Auditory Filter-Based Binaural Loss

Dor Shamay

1 Introduction

In previous work, interaural cues are extracted from the time-domain binaural signal using an auditory front-end based on the framework presented in [1], which also forms the foundation of the binaural audio quality metric introduced in [2]. The auditory processing employed in this work adopts a similar structure, consisting of monaural processing applied independently to the left and right channels, followed by binaural processing to extract interaural cues.

Monaural processing begins by modeling the middle ear transfer characteristics with a first-order bandpass filter in the range 500–2000 Hz. This is followed by basilar membrane filtering, approximated using a third-order gammatone filter bank comprising 29 frequency bands distributed from 50 to 6000 Hz, with center frequencies spaced one equivalent rectangular bandwidth (ERB) apart. Cochlear compression is applied via instantaneous compression with an exponent of 0.4 to the gammatone filter outputs. Hair cell transduction is then modeled through half-wave rectification followed by a fifth-order low-pass filter with a cutoff frequency of 770 Hz.

In the binaural processing stage, ILD is extracted by applying a second-order Butterworth low-pass filter with a cutoff frequency of 30Hz to the complex monaural outputs of the cochlear compression mechanism. The filtered outputs are denoted as $a^{l,r}(t, f_{c_k})$. A de-compression with the exponent is applied, and ILD is calculated as:

$$\mathrm{ILD}(t, f_{c_k}) = \frac{20}{0.4} \log_{10} \left(\frac{|a^r(t, f_{c_k})|}{|a^l(t, f_{c_k})|} \right). \tag{1}$$

In a second computation, interaural temporal disparities are derived by applying a second-order gammatone filter to the hair cell transduction monaural outputs, yielding the left and right time-domain complex signals $g^{l,r}(t, f_c)$ in each frequency band f_c. The interaural transfer function (ITF) was calculated for each band as:

$$\mathrm{ITF}(t, f_c) = g^l(t, f_c) \cdot \overline{g}^r(t, f_c), \tag{2}$$

where $\overline{g}^r$ denotes the complex conjugate of g^r. The IPD is then extracted by computing:

$$\mathrm{IPD}(t, f_c) = \arg\left(\mathrm{ITF}(t, f_c)\right). \tag{3}$$

Since sensitivity of the human auditory system to fine temporal structure becomes negligible above approximately 1.4 kHz [3], only frequency bands below this limit are used for IPD analysis.

Finally, interaural vector strength (IVS), introduced in [1] as a measure of interaural coherence (IC) derived from the ITF in the time domain:

$$\mathrm{IVS}(t, f_{c_k}) = \frac{\left|\int_0^\infty d\tau\, \mathrm{ITF}(t-\tau, f_{c_k}) e^{-\tau/\tau_s}\right|}{\left|\int_0^\infty d\tau\, \mathrm{ITF}(t-\tau, f_{c_k}) e^{-\tau/\tau_s}\right|}, \tag{4}$$

where τ_s denotes the integration time constant, set as a multiple of the cycle duration corresponding to the center frequency of the respective gammatone filter band. A value of $\tau_s = 5$ was found to be optimal for localization performance [1], and was also adopted in the binaural audio quality metric in [2].

2 Binaural Loss Incorporating Auditory Filters

To leverage ILD, IPD, and IVS as loss functions, we compute the mean squared error (MSE) between each interaural cue of the target and the estimated signal:

$$\mathcal{L}_{\mathrm{ILD}} = \frac{1}{29T} \sum_{t=1}^{T} \sum_{k=1}^{29} \left(\mathrm{ILD}^{\mathrm{tgt}}(t, f_{c_k}) - \mathrm{ILD}^{\mathrm{est}}(t, f_{c_k})\right)^2, \tag{5}$$

$$\mathcal{L}_{\mathrm{IPD}} = \frac{1}{17T} \sum_{t=1}^{T} \sum_{k=1}^{17} \left(\mathrm{IPD}^{\mathrm{tgt}}(t, f_{c_k}) - \mathrm{IPD}^{\mathrm{est}}(t, f_{c_k})\right)^2, \tag{6}$$

$$\mathcal{L}_{\mathrm{IVS}} = \frac{1}{29T} \sum_{t=1}^{T} \sum_{k=1}^{29} \left(\mathrm{IVS}^{\mathrm{tgt}}(t, f_{c_k}) - \mathrm{IVS}^{\mathrm{est}}(t, f_{c_k})\right)^2, \tag{7}$$

where $\mathrm{ILD}^{\mathrm{tgt}}, \mathrm{IPD}^{\mathrm{tgt}}$, and $\mathrm{IVS}^{\mathrm{tgt}}$ denote the target binaural cues, and $\mathrm{ILD}^{\mathrm{est}}$, $\mathrm{IPD}^{\mathrm{est}}, \mathrm{IVS}^{\mathrm{est}}$ denote the corresponding estimated cues, all computed using (1), (3), and (4), respectively. In these equations, T is the number of time frames and the constants represent the number of frequency bands. The ILD loss $\mathcal{L}_{\mathrm{ILD}}$, the IPD loss $\mathcal{L}_{\mathrm{IPD}}$, and the IVS loss $\mathcal{L}_{\mathrm{IVS}}$ are combined to obtain the overall binaural loss:

$$\mathcal{L}_{\mathrm{binaural}} = \delta\mathcal{L}_{\mathrm{ILD}} + \lambda\mathcal{L}_{\mathrm{IPD}} + \kappa\mathcal{L}_{\mathrm{IVS}}, \tag{8}$$

where δ, λ and κ denote the weights applied to each interaural cue loss component.

References

[1] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model-based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[2] T. Biberger, H. Schepker, F. Denk, and S. D. Ewert, "Instrumental quality predictions and analysis of auditory cues for algorithms in modern headphone technology," *Trends Hear.*, vol. 25, p. 23312165211001219, 2021.

[3] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, vol. 62, no. 1, pp. 157–167, 1977.