

# Concept and Evaluation of a Downward-Compatible System for Spatial Teleconferencing using Automatic Speaker Clustering

Alexander Raake, Sascha Spors, Jens Ahrens, Jitendra Ajmera

Deutsche Telekom Laboratories, Berlin University of Technology, Berlin, Germany

{alexander.raake, sascha.spors, jens.ahrens, jitendra.ajmera}@telekom.de

## Abstract

In multi-party teleconferencing, the transport of separate speech streams to a particular user and the subsequent spatial rendering of the different streams enables a more efficient communication. A simple means of spatial presentation at client side is that of binaural rendering and headphone presentation. For downward-compatibility, e.g. when the transport mechanism does not support multiple parallel downlink streams, a system is proposed that combines an automatic speaker classification mechanism with a spatial rendering of the segregated streams. The combined system aims at a better separability of the speakers than conventional systems. The paper details the two basic components, namely automatic speaker classification, and binaural rendering. Based on a first evaluation of the approach, a proof of concept is provided, and directions for further improvement are discussed.

**Index Terms:** speech communication, audio systems, teleconferencing, clustering methods, Gaussian distributions

## 1. Introduction

We consider a multi-party communication scenario, where the communication between the participants is performed via telephone networks. Such a scenario is also known as teleconferencing. Problems that may arise with standard telephone equipment are loss of intelligibility, comfort and task efficiency compared to a natural multi-party communication. The problems of conventional teleconferences are mainly due to the loss of the natural spatial auditory cues and the reduced bandwidth. Spatial sound reproduction can improve intelligibility due to the cocktail-party effect [1], can increase quality already due to a natural wideband sound reproduction [2] and was shown to lead to an increased speaker recognition efficiency for simultaneous talkers [3]. Spatial reproduction of a conference call requires to transmit all voice streams of all participants to all local terminals. However, in most traditional conference call systems the streams are mixed together to one stream in the telephone network.

The basic idea of this paper is to apply an automatic speaker change detection and clustering algorithm to segregate the mixed voice signal into streams. The separated streams are then spatially distributed in a virtual auditory environment by an audio rendering system in the local terminal, recreating the auditory spatial cues of a natural communication situation. Figure 1 illustrates the proposed system.

Two methods for spatial sound reproduction can principally be differentiated: (1) recreation of the wave field within a limited listening area, and (2) recreation of the wave field at the listeners ears. Methods of type (1) use loudspeakers for reproduction, methods according to (2) typically headphones. In this paper,

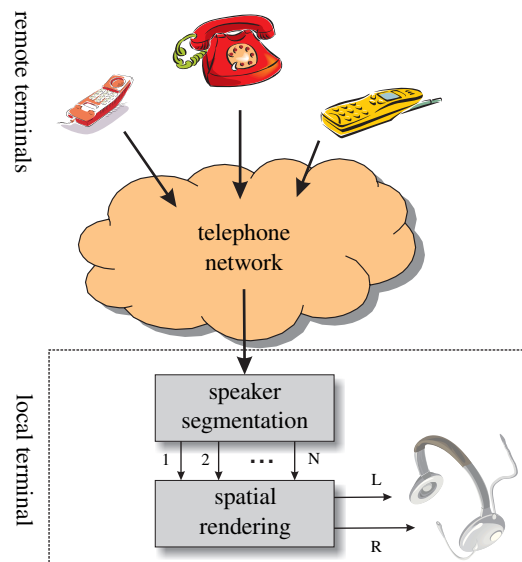


Figure 1: Multi-party communication scenario considered for this paper.

we focus on the second approach which is also known as *binaural reproduction*. However, the binaural rendering component can be exchanged by almost any other spatial rendering technique available.

The paper is organized as follows: Section 2 introduces the speaker clustering algorithm used for voice stream segregation, Sec. 3 presents the binaural sound reproduction system and Sec. 4 gives first evaluation results for the proposed combination of the two approaches.

## 2. Speaker change detection and clustering

We use the algorithm proposed in [4] to detect the speaker change points. Specifically, if we wish to find if there is a speaker change point at time  $t$ , two neighboring windows of relatively small size are considered (Figure 2). The contents of these windows are feature vectors extracted from the speech signal. In this work, 13 mel frequency cepstral coefficients (MFCC) [5] are extracted every 10 ms and used as feature vectors. In Figure 2, these sequences are denoted as  $X = \{x_1, x_2 \dots x_{N_x}\}$  and  $Y = \{y_1, y_2 \dots y_{N_y}\}$ , where  $N_x$  and  $N_y$  are the numbers of feature vectors in these two windows, respectively.  $Z$  represents the combination of these two sequences, with  $N_z = N_x + N_y$  denoting the total number of feature vectors. With this formulation of the problem, a speaker

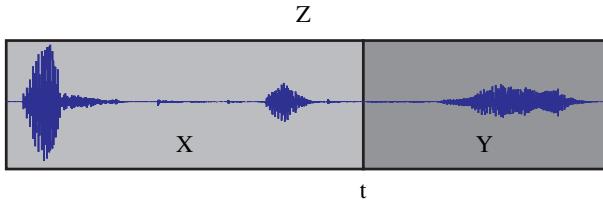


Figure 2: Two neighboring windows  $X$  and  $Y$  around time  $t$  to decide if there is a speaker change point or not.

change at time  $t$  is found if

$$\log p(X|\theta_x) + \log p(Y|\theta_y) > \log p(Z|\theta_z), \quad (1)$$

where  $\theta_x$  and  $\theta_y$  are parameters of single Gaussian densities estimated from  $X$  and  $Y$ , respectively.  $\theta_z$  are parameters of a Gaussian mixture model (GMM) with 2 Gaussian components, estimated from the data-set  $Z$ .

This search is performed for all time instants in the window shown in Figure 2. If more than one point satisfies the condition given by Eq. 1, the point maximizing the difference between the terms on the right and the left hand side of Eq. 1 is considered to be the speaker change point. If a change point is found in the window, a new window is initiated starting from the change point. If no change point is found in the entire window, the window size is increased by appending a few more feature vectors, and this process is repeated.

Once the speaker change points are found, the next step is to assign speaker labels to these speaker segments. This process is commonly referred to as speaker clustering. The algorithm used here for speaker clustering is similar to the algorithm proposed in [6], however, modified to run in an online fashion. If  $S_x$  and  $S_y$  are two speaker segments detected by the speaker change detection algorithm described above, they are compared to determine if they belong to the same speaker or not. To achieve this, we consider two GMMs with parameters  $\theta_x$  and  $\theta_y$ , estimated over the two segments. The number of Gaussian components in these GMMs,  $M_x$  and  $M_y$  are proportional to the length of these segments. In addition, another GMM with parameter set  $\theta$  is used to model the union of two segments  $S$ . The parameter set  $\theta$  is trained using data from both the segments and number of Gaussian components in this GMM.  $M$  is kept equal to the sum of the numbers of Gaussian components in two individual GMMs mentioned above, i.e.  $M = M_x + M_y$ . With these notations, two segments are considered to belong to different speakers if

$$\log p(S_x|\theta_x) + \log p(S_y|\theta_y) > \log p(S|\theta). \quad (2)$$

Eq. 2 is basically similar to Eq. 1. The important difference is that the segments considered for clustering are generally much larger compared to the windows (as shown in Figure 2) considered for speaker change detection. As the size (number of feature vectors) of the segments grows, more and more parameters are required to model the speaker characteristics. Therefore, Eq. 2 is based on GMMs (with the number of components proportional to the size of the segments) instead of single Gaussian densities. Each new segment resulting from the speaker change detection process is compared with all previous segments by using Eq. 2. If no match for this segment is found, a new speaker is hypothesized and a new speaker label is provided. The data for this new speaker is saved in a buffer. If a match from a previous speaker is found, then the union of the two data-sets  $S$  is

saved in a buffer for this particular matching speaker.

The computational complexity of the proposed algorithm allows it to run in real-time.

### 3. Binaural sound reproduction

Binaural sound reproduction techniques aim at recreating the wave field of a virtual acoustic scene at the entrances of the listeners ears. If optimally performed, the listener will have the impression of residing in the desired acoustic scene. The human auditory system is essentially based on analyzing the acoustic cues created by the scattering performed by the upper body and the head, and the acoustic properties of the pinna [7]. These cues depend mainly on the position of the listener in the virtual scene and the orientation of the listeners head with respect to his shoulders. However, due to inter-individual anatomical differences there is also a considerable inter-individual variation of these cues.

A straightforward realization of binaural sound reproduction is to place small probe microphones in the listeners ear canals or in an artificial (dummy) head, record the sound and reproduce it via headphones. This approach is also known as *dummy-head stereophony*. However, this simple approach is not very flexible since almost all degrees of freedom are fixed by the recording setup.

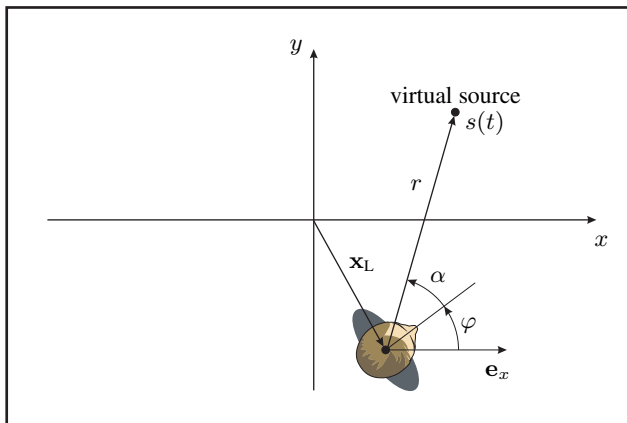
More flexibility can be reached by using sets of impulse responses. For the following discussion, the reproduction of one virtual point source for a listener residing at one fixed position is considered (Figure 3). The desired auditory cues are captured by the impulse responses corresponding to the acoustic transmission path from the virtual source position to the listeners ears. If captured in a reverberant environment, these impulse responses are often referred to as *binaural room impulse responses* (BRIR), if captured under free-field acoustic conditions, as *head-related impulse responses* (HRIR). Now, a flexible reproduction of (synthetic) virtual scenes can be achieved with a database of BRIRs captured for all desired listener positions and head orientations.

The binaural reproduction of a virtual source with headphones is performed by convolving the source signal  $s(t)$  with the appropriate BRIRs as follows

$$q_{L,R}(t) = h_{L,R}(\mathbf{x}_L, \varphi, \delta, \alpha, \beta, r, t) * s(t), \quad (3)$$

where  $\mathbf{x}_L$  denotes the position of the listener,  $\varphi$  and  $\delta$  the azimuth and elevation of the listeners head,  $\alpha$ ,  $\beta$  and  $r$  the azimuth, elevation and distance of the virtual source relative to the listener's head, and  $h_{L,R}(\cdot)$  the impulse responses from the virtual source position to the left/right ear of the listener, respectively. The geometric parameters are illustrated in Fig. 3. To select the appropriate BRIRs from the database requires information about the head orientation of the listener. When the orientation of the head or the virtual source position is changed, the current BRIRs need to be replaced.

We implemented a real-time PC-based binaural sound reproduction system using the principles outlined above. The system is based on a real-time convolution engine that is fed with the appropriate BRIRs derived from a user-defined database. The head-orientation is tracked by a commercially available orientation tracking system. The virtual scene is controlled by a graphical user interface that is operated via a touch screen. Sound reproduction is done via high quality headphones.



virtual room

Figure 3: Illustration of the coordinate system used for binaural sound reproduction. Only the horizontal plane ( $z = 0$ ) is shown.

## 4. Evaluation

In order to provide a proof of concept and to evaluate the system implementation we carried out an instrumental verification of the speaker segmentation algorithm as well as a first evaluation with human test subjects.

We utilized the VeriDat database [8] to compile the test items. The database is primarily intended to serve for speaker identification research in mobile networks. It contains speech from a large variety of female and male speakers. We used uttered German digits from this database to suppress contextual semantic cues in the speaker identification test. The Digit strings from various speakers were concatenated and long segments of silences were removed. We compiled five test-sequences: One test sequence with two speakers and two sequences for three and four speakers (40 s – 1 min duration). To emphasize the downward-compatibility to traditional telephony, the test sequences are sampled at 8 kHz.

### 4.1. Instrumental evaluation of speaker detection and clustering

There are, in total, 48 speaker changes in the five test samples. Many of these changes occur within 2 seconds. The speaker change point detection algorithm found all 48 of them correctly, while finding 12 false change points. The subsequent clustering algorithm discarded most of the false alarms. After clustering, there are 46 change points found with 2 false change points. The efficiency of the clustering algorithm is 88.20%, i.e. 88.20% of the speech frames were correctly associated to the speakers. The majority of the clustering error comes from errors when the found number of speakers is higher than the actual number of speakers. The number of speakers found for the speech files with 2, 3, 3, 4 and 4 speakers were 2, 3, 4, 5 and 5 respectively.

### 4.2. Auditory evaluation of proposed system

#### 4.2.1. Test setup and procedure

The number of speakers as well as the representation method were varied between the test items. The spatial presentation methods were (1) diotic (“mono”), and one binaural presentation each with (2) automatically segmented (“auto”) and (3) ide-

ally segmented (“ideal”) voice streams. The latter one serves as a reference. The speakers were arranged symmetrically based on the order of their first occurrence, using the angles  $\alpha = \{60^\circ, -60^\circ, 0^\circ, 30^\circ, -30^\circ\}$ ,  $\beta = 0^\circ$  relative to the listener (at a distance of  $r = 2$  m). With the three presentations methods this results in 15 test sequences (1·3+2·3+2·3). In order to avoid effects due to inaccurate head-tracking, we chose a static presentation for the experiments, and the subjects were instructed not to turn their head. In order to process the test items for the spatial representation, the segmented voice streams were convolved by BRIRs measured in a low-reverberant studio. Its early reflections support the localization [7] and lead to a more natural presentation than anechoic HRTFs. We used a AKG K240 DF headphone for the experiments. 16 native German subjects (7 female, 9 male) participated in the test. The test items were presented according to a  $15 \times 12$  diagram-balanced square design [9]. The three 2-speaker items were presented as training material in a randomized fashion at the beginning of each listening session, yielding the targeted 15 presentations. The test subjects had to report the speakers and the speaker change points via a graphical user interface operated on a touch screen. After each test item the subjects had to give two judgments using a slider: (1) “How did you perceive the audio reproduction?” and (2) “How complicated was the speaker assignment task?”. The extreme points of the sliders were labeled as “pleasant”, “unpleasant” (1) and “difficult”, “easy” (2). The position of the slider was internally mapped to a variable ranging from 0 to 100.

#### 4.2.2. Results

The results of the subjective evaluation of the proposed system have been verified by an analysis of variance (ANOVA) to prove the statistical significance of the (here used as fixed) factors “presentation mode” and “number of speakers”. Figure 4 shows the performance of the test subjects to detect the speaker changes in terms of correctly detected, substituted (detected speaker change but identified wrong speaker) and deleted (missed) speaker changes for the different representation methods and number of speakers. Figure 4(a) shows that for the three and four speaker case the spatial representation considerably increases the ability of the subjects to correctly detect the speaker changes. In both cases their performance is best for the ideal segmented spatial representation and worst for the “mono” representation. The performance of the subjects for the automatically segmented spatial representation is in between these. Figure 4(b) and 4(c) show similar tendencies for the insertion and deletion of speaker changes by the subjects. Figure 5 illustrates the results of the task difficulty ratings given by the subjects. As in the case of the speaker change detection, the ratings for the ideal segmented spatial representation are best, and worst for the mono representation. The ratings for the automatic segmented spatial representation are somewhat in between these two representation methods. Figure 6 gives the results of the pleasantness ratings given by the subjects. Here, the performance of the automatic segmented spatial representation for the three and four speaker case performs worst, followed by the mono representation and the ideal segmented spatial representation. According to the comments given by the test subjects, the misclassifications of the automatic speaker clustering have been perceived as very annoying, since they lead to changes in perceived location during ongoing utterances. These errors are thought to be the cause for reduced speaker identification performance in this case (“auto”).

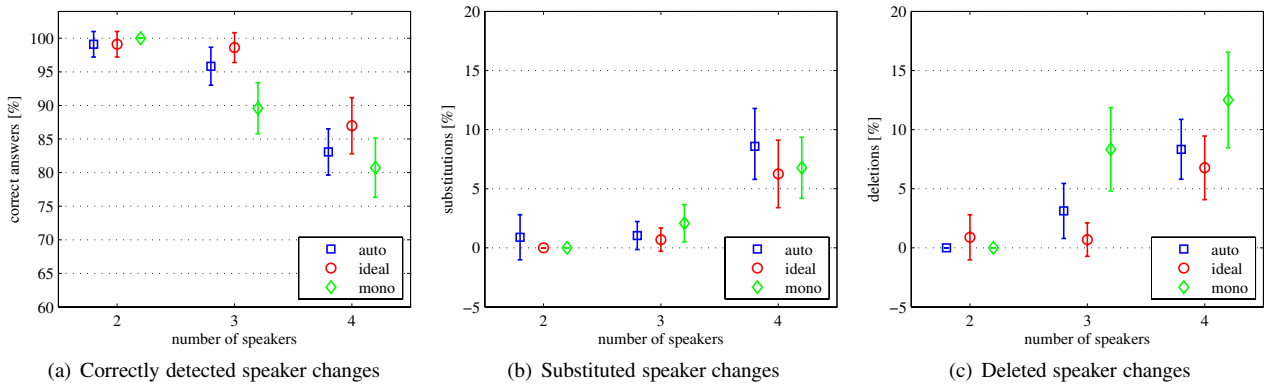


Figure 4: Performance of the test subjects to detect the speaker changes. The bars denote the 95% confidence intervals.

## 5. Conclusions

The proposed system for the spatial representation of multi-party teleconferences provides downward compatibility to traditional teleconference systems, where the voice streams of the different participants are mixed in the telephone network. The results of the first instrumental and auditory evaluation prove that spatial reproduction of multi-party communications highly alleviates teleconferencing applications. However, the results also show that the automatic speaker clustering in the local terminal has to be improved in order to reduce the impact of location changes of the spatially rendered speech signals. Future work will focus on improvements of the two components, and on other evaluation paradigms more realistically addressing conversations as in practical conferencing applications.

## 6. References

- [1] A. Bronkhorst, "The Cocktail Party phenomenon: A review of research on speech intelligibility in multi-talker conditions," *Acta Acustica utd w. Acustica*, vol. 86, pp. 117–128, 2000.
- [2] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*. Wiley, 2006.
- [3] B. Drullman and A. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am.*, vol. 107, no. 4, pp. 2224–2235, 2000.
- [4] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–652, 2004.
- [5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, Signal Processing*, pp. 357–366, 1980.
- [6] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," *IEEE Automatic Speech Recognition and Understanding workshop (ASRU)*, pp. 357–366, 2004.
- [7] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.
- [8] U. Turk and F. Schiel, "Speaker verification based on the German VeriDat database," in *Eurospeech*, 2003, pp. 3025–3028.
- [9] W. A. Wagenaar, "Note on the construction of digram-balanced Latin Squares," *Psychological Bulletin*, vol. 72, pp. 384–386, 1969.

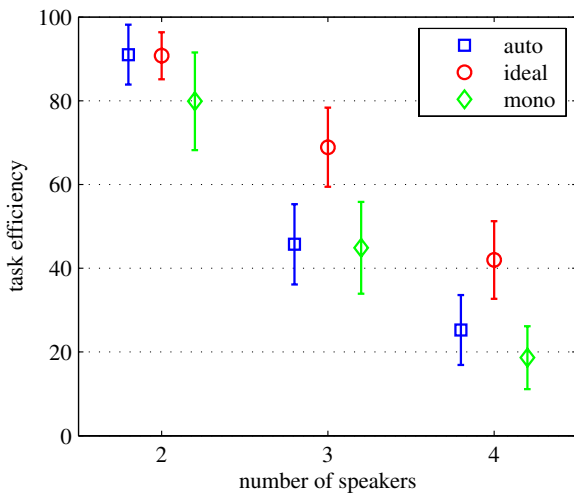


Figure 5: Task difficulty (bars: 95% confidence intervals).

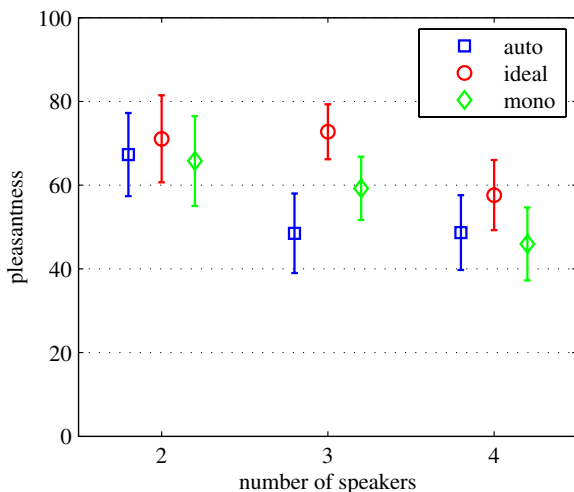


Figure 6: Pleasantness of the representation as perceived by the subjects (bars: 95% confidence intervals).