Ambient Human-to-Human Communication

Aki Härmä

1 Introduction

In the current technological landscape colored by environmental and security concerns the logic of replacing traveling by technical means of communications is undisputable. For example, consider a comparison between a normal family car and a video conference system with two laptop computers connected over the Internet. The power consumption of the car is approximately 25kW while the two computers and their share of the power consumption in the intermediate routers in total is in the range of 50W. Therefore, to meet a person using a car at an one hour driving distance is equivalent to 1000 hours of video conference. The difference in the costs is also increasing. An estimate on the same cost difference between travel and video conference twenty years ago gave only three days of continuous video conference for the same situation [29]. The cost of video conference depends on the duration of the session while traveling depends only on the distance. However, in a strict economical and environmental sense even a five minute trip by a car in 2008 becomes more economical than a video conference only when the meeting lasts more than three and half days.

It is often suggested that the transportation and telecommunications are strongly coupled, or complementary. The telephone is an important means of making appointments and therefore enabling and justifying more travel, but at the same time travel increases the need for communications. For example, Short et al [97] give an example where the opening of a bridge in the UK led to a significant increase in the telephone traffic between the two previously separated areas. In fact, a recent analysis by Choo and Mokhtarian [23] on travel and telephony data in the USA since 1950's indicates that travel triggers more need for telecommunication than telephony leads to travel.

Härmä
Philips Research Europe, Eindhoven, The Netherlands. e-mail: aki.harma@philips.com

H. Nakashima et al. (eds.), *Handbook of Ambient Intelligence and Smart Environments*,
DOI 10.1007/978-0-387-93808-0_30,

The analysis of travel and telephony statistics in the USA since 1950's till early 2000's [23] draws probably a quite accurate picture of the complementarity because both travel (mainly cars) and telephony in early 2000's were actually very similar to 1950's. However, in the last few years travel has became even less attractive due to increasing energy prices, environmental, and security concerns, while in telephony there are suddenly many new possibilities related to the availability of broadband communications, the Internet and Voice-over-IP (VoIP) technologies. However, it seems that there are still some aspects missing from current telecommunications which makes it only partially acceptable replacement for travel. A realistic face-to-face experience, and the possibility to touch the other are certainly contributing factors but the experience of a common space and control of the involvement in a communication session may also play a significant role.

The topic of this chapter is ambient communication technologies. We may define an ambient communication system as a spatially distributed system of connected terminal device which enable dynamic migration of a communication session from one location in the environment to another and spatial capture and rendering for multiple simultaneous sessions. The goal of this chapter is to give a broad picture of the elements of ambient communication systems and give more detailed examples of the architectures and specific solutions needed in distributed voice-only speaker telephony, that is, ambient telephony. The main principles of ambient communications are introduced in Sections 2 and 3. In the Section 4 we give an overview of some of the existing solutions and research challenges related to the development of a full-scale ambient telephone system. Later, in Section 7 we give also some ideas on how the same concept can be extended to the visual communication technologies. We also review some approaches on user tracking and calibration of ambient communication systems in Sections 5 and 6, respectively. Finally, we discuss several additional topics in Section 8 and give a summary of areas where more research is needed.

2 The Long Call

When the first commercial telephone service was started in New Haven NY, USA, 130 years ago, the early telephones were leased in pairs [19]. There was a fixed wiring between the two devices and the connection was always open, and there were no phone bills because the call counter had not been invented yet. The modern scenario is very similar: the Voice-over-IP (VoIP) systems are also based on a peer-to-peer connection over the internet and there are no counters or phone bills.

If the connection time is not counted one could expect that the call durations increase. Recent telecom statistics [33] show that the average call durations in the traditional telephone and IP telephony are 163 and 379 seconds, respectively. What is interesting in this change is that not only the mean (or median) call duration increases but there is a new category of very long duration calls. For example, the percentage of traditional PSTN calls lasting over one hour was 0.7% [33]. The re-

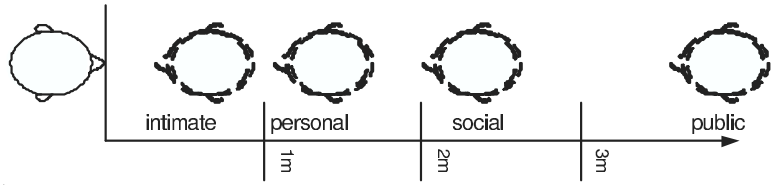

Fig. 1 Hall's classification of the social interpersonal distance as a function of the physical interpersonal distance.

sults from a Skype traffic measurement [42] indicates that 4.5% of the calls were longer than one hour and that 0.5% of the calls took more than three hours. It is known that many people take voip calls that last hours or days. These can no longer be considered as traditional telephone calls: the voip phone is used as a awareness system [75] eliciting the experience of connectedness to the other in the user's environment [1]. However, that statistics cited above show that currently few people make very long calls even if it was free. One possible reason is that the current terminal technologies do not support such use.

The typical speech rate in a traditional telephone call is around 120 words per minute. In a long duration call, it can be expected that the words per minute rate fluctuates in a similar way as in natural interaction between people who are in the same room. That is, the phone call becomes a fragmented sequence consisting of interactions and silent periods. In many ways, the form of social interaction in a continuously open telephone line could be similar to the interaction between people who, for example, live together.

If the model for the next generation telephony is taken from the natural interaction between the people who are physically present, we have to take into account also the fluctuations in the inter-personal distance. The theory of *proxemics* by Hall [44], see Fig. 1, suggests that the social distance, or the intensity of interaction, between people correlates with the physical distance. The possibility to control the interpersonal distance depending on situational, social, and emotional context is one of the missing aspects in traditional communication technologies.

For natural one-to-one conversation the typical distance is the *personal* distance around one meter. There are clear biases depending on the cultural background, personality, gender [90], and the relation between the persons, see, e.g., [97, 68] for a review. However, it seems that there is often an *optimal* interpersonal distance for communication. If the other talker is too close, e.g., closer than one meter, it may be perceived inconvenient or arousing. On the other hand, in one study [100], when the nose-to-nose distance was larger than 1.7 meters, familiar subjects in one-to-one conversation tended to search for a seating position closer to each other.

A speakerphone system aiming at mimicing physical presence should be able to support the fluctuations in the interpersonal distance. In Section 4.4 we give an overview of some of the technologies to control the distance in audio telephony.

It is clear that the traditional handset is not an optimal terminal device for a long call with a fluctuating conversational state. Holding a phone, or sitting in front of

a videoconference device for more than an hour causes fatigue and makes the user unavailable for other communication with local and other remote people. Also, if the users leave the terminal devices, it would be difficult to know when the other is next time close to the phone and available to continue the conversation. Naturally one way of staying in reach is to wear the terminal continuously, for example, a bluetooth earpiece or a head-mounted display. Advanced techniques to combine the local acoustic scene and remote persons voices using special headsets have been proposed by several authors, see, e.g, [50, 55]. However, long-term use of body-worn appliances is inconvenient for various reasons. In this chapter we focus on technologies where the same service is provided by a network of terminal devices distributed in the environment. This scenario facilitates a flexible control of spatial attributes of the presence of the remote person, which may possibly solve most of the terminal problems mentioned above.

3 Spatial Attributes Of Presence

The word presence implies that the other has some location in the user's environment. Let us first make a distinction between two extremes: telepresence and social presence [59, 13]. Telepresence is often characterized by the experience of *being there*. Telepresence technology aims at providing an experience of being present in another location, e.g., typically in virtual reality with the help of a head-mounted display and data gloves, or an immersive environments such as the CAVE environments installed in several research laboratories. Collaborative virtual environments have been studied extensively in business, conference, and e-learning applications [24], but not usually in the context of home communications.

Social presence may be characterized as an experience of the *other being here*. In ambient communication the users remain in their natural environment and therefore the focus is more specifically in social presence than in telepresence, although, it is usually not possible to make a complete separation. For example, while a high-quality speech communication system may give a social presence experience of having the other *here*, the background sounds heard in any telephony system may give also an experience of being there. The mixing of the two main branches of the presence technology in practical systems is partly due to the limitations in technology, i.e., in segmentation, transmission, and rendering of the audiovisual or multi-modal representations of the other person.

A typical use case for social presence technology is illustrated in Fig. 2, where the desired experience is the social presence of the other in the user's own natural environment. In this scenario, a representation of a remote person, the other, is virtually transferred to and rendered in the natural environment of the user. In other words, the user's environment is augmented by a mediated presence of the other. The *Augmented Reality*, AR, is produced by adding synthetic objects into the real environment [20]. Many traditional voice-only teleconference systems built in dedicated rooms already starting from 1950's can be considered as augmented reality

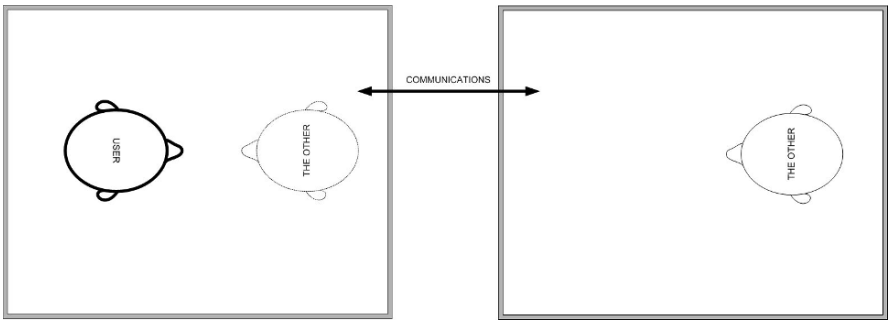

Fig. 2 *Augmented presence.*

systems, see e.g. [82, 111], for a review of historical systems. In these voice-only teleconference systems the microphone is typically placed on a meeting table and one or more loudspeakers also on the table or in some cases in in the ceiling or walls of the room. In the Remote Meeting Table system used in the 1970's by the UK Civil Service Department is a very clear example. In the reported system the voices of remote participants in a meeting room teleconference application were played from individual loudspeakers with a name sign and a activity lamp placed on a meeting table in positions where the remote person would be seated if present [97]. More recently similar systems including a small camera and a video screen have been proposed, for example, by [96, 114]. The same idea can be also taken even further by including an anthropomorphic robotic interface such as in the TELENOR system introduced in [104].

Mixed reality (MR) includes Virtual Reality (VR), AR, and a continuum between them [78, 105]. One example of a mixed reality system is illustrated in Fig. 3, where the user's physical environment is extended by an opening to the physical environment of the other. This configuration is essentially the spatial model for all videoconference systems, see [82, 62], for a review.

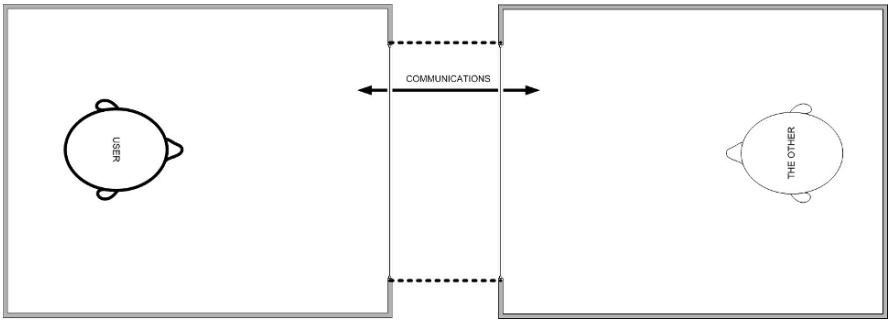

Fig. 3 *Extended space by a virtual opening.*

The opening between the two rooms in Fig. 3 resembles an open window between the two environments. Conceptually, the environment of the other is virtually moved to the neighborhood of the user and the two walls are co-positioned and removed. Communication systems based on the concept of an acoustic window [99, 46, 43], or audio-visual window [17, 54] have been proposed by many authors.

Naturally, one may also consider a mixed reality system where the remote environment, or several distant environments, are mixed with the local environment, see Fig. 4. This does not necessarily mean that all multi-modal inputs from two or more environments are mixed into one complex scene. The idea of overlapping the home floor plans is typically used in order to position representations of the remote people in an intuitive way. Grivas introduced an ambient communication system where a small number of similar devices at two homes were represented in the home of another person by colored light sources [41]. For example, if a user switched on a coffee machine, it turned on a light in the other user's home in a location corresponding to the coffee machine.

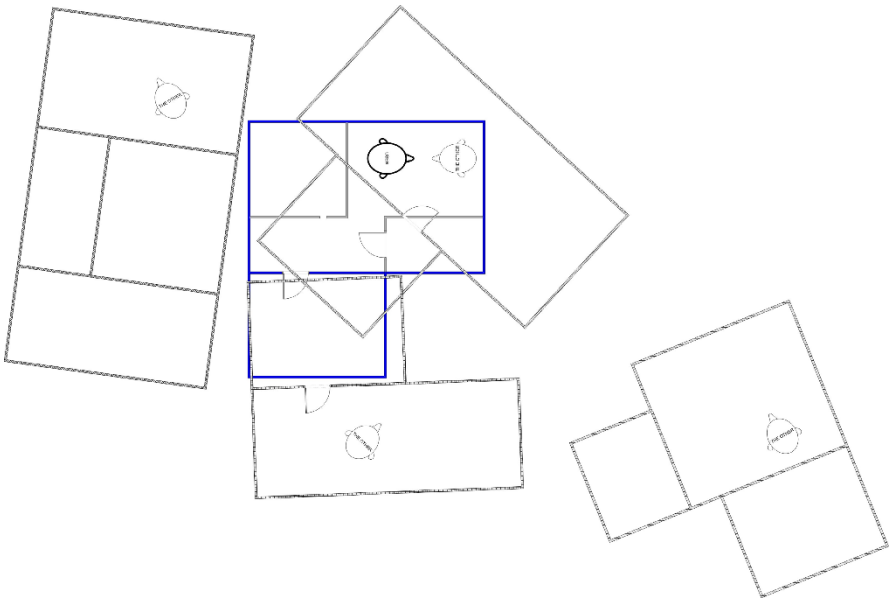

Fig. 4 *The physical spaces of the user and the others can be positioned freely, e.g., on top of each other, or as virtual neighbors.*

The technologies for real-time communication can be characterized by the map of Fig. 5. The conventional telephony falls into the left bottom corner of the map. It is *session-based* technology for which the characteristic model for interaction is the call. The call is a session which is started and terminated, and it has a high intensity at the level of 120 words per minute during the session. In the visual communication

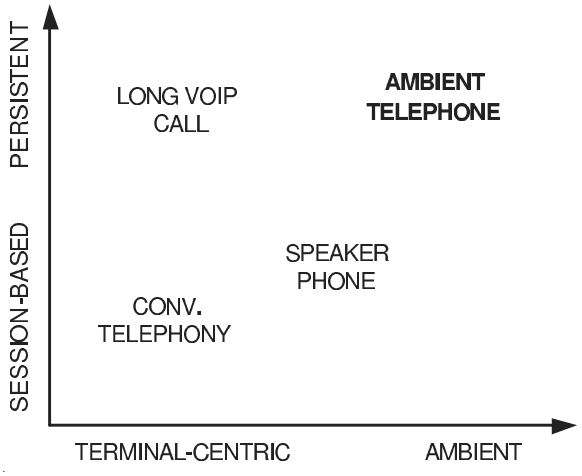

Fig. 5 The map of telephones.

there is a continuous eye-contact between the participants. A long voip call is an example of a more *persistent* form of communication where the concept of the call session is vanishing and it is replaced by a continuous acoustic presence of the other. Moreover, the traditional telephone is an example of *terminal-centric* technology where the attention of the user to the terminal itself is needed to use the system. The speakerphone is a step towards *ambient* communications where the user may carry on conversation in the neighborhood of the device.

The concept of *ambient telephony*, in the right top corner of the map of Fig. 5, is introduced in detail in the following sections. ambient telephone is a speakerphone which supports the persistent use and it is ubiquitously available in the home environment [48]. The ambient communication is essentially a service provided by the networked infrastructure in home. In some sense, this service is similar to heating, air conditioning, or lighting which are also infrastructure services provided everywhere and continuously in the modern home environment.

4 Ambient Speech Technology

The *ambient telephone* is a speakerphone system based on arrays of loudspeakers and microphones, which are distributed in the home environment and are connected to each other via a home network. The audio rendering and capture in the ambient communication system are performed in a spatially selective way. Therefore, it is possible to position the voice of a remote person and the capture position of the local speech, in principle, at any location in the environment or move it from one room to another with a local user.

The possibility to move the call from one device and one spatial location to another is one of the central features of the ambient telephone. The free mobility of the remote and local participants in the environment is one of the elementary properties of real physical presence and it is essential for the control of the interpersonal distance. The mobility also makes it possible to carry on with other activities while having a call.

Another benefit of the ambient telephone is the easy management of multiple simultaneous calls. For example, receiving a new call or making a new call while another call is still open can be performed very naturally because of the possibilities to move talkers to distinct spatial positions or leave calls open, for example, in different rooms.

4.1 Ambient Telephone Architecture

One possible architecture for an ambient telephone system is shown in Fig. 6. The system has one master phone, which is the main gateway to external communication channels such as the Internet, mobile phone network, or the public switched telephone network, PSTN. The master phone is typically the central control point of the system. The number of individual telephone units, or *phonelets*, in the system may be arbitrary and they can be placed freely. In this chapter we use the term phonelet for individual units of a distributed communication system. Note that the term has not necessarily relation to the *phonelets* used in Java-based telephone software development [67].

The communication between the phonelets and the master phone include audio streams and different types of control messages. The network platform for an ambient telephone system can be TCP/IP network where individual devices may be connected in different ways including wired, wireless, and powerline connections. However, a similar functionality can also be built on top of other network platforms including the DECT/CAT-IQ cordless telephone network [27].

In order to implement a spatially selective capture and rendering of audio it is necessary to know the physical locations of the phonelets in the environment. The configuration and calibration of the ambient communication system is discussed in Section 6.

A typical block diagram of a phonelet is shown in Fig. 7. For each incoming call each phonelet initializes a *caller instance* which contains the implementations of certain signal processing algorithms for capture and rendering needed in the distributed speakerphone. The master phone may be similar to the other phonelets but it has certain additional functions mainly related to the control of the entire phone system as illustrated in Fig. 8. The master phone also creates a new software instance for each incoming call and it is responsible for controlling the rendering and capture position of the call, and monitor the activity status of the call.

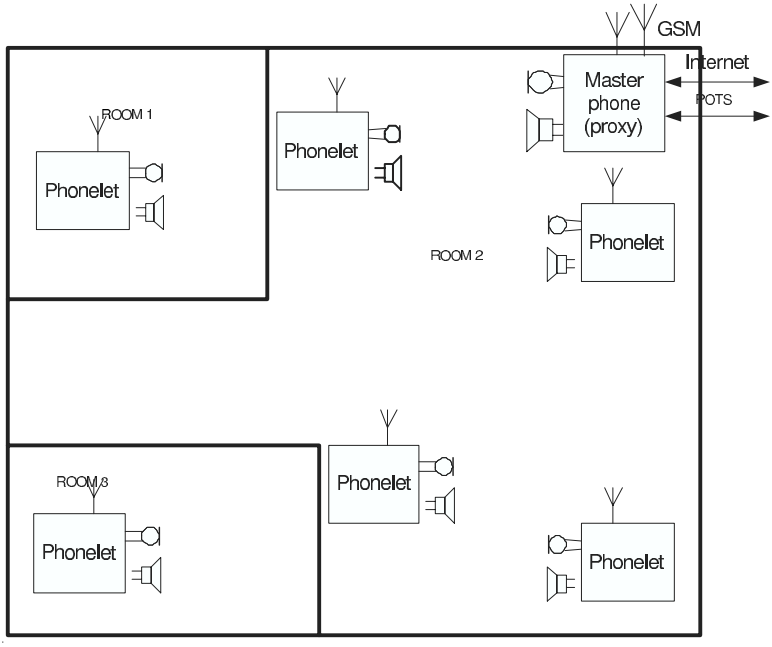

Fig. 6 A distributed ambient telephone system.

4.2 Speech Capture And Enhancement

The fundamental problem in speech capture is to maximize the signal-to-interference ratio for the desired talker. The interferences are caused by other sound sources in the environment, reverberation, and the speech of the remote person rendered to the environment.

In the ambient telephone the distance between the talker and the microphone device is often beyond the echo radius, which is the distance of the microphone from the talker where the energies of the direct sound and room reverberation are at the same level. To reduce the amount of reverberant sound it is beneficial to use microphone arrays combined with beamforming to maximize the amplitude of the direct sound [107]. The use of beamforming requires tracking of the users which is briefly discussed below. It is also often necessary to try to actively cancel unwanted sound sources using side-lobe cancellation techniques, see, [40, 61], and active noise suppressions techniques [30]. The suppression of reverberantion can also be treated separately, see, e.g., [106]. Several powerful algorithms exist for the speech enhancement using microphone arrays, see, e.g., [9, 108] for a review. However, the dynamic hand-over of speech capture in an ambient telephone system of two or more spatially separated arrays contains new challenges for speech enhancement algorithms.

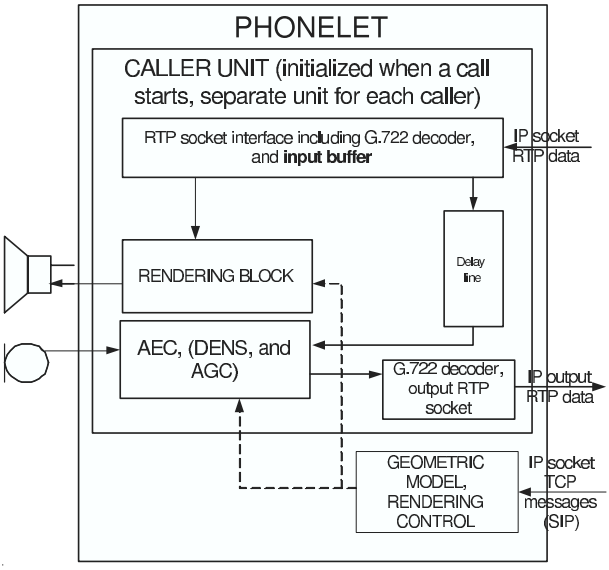

Fig. 7 A typical architecture of a phonelet.

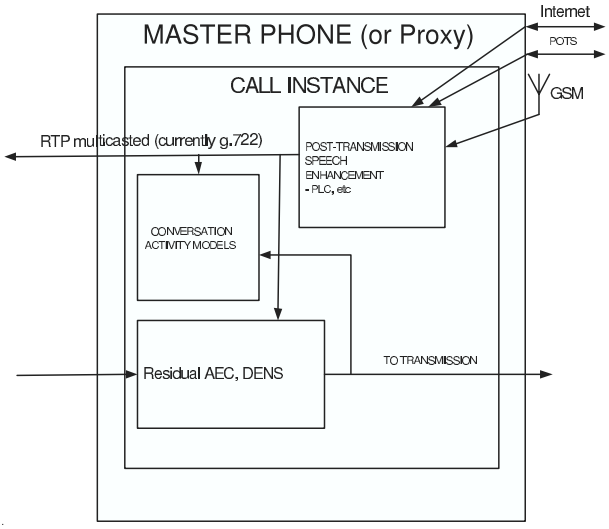

Fig. 8 A typical architecture of a master phone.

Most work for speech capture in stereophonic and multichannel audio communication have been performed in the extended space configuration of Fig. 3. The benefits of stereophonic representation of the audio are usually linked to the cocktail

party effect [21] which in this case translates to better separation and identification of far-end talkers and higher intelligibility in noisy conditions.

The problem of acoustic echo arises from the acoustic propagation of the sound played back from a loudspeaker to the microphone capturing the local speech. In the first stereophonic speakerphones [15, 3] this was solved by using half-duplex communication solution where the near-end microphones were muted when the far-end talker was active. Better solutions based on adaptive echo cancelation [101] and later adaptive multichannel echo cancelation [8] have been developed to make communication systems full-duplex. However, the near-end speech activity detection is still necessary to control the adaptation of the cancelation filters during double-talk and the residual echo attenuation. Basically all solutions are based on adaptive cancelation filters and dynamic attenuation of the remaining echo.

In a multichannel audio communication systems there is a separate acoustic propagation path from each active loudspeaker to each active microphone. The task of the echo canceller is to synthesize copies of the echo path signals using typically frequency-domain adaptive filters [98] for modeling each echo path. The generated synthetic signals are then subtracted from the microphone signals. This is illustrated in Fig. 9 where $\tilde{\mathbf{H}}$ represents a matrix of acoustic transfer functions from each loudspeaker to each microphone. The task of the multichannel echo canceller is to use a truncated estimate of the acoustic transfer functions $\mathbf{H}$, as illustrated in Fig. 9 to cancel those parts from the microphone signals. The stereophonic echo problem is significantly more challenging than the single-channel echo problem. The main problem is related to the fact that the two signals played from the loudspeakers may be, and usually are, correlated. Therefore, there is no unique way to determine which signal components observed in the microphone signal are arriving from which speaker in the room. Consequently, there is no unique solution for the normal equations needed to solve the coefficients of the adaptive filters modelling the acoustic paths. In addition, it can be shown that most possible solutions depend on the acoustic transfer functions from the talker to the microphone in the far-end room. The two most common solutions are to decorrelate the loudspeaker signal by using non-linear distortion to one of the signals [7], or to control the estimation of the adaptive filter coefficients dynamically in such a way that the filters are only adapted in frequency and time slots where the two signals are uncorrelated [94].

In principle, the use of multichannel acoustic echo cancelation of Fig. 9 is necessary when the audio is being transmitted in a stereo or multichannel representation. This is typically the case in extended space scenario [15, 17, 18, 43]. However, in the ambient telephone system for the home environment we may usually assume that the speech signals transmitted from the far-end are monophonic. Therefore, it is possible to simplify the system by at least two possible ways. Fig. 10 gives an example of a system of *multiple single-channel* echo cancelers and in Fig. 11 shows even a simpler system with only one echo canceler between the received far-end signal before spatial reproduction and the transmitted single-channel speech signal derived from the multi-microphone input. The obvious problem with the solution in Fig. 10 is that the echo path being modeled by the adaptive filter will also model spatial sound rendering method $\mathbf{R}$, which can be usually represented as a linear slowly-

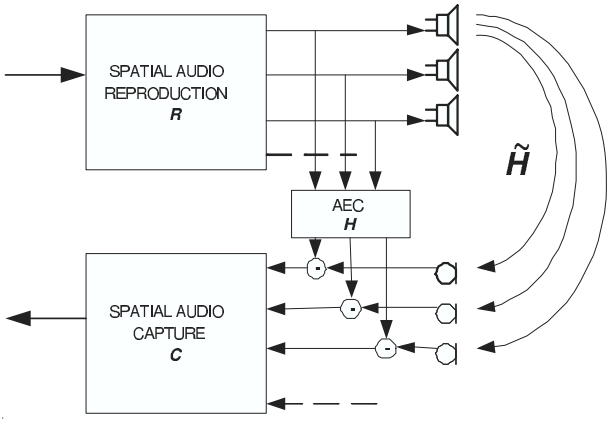

Fig. 9 A multichannel echo cancelation system.

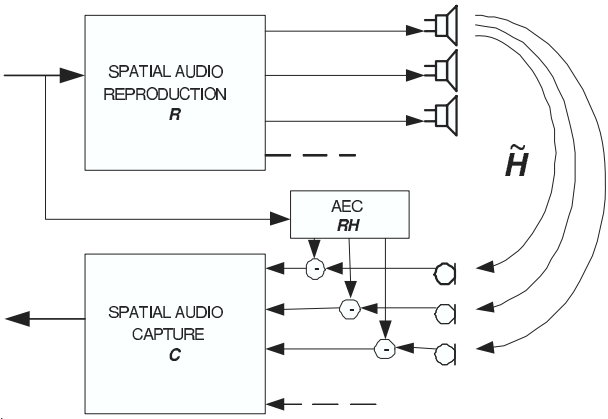

Fig. 10 A multiple single-channel echo cancelation system.

varying matrix filtering operation. In the system of Fig. 11, the modeled echo path will also contain the processing of the microphone signals, that is, the echo canceler can be characterized as a system given by **RHC** [22]. In general, it is difficult to isolate the estimation of the filter matrix representing the pure acoustic paths $\tilde{\mathbf{H}}$ in the systems of Figs. 10-11. Reed, Hawksford, and Hughes [88] have proposed a modification of an AEC algorithm where the rendering can be separated in the case of using amplitude panning in spatial sound reproduction.

The implementation of the ambient telephone system using network-connected phonelets in the way shown in Fig. 6 poses additional requirements for the algorithmic architecture. In principle it is necessary that all echo paths in the system of Fig. 9 which is a computationally expensive task. [18] and [18] and [84] have demonstrated that the complexity of the multi-channel echo canceller can be signifi-

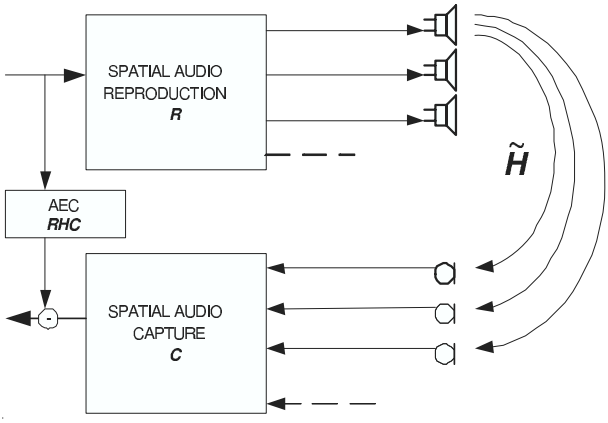

Fig. 11 A single channel echo cancelation system.

cantly reduced in the case where audio reproduction is based on wavefield synthesis of a single source signal. This can also be seen as a method for the separation of **R** from the estimation of the canceller filter. The adaptation of set of filter coefficients **H** in Fig. 9 requires that all loudspeaker and microphone signals are available in one device, which translates to a high continuous load for the local network. In addition, since there is a separate echo canceler array for each incoming caller the multi-channel echo cancelation solution quickly leads to large computational load. In the multiple single-channel solution of Fig. 10 the data transmission can be kept relatively low and there is only a single echo canceler for each caller which makes it somewhat easier to scale the solution for different numbers of phonelets. The speech enhancement and the decision to transmit the microphone data to the master phone can also be performed locally in each phonelet which reduces the network load. This is essentially the solution illustrated in Fig.6 but there is additionally a secondary echo canceler in the master phone. A similar solution has been proposed earlier, for example, by [63] but not in a context of a network-distributed telephone system.

The system of Fig.11 is also a usable solution. The main additional problem in this configuration is that also the dynamic microphone preprocessing **C** is now inside the echo cancellation loop. Modified AEC algorithms for the case of an adaptive beamformer in the echo loop have been proposed, e.g., in [63, 45]. In the real-time 16-channel hard-wired ambient telephone system installed in the HomeLab [92] of Philips Research, Eindhoven, The Netherlands in 2007, the echo cancelation solution was a single-channel echo canceler where the echo-loop contained spatial rendering and speech capture which was relatively simple microphone selection method based on user tracking. In practical tests the performance of this solution was found acceptable, although, some residual echo can be heard at the remote end especially during the periods when the rendering position of the caller is being moved from one room to another. The movement of the rendering and capture position is per-

formed by changing the rendering and capture processing, **R** and **C**, Any changes in those also naturally affect the acoustic transfer functions in **H**.

The speech capture should be controlled by speech activity detection [108]. The fluctuating activity level of long calls sets high requirements for the detection of the speech activity. In a multi-user environment, the VAD should also be combined with speaker recognition. Moreover, the problem of near-end interference is emphasized in ambient telephony because of *acoustic multitasking*, e.g., listening to music or watching television during a call, which are a natural activities in the case of real physical presence with people.

4.3 Speech Transmission

There are a large number of standardized speech coders which are used in different speech communication applications, see, for example, [102, 66], for overview. In VOIP applications standardized ITU-T G.711, G.729/G.729A, and G.723.1 coders are commonly used but there is also a large number of proprietary coders used in popular VOIP applications. In wideband speech communication the G.722 [76] and AMR-WB [11] coders are also popular. In the ambient telephone application the transmission content can be based on many different coders depending on whether the incoming call is received through a mobile or VoIP network. When the speech signal of the remote caller is played through a loudspeaker system the requirements for the voice quality are high. For example, a low-quality mobile phone speech played through a loudspeaker system may not sound acceptable. Therefore, it may be necessary to develop techniques for post-transmissions speech enhancement to alleviate the quality problems and preferable make every caller sound as good as possible. For example, in the system illustrated in Fig. 6 this is performed in the master telephone, see, Fig. 8. Typically post-transmission speech enhancement should contain algorithms for speech bandwidth extension, see [73], and packet-loss concealment in the case of VoIP transmission [108].

The requirements for the echo cancelation are coupled with the choice of the audio transmission format. For example, in the network-distributed ambient telephone system illustrated in Figs. 6-8 the residual echo cancellation is performed in the master phone. The speech signal from a phonelet should be transmitted over the local network preferably in encoded form. It is known that many sophisticated coders such as AMR-WB introduce non-linear distortions to the microphone which severely degrade the performance of the adaptive filters used in echo cancellation [47]. For this reason, the system illustrated in the figures actually uses a simple G.722 coder for in-home transmission because there the problem is much smaller.

4.4 Spatial Speech Reproduction

One of the challenges for spatial speech production is to create positioned sound sources in a very large listening area. The rendering of the speech of the other person can be performed using any of spatial audio rendering techniques including amplitude panning [86, 74], various types of holophonic reproduction methods such wave field synthesis (WFS) [10, 110], adaptive methods such as transaural reproduction [65], or adaptive wave field synthesis [36]. In all cases the listener is assumed to stay more or less in an optimal listening position, the sweet spot, or at least within a restricted listening area inside a volume enclosed by the loudspeakers. Furthermore, the sources are typically restricted to the space outside this volume. Some of these restriction should be relaxed in the ambient telephone scenario because one cannot necessarily assume that listener is seated in the *sweet spot* and this configuration does not allow flexible control of the interpersonal distance.

4.4.1 Follow-Me Effect For A Large Listening Area

In ambient telephony the user is free to move in the environment, for example, to walk from one room to another or to go outside of the listening area. In theory, it is possible to build a sound field reproduction system which can create a physically plausible approximation of the actual acoustic wavefield produced by a human talker in the environment [10]. However, this is not always feasible or desired, e.g, in the home environment, because the required number of loudspeakers increases rapidly as a function of the area or the volume of the listening space.

In practice, we are limited to a sparse and widely distributed arrays of loudspeakers. The most familiar example of such an *ambient* audio reproduction system is the public announcement system of a railway station or an airport. While the public announcement system is designed to deliver the same message to everyone in the area the goal of an ambient telephone system is to produce localized voice of a remote person to a tracked individual.

Clearly a sparse array of speakers does not allow for an exact reproduction of the desired wave field for most listening positions. In fact, it turns out that the reproduced sound field is very different from the desired physical sound field. However, in [51] it was demonstrated that this type of sparse array can actually create a very strong illusion that the sound source is moving with the listener when the listener walks, for example, through a hallway or from one room to another. This can be done simply by playing the same speech signal synchronously from all the loudspeakers. The produced effect is similar to the visual phenomenon that sun always appears moving with an observer. Fig. 12 shows the predictions of the direction of a sound source for a person walking by a line array of loudspeakers. The directional estimates were computed using a model of binaural hearing detailed in [51].

The simple method of playing identical signals from a line array of loudspeakers to create the follow-me effect for a moving observer has many shortcoming. First, for a listener who is not in front of the array but outside of the range spanned by the

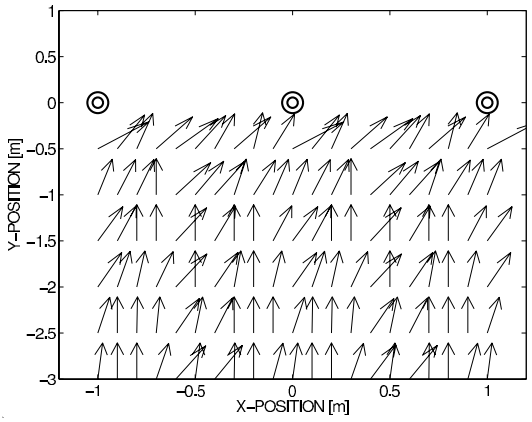

Fig. 12 A prediction of the binaural localization model of the perceived direction of the source in positions in front of a uniform line array with listeners facing parallel to the array.

two ends of the array, the sound appears severely colored due to the comb filtering effect. In hands-free telephony, it is desired that the sound is following the user. The sound source should appear localized close to the user also for a second person with a different state of motion. For a second person who is not using the application, the stationary follow-me effect may appear disturbing.

A simplest method for localized playback is to play the voice of the remote person always only from the nearest loudspeaker. Typically that transition form one speaker to another should be performed smoothly to avoid artifacts related to the onset and offset of audio. However, this method has the limitation that the voice appears jumping from one device to another.

It is possible to use multiple loudspeakers simultaneously to create localized phantom sound sources. Several dynamic spatial rendering algorithms were compared in [51]. The algorithms were compared in a listening experiment illustrated in Fig. 13. The listening test was carried out in a quiet and relatively damped listening room with a line array of eight small loudspeakers. In the experiments the subjects were asked to walk back and forth the path marked in Fig. 13. The reproduction algorithms were controlled in real-time based on camera-based tracking of the position of the subject. The subjects gave grades for various attributes related to the perceived position and movement of a sound source.

It actually turned out that it is relatively easy to create a convincing illusion that the remote talker is moving with the user. The movement in most algorithms is smooth and the coloration of sound is at a low level. However, larger differences were found for a second observer who is not moving. Therefore the authors of [51] recommended the use of a rendering technique based on the secondary source method. In the secondary source method the sound is delayed and attenuated in individual loudspeakers to simulate the actual propagation of sound from the desired source position to the position of the loudspeaker. A similar method is used in con-

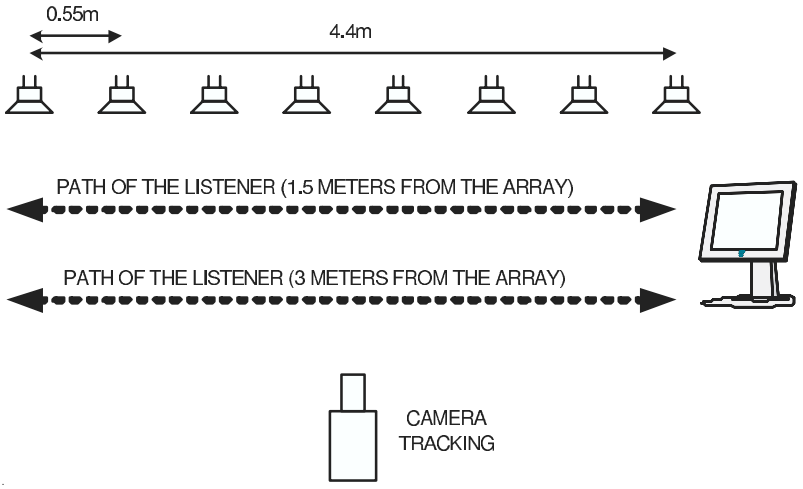

Fig. 13 The listening test setup.

ference and lecture hall reinforcement systems to improve the localization of the local talkers [26]. In the system of Figs. 6 the secondary source method can be implemented in a distributed way such that each incoming speech signal is multicasted to all phonelets where the signal is modified for reproduction in the rendering block indicated in Fig. 7 using the dynamic geometric model maintained by the master phone of Fig. 8.

4.4.2 The Control Of Interpersonal Distance

The secondary source method described above has a low complexity and it produces a clearly localized ambient speech source relatively independently of the position of the observer. However, the perceived distance of the remote person is always at the minimum at the distance of the nearest loudspeaker.

Several acoustical properties that depend on the distance between source and listener are known to affect the perception of distance. In the free field the source intensity at the position of the listener is inversely proportional to the source distance. Various studies have shown that for familiar sources the intensity correlates with perceived distance [25, 103]. The ratio between the direct and reverberant sound is inversely proportional to the source distance and provides a potential cue for distance perception [77] and the temporal structure of early reflections from walls is know to contribute [64, 16]. Other acoustical cues that can contribute to perceived distance at near distances close the listener are spectral cues due to head scattering resulting in a relative increase in low frequency energy [31], interaural level difference cues, and to a lesser extent interaural time delay cues, while at very large

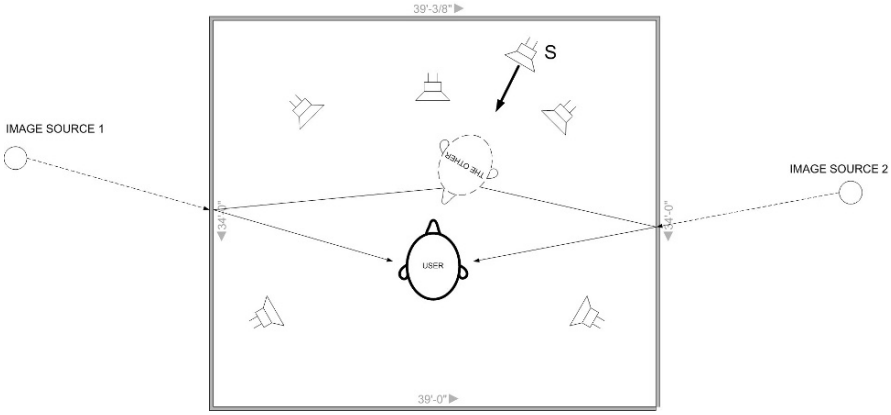

Fig. 14 The listening test setup.

distances the attenuation of high frequencies may provide cues for distance perception (cf. [25]). The dynamic distance cues such as the Doppler effect, see e.g. [49] for a review, are probably not relevant in ambient telephony because the effect becomes only audible when the speed of the sound source is high.

It is relatively easy to reproduce sound of a remote talker such that it is at any distance farther away than the nearest loudspeaker [37] but it is difficult to create an illusion that the other is at the intimate or even personal range if the loudspeakers are farther away that two meters. One of few potential technologies for creating a virtual sound source very close to a listener is an exotic parametric array reproduction technique known as the audio spotlight [116, 85]. A somewhat similar illusion can also be created using by focussed beam-shaping with an array of loudspeakers [69].

There are also possibilities to use highly directional loudspeakers as a part of the sound reproduction system for the ambient telephone to control the distance effect. A loudspeaker configuration combining a highly directional panel loudspeaker and a surround audio system was introduced in [52]. The system is shown in Fig. 14 where *S* is a directional speaker and the surrounding loudspeaker system is used to create virtual image sources [2] that correspond to the desired virtual source position. It was demonstrated in [52] that with the careful control of sound reproduction with a highly directional loudspeaker aiming at the subject and the level of simulated early reflections it is possible to create a controllable effect of having a remote talker's voice in a position between the listener and the nearest loudspeakers. However, this method again requires knowing the position of the subject and some means of controlling the radiation pattern of the directional loudspeaker.

5 Tracking And Interaction

The control of spatial rendering and capture usually requires that the position of the local talker is known. For example, in the follow-me scenario where the virtual speakerphone follows a user from one room the position of the local user should be tracked. The technologies for localization of users are reviewed in another chapter of this handbook (Part VI, Chapter 6).

The three major techniques for location-sensing are *triangulation*, *proximity* and *scene analysis*, see [56] for the taxonomy. Triangulation is based on multiple distance measurements between know points, in proximity techniques the distance to a known set of points is measured, and in scene analysis a view from a particular vantage point is examined. In many tracking methods the user is required to carry a *beacon* which may be a transmitter or a receiver device. Active RFID tags are a commonly used for tracking based on either proximity or triangulation. However, the preferred solution for ambient communication systems is *beaconless* tracking where the location is determined by the user's voice, moving body, radiated heat, or some other measurable quantity related to the presence of the user. The most commonly used beaconless tracking techniques are based on localization of active talkers using microphone arrays [28] or cameras, or a combination of both microphone and camera data, see, e.g., [35]. In the case where there are several potential users in the environment it is also necessary to identify the users in order to determine how the speech signal(s) should be rendered in the environment. This requires development of algorithms that are capable to both localize and identify the users.

Most likely a functional solution to tracking in ambient communication applications should be based on multiple modalities and high-level reasoning and learning algorithms. Software architectures for ubiquitous tracking in network environment have been proposed, see, e.g., [57, 93].

In this chapter we have mainly focused on the media technology of the ambient communication. Various topics related to the user interaction with the system and privacy need further study. The privacy issues in ubiquitous systems such as ambient telephone have been recently reviewed, e.g., by [72]. The user interaction aspects of ambient intelligence has a dedicated chapter in this book (Part III).

6 Calibration And Configuration

It is easy to build an ambient telephone setup in a laboratory environment. The loudspeakers and microphones can be put in optimal positions and the rendering and capture can be controlled by a known geometric model. When a user installs such a system in the home environment the geometry is unknown. Therefore, the system should be able to calibrate itself. The networked nodes of the system can find each other via some middleware such as UPNP [60]. However, the network discovery of devices does not provide information about the physical locations of the devices in the environment. When the devices are connected via a wireless network it is possi-

ble to find their relative locations by measuring the signal strength, time-of-arrival, or angle-of-arrival of radio frequency signal between each transmitter and receiver, see [83]. However, the RMS localization errors in WLAN and DECT are typically above 2 meters [109]. In the Ultra-Wideband (UWB) wireless technology the localization errors are well below one meter [38] which is probably sufficient for most ambient communication scenarios. The RF measurement does not provide accurate information to control audiovisual capture and rendering applications in a multi-room application. It may happen that there is no line-of-sight between two devices but they are still in the same room, or two devices may be in different rooms but only on opposite sides of a wall. For example, drywall is acoustically and visually an efficient isolator between the rooms but has a low tangent loss for UWB radio propagation [80]. Therefore, it is necessary to use also acoustic or vision techniques in the calibration of the ambient communication system.

The acoustic calibration can be performed using a separate measurement session, see, e.g. [79]. Many modern high-end surround audio systems feature automatic off-line calibration of the loudspeaker setup. Usually those are based on playing test sequences from loudspeakers and comparing those to a microphone signal captured at the central listening area. In [112], it has been demonstrated that an ad hoc network of laptops can be used as a sensor array for capturing speech from participants of a meeting and such arrays can also be calibrated automatically using off-line measurements [87]. Controlled off-line measurement provides accurate measurement data on an acoustic path. However, a separate measurement session needs to be repeated every time a new device is added to the system, or devices are moved from one place to another. The orchestration of such measurements between various devices in a dynamic environment becomes very difficult.

An alternative way of measuring the acoustic paths between devices is to perform it continuously during the normal operation of the system. These techniques are typically based on adaptive system identification where the acoustic path is modelled as a high order FIR filter and the coefficients of that filter are estimated by comparing the original signal to a captured microphone signal. For example, [71] reported that a standard FIR LMS algorithm converged to a useful solution in one room in few minutes of playing typical audio material. Adaptive estimation can be also performed simultaneously to several positions in the listening area [32]. It is also possible to embed low-level test signals into audio signals and use those in the identification of the path [81]. One potential method for the automatic measurement of acoustic paths in a system of networked audio devices was recently introduced in [47].

The calibration should ideally give a physical model of the environment the devices there which could be then used in the control of *focus* and *nimbus* [89], or the audibility and visibility of the remote and local participants. The geometric model should be the basis for the software architecture used to control an ambient communication system. The free mobility of the call in the user's environment is a challenge for the software architecture because it requires decoupling of the communication application from the individual devices such as phonelets running the software. For example, the migration of the call from one device to another be performed with the help of the Session Initiation Protocol (SIP) [58] but it essentially closes the call in

one terminal opens it in another terminal. A middleware solution more suitable for the control of ambient communication applications has been recently proposed in [93].

7 Ambient Visual Communication

Sound and light are fundamentally different in the sense that the diffraction around obstacles and reflection from surfaces largely preserves the acoustic information while visual information is usually completely lost. Therefore, direct line-of-sight is needed both in the capture of the visual representation of a subject and rendering of the received image at the receiving end.

In the extended space scenario of Fig. 3, e.g., the traditional video-telephony, the geometric properties of localized audio and video are similar. In both visual and acoustic domains the goal is to capture and reproduce a view through a window connecting the two locations. The experience of interpersonal distance is naturally an essential aspect of video communication. It can also be used, for example, as a part of the user interface of the system [91]. In the ambient communication scenario of Fig. 2, however, there is not clear matching paradigm in visual reproduction for the spatial localized audio. For example, true holographic projection systems capable of creating visible 3D characters in open air probably remain science fiction in the foreseeable future. Three-dimensional autostereographic display technologies are already mature [70] and provide interesting possibilities for the control of the interpersonal distance in conversation. However, the viewing area is still limited in front of the flat display, thus making the communication mode terminal-centric. Various volumetric display systems, see, e.g, [14] for a review, are not limited to the 2D display but the visual representation is projected inside an enclosed volume that can only be observed from outside.

In a large multi-room environment continuous capture of a freely moving subject can only be performed with a network of cameras. Multi-camera techniques for tracking and capturing of a human character have been reviewed in another chapter of this book (Part VII, Chapter 1) and, e.g., in [53].

For the rendering of the visual representation there are currently three solutions available: the use of multiple display devices, video projection [12] or pixelated light techniques integrated in lighting. In multi-display reproduction separate display devices are placed in different rooms. In the follow-me scenario they can be controlled in such a way that the current audio-visual call migrates from one display device to another following the moving user. This is technically feasible and supported by middleware solutions for networked home, however, intelligent power management techniques are needed in the displays to make the application environmentally acceptable. One possibility is to use very low-power electrophoretic display techniques (e-paper), for example, integrated in a wall paper or a piece of furniture. The video projection techniques provide interesting new opportunities due

to the progress in solid-state lighting and liquid crystal active optics which makes it possible to build power-efficient projection devices.

In e-paper and video projection on arbitrary surface it is difficult to create a visually accurate representation of the remote person, although, full-colour e-paper technology will be available in the near future and there exist also various techniques for the photometric compensation in video projection, see, e.g., [5, 34]. It is possible to render a reduced visual representation which may be sufficient for determining the pose and current activity of a remote person and also support audio telephony with hand and body gestures. For example, the remote person can be represented as an animated character or an avatar [95] which is optimized for the limited visual rendering. In video projection it is also possible to render silhouette images as shadows representing the remote people [4, 115].

8 Future Research

The topic of this chapter is ambient communication systems which are defined as distributed systems of connected terminal devices which enable dynamic migration of a communication session from one location in the environment to another and spatial capture and rendering for multiple simultaneous sessions. In this chapter we give a broad overview of existing research and challenges in the area of ambient communications rather than the algorithmic details.

If the idea of ambient human-to-human communication modeling the real physical presence of remote people in the users natural environment is taken literally, the list of open research questions is very long. In voice-only telephony the main challenge is the clean capture of speech in a noisy and reverberant environment. In the audio rendering the accurate control of the perceived interpersonal distance requires more research. In visual communication the challenges are in the capture of the character of a person from a large area in varying lighting conditions. The capture is tightly coupled with the rendering of the image in the receiving end. For ambient visual rendering there are several solutions available but they are limited by the terminal-centricity or quality and visibility.

In addition to the continuous development of algorithms for audio-visual communication there is also very interesting new developments in materials and hardware technologies. For example, the pressure-sensitive paint introduced in [39] may evolve into a usable solution for true ambient capture of speech. There the microphones can be literally painted on any surface and the pressure signal impinging the surface can be read using a laser scanner from another location in the room. For the audio reproduction the development of ferroelectret materials [6] and new advances in nano-technology [113] may offer in the future thin free-form loudspeaker panels that can be integrated in furniture or wall paper.

9 Acknowledgements

The author would like to thank S. van de Par and W. de Bruijn for some of the material reprinted here from [51, 52], and B. den Brinker, M. Bulut, and T. Määttä for valuable comments on an early manuscript.

References

[1] Ackerman MS, Starr B, Hindus D, Mainwaring SD (1997) Hanging on the 'wire': a field study of an audio-only media space. ACM Transactions on Computer-Human Interaction 4(1):39–66
[2] Allen JB, Berkley DA (1979) Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 65(4):943–950
[3] Aoki A, Koizumi N (1987) Expansion of listening area with good localization in audio conferencing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '87), pp 149 – 152
[4] Apperley M, McLeod L, Masoodian M, Paine L, Phillips M, Rogers B, Thomson K (2003) Use of video shadow for small group interaction awareness on a large interactive. In: Proc. of Fourth Australasian User Interface Conference (AUIC2003)
[5] Ashdown M, Okabe T, Sato I, Sato Y (2006) Robust content-dependent photometric projector compensation. In: Proc. of ProCams 2006: International Workshop on Projector-Camera Systems, pp 60–67
[6] Bauer S, Gerhard-Multhaupt R, Sessler GM (2004) Ferroelectrets: Soft electroactive foams for transducers. Physics Today 57:37–43
[7] Benesty J, Morgan DR, Sondhi MM (1998) A better understanding and an improved solution to the specific problems of stereophonic echo cancellation. IEEE Trans Audio Speech Processing 6:156–165
[8] Benesty J, Gaensler T, Eneroth P (2000) Multi-channel sound, acoustic echo cancellation, and multi-channel time-domain adaptive filtering. In: Acoustic Signal Processing for Telecommunications, Kluwer Academic Publ., Boston, USA, chap 6, pp 101–120
[9] Benesty J, Makino S, Chen J (eds) (2005) Speech enhancement. Springer
[10] Berkhout AJ, de Vries D, Vogel P (1993) Acoustic control by wave field synthesis. J Acoust Soc Am 93(5):2764–2778
[11] Bessette B, Salami R, Lefebvre R, Jelinek M, Rotola-Pukkila J, Vainio J, Mikkola H, Järvinen K (2002) The adaptive multirate wideband speech codec (amr-wb). IEEE Trans Speech Audio Processing 10(8):620 – 636
[12] Bimber O, Emmerling A, Klemmer T (2005) Embedded entertainment with smart projectors. IEEE Computer Society pp 48–55
[13] Biocca F, Harms C, Burgoon JK (2003) Toward a more robust theory and measure of social presence: Review and suggested criteria. Presence: Teleoperators and virtual environments 12(5):456–480

[14] Blundell BG, Schwarz AJ (2002) The classification of volumetric display systems: Characteristics and predictability of the image space. IEEE Trans Vis Computer Graphics 8(1):66–75
[15] Botros R, Abdel-Alim O, Damaske P (1986) Stereophonic speech teleconferencing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86), pp 1321 – 1324
[16] Bronkhorst A, Houtgast T (1999) Auditory distance perception in rooms. Nature 397:517–520
[17] de Bruijn W (2004) Application of wave field synthesis in videoconferencing. PhD thesis, TU Delft, Delft, The Netherlands
[18] Buchner H, Spors S, Kellermann W (2004) Wave-domain adaptive filtering: acoustic echo cancellation for full-duplex systems based on wave-field synthesis. In: Proc. Int. Conf. Acoustics, Speech, Signal Process. (ICASSP'04), Montreal, Canada, vol IV, pp 117–120
[19] Casson HN (1910) The History of the Telephone, 1st edn. A. C. McClurg & Co., Chicago, USA, univ. Virginia Library Electr. Text Center
[20] Caudell T, Mizell D (1992) Augmented reality: an application of heads-up display technology to manual manufacturing processes. In: Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, IEEE, Hawaii, vol II, pp 659 –669
[21] Cherry EC (1953) Some experiments on the recognition of speech, with one and two ears 25(5):975–979
[22] Chiucchi S, Piazza F (2000) A virtual stereo approach to stereophonic acoustic echo cancellation. In: Proc. Int. Symp. Circuits and Systems (ISCAS'00), Geneva, Switzerland, vol IV, pp 745–748
[23] Choo S, Mokhtarian PL (2005) Do telecommunications affect passanger travel or vice versa? Transportation Research Record (1926):224–232
[24] Churchill EF, Snowdon DN, Munro AJ (eds) (2001) Collaborative Virtual Environments: Digital Places and Spaces for Interaction. Springer-Verlang London Ltd
[25] Coleman P (1963) An analysis of cues to auditory depth perception in free space. Psychol Bull 60:302–315
[26] Davis D, Davis C (1997) Sound System Engineering, 2nd edn. Focal Press, Boston, MA, USA
[27] DECT (2008) Cordless advanced technology - internet and quality. http://www.dect.com
[28] Di Biase JH, Silverman HF, Brandstein MS (2001) Robust localization in reverberant rooms. In: Brandstein MS, Ward D (eds) Microphone arrays: Signal Processing Techniques and Applications, Springer-Verlag, chap 7, pp 131–154
[29] Dickson E, Bowers R (1973) The videotelephone: a new era in telecommunications. Tech. rep., Cornell University
[30] Diethorn E (2005) Foundations of spectral-gain formulae for speech noise reduction. In: Proc. Int. Workshop Acoustic Echo and Noise Control (IWAENC'2005), Eindhoven, The Netherlands

[31] DS Brunghart WR NI Durlach (1999) Auditory localization of nearby sources. II Localization of a broadband source. J Acoust Soc Am 106:1956–1968
[32] Elliott SJ, Nelson PA (1989) Multiple-point equalization in a room using adaptive digital filters. J Audio Eng Soc 37(11):899–907
[33] FCC (2007) Trends in telephone service. Tech. rep., Federal Communication Commission, Washington DC, USA
[34] Fujii K, Grossberg M, Nayar S (2005) A projector-camera system with real-time photometric adaptation for dynamic environments. In: In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, VPR 2005, vol 2, pp 1180–1187
[35] Gatica-Perez D, Lathoud G, Odobez JM, McCowan I (2007) Audiovisual probabilistic tracking of multiple speakers in meetings. IEEE Trans Audio, Speech, Language Processing 601(15):2
[36] Gauthier PA, Berry A (2006) Adaptive wave field synthesis with independent radiation mode control for active sound field reproduction: Theory. J Acoust Soc Am 119(5):2721–2737
[37] Gerzon MA (1992) The design of distance panpots. In: 92nd AES Convention preprint 3308, Vienna, Austria
[38] Gezisi S, Tian Z, Giannakis GB, Kobayashi H, molisch AF, Poor HV, Sahinoglu Z (2005) Localization via ultra-wideband radios. IEEE Signal Process Mag 22(4):70–84
[39] Gregory JW, Sullivan JP, Wanis SS, Komerath NM (2006) Pressure-sensitive paint as a distributed optical microphone array. JAcoust Soc Am 119(1):251–261
[40] Griffiths LJ, Jim CW (1982) An alternative approach to linearly constrained adaptive beamforming. IEEE Trans Antennas and Propagation AP-30(1):27–34
[41] Grivas K (2006) Digital selves: Devices for intimate communications between homes. Pers Ubiquit Comput 10:66–76
[42] Guha S, Daswani N, Jain R (2006) An experimental study of the skype peer-to-peer voip system. In: Proc. The 5th Int. Workshop on Peer-to-Peer Systems (IPTPS '06), Santa Barbara, CA, USA, pp 1–6
[43] Haapsaari T, de Bruijn W, Härmä A (2007) Comparison of different sound capture and reproduction techniques in a virtual acoustic window. In: 122nd AES Convention, paper 6995, Vienna, Austria
[44] Hall ET (1963) A system for the notation of proxemic behavior. American Anthropologist 65:1003–1026
[45] Hämäläinen M, Myllylä V (2007) Acoustic echo cancellation for dynamically steered microphone array systems. In: Proc. IEEE Workshop Appl. Digital Signal Process. Audio Acoustics (WASPAA'2007), New Paltz, NY, USA, pp 58–61
[46] Härmä A (2002) Coding principles for virtual acoustic openings. In: Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland

[47] Härmä A (2006) Online acoustic measurements in a networked audio system. In: 120th AES Convention Preprint 6666, Paris, France
[48] Härmä A (2007) Ambient telephony: scenarios and research challenges. In: Proc. INTERSPEECH 2007, Antwerp, Belgium
[49] Härmä A, van de Par S (2007) Spatial track transition effects for headphone listening. In: Proc. 10th DAFx, Bordeaux, France
[50] Härmä A, Jakka J, Tikander M, Karjalainen M, Lokki T, Hiipakka J, Lorho G (2004) Augmented reality audio for mobile and wearable appliances. J Audio Engineering Society 52
[51] Härmä A, van de Par S, de Bruijn W (2007) Spatial audio rendering using sparse and distributed arrays. In: 122nd AES Convention, paper 7056, Vienna, Austria
[52] Härmä A, van de Par S, de Bruijn W (2008) On the use of directional loudspeakers to create a sound source close to the listener. In: 124nd AES Convention, paper 7328, Amsterdam, The Netherlands
[53] Hartley RI, Zisserman A (2004) Multiple view geometry in computer vision, 2nd edn. Cambridge University Press
[54] Heeter C, J Gregg FBDD J Climo (2003) Being There: Concepts, effects and measurement of user presence in synthetic environments, Ios Press, Amsterdam, The Netherlands, chap 19: Telewindows: Case Studies in Asymmetrical Social Presence
[55] Higa K, Nishiura T, Kimura A, Shibata F, Tamura H (2007) A system architecture for ubiquitous tracking environments. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR'07), Nara, Japan
[56] Hightower J, Borriello G (2001) Location systems for ubiquitous computing. In: Dimitrova N (ed) Trends, Technologies & Applications in Mobile Computing, IEEE Computer Society, special report 5, pp 57–66
[57] Huber M, Pustka D, Keitler P, Echtler F, Klinker G (2007) A system architecture for ubiquitous tracking environments. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR'07), Nara, Japan
[58] IETF (2007) Session Initiation Protocol. http://www.cs.columbia.edu/sip/
[59] IJsselsteinjn WA (2005) 3D Video Communication, John Wiley & Sons, chap History of telepresence, pp 7–21
[60] Jeronimo M, Weast J (2003) UPnP design by example - software developers guide to Universal Plug and Play. Intel Press
[61] Johnson DH, Dungeon DE (1993) Array Signal Processing: Concepts and Techniques. Prentice Hall, Englewood Cliffs, NJ, USA
[62] Kauff P, Schreer O (2005) 3D Video Communication, John Wiley & Sons, chap Immersive videoconferencing, pp 76–90
[63] Kellermann W (1997) Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. In: P. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'97), Munich, Germany, pp 219–222
[64] Kendall G, Martens W (1984) Simulating the cues of spatial hearing in natural environments. In: Proc. International Computer Music Conf., Paris, pp 111–125

[65] Kirkeby O, Nelson PA, Orduna-Bustamante F, Hamada H (1996) Local sound field reproduction using digital signal processing. J Acoust Soc Am 100:1584–1593
[66] Kleijn WB, Paliwal KK (eds) (1995) Speech Synthesis and Coding. Elsevier Science Publ.
[67] Klinner III KV, Walker DB (2001) Building a telephone/voice portal with java. Java developers journal pp 74–82
[68] Knapp ML, Hall JA (2006) Nonverbal communication in human interaction, 6th edn. Thomson Wadsworth
[69] Komiyama S, Morita A, Kurozumi K, Nakabayshi K (1991) Distance control system for a sound image. In: AES 9th Int. Conf., Detroit, USA
[70] Kubota A, Smolic A, Magnor M, Tanimoto M, Chen T, Zhang C (2007) Multiview imaging and 3DTV. IEEE Signal Processing Magazine 10:10–21
[71] Kuriyama J, Furukawa Y (1989) Adaptive loudspeaker system. J Audio Eng Soc 37(11):919–926
[72] Langheinrich M (2005) Personal privacy in ubiquitous computing: Tools and system support. PhD thesis, Swiss Fed. Inst. Tech. Zurich, Zurich, Switzerland
[73] Larsen E, Aarts RM (2004) Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. Wiley
[74] van Leest AJ (2005) On amplitude panning and asymmetric loudspeaker setups. In: AES 119th Convention preprint 6613, New York, USA
[75] Markopoulos P, IJsselsteijn WA, Huijnen C, de Ruyter B (2005) Sharing experiences through awareness systems in the home. Interacting with Computers 17 pp 506–521
[76] Mermelstein P (1988) G.722, a new CCITT coding standard for digital transmission of wideband audio signals. IEEE Communications Magazine 26:8–15
[77] Mershon D, King E (1975) Intensity and reverberation as factors in the auditory perception of egocentric distance. Percept Psychophys 18:409–415
[78] Milgram P, Kishino F (1994) A taxonomy of mixed reality visual display. IEICE Trans Information and Systems E77-D(12):1321–1329
[79] Mourjopoulos J (1988) Digital equalization methods for audio systems. In: Proc. 84th Conv. Audio Engineering Society
[80] Muqaibel A, Safaai-Jazi A, Bayram A, Attiya AM, Riad SM (2005) Ultrawideband through-the-wall propagation. IEE Proc Microwaves, Antennas and Propagation 9:581–588
[81] Nielsen JL (1996) Maximum-length sequence measurement of room impulse responses with high level disturbances. In: AES 100th Convention Preprint 4267, Copenhagen, Denmark
[82] Olgren CH, Parker LA (1983) Teleconference technology and applications. Artech House, Inc., Dedham, MA, USA
[83] Patwari N, Ash JN, Kyperountas S, III AOH, Moses RL, Correal NS (2005) Locating the nodes. IEEE Signal Process Mag 22(4):54–69

[84] Peretti P, Palestini L, Cecchi S, Piazza F (2008) A subband approach to wave-domain filtering. In: Proc. HSCMA'2008, Trento, Italy, pp 204–207
[85] Pompei FJ (1998) The use of airborne ultrasonics for generating audible sound beams. In: 105th AES Conv. Preprint 4853, San Francisco, USA
[86] Pulkki V (1997) Virtual source positioning using vector base amplitude panning. J Audio Eng Soc 45(6):456–466
[87] Raykar VC, Kozintsev IV, Lienhart R (2005) Position calibration of microphones and loudspeakers in distributed computing platforms. IEEE Trans Audio and Speech Processing 13(1):70–83
[88] Reed MJ, Hawksford MO, Hughes P (2005) Acoustic echo cancellation for stereophonic systems derived from pairwise panning of monophonic speech. Proc IEE Vis Image Signal Process 152(1):122–128
[89] Rodden T (1996) Populating the application: A model of awareness for cooperative applications. In: Proc. ACM 1996 Conf. Computer Supported Cooperative Work, Boston, MA, USA, pp 87–96
[90] Rosegrant T, McCroskey JC (1975) The effects of race and sex on proxemic behavior in an interview setting. The Southern Speech Communication Journal 40:408–419
[91] Roussel N, Evans H, Hansen H (2004) Proximity as an interface for video communication. IEEE Multimedia 11(3):12–16
[92] de Ruyter B, Aarts E (2005) Ambient intelligence: visualising the future. In: Proc. Advanced Visual Interfaces (ACM) (AVI 2005)
[93] Schmalenstroeer J, Leutnant V, Haeb-Umbach R (2007) Amigo context management service with applications in ambient communication scenarios. In: Proc. European conference on Ambient intelligence (AMI'07), Darmstad, Germany
[94] Schobben DWE, Sommen PCW (1999) A new algorithm for joint blind signal separation and acoustic echo canceling. In: Proc. 5th Int. Symp. Signal Processing App., Brisbane, Australia, pp 889–892
[95] Schreer O, Englert R, Eisert P, Tanger R (2008) Real-time vision and speech driven avatars for multimedia applications. IEEE Trans Multimedia 10(3):352–360
[96] Sellen A, Buxton B (1992) Using spatial cues to improve teleconferencing. In: Proc. CHI'92
[97] Short J, Williams E, Christie B (1976) The social psychology of telecommunications. John Wiley & sons
[98] Shynk JJ (1992) Frequency-domain and multirate adaptive filtering. IEEE Signal Processing Magazine pp 14–37
[99] Snow WB (1934) Auditory perspective. Bell Laboratories Record 12(7):194–198
[100] Sommer R (1961) Leadership and group geometry. Sociometry 24:99–110
[101] Sondhi MM (1967) An adaptive echo canceller. Bell Syst Tech J XLVI:497–510
[102] Spanias AS (1994) Speech coding: a tutorial review. Proc IEEE 82(10):1541–1582

[103] Strybel T, Perrott D (1984) Discrimination of relative distance in the auditory modality: The success and failure of the loudness discrimination hypothesis. J Acoust Soc Am 76:318–320
[104] Tadakuma R, Asahara Y, Kawakami HKN, Tachi S (2005) Development of anthropomorphic multi-d.o.f. master-slave arm for mutual telexistence. IEEE Transactions on Visualization and Computer Graphics 11(6):626–636
[105] Tamura H, Yamamoto H, Katayama A (2001) Mixed reality: Future dreams seen at the border between real and virtual worlds. IEEE Computer Graphics and Applications 21(6):64–70
[106] Triki M, Slock DTM (2008) Robust delay-predict equalization for blind SIMO channel dereverberation. In: Proc. HSCMA'2008, Trento, Italy, pp 248–251
[107] Van Veen BD, Buckley KM (1988) Beamforming: A versatile approach to spatial filtering. IEEE ASSP Magazine 5(2):4–24
[108] Vary P, Martin R (2006) Digital Speech Transmission: Enhancement, Coding and Error Concealment. John Wiley & Sons, Chichester, England
[109] Vossiek M, Wiebking L, Gulden P, Wieghardt J, Hoffmann C, Heide P (2003) Wireless local positioning. IEEE Microwave Magazine pp 77–86
[110] Ward DB, Abhayapala TD (2001) Reproduction of plane-wave sound field using an array of loudspeakers. IEEE Trans Speech, Audio Processing 9(6):697–707
[111] Watanabe K, Murakami S, Ishikawa H, Kamae T (1985) Audio and visual augmented teleconferencing. Proc IEEE 73(4)
[112] Wehr S, Kozintsev I, Lienhart A, Kellermann W (2004) Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation. In: Proc. IEEE Sixth Int. Symp. Multimedia Software Eng. (ISMSE'04)
[113] Xiao L, Chen Z, Feng C, Liu L, Bai ZQ, Wang Y, Qian L, Zhang Y, Li Q, Jiang K, Fan S (2008) Flexible, stretchable, transparent carbon nanotube thin film loudspeakers. Nano Letters
[114] Yankelovich N, Simpson N, Kaplan J, Provino J (2007) Porta-person: Telepresence for the connected conference room. In: Proc. CHI'2007, San Jose, CA, USA
[115] Yasuda S, Hashimoto S, Koizumi M, Okude N (2007) Teleshadow: Shadow lamp to feel other presense. In: Proc. SIGGRAPH'2007, San Diego, CA, USA
[116] Yoneyama M, Fujimoto JI (1983) The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. J Acoust Soc Am 73(5):1532–1536