project is trialing remote language teaching over the Mbone, using the tools RAT [13], vic [14], and wb [15] / nt [16], and has produced an integrated interface for novice computer users (figure 3). Integration of audio and video information was achieved by using a conference bus, rather than a more conventional method of inter-process communication. A conference bus transmits multicast information using a ttl of zero, which retains all the benefits of multicast - multi-way communication, and dynamic group membership - while restricting the traffic inside the workstation [17].

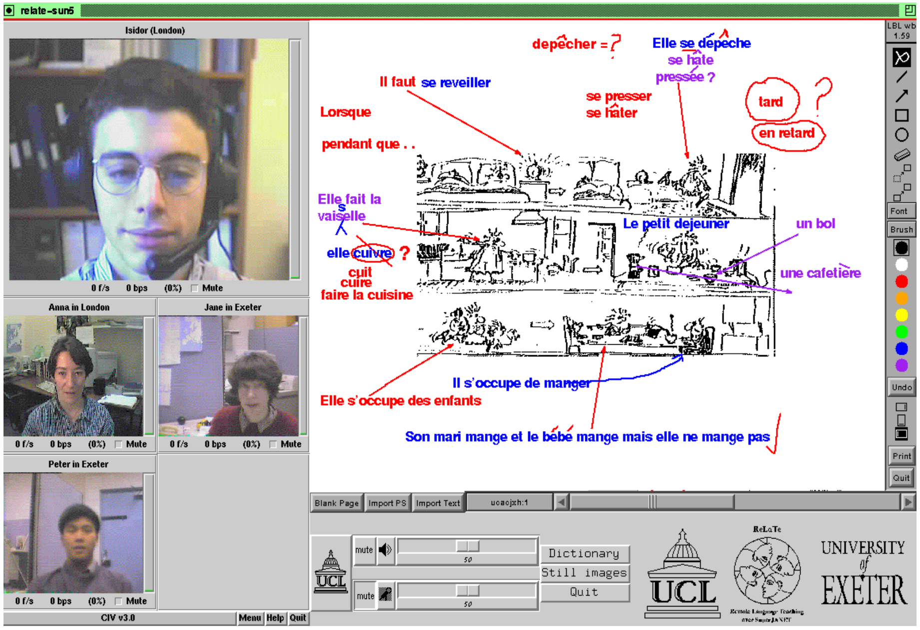


Figure 3: The ReLaTe Interface

2.5 Trial Observations

Informal evaluation in the MICE project identified audio as the most important medium. This was borne out by formal evaluation during the ReLaTe project, from language teachers, HCI and technical experts, which also identified poor audio quality as being the major problem with the system. A detailed description of the evaluation can be found in [18].

3. Networked Enhanced Reality Audio

Audio was identified early on in the ReLaTe project as being the primary area of concern. The problem can be broken into a number of symptoms that are true of most Mbone multimedia conferences compared to face-to-face situations:

- ***Audio Gaps***

  Audio packet loss results in gaps in the output speech, and is very detrimental to speech intelligibility above 10% loss. Gaps in the output speech can also occur because of the lack of real-time support on general purpose operating systems, such as UNIX; samples have to be regularly fed to the output device, and a missed dead line results in a gap.

- ***Lack of Hands-free Operation***

  The lack of hands-free operation was due to the use of the ‘push-to-talk’ facility. Use of the ‘push-to-talk’ mechanism reduced the spontaneity and interactivity of the conversation [18].

- ***Reduced Speech Intelligibility***

  Restricted speech intelligibility of an isolated speaker is a result of using telephone quality speech. Monaural hearing reduces the intelligibility of a speaker when another person speaks (cocktail party effect [19]).

- ***Speaker Identification Difficult***

  Speaker identification is difficult in most audio tools, since a listener has to look down the list of participants for a high-lighted name (see figure 2). The ReLaTe system (figure 3) improved the situation substantially by associating individual audio power meters with video images, but this technique is still artificial.
- ***Lack of Distance Cues***

  The audio channel provided by existing audio tools sounds acoustically 'dead', since there are no distance clues in the audio; speech is provided over head-phones, and the audio sounds as if it is coming from inside the head. The lack of distance clues means that some speakers tend to shout, while others speak quietly.

RAT has been developed as a platform for multicast audio research, and the intention is to address the problems identified above.

Mechanisms to repair packet loss over the network, and to cope with the lack of real-time support on general purpose operating systems have been developed at UCL [20][21], and initial investigations show that significant success has been attained. The lack of hands-free operation has also been addressed by the development of an improved silence detection mechanism. Further work in all of these areas is planned under Project RAT [3].

It is well known that wide-band speech (16kHz, rather than 8kHz sampling rate) provides greater speaker intelligibility, and retains more of the individual characteristics of a voice, which allows easier participant identification. A reduced complexity wide-band speech codec [22] is being developed for RAT as part of an MRES project at UCL.

Sound localisation [8] has the ability to both position sound sources out in space (which facilitates easier participant identification), and provide some distance information (which will allow a speaker to subconsciously determine how loudly to speak). Sound localisation is usually provided in both multimedia conferencing systems and virtual reality systems by special purpose hardware, such as the convolvatron [23] and beachtron [24] cards. The DSP-based hardware has the processing power to artificially convert a mono sound source into localised sound using the Head Related Transfer Functions (HRTF - [25]), and reverberation [8]. HRTFs simulate the inter-aural intensity and phase differences of sound arriving at the two ears, together with the effect of head shadows, the effect of the shoulders, and reflections from the pinnae [23]. In order to successfully externalise the sound, a minimum sampling frequency of 16-32kHz is required [25]. Uncompressed transmission of this bandwidth would result in excessive network bandwidth consumption, which should be avoided if at all possible. The 16kHz wide-band speech compression algorithm [22] can be used to support sound localisation, and the algorithm produces an output bit-rate of 64kbps, which is suitable for transmission over the Mbone. The codec has not been developed for use over packet networks, but current work at UCL is investigating this problem. The codec provides good enough quality audio for use with sound localisation techniques. The provision of sound localisation via the HRTFs in RAT is being developed at UCL.

In the telephone network, distance cues are provided to a speaker by the use of side-tone[26] (leakage between the microphone and loudspeaker). The provision of side-tone and automatic gain control [27] are also being developed for RAT.

4. A Robust-Audio Tool

The robust-audio tool (RAT) emerged at UCL as a result of experiences gained from applications piloting activities (projects MICE and ReLaTe). RAT is based on existing audio tools, such as vat [12], but it also includes new techniques developed at UCL which improve the quality of the audio. RAT is currently available for a wide range of platforms: SunOS, Solaris, IRIX, and Win95, and is currently being ported to HP, DEC Alpha, Linux, and FreeBSD.

A block diagram of RAT can be seen in the diagram (figure 4). The audio tool has been designed for use with either headsets or a microphone and loudspeaker. At the transmitter, speech samples collected from the microphone are used to fill a packet, which is then transmitted over the network. At the receiver, the samples are retrieved from the packet, and sent to the audio device driver, where they are played out to the headphones or loudspeaker.

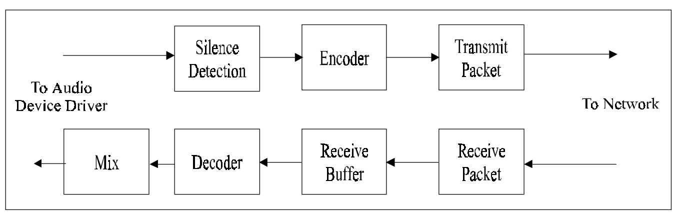


Figure 4: Block diagram of RAT

4.1 Silence Detection

Silence detection [27] is commonly used in packet network speech systems to reduce the bandwidth consumed. Silence is not transmitted, and since at least half of speech is 'silence', this represents a significant saving. In a multi-way environment, usually only one person is speaking at once, and silence detection then has the potential to reduce the bandwidth consumed to approximately half of one normal telephone channel. Silence detection has implications for 3D spatial audio, since if only one person is speaking at once, then only one sound source needs to be manipulated. Work at UCL as part of the RAT project will improve the quality of the silence detection algorithm.

4.2 Real-Time Protocol (RTP)

Real-Time Protocol (RTP) was developed for the transmission of real-time audio and video over the Mbone[28]. RTP provides facilities such as time-stamping, session control etc. Initial proposals from others at UCL, to develop networked distributed virtual reality protocols for use in a shared multicast environment propose the use of RTP, and can be found in [29].

4.3 Speech Compression

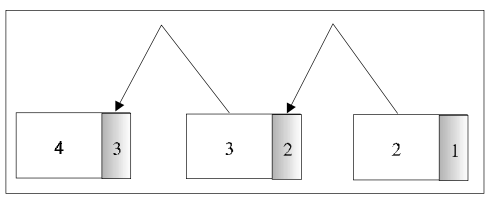
Speech compression algorithms are used to reduce the packet rate, which improves the packet loss rate. Existing audio tools transmit telephone quality speech (8kHz sampling rate). 16kHz sampled speech can be coded using the wide-band coding algorithm [22], and the bandwidth consumption restricted to reasonable levels. Wideband speech improves the quality of audio, such that individual speakers can be recognised. The sampling rate offered by wideband speech (16kHz) is also the minimum that is required in order to externalise 3D sound [25].

4.4 Packet Loss

Packets may be lost over the network, since routers throw away packets when they get congested. The impact of packet loss on speech quality can be very severe, especially given the packet sizes commonly used over the Mbone [20]. Work at UCL as part of the RAT project [3] is developing techniques to improve packet loss robustness. Speech from existing audio tools is intelligible for packet loss up to 10%, whereas from RAT it is intelligible for up to 30% [13].

4.5 Packet Loss Robustness in RAT

At high loss rates, and for the length of packets commonly used over the Mbone, packet loss can only be successfully repaired by the use of redundancy [20]. Redundancy transmits a low bandwidth



synthetic version of the speech on the packet behind, which can be used in the event of single packet loss to repair the speech at the receiver (figure 5).

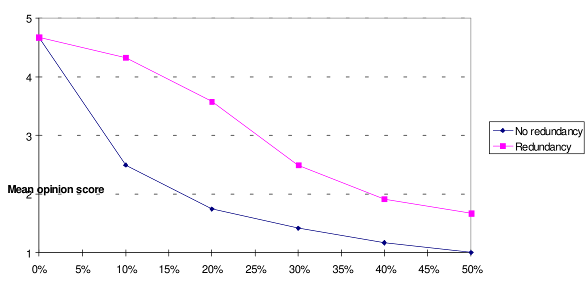


Figure 5: Redundancy for the audio is carried in the packet behind

This mechanism provides satisfactory speech quality for up to 30% loss (figure 6).

Figure 6: Subjective Assessment of Packet Loss Repaired Using Redundancy

Bursts of packet loss can be repaired by using multiple copies of redundancy, although what is really required is to reduce the offered load on the network, by increasing the compression level [20]. This work is currently being accomplished under Project RAT [3].

4.6 Adaptive Workstation Real-time Performance Management

The lack of support for real-time applications on general purpose computers may produce gaps in the output audio. In RAT, the adaptive cushion algorithm [21] estimates the current work-load of the host, and uses the device driver buffer to minimise the occurrence of gaps in the audio, while minimising the delay. The results show both a decrease in the number and length of audio gaps, and a decrease in the mean and variance of the perceived end-to-end delay, compared to existing audio tools.

5. Networked 3D Spatial Audio

Networks usually only transmit one channel at low bandwidth to minimise costs. A monaural signal is known to affect many aspects of the perception of audio (such as speaker identification difficult, and the lack of distance cues). The provision of natural two channel sound is obviously desirable.

5.1 3D Audio

A mono signal can be transformed into an apparently spatial sound source, by duplicating the mono signal, and then manipulating the two channels individually to simulate spatialised sound.

Intuitively, one might think that free-field loudspeakers could be used to provide 3D sound, but the problems of cross-talk (signals from the left loudspeaker reaching the right ear and vice versa), make this problem more complex (cross-talk cancellation algorithms are required) than when using headsets. In contrast, cross-talk is not a problem for headsets, but the use of head-tracking devices is still required, or else the location of a sound object does not change position relative to the user's head. In multimedia conferencing, headsets are commonly used with audio tools, such as RAT, to ease problems with hands-free operation, and so the following descriptions apply to the use of headsets only.

There have been two approaches to providing spatialised audio: lateralisation and localisation.

- ***Lateralisation***

  Speech based on mono signals is perceived as originating from the midpoint between the two ears, and usually more at the back of the head. Lateralisation separates conference participants out in space, although all the sounds appear to still come from inside the head.

  To achieve pure lateralisation, the application of the duplex theory [30] is sufficient; the channel being sent to the far ear has to experience a delay and a dampening of the amplitude relative to the channel being sent to the near ear (figure 7).

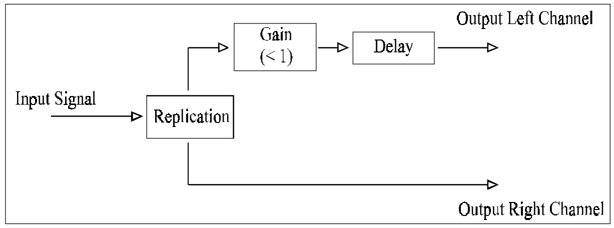


Figure 7: Lateralisation from a mono input channel.

However, this simple approach to lateralisation does not take into account the frequency dependency of the duplex cues.

- ***Localization***

Localization aims to transform the incoming signal in a way that is faithful to the real life situation; this is what the head-related transfer function (HRTF) does [25], and since sampled speech is a sequence of impulses, it can be spatialised by convolving the HRTF data set with the input audio.

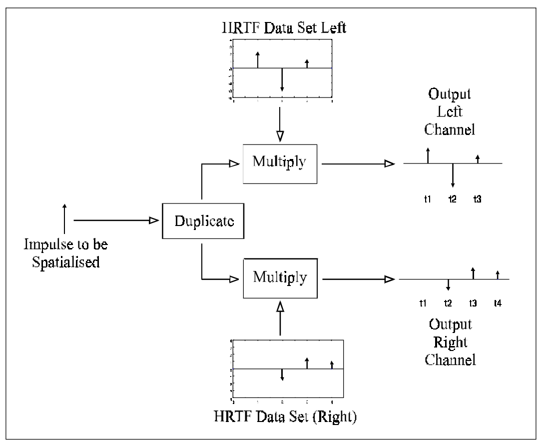


Figure 8: Using the HRTF data sets to produce Localisation.

In contrast to the operations necessary for achieving mere lateralisation, convolutions are costly in terms of computing resources. In the time domain the complexity is O(n*m), in the frequency domain O(n*log2n).

5.2 RAT and 3D Audio

Sound localisation is being developed for RAT in order to address a number of problems identified during ReLaTe [2] and MICE [1] trials. The environment that this system has to work in means that localisation blur [25] is not as critical as externalising the sound, since we wish to provide a natural acoustic environment for conference participants, and improve the intelligibility of what is being said, by separating the sources out in space. Maintaining the low-cost strategy employed in these projects, necessarily involves making some compromises in the 3D system design. The performance of the enhanced RAT audio system will be reduced compared to that achievable using dedicated hardware, but we intend to back up this development with extensive human perception analysis, in order to provide a system that is adequate for the requirements of multi-way networked enhanced reality multimedia conferencing.

The low-cost constraints of this development are as follows:

- real-time processing on modern general purpose multi-tasking operating systems (no DSP boards or separate special-purpose hardware)
- minimal use of computing power (other multimedia tools need processor time)
- minimal network bandwidth (currently this means using low sampling rates, 16 kHz instead of 32kHz)

As can be seen from the above, the criteria for a low-cost solution are different from the goals pursued in current VR and perceptual psychology research, where highly accurate localisation is sought [23]. Within our low-cost strategy, the desired improvement is a qualitative step from a mono signal being perceived inside the head to externalised sound appearing outside the head, and with reasonable azimuth separation. The externalisation is further helped in scenarios that include the visual as well as the auditory sense, as is the case in multimedia conferencing. Visual awareness of the apparent sound sources (i.e. conference participants) improves externalisation, and reduces the occurrence of front/back reversals [31].

In the light of the different circumstances and objectives compared to mainstream virtual reality research, a low-cost strategy for the general design of the module can be pursued.

5.2.1 Design and Implementation

Localisation is known to consist of two classes of effects: diffractions caused by the upper torso, and the duplex cues [31]. The HRTF data set [32] was empirically measured, and therefore includes both of these effects. Analysis of the HRIR (impulse responses) shows that the delay and attenuation of the signal can be identified. Since convolution is very costly in terms of computing power, complexity reduction techniques were sought for implementation in RAT.

In the RAT 3D spatial audio system, the HRTF transformation has been split up into two distinctive stages:

- effects that are not present in the pure mono signal, but make the signal appear to have been transferred from the free-field to the eardrum
- duplex cues, i.e. delay and gain

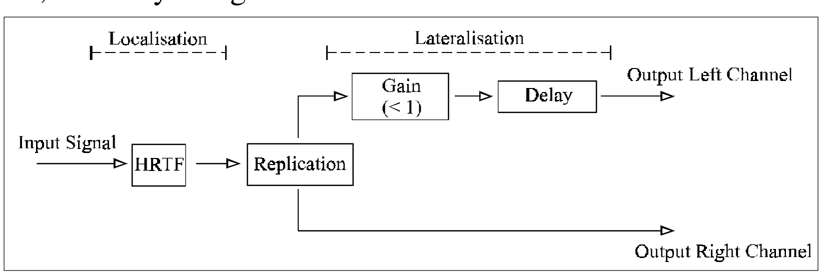


Figure 9: Use of the HRTF Data Set in RAT

Figure 9 shows that the first stage is a convolution with an HRTF data set which results in the externalization of the signal. The second stage implements the duplex cues (lateralisation). Since the delay and gain operations consume negligible processing power compared to the convolution, this design requires roughly half the computational power that would be used in the approach outlined in Figure 8. This would be well within the real-time capabilities of general purpose workstations.

The RAT architecture has been designed to allow relatively painless enhancement. Figure 10 illustrates the integration of the sound localisation module into RAT.

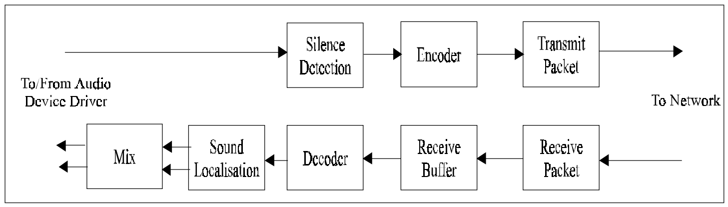


Figure 10: The Integration of 3D Audio into RAT

Upon joining a multicast group, each participant is assigned an azimuth on the horizontal. Using the relevant HRTF data set for that azimuth, appropriate delay and gain values are calculated. After the received packets are buffered and decoded, the speech samples are convolved with a chosen data set, and the result replicated. The far ear samples are then multiplied with the gain factor (IID), and delayed by a number of samples equivalent to the calculated delay (IDT).

5.2.2 Selection and Preparation of HRTF Data Sets

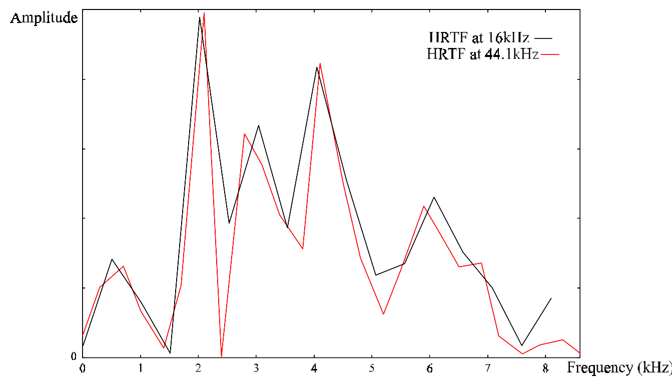
Due to the enormous amount of work needed to measure an adequate number of related HRTF data sets, only one has so far been released into the public domain (data recorded at MIT using a KEMAR dummy [32]), although others exist [25]. The MIT data set was recorded at a sampling rate of 44.1kHz, and two versions are provided: the recorded sets and the diffuse field sets. The recorded data sets contain some unwanted components: inaccuracies of the measurement set-up, and most importantly, the ear-canal response. Using headsets means that an ear-canal response will be included naturally, so the diffuse-field version must be used instead.

The data sets were re-sampled to 16 kHz using a program based on the sinc-pulse function to restore the continuous time signal [33]. This reduces the filter length from 512 values to (512 * 16/44.1 =) 185 values. Using the fact that the 64 most significant values of a 512 value data set produce nearly the same frequency response as the complete set [Begault], the (64 * 16/44.1 =) 24 most relevant values are extracted from the 185 resulting from the resampling operation.

In the implementation of this operation in RAT, we use a sliding window to isolate each possible combination of 24 consecutive taps, and sum the absolute values. The greatest resulting value is taken to mean that we have identified the ‘good taps’ [31] in each filter. This is done for both data sets relating to a particular azimuth. The next state is to identify the delay and gain cues from the two filters. This is achieved by calculating the delay from the separation between the ‘good taps’ for each filter, and by evaluating the power of near ear divided by the power of far ear to give the gain.

5.2.3 Preliminary Results

The RAT interface was enhanced to allow the user to switch between monaural sound, and filtered binaural sound. Preliminary informal listening tests revealed that listeners were able to report externalisation of the sound source. Plots of the frequency responses of the near ear HRTF data sets at



44.1kHz and 16kHz can be seen below.

Figure 11: Comparison of the Frequency Responses of an HRTF Data Set Used in RAT, sampled at 16kHz with identification of 32 ‘Good Taps’ to the original data set, sampled at 44.1 kHz and 512 taps.

The graph (figure 11) shows that down-sampling and extracting a sub-set of significant taps does not affect the general shape of the frequency response.

6. Conclusion

The paper has shown approaches to solve audio problems that invariably appear with multimedia conferences over ‘shared’ networks, and that use only general purpose hardware. The approaches described will provide higher acceptance of an audio tool like RAT, because they address the two main functions of the human auditory system.

Solutions were presented that tackle the problems of audio, such as gaps in the output stream (resulting from packet loss and the lack of real-time support on general purpose operating systems), and the lack of hands-free operation, that initially proved to be such problems for our applications piloting projects. These solutions now provide a greater robustness to the audio output, and therefore to the communication between remote participants.

Later evaluation of the piloting projects, identified further potential areas of improvement, specifically providing a more immersive environment. The techniques researched and integrated into RAT employ some novel complexity reduction methods that enable real-time operation in a multi-tasking environment, using exclusively general purpose hardware.

7. Future Work

The success of the pre-split HRTF design as outlined in section 5.2.1. crucially depends on the quality of the data set applied. The use of a down-sampled individual data set can only be regarded as a starting point.

Work in the near future will comprise:

- the use of generalised HRTF data that exploits the common features of individual data sets.
- informal 'playing' with the data sets accompanied by human perception evaluation with regards to the externalisation and reverberation achieved.

Work in the medium future might include the implementation of proper distance cues (ratio of direct to reverberation signals and dampening of high frequency components).

Acknowledgments

We would like to acknowledge the following, who work together to develop RAT using a wide variety of backgrounds and specialist knowledge: Angela, Isidor, Anna, Colin, Orion, Mark, Jon, and Peter. We would also like to acknowledge Isidor, Anna, Mel and Anthony for their comments on this paper.

Biographies

Vicky Hardman is a newly appointed lecturer in the Multimedia Group, Department of Computer Science, UCL. She has worked on a number of national and international research projects in multimedia conferencing, such as ReLaTe (BT/JISC - where she was the research fellow on the project), MICE (ESPRIT - which pilots multimedia conferencing systems), and Unison (ALVEY - which developed a multimedia packet network). Recently, she has been made co-investigator of the continuation of the ReLaTe project (JISC), and co-investigator of an EPSRC project (RAT #GR/K72780). Current research areas are speech over packet networks, and enhanced reality systems for use in multicast packet networks. Vicky has a PhD in Speech over Packet Networks, from Department of Electronic and Electrical Engineering, Loughborough University of Technology. She has also worked for voice switching manufacturers for 3 years, on CCS7 and CAS voice networks.

Marcus Iken is a final year student at Stuttgart University. He is currently undertaking his final year research project at UCL with Vicky Hardman. Marcus hopes to study a PhD at UCL in sound localisation.

References

[1]Handley M.,Kirstein P.,Sasse M.A., 'Multimedia Integrated Conferencing for European Researchers (MICE): piloting activities' Computer Networks and ISDN Systems,26(3),275-290, 1995.

[2] Buckett J. Campbell I. Watson T.J. Sasse M.A. Hardman V.J. Watson A. 'ReLaTe: Remote Language Teaching over SuperJANET' Proceedings of UKERNA 95, Networkshop, March 1995.

[3] 'Multi-way Multicast Speech for Multimedia Conferencing over Heterogenous Shared Packet Networks (RAT Robust-Audio Tool)', EPSRC Research Grant #GR/K72780.

[4] 'Coven: Collaborative Virtual Environments', ACTS 4th Framework Project.

[5] 'DEVRL: Distributed Extensible Virtual Reality' EPSRC:ROPA Project #AC040.

[6] Hardman V.J., 'Low-cost Multi-way Distance Learning' Proceedings of the Coseners Network Shop, 1996.

[7] Yngvesson J., Kvarnstrom B., 'Telepresens in Conference Applications' MultiG Workshop 1992.

[8] Strauss, Blauert J., 'Virtual Auditory Environments', Proceedings of the Conference of the FIVE Working Group (ESPRIT 9122), QMW, University of London, 18-19 December 1995.

[9] Benford, S., Snowdon D., Greenhalgh C., Ingram R., Knox I., Brown C., 'VR-VIBE: A Virtual Environment for Co-operative Information Retrieval' Eurographics 95, Publ. Blackwell.

[10] Deering S., 'Host Exetensions for IP Multicasting. Request for comments RFC 1112, Internet Engineering Task Force, August 1989.

[11] Zhang L., Deering S., Estrin D., Shenker S., Zappala D.,'RSVP: A new Resource Reservation Protocol', IEEE Net, Sept. 1993., Volume 7., No. 5, pp.8-18.

[12] Jacobson V. ‘VAT manual pages’, Lawrence Berkeley Laboratory (LBL) February 1992.
[13] Hardman V., Kouvelas I., Sasse M.A., Watson A., ‘A Packet Loss Robust-Audio Tool for Use over the Mbone’ Department of Computer Science Research Note, RN/96/8
[14]McCanne S., Jacobsen V., ‘Vic: A Flexible Framework for Packet Video. Proceedings of ACM Multimedia ‘95, November 1995.’
[15]McCanne S., ‘A Distributed Whiteboard for Network Conferencing’, May 1992, UC Berkeley CS 268 Computer Networks term Project.
[16] Handley M., ‘A Shared Text Editor’ URL: http://www.cs.ucl.ac.uk/mice/mice_home.html
[17] Handley M.J., Wakeman I., ‘CCCP: Conference Control Channel Protocol, A Scaleable Base for Building Conference Control Applications’ Conference Proceedings ACM SIGCOMM95, Cambridge Massachusetts, 1995.
[18] Watson A., Sasse M.A., ‘Evaluating Audio and Video Quality in Multimedia Conferencing Systems’ To appear in Interacting with Computers Journal.
[19] Cherry C.E., ‘Some Experiemnts on the Recognition of Speech, with One and Two Ears’ Journal of the Acoustical Society of America, Vol. 25, No. %, pp. 975-979, USA.
[20] Hardman V., Sasse M.A., Handley M., Watson A., ‘Robust Audio for Use over the Internet’, INET 95, Honalulu, Oahu, Hawaii, June 1995.
[21] Kouvelas, I., Hardman V.J.’Overcoming Workstation Scheduling Problems in a Real-Time Audio Tool’ Dept. Computer Science, UCL, Research Note No. RN/96/44.
[22] Hardman V.J., ‘Wide-band Speech Teleconferencingover an Integrated Network’, PhD Thesis, Dept. Electronic and Electrical Engineering, Loughborough University of Technology, Loughborough, Leicestershire, November 1993.
[23] Wenzel E.M.‘Localisation in Virtual Acoustic Displays’ Published in Presence,Volume1,Number 1, pp80 - 107.
[24] ‘Beachtron Card: Crystal River Engineering’, URL http://www.cre.com/index.html
[25] Wightman F.L, Kistler D.J., ‘Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis’ Journal of the Acoustical Socity of America, Vol. 85, No. 2, 1989, pp. 858-867, USA.
[26] Richards D.L.’ Telecommunication by Speech’ Pub. Butterworth and Co. 1973.
[27] Rabiner L.R., Schafer R.W., ‘Digital processing of Speech Signals’ Publ. Prentice Hall, 1978.
[28] ‘RTP: A Transport Protocol for Real-Time Applications’, Audio-Video Transport WG, rfc 1889.
[29] Crowcroft J., Handley M., ‘Network Support for Scaleable Distributed Multiparty Virtual Reality’ Conference EPFL-UCB, Workshop on Multimedia Networking 96, Lausanne, Switzerland, July 11-12, 1996.
[30] Blauert J., ‘Spatial Hearing, The Psychophysics of Human Sound Localisation’, The MIT Press, Massachusetts, USA, 1983.
[31] Begault D.R. ‘3D Sound for Virtual Reality and Multimedia’ Published by Academic Press, 1994.
[32] Gardner B., Martin K., ‘HRTF Measurements of a KEMAR Dummy-head Microphone’ Technical Paper No. 280, MIT Media Lab. Perceptual Computing, Dept. Electrical Engineering and Computer Science, USA, 1994.
[33] Van Den Enden A.W.M., Verhoeckx N.A.M.’Discrete Signal Processing - An Introduction’, Pub. Prentice Hall, 1989.