VARIATION IN TIMBRE DESCRIPTORS DUE TO OUTER EAR FILTERING

Georgios Marentakis
Department of Information Technology
Østfold University College
georgios.marentakis@hiof.no

Charalampos Saitis
Centre for Digital Music
Queen Mary University London
c.saitis@qmul.ac.uk

ABSTRACT

The psychoacoustic investigation of timbre traditionally relies on audio descriptors extracted from anechoic or semi- anechoic recordings of musical instrument sounds, which are presented to listeners in diotic fashion. As a result, the extent to which spectral modifications due to the outer ear interact with timbre perception is not fully understood. As a first step towards investigating this research question, we examine here whether timbre descriptors calculated using HRTF filtered instrumental sounds deviate across ears and from values obtained from the same sounds without HRTF filtering for different listeners. The sound set comprised isolated notes played at the same fundamental frequency and dynamic from a database of anechoic recordings of modern orchestral instruments and some of their classical and baroque precursors. These were convolved with anechoic high spatial resolution HRTFs of human listeners. We present results and discuss implications for research on timbre perception and cognition.

1. INTRODUCTION

Timbre is a particularly important auditory perceptual attribute, which is determined by both spectral and temporal aspects of sound in a complex way. Timbre is known ot be multidimensional and there have been significant efforts to identify its perceptual dimensions but also find appropriate acoustic correlates [1–3]. However, most of these research efforts have employed on monophonic sounds often recorded in anechoic environments. As a result the influence of the transmission path to the listener has received less attention.

Sounds that propagate to the ear from a musical instrument or a loudspeaker typically undergo a number of transformations due to the influence of room acoustics but also the listener's torso, head, and outer ear(s). The influence of room acoustics on timbre, in particular sound quality, has received some attention in the literature, for example in [4–7]. The impact of the outer ear filtering on timbre perception has been less investigated.

Processing by the outer ear is particularly important for sound localization, in particular elevation and front-back discrimination. The transformation due to the outer ear can be described as a filtering operation between a sound signal and the Head Related Transfer Functions (HRTFs) [8]. Filtering with HRTFs typically results in significant spectral effects.

Listening is binaural and the signals that arrive at the ears are markedly different due to the difference in the propagation path length, head shadowing, and deviations in the shape of both ears. If the signal of one ear was contrasted to the other, it is not unlikely that certain perceptual qualities of the resulting sound will be different.

The motivation behind the article is to understand better the differences in the timbre of a monophonic signal presented diotically and this of a binaural signal. More specifically, we want to understand how does the timbre of perceived sound objects relates to the timbre of the individual signals in each ear as these emerge through HRTF processing.

Our immediate perception would let us think that processing by the outer ear should not affect the recognition of sound timbre. Perceptual mechanisms that compensate for the effects of the channel may help maintain constancy of timbre and ensure that a sound is recognized despite spectral modifications caused by transmission [9]. However, spectral modifications do affect timbre and eventually sound quality even if they do not render sound unrecognizable.

As a first step towards a better understanding of the aforementioned questions, we estimate the extent to which HRTF filtering might result in systematic deviations in the values of acoustic correlates in each ear. We consider the two primary perceptual dimensions of timbre: spectral centroid and (log) attack time. Furthermore, we speculate about how these may be combined to yield a single timbre percept.

2. BACKGROUND

HRTFs contain the influence of a multitude of interactions of sound with human body. These contain reflections from the torso and the shoulders, head shadowing, and the influence of the outer ear and of the ear canal, not all of which are direction dependent. The influence of the outer ear is typically found in high frequencies typically above 2kHz. However, reflections from shoulder and body, and head diffraction and reflection affect the spectrum significantly pretty much throughout the audible spectrum [8]. Spectral cues are known to be important for localization in particular elevation perception and front-back discrimination [10]. A spectral peak around 1kHz is associated with sounds coming from the back [10–13]. A spectral notch with central frequency between 6 and 9 kHz is a stable cue for elevation [14]. Head shadowing results in that high frequencies are attenuated in the contralateral ear and is responsible for the interaural level differences (ILDs) that play a crucial role in localization in the horizontal plane.

Although HRTF filtering alters the sound spectrum, it does not introduce new frequencies. Rather, it changes the relative level of existing frequency components. As a result the amplitude of peaks or notches in the signal may be shifted. Spectral peaks and notches modify the timbre of the resulting reproduction, often leading to coloration [4–7]. The audibility of resonances or anti-resonances depend on their Q factor, center frequency, and amplitude. Changes in the amplitude of spectral notches or peaks as low as 2-4 dB may be detected, thresholds increase, however, as notch bandwidth decreases [15]. Peaks and notches are not detected in the same way. Peaks are easier to detect when bandwidth is limited in comparison to notches as in the latter case energy from neighbouring bands leaks in the excitation pattern [16]. This coloration should be relatively easy to hear, as peaks and notches in HRTFs may easily exceed 20dB.

A side effect of HRTF filtering may also be that the center frequency of existing peaks and notches may be shifted, for example, due to sound or listener movement. Furthermore, the redistribution of spectral energy may result in changes in timbral descriptors, such as the spectral centroid. Changes in the resonant frequency of a second-order filter can be discriminated if they exceed 8% the centre frequency, or even less for Q>1, for centre frequencies between 300 and 2kHz [17]. Furthermore, changes of around 1% centre frequency for an 1 and 8kHz noise band with a bandwidth of 0.125 × centre frequency were discriminated above threshold [15]. These values compare well to these obtained for the frequency discrimination of simple tones, which can be performed at changes of about 1% tone frequency (increasing to 3% above 4kHz) [18].

HRTF filtering introduces frequency-dependent time delay as the path to each ear has a different length. This is typically modelled as a frequency independent time-delay and a frequency dependent phase difference. However, the extent to which phase information affects the perception of timbre is the matter of a long-standing debate.

3. SIMULATIONS

3.1 Anechoic Monaural Recordings

Single notes from 19 common orchestral instruments were selected from two extensive databases of anechoic instrument recordings: bassoon (BSN), clarinet (CLN), flute (FLT), English horn (EHN), oboe (OBO), harp (HRP), acoustic guitar (GTR), double bass pizzicato (DBP), bowed cello (CLB), bowed violin (VLB), violin pizzicato (VLP), vibraphone (VIB), marimba (MBA), and piano (PNO). The last four notes were taken from the University of Iowa Musical Instrument Samples database [1] ; all other came from the TU Berlin Database of Anechoic Microphone Array Measurements of Musical Instruments [19]. All notes are played forte at 311 Hz (Eb4) without vibrato.

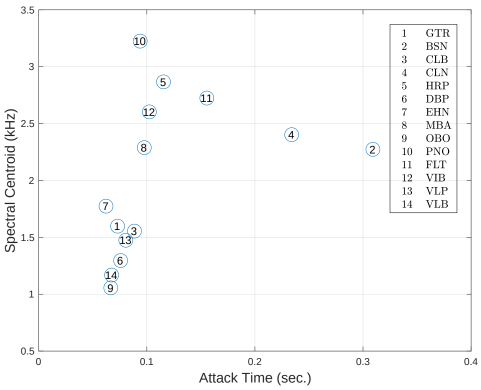


Figure 1: The timbre space of the analyzed anechoic instruments recordings focusing on the two primary dimensions: attack time and spectral centroid.

The Berlin recordings were made using a quasi-spherical 32-microphone array. For the purposes of the present analysis, only one of the 32 channels was used from each recording. Calculating a sum of the channels was not considered to avoid comb filter effects. Instead, for each instrument we selected that channel which most often exhibited the highest RMS signal level over all recorded notes as the principal channel (i.e., as the principal direction of sound radiation). The same approach was applied to the stereo recordings of the Iowa database.

3.2 Signal Pre-processing and HRTF Convolution

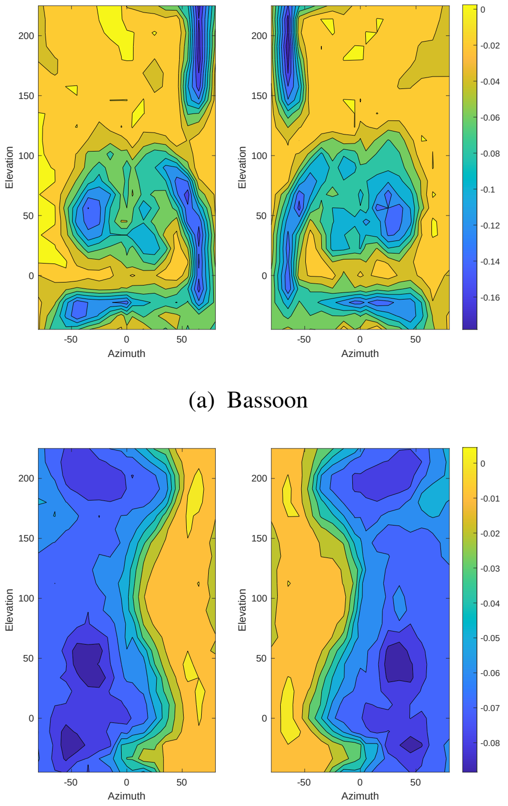
Both databases comprise recordings made at 44.1 kHz but with different numbers of bits per sample. We resampled the Berlin sounds from 32 to 24 bits to match the resolution of the Iowa database. Sounds were normalized using their RMS value. All sounds were cropped mildly in the beginning. This involved selecting as a starting point the moment the sample attained 5dB SNR. Furthermore, only the first 500 ms of the sound were used.

Subsequently, samples were filtered with anechoic HRTF sets from the CIPIC database [20]. Not all HRTF azimuths were used. We focused on N = 17 azimuths: −80°, −65°, −55°, −45°, −35°, −25°, −15°, −5°, 0°, 5°, 15°, 25°, 35°, 45°, 55°, 65°, 80° and N=25 elevations that spanned −45° and 225° with a step of 11.25°. This yielded a total of 17 × 25 = 425 HRTF measurements per ear (34% of available grid). The operation was performed on the first thirty subjects.

3.3 Audio Descriptors

We focus on two acoustic features that are related to salient dimensions of timbre perception, namely log attack time and spectral centroid. These were extracted from the audio

[1] `http://theremin.music.uiowa.edu/MIS.html`



(a) Bassoon

(b) Clarinet

Figure 2: Deviation in the attack time between HRTF-filtered and anechoic input signals for the (azimuth,elevation) pairs in the simulations averaged across subjects. Deviation is spread to larger areas of space for the clarinet in comparison to the bassoon. Left panel is left ear, middle right ear.

signals using the Timbre Toolbox [2], first for the monophonic anechoic recording and then for each of the binaural HRTF-filtered signals for each position and HRTF set.

The log-attack time is defined as the (logarithm) of the duration between the onset of a sound and its more stable part. Attack time is a global feature of the signal computed from its temporal envelope. Attack time was calculated from the temporal envelope based on the weakest-effort method [2].

The spectral centroid is defined as the amplitude-weighted mean frequency of the sound spectrum and can be interpreted as the center of gravity of the spectral envelope or the frequency that divides the spectrum into two regions with equal energy. It has been shown to correlate with brightness ratings of musical instrument tones across different psychoacoustical tasks [21]. Spectral centroid is time-varying; it was computed for each 25 ms time frame. Spectra were derived using an ERB-spaced gammatone filter bank decomposition of the signal [22, 23]. The median value is taken into consideration in the analysis.

	Attack Time (s)			
	Min	Anechoic	Max	Range
Bassoon	0.104	0.309	0.319	0.214
Clarinet	0.120	0.234	0.270	0.149
Harp	0.102	0.115	0.212	0.109
Flute	0.094	0.156	0.184	0.091
Vibraphone	0.084	0.102	0.143	0.059
Marimba	0.080	0.098	0.131	0.051
Cello	0.084	0.089	0.120	0.036
Violin (P)	0.059	0.081	0.083	0.024
Piano	0.091	0.094	0.096	0.005
Horn (E)	0.060	0.062	0.063	0.004
Guitar (A)	0.072	0.073	0.075	0.003
Bass (D)	0.075	0.076	0.078	0.003
Oboe	0.066	0.067	0.067	0.002
Violin (M)	0.067	0.067	0.068	0.001

Table 1: Variation in the attack time estimators. Minimum and maximum estimator values for HRTF-filtered input signals are shown and the associated range.

4. RESULTS

The spectral centroid and attack time of the anechoic monophonic signals as were calculated using the timbre toolbox are plotted in Figure 1. We see that attack time ranges from 60 msec to 300 msec for the instruments under consideration. Spectral centroid ranges with 1.06 to 3.22 kHz. The variation in attack time is somewhat limited because all analyzed sounds were played forte.

4.1 Attack Time

Table 1 shows the attack time for the anechoic monophonic samples as well as the maximum and minimum values attained after HRTF filtering with the 30 HRTF sets and the respective range.

It is clear that HRTF filtering may not only reduce but also increase attack time. However, not all instruments are affected in the same way by HRTF filtering. For the last six instruments in Table 1 the range of the deviation of attack time from the monophonic recordings was below 10ms. On the other hand, for the first six instruments attack time after HRTF filtering deviates more than 50ms from the anechoic monophonic recordings and up to 100ms shifts from the monophonic sample attack time are seen. The deviations are relatively systematic across the different areas of space. The size of the affected area depends on the instrument, for some it is large, as for example for the clarinet, in Figure 2b, while for others smaller, as for the bassoon in Figure 2a, or even insignificant. Interestingly, for some instruments the deviation is higher for the contralateral and negligible for the ipsilateral ear, as for the clarinet. For others, both ears are affected in a similar way, as for the bassoon (Figure 2a).

	Spectral Centroid (kHz)			
	Min	Anechoic	Max	Range
Piano	2.07	3.22	4.11	2.04
Vibraphone	1.74	2.60	3.39	1.65
Flute	1.92	2.72	3.46	1.54
Harp	2.13	2.87	3.59	1.47
Horn(E)	1.21	1.77	2.54	1.33
Marimba	1.68	2.29	2.95	1.27
Clarinet	1.80	2.40	2.98	1.18
Bassoon	1.76	2.27	2.88	1.12
Guitar(A)	1.23	1.60	2.16	0.93
Violin (P)	1.22	1.47	1.95	0.73
Cello	1.32	1.55	2.00	0.67
Bass(D)	1.13	1.30	1.65	0.52
Violin (M)	1.02	1.17	1.52	0.51
Oboe	0.99	1.06	1.29	0.30

Table 2: Variation in spectral centroid estimators. Minimum and maximum estimator values for HRTF-filtered input signals are shown and the associated range.

4.2 Spectral Centroid

Table 2 shows the spectral centroid for the anechoic monophonic samples as well as the maximum and minimum centroids attained after HRTF filtering. As observed for the attack time, the spectral centroid after HRTF filtering may exceed but also fall short of the value attained for the monophonic anechoic sound. The spectral centroid is displaced significantly for all instruments due to HRTF filtering. Again, however, the range varies from 2.04 kHz for the Piano to 0.3 kHz for the oboe.

Figure 3 shows the deviation in terms of spectral centroid for the azimuth and elevation combinations used in the simulations and the instrument with the highest range in the log-attack values in the simulations (Piano). Interestingly, the deviation patterns are highly symmetric and an increase in spectral centroid frequency for one ear is accompanied by a fall in the spectral centroid frequency for the other. Spectral centroid moves towards lower frequencies for the contralateral ear while it increases for the ipsilateral relative to the value attained for the anechoic monophonic sample. This pattern holds approximately for all elevations, the exact turning point depends on the elevation.

If the deviations from the anechoic monophonic spectral centroid of the two ears were to be added, we see that the spectral centroid tends to increase in the area of space between ±30° azimuth and between -30° and 80° elevation, while it decreases for the rest of the space. The shift amount is smaller compared to its value for each contributing ear.

5. DISCUSSION

The goal of this study was to estimate the extent to which established acoustic correlates of the perceptual dimensions of timbre are affected by HRTF filtering. To this end,

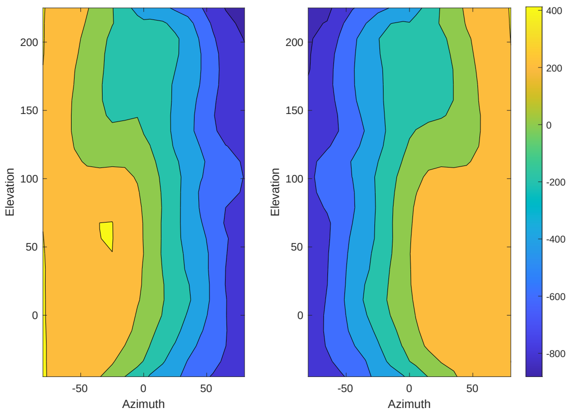


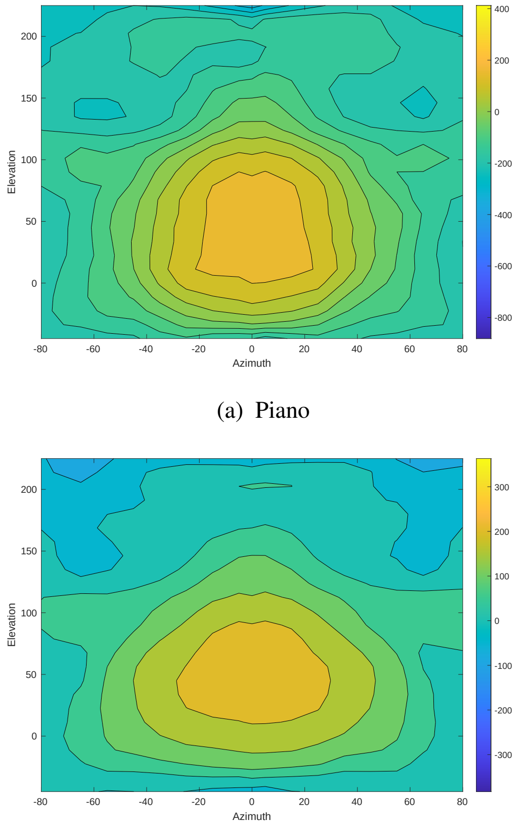
Figure 3: Deviation in the spectral centroid between HRTF-filtered and anechoic input signals for the (azimuth,elevation) pairs in the simulations averaged across subjects. The case of the piano is shown. Left panel is left ear, right is the right ear.

the attack time and the spectral centroid were calculated for monophonic anechoic samples and values were compared with these obtained after samples were HRTF filtered. The results indicate the both features examined here were affected by HRTF filtering.

For spectral centroid, a symmetry with respect to the median plane between the two ears was observed. The spectral centroid is reduced due to HRTF filtering on the ipsilateral ear but is increased in the contralateral ear. This points towards a, perhaps dominating, influence of head shadowing in the results. The increase in spectral centroid on the ipsilateral ear may be explained by energy boost in high frequency regions due to HRTF filtering, while the decrease in the spectral centroid in the contralateral ear is due to the attenuation of high frequencies by head shadowing.

The results for the attack time are not as clear-cut. When symmetry with respect to the median plane is observed, attack time decreases at the ipsilateral ear and remains stable on the contralateral ear. In other cases, the decrease is associated with specific areas of space. A straight forward explanation based on head shadowing is not as easy to make.

An interesting observation in the results is that the attack time and spectral centroid estimators differ for both ears. To understand how this may affect the perception of timbre, we draw on an assumption formulated in [24], according to which the perception of time-invariant patterns in the timbre of complex tones is related to the relative level produced at the output of each auditory filter [24]. For binaural sounds, this would imply combining energy levels from the two ears, perhaps in a similar way as is already done for loudness [25]. Even though such an investigation is outside the scope of this article, we illustrate in Figure 4 a hypothesis about the frequency of a binaural spectral centroid based on summing estimates for each ear, as the ones in Figure 3 for the piano. It is clear that the symmetry with respect to the median plane discussed above results in



(a) Piano

(b) Clarinet

Figure 4: The sum of spectral centroid estimator of each ear for the (azimuth, elevation) pairs in the simulations averaged across subjects for the piano and the clarinet. Centroid is higher for frontal incidence.

that deviations are smoothed out. Nevertheless an area of higher spectral centroid remains for sounds within a narrow area in front of a listener; an interesting hypothesis to be tested experimentally in a perceptual experiment.

6. CONCLUSION

Timbre is usually investigated using anechoic monophonic recordings and the influence of the spectral modifications due to the torso, head, and outer ear is often neglected. Even if such modifications do not render a sound unrecognizable, they arguably affect its sound colour or timbre. To identify whether this is the case we estimated the extent to which acoustic correlates to the two primary timbre dimensions vary across ears and in relation to an anechoic sample. The results indicate that most often the values of the acoustic in each ear deviate significantly both between each other but also from the values obtained using a monophonic anechoic recording. Furthermore, even after combining the descriptors according to a simplified binaural summation process, systematic differences to the monophonic sound remain.

7. ACKNOWLEDGEMENTS

We would like to acknowledge the support of Leif Johannsen from TU Berlin in preparing the sound recordings. Parts of the work performed while C. Saitis was a Humboldt research fellow at TU Berlin and Georgios Marentakis a teaching and research fellow at TU Graz.

8. REFERENCES

[1] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, “Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes,” *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.

[2] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, “The Timbre Toolbox: Extracting audio descriptors from musical signals,” *The Joutnal of thr Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[3] S. McAdams, “The perceptual representation of timbre,” in *Timbre: Acoustics, Perception, and Cognition*, pp. 23–57, Springer, 2019.

[4] S. E. Olive and W. L. Martens, “Interaction between loudspeakers and room acoustics influences loudspeaker preferences in multichannel audio reproduction,” in *Audio Engineering Society Convention 123*, Audio Engineering Society, 2007.

[5] S. E. Olive, P. L. Schuck, S. L. Sally, and M. E. Bonneville, “The effects of loudspeaker placement on listener preference ratings,” *Journal of the Audio Engineering Society*, vol. 42, no. 9, pp. 651–669, 1994.

[6] S. E. Olive, P. L. Schuck, S. L. Sally, and M. Bonneville, “The variability of loudspeaker sound quality among four domestic-sized rooms,” in *Audio Engineering Society Convention 99*, Audio Engineering Society, 1995.

[7] F. E. Toole and S. E. Olive, “The modification of timbre by resonances: Perception and measurement,” *Journal of the Audio Engineering Society*, vol. 36, no. 3, pp. 122–142, 1988.

[8] D. R. Begault, *3-D sound for virtual reality and multimedia*. Morgan Kaufmann Pub, 1994.

[9] C. D. Pike, *Timbral constancy and compensation for spectral distortion caused by loudspeaker and room acoustics*. PhD thesis, University of Surrey (United Kingdom), 2016.

[10] J. Blauert, *Spatial hearing*. Cambridge, MA: MIT Press, 1997.

[11] J. Blauert, “Sound localization in the median plane,” *Acta Acustica united with Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

[12] C.-J. Tan and W.-S. Gan, "User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems," *Electronics letters*, vol. 34, no. 25, pp. 2387–2389, 1998.

[13] R. H. So, N. Leung, A. B. Horner, J. Braasch, and K. Leung, "Effects of spectral manipulation on non-individualized head-related transfer functions (hrtfs)," *Human factors*, vol. 53, no. 3, pp. 271–283, 2011.

[14] B. Zonooz, E. Arani, K. P. Körding, P. R. Aalbers, T. Celikel, and A. J. Van Opstal, "Spectral weighting underlies perceived sound elevation," *Scientific reports*, vol. 9, no. 1, p. 1642, 2019.

[15] B. C. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 khz," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 820–836, 1989.

[16] M. G. Heinz and C. Formby, "Detection of time-and bandlimited increments and decrements in a random-level noise," *The Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 313–326, 1999.

[17] J.-P. Gagné and P. Zurek, "Resonance-frequency discrimination," *The Journal of the Acoustical Society of America*, vol. 83, no. 6, pp. 2293–2299, 1988.

[18] A. Sek and B. C. Moore, "Frequency discrimination as a function of frequency, measured in several ways," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2479–2486, 1995.

[19] S. Weinzierl, M. Vorländer, G. Behler, F. Brinkmann, H. von Coler, E. Detzner, J. Krämer, A. Lindau, M. Pollow, F. Schulz, *et al.*, "A database of anechoic microphone array measurements of musical instruments," 2017.

[20] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102, IEEE, 2001.

[21] C. Saitis and K. Siedenburg, "Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2256–2266, 2020.

[22] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[23] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception* (Y. Cazals, L. Demany, and K. Horner, eds.), (Oxford), pp. 429–446, Pergamon Press, 1992.

[24] B. C. Moore, "Loudness, pitch and timbre," *Blackwell handbook of sensation and perception*, pp. 408–436, 2005.

[25] B. C. Moore and B. R. Glasberg, "Modeling binaural loudness," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1604–1612, 2007.