Automatic Mixing for Immersive Teleconferencing Systems

Christian Schörkhuber[1], Matthias Frank[1], Franz Zotter[1], Robert Höldrich[1], Peter Grosche[2]
[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz*
[2] *Huawei, European Research Centre, Munich*

Introduction

The aim of immersive teleconferencing is to convey a realistic sound field impression to a remote participant. To this end, the spatial distribution of talkers as well as room information needs to be captured by the near-end system and accurately reproduced on the far-end. As illustrated in Figure 1, we consider a setup where high speech quality is obtained by means of several close microphones (*spot microphones*) and spatial information is captured with a small circular or spherical microphone array in the centre of the acoustic scene. The proposed automatic mixing system robustly estimates the directions of multiple active talkers and mixes the close-microphone signals with the room information gathered by the central microphone array, whereas the spot microphones need not be synchronized with the microphone array and their positions are assumed to be unknown. Furthermore, we propose a novel automatic gain control method that keeps natural speech dynamics while equalizing speech level fluctuations due to unintentional changes of the talker-microphone distance. To allow for maximal flexibility concerning the reproduction system on the far-end (e.g. different loudspeaker setups or binaural reproduction for headphones), the sound field is encoded in higher-order Ambisonics. Listening experiments of our concluding evaluation indicate the optimal settings for recorded multi-talker scenarios using both headphone- and loudspeaker-based reproduction. With

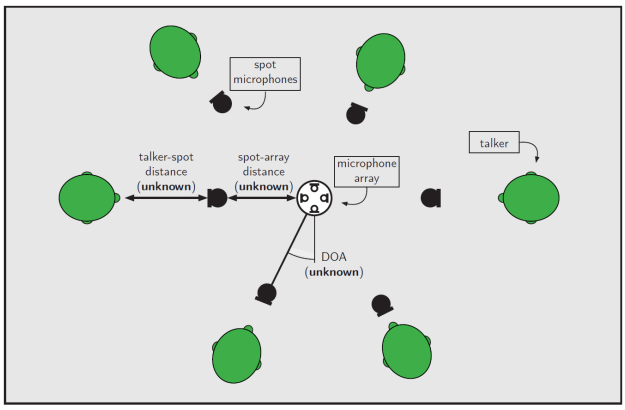

Figure 1: Illustration of the considered setup.

reference to Figure 2, the three main building blocks of the proposed system are the *parameter estimation stage*, the *mixing/encoding* stage and the *decoding/rendering stage*. The estimated parameters, i.e. the optimal gain of each spot microphone and the direction-of-arrival (DOA) of all active speakers provide information for the mixing and encoding stage on how to embed the spot microphone signals into the spatial recording of the central microphone array.

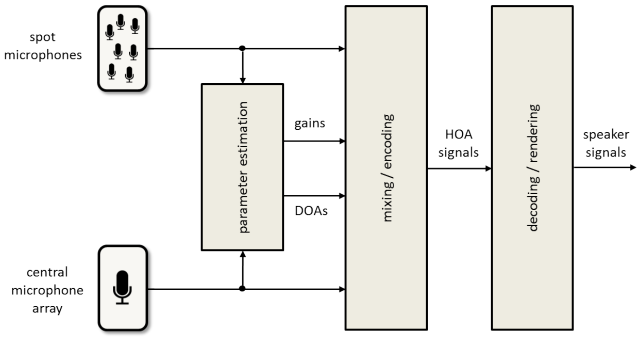

Figure 2: Main building blocks of the proposed method.

Mixing Parameter Estimation

As depicted in the block diagram in Figure 3, the mixing parameters g_i (*gains*) and φ_i (*DOAs*) for a spot microphone i are estimated on a block-by-block basis where each signal block is firstly transformed in the frequency domain, where $X_i(k, n)$ refers to the i^{th} spot microphone signal at frequency bin k and time frame n. Transform parameters, i.e. window size, hop size and transform size (zero padding) are not critical as the transformed signals need not be transformed back into the time domain. Furthermore, the parameter estimation stage does not introduce audio latency since the estimated mixing parameters are applied to the unprocessed audio stream. To further relax the requirements concerning the hardware setup, only the squared magnitude spectra of the input signals are utilized in the estimation process, such that the microphones do not need to be tightly synchronized, i.e. microphone signals of mobile devices could be used rather than pre-installed microphones. The close talker

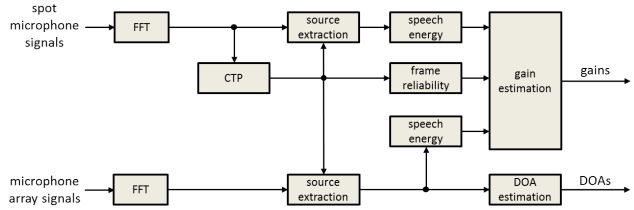

Figure 3: Block diagram of the parameter estimation stage.

probability (CTP), subsequently used as speech extraction filters, are constructed based on two assumptions. Firstly, we assume that on average one time-frequency slot of each microphone signal is mostly dominated by

a single talker (spectral disjointness)[1]. Secondly, we assume that the observed speech energy corresponding to talker i is highest for microphone i (i.e. the corresponding close microphone). From these assumptions we conclude, that a filter that extracts the speech of talker i can be derived from the short-time spectral power ratios between the signals of microphone i and all remaining spot microphones. We compute the close-talker probability, defined as the probability that a talker closest to spot microphone i is active in time frame n and frequency band k, as

$$P_i(n,k) = \frac{\gamma\left(|X_i(n,k)|^2 - \max_{j\neq i}\left(|X_j(n,k)|^2\right)\right)}{|X_i(n,k)|^2}, \quad (1)$$

with

$$\gamma(z) = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $0 \leq P_i(n,k) \leq 1$ can be used as a spectral mask, i.e.

$$Y_{ij}(n,k) = P_i(n,k)X_j(n,k), \quad (3)$$

such that $Y_{ij}(k)$ approximates the speech signal of talker i as picked up by microphone j.

Automatic Gain Control

Level variations in recorded speech signals stem from two independent causes, namely intentional variations by the talker and unintentional level fluctuations due to variations of the talker-microphone distance. More formally, the average speech energy of talker i in time frame n as recorded by microphone j is given by

$$s_{ij}(n) = m_j^s m_j^g e_i(n) a_{ij}(n), \quad (4)$$

where the sensitivity m_j^s and the initial gain m_j^g of microphone j can be assumed to be time invariant, whereas intentional variations of the emitted speech energy e_i (natural speech dynamics) and the attenuation $a_{ij} = f(d_{ij})$, which is a function of the euclidean distance d_{ij} between talker i and microphone j, causing unintentional level variations, are time-variant. Since our goal is to convey a realistic soundfield impression to the listener on the far-end, we strive to compensate for the unintentional level fluctuations while fully preserving natural speech dynamics. This is in contrast to traditional automatic gain control systems which do not differentiate between the two causes of level variations.To this end, we compute the instantaneous gain estimate as

$$\tilde{g}_i(n) = c\frac{\tilde{s}_{iQ}(n)}{\tilde{s}_{ii}(n)} = c\frac{m_Q^s m_Q^g e_i(n) a_{iQ}(n)}{m_i^s m_i^g e_i(n) a_{ii}(n)}, \quad (5)$$

where we refer to $\tilde{s}_{ii}$ and $\tilde{s}_{iQ}$ as estimations of the close and distant speech energy, respectively, Q is the index of the reference microphone, and c is a scaling factor. If the reference microphone is chosen such that $d_{iQ} \gg d_{ii}$, the contribution of unintentional level variations to s_{iQ} due to small changes of the position of talker i are negligible, hence

$$\tilde{g}_i(n) \approx \frac{cq}{a_{ii}(n)}, \quad (6)$$

with

$$q = a_{iQ}\frac{m_Q^s m_Q^g}{m_i^s m_i^g}, \quad (7)$$

being time-invariant. A reasonable choice for the reference microphone is some (real or virtual) microphone of the central array since it can be assumed that it is sufficiently far away from all talkers while exhibiting acceptable signal-to-noise ratios. We estimate the close and distant speech energies of talker i as picked up by microphone j with

$$\tilde{s}_{ij} = \mathbf{p}_i^T \operatorname{diag}(\mathbf{w})\mathbf{x}_j, \quad (8)$$

with the close-talker probability vector

$$\mathbf{p}_i = \left[P_i(1), P_i(2), \cdots, P_i(K)\right]^T, \quad (9)$$

the frequency weighting vector

$$\mathbf{w} = \left[w(1), w(2), \cdots, w(K)\right]^T, \quad (10)$$

and the squared magnitude vector

$$\mathbf{x}_j = \left[|X_j(1)|^2, |X_j(2)|^2, \cdots, |X_j(K)|^2\right]^T, \quad (11)$$

where $[\cdot]^T$ denotes the transpose of a vector or matrix, and K is the number of frequency bins. The weights $0 < w(k) < 1$ are used to de-emphasize less important frequency ranges, e.g. very low or very high frequencies. The instantaneous gain estimates $\tilde{g}_i$ are subsequently smoothed by means of a one-tap IIR filter, i.e.

$$\hat{g}_i(n) = \mu_i(n)\tilde{g}_i(n) + [1 - \mu_i(n)]\hat{g}_i(n-1), \quad (12)$$

where we refer to the time-varying coefficient $\mu_i(n)$ (adaptation step size) as the frame reliability corresponding to spot microphone i. The frame reliability is related to the probability that a talker i is active in time frame n and is derived from the close-talker probability vector as

$$\mu_i(n) = \frac{\mathbf{w}^T\mathbf{p}_i(n)}{\mathbf{w}^T\mathbf{w}} \leq 1. \quad (13)$$

Close-Talker Activity Detection

Since the gain of spot microphones corresponding to inactive talkers should be set to a low value, we implement a simple close-talker activity detection scheme based on the frame reliability $\mu_i(n)$. We compute the binary close talker activity

$$A_i(n) = f_{ST}\left(\tilde{A}_i(n), \tilde{A}_i(n-1)\right) \quad (14)$$

$$\tilde{A}_i(n) = h_{LP} \star f_G\left(\mu_i(n)\right), \quad (15)$$

where

$$f_{ST}(x_1, x_2) = \begin{cases} 1 & \text{if } ((x_1 \geq \xi_u) \wedge (x_1 > x_2)) \\ & \vee((x_1 < \xi_l) \wedge (x_1 \leq x_2)) \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

implements a Schmitt trigger with lower and upper thresholds ξ_l and ξ_u, respectively,

$$f_G(x) = \begin{cases} x & \text{if } x \geq \xi_g \\ 0 & \text{otherwise,} \end{cases} \quad (17)$$

implements a gating function with threshold ξ_g, h_{LP} is a simple low-pass filter and $\star$ denotes linear convolution.

Final Gain Computation

If the close talker activity $A_i(n)$ for microphone i is 0, $\hat{g}_i(n)$ is set to a low, predefined level G_{inactive}. The resulting gain is again low-pass filtered to avoid abrupt gain changes in the final gain estimate $g_i(n)$. In Figure 4 the block diagram of the entire gain estimation stage is depicted.

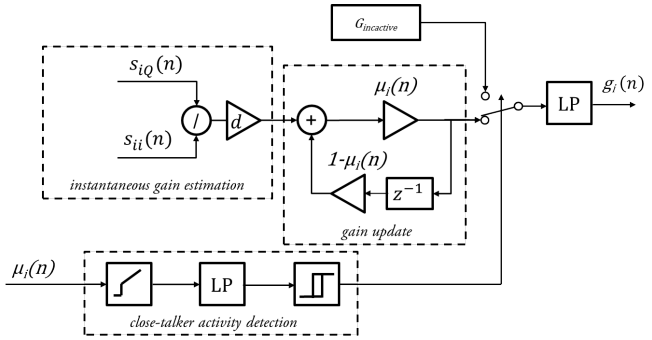

Figure 4: Overall block diagram of the automatic gain control stage.

Direction-of-Arrival Estimation

Estimating the direction-of-arrivals of multiple talkers with a small microphone array in reverberant conditions is a challenging task, even more so when the number of talkers is unknown. However, we reduce the problem of estimating DOAs of multiple concurrently active talkers to multiple single source DOA estimations by sequentially extracting the speech signal of talker i as observed by the microphones of the central array using the time-frequency masks defined in Equation (1). Since the resulting microphone signals approximately contain the speech signal of only one talker, any single-source DOA estimation method could be used. As we used first-order spherical microphone arrays in our experiments, we used a simple pseudo-intensity vector based approach[2].

Estimation of Time Delays

To avoid comb-filtering effects when the spot microphone signals are mixed with the array signals, the time delays between all spot microphones and the central array should be estimated and equalized. However, we found that for typical array-spot distances these effects are not audible since the direct signals picked up by the array act as early reflections in the mixture. Therefore, we omitted time-alignment in our experiments, however, due to the source extraction step, time-delay estimation is straightforward and easily implemented with some correlation-based method[3].

Mixing and Encoding

The DOAs detected for each spot microphone are employed to represent its signal in the corresponding playback direction using an Ambisonic encoder. Encoding to the Ambisonic representation allows to playback the spatialized spot microphone with relatively high flexibility concerning the playback facilities: it could either be a horizontal (surround, stereo, or more general setups), spherical loudspeaker arrangement, or headphone playback. For horizontal playback, encoding feeds each gain-weighted spot microphone signal $g_i(t)\,x_i(t)$ on a bus of $2\mathrm{N}+1$ Ambisonic signals using the circular harmonics $\boldsymbol{y}_{\mathrm{N}}(\varphi)$ evaluated at the DOA $\varphi_i(n)$ detected for the i^{th} spot microphone[4]:

$$\boldsymbol{\chi}_{\mathrm{N}}^{\mathrm{spot}}(t) = \sum_{i=1}^{I} \boldsymbol{y}_{\mathrm{N}}(\varphi_i(t))\, g_i(t)\, x_i(t), \tag{18}$$

where N is the order of the circular harmonic decomposition and we used $\mathrm{N}=3$ in our experiments. The parameters $\varphi_i(t)$, $g_i(t)$ are obtained from the respective frame-wise estimation using simple value repetition. Similarly, the raw signals of the microphone array channels are encoded in first-order Ambisonic signals and zero-padded to match the size of the encoded spot microphones, i.e.

$$\boldsymbol{\chi}_{\mathrm{N}}^{\mathrm{arr}}(t) = \big[0,\,\ldots,\,0,\,\chi_{-1}^{\mathrm{arr}}(t),\,\chi_{0}^{\mathrm{arr}}(t),\,\chi_{+1}^{\mathrm{arr}}(t),\,0,\,\ldots,\,0\big]^{\mathrm{T}}, \tag{19}$$

where $\chi_{0}^{\mathrm{arr}}(t)$, $\chi_{-1}^{\mathrm{arr}}(t)$, $\chi_{+1}^{\mathrm{arr}}(t)$ are the omnidirectional (W) and the two orthogonal figure-of-eight (Y,X) components of the 2-dimensional B-format signal, respectively. The output signal of the mixing/encoding stage is obtained by

$$\boldsymbol{\chi}_{\mathrm{N}}(t) = G_{\mathrm{mix}} \cdot \boldsymbol{\chi}_{\mathrm{N}}^{\mathrm{arr}}(t) + (1 - G_{\mathrm{mix}}) \cdot \boldsymbol{\chi}_{\mathrm{N}}^{\mathrm{spot}}(t), \tag{20}$$

where $G_{\mathrm{mix}} \in (0,1)$ defines the balance between the spot microphone (direct signals) and the microphone array (ambient signal). The choice of G_{mix} defines the trade-off between maximal intelligibility ($G_{\mathrm{mix}}=0$) and naturalness ($G_{\mathrm{mix}}=1$) of the presented sound field. The preferred values for G_{mix} using different playback setups have been determined by means of listening experiments.

Decoding and Rendering

On the far-end, the signals $\boldsymbol{\chi}_{\mathrm{N}}(t)$ of the Ambisonic bus are fed to a decoder that takes into account the $\max_{\boldsymbol{r}_{\mathrm{E}}}$ weights[5] and the directions of the loudspeakers or HRIR/HRTF dataset $\{\phi_l\}$. For a regular horizontal set of loudspeaker/HRIR directions, decoding is achieved by the transpose of the matrix

$$\boldsymbol{Y}_{\mathrm{N}} = \big[\boldsymbol{y}_{\mathrm{N}}(\phi_1),\,\ldots,\,\boldsymbol{y}_{\mathrm{N}}(\phi_{\mathrm{L}})\big] \tag{21}$$

that encodes the set of L directions.

The decoded signals $\boldsymbol{s}$ driving either the loudspeakers or the HRIR convolver are obtained by the decoding equation:

$$\boldsymbol{s}(t) = \boldsymbol{Y}_{\mathrm{N}}^{\mathrm{T}}\, \boldsymbol{\chi}_{\mathrm{N}}(t). \tag{22}$$

Evaluation Experiment

To compare different settings for the mixing parameter G_{mix} for both surround loudspeaker and headphone based playback, we conducted a listening experiment using a set of real conference recordings including single

talk, double talk and triple talk scenarios with both static and moving talkers. The data has been recorded at the IEM CUBE, a $11 \times 10 \times 5.5$m room with about $T_{60} \approx 0.75$s, and the entire recording setup comprises of 7 table-mounted spot microphones with cardioid directivity placed on a semi-circle (radius = 3 m) around a 4-channel tetrahedral microphone array (Soundfield ST-450 MK2), and one head-mounted microphone. Listeners were asked for a *preference* rating considering the *listening effort, speech intelligibility, and naturalness* of the different conditions. The MUSHRA-like GUI offered a sorting function to ascendingly sort the conditions according the individually entered rating to facilitate consistent ranking. There was a set of 8 trials in total, emerging from the combinations of three possible binary settings: (a) 2 conference scenarios (static and moving talkers), (b) playback with and without video, as well as (c) loudspeaker or headphone playback. The 8 conditions of each of the eight trials comprised playback of a mono version of the main microphone signal, a mono mix of the spot microphone signals (with automatic gain control), and a mixture of automatically panned and levelled spot microphones with the surround version of the main microphone using different mixing rations.

Results and Discussion

Figure 5 provides an overview of the preference ratings for the different mixing conditions using both headphone- and loudspeaker-based playback, involving preference of both scenarios. The mono mix of the Ambisonic main microphone is always least preferred, followed by the mono mix of the spot microphones. The reason for the better performance lies in the fact that the spot microphone signals are less reverberant and therefore better to understand when mixed to mono. Conditions 3 to 7 involve the main Ambisonics microphone including its first-order directional resolution and the automatically panned and gain-compensated spot microphones in a mixing level difference of $-\infty$ dB (Main only), -6 dB, 0 dB, 6 dB, ∞ dB (Spots only). The third condition playing back main Ambisonics microphone including its first-order directional resolution is superior to the mono mix of spot microphones, despite it is more reverberant. The spatial resolution appears to resolve some of the difficulty in understanding of the talkers of the multi-talk scenario. The automatically mixed and panned spot microphone signals could be assumed to be the best when it comes to speech intelligibility and spatial focus, as they are the least reverberant and carry the cleanest talker signals, and their high directional resolution is due to the above-described automatic directional mixing approach. Nevertheless, they are significantly less preferred with headphone playback due to the unnatural acoustic situation.

Conclusion

We proposed a method to automatically embed unsynchronized spot microphone signals into the spatial recording of a microphone array for multi-talker conference scenarios with unknown positions of both talkers and spot microphones. The mixing parameters were estimated from the separated speech signals using a soft time-frequency mask derived from energy ratios among spot microphones. Furthermore, a novel automatic gain control method has been proposed that compensates for level variations due to talker movements but keeps the natural speech dynamics unaltered.

Parts of the proposed method have been filed for patent.

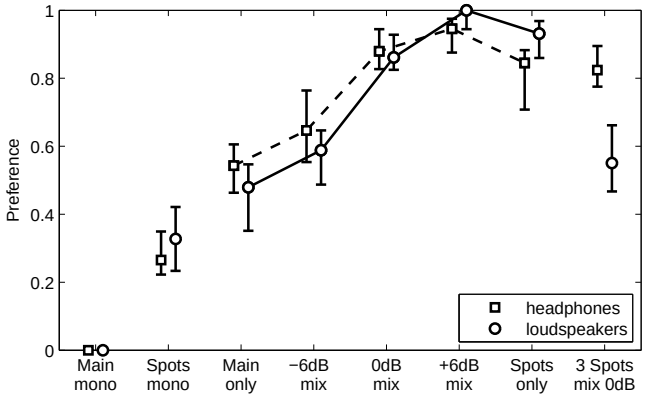

Figure 5: Median values and corresponding 95% confidence intervals of preference ratings for all conditions and headphone/loudspeaker playback averaged over all scenes and visual/non-visual presentation.

References

[1] Ö. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *Signal Processing, IEEE Transactions on*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, pp. 503–516, 2007.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[4] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[5] J. Daniel and S. Moreau, "Further study of sound field coding with higher order ambisonics," in *Audio Engineering Society Convention 116*, Audio Engineering Society, 2004.