1-1-2008

Spatialized teleconferencing: recording and 'Squeezed' rendering of multiple distributed sites

Eva Cheng
University of Wollongong, ecc04@uow.edu.au

Bin Cheng
University of Wollongong, bc362@uow.edu.au

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

I. Burnett
Royal Melbourne Institute of Technology, ianb@uow.edu.au

Recommended Citation
Cheng, Eva; Cheng, Bin; Ritz, Christian H.; and Burnett, I.: Spatialized teleconferencing: recording and 'Squeezed' rendering of multiple distributed sites 2008, 411-416.
https://ro.uow.edu.au/infopapers/1581

Spatialized teleconferencing: recording and 'Squeezed' rendering of multiple distributed sites

Abstract

Teleconferencing systems are becoming increasing realistic and pleasant for users to interact with geographically distant meeting participants. Video screens display a complete view of the remote participants, using technology such as wraparound or multiple video screens. However, the corresponding audio does not offer the same sophistication: often only a mono or stereo track is presented. This paper proposes a teleconferencing audio recording and playback paradigm that captures the spatial location of the geographically distributed participants for rendering of the remote soundfields at the users' end. Utilizing standard 5.1 surround sound playback, this paper proposes a surround rendering approach that `squeezes' the multiple recorded soundfields from remote teleconferencing sites to assist the user to disambiguate multiple speakers from different participating sites.

Keywords

Spatialized, teleconferencing, recording, Squeezed, rendering, multiple, distributed, sites

Disciplines

Physical Sciences and Mathematics

Publication Details

E. Cheng, B. Cheng, C. H. Ritz & I. S. Burnett, "Spatialized teleconferencing: recording and "Squeezed" rendering of multiple distributed sites," in Australian Telecommunication Networks and Applications Conference, 2008, pp. 411-416.

Spatialized Teleconferencing: Recording and 'Squeezed' Rendering of Multiple Distributed Sites

Eva Cheng[1], Bin Cheng[1], Christian Ritz[1], Ian S. Burnett[2]

[1]Whisper Laboratories
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong NSW Australia 2522
{ecc04, bc362, critz}@uow.edu.au

[2]School of Electrical and Computer Engineering
Royal Melbourne Institute of Technology, Melbourne, VIC Australia 3000
ian.burnett@rmit.edu.au

Abstract-**Teleconferencing systems are becoming increasing realistic and pleasant for users to interact with geographically distant meeting participants. Video screens display a complete view of the remote participants, using technology such as wraparound or multiple video screens. However, the corresponding audio does not offer the same sophistication: often only a mono or stereo track is presented. This paper proposes a teleconferencing audio recording and playback paradigm that captures the spatial location of the geographically distributed participants for rendering of the remote soundfields at the users' end. Utilizing standard 5.1 surround sound playback, this paper proposes a surround rendering approach that 'squeezes' the multiple recorded soundfields from remote teleconferencing sites to assist the user to disambiguate multiple speakers from different participating sites.**

I. Introduction

Teleconferencing is an efficient and effective technology for connecting geographically distributed participants in meetings for business, education, or for connecting remote communities. Commercial teleconferencing systems currently available, although offering sophisticated video stimulus of the remote participants, commonly employ only mono and stereo audio playback for the user; however, telepresence can be greatly improved by spatializing the audio (using headphones or loudspeakers) to assist listeners to distinguish between (concurrent) participating speakers [1][2][3].

A recent system that addresses spatialized teleconferencing audio uses online avatars to co-locate remote participants over the Internet in virtual space with (binaural) audio spatialized over headphones [4]. Vocal Village [4] adds speaker location cues to monaural speech to *create* a user-manipulable soundfield that matches the avatar's position in the virtual space; in contrast, the proposed approach in this paper 'squeezes' the *original* recorded meeting speech soundfield into 'sectors' of the users listening soundfield (where the sector width depends on how many remote meetings need to be spatially disambiguated). A different approach was introduced in [5], which applied the Directional Audio Coding (DirAC) technique to record, efficiently transmit, and render the remote spatial soundfield; however, the DirAC approach did not address the spatialization of multiple remote sites, and required specific Ambisonic recording hardware, which can be expensive.

To improve the users' feel of telepresence, this paper proposes a teleconferencing recording and playback system that spatially records and unambiguously renders multiple remote auditory soundfields. For maximum flexibility, the system proposed in this paper utilizes a standard 5.1 playback system for rendering and does not require specific recording hardware, analysis algorithms or software at participating sites: only a mono speech stream accompanied by speaker azimuth metadata is required for spatial rendering in 5.1 surround. This paper merges multiple remote soundfields unambiguously into a 5.1 surround setup at the users' end: a novel algorithm to 'squeeze' multiple soundfields together is introduced, adopted from the authors' Spatially Squeezed Surround Audio Coding (S^3AC) technique [6].

In the remainder of this paper, Section II describes the proposed system and the core technologies required for spatial teleconferencing speech recording and the proposed spatial rendering of participants at remote sites at the users' end. Section III details the simulations and speech recordings used to demonstrate the proposed system, with the results presented in Section IV. Section V thus concludes this paper.

II. Proposed System

Fig. 1 illustrates the proposed teleconferencing recording and playback system. With N geographically distributed sites concurrently participating in the teleconference of Fig.

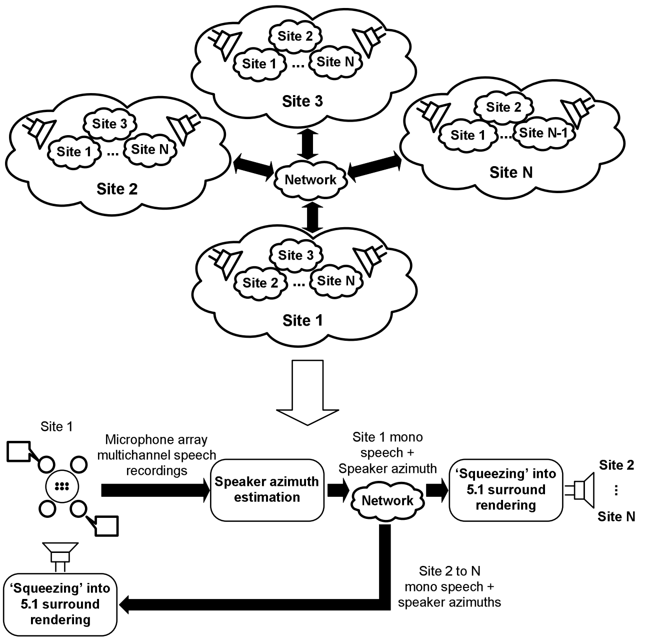

Site 2
Site 1
...
Site N
Site 3
Site 3
Site 1
...
Site N
Site 2
Network
Site 2
Site 1
...
Site N-1
Site N
Site 3
Site 2
...
Site N
Site 1
Site 1
Microphone array multichannel speech recordings
Speaker azimuth estimation
Site 1 mono speech + Speaker azimuth
Network
'Squeezing' into 5.1 surround rendering
Site 2
⋮
Site N
Site 2 to N mono speech + speaker azimuths
'Squeezing' into 5.1 surround rendering

Fig. 1. Proposed teleconferencing system

1, each site must thus unambiguously spatialize N - 1 remote sites. The two main components of the proposed system are: (spatial) recording and efficient transmission of speech and spatial metadata between sites e.g., over the Internet, and merging the $N - 1$ remote soundfields at each site using the proposed 'squeezing' approach adopted from S^{3}AC.

A. *Spatial Meeting Speech Recording*

Multiparty meetings are generally recorded with multiple (omnidirectional) microphones, arranged in an array for signal enhancement and processing e.g., beamforming, localization, etc. For the system proposed in this paper, to spatially render and merge multiple soundfields from remote sites, the only recording requirements of participating sites are a mono speech stream transmitted with the speaker azimuth metadata. Thus, any recording hardware setup and speaker azimuth estimation algorithm can be employed: without loss of generality the sites in this paper each employ a four-element array of omnidirectional microphones, with the speaker azimuths estimated using the Steered Response Power with PHAse Transform (SRP-PHAT [7]) algorithm. SRP-PHAT is widely used for speech source localization, as it has been shown to accurately localize (multiple) speakers utilizing short analysis frames and in reverberant acoustic environments (e.g., most meeting rooms) [6].

SRP-PHAT builds upon the Generalized Cross Correlation with PHAT (GCC-PHAT) algorithm, a well known time-delay estimation (TDE) technique shown to reliably estimate TDE with reverberant speech (due to the PHAT weighting function) [8]. The performance of GCC-PHAT improves with longer analysis frames, which is suboptimal for real-time or delay-sensitive applications such as teleconferencing. Furthermore, GCC-based techniques cannot estimate TDE from multiple concurrent speakers; rather, TDE techniques detect the strongest speaker in each analysis frame [6].

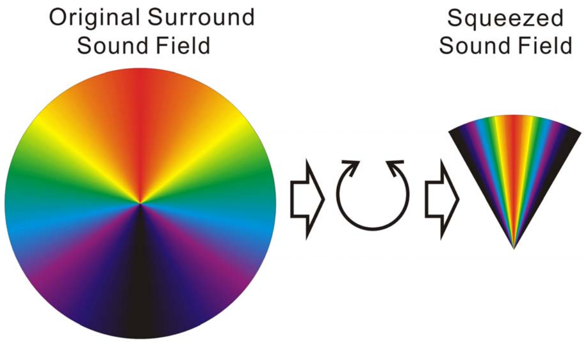


Fig. 2. The squeezing approach of S^{3}AC [6]

SRP-PHAT overcomes the shortcomings of GCC-PHAT by employing the PHAT weighting to a delay-and-sum beamforming approach for speech source azimuth estimation. For the microphone pair between channels m and n with TDE τ_{mn}, the TDE $\hat{\tau}$ estimated by GCC-PHAT is given by:

$$\hat{\tau} = \arg\max_{\tau_{mn}} \left(\int_{-\infty}^{+\infty} \frac{X_m(\omega)\cdot X_n^*(\omega)}{\left|X_m(\omega)\cdot X_n^*(\omega)\right|} e^{j\omega\tau_{mn}} d\omega \right) \qquad (1)$$

where the Discrete Fourier Transform (DFT) of the m^{th} microphone channel $x_m(n)$ is denoted by $X_m(\omega)$. SRP-PHAT thus employs GCC-PHAT in a delay-and-sum beamformer to calculate the SRP, $P(\mathbf{q})$:

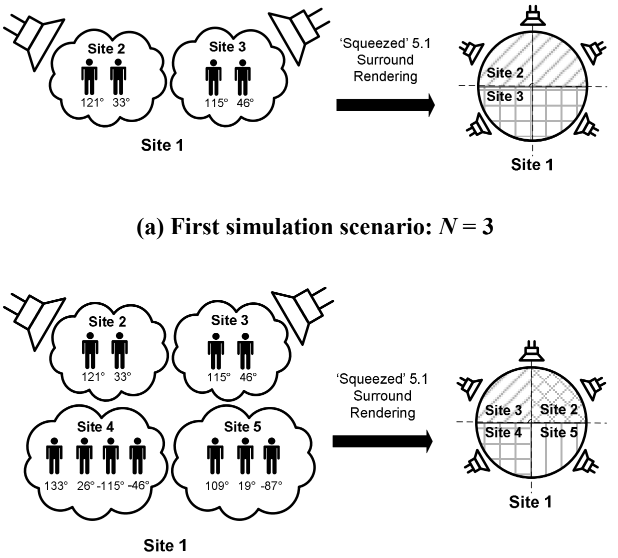
$$P(\mathbf{q}) = \left(\sum_{n=1}^{C}\sum_{m=1}^{C} \int_{-\infty}^{+\infty} \frac{X_m(\omega)\cdot X_n^*(\omega)}{\left|X_m(\omega)\cdot X_n^*(\omega)\right|} e^{j\omega\Delta_{mn}(\mathbf{q})} d\omega \right) \qquad (2)$$

where C is total number of microphone channels, $\Delta_{mn}(\mathbf{q})$ is the steering delay between each candidate source location $\mathbf{q}$ of the SRP search space and microphone pair between channels m and n. It has been shown that the SRP $P(\mathbf{q})$ in (2) can be formed by summing the GCC from all possible microphone pairs time-shifted by the steering delays for each location $\mathbf{q}$ [6]. The estimated source location $\hat{\mathbf{q}}$ is thus computed as the candidate location $\mathbf{q}$ that maximizes $P(\mathbf{q})$:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} P(\mathbf{q}) \qquad (3)$$

Such an exhaustive search of all $\mathbf{q}$ defined a priori in the SRP search space can be computationally expensive; however, recent work in search space reduction and search optimization has enabled real-time implementations of SRP-PHAT [9][10]. SRP-PHAT thus requires knowledge of the microphone array geometry and room dimensions to generate the SRP search space, but it is assumed that this will generally be known (or easily calculated) for teleconferencing rooms.

In addition, echo cancellation at each site must be performed to remove the 5.1 surround playback of remote sites from the microphone array recordings at each site. This paper does not implement echo cancellation as experiments



(a) First simulation scenario: *N* = 3

(b) Second simulation scenario: *N* = 5

Fig. 3. Simulation scenarios

simulate the remote site recordings and re-spatialized 'squeezed' soundfield at the user's site; however, any echo cancellation approach may be employed e.g., directional-nulling as used in [5] (since the 5.1 speaker locations are known).

The system proposed in this paper requires the speaker azimuth location estimate to be transmitted accompanying a mono meeting speech signal e.g., one of the microphone channels or an enhanced speech signal as derived from the array. Without loss of generality, this paper spatialized and transmitted channel one with the SRP-PHAT estimated speaker azimuth. Although not implemented in this paper for simplicity, further transmission bandwidth savings can be achieved by compressing the transmitted speech using any standard speech coding techniques e.g., AMR-WB [11], Speex [12].

B. S^{3}AC

Spatially Squeezed Surround Audio Coding (S^{3}AC) was originally proposed as an efficient compression technique for 5.1 multi-channel spatial audio coding [6]. The main goal in designing this technique is to achieve highly accurate localization of spatial sound objects. The core principle of S^{3}AC is to maintain the equivalence between an original large soundfield (360°) and a 'squeezed' soundfield in a psychoacoustic manner. To achieve this, S^{3}AC exploits a psychoacoustic phenomenon called 'localization blur', where human ears have limited resolution ability in precisely locating sound source [13]. Generally, to compress a 5.1 multi-channel signal, S^{3}AC applies an azimuth estimation algorithm based on inverse amplitude panning in the frequency domain; the resulting frequency domain virtual sound source is squeezed into a smaller soundfield, as illustrated in Fig. 2. Due to the limited localization resolution of human ears, the source localization resolution information saved in the squeezed soundfield is sufficient for recovering a full 360° soundfield without any perceptual localization distortion [6].

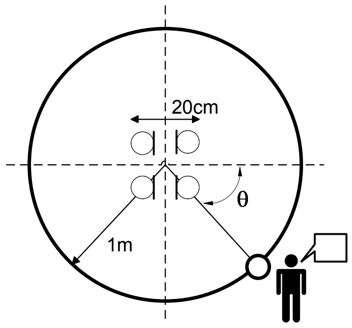


Fig. 4. Simulated recording setup for each meeting

For the teleconferencing application of this paper, the S^{3}AC technique is used to reproduce the 'squeezed' soundfield representing multiple remote teleconference sites. As illustrated in Fig. 3, two speakers at different sites may be located too close to be disambiguated if spatialized with the original speaker azimuths at a third site. To enhance discriminated speaker localization between different conference sites, soundfield information transmitted from each remote site containing full 360° localization information is squeezed into a unique sector for the user. This is achieved by applying a bijective azimuth mapping function, Θ_n , on the transmitted azimuth of each remote site:

$$A_n = \Theta_n(a_n) \tag{4}$$

where A_n and a_n are the squeezed and original azimuths from the n^{th} site, respectively, and the azimuth mapping function Θ_n is adaptively defined depending on the number of sites and number of participants per site to be spatially rendered. For example, while 'squeezed' sectors of equal widths are allocated to remote sites in Fig. 3, the azimuth mapping function can be modified such that remote sites with a large number of speakers can be assigned a larger sector for unambiguous rendering between speakers from this site. In this squeezing process, while speakers from different remote sites are displaced, the spatial relationship between speakers at each site remains intact.

The transmitted speech stream from each remote site is then rendered by the S^{3}AC amplitude panning process to the squeezed sector, using the two loudspeakers closest to each mapped azimuth. This processed is performed in the frequency domain, where time-frequency transform can be achieved by any modern filter e.g., STFT or QMF, by:

$$\begin{aligned} LS_1(t,k) &= S(t,k)\cdot\left[\tan(\eta)+\tan(A_n(t,k))\right] \\ LS_2(t,k) &= S(t,k)\cdot\left[\tan(\eta)-\tan(A_n(t,k))\right] \end{aligned} \tag{5}$$

where $LS_1(t,k)$ and $LS_2(t,k)$ are the two loudspeaker signals, $S(t,k)$ is the transmitted mono speech, η is the azimuth separation between the two loudspeakers, $A_n(t,k)$ is the mapped speech azimuth in the squeezed sector obtained by (4), and t and k are frame and frequency indexes, respectively. $LS_1(t,k)$ and $LS_2(t,k)$ are then transformed back to time-domain to form the loudspeaker feed signal.

III. Simulations

To illustrate the proposed teleconferencing system, simulations were conducted from the point of view of a teleconference with *N*-1 remote participating sites. That is, there are *N* teleconference sites in total: *N*-1 remote sites plus the user site spatializing the *N*-1 remote sites. Two simulation scenarios were thus conducted with this paradigm, from the point of view of Site 1 (as shown in Fig. 3): firstly, two remote sites of two participants each (*N*=3 as shown in Fig. 3a); secondly, with four remote sites, two from the first simulation scenario plus two more of three and four participants each (*N*=5, as shown in Fig. 3b).

Ground-truth speaker azimuths (as measured from the positive *x*-axis) are shown underneath each speaker in Fig. 3. Speakers at the four remote sites were placed at similar azimuths to maximally illustrate the advantage of 'squeezing' soundfields that would otherwise overlap if remote site soundfields were simply resynthesized using the original speaker azimuths.

Meeting recordings at each site were simulated using anechoically recorded speech; all sites spatialized speech to a meeting room of dimensions 3m×3m×3m. Reverberation times (RT60) from 0s (anechoic) to 0.5s were modeled using Allen & Berkeley's image method [14]. To record the 'meeting' speech at remote sites, each site modeled four omnidirectional microphones placed 20cm apart centred around the origin, with speakers located on the unit circle; this recording setup is shown in Fig. 4.

A total of eleven different speakers were thus required for the two simulated teleconferencing scenarios. Each teleconference site played out each speaker in turn, without any speaker overlap. Eleven anechoic speech sentences from different speakers, six female and five male, each approx. 5s in duration were sourced from the Australian National Database of Spoken Languages (ANDOSL) [15]. Speech sentences were normalized and downsampled from 20kHz to 16kHz, and stored at 16 bits/sample.

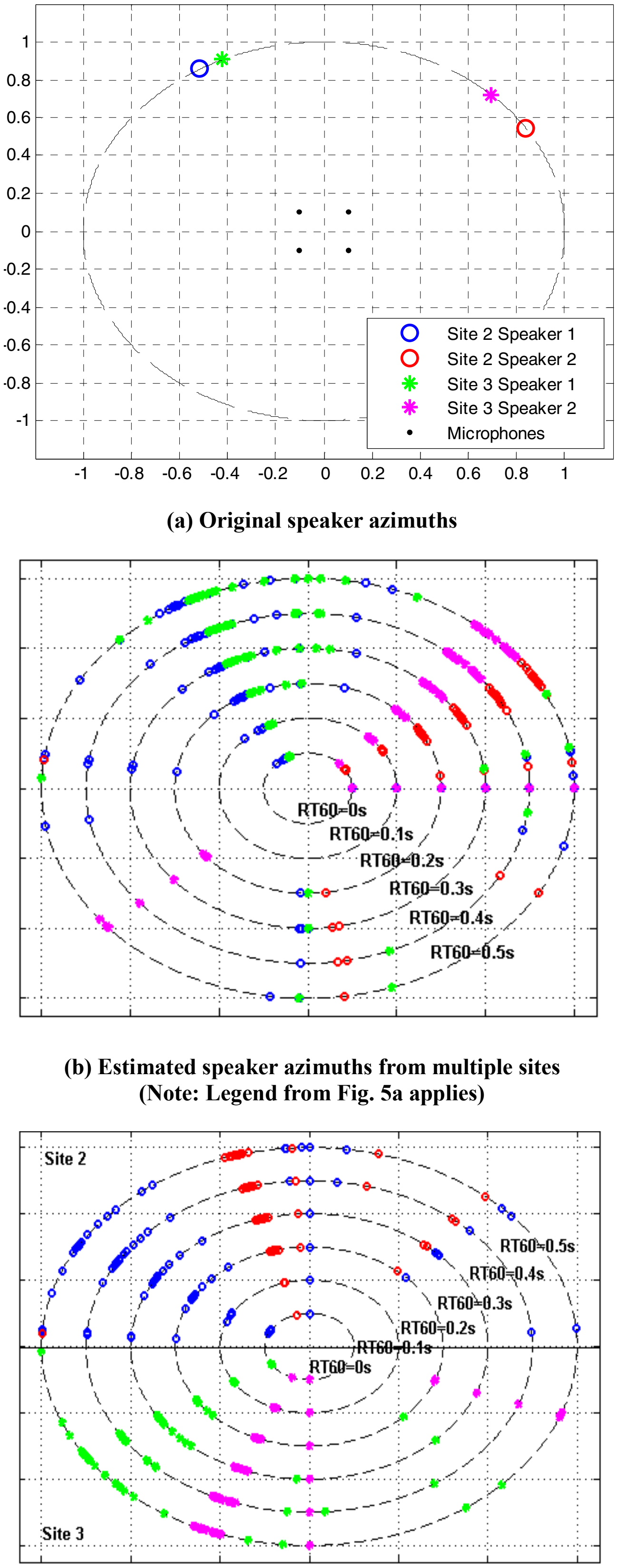
IV. Results

For both simulation scenarios (*N*=3 and *N*=5), SRP-PHAT analysis frames were chosen to be 32ms in length and Hamming-windowed with 50% overlap. Thus, an azimuth estimation is given and thus re-spatialized at the user's end every 16ms.

For each of the two simulation scenarios, results are presented as graphical plots of the speaker azimuths from all participating teleconference sites as estimated from SRP-PHAT (i.e., original azimuth) and after 'squeezing' into the user's soundfield (i.e., Site 1 in Fig. 3) for site and speaker disambiguation. To illustrate the effect of increasing reverberation time, the speaker azimuths are plotted in concentric circles of increasing reverberation time (RT60=0s to 0.5s in 0.1s increments) with increasing circle radius.

A. First simulation scenario (N=3)

Fig 5 shows the results obtained from spatializing two remote sites to a third site (see Fig. 3a). Fig 5a illustrates the



(a) Original speaker azimuths

(b) Estimated speaker azimuths from multiple sites (Note: Legend from Fig. 5a applies)

(c) 'Squeezed' speaker azimuths from multiple sites (Note: Legend from Fig. 5a applies)

Fig. 5. Simulation scenario 1 results

ground truth speaker azimuths for both remote sites, with the azimuths estimated from SRP-PHAT shown in Fig. 5b (note that the legend from Fig. 5a also applies to Figs. 5b and 5c). It can clearly be seen from Fig. 5b that simply respatializing the speakers to their original azimuths will cause spatial overlap for the user at Site 1, where the user will not be able to easily disambiguate between speakers 1 or 2 from either Site 2 or 3.

Fig. 5c shows the azimuths 'squeezed' by the approach proposed in this paper. Site 2 has been squeezed to the top half of the listening circle, whilst Site 3 is squeezed to the bottom half. The speakers within each site and between sites are clearly spatially separated, even in higher reverberation times where the azimuth estimations from SRP-PHAT exhibit greater variance due to the reverberant signal degradation.

B. *Second simulation scenario (N=5)*

The results of the first simulation scenario in Fig. 5 showed that the proposed squeezing approach can spatially disambiguate speakers from within a site as well as between sites; however, this was only a simple scenario with two remote sites with two participants each. This second simulation aims to explore the squeezing approach with more remote sites and more participants at a remote site.

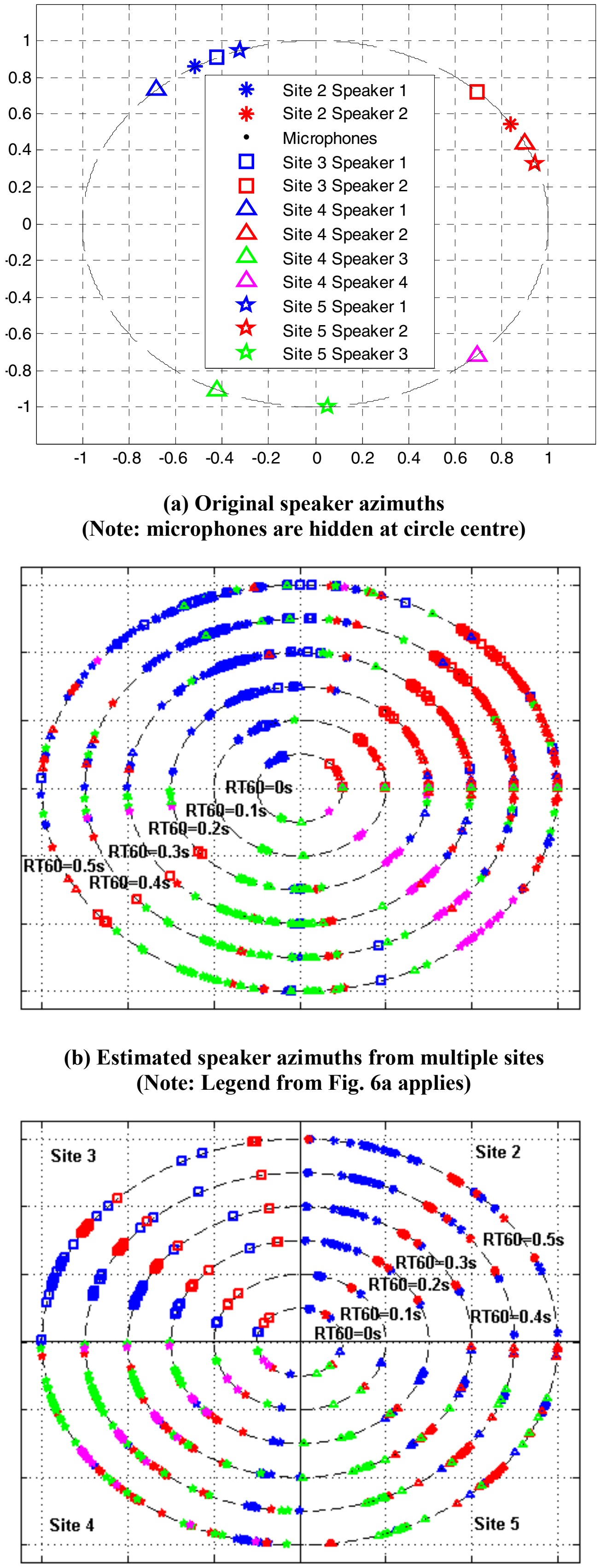
Fig. 6 exhibits the results obtained from the second simulation scenario with four remote sites of two to four participants (see Fig. 3b). Similar to Fig. 5a, Fig. 6a shows the ground truth speaker azimuths for all four sites; the legend in Fig. 6a also applies to Figs. 6b and 6c, and differentiates between remote sites with different plot point symbols whilst speakers at the same site are differentiated by colour. Fig. 6b shows the speaker azimuths for all remote sites as estimated by SRP-PHAT, and similar to Fig. 5b it can clearly be seen that with more participants the spatial separation of speakers between sites is ambiguous.

Fig. 6c thus shows the re-spatialized speaker azimuths as rendered by the squeezing approach proposed in this paper. The four remote sites were squeezed to:

- Site 2 (two participants): top right quadrant;
- Site 3 (two participants): top left quadrant;
- Site 4 (four participants): bottom left quadrant;
- Site 5 (three participants): bottom right quadrant.

The four quadrants of sites and speakers in Sites 2, 3, and 5 are clearly spatially separated, even with the greater variance in SRP-PHAT azimuth estimates at higher reverberation times. However, the four speakers of Site 4 in the bottom left quadrant are more ambiguously placed, owing to the larger number of speakers squeezed into the equally-sized site sectors.

A second spatialization result employing a different squeezing function is illustrated in Fig. 7, where the squeezed sector sizes are adjusted according to the number of speakers per site to be spatialized. Fig. 7 shows that allowing for smaller sectors for sites with fewer participants (Site 2, 3) does not ambiguously reduce speaker spatial separation within the site, whilst sites with more participants (Site 4) clearly benefit with greater spatial separation of its speakers.



(a) Original speaker azimuths (Note: microphones are hidden at circle centre)

(b) Estimated speaker azimuths from multiple sites (Note: Legend from Fig. 6a applies)

(c) 'Squeezed' speaker azimuths from multiple sites (Note: Legend from Fig. 6a applies)

Fig. 6. Simulation scenario 2 results

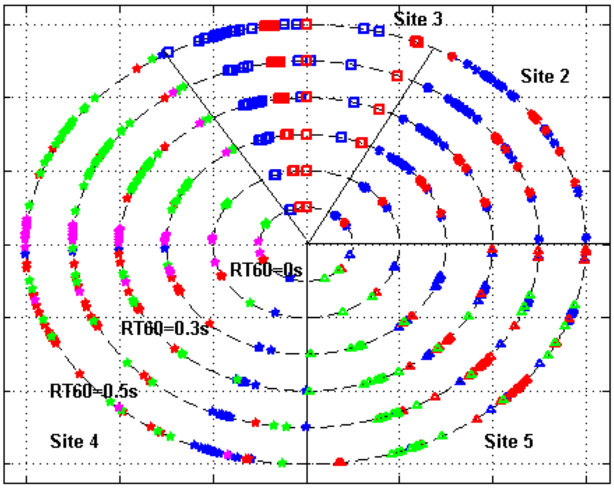


Fig. 7. Simulation scenario 2 with unequal ‘squeezed’ sectors

V. Conclusion

This paper proposed a teleconferencing system that ‘squeezes’ the original speech soundfields from multiple distributed remote sites to unambiguously spatially merge sites together for the user’s 5.1 surround playback. Simulation results presented show that the proposed squeezing approach spatially separates speakers of a remote site and between sites. However, remote sites with a greater number of participants can exhibit spatial overlap between speakers, thus ‘squeezed’ sectors that are sized according to the number of participants at each site achieve improved intra-site speaker spatial separation, whilst maintaining inter-site spatial disambiguation.

Currently, user listening tests are being conducted in addition to investigations into ‘squeezed’ rendering that can disambiguate multiple active talkers at the same remote site. The authors also intend to implement the proposed ‘squeezing’ approach for surround rendering over headphones, and compare the speaker and remote site spatial disambiguation of binaural versus 5.1 surround loudspeaker rendering.

References

[1] J. J. Baldis, “Effects of spatial audio on memory, comprehension, and preference during desktop conferences,” in proc. ACM SIGCHI Conference on Human factors in Computing Systems, pp.166-173, Washington, USA, March 2001.

[2] M. J. Evans, A. I. TEW, J. A. S. Angus, “Perceived performance of loudspeaker-spatialized speech for teleconferencing,” Journal of the Audio Engineering Society, vol. 48, no9, pp. 771-785, 2000.

[3] D. B. Ward, G. W. Elko, “Robust and adaptive spatialized audio for desktop conferencing,” Journal of the Acoustical Society of America, vol. 105, no. 2, p. 1099, Feb. 1999.

[4] R. Kilgore, M. Chignell, P. Smith, “Spatialized audioconferencing: what are the benefits?” in proc. 2003 IBM Conference of the Centre for Advanced Studies on Collaborative Research, pp. 135-144, Ontario, Canada, 2003.

[5] J. Ahonen, V. Pulkki, T. Lokki, “Teleconference Application and B-Format Microphone Array for Directional Audio Coding,” AES 30th Int. Conf: Intelligent Audio Environments, Finland, March 2007.

[6] B. Cheng, C. Ritz and I. Burnett, “A Spatial Squeezing Approach to Ambisonic Audio Compression”, in Proc. IEEE ICASSP 2008, Las Vegas, USA, Mar. 2008.

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays: Techniques and Applications*, M. Brandstein and D. Ward, Eds., Berlin: Springer-Verlag, 2001, pp. 157–180.

[8] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” IEEE Trans. Acoust., Speech, Signal Process., vol. 24, no. 4, pp. 320–327, Aug. 1976.

[9] A. Johansson, N. Grbic and S. Nordholm, "Speaker Localisation using the far-field SRP-PHAT in conference telephony," IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Taiwan, Nov. 2002.

[10] D. Hoang, H. Silverman, Y. Ying, “A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array,” in proc. ICASSP 2007, vol. 1, pp. I-121 - I-124, Hawaii, April 2007.

[11] 3GPP Technical Standard (TS) 26.171 (version 7.0.0 release 7), “Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description,” June 2007.

[12] Speex: A Free Codec For Free Speech [Online]
Available: http://www.speex.org

[13] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Cambridge: MIT Press, 1997.

[14] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” Journal of the Acoustical Society of America, vol. 65, no. 4, pp. 943-950, April 1979.

[15] ANDOSL: Australian National Database of Spoken Language [Online]
Available: http://andosl.anu.edu.au/andosl