Effects of Binaural Synthesis on Speaker Recognition

Lars-Erik Riechert, Laura Fernández Gallardo and Dennis Guse

TU Berlin, Quality and Usability Lab, 10587 Berlin, Germany, E-Mail: lars-erik.riechert@alumni.tu-berlin.de

Abstract

For participants of multi-party teleconferences, it can be challenging to attribute what was said to the individual talkers. Spatial audio reproduction may help to overcome this issue. Our work investigates the potential improvement for speaker recognition using binaural synthesis. Here, the individual talkers are distributed to different simulated positions around the listener. In a listening-only test, we evaluated the effects of binaural synthesis and additional reverberation compared to a mono-diotic representation on speaker recognition. The results show a positive effect of the binaural synthesis compared to a mono-diotic condition. In addition, binaural synthesis with low levels of reverberation resulted in the highest speaker recognition performance compared to the anechoic settings. Nevertheless, a slight negative impact of the binaural synthesis was observed if all speakers appear from the same direction and thus no spatial information is provided.

Introduction

The question of *who said what* is of great importance for participants of teleconferences. How easily participants can answer this question depends on the composition of talkers and presumably the design of the teleconferencing system. In two user studies, we first identified a set of talkers that are not easily distinguished from each other. Second, we assessed the influence the system attributes of spatial representation and added reverberation have on speaker recognition.

Factors for speaker recognition

Various research has been performed under the topic termed cocktail party effect [7], which describes *speaker segregation* in a mixture of multiple talkers. For stream segregation the selective attention to a specific source is the key [8]. In a teleconference, however, information should be gathered from all interlocutors and correctly attributed. In the following, the term *speaker segregation* will be used in the simultaneous case in contrast to *speaker recognition* in the sequential case.

Speaker recognition is a problem that is solved on a multifactorial basis. Some of these factors, such as visual cues, are not available in teleconferences. This may lead to incorrect attribution of content to speakers. From the cues available in teleconferences, the individual pitch offsets and pitch contours are the major contributors for speaker segregation [9] and recognition [12,13,14]. The perceptual system even overemphasizes the importance of the pitch over other available cues in the spectrum (e.g., shimmer, jitter, or dispersions between certain formants [3]). However, the pitches of two voices are not stable and may cross at some point of time, which can lead to confusions [9].

The potential gain of reintroducing cues (as spatial information), is not easily predictable as recognition cues are weighted differently depending on the situation. For example, male and female voices are accessed differently [13,3] also cues are reweighted for every talker [12] and by every listener [14]. This reweighting may even observable on a cognitive level as different talkers are recognized on individual neural pathways [15].

It has been argued that monaural cues are more important than spatial cues for speaker segregation [16]. However, spatial cues provide advantages for two or more concurrent talkers [11, 18]. In a teleconferencing system with three talkers, spatial cues provided even a larger advantage than an increased signal bandwidth (which transports the spectral cues besides the pitch) for speaker recognition [2]. As speaker recognition is based on a set of cues, these may individually not provide much insight but interact with each other [19], i.e., if localization performance in a system is poor, speaker recognition can still be improved by spatial representation [20].

Effects of reverberation on speaker recognition

If sound is reflected in a room, information about its geometry as well as the source and receiver positions are encoded in the reverberations. This explains why perceived distance (externalization) depends on reverberations [21]. From this finding one may argue that reverberations may enhance a spatial teleconferencing system, as the lack of externalization would inhibit a vivid spatial perception of the talkers around the listener.

However, reverberation has a negative effect on spatial segregation of sounds [22], even if this effect was shown to be smaller for voice than for noises [18]. Additionally, spectral cues are diminished by reverberation [22].

Spatial teleconferencing systems

Spatial teleconferencing systems implemented with multiple physical loudspeakers were found to enhance speaker recognition performance [1], as well as measures of perceived quality [23]. For a system using binaural synthesis, this might depend by the specific implementation as well as the HRTFs. For example, a spatial teleconferencing system implemented using only changes in ITDs and ILDs (providing little to no externalization) was found to be preferred over mono although it did not improve the speaker recognition performance [24]. Findings made for physical setups should in general be transferable to simulated ones, as localization performance in the horizontal plane was shown to be similar for physical loudspeakers and binaural synthesis [25]. If a room specific set of HRTFs is used, binaural synthesis can even be perceptual transparent regarding localization performance [26]. Evidence shows that speaker segregation is not or only barely affected by the simulation [17,18]. However, detection performance of sound utterances can be reduced by the simulation [25].

If non-individualized HRTFs are used, localization performance in the horizontal plane is not affected [20]. Nevertheless, the perception of different heights, which is also determined by the spectral filtering of the pinna [27], is limited. Additionally, non-individualized HRTFs are prone to front-back confusions [28]. Head tracking is a potential solution while providing a potential gain of elevation without improving the azimuth accuracy [21]. This corresponds to the observation that speaker recognition rates in a binaural conferencing system are not altered by head tracking if all talkers are presented in the frontal hemisphere at equal height [2]. So far, it was not shown that speaker recognition can be increased if talkers are distributed at wider angles than required for localization, but a wider distribution may be preferred by users [1, 2]. Finally, the positions of the talkers should remain constant as a priori knowledge of the locations is important to segregate talkers [11].

Experiment 1: Similar-sounding Talkers

Certain combinations of speakers provide enough spectral and other cues to be easily distinguished from each other. As we wanted to focus on cases where monaural cues are not sufficient anymore, an experiment was conducted to determine a set of similar speakers.

Experimental Design

Subjects: 10 normal hearing participants were of age 20 to 41 (median age: 27), which were not familiar with the voices in the stimuli to be presented.

Stimuli: A database of 31 non-professional speakers (16 female, 15 male, all German) was used [29], featuring dialogues, 25 sentences and 52 single words. The content was identical for all speakers. The stimuli were edited semi-automatically to remove upfront respiratory noise and other artefacts. The first experiment employed 18 declarative sentences from 14 female and 5 male speakers.

Procedure: The experiment featured 105 pairwise comparisons of all 14 female speakers with each other (and themselves as control condition). For each trial two different sentences from two talkers were played back via headphones with a 2s gap between them. Participants were asked to assess similarity of the two speakers on a 7-point Likert with the endpoints “very similar” and “very dissimilar”. The order of speakers and sentences was randomized for every participant and each sentence was used at most once per talker. The study was preceded by a training of 15 comparisons of the 5 male speakers.

Results

The distance ratings were mapped to an interval from 0 (very similar) to 1 (very dissimilar). The mean value over all participants and speaker combinations was 0.56 (σ: 0.32). 2002 different groups of 5 talkers can be formed out of 14 talkers. The average distance of talkers within all these groups was calculated. The group with the lowest average distance (talkers: 5, 7, 12, 13, 14) featured a value of 0.38 (σ: 0.23), compared to a value of 0.77 (σ: 0.23) in the group with the highest average distance (talkers: 2, 4, 8, 9, 14). The distances of the 14 speakers span a 13-dimensional space, where each dimension resembles the distance to another speaker. To visualize this space and therefore reduce dimensionality non-metric *Multidimensional Scaling* (MDS) was used (stress = 0.15).

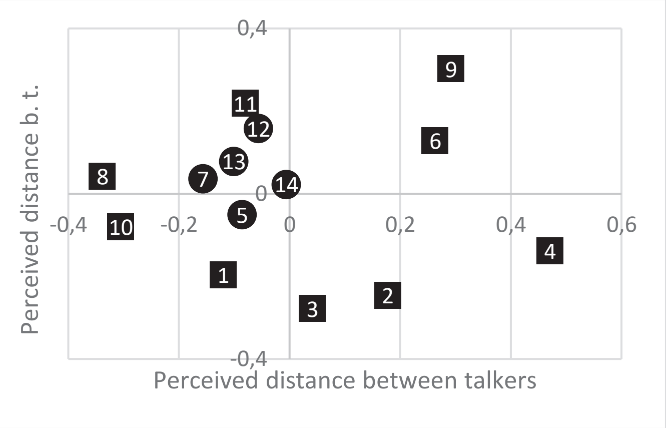

Figure 1: All 14 speakers fitted in a 2-D space using MDS; Circles represent the most similar group.

Experiment 2: Effects on Speaker Recognition

The influence of spatial *representation* and *reverberation* on speaker recognition performance was assessed in the 2nd experiment. Four similar sounding female voices were presented sequentially either as *spatial:* binaural synthesis with four virtual positions at ±12° & ±30°; *frontal:* binaural synthesis with identical positions at 0°; or using *diotic* representation. The frontal and spatial condition were presented in an *anechoic* and a *reverberant* condition.

Experimental design

Subjects: 42 normal hearing participants were of age 21 to 39 (median age: 29) and were not familiar with the voices they heard.

Stimuli were taken from the same database, as in the prior experiment, but only the 5 most similar female talkers were used. This time, all 25 sentences and 52 single words were utilized. For each participant, a random set of 4 from 5 speakers was chosen. Two sets of HRTFs that have been recorded with the same KEMAR 45BA artificial head were used. One has been recorded under anechoic conditions [5] and the other in a box shaped room featuring a reverberation time of 270ms [6].

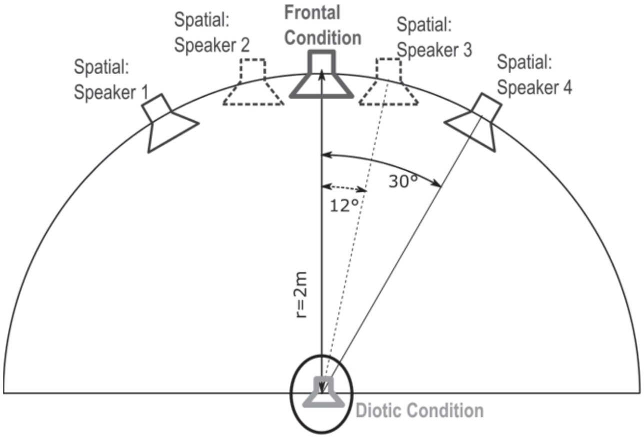

Figure 2: Physical positions of loudspeakers in the HRTF set used. Positions 1, 4, and the front were equipped with physical speakers. Positions 2 and 4 use the frontal speaker and a head movement of 12°.

Procedure: During the main phase of the experiment, a single stimulus was played back and participants had to decide who the talker was and press the corresponding button. Initially, all talkers were presented in a training (10min). During this training, pressing the buttons led to the playback from corresponding talkers. All conditions were trained for an equal length. Participants were allowed to take notes. In order not to confuse participants about spatial information being available or not, the experimental conditions were not randomized entirely but separated in 3 blocks for the 3 representation conditions. Additionally, an icon informed the participants about current experimental condition. The ordering of blocks remained identical for training as well as the main phase. The order was balanced over all participants. The target talkers and stimuli of both lengths were randomized within the blocks. Two randomized, but non-overlapping sets of stimuli were used for training and the main phase.

Results

For analysis, a *Generalized Linear Mixed Effect Model* (GLMM) with logit link was used. Results are given as odds (1) for the reference category and odd ratios (OR) for every independent variable (2). Probability of a correct answer can be calculated from odd ratios (3).

$$Odd(Factor_i) = \frac{Correct\ identifications|Factor_i}{Incorrect\ Identifications|Factor_i} \quad (1)$$

$$OR(Factor_i) = \frac{Odd(Factor_i)}{Odd(\neg Factor_i)} \quad (2)$$

$$P = \frac{Odd(Reference) * OR(Factor_i) \ldots * \ OR(Factor_n)}{1 - \ Odd(Reference) * OR(Factor_i) \ldots * OR(Factor_n)} \quad (3)$$

The results show a positive effect for the spatial representation and a smaller negative effect of the frontal representation compared to the diotic condition. A significant positive effect for added reverberation was found. If the stimulus was a sentence, the performance increased. Participants made fewer errors for the target speakers 2 & 3 and more errors for speakers 4 & 5 compared to speaker 1. If the target speaker was identical to the previous stimulus, significantly more errors occurred. Finally, there is a significant increase in performance from trial to trial. Two significant interaction effects were found: Speaker 2 was more often correctly recognized in the frontal condition and the increase of performance from trial to trial was more pronounced in the spatial condition.

Discussion

Binaural synthesis itself, as seen in the frontal condition, seems to compromise some mono-aural cues and therefore speaker recognition performance. However, this effect was found to be rather small compared to the gain if spatial information is provided. It was less difficult to recognize speakers under certain conditions (long stimuli, certain target speakers). However, the additional performance gain due to the availability of spatial cues was seemingly not influenced by the difficulty of the task. This finding suggests that the potential benefit of binaural synthesis is not limited to the case when target speakers are very similar.

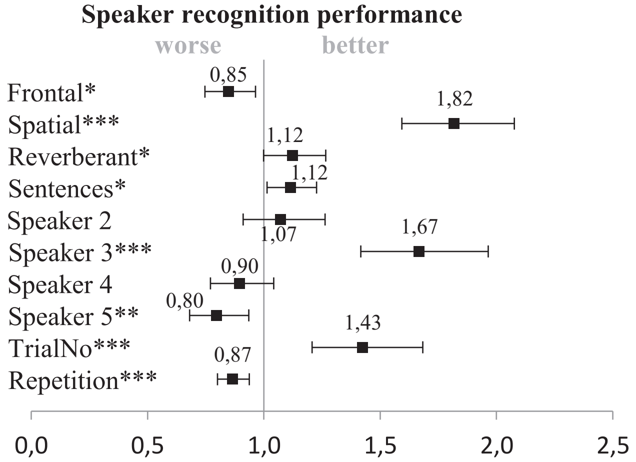

Figure 3: Odd ratios and confidence intervals of main effects for correct speaker recognition compared to the reference category (diotic + anechoic + single word + speaker 1 + first trial + no repetition) with an odd of 1.23 (55% correct answers); (*p<0.05, **p<0.01, and ***p<0.001).

The overall effect of reverberations is positive while relatively small. The negative effects of reverberation on spatial and spectral identifications cues seem not to outweigh the positive effect of improved externalization.

The higher number of incorrect identifications if the target speaker was repeated in the next trial may correspond to a bias of human decision making called repetition avoidance. People tend to underestimate the occurrence of repetitions in a series of random events [10].

Even if no feedback was provided to the listeners, their performance improved during the experiment, especially in the spatial condition. This may exemplify the ongoing refinement of the auditory system in selecting and weighting the cues which are best suited in the current situation. Also, spatial representation may be increasingly helpful in longer conversations.

Conclusion

A positive effect of binaural synthesis for speaker recognition performance was found, even if talkers are distributed very narrowly in the horizontal plane. In addition, binaural synthesis with low levels of reverberation resulted in the best speaker recognition performance compared to the anechoic settings. Nevertheless, a slight negative impact of the binaural synthesis was observed if no spatial information was provided. This indicates that the processing weakens some mono-aural cues. However, this negative effect is smaller than the potential gain for speaker recognition performance if talkers are virtually distributed around the listener

Outlook

The small, positive effect of added reverberation on speaker recognition may be increased if the properties of the room the listener is in are represented well. Additionally, only early reflections could be used, which have been shown to be sufficient for externalization [21]. Both these requirements could be met if room simulation is used. It remains to be investigated if room simulation provides substantial benefits in a spatial teleconferencing system.

References

[1] Baldis, J. (2001): Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: the SIGCHI conference, S. 166–173.

[2] Raake, A.; Schlegel, C.; Hoeldtke, K.; Geier, M.s; Ahrens, J. (2010): Listening and conversational quality of spatial audio conferencing. In: AES 40th international conference.

[3] Baumann, O.; Belin, P. (2010): Perceptual scaling of voice identity: common dimensions for different vowels and speakers. In: Psychological research 74 (1), S. 110–120.

[4] Skowronek, J.; Raake, A. (2015): Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio.

[5] Wierstorf, H., Geier, M., Raake, A., Spors, S. (2011) A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances, 130th AES Convention, eBrief 6.

[6] Wierstorf, H.(2016). Binaural room impulse responses of a 5.0 surround setup for different listening positions http://doi.org/10.5281/zenodo.160761

[7] Cherry, C. E. (1953): Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, v. 25, no. 5.

[8] Ebata, M. (2003): Spatial unmasking and attention related to the cocktail party problem. In: Acoustical Science and Technology 24 (5).

[9] Brokx, J. P. L.; Nooteboom, S. G. (1981): Intonation and the perceptual separation of simultaneons voices. In: Institute for Perception Research.

[10] Schilling, M. F. (1990): The longest run of heads. In: College Math. J 21 (3), S. 196–207.

[11] Brungart, D. S.; Ericson, M. A.; Simpson, B. D. (2002): Design considerations for improving the effectiveness of multitalker speech displays. In: Proceedings of the 2002 International Conference on Auditory Display.

[12] v. Dommelden, W. (1990): Acoustic parameters in human speaker recognition. In: Language and Speech 33.

[13] Murry, T. (1980): Multidimensional analysis of male and female voices. In: J. Acoust. Soc. Am. 68 (5), S. 1294.

[14] Kuwabara, H.; Ohgushi, K. (1984): Experiments on voice qualities of vowels in males and females and correlation with acoustic features. In: Language and Speech 27.

[15] Formisano, E.; Martino, F.; Bonte, M.; Goebel, R. (2008): "Who" is saying "what"? Brain-based decoding of human voice and speech. In: Science (New York, N.Y.) 322 (5903), S. 970–973.

[16] Ericson, M. A.; McKinley, R.L. (2001): The intelligibility of multiple talkers separated spatially in noise. Air Force research lab Wright-Patterson AFB OH human effectiveness directorate.

[17] Nelson, W. T.; Bolia, R. S.; Ericson, M. A.; McKinley, R. L. (1999): A Comparison of Free Field and Virtual Acoustic Environments. In: Proceedings of the human factors and ergonomics society 43rd annual meeting

[18] Hawley, M. L; Litovsky, R. Y.; Culling, J. F. (2004): The benefit of binaural hearing in a cocktail party: Effect of location and type of interferernce. In: The Journal of the Acoustical Society of America 115 (2), S. 833–843.

[19] Shinn-Cunningham, B.G. (2005): Influences of spatial cues on grouping and understanding sound.

[20] Drullman, R.; Bronkhorst, A. W. (2000): Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. In: J. Acoust. Soc. Am. 107.

[21] Begault, D.R.; Wenzel, E.M.; Anderson, M.R. (2001): Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. In: Journal of the Audio Engineering Society 49 (10), S. 904–916.

[22] Culling, J. F.; Hodder, K. I.; Toh, C. Y. (2003): Effects of reverberation on perceptual segregation of competing voices. In: J. Acoust. Soc. Am. 114 (5), S. 2871.

[23] Evans, M. J.; Tew, A. I.; Angus, J. A. S. (2000): Perceived Performance of Loudspeaker-Spatialized Speech for Teleconferencing. In: J. Audio Eng. Soc 48 (9), S. 771–785.

[24] Kilgore, R.; Chignell, M.; Smith, P. (2003): Spatialized Audioconferencing: What are the Benefits? In: Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative Research.

[25] Yost, W. A.; Dye, R. H.; Sheft, S. (1996): A simulated “cocktail party” with up to three sound sources. In: Perception & Psychophysics 58 (7), S. 1026–1036

[26] Wierstorf, H.; Spors, S.; Raake, A. (2010): Perception and evaluation of sound fields. In: 59th Open Seminar on Acoustics.

[27] Blauert, J. (1969/70): Sound Localization in the Median Plane. In: Acustica 22, S. 205–213.

[28] Wenzel, E. M. (1993): Localization using non-individualized head-related transfer functions. In: J. Acoust. Soc. Am. 94 (1), S. 111.

[29] Fernandez Gallardo, L. (2006): Recording a High-Quality German Speech Database for the Study of Speaker Personality and Likability. In: 12. Tagung Phonetik und Phonologie im deutschprachigen Raum, S. 43-46.