Enabling Composition-Based Video-Conferencing for the Home

Jack Jansen, Pablo Cesar, Dick C.A. Bulterman, Tim Stevens, Ian Kegel and Jochen Issing, *Member, IEEE*

***Abstract*— This paper describes a videoconferencing system that meets performance constraints and functional requirements for use in consumer homes. Our system improves existing home technologies (such as video chat) by providing high-quality audiovisual communication, efficient encoding mechanisms, and low end-to-end delay. Moreover, the system includes a control interface that is capable of dynamically manipulating and compositing audiovisual content streams. This innovative architectural component is required for a domestic setting, where the television acts as the main screen and multiple people gather around it. Apart from the requirements and architecture, this paper analyses the performance of our system. The results validate our architectural decisions and provide a valuable input for further research in domestic videoconferencing.**

***Index Terms*— Videoconferencing and Collaboration Environments, Presentation of Content in Multimedia Sessions, Compression and Coding, Standards and Related Issues, Consumer Electronics and Entertainment.**

This work was supported in part by funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2007-214793.
Jack Jansen, Pablo Cesar and Dick C.A. Bulterman are with the Centrum voor Wiskunde & Informatica, Amsterdam, the Netherlands (phone: +31 20 5924303; Jack.Jansen@cwi.nl, p.s.cesar@cwi.nl, dick.bulterman@cwi.nl).
Tim Stevens and Ian Kegel are with BT Research & Technology, Ipswich, UK.
Jochen Issing is with the University of Erlangen-Nuremberg, Germany.

I. INTRODUCTION

The application of high-quality videoconferencing has been typically restricted to the office setting, mainly in the form of carefully architected shared meeting rooms [1] with dedicated networking hardware and software. The growth in networking bandwidth and compute power available at home is beginning to enable high-quality videoconferencing in a domestic setting. By mid-2010, key players like Skype and Cisco had announced systems operating via the Internet, using available home devices such as a high-definition television and requiring at least 1.2 Mbits/s of symmetric bandwidth.

It is tempting to think of home videoconferencing as primarily a problem of supplying reliable network bandwidth. We think, however, that the qualitative difference between the way people interact within a business setting and at home [2] requires support for new forms of conferencing dynamism that are not yet adequately supported in office settings. In a meeting, joint interaction is normally centered on a focused presentation metaphor, augmented by incidental verbal communication/discussion. Users sit in relatively fixed positions for the duration of the conference. In this environment, it is not unusual that people interact within this context for several hours. In most households, the position of participants is more dynamic and movable, with any interaction being centered around a physical shared activity. In addition, a home living room - unlike a dedicated meeting room - is inherently informal and multi-functional, which puts constraints on core conferencing aspects, such as the number (and placement) of screens, cameras and microphones. Fig. 1 illustrates the problems with transferring basic videoconferencing technology to the

home: rather than having the users adapt to the meeting room, the meeting must adapt to a heterogeneous user space.

Despite current limitations, videoconferencing between households is filtering into everyday use. Recent studies on human factors have identified common practices and restrictions, when using Skype at home [3][4][5]. Two of the findings are of special interest: the need for an acceptable level of end-to-end performance and the need for supporting multiple people in front of the camera. Based on these findings, this article argues that supporting videoconferencing in the home requires more than a simple audiovisual link between households. A control substrate is required that can support dynamic manipulations on the audio/video content streams, so that movements of the participants, non-verbal interactions, and changes in the common activity can be captured and shared. In particular, this article identifies a number of performance and functional requirements for next-generation domestic videoconferencing.

Figure 1: Issues with using videoconferencing in the home.

Effective home-to-home videoconferencing systems require a bounded end-to-end delay similar to that in office systems. In terms of system performance, it is commonly assumed that one-way audio delays below 100 ms cannot be distinguished from fully synchronized video [6], while 150ms is considered to be the acceptable lower bound for delay in verbal communication [7]. Nevertheless, according to some experiments we conducted with commercial systems (as shown in section II) it seems that even longer delays are considered acceptable – up to 350 ms; meaning that effective conferencing will not be dependent on bandwidth alone.

Instead, the main problem to be solved is in (dynamically) constructing a stream of audio and video content among participants: the continuous video stream from a single camera used in commercial video conferencing rooms will not suffice for the living room. The wall-of-screens model that works well in an office-based videoconferencing system is not feasible for installation in most living rooms, where a single widescreen monitor that also serves as an actual television is the norm. A resulting functional requirement is the need to support changes in visual data (such as changes in camera angle and crop). This allows the social communication to better focus on the 'active' content while still requiring only moderate bandwidth. Some of the changes in the visual content are required to support social interaction [8], while other changes will be required to adapt to the time-variant availability of network resources to and from each home. Audio presents a similar set of challenges. If there are multiple people in the living room, the monaural channel provided by tools like Skype is not precise enough to allow speaker differentiation, and the office-based solution of providing a microphone for each participant is neither practical nor natural in the home setting. High-Quality directional audio through the use of a small, easily concealed, microphone array should solve this issue. Additionally, since home communication is often structured around a common activity (such as sharing media or playing a game), a relatively dynamic control of the content within the audiovisual chain (including seamless screen composition) is essential.

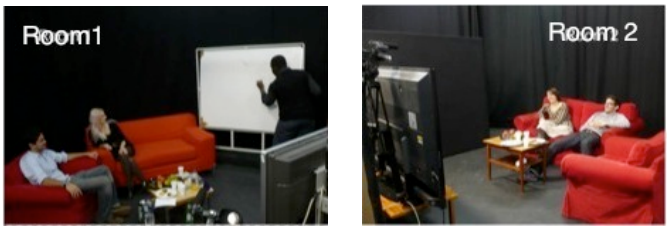

Figure 2. Groups in two locations playing Pictionary. Room 1 and Room 2 have a similar multi-camera setting.

Figure 3. Views from the four cameras in room 1. 1A: close-up of the board; 1B: long shot of the board; 1C: long shot of the couch; 1D: close-up of the couch.

In a preliminary lab study run in London [9], two groups of people remotely played Pictionary (Fig. 2). In order to integrate the shared activity (Pictionary playing) with group-to-group social interaction, four cameras were placed in the rooms (Fig. 3). For capturing what is happening in the rooms, it is necessary to provide video composition based on low-level cues, some of them related to the game (e.g., hand movement) and others related to the social interaction (e.g., eye-gaze). These video composition policies might include cropping a video, selecting a different camera, or including synthetic graphics on the screen.

Our targeted hardware platform is representative of the maximum that can be expected in private homes within a three-year window, given technical, financial and social constraints. The visual hardware in the home consists of an HD television and audio speakers for output, and three HD cameras and a microphone array for input. While video connectivity is HD 720p the central camera is a full 1080p HD camera from which a 720p image is cropped or scaled. This allows a large number of viewpoints to be selected for display at the remote side. Depending on the activity, there may be auxiliary screens or other hardware involved too. Invisible hardware is a 10 Mbit/s symmetrical network connection and 2-3 desktop-class machines.

Domestic videoconferencing can be studied from different perspectives: human-centred studies and comparisons, real-time content analysis and cue detection, semantic decision-making algorithms, and experimental technical issues. This paper focuses on the underlying infrastructure and the *mechanisms*, the technical issues that make composition-based videoconferencing for domestic use possible. Results on content analysis and cue detection [10], high-level interaction patterns detection [11], and semantic decision-making algorithms for dynamic composition (the *policy*) [9] are described elsewhere and are out of the scope of this paper.

The contribution of this paper is a control substrate that supports dynamic manipulations on the audio/video content streams. It includes two innovative software components for home videoconferencing systems: Audio Communication Engine (ACE) and Visual Composition Engine (VCE). The ACE is capable of high-quality audio communication providing multi-channel echo cancelation and directional audio, with an estimated delay below 50 ms. The VCE provides high-definition video communication (720p25) with a delay of 350 milliseconds. While this delay is high, it has been shown to be acceptable to end-users as will be shown in section II. While meeting the performance requirements for home-to-home communication, these two components are capable of dynamically

reacting to low-level cues (e.g., voice activity, eye-gaze, movements) [10], conversational patterns [11], and game play.

This paper is structured as follows. Section II describes related work, while Section III discusses the requirements based on a number of recent findings about videoconferencing usage at home. Section IV introduces our architecture and Sections V and VI report on the innovative components and their performance compared to previous solutions. Section VII discusses the next steps and provides a number of insights on the implications of bringing videoconferencing to the home.

II. Related Work

This section describes the state of the art of end-to-end conferencing systems from two perspectives. First, it discusses existing architectural solutions related to our system. Then, it reports on the performance of current commercial videoconferencing systems and research prototypes.

A. Architectural Solutions

Immersive end-to-end communication systems normally include one architectural block that it is not present in video chatting applications, an audiovisual composition component. Video chatting requires efficient transmission mechanisms and optional content stream manipulations just after capture (e.g., iChat effects). More immersive systems, on the other hand, require complex content stream manipulations and camera control [12][13][14] where the orchestration of the scene can be done manually – as in commercial videoconferencing systems – or automatically, as is the goal of our system.

In commercial systems such as those developed by Lifesize Communications[1], all participants can manually control how a composite image from a videoconference is displayed locally based on a number of predefined layout options. Polycom[2] provides similar capabilities with the addition of a 'green screen' mode, which enables background subtraction from one source to superimpose it over external sources. Finally, Tandberg's[3] C90 codec provides more sophisticated orchestration alternatives, enabling one skilled in the art to rapidly build complex screen layouts.

Research solutions in the past have explored different video composition mechanisms. Orchestrated dancing in 3D tele-immersive environments [14] uses multiple cameras to capture a scene, and real-time 3D reconstruction to display it, where multi-stream coordination for synchronized reconstruction is a requirement [15]. Office-oriented solutions like Coliseum use background subtraction to isolate users from their surroundings [12]. They generate 3D models of the users, which can then be placed in virtual spaces for collaboration. Educational systems allow for compositing the videos of various students in several screens [13].

While the majority of the research on videoconferencing has focused on the office environment, some recent work has considered videoconferencing within the context of home. Social games such as Mafia [2] provide one example, in which users valued the ability to do things together with remote parties. A recent study on video-mediated free

[1] http://www.lifesize.com/
[2] http://www.polycom.com/
[3] http://www.tandberg.com

play between children found that different kinds of views lead to different types of play [16]. This study corroborates previous results on the importance of being able to manipulate and manage components within a set of video streams [17]. Our particular focus is in providing support for television techniques and rules - switch camera, pan, zoom - or enriching the shared experience. Recent experiments suggest that this approach provides higher user satisfaction for recorded meetings [18] and more vivid recorded lectures [19]. Other experiments demonstrate that good framing techniques improve social communication [20].

Effective social communication in mediated environments is dependent on effective gathering of information about the participants. Research on social communication has shown the importance of non-verbal cues, such as eye-gaze, head position and movement [8]. Other research has demonstrated the importance of where the participants are located in the room [21] and of verbal cues [10].

Based on the identified functional requirements, our system provides audiovisual composition that combines the remote view with the shared activity. Analogous to commercial systems and research solutions, we provide a control component that is capable of dynamically manipulating the audiovisual streams offering just-in-time composition and directional audio rendering. Unlike previous research, our system targets a domestic environment where one high-definition television is used as the only display with multiple people in each location.

B. Performance of Videoconferencing Systems

In order to establish a baseline for home performance, we conducted a set of experiments to gain insight on the performance of current commercial systems. We placed in view of one system's camera (the 'local' camera) a generated image of a millisecond counter that was displayed on a 'local' screen as a self-view. The system was connected to another in the same room (the 'remote' system) that displayed the transmitted image of the counter on the 'remote' screen. A digital SLR camera was then used to photograph the 'local' and 'remote' screens in the same frame, using a high-speed shutter (up to $1/1000^{th}$ of a second). By examining the photographs we could calculate the difference between the two counter images. By repeating the experiments on four occasions and using several consecutive shots, we obtained an estimate of the delay. Surprisingly enough, the results indicate that commercial videoconferencing systems do not meet the 100ms and 150ms guidelines. In particular, we measured the end-to-end delay between the Lifesize Room and Lifesize Passport products to be about 250ms, and approximately 340ms between the Tandberg Profile and Polycom HDX products.

It is important to note that this approach only provides an approximation, and cannot substitute accurate measurement of the individual system components – however for commercial equipment it is generally not possible to obtain such measurement without detailed co-operation from its vendors.

Another more programmatic method for calculating end-to-end delay is to use timestamps encoded as bar codes, as used by VDelay [22]. The capture-to-display delays obtained for Skype (640x480), Aim (240x180), and Gtalk (about 512x300) were 238ms, 147ms, and 99ms respectively. We aim to obtain a similar performance taking into account that our environment transmits high-definition video (1280x720).

For comparison purposes we have also measured the performance of Skype and iChat, but this time using an automated tool with which we also measured our own system, described in section VI.B. The systems under test

were reasonably well connected, through an 802.11n network and Gigabit Ethernet. Neither Skype nor iChat allows control over resolution and framerate, but we believe this was 640x480 at 30Hz. The average end-to-end delay measured was 186ms for iChat (with a standard deviation of 26ms) and 312ms for Skype (standard deviation 67ms). Our tool is probably more susceptible to image quality than vDelay [22], as we use more complex synthetic images and an extra camera. The minimum delay we measured for Skype was 260ms, and if we factor in the fact that vDelay uses screen capture at the receiving Skype machine whereas our tool incurs the extra display delay the numbers become very similar.

In general bandwidth requirements and delays play an essential role in the overall Quality of Experience. From a research perspective, there are studies available [23] that compare the delay and bandwidth requirements for different multipoint topologies in the context of domestic videoconferencing, but these only consider 640x480 resolution. Quality assessment of high-definition video can be found in [24], but it does not target videoconferencing. The numbers in two papers match up, however, and suggest that transmitting 1280x720 at 30fps will require 2-3Mbit/s of bandwidth to be perceived as good quality.

III. REQUIREMENTS

In order to better understand user requirements for domestic videoconferencing, we conducted interviews within sixteen families across four countries (UK, Sweden, Netherlands and Germany). The households interviewed had children aged between about 6 and 25. A full description of the methodology and a comprehensive report on the results is out of the scope of this paper and can be found elsewhere [5]. As expected, the majority of the participants had a television set in the living room, and the living room was considered as the central area for contact within the family. In Sweden the families did not use video communication at all, in the United Kingdom only one family used Skype video, as did two of the families in Germany and the Netherlands. Even though it is not statistically representative, there are a number of common issues that we have been able to obtain from the interviews.

Based on the interviews can conclude that the use of video communication was not found to be compelling because of the technology. Interviewees often complained about the quality of the video and the time gap between the video and the audio communication. Similar results have been obtained in recent studies conducted elsewhere (using Skype for communicating between families). “Families frequently encounter technical difficulties even after the call is established: unreliable Internet connections, microphones with feedback, video lag or visual artifacts, frozen screens, and crashed applications were all common” [3].

This imposes a set of *performance* requirements:

- Video quality that improves the visual communication, and audio quality that allows for speaker identification
- Low delay and synchronization comparable to commercial systems (up to around 350ms as shown in section II)

According to our interviews, playful activities and games were considered as a common part of the way families interacted in person; as a common activity that was enjoyed during social gatherings. Some of the interviewees agreed that they did not see the games as an end in themselves, but tended to play them rather as a convenient excuse for getting together physically. Other studies on home videoconferencing have reported that: “The systems used in the homes we observed were often used by multiple people… This suggests a need to develop a home appliance for

multiparty viewing and use" [4]. This imposes a set of *functional* requirements:

- Video communication should be bounded and reactive to social activities.
- The system should support a group of people in front of the camera.

Based on our studies, on recent studies on people using Skype for videoconferencing between families, and on computer-mediated social communication research [8][20], two sets of key requirements need to be fulfilled for bringing videoconferencing to the home. *Functional requirements* include dynamic control of the audiovisual streams and directional audio for supporting multiparty viewing and use, and for supporting shared activities; and *Performance requirements* such as reliability, bounded end-to-end delay and high-quality communication.

IV. Architecture and Infrastructure

The need for dynamic composition, as outlined in the introduction and Section III, has implications for the overall architecture of our system, as there needs to be a mechanism whereby the audiovisual presentation can be modified, while maintaining the performance requirements. In addition, the requirement to support multiple activities - and extensibility in general - has led us to a design of loosely-coupled modules. The overall architecture is sketched in Fig. 4. This is a simplified diagram, which leaves out some auxiliary modules for handling clock synchronization, communication, session setup and discovery.

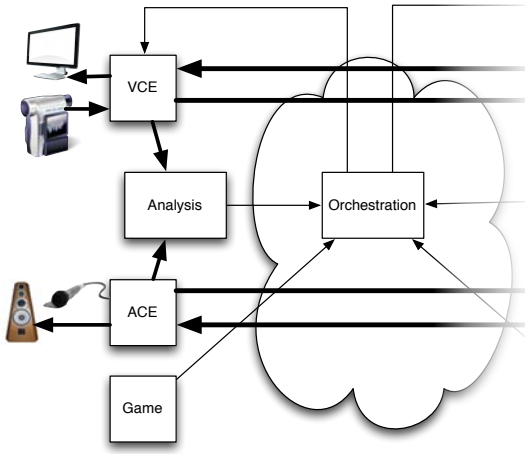

Figure 4: Global architecture of the system.

The relevant components of each node in the system are:

- a *game*, or, more generally, the activity around which the videoconference is centered.
- a *visual composition engine* (VCE), a separate hardware unit to manage the visual aspects of video selection and composition,
- an *audio communication engine* (ACE), a separate hardware unit which digitizes, transmits and renders multichannel audio,
- an *analysis engine*, which processes the relevant video and multichannel audio data to identify verbal (e.g., voice activity, speaker identification) and non-verbal (e.g., eye-gaze, head pose, movement) cues [10],
- an *orchestration engine*, which gathers the cues from the analysis engine and takes decisions on the audiovisual

Figure 5: Videoconferencing Rooms in Antwerp (Belgium) and Amsterdam (The Netherlands).

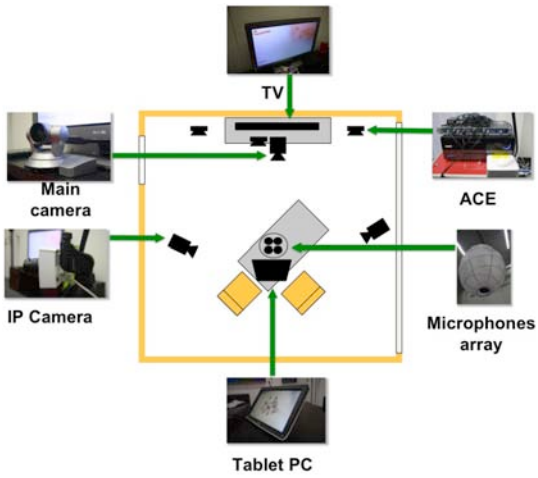

Figure 6: Sketch of the components used in each location.

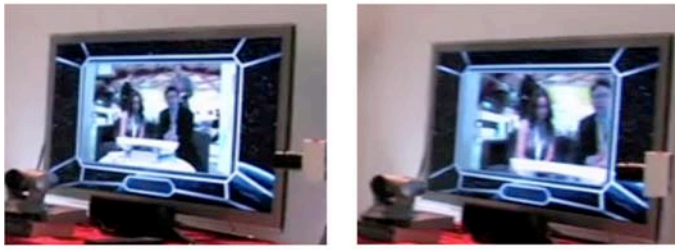

Figure 7: Composition-Based video-conferencing in action. The video as it captured by the main camera (top), a full crop composition with synthetic graphics (left), and a partial crop of the video framing one person (right).

composition policies to be applied (e.g., which video shots to present to the participants) [9].

Of these components, orchestration is functionally centralized while the others are distributed. In other words, there is a single orchestration engine while the other components have an instantiation per household.

Fig. 5 and Fig. 6 show the infrastructure we are using for enabling orchestrated video conferencing at home. Two instances of the system (one in Antwerp and one in Amsterdam) are illustrated in Fig. 5. Fig. 6 sketches the hardware configuration at each of the locations. To achieve composition-based video conferencing, each room has a TV screen acting as the window between the rooms, where communication and game play happens; the ACE that provides high-quality audio communication; a main camera and two IP cameras needed for supporting multiple people in one location and integrating activities like playing a game together; a microphone array hidden in a lamp for directional audio; and a tablet PC where the common activity runs.

The system presented in this article is a functional prototype. Fig. 7 shows screenshots of orchestrated communication using a similar system, captured at the ICT event at Brussels (September, 2010)[4].

While a detailed description of the analysis and orchestration components falls outside the scope of this paper, it is important to note their overall functionality and architectural implications. These components monitor (analysis) everything that happens at one location, and implement the visual composition policy (orchestrator) best suited to the current situation. The mechanism for implementing this policy is then provided by the VCE. It is important to note that the policy components get a feed of the video and audio streams as they are captured by the hardware, and can control both video framing at the source and general composition of video streams and other media at the destination.

[4] http://www.youtube.com/ta2project#p/u/0/2wnjfGpDlAs.

This allows a tradeoff of bandwidth (framing at the source) against reactivity (switching at the destination).

The architecture is not restricted to using the sketched analysis and orchestration components, and for the purpose of our discussion this functionality could just as well be provided by a human director controlling composition by pressing buttons on a mixing console. More important, completely different policies are possible, such as having the system react to changes in the operating environment, such as available bandwidth.

All components function independently to enable the architecture to meet the performance requirements. However, this raises the issue of synchronization, especially if components reside on different machines. This is addressed by adding timestamps to all media data, and synchronizing all local machines using a local NTP server. Moreover, the APIs of the ACE and VCE allow the audio output to be delayed to match the current video delay, to ensure audio and video lip-sync.

The components in the architecture have been implemented as independent applications primarily communicating using standard protocols such as XML-RPC for control and RTP/RTSP for media transfer. This enables the afore-mentioned replacement of policy components by different ones, and also allows us to do load balancing, by assigning components to different machines. Additionally, the separation of functionality helps the development process, as different teams are responsible for the different components.

V. Audio Communication Engine

A. Description

For natural communication, audio quality and bandwidth are particularly important. To give the remote participants the feeling that they are in the same room, the audio quality must be as good as if they really were in the same room. Meeting this requirement is still challenging to implement when low delay, low bit-rate, hands-free operation, and multichannel capability are considered. It can only be fulfilled by a super wide band audio codec supporting multichannel to reproduce the original sound image at the remote site.

The Audio Communication Engine (ACE) is built to meet these goals and covers the complete audio chain from the microphone over the IP transport to the remote speaker. Its functionality is separated into three categories: codec, IP transport and acoustic interface. Fig. 8 shows these categories and their modules, which will be described in the following sections.

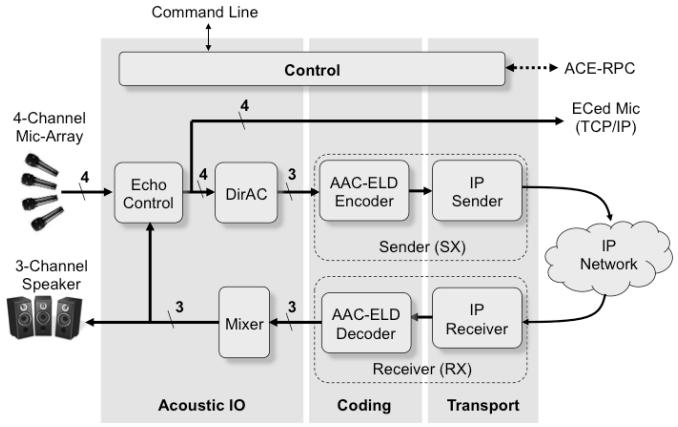

Figure 8: Components of the Audio Communication Engine (ACE).

1) Audio Coding

Advanced Audio Coding (AAC) is already in widespread use today. Though AAC offers excellent audio quality at stereo bit-rates of 128 kbit/s, it has high algorithmic delay. For this reason, MPEG has standardized variations of AAC with reduced delay. The most recent extension of AAC in this regard is AAC Enhanced Low Delay (AAC-ELD) [25]. The main features of ELD are super wide band audio bandwidth (up to 48 kHz sampling rate), 15-36 ms algorithmic delay, typical bit-rates of 24-64 kbit/s per channel, and 10 or 20 ms frame length. For a realistic replication of the remote acoustic image, the AAC-ELD codec typically encodes three channels (left, center, right) at 192 kbit/s total.

Starting from AAC-ELD as the core of the ACE, the scope is extended towards two directions. On the one hand, the Access Unit (AU: compressed audio frame) has to be transmitted through the IP-network. On the other hand, the PCM samples that form the input to the encoder and output of the decoder have to be recorded and rendered. These two aspects are described in the following subsections.

2) IP Transport

The transport over the IP-network includes the encapsulation of AUs into the Real-time Transport Protocol (RTP) [26] and transmission over the User Datagram Protocol (UDP) [27]. The receiver side employs adaptive playout to compensate delay variations on the network. Adaptive playout allows reducing the amount of time that a packet is buffered before decoding and therefore reduces end-to-end delay. The required time modification is realized by exploiting the properties of the AAC-ELD codec at low complexity and good quality. For further information on adaptive playout in the ACE we refer to [28]. The support for multipoint conferencing between more than two locations is based on the architecture of a centralized Multipoint Conferencing Unit (MCU) [29].

3) Acoustic Interface

The acoustic interface refers to the physical setup of microphones and loudspeakers as well as additional signal processing between those and the audio codec. As illustrated in Fig. 8, the ACE provides advanced features for spatial audio with more than 2 channels. A microphone array is used to capture the spatial sound field and allow basic beam forming. The array consists of four omnidirectional boundary condenser microphone capsules arranged at the corners of a square with 4.4 cm lateral length. The microphone array is mounted in a horizontal plane, either embedded in the surface of a table or into a lamp hanging above the table. The performance of the microphone array

has been analyzed in detail in [30][31].

Note that the number of output channels in Fig. 8 is not equal to the number of microphones. Instead, Directional Audio Coding (DirAC) is used to represent the multichannel audio signal in a more flexible way. DirAC provides a parametric representation of the sound field by using a mono downmix of the audio signal and additional side information, namely the direction-of-arrival (DOA) and diffuseness of the sound in each time- and frequency bin [32]. On this basis, the output signals for an arbitrary loudspeaker setup can be determined allowing for a perceptually accurate spatial rendering of the original or manipulated sound scene. This representation is not necessarily more compact, but allows for flexible signal manipulations, e.g., repositioning of users or activity with regard to orchestration rules.

Before DirAC processing, it is necessary to apply the Echo Control (EC) as required for hands-free operation. The EC takes the characteristics of human perception into account, rather than blindly following estimation theory principles. It supports multiple channels at low complexity and is very robust against changes in the echo path and nonlinearities. The EC provided in the ACE is further described in [33][34]. In Fig. 8, all four microphone signals are echo controlled ("ECed") based on the input of the three speaker channels. The ECed input signal is split up and forwarded to the analysis engine for voice activity detection and speaker identification.

The Mixer shown in Fig. 8 finally mixes all audio for the local loudspeakers and sends a copy of it to the echo control, to cancel the feedback signal in the microphone. The ECed microphone is sent to the analysis engine for object detection.

B. Evaluation and Results

The following sections present objective and subjective test results of some ACE components. The first part covers audio codec evaluation. The second part presents listening test results assessing the audio quality of ACE's playout adaptation. Finally, the user experience is evaluated.

1) AAC-ELD Evaluation

The key to reducing the acoustic distance between two domestic environments is based on the quality of the audio codec. Apart from that, audiocoding artifacts may be amplified by further signal processing, e.g., in case of dynamic composition. Thus, the audio codec must be chosen not only with regard to low delay, but also towards low bit-rate and the support of excellent audio quality. The low delay capability of AAC-ELD was demonstrated during the formal MPEG Verification Test [35]. In this test, the performance of AAC-ELD was compared to that of the latest available super wide band ITU codec, G.722.1-C.

The report revealed that AAC-ELD has a nearly 25 percent lower algorithmic delay when compared to the ITU codec at 32 kbit/s (32 vs. 40 ms). When increasing the bit rate up to 64 kbit/s, AAC-ELD can take advantage of this by further reducing the delay down to 15 ms while the delay of G.722.1-C stays fixed at 40 ms. [35] assesses the subjective audio quality of the investigated codecs through the MUSHRA (Multiple Stimuli Hidden Reference and Anchor) test methodology [36], where the codecs are compared to a hidden reference, other codecs under test and at least one anchor (3.5 kHz band limited signal). The items are rated on a score between 0 and 100, ranging from *Bad*

to *Excellent*.

Five companies contributed to this exercise as listening test laboratories with a grand total of 152 subjects. The results indicate that AAC-ELD at 24 kbit/s can achieve comparable audio quality as G.722.1-C at 32 kbit/s. Furthermore, there is a significant performance difference at 32 kbit/s of 23 MUSHRA scores.

Table I: Minimum bit-rates required for Excellent quality of mono signals according to [37].

Codec \ Bit Rate [kbit/s]	Audio Item 1	2	3	4	5	6
AAC-ELD	32	24	48	32	32	32
AAC-LD	48	48	32	48	48	48
CELT	64	48	64	48	48	48
G.718	48	-	-	-	-	32
G.719	32	32	48	48	32	48
G722.1 -C	48	32	-	24	32	-
G722.2	-	-	-	-	-	-
G.722	-	-	-	-	-	-
SILK	40	-	40	-	-	40
Speex	-	-	-	-	-	-

Further evidence for the superior quality of AAC-ELD has been given by the Deutsche Telekom AG with results of an independent listening test submitted to the 59th SA4 Meeting of 3GPP [37]. The set comprises MPEG codecs (AAC-LD, AAC-ELD), ITU codecs (G.718, G.719, G.722.1-C, G.722.2), the audio codec used by Skype (Silk) as well as open source codecs (CELT, Speex). The report concludes that excellent quality (MUSHRA score > 80) can only be achieved with a subset of the codecs under test (AAC-ELD, AAC-LD, CELT, G722.1-C and G.719). Table I shows the minimum bit-rate at which excellent audio quality can be achieved for mono signals. Only AAC-LD/ELD, G.719 and CELT achieve excellent quality consistently over all six test items. While AAC-ELD can provide this quality at about 32 kbit/s, G.719 needs between 32kbit/s and 48kbit/s at 40 ms algorithmic delay, and CELT requires 48-64 kbit/s.

2) Adaptive Playout Evaluation

To maintain low delay during the whole conversation, adaptive playout (AP) estimates the network jitter and schedules the playout of each audio frame by performing time stretching and shrinking algorithms. Packet delay is often described by a random floor with single delay bursts, also called delay spikes [38]. To provide low delay, AP allows a certain amount of late loss, to cut off such delay spikes. Thereby, time-scaling artifacts and mean delay can be significantly reduced. The accepted loss, however, leads to burst concealments during the spikes. Fig. 9 shows the MUSHRA score of such concealment bursts as well as uniform concealments over different late loss rates. The bursts were modeled using a fixed burst event rate of 1% and varying burst lengths of {10, 50, 100, 200} ms. The uniform concealments are single losses equally distributed over time.

As can be seen from Fig. 9, burst concealments are more critical for human perception than individual losses. To keep excellent quality, the accepted late loss rate of the ACE is limited to less than 1%.

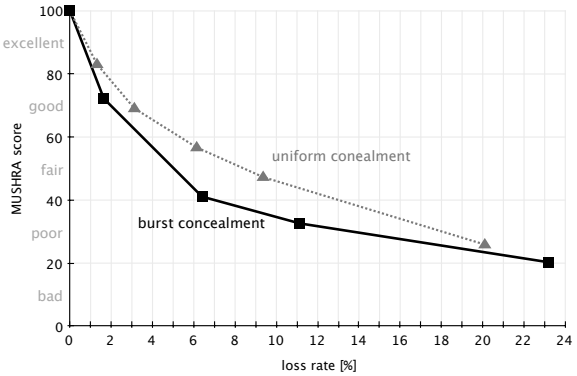

Figure 9: MUSHRA scores for burst and uniform distributed concealments.

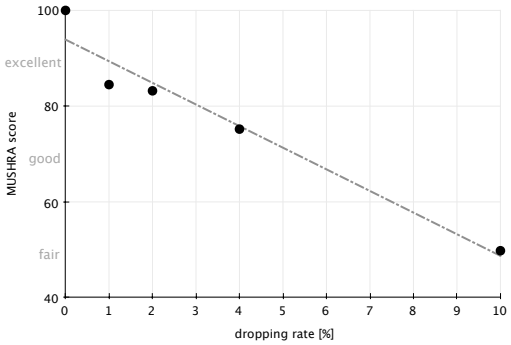

Figure 10: MUSHRA score of AAC frame dropping.

To reduce the buffer size, the ACE exploits the overlap-add structure of AAC and therefore simply skips single audio frames at the decoder. I.e. instead of the full sequence *1,2,3,4*, the ACE drops frame *3*, decodes *1,2,4*, and thus shrinks the audio. The results of listening tests are shown in Fig. 10. To retain excellent audio quality while dropping, the ACE drop rate is limited to 2-3%. By the use of AP, the ACE's delay can be reduced to 50 ms, which leaves enough space for additional delay to the limit of voice communication [39] [40].

3) User Experience Evaluation

While the above tests evaluate specific aspects of audio quality in isolated test conditions, it is still an open question if the demonstrated benefits also affect the overall user experience. Therefore, user experience (UX) tests have been conducted with families and friends when being involved in an interactive group activity. For this purpose, the ACE was installed in two rooms for mediated audio communication. In addition, an HDvideoconference and coffee-table with a touch-screen and simple board-games such as "Memory" or "Ludo" were installed. The games are commonly known and people are used to playing them in the living room in general, with lots of interaction and chatting. Thus, users react more sensitively to degradation of the communication system.

The group was exposed to "high" and "low" audio quality for a duration of about 5 min. each and then asked to express their agreement to a statement such as "The communication was natural and without problems" on a 5-point rating scale ranging from 1 (strongly agree) to 5 (strongly disagree). While the "high" audio quality is characterized by high bandwidth (44.1 kHz sampling rate), spatial audio (left, center, right), and low delay (80 ms), the "low" audio quality is characterized by narrowband (8 kHz sampling rate), mono (routed to all three speakers), and high delay (600 ms). From the 25 people who participated in the test, only 12 were above the age of 12 and allowed to vote. Since each test condition was presented twice, the results are based on 24 votes.

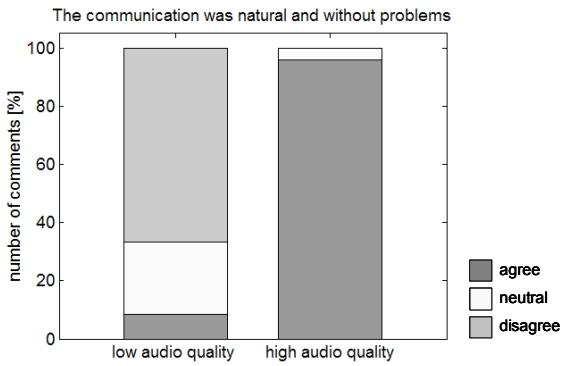

Figure 11: User experience during interactive group activity for low and high audio quality.

Fig. 11 illustrates the results when mapping the scores 1&2 to a general "disagree" and scores 4&5 to a general "agree". A score of 3 is interpreted as "neutral". As can be seen, the agreement raises from below 10% to above 90% when improving the audio quality from low to high. Though further UX tests are needed to validate these results and to investigate which parameters of the audio chain influence the user experience most, the initial results provide a good indication for the importance of audio quality in group-to-group communication.

VI. Visual Composition Engine

Where the ACE is responsible for the audio communication chain, the VCE is not only responsible for the video communication chain but also for display of pre-recorded media: text, graphics, movies and even audio clips. The VCE should allow compositing of all these items in an aesthetic way. From the previous sections it follows that the VCE should be designed in such a way that it can provide the required dynamic composition goals, while still meeting the performance requirements.

A. Description

The VCE architecture is sketched in Fig. 12. The lower, shaded, part of the figure is a rather traditional video pipeline for videoconferencing, the upper part shows the novel composition engine and the control module.

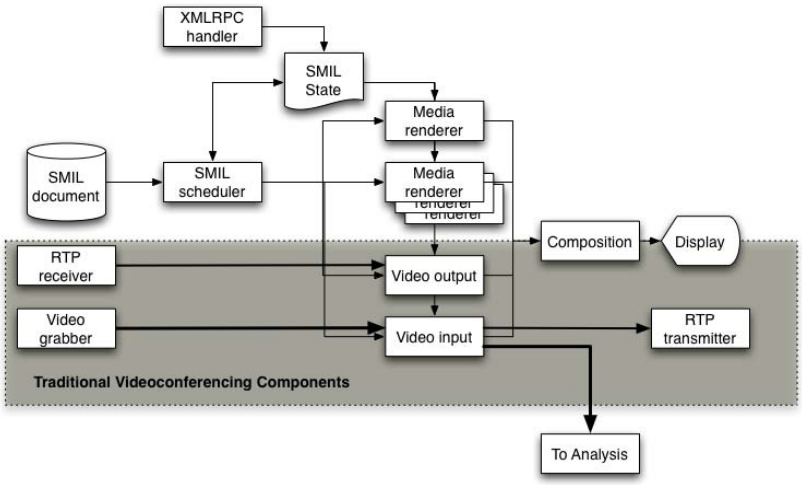

Figure 12: VCE Architecture.

The control module of the VCE is a SMIL scheduler. The SMIL language [41] not only provides visual composition but also has a rich set of temporal composition operators [42]. SMIL has traditionally been used to allow multimedia presentations to be rendered at viewing time, as opposed to rendering at authoring time, as is the case for formats like QuickTime or MPEG-4. This feature enables authors to let presentations adapt to user interaction, bandwidth availability, and the environment [43]. For playback of multimedia presentations, it has been shown that a declarative composition language like SMIL has advantages over static container formats. With the VCE we explore these advantages in the realm of videoconferencing and game playing, by allowing real-time modifications of running presentations.

The use of SMIL enables a clean separation of composition mechanism and policy, thereby isolating the orchestration component from low-level timing and rendering issues: orchestration can schedule compositions to happen in the future without having to worry about real-time issues. In addition, SMIL Animation and Transitions [41] provide mechanisms for implementing "eye candy" aesthetics in a relatively easy way.

In addition, the VCE makes use of SMIL State, a feature introduced in SMIL 3.0. Traditionally, all adaptations a SMIL presentation could do at presentation time had to be determined at authoring time, but SMIL State introduces a mechanism whereby external agents can change and add media items and rendering parameters at rendering time. This is somewhat analogous to how DOM access allows changes to an HTML document at display time (albeit through a completely different mechanism).

The SMIL document models the videoconferencing communication (real-time transmission of the video) and the current activity (the game), and has been created by the designer of the activity. This ensures that the "look and feel" of the videoconferencing matches that of the game, and that the interaction points between VCE and other components such as orchestration are well-defined. Upon start of the activity, the VCE scheduler creates a time graph based on the SMIL document, and then interprets this time graph. Whenever a media item has to be rendered, the scheduler determines which renderer is appropriate for the media item and creates it. From this point on, the renderer is semi-autonomous, which makes it relatively easy to ensure the parallelism needed to ensure good video performance. For the SMIL scheduler, composition engine, media renderers and support components such as XML

parser we have used the open source Ambulant SMIL player[5]. Ambulant is multithreaded, with locking at a level that allows us to make use of multiple cores in the renderers. It is also easily extended with the special-purpose renderers we need (video grabbing, live video display, etc).

Using SMIL for our visual input and output ensures we have a good basis for specifying (and implementing) our synchronization requirements, and with SMIL State we have a mechanism that enables interaction between the SMIL scheduler, renderers and composition on the one hand and external components on the other hand. SMIL State adds a structured data store to SMIL documents [44]. Values in this data store can be modified at rendering time, and these values can be used to control parameters in the presentation – to enforce the policies on video composition dictated by the orchestration engine. Moreover, changing a value in the data store can raise an event that can be used as a trigger to start or stop media rendering. Value changes are propagated immediately similar to the way a spreadsheet works.

In the VCE, we use a small XML-RPC server that listens for incoming commands, and modifies variables in the SMIL State data store. The new value is then propagated to the renderer component. The API exported by the XML-RPC server is tailored to the needs of the Orchestration and the game components, implementing the mechanisms to enable its policy decisions.

In particular the current composition-based functionality provided by the VCE include:

- Switch camera: to seamlessly switch between one of the three cameras in the room for better conveying social communication between locations
- Crop Video: to crop the transmitted video for better framing a person, an action, or a communication pattern (e.g., a conversation in the room)
- Face Tracking: to track the face of one specific person for specific purposes (e.g., for highlighting that is his/her turn in the game)
- Video transmission: to capture, encode, and transmit the video stream meeting minimal end-to-end delays
- Synthetic Graphics insertion: to synchronize auxiliary graphics (e.g., images, text, other videos) for integrating the game play - for example by providing status related to the game in the television screen

As an example of this, let us examine the video input component. The video grabber is started from the timegraph, and grabs 1080i frames, optionally crops them, scales them to the required size and transmits them. The crop coordinates and transmission size are tied to state variables, so whenever the orchestration component changes crop coordinates, the video input component will be notified of the new values and the next frame will be cropped to this new rectangle.

Video input and output merit special attention because they are the most critical when it comes to performance. They have been designed for minimal dependency on the rest of the system. On the input side, we use the Microsoft DirectShow infrastructure to grab UYVY frames. Each frame is immediately put in a queue for transmission to the Analysis engine, then cropped and scaled to fit the HD TV resolution and deposited in two more queues, one for local playback, one for RTP transmission. Events are used to wake up the transmitters, which run at high priority,

[5] http://www.ambulantplayer.org

and local playback is handled at a lower priority by the composition redraw code. The analysis transmitter will scale the raw frame to quarter size and send it out. The RTP transmitter feeds the frame to the x264 H.264 encoder and sends the result out as an RTP packet. The encoder parameters are set for minimal delay on sending and receiving side.

Video output is received by an RTP receiver thread running as a DirectShow input filter; H.264-decoded and displayed on the screen directly, bypassing the composition redraw code. The position and size are controlled by the composition code, however. This method of display has the disadvantage that displaying other media items on top of the video is no longer possible, but it significantly lowers video delay.

Video feeds between VCEs are initially set up using RTSP: the video input component has a small RTSP server and the output component is a client. This client is also used to connect two more video output renderer instances to the self-contained IP cameras which provide alternative side views.

B. Measurements

In order to validate the performance requirements we have performed delay measurements over the Internet. As video delay measurements are tedious to do manually we have created a standalone tool, running on a separate machine, that generates a barcode representing the current system time on-screen, and then measures the delay until this barcode is detected by the camera on this same machine. We point the local VCE camera at the measurement system screen, the remote VCE camera at its own screen and the measurement system camera at the local VCE screen. This allows us to obtain round trip delay times. The idea behind the measurement system is loosely based on [22], but for the current set of measurements we have opted for doing generation and detection on the same machine. This has the advantage that we have no problem with clock differences and clock drift between generator and detector. We have measured both the delay of a continuous video stream and the delay of implementing a change in composition, i.e. switching the output view from one video stream to another.

C. Evaluation and Results

The first measurement we did was to measure the delay of the test system itself, by pointing the camera at its own screen and measuring the delay. The average delay over 420 measurements was 68ms, with a standard deviation of 37ms.

Next, we did the delay test of our system over the internet. We took 450 round-trip measurements between a room in Antwerp, Belgium and a room in Ipswich, UK. We transmitted H.264-encoded 30fps 1280x720 video (scaled from a 1920x1080 video source) via a 40 MBit/s VPN connection over the Internet, of which approximately 2.5Mbit/s was used. For reasons of simplicity we have used VBR encoding with the x264 quantization parameter *qp* set to 30. Since VBR is the norm for current videoconferencing, determining the effects of using CBR is left as future work. The average raw round-trip delay was 770ms with a standard deviation of 56ms. As the measurement system and the system under test are independent we can subtract the averages, and add the variances. This gives a round-trip delay of approximately 700ms (with a standard deviation of 67ms). The raw measurements are displayed in Fig. 13, the distribution in Fig. 14 (with the corresponding normal distribution curve overlaid).

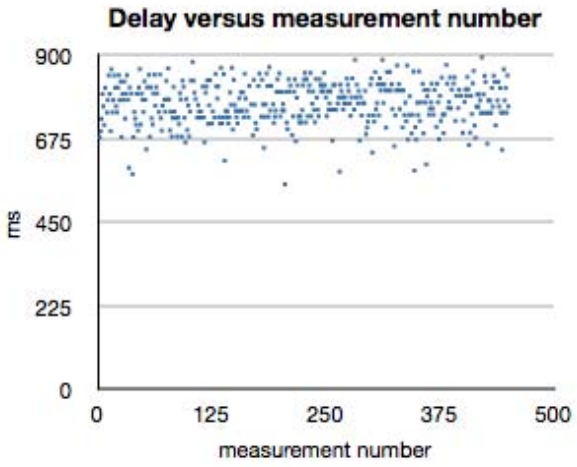

Figure 13: Video round trip delay measurements.

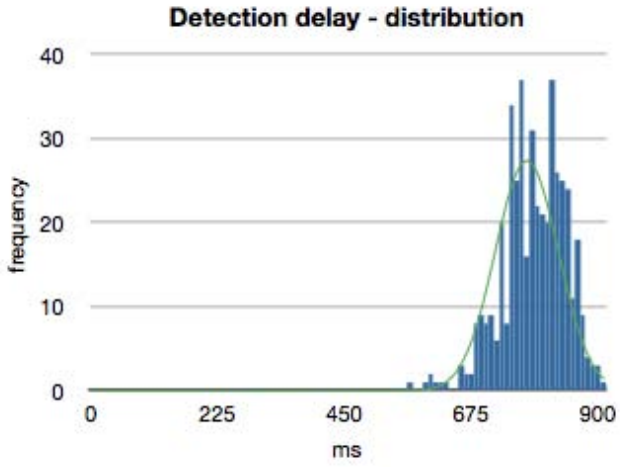

Figure 14: Video round trip delays, frequency distribution and corresponding normal distribution.

Apart from video communication, the VCE provides the functional requirements identified in Section III: camera switch, video crop, face tracking, and addition of synthetic graphics and extra media. In order to be effective, these video composition mechanisms should be efficient. In order to determine the impact of composition on delay figures, we have also performed measurements that compare cropping a 1280x720 section from the original 1920x1080 video input to scaling the full video to the required size. The results were nearly identical (401ms versus 408ms average, for about 500 measurements). This means that the impact of cropping and scaling is minimal. Due to practical constraints these measurements were made with end-to-end communications remaining on the local network, i.e. with the second VCE system collocated with the first one. Hence, these numbers can be compared to each other, but not to those of the previous paragraph.

Finally, we measured the delay in implementing a composition change of switching between cameras. The system under test was receiving two live video feeds, with one being displayed at a time. We measured the time it took from issuing the command to switch to the other video feed until the new video feed was detected by our measurement system. The switch command was issued by the measurement system, using the same network-based XML-RPC API that the orchestration component would use in a production system.

Unfortunately the two source cameras had completely different delay characteristics, so the average for this measurement is difficult to interpret. Instead, we measured the minimum delay in implementing the camera switch, which was about 90ms longer than the minimum video delay during normal display. Most this is attributable to the XML-RPC delay (for which we measured a round-trip time with a minimum of 100ms).

VII. Discussion and Conclusions

In this paper we postulate that a videoconferencing system for domestic use has qualitatively different requirements than a traditional conferencing system for professional use, due to the different social setting, interaction patterns and hardware constraints. We validate that a system can be constructed that meets the functional requirements as well as the performance requirements.

The social setting requires minimal intrusiveness of hardware used, and through the use of a small, easily hidden,

microphone array and a single frontal HD camera we feel we have met this requirement. Additional cameras may be used for alternative side angles, and they may be omitted if wanted. The domestic setting also requires that the communication tool melds into the background, with the game and the social interaction being the prime focus of the users. Dynamic composition provides the mechanism for this.

Capturing and transmitting the dynamic interaction of people in a home setting is catered for through the use of spatial audio and dynamic adaptation of the video shots. This allows participants to discriminate between the different people at the remote location, and moreover the system can adapt the video feed based on the interaction patterns detected. High quality video is also of importance here.

The hardware constraints limit the system in both compute power and bandwidth used. Video is the bottleneck here, and the testbed system used a quad-core 2.83Ghz Intel Xeon system to run the VCE. While this type of system is more powerful than what is available in the average home today we feel that advances in the field should have similar configurations generally available in 1-2 years, in a form that is suitable for home use (single small enclosure, no loud fans). Similar for bandwidth: 2-3Mbit/s uplink bandwidth is not common today, but if the current trend continues it should be reasonably common in that same timeframe. The test bed system used multiple machines for the other components (ACE, orchestration, analysis), but this has been done due to practical reasons: performance-wise these services are not as compute-heavy as video encoding and composition.

Our system meets the acceptable video performance of 350 ms end-to-end delay (similar than that of commercial videoconferencing systems), but it falls short of the optimal 100-150ms. We plan to tackle this in the future, after determining where the delays are actually incurred. Our measurement tool can be adapted for doing such in-the-box measurements. The audio performance requirements are easily met by our system.

The system as it runs today is complete and functional, including the analysis and orchestration components that are described elsewhere [9][10][11]. This paper focuses on the development of the audiovisual composition engines, and reports on a number of experiments. The goal of the experiments is to validate that audiovisual composition works and that the delays are within acceptable limits. Reliability and performance of the analysis component, algorithmic solutions regarding the orchestration component, and human-centered studies around the social communication benefits of audiovisual composition are out of scope; and they will be reported in subsequent articles.

If we compare our system to existing video chat solutions for the home such as Skype, there is a quantitative difference in quality (HD video versus QCIF, spatial audio versus mono). More important, however, is the architectural difference of providing dynamic audiovisual composition. This allows integration of audio and video communication with other activities such as a game, something that is not possible today with commercial systems like Gtalk, Skype, and iChat.

The main contribution of this paper is that the domestic setting requires a different architecture (as shown in section IV) for doing videoconferencing than the meeting room, and directional audio and dynamic composition are two of the features required. This is needed for supporting activities that happen while videoconferencing occurs, and for allowing group-to-group communication. The separation of dynamic composition in mechanism and policy, with a clearly defined interface between them, not only enables adaptation to social communication, as demonstrated in

this paper, but the mechanism can also be reused to adapt to environmental changes, such as available bandwidth.

We want to extend our system to conferences of more than two locations, and we expect these features to become even more important then. Our architecture allows us to experiment with variants of dynamic composition, to determine the difference between switching video at the source and at the destination. There is a tradeoff here between bandwidth usage and reactivity.

REFERENCES

[1] Poltrock, S. E. and Grudin, J. 2005. Videoconferencing: Recent Experiments and Reassessment. In *Proceedings of the Annual Hawaii international Conference on System Sciences*, Volume 04, 104.1.

[2] Batcheller, A. L., Hilligoss, B., Nam, K., Rader, E., Rey-Babarro, M., and Zhou, X. 2007. Testing the technology: playing games with video conferencing. In *Proceedings of ACM CHI*, 849-852.

[3] Ames, M. G., Go, J., Kaye, J., and Spasojevic, M. 2010. Making love in the network closet: the benefits and work of family videochat. In *Proceedings of ACM CSCW*, 145-154.

[4] Kirk, D. S., Sellen, A., and Cao, X. 2010. Home video communication: mediating 'closeness'. In *Proceedings of ACM CSCW*, 135-144.

[5] Williams, D., Ursu, M. F., Cesar, P., Bergström, K., Kegel, I., and Meenowa, J. 2009. An emergent role for TV in social communication. In *Proceedings of EuroITV*, 19-28.

[6] Baldi, M. and Ofek, Y. 2000. End-to-end delay analysis of videoconferencing over packet-switched networks. *IEEE/ACM Transactions on Networking*, 8(4): 479-492.

[7] Roberts, D., Duckworth, T., Moore, C., Wolff, R., and O'Hare, J. 2009. Comparing the End to End Latency of an Immersive Collaborative Environment and a Video Conference. In *Proceedings of the EEE/ACM international Symposium on Distributed Simulation and Real Time Applications*, 89-94.

[8] Gatica-Perez, D. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Computing*, 27(12): 1775-1787.

[9] TA2 project. 2010. Interaction Modelling Language and Reasoners for Interaction Moderation and Content Capture. *Deliverables D7.6 & D7.7*.

[10] Duffner, S., Motlicek, P., and Korchagin, D. 2011. The TA2 Database - A Multi-Modal Database from Home Entertainment. In *Proceedings of the International Conference on Signal Acquisition and Processing*.

[11] Kaiser, R., Torres, P., and Martin Hoffernig. 2010. The interaction ontology: low-level cue processing in real-time group conversations. In *Proceedings of the ACM international workshop on Events in multimedia*, 29-34.

[12] Baker, H. H., Bhatti, N., Tanguay, D., Sobel, I., Gelb, D., Goss, M. E., Culbertson, W. B., and Malzbender, T. 2005. Understanding performance in coliseum, an immersive videoconferencing system. *ACM TOMCAP*, 1(2): 190-210.

[13] Chen, M. 2001. Design of a virtual auditorium. In *Proceedings of the ACM International Conference on Multimedia*, 19-28.

[14] Yang, Z., Wu, W., Nahrstedt, K., Kurillo, G., and Bajcsy, R. 2010. Enabling multi-party 3D tele-immersive environments with ViewCast. *ACM TOMCCAP*, 6(2): article number 7.

[15] Ott, D.E., and Mayer-Patel, K. 2004. Coordinated multi-streaming for 3D tele-immersion. In *Proceedings of the ACM International Conference on Multimedia*, 596–603.

[16] Yarosh, S., Inkpen, K.M., and Brush, A.J.B. 2010. Video playdate: toward free play across distance. In *Proceeding of ACM CHI*, 1251-1260.

[17] Gaver, W., Sellen, A., Heath, C., and Luff, P. 1993. One is not enough: multiple views in a media space. In *Proceedings of ACM CHI*, 335-341.

[18] Ranjan, A., Birnholtz, J., and Balakrishnan, R. 2008. Improving meeting capture by applying television production principles with audio and motion detection. In *Proceeding of ACM CHI*, 227-236.

[19] Lampi, F., Kopf, S., and Effelsberg, W. 2008. Automatic lecture recording. In *Proceeding of the ACM international Conference on Multimedia*, 1103-1104.

[20] Nguyen, D.T. and Canny, J. 2009. More than face-to-face: empathy effects of video framing. In *Proceedings of ACM CHI*, 423-432.

[21] Yamashita, N., Hirata, K., Aoyagi, S., Kuzuoka, H., and Harada, Y. 2008. Impact of seating positions on group video communication. In *Proceedings of ACM CSCW*, 177-186.

[22] Boyaci, O., Forte, A., Baset, S. A., and Schulzrinne, H. 2009. vDelay: A Tool to Measure Capture-to-Display Latency and Frame Rate. In *Proceedings of IEE ISM*, 194-200.

[23] Han, I., Park, H., Choi, Y., and Park, K. 2008. Four-way video conference and its session control based on distributed mini-MCU in home server. In *Proceedings of the IEEE International Conference on Consumer Electronics*, 233-234.

[24] Pinson, H.M., Wolf, S. and Cermak, G. 2010. HDTV Subjective Quality of H.264 vs. MPEG-2, With and Without Packet Loss. In *IEEE Transactions on Broadcasting*, 56(1): 86-91.

[25] Schnell, M., et al. 2007. Enhanced MPEG-4 Low Delay AAC – Low Bitrate High Quality Communication. In *122th AES Convention*.

[26] Schulzrinne, H., et al. 2003. RFC 3550 – RTP: A Transport Protocol for Real-Time Application. http://www.ietf.org/rfc/rfc3550.txt

[27] IETF STD 0006. 1980. User Datagram Protocol. Postel J., August 1980.

[28] Issing, J., at al. 2008. Adaptive Playout for VoIP based on the Enhanced Low Delay AAC Audio Codec. In *124th AES Convention*.

[29] ITU-T Recommendation H.231. 1997. Multipoint control units for audiovisual systems using digital channels up to 1920 kbit/s.

[30] Schultz-Amling, R., et al. 2008. Planar Microphone Array Processing for the Analysis and Reproduction of Spatial Audio using Directional Audio Coding. In *124th AES Convention*.

[31] Kallinger, M. 2008. Analysis and Adjustment of Planar Microphone Arrays for Application in Directional Audio Coding. In *124th AES Convention*.
[32] Pulkki, V. 2007. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6): 503–516.
[33] Kuech, F., et. al. 2008. Acoustic Echo Supression Based on Separation of Stationary and Non-Stationary Echo Components. In *Proc. Intl. Works. on Acoust. Echo and Noise Control* (IWAENC).
[34] Favrot, A., et. al. 2008. Acoustic echo control based on temporal fluctuation of short-time spectra. in *Proc. Intl. Works. on Acoust. Echo and Noise Control* (IWAENC).
[35] ISO/IEC JTC1/SC29/WG11, N10032. 2008. Report on the Verification Test of MPEG-4 Enhanced Low Delay AAC.
[36] ITU-R Recommendation BS.1534, Method for the subjective assessment of intermediate quality levels of coding systems.
[37] 3GPP Document S4-100479. 2010. Listening tests concerning reference codecs for EVS. From Deutsche Telekom AG, TSG-SA4#59 meeting, 21-24 June 2010, Prague, Czech Republic, available at ftp://ftp.3gpp.org/tsg_sa/WG4_CODEC/TSGS4_59/Docs/S4-100479.zip
[38] Markopoulou, A.P., Tobagi, F.A., and Karam. M.J. 2003. As- sessing the Quality of Voice Communications over Internet Backbones. *IEEE/ACM Transactions On Networking*, 11(5): 747–760.
[39] G.107. 2000. The E-model, a computational model for use in transmission planning. *ITU-T Recommendation*.
[40] G.114. 2003. One-way transmission time. *ITU-T Recommendation*.
[41] Bulterman, D. et al. 2008. Synchronized Multimedia Integration Language (SMIL 3.0). *W3C*. URL=http://www.w3.org/TR/SMIL/
[42] Rogge, B., Bekaert, J., and Van de Walle, R. 2004. Timing Issues in Multimedia Formats: Review of the Principles and Comparison of Existing Formats. *IEEE Transactions on Multimedia*, 6(6): 910-924.
[43] Kernchen, R., et al. 2010. Intelligent Multimedia Presentation in Ubiquitous Multidevice Scenarios. *IEEE MultiMedia*, 17(2): 52-63.
[44] Jansen, J. and Bulterman, D. C. 2008. Enabling adaptive time-based web applications with SMIL state. In *Proceeding of ACM DocEng,* 18-27.

Jack Jansen is a researcher at Centrum Wiskunde en Informatica (CWI), with over 25 years of experience in multimedia and distributed systems. Empowering people to put available technology to a use they themselves envision is his driving principle. This results in activities ranging from languages, such as Python, via web standardization work (SMIL, Rich Web Application Backplane) to implementing systems for accessible and reusable multimedia (Ambulant). Recently, he has finally started to pursue a PhD. Webpage: http://homepages.cwi.nl/~jack/

Pablo Cesar is a tenure track researcher at CWI (The National Research Institute for Mathematics and Computer Science in the Netherlands). He received his PhD from the Helsinki University of Technology in 2006. He has (co)authored over 40 articles (conference papers and journal articles) about multimedia systems and infrastructures, social media sharing, interactive media, multimedia content modeling, and user interaction. He is involved in standardization activities (e.g., SMIL from W3C) and has been active in a number of European projects such as Passepartout, SPICE, iNEM4U, and Ta2. He is coeditor of the book “Social Interactive Television: Immersive Shared Experiences and Perspectives” and has given tutorials about multimedia systems in prestigious conferences such as ACM Multimedia and the WWW conference. Webpage: http://homepages.cwi.nl/~garcia/

Dick Bulterman is head of distributed and interactive systems research at CWI, the Dutch national center for mathematics and computer science in Amsterdam. He is also a professor of computer science at the VU University in Amsterdam. Dr. Bulterman received his Ph.D. in computer science from Brown University in Providence RI (USA) in 1981. He has been co-chair of the W3C working group on synchronized multimedia since 2007; this group released the SMIL 3.0 Recommendation in late 2008. Bulterman has been active in the Document Engineering community since 2005. He is past program chair and past general chair of the ACM DocEng Symposium. He is also past chair of ACM Multimedia of and IEEE ISM. Dick Bulterman lives in Amsterdam with his wife and two children. Webpage: http://homepages.cwi.nl/~dcab/

Tim Stevens joined BT after graduating with a BSc in physics in 1986. After writing software for embedded systems and work on the objective and subjective measurement of analogue telephony systems, he moved into research into digital media and metadata, developing software tools to generate and distribute interactive content for broadband and broadcast

distribution, for which he gained several patents. Recently, he has developed software for real-time video capture, switching & recording as part of the EU FP7 TA2 project. Tim is a Chartered Engineer and member of the UK's Institution of Engineering and Technology

Ian Kegel heads the Future Content Group, whose role it is to supply BT with the product ideas, technology and foresight which it needs to help its customers take full advantage of the world of digital content. Having studied Electrical and Information Sciences at the University of Cambridge, Ian has worked in both the defence and telecommunications industries on projects ranging from radar signal processing to multimedia delivery, and has spent the last 10 years undertaking content-related research within BT. Ian co-ordinates a programme of work focused on audiovisual entertainment and multimedia communications, and seeks to develop compelling new applications and services for the digital home. He has collaborated with partners from industry and academia from across Europe in a variety of projects and initiatives including the DTI/TSB and EU Framework Programmes.

Jochen Issing is a PhD student at the Department of Computer Science 7 of Friedrich-Alexander-University, Erlangen-Nuremberg. He received his degree in electronic engineering from University of Applied Science Amberg-Weiden in 2002. Between 2002 and 2008 he worked as a research associate at the Multimedia Transport Group of Fraunhofer IIS, Erlangen. His research interests include multimedia streaming, human perception, robustness and simulation.