Ambient telephony: scenarios and research challenges

Aki Härmä

Philips Research Laboratories, Eindhoven, The Netherlands

aki.harma@philips.com

Abstract

Telecommunications at home is changing rapidly. Many people have moved from the traditional PSTN phone to the mobile phone. Now for increasingly many people Voice-over-IP telephony on a PC platform is becoming the primary technology for voice communications. In this tutorial paper we give an overview of some of the current trends and try to characterize the next generation of home telephony, in particular, the concept of ambient telephony. We give an overview of the research challenges in the development of ambient telephone systems and introduce some potential solutions and scenarios.

Index Terms: hands-free telephone, networking

1. Introduction

When the first commercial telephone service was started in New Haven NY, USA, 130 years ago, the early telephones were leased in pairs [1]. 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 [2] shows that the average call durations in the traditional telephone and IP telephony are 163s and 379s, 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%[2]. The results from a Skype traffic measurement [3] 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 [4] eliciting the experience of connectedness to the other in user's environment [5].

The typical speech rate in a traditional telephone call is around 120 words per minute. In a long 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 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 [6], see Fig. 1, suggests that the social distance, or the intensity of interaction, between people correlates with the physical distance. Naturally a speakerphone system aiming at following this model should be able to support also the fluctuations in the interpersonal distance.

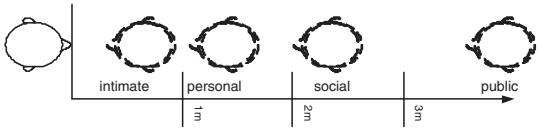

Figure 1: Hall's classification of the social interpersonal distance as a function of the physical interpersonal distance.

It is clear that the traditional handset is not an optimal terminal device for a long call. For example, it would be difficult to know when the other is close to the phone and available for conversation. Naturally one way of staying in reach is to wear continuously, for example, a bluetooth earpiece. However, in this paper we focus on technologies that are based on the speakerphone concept where the speech is captured and reproduced using microphones and loudspeakers installed in the home environment and the user is not wearing the technology.

The technologies for the speech communication can be characterized by the map of Fig. 2. 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. The long voip call is an example of a more *persistent* form of telephony where the concept of a telephone 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 no direct attention to the device itself is needed, although, it may still be required to stay close to the phone.

The concept of *ambient telephony*, in the right top corner of the map of Fig. 2, is the topic of the current article. Ambient telephone is a speakerphone which supports the persistent use and it is ubiquitously available in the home environment. The ambient telephone 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.

In the following sections we give an overview of research challenges related to the development of a full-scale ambient telephone system.

10.21437/Interspeech.2007-250

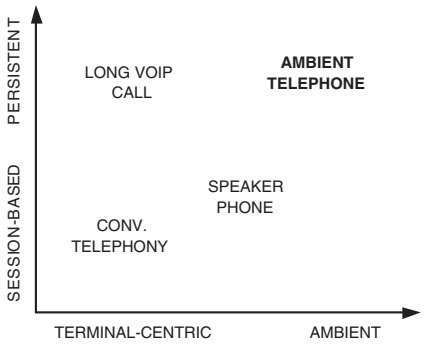

Figure 2: The map of telephones.

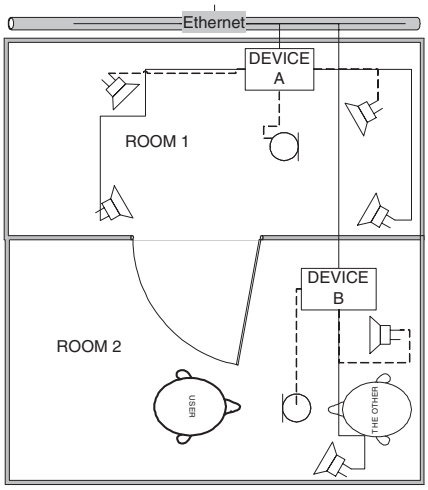

Figure 3: An example of an ambient telephone system.

2. Ambient speech technology

For the purposes of this paper, 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 system can receive calls from any source via a central device connected to the Internet, cellular phone network, and possibly the traditional land line. The audio rendering and capture can be performed in a spatially selective way. For example, a user can carry on conversation with the other such that the other appears moving smoothly with the talker from one room to another, or such that there are several simultaneous connections open and the contacts are rendered in spatially separate positions in the home environment.

Figure 3 gives one example of an ambient telephone system. Let us study the following scenario. The user is in Room 2, for example, working with a PC and having a conversation with a remote person using an IP telephone system integrated in the PC. At some point the user wants to move to Room 1 to start watching a movie in a home theater system, but continue the conversation with the other without interruption.

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 discussed in this paper. Basically, the migration of the call from Device B in Room 2 to Device A connected to the audio system in Room 1 can be performed using the Session Initiation Protocol (SIP) [7]. There the call is terminated in Device B and a new session is started between Device A and the other. Currently, SIP (or other common middleware) does not support a smooth transition of the call session from one device to another. However, it is possible to use one device as the control point for the ambient telephone and others as external speaker and microphone devices that are controlled dynamically.

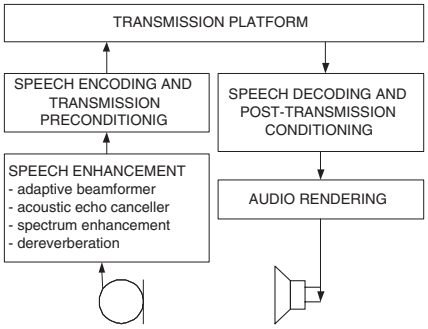

Figure 4: A generic processing model for a speakerphone.

2.1. Post-transmission speech enhancement

A generic block diagram of a speakerphone is shown in Fig. 4. The deterioration of the decoded speech signal can be very different depending on the transmission medium. Ideally, the speech quality should be similar to the quality of voices of the people physically present in the environment. Therefore, it could also be desired to normalize certain properties of speech signals in such a way that they appear subjectively similar independently of the transmission medium and the far-end terminal device (e.g., another speakerphone, or a mobile telephone). The largest differences are the bandwidth reduced by speech coding algorithms, and the level of background noise. The missing part of the spectrum can be synthesized using techniques of bandwidth extension (BWE), see, e.g. [8]. In IP telephony the packet losses is a fundamental problem and several packet-loss concealment (PLC) algorithms have been introduced. Since the aim of the BWE is to fill the missing spectrum information and PLC aims at filling missing time segments of the speech signals it is convenient to see them as components of a generic *post-transmission speech enhancement* module shown in Fig. 4. Naturally, the basis for both should be a unified model of a speech signal.

2.2. Spatial speech reproduction

The rendering of the speech of the other person can be performed using any of spatial audio rendering techniques including amplitude panning [9], various types of holophonic reproduction methods such wave field synthesis (WFS) [10], adaptive methods such as transaural reproduction [11], or adaptive wave field synthesis [12]. The follow-me effect where the other appears moving with the user from one room to another can be created using many different techniques. It was demonstrated in a recent paper [13] that the follow-me effect can be produced relatively easily even with a sparse array of loudspeakers.

The biggest challenge for spatial speech production is to

create positioned sound sources in a very large listening area. In addition, the reproduction methods should also be able to support different interpersonal distances illustrated in Fig. 1. It is relatively easy to render the other such that it is at the *public* distance 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 the few know technologies is an exotic parametric array reproduction known as the audio spotlight [14].

2.3. 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 other sound sources in the environment, the speech of the other rendered to the environment, and the reverberation.

In the ambient telephone the distance between the talker and the microphone device is often beyond the echo radius. Therefore, it is necessary to use microphone arrays typically combined with beamforming to capture the voice of the talker as cleanly as possible. It is also often necessary to try to actively cancel unwanted sound sources using side-lobe cancellation techniques [15]. Several powerful algorithms exist for the speech enhancement using one microphone array [16]. However, the dynamic hand-over of speech capture in an ambient telephone system of two or more spatially separated arrays contains some new challenges for speech enhancement algorithms.

The speech sounds rendered in the same environment produces an echo for the far-end talker. Therefore the first stereophonic speakerphones [17, 18] were based on half-duplex communication where the microphone of the other was muted when the near-end talker was active. Techniques for adaptive echo cancellation [19] and later adaptive multichannel echo cancellation [20] have been studied widely in the literature. However, the near-end speech activity detection is still usually necessary, at least, in limiting the adaptation of the cancellation filters during double-talk. The double-talk problem is emphasized in ambient telephony because people may want to listen to music or watch TV while talking to their friends. We will return to the problems of multi-tasking at the end of this article.

In the ambient telephone there are several possible places in the signal path where the echo cancellation could be performed. The network of Fig. 4 is one alternative where the echo cancellation is performed between the outputs and inputs of individual loudspeakers and microphones. This leads to a multichannel echo cancellation formulation. However, the echo cancellation can also be performed between the monophonic signal before the spatial rendering and the individual microphone signals, or even the speech signal received after the beamformer. In this configuration the spatial rendering technique and the (adaptive) beamformer effectively becomes a part of the echo path to be cancelled but the benefit is that one can use a monophonic echo canceller.

The fluctuating activity level of long calls sets high requirements for the detection of the speech activity. The voice activity detection (VAD) from a large distance in a reverberant environment is difficult. In a multi-user environment, the VAD should also be combined with speaker recognition.

2.4. Tracking and interaction

The use of beamforming for the speech capture requires that the position of the talker is known. Also the follow-me scenario where the virtual speakerphone follows a user from one room to another requires tracking. The use of RFID tags is currently a popular technology for tracking, however, it is not very accurate and also requires that the user is carrying the tag. The most commonly used beaconless tracking techniques are based on localization of active talkers using microphone arrays [21] or cameras, or a combination of both microphone and camera data, see, e.g., [22]. 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.

2.5. 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 calibration can be performed using a separate measurement session, see, e.g. [23]. 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 [24], 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 [25]. 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, Kuriyama *et al.* reported that a standard FIR LMS algorithm converged to a useful solution in one room in few minutes of playing typical audio material [26]. Adaptive estimation can be also performed simultaneously to several positions in the listening area [27]. It is also possible to embed low-level test signals into audio signals and use those in the identification of the path [28]. One potential method for the automatic measurement of acoustic paths in a system of networked audio devices was recently introduced in [29].

3. Multi-tasking

It is not a problem to use the traditional handset and watch movie, or listen to music simultaneously. However, for the ambient telephone *multi-tasking* makes many things including speech capture and user tracking more complicated. All additional sound sources are essentially noise sources degrading the quality of the captured speech. Moreover, a recent report [30] revealed that the age group 8-18 years has accustomed to use several media simultaneously. This predicts that the problem of noise is only going to get worse in the future home environment.

4. Conclusions

One interesting direction for the development of home telephony is to make it an ambient service available everywhere in the home environment. In this tutorial paper we give an overview of the technical challenges in developing a full-scale *ambient telephone* for the home environment. Generally, technologies already exist. However, the home environment with possibly several potential users and the presence of other simultaneous media applications sets high requirements especially for clean capture of desired speech signals.

5. References

[1] H. N. Casson, *The History of the Telephone*. Chicago, USA: A. C. McClurg & Co., 1 ed., 1910. Univ. Virginia Library Electr. Text Center.

[2] FCC, "Trends in telephone service," tech. rep., Federal Communication Commission, Washington DC, USA, February 2007.

[3] S. Guha, N. Daswani, and R. Jain, "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, February 2006.

[4] P. Markopoulos, W. A. IJsselsteijn, C. Huijnen, and B. de Ruyter, "Sharing experiences through awareness systems in the home," *Interacting with Computers 17*, pp. 506–521, 2005.

[5] M. S. Ackerman, "Hanging on the wire: a field study of an audio-only media space," *ACM Transactions on Computer-Human Interaction*, pp. 39–66, 1997.

[6] E. T. Hall, "A system for the notation of proxemic behavior," *American Anthropologist*, vol. 65, pp. 1003–1026, 1963.

[7] IETF, "Session Initiation Protocol." http://www.cs.columbia.edu/sip/, 2007.

[8] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. Wiley, 2004.

[9] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[10] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, pp. 2764–2778, May 1993.

[11] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local sound field reproduction using digital signal processing," *J. Acoust. Soc. Am.*, vol. 100, pp. 1584 – 1593, 1996.

[12] P.-A. Gauthier and A. Berry, "Adaptive wave field synthesis with independent radiation mode control for active sound field reproduction: Theory," *J. Acoust. Soc. Am.*, vol. 119, pp. 2721–2737, May 2006.

[13] A. Härmä, S. van de Par, and W. de Bruijn, "Spatial audio rendering using sparse and distributed arrays," in *AES 122nd Conv. Preprint*, (Vienna, Austria), May 2007.

[14] F. J. Pompei, "The use of airborne ultrasonics for generating audible sound beams," in *105th AES Conv. Preprint 4853*, (San Francisco, USA), September 1998.

[15] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. AP-30, pp. 27–34, January 1982.

[16] J. Benesty, S. Makino, and J. Chen, eds., *Speech enhancement*. Springer, 2005.

[17] R. Botros, O. Abdel-Alim, and P. Damaske, "Stereophonic speech teleconferencing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, pp. 1321 – 1324, April 1986.

[18] A. Aoki and N. Koizumi, "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, April 1987.

[19] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. XLVI, pp. 497–510, March 1967.

[20] J. Benesty, T. Gaensler, and P. Eneroth, "Multi-channel sound, acoustic echo cancellation, and multi-channel time-domain adaptive filtering," in *Acoustic Signal Processing for Telecommunications*, ch. 6, pp. 101–120, Boston, USA: Kluwer Academic Publ., 2000.

[21] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: Signal Processing Techniques and Applications* (M. S. Brandstein and D. Ward, eds.), ch. 7, pp. 131–154, Springer-Verlag, 2001.

[22] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Language Processing*, vol. 601, p. 2, February 2007.

[23] J. Mourjopoulos, "Digital equalization methods for audio systems," in *Proc. 84th Conv. Audio Engineering Society*, May 1988.

[24] S. Wehr, I. Kozintsev, A. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed adhoc audio networks and its use for blind source separation," in *Proc. IEEE Sixth Int. Symp. Multimedia Software Eng. (ISMSE04)*, 2004.

[25] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Audio and Speech Processing*, vol. 13, pp. 70–83, January 2005.

[26] J. Kuriyama and Y. Furukawa, "Adaptive loudspeaker system," *J. Audio Eng. Soc.*, vol. 37, pp. 919–926, November 1989.

[27] S. J. Elliott and P. A. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *J. Audio Eng. Soc.*, vol. 37, pp. 899–907, November 1989.

[28] J. L. Nielsen, "Maximum-length sequence measurement of room impulse responses with high level disturbances," in *AES 100th Convention Preprint 4267*, (Copenhagen, Denmark), May 1996.

[29] A. Härmä, "Online acoustic measurements in a networked audio system," in *120th AES Convention Preprint 6666*, (Paris, France), May 2006.

[30] D. F. Roberts, U. G. Foehr, and V. Rideout, "Generation M: Media in the lives of the 8-18 year-olds," tech. rep., Kaiser Family Foundation, March 2005.