DYNAMIC BINAURAL CUE ADAPTATION

Sebastian Nagel, Peter Jax

Institute of Communication Systems (IKS), RWTH Aachen University, Germany

{ nagel, jax } @ iks.rwth-aachen.de

ABSTRACT

Binaural recording using artificial heads is a common method to record acoustic scenes spatially. A drawback of this method has been its inability to incorporate head movements by listeners during playback, with detrimental effects such as in-head localization. Adapting recorded acoustic scenes to head movements has so far required different recording techniques involving microphone arrays and more than two channels. This contribution presents a new approach to adapt conventional two-channel binaural recordings to head movements by analyzing and modifying the cues contained in the binaural signal. According to a listening test, the proposed algorithm works with signals recorded in reverberant single-source as well as multi-source scenes. Therefore, it opens up the possibility to improve localization, externalization and realism of traditional binaural recordings.

Index Terms— Spatial Audio, Binaural Recording, Headphone Reproduction, Head Tracking, Artificial Heads

1. INTRODUCTION

Binaural recordings made with artificial heads or in-ear microphones and played back with headphones are a well-established technique to capture and recreate many of the spatial cues perceived by humans in an acoustic scene [1, 2]. However, with normal headphone listening, signals do not change with head movements, so that sound sources in the acoustic scene seem to rotate with the listener’s head. Several studies have highlighted the detrimental effects of this, such as limited source localization [3, 4, 5, 6] and lack of externalization, i.e. the perception of sources as inside the head [7, 8, 9].

To improve externalization, localization and realism, head movements can be tracked with a head tracking device. Then, headphone signals adapted to head movements can be generated so that sources appear fixed in space as in reality. For artificially generated scenes, this can be achieved, e.g., with dynamic binaural synthesis using Head-Related Transfer Functions (HRTFs) and anechoic source signals [10, 11]. For real scenes, different recording techniques have been proposed. Generally, microphone arrays with more than two microphones are used to sample the sound field during recording. Approaches then range from head orientation-dependent interpolation between the signals of different microphones [12, 13] over beamforming with beamformer characteristics that resemble HRTFs [14] to binaural synthesis of signals processed in the spherical harmonics-domain [15, 16, 17, 18, 19]. In contrast to these methods, in this paper a new approach is proposed that can dynamically adapt static binaural recordings made with an artificial head without additional microphones to head movements.

In the new approach, the cues contained in the two binaural microphone signals are analyzed to estimate the perceived source directions in the original scene. Then, the cues are modified to adapt the source directions to the listener’s head orientation. The method presented in this paper can rotate a scene in a contrary direction to the rotation of a listener’s head, so that sources appear fixed with regard to the environment. It is computationally simple and can be applied in real-time, yet according to a listening test, it works with reverberant conditions and single-source as well as multi-source scenes. Therefore, it opens up the possibility to improve localization, externalization and realism of traditional binaural recordings.

The paper is structured as follows. In Sec. 2 we introduce the binaural signal model and assumptions underlying the proposed method. In Sec. 3 the system and its elements are presented in detail, followed by an evaluation with a listening test in Sec. 4.

2. BINAURAL HEARING AND SIGNAL MODEL

The ability of the human auditory system to localize sound sources has been subject of extensive research [20]. According to the Duplex theory, humans evaluate Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) to estimate the lateral sound source direction. At low frequencies, ILDs are small while ITDs dominate. At high frequencies, ITDs become ambiguous for source localization and ILDs become the dominant cues. Although evidence exists that Interaural Time Differences (ITDs) determined from signal envelopes can also be relevant at high frequencies [21], in this contribution only ITDs and ILDs are regarded for simplicity.

These cues are often ambiguous, with different source position on a so-called Cone of Confusion leading to the same ITDs and ILDs. The human auditory system uses additional cues, e.g. obtained from head movements, to determine source elevations or to distinguish between front and back [10]. In this contribution, we avoid ambiguities of binaural cues by considering all sources as in front of the listener and on the horizontal plane. This constraint is reasonable for many realistic applications, such as teleconferencing or concert recording.

Using the frequency analysis described in Sec. 3.4 for a single source in a frequency bin at a time instant, the binaural ear signals and right ears with can be described using a source signal as

\[ X_l(\lambda, \mu) = H_l(\lambda, \mu) \cdot S(\lambda, \mu) \]  

(1)

\[ X_r(\lambda, \mu) = H_r(\lambda, \mu) \cdot S(\lambda, \mu) \]  

(2)

where \( H_i(\lambda, \mu) \) with \( i \in \{1, r\} \) represent HRTFs for the left and right ear, which can be divided into magnitude and phase components as

\[ H_i(\lambda, \mu) = |H_i(\lambda, \mu)| \cdot e^{j\phi_i(\lambda, \mu)}. \]  

(3)

The HRTFs describe the modification of the signal depending on the direction of sound incidence. According to our previous assumptions, the auditory system evaluates them in terms of a relative transfer function

\[ \frac{X_l(\lambda, \mu)}{X_l(\lambda, \mu)} = H_l(\lambda, \mu) / H_l(\lambda, \mu). \]  

With this, the...
binural cues perceived with the source signal $S(\lambda, \mu)$ can be expressed using \(5\) as

$$ILD(\lambda, \mu) = \left| \frac{X_r(\lambda, \mu)}{X_l(\lambda, \mu)} \right| - \left| \frac{H_r(\lambda, \mu)}{H_l(\lambda, \mu)} \right|, \tag{4}$$

$$IPD(\lambda, \mu) = \arg \left( \frac{X_r(\lambda, \mu)}{X_l(\lambda, \mu)} \right) - \phi_l(\lambda, \mu) - \phi_l(\lambda, \mu). \tag{5}$$

The relationship between the perceived cues and sound source directions can be described as a codebook, which can be obtained from a database of HRTFs $\hat{H}_l(\mu, \varphi)$ for direction azimuths $\varphi$. In the codebook, the cues corresponding to the direction $\varphi$ are given by

$$\hat{ILD}(\mu, \varphi) = \left| \frac{\hat{H}_r(\mu, \varphi)}{\hat{H}_l(\mu, \varphi)} \right|,$$

$$\hat{IPD}(\mu, \varphi) = \arg \left( \frac{\hat{H}_r(\mu, \varphi)}{\hat{H}_l(\mu, \varphi)} \right) - \hat{\phi}_l(\mu, \varphi) - \hat{\phi}_l(\mu, \varphi). \tag{7}$$

Trying to model the process by which the auditory system selects alternatives to obtain the codebook: an analytic spherical head model [23] as used in [22] and measured HRTFs from [24] of the artificial head used for the recording of the evaluation signals.

3. PROPOSED SYSTEM ARCHITECTURE

The aim of the proposed system is to keep the apparent source positions in the room stable when the listeners rotate their head. To achieve this, the perceived source directions $\varphi_{\text{orig}}(\lambda, \mu)$ need to be adapted to the head movements as illustrated in Fig. 1 if the listener rotates the head by an angle $\Delta \varphi$, a source that is perceived at $\varphi_{\text{orig}}$ in the unmodified recording should be perceived at $\varphi_{\text{dest}} = \varphi_{\text{orig}} - \Delta \varphi$ in the adapted signal. This corresponds to the operation of rotating the acoustic scene by $\Delta \varphi$ in the opposite direction of the head rotation.

A block diagram of the proposed system is given in Fig. 2. A cue-to-direction codebook is used for estimating the perceived source directions in the original scene $\varphi_{\text{orig}}(\lambda, \mu)$ and determining the cue modifications required to rotate them to $\varphi_{\text{dest}}(\lambda, \mu)$. When the concept is applied, a headtracker can be used to determine the head rotation $\Delta \varphi$ relative to a reference orientation. Two possibilities to obtain the cue-to-direction codebook are introduced in Sec. 3.2. The source direction estimation and cue modification blocks are presented in Sec. 3.3 and the frequency domain filtering framework used to apply the modifications is described in Sec. 3.4.

![Fig. 1: Adapting source directions to head rotations.](image)

3.1. Cue-to-Direction Codebooks

The cue-to-direction codebook is obtained by inserting HRTFs for different source directions into \(6\) and \(7\). As one alternative, we use measured HRTFs of a Neumann KU100 artificial head, measured in the horizontal plane in increments of 1°, from [24].

As another alternative, an analytic model of the head as a sphere [23] is used for discrete angles with increments of 1°. The HRTFs in this model are given by

$$\hat{H}_l(\mu, \varphi) = \frac{1 + j \frac{2 \pi f (\mu)}{2 \beta c}}{1 + j \frac{2 \pi f (\mu)}{2 \beta c}} e^{-j 2 \pi f (\mu) \tau(\varphi)}, \tag{8}$$

where $\theta_i$ with $-180^\circ \leq \theta < 180^\circ$ is the angle of sound incidence relative to the ear $i$, so that for the right ear $\theta = 90^\circ + \phi$ and for the left ear $\theta = 90^\circ - \phi$. $f(\mu)$ is the frequency corresponding to the frequency bin $\mu$ and $f_b = \frac{c}{\tau}$ with $c$ the speed of sound and $\tau$ the head radius. The parameters $\gamma(\theta_i)$ and $\tau(\varphi)$ are given by

$$\gamma(\theta_i) = \left( 1 + \frac{\beta_{\min}}{2} \right) + \left( 1 - \frac{\beta_{\min}}{2} \right) \cos \left( \frac{\theta_i}{\beta_{\min}} - 180^\circ \right), \tag{9}$$

$$\tau(\varphi) = \frac{1}{2} \cos(\varphi) \quad \text{if} \quad 0^\circ \leq |\theta| < 90^\circ, \quad \text{or} \quad 90^\circ \leq |\theta| < 180^\circ. \tag{10}$$

with $\beta_{\min} = 150^\circ$ and $\beta_{\min} = 0.1$. In the following, we assume a speed of sound of $c = 340$ m/s and a head radius of $\alpha = 8.75$ cm.

Figure 3 compares the codebooks determined using \(6\) and \(7\) from the measured HRTFs and the analytic model for different azimuth angles $\varphi$. It can be seen that the spherical head model yields a somewhat smoothed approximation of the measured cues. Also, a disturbance of the measured cues around 8 kHz is visible which can be explained by a pinna notch of the artificial head. If the binaural signals contain energy in the affected frequency bands, this disturbance is a potential source of artifacts in the cue modification.

3.2. Source Direction Estimation

To determine the source direction $\varphi_{\text{orig}}(\lambda, \mu)$ from the codebook that best fits the binaural cues in the signal $ILD(\lambda, \mu)$ and $IPD(\lambda, \mu)$, we apply a binaural localization algorithm from [22]. The source direction $\varphi_{\text{orig}}(\lambda, \mu)$ is obtained by minimizing a cost function

$$J(\lambda, \mu, \varphi) = \left| \frac{X_r(\lambda, \mu)}{H_r(\lambda, \mu)} \right|^2 \left| \frac{X_l(\lambda, \mu)}{H_l(\lambda, \mu)} \right|^2.$$

which aims to reconstruct the most consistent source signal $S(\lambda, \mu)$. As shown in [22], the minimization of the cost function can be algebraically simplified to determining the direction $\varphi_{\text{orig}}$ which minimizes the ILD and IPD deviation between the codebook and the perceived cues according to

$$\varphi_{\text{orig}}(\lambda, \mu) = \arg \min_{\varphi} \frac{ILD(\lambda, \mu)}{ILD(\lambda, \mu)} + \frac{ILD(\lambda, \mu)}{ILD(\mu, \varphi)} - 2 \cos \left( IPD(\lambda, \mu) - IPD(\mu, \varphi) \right). \tag{11}$$

3.3. Cue Modification

To rotate the acoustic scene by an angle $\Delta \varphi$, the binaural cues are modified so that a source with an apparent position $\varphi_{\text{dest}} = \varphi_{\text{orig}} - \Delta \varphi$ in the output signal, which in the given architecture is
which yields the complex coefficients’ phases as

\[ \phi_{\text{orig}}(\lambda, \mu) \]

The IPD modification can be equally distributed to both channels, which yields the complex coefficients’ phases as

\[ G_i(\lambda, \mu) = G_i^{\text{ILD}}(\lambda, \mu) \cdot e^{j \Delta \phi_{\text{IPD}}(\lambda, \mu)} \] (12)

The required IPD modification for the rotation of a source from \( \phi_{\text{orig}}(\lambda, \mu) \) to \( \phi_{\text{dest}}(\lambda, \mu) \) is obtained from the cue-to-direction codebook (7) as

\[ \Delta \phi_{\text{IPD}}(\lambda, \mu) = \int \overline{PD}(\mu, \phi_{\text{dest}}(\lambda, \mu)) - \overline{PD}(\mu, \phi_{\text{orig}}(\lambda, \mu)). \]

The IPD modification can be equally distributed to both channels, which yields the complex coefficients’ phases as

\[ G_i^{\text{IPD}}(\lambda, \mu) = e^{\frac{\Delta \phi_{\text{IPD}}(\lambda, \mu)}{2}}. \] (13)

\[ G_i^{\text{IPD}}(\lambda, \mu) = e^{\frac{-\Delta \phi_{\text{IPD}}(\lambda, \mu)}{2}}. \] (14)

The magnitude of the complex coefficient modifies the ILDs and is obtained directly from the HRTFs used in the codebook according to

\[ G_i^{\text{ILD}}(\lambda, \mu) = \frac{|\hat{H}_i(\mu, \phi_{\text{dest}}(\lambda, \mu))|}{|\hat{H}_i(\mu, \phi_{\text{orig}}(\lambda, \mu))|}. \] (15)

Note that the modification gets less invasive for smaller values of \( \Delta \phi \) and the signal is not modified at all for \( \Delta \phi = 0 \).

3.4. Frequency Domain Processing

In the remainder of this paper, a time domain signal \( y_i(k) \) with sample index \( k \) and sampling rate \( f_s = 32 \text{ kHz} \) is divided into frames of length \( N = 1024 \) samples with an overlap of \( W/2 = 448 \) samples. The frames are multiplied with an analysis window \( w(k) \) of length \( N \), in which an effective window of length \( W = N - 2M \) samples is padded with \( M = 64 \) zeros each at the beginning and end, as suggested in [25]. We use the square root of a Hann window as effective window, so that

\[ w(k) = \begin{cases} \sin\left(\frac{(k-M)\pi}{W}\right) & \text{for } M \leq k < N - M \\ 0 & \text{otherwise}. \end{cases} \] (16)

The double-sided zero padding allows for the time-shifting that results from the IPD modifications. A Discrete Fourier Transform (DFT) of length \( N \) is applied to each windowed frame with index \( \lambda \) to obtain \( X_i(\lambda, \mu) \), so that \( \mu \in \{0, \ldots, N - 1\} \). After the modification, the modified signal \( Y_i(\lambda, \mu) \) is transformed back to the time domain by applying the Inverse Discrete Fourier Transform (IDFT), multiplying again with \( w(k) \) as a synthesis window and performing an overlapping addition of the frames with a frame shift of \( W/2 = 448 \). The combination of analysis and synthesis window allows for perfect reconstruction of an unmodified signal and also minimizes artifacts caused by the time-variant filtering.

4. EVALUATION

The evaluation of the method is conducted in a Multiple Stimulus with Hidden Reference and Anchor listening test according to [26] with eight participants. This number is similar to other experiments, e.g., in [14, 17]. For the evaluation, binaural signals are recorded with a Neumann KU100 artificial head placed on a turntable in a listening room with a reverberation time of 0.25 s. Acoustic scenes are created by playing individual source signals on loudspeakers. Four different scenes are recorded with the artificial head facing forward (\( \varphi_{\text{head}} = 0^\circ \)), a frontal single speaker at \( \varphi = 0^\circ \), a single speaker at \( \varphi = 45^\circ \), two simultaneous speakers at \( \varphi = 30^\circ \) and \( \varphi = -30^\circ \) and a music signal consisting of female vocals, drums, guitar and bass at \( \varphi = 60^\circ \), \( \varphi = 30^\circ \), \( \varphi = -30^\circ \), and \( \varphi = -60^\circ \). EBU SQAM speech samples #53 and #54 [27] are used for the speakers.

In order to use a stable reference and to eliminate the influence of different head movements made by different subjects, no head tracking is used in the listening test. The proposed system is employed to artificially rotate the binaurally recorded scenes by fixed angles \( \Delta \varphi \). Six items are chosen for the listening test: both single-source

Fig. 2: Block diagram of the proposed system.

Fig. 3: Comparison of codebook ILDs (top) and IPDs (bottom) determined from measured HRTFs (left) and analytic model (right).
Box plots of the results, aggregated over all subjects and items, are shown in Fig. 5. The results indicate that the system is indeed able to change the perceived source directions and rotate binaural scenes. No significant difference is found between using measured HRTFs and the analytic codebook. The large variance for the medium anchor is expected to depend on the value of $\Delta \varphi$ due to the varying angular distance to the reference. The proposed system’s output should always receive higher ratings than the anchors.

While the mismatch between model and reality may be larger for the analytic codebook, audible artifacts caused by the influence of the pinna notch visible in Fig. 5 may have lowered the ratings for the measured codebook. However, the subjective quality is rated on average similar to the medium anchor, which means that the quality is acceptable. A preprocessing of the measured HRTFs could potentially reduce some artifacts. The nature and origins of the remaining differences, their effect on the plausibility of the scenes and methods to reduce them need to be further investigated.

5. CONCLUSION

A new approach which allows to adapt static binaural recordings to dynamic head movements was presented. It works with unmodified common binaural recording devices such as artificial heads and can also be applied to existing binaural recordings. No source separation is performed and the approach is transparent when the listener’s head orientation matches the orientation of the recording device. The proposed system is computationally simple and can be used for real-time processing. According to the conducted listening test, it can adapt the perceived direction of sound sources to head movements even for complex acoustic scenarios involving reverberation and multiple sources. Therefore, the presented approach bears good prospects to improve localization, externalization and realism of traditional binaural recordings.
6. REFERENCES


