Binaural Techniques with Cross-Talk Canceled Loudspeakers

Yangang Cao

March 6, 2019

Binaural recordings are meant to be played back in such a way that the sound which originates from the left ear is played back only to the left ear, and correspondingly with the right ear. If such a recording is played back with stereophonic setup of loudspeakers, the sound from the left loudspeaker also travels to the right ear, and vice versa, called cross-talk, which ruins the spatial audio quality.

In order to be able to listen to binaural recordings over two loudspeakers, some methods have been proposed. In these methods, the loudspeakers are driven in such a way that in practice the cross-talk is canceled as much as possible.

A system can be formed to deliver binaurally recorded signals to the listener's ears using two closely spaced loudspeakers with cross-talk cancellation. The binaural signals are represented as a 2×1 vector in $\mathbf{x}(n)$, and the produced ear canal signals also as 2×1 vector $\mathbf{d}(n)$. The system can be formulated in the z-domain

$$\mathbf{d}(z) = \mathbf{C}(z)\mathbf{H}(z)\mathbf{x}(z),$$

where $\mathbf{C}(z) = \begin{bmatrix} C_{11}(z) & C_{12}(z) \\ C_{21}(z) & C_{22}(z) \end{bmatrix}$ contains the electro-acoustical responses of the loudspeakers measured in the ear canals, and $\mathbf{H}(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix}$ contains the responses for performing inverse filtering to minimize the cross-talk. Ideally, $\mathbf{x}(z) = \mathbf{d}(z)$, which can be obtained if $\mathbf{H}(z) = \mathbf{C}(z)^{-1}$. Unfortunately, the direct inversion is not feasible due to unidealities of the loudspeakers and the listening conditions. A regularized method to find an optimal $\mathbf{H}_{opt}(z)$ has been proposed

$$\mathbf{H}_{opt}(z) = \left[\mathbf{C}^T\left(z^{-1}\right)\mathbf{C}(z) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^T\left(z^{-1}\right)z^{-m},$$

where β is a positive scalar regularization factor, and z^{-m} models the time delay due to the sound reproduction system. If β is selected very low, there will be sharp peaks in the resulting time-domain inverse filters, which may exceed the dynamic range of the loudspeakers. If β is selected to be higher, the inverse filter will have longer duration in time, which is less demanding on the loudspeakers, but unfortunately the inversion is also less accurate.

A **MATLAB** example is provided in the following to compute inverse filters for a cross-talk canceling system:

- The responses in $\mathbf{C}$ are moved into the frequency domain with discrete Fourier transform (DFT) with the desired length of time window.
- The filter responses are computed by $\mathbf{H}_{opt}(k) = \left[\mathbf{C}^H(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^H(k)$, where k presents the frequency bin indexes and $\mathbf{H}$ Hermitian transposition.
- The inverse DFT is taken of $\mathbf{H}$, resulting in the inverse filters for cross-talk cancellation.
- A circular shift of half of the applied time-window length is implemented on the inverse filters.

```
% Simplified cross-talk canceler

theta = 10;  % spacing of stereo loudspeakers in azimuth
Fs = 44100; % sample rate
b = 10^-5;   % regularization factor
% loudspeaker HRIRs for both ears (ear_num,loudspeaker_num)
% If more realistic HRIRs are available, pls use them
HRIRs(1,1,:) = simpleHRIR(theta/2,Fs);
HRIRs(1,2,:) = simpleHRIR(-theta/2,Fs);
HRIRs(2,1,:) = HRIRs(1,2,:);
HRIRs(2,2,:) = HRIRs(1,1,:);
Nh = length(HRIRs(1,1,:));
%transfer to frequency domain
for i = 1:2
    for j = 1:2
        C_f(i,j,:) = fft(HRIRs(i,j,:),Nh)
    end
end
% Regularized inversion of matrix C
H_f = zeros(2,2,Nh);
for k = 1:Nh
    H_f(:,:,k) = inv((C_f(:,:,k)'*C_f(:,:,k)+eye(2)*b))*C_f(:,:,k)';
end
% Moving back to time domain
for k = 1:2
    for m = 1:2
        H_n(k,m,:) = real(ifft(H_f(k,m,:)));
        H_n(k,m,:) = fftshift(H_n(k,m,:));
    end
end
% Generate binaural signals.  Any binaural recording shoud also ...
    be ok
binauralsignal = simplehrtfconv(70);
%binauralsignal = wavread('road_binaural.wav');
% Convolve the loudspeaker signals
loudspsig = [conv(reshape(H_n(1,1,:),Nh,1), binauralsignal(:,1)) ...
    + ...
conv(reshape(H_n(1,2,:),Nh,1), binauralsignal(:,2)) ...
conv(reshape(H_n(2,1,:),Nh,1), binauralsignal(:,1)) + ...
conv(reshape(H_n(2,2,:),Nh,1), binauralsignal(:,2))];
soundsc(loudspsig,Fs)        % play sound for loudspeakers
```

In practice, this method works best with loudspeakers close to each other, as a larger loud- speaker base angle would lead to coloration at lower frequencies. The listening area in which the effect is audible is very small, as if the listener

departs from the mid line between the loudspeakers by about 1–2 cm, the effect is lost.

A nice feature of this technique is that the sound is typically externalized. This may be due to the fact that head movements of the listener produce somewhat relevant cues, and since the sound is reproduced using far-field loudspeakers generating plausible monaural spectral cues. However, although the sound is externalized, a surrounding spatial effect is hard to obtain with this technique. With a stereo dipole in the front, the reproduced sound scene is typically perceived only at the front.

The technique also is affected by the reflections and reverberation of the listening room. It works best only in spaces without prominent reflections. To get the best results, the HRTFs of the listener should be known, however already very plausible results can be obtained with generic responses.