DOI: 10.61782/fa.2025.0809

FORUM ACUSTICUM
EURONOISE 2025

SPATIAL AUDIO MODELS' INVENTORY TO COVER THE ATTRIBUTES FROM THE SPATIAL AUDIO QUALITY INVENTORY

Pedro Lladó[1*] **Annika Neidhardt**[1] **Fabian Brinkmann**[2] **Enzo De Sena**[1]

[1] Institute of Sound Recording, University of Surrey, United Kingdom
[2] Audio Communication Group, Technische Universität Berlin, Germany

ABSTRACT

The Spatial Audio Quality Inventory (SAQI, Lindau *et al.* 2014 [1]) defines a comprehensive list of attributes for quality assessment of spatial audio. These attributes are traditionally used in perceptual experiments. However, automatic evaluation is a common alternative to assess spatial audio algorithms by means of acoustic recordings and numerical methods. This study aims at bridging the gap between perceptual evaluation and automatic assessment methods. We performed a focused literature review on available auditory models and proposed a list to cover the attributes in SAQI based on self-imposed selection criteria, such as binaural compatibility. The selected models are publicly available and ready to be used in automatic assessment methods. This Spatial Audio Models' Inventory (SAMI) could serve as relevant metrics to train and/or optimise machine-learning and deep-learning algorithms when the objective is to improve the perceived quality of reproduction in spatial audio applications. Moreover, SAMI composes a benchmark to challenge novel models.

Keywords: *Spatial audio, auditory models, inventory*

1. INTRODUCTION

Spatial audio reproduction aims at delivering three-dimensional characteristics of an acoustic scene to the listener [2]. Spatial features are typically achieved by means of a loudspeaker array around the listener or by adding localisation cues, from the filtering of the head and torso, to headphone reproduction. Adoption of spatial audio is key to enhance immersion, e.g. for virtual and augmented reality experiences, and spatial awareness when hearing aids or hearing protectors are worn [3, 4].

The Spatial Audio Quality Inventory (SAQI) [1] provides a comprehensive list of perceptual attributes relevant to evaluating spatial audio quality. The SAQI attributes result from an exhaustive expert group study and include a number of attributes traditionally used in psychoacoustic experiments. However, with the increasing interest in automated evaluation methods, there is a growing need for models that predict perceptual assessment with computational approaches. A recent study identified relevant models to estimate sound quality attributes of mono signals [5]. Adopting a similar approach, this study identifies models for spatial audio quality assessment that align with SAQI attributes. The result is the Spatial Audio Models' Inventory (SAMI), which provides a catalogue of models to capture essential perceptual attributes in automated workflows. SAMI aims at facilitating automated spatial audio quality assessment.

2. SPATIAL AUDIO ATTRIBUTES AND MODEL SELECTION METHODOLOGY

The 36 SAQI attributes are listed in Table 1, together with a summary of the outcome of this review. For each attribute, we selected a model to facilitate the introduction of auditory models in numerical assessment methods. This model selection aims at leveraging the model performance with its ease of use; it does not aim at assessing which is the best model for each attribute. The

*__Corresponding author__: p.llado@surrey.ac.uk.

model selection is performed based on four criteria. First, preference is given for models for which code is *publicly available*. Second, it must have been *validated* against psychoacoustic data. Third, preference is given to models that are able to process *binaural signals*. And fourth, preference is given to *simple models*, since they are generally preferred unless the increased complexity is needed to fit the purpose [6]. We additionally ranked the level of suitability of the models using a numerical scale:

- Lvl. 0 - no model available;
- Lvl. 1 - no implementation available but the literature suggests how to compute it;
- Lvl. 2 - the model provides means to estimate the attribute, but it is not explicitly estimated;
- Lvl. 3 - the model output captures the attribute but may not isolate it from other attributes;
- Lvl. 4 - the model explicitly estimates the attribute.

The SAMI focuses on single-source static scenes and comparisons against an explicit reference, thereby excluding models that consider listener or source motion, or comparisons to implicit, inner references. In relation to the latter, the artifact-related attributes were excluded. Determining whether a difference is due to an artifact (i.e. which are defined as *unintended* in [1]) relies on a high-level understanding of the signal that auditory models may lack. Moreover, multi-modal and/or highly multi-dimensional attributes, such as 'responsiveness', 'tactile vibration', 'naturalness', 'presence' and 'degree-of-liking', and 'other' were also excluded.

3. SAMI MODELS

The following paragraphs introduce the selected models for each attribute, grouped by the SAQI categories shown in the first column of Table 1. Some of the attributes were grouped to avoid redundancy in the definitions.

3.1 Difference

The "difference" is intended as an integrative audio quality measure that characterises dissimilarities between audio samples. The presence of differences enables studying the contribution of the more specific attributes. Several models address such integrative measures by aggregating monaural and binaural cues to estimate timbral and spatial quantities. In [23], based on [24], a data driven approach was introduced which computes spatial and timbral quality separately and weights them according to a perceptually-motivated measure. In [7], an auditory model with one monaural and one binaural path estimates the overall difference as the largest when comparing the

Table 1: Summary of the models suggested for the SAQI attributes. If the code is included in a toolbox, it is written in parenthesis. The last column refers to the level of suitability according to Sec. 2. The greyed-out attributes are considered out of scope (see Sec. 2). (*) Tested for RIRs only. (M) The model uses monaural inputs.

	SAQI ATTRIBUTE	AUDITORY MODEL	Lvl.
	Difference	eurich2024 [7]	4
Timbre	Tone color (bright dark)	mckenzie2025 [8] (AMT)	3
	High-frequency tone color		3
	Mid-frequency tone color		3
	Low-frequency tone color		3
	Sharpness	DIN 45692:2009-08 [9] (SQAT)	4 (M)
	Roughness	daniel1997 [10] (SQAT)	4 (M)
	Comb-filter colouration	mckenzie2025 [8] (AMT)	3
	Metallic tone color		3
Tonalness	Tonalness	terhardt1982 [11] (SQAT)	4 (M)
	Pitch	kim2018 [12]	4 (M)
	Doppler effect		0
Geometry	Horizontal direction	dietz2011 [13] (AMT)	4
	Vertical direction	baumgartner2014 [14] (AMT)	4
	Front-back position		4
	Distance	georganti2013 [15] (AMT)	4
	Depth		0
	Width	klockgether2014 [16]	1
	Height		0
	Externalization	li2020 [17] (AMT)	4
	Localisability	barumerli2023 [18] (AMT)	2
	Spatial disintegration		0
Room	Level of reverberation	klockgether2014 [16]	1
	Duration of reverberation	osses2017 [19]	4 (M)
	Envelopment	klockgether2014 [16]	1
Time Behaviour	Pre-echoes	jüterbock2023 [20]	2*
	Post-echoes		2*
	Temporal disintegration		0
	Crispness	moore2018 [21] (AMT)	2
	Speed		0
	Sequence of events		0
	Responsiveness		0
Dynamics	Loudness	moore2018 [21] (AMT)	4
	Dynamic range		2
	Dynamic compression effects		2
Artifacts	Pitched artifact		0
	Impulsive artifact		0
	Noise-like artifact		0
	Alien source		0
	Ghost source		0
	Distortion		0
General	Tactile vibration		0
	Clarity	C_{50} / C_{80}	2 (M)
	Speech intelligibility	lavandier2022 [22] (AMT)	4
	Naturalness		0
	Presence		0
	Degree-of-liking		0
	Other		0

two paths. Both [23] and [7] are equally good candidates from a performance point of view. The latter was selected since it is available in the Auditory Modelling Toolbox (AMT) [25], toolbox that contains several models selected for the SAMI.

3.2 Timbre-related attributes

Timbre is a multidimensional psychoacoustic attribute that allows differentiating two sounds with the same pitch, loudness,spatial location, and duration [26].

Tone colour: Six SAQI attributes are included under "tone colour": 1) tone colour bright-dark, 2) high-, 3) mid- and 4) low-frequency tone colour, 5) comb-filter colouration and 6) metallic tone colour. Despite the differences in their definitions, all these attributes refer to spectral differences between a target and a reference sound.

Models that estimate colouration focus on quantifying differences in the magnitude spectrum, in most cases processing the target and the reference sounds with a peripheral auditory model. Most binaural colouration models are based on the Composite Loudness Level (CLL) model introduced in [27]. This CLL model computes the specific loudness of the target sound, i.e. the loudness associated with each auditory filter, e.g. at the output of a Gammatone Filterbank [28]. Then, it compares it to the specific loudness of the reference for each frequency band. The colouration is then estimated by integrating the differences over frequency bands and over ears. A model introduced recently outperformed previous colouration models on predicting perceptual data [8]. The model will be available in the next release of the AMT [25].

Sharpness: Defined as the "timbral impression which e.g., is indicative of the force with which a sound source is excited" [1]. Sharpness is associated with the center of gravity of the spectrum of a sound, with sharper sounds associated with a higher center of gravity [29]. The sound level also seems to affect the sharpness, but not prominently. A model proposed in [29] measures the sharpness by a weighted integration of the specific loudness over frequency bands divided by the total loudness. This model was adopted in the DIN 45692:2009-08 standard [9]. Despite the simplicity of this approach, it was shown to provide fairly accurate estimates [30]. Its implementation is available in the Sound Quality Analysis Toolbox (SQAT [1]) [5,31]. Note that this model is monaural.

Roughness: Defined as "timbral impression of fierce or aggressive modulation/vibration" where "individual oscillations are hardly distinguishable" [1]. Conceptually, roughness is easily explained for a pair of tones that are within a frequency distance of about $15 < \Delta f < 300$ Hz, which cause an unpleasant amplitude modulation [29]. Models of roughness typically compute the modulation depth of the acoustic signal for each auditory frequency band, and integrate over frequency to estimate the roughness [10, 32, 33]. The model from [32] seems to correlate well with perceptual roughness according to [10]. A MATLAB implementation is included in SQAT [5, 31]. Note that this model is monaural.

3.3 Tonalness-related attributes

Tonalness is the attribute that is rated low for noisy sounds and increases for sounds with prominent tones [30].

Tonalness: Defined as "perceptibility of a pitch in a sound" [1]. A model introduced in [34] estimates the tonalness from the estimated bandwidth, central frequency and level of each tonal component identified using the algorithm proposed in [11]. The model implementation is available in SQAT [5, 31]. Note that this model is monaural.

Pitch: Defined as "The perception of pitch allows arranging tonal signals along a scale 'higher – lower' " [1]. Two theories of pitch perception have been traditionally studied [35]. The first one assumes that the pitch is extracted from a spectral analysis of the acoustic information (place theory, referring to the place of maximum excitation of the basilar membrane). The second hypothesis assumes that the pitch is extracted from the periodicity information (temporal theory). The temporal theory is currently more widely accepted, since the place theory does not account for certain pitch-related phenomena that occur within a single auditory frequency band [36]. The simplest model adopting the temporal theory is based on the auto-correlation after applying peripheral processing [35]. Several data-driven models have shown to be a robust alternative [12, 37]. An open source implementation of [12] is available [2], which outperforms the state of the art and is selected here.

3.4 Geometry-related attributes

Geometry, in this context, refers to the physical properties of sound sources in relation to their position in the

[1] https://github.com/ggrecow/SQAT

[2] https://github.com/marl/crepe

space and their size. Note that the focus is on estimating perceived geometrical properties and not true physical properties of the sound sources.

Horizontal direction (left-right direction) Defined as "direction of a sound source in the horizontal plane" [1]. Since there is one specific attribute about front-back position (see next attribute), the interpretation of the horizontal direction here is limited to the range from -90° (right) to 90° (left), and we will refer to it as the left-right direction. The perception of left-right direction is possible mainly due to the interaural cues, i.e. interaural time and level differences (ITDs and ILDs, respectively) that depend on the location of the source.

The most common models of left-right direction are based on the coincidence-detection model [38]. Such models estimate the ITDs at the output of a binaural peripheral model by finding the lag that maximises the interaural cross-correlation (IACC) at each frequency channel. ILDs are often computed as the energy ratio between the left and the right ear for each frequency channel [13]. The direction of arrival can be then computed using a look-up table that links the lateral angles to their corresponding interaural differences [13]. The model in [13] is particularly interesting due to the trade-off between its simplicity, its accuracy, the quality of its documentation and its verified code, which is available in [25]. However, if complex phenomena need to be modelled, e.g. precedence effect, more complex models may be needed.

Vertical direction and front-back position (polar direction) The vertical direction is defined as "direction of a sound source in the vertical plane" [1]. The front-back position is defined as "impression of a position difference of a sound source caused by 'reflecting' its position on the frontal plane." [1]. These two attributes have been merged here since both are related to resolving the direction of the sound in the cone of confusion. Localisation of the polar direction in the cone of confusion (or in a sagittal plane, if the distance is fixed) is possible due to the monaural spectral cues [39]. These spectral cues are systematic changes in the spectrum of the sound due to the filtering of the torso, head and pinnae, which imprints systematic spectral changes to the arriving sound.

Models for the localisation in the polar direction are often based on a template-matching strategy. They compute the spectral cues of the target sound and compare them with an internal template, obtained by computing the spectral cues for all the directions available in the head-related transfer functions (HRTFs) of the listener. The sagittal-plane localisation model introduced in [14] estimates the probability for the sound coming from each direction in the polar dimension. The vertical direction and the front-back position can be computed by restricting the range of available polar angles in the response range for the model output. The model is available in [25]. It is worth mentioning that the models introduced in [18, 40] might be interesting alternatives to estimate direction by combining both lateral and polar directions.

Distance: Defined as "perceived distance of a sound source" [1]. Multiple factors seem to influence the perceived distance, such as the sound-pressure level, the direct-to-reverberant ratio (DRR), and spatial cues [15].

A model of distance proposed in [15] computes interaural cues, i.e. ITDs, ILDs and interaural coherence (IC), at the output of a peripheral auditory model. Several statistics are computed from these interaural cues, which are fed into a gaussian-mixture-model (GMM) classifier to estimate the distance. This model seems to provide promising results when the training and testing environment have a reverberation time of similar order. Its implementation is available in [25].

Width: Defined as "perceived extent of a sound source in horizontal direction" [1]. The term apparent source width (ASW) is often used, especially in the context of concert hall acoustics. The ASW is one of the two aspects of the attribute *spaciousness* introduced in [41], together with the envelopment. Changes in ASW seem to be related to the IC and fluctuations in interaural cues, which can occur due to lateral early reflections [42]. The ASW has been traditionally measured as 1-IACC, or more precisely, as 1-IACC_{E} (E for the early part of the room impulse response), which reflects the connection between a drop in IC and a wider ASW [43]. This strategy was adopted in [16, 44], where they estimated the ASW from the 1-IACC at the output of a peripheral auditory model. In [16], the energy ratio between a temporal segment and the previous one is used to determine the early part of the signal. The latter is selected but, despite a detailed description, its implementation is not publicly available.

Height and Depth: Defined as "perceived extent of a sound source in vertical (height) and radial (depth) direction" [1]. To the best of our knowledge, no available model covers these attributes.

Externalisation: "Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance" [1]. Similarly to distance perception, externalisation perception seems to depend on several acoustic attributes, such as interaural cues, monaural spectral cues and DRR [17,45]. Two models de-

veloped in parallel, introduced in [17] and [46], estimate the externalisation from a binaural signal by combining the estimated ILDs, ILD fluctuations and monaural spectral cues. The model in [17], available in [25], includes a long-term and a short-term memory paths to account for the effect of reverberation in externalisation, and is selected here for this reason.

Localisability: "Often associated with high/low perceived extent of a sound source. If localisability is low, spatial extent and location of a sound source are difficult to estimate, or appear diffuse. If localisability is high, a sound source is clearly delimited." [1]. This attribute may be related to the precision in the localisation responses. Uncertainty measures, such as the estimated quadrant error rate [14], are potentially useful to estimate the localisability. The model introduced in [18], available in [25], seems especially relevant, since it covers the full sphere and provides a probabilistic output. The model computes ITDs, ILDs and monaural spectral cues of the target signal and compares it with those of a template derived from the listener's HRTFs. The model estimates the probability of the target sound coming from each direction in the sphere (discretised) and computes the maximum a posteriori to give a response. Localisability could potentially be estimated from the posterior.

Spatial disintegration: Defined as "sound sources, which - by experience - should have a united spatial shape, appear spatially separated" [1]. This concept is related to inconsistency of localisation cues resulting in the perception of multiple sources instead of one (e.g. different frequency ranges of interaural cues containing substantially different ITD and ILD values). The model of localisation in the lateral angle introduced in [13] has demonstrated the capability of localising more than one source, corresponding to the perceptual responses obtained in listening experiments. The same concept could be applied for the localisation model introduced in [18], which covers the full sphere. However, spatial disintegration is a phenomenon that may occur due to early reflections (see categories 1 and 2 in [47]), for which more complex mechanisms should be considered. In any case, to the best of the authors' knowledge, there is no established modelling strategy to making the decision about the disintegration.

3.5 Room-related attributes

Room-related attributes arise from the existence of reflections in the sound. While reflections may influence all other attributes in SAQI, this section refers only to attributes that appear exclusively from room reflections.

Level of reverberation and envelopment by reverberation: Level of reverberation is defined as "the perception of a strong reverberant sound field, caused by a high ratio of reflected to direct sound energy. Leads to the impression of high diffusivity in case of stationary excitation" [1]. This definition suggests that this attribute refers to the perception of DRR. While this attribute could be addressed also from a monaural perspective, in binaural settings, it is likely related to concepts such as ASW and envelopment, or to the spaciousness, term used in concert hall acoustics [41]. Envelopment is defined here as the "sensation of being spatially surrounded by the reverberation" [1]. Disentangling this attribute from the ASW is not always easy. However, the ASW depends on the early reflections, while the envelopment may be more related to the late reverberation [42]. The level of reverberation and the envelopment are expected to correlate with the IC of the binaural signal. In [16], the envelopment is estimated from the IACC for the reverberant part of the signal, which is computed from the energy ratio between a temporal segment and the previous one. This model is selected but, despite a detailed description in their manuscript, its implementation is not publicly available.

Duration of reverberation Defined as "Duration of the reverberant decay. Well audible at the end of signals" [1]. A model introduced in [19] computes peaks and dips at the output of a peripheral model for each ear, aiming at segregating the direct sound and the reverberant field. The perceived reverberance is estimated by combining the reverberant part of the signals for each ear. Although the model output varied significantly as with the early decay time, they also correlated to the reverberation time (RT30). Its implementation is available online [3].

3.6 Time-behaviour-related attributes

Time-behaviour-related attributes refer to characteristics of auditory events in terms of their timing order, the perception of copies of the sound or speed of reproduction.

Pre- and post-echos: Defined as "copies of a sound with mostly lower loudness prior to / after the actually intended starting point of a sound" [1]. These echoes may potentially affect multiple attributes, such as localisation, ASW, timbral aspects, etc. [48]. Since some of the effects of these echoes are addressed in the specific attributes (e.g. localisation, ASW, timbral aspects...), we focus here on

[3] https://github.com/aosses-tue/binaural-auditory-model-RAA

the echo thresholds, i.e. the audibility of these echoes. A model of echo thresholds that includes masking was proposed in [20], but it is meant to analyse spatial room impulse responses. Thus, while this model might generally be useful for predicting the audibility of such pre- and post-echoes, it has no stage to predict if they are stronger or weaker compared to a reference.

Temporal disintegration: Defined as "Sound sources, which - by experience - should have a united temporal shape, appear temporally separated" [1]. To the best of our knowledge, no available model is suitable to estimate this attribute.

Crispness: Defined as "perception of the reproduction of transients" [1]. To the best of our knowledge, no available model is suitable to estimate this attribute. However, models of time-varying loudness that track instantaneous loudness, such as [21] (see Section 3.7) could provide evidence of differences in the perception of transients between a target and a reference.

Speed and sequence of events: Speed is defined as "identical in content and sound, but evolves faster or slower. Does not have to be accompanied by a change in pitch" [1]. The sequence of events is defined as the "order or occurrence of scene components" [1]. To the best of our knowledge, no available model is suitable to estimate these attributes.

3.7 Dynamics-related attributes

Dynamics-related attributes refer to the perceived sound level and how it varies over time.

Loudness, dynamic range and dynamic compression effects: Loudness is defined as the perceived volume of sound that increases from the threshold of hearing to the threshold of pain [30]. The dynamic range is defined in [1] as "amount of loudness differences between loud and soft passages". The dynamic compression effects are defined as "sound changes beyond the long-term loudness" [1].

The "Cambridge" loudness model has been constantly evolving since 1996, and multiple updates have been proposed to include new functionalities and account for new experimental data [49]. The current implementation of the model [21], available in [25], accounts for binaural, time-varying integration, and computes both long- and short-term loudness, being suitable to address the three attributes included in this section. Thus, loudness can be estimated as instantaneous or as long-term averaged, while the dynamic range can be computed as the range in which the short-term loudness varies. The dynamic compression effects are expected to be represented in short- and long-term loudness estimates.

3.8 General attributes

Clarity: Defined as "impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected" [1]. In room acoustics, this attribute has been used to measure how increasing reverberation affects the ability to differentiate among different instruments [41], but it is intended to be more general in the context of SAQI. The *traditional* clarity is often measured as the energy ratio between the early and the late parts of the sound, with the threshold often defined at 50 ms for speech and or 80 ms for music in concert halls (C_{50} and C_{80} respectively) [41]. To the best of the authors' knowledge, there is no model that addresses this attribute in the strict SAQI sense, and C_{50}/C_{80} may be the closest to quantify clarity, but from the room acoustics interpretation.

Speech intelligibility: Defined as "impression of how well the words of a speaker can be understood" [1]. Speech intelligibility is arguably one of the most important attributes, since it is central to human communication. This attribute is often measured by the rate of correct responses when reporting a heard letter, number, word or sentence. One of many models of speech intelligibility, introduced in 2010, has been continuously extended to account for different phenomena related to binaural hearing, such as better ear listening, binaural unmasking, etc. [22]. Different versions of the model include stages for normal-hearing or hearing-impaired listeners. The model is available in [25].

4. SUMMARY

This manuscript presents a focused literature review on available auditory models to cover most attributes in SAQI. The selected models were chosen as a compromise between easy to use and performance. Thus, it is possible that the review and the resulting list is not exhaustive, but a selection in accordance with the objectives of this work. Otherwise, these models provide a comprehensive benchmark for perceptual models and can be used to assess and/or optimise spatial audio algorithms.

5. ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/X032914/1, project Challenges in Immersive Audio Technology.

6. REFERENCES

[1] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acust. united Ac.*, vol. 100, no. 5, pp. 984–994, 2014.

[2] F. Rumsey, *Spatial audio*. Routledge, 2012.

[3] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, "Defining immersion: Literature review and implications for research on immersive audiovisual experiences," *Journal of Audio Engineering Society*, vol. 68, no. 6, pp. 404–417, 2019.

[4] M. A. Akeroyd and W. M. Whitmer, "Spatial hearing and hearing aids," *Hearing aids*, pp. 181–215, 2016.

[5] G. F. Greco, R. Merino-M., A. Osses, and S. C. Langer, "SQAT: A MATLAB-based toolbox for quantitative sound quality analysis," in *Inter-Noise Proceedings*, vol. 268, pp. 7172–7183, 2023.

[6] L. Chwif, M. R. P. Barretto, and R. J. Paul, "On simulation model complexity," in *Winter Simulation Conference Proceedings*, vol. 1, pp. 449–455, IEEE, 2000.

[7] B. Eurich, S. D. Ewert, M. Dietz, and T. Biberger, "A computationally efficient model for combined assessment of monaural and binaural audio quality," *J. Audio Eng. Soc.*, vol. 72, no. 9, pp. 536–551, 2024.

[8] T. McKenzie and F. Brinkmann, "Toward an improved auditory model for predicting binaural coloration," *J. Audio Eng. Soc.*, vol. 73, pp. 115–126, 2025.

[9] DIN-45692, "Measurement technique for the simulation of the auditory sensation of sharpness," 2009.

[10] P. Daniel and R. Weber, "Psychoacoustical roughness: Implementation of an optimized model," *Acta Acust. united Ac.*, vol. 83, no. 1, pp. 113–123, 1997.

[11] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, no. 3, pp. 679–688, 1982.

[12] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, IEEE, 2018.

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

[14] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802, 2014.

[15] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos, "Extracting sound-source-distance information from binaural signals," *The technology of binaural listening*, pp. 171–199, 2013.

[16] S. Klockgether and S. van de Par, "A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception," *Acta Acust. united Ac.*, vol. 100, no. 8, pp. 964–971, 2014.

[17] S. Li, R. Baumgartner, and J. Peissig, "Modeling perceived externalization of a static, lateral sound image," *Acta Acustica*, vol. 4, no. 5, p. 21, 2020.

[18] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A bayesian model for human directional localization of broadband static sound sources," *Acta Acustica*, vol. 7, p. 12, 2023.

[19] A. Osses Vecchi, A. Kohlrausch, W. Lachenmayr, and E. Mommertz, "Predicting the perceived reverberation in different room acoustic environments using a binaural auditory model," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. 381–387, 2017.

[20] T. Jüterbock, F. Brinkmann, H. Gamper, N. Raghuvanshi, and S. Weinzierl, "Spatio-temporal windowing for encoding perceptually salient early reflections in parametric spatial audio rendering," *J. Audio Eng. Soc.*, vol. 71, no. 10, pp. 664–678, 2023.

[21] B. C. Moore, M. Jervis, L. Harries, and J. Schlittenlacher, "Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition," *J. Acoust. Soc. Am.*, vol. 143, no. 3, pp. 1504–1513, 2018.

[22] M. Lavandier, T. Vicente, and L. Prud'homme, "A series of SNR-based speech intelligibility models in the Auditory Modeling Toolbox," *Acta Acustica*, vol. 6, p. 20, 2022.

[23] Y. Zheng, J. Yao, X. Deng, Y. Yang, R. Liao, W. Tu, and C. Lin, "HAPG-SAQAM: Human auditory perception guided spatial audio quality assessment metric," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[24] P. Manocha, A. Kumar, B. Xu, A. Menon, Gebru, V. K. Ithapu, and P. Calamia, "SAQAM: Spatial audio quality assessment metric," in *Interspeech*, pp. 649–653, 2022.

[25] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1. x: A toolbox for reproducible research in auditory modeling," *Acta Acustica*, vol. 6, p. 19, 2022.

[26] ANSI/ASA:S1.1-2013, *Acoustical Terminology*. Washington, DC, USA: American National Standards Institute, 2013.

[27] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing virtual sound source attributes using a binaural auditory model," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 203–217, 1999.

[28] R. D. Patterson, "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Am.*, vol. 96, no. 3, pp. 1409–1418, 1994.

[29] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2006.

[30] V. Pulkki and M. Karjalainen, *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, 2015.

[31] G. Greco, R. Merino-M., and A. Osses, "SQAT: A sound quality analysis toolbox for MATLAB," 2023.

[32] W. Aures, "Ein Berechnungsverfahren der Rauhigkeit," *Acta Acust. united Ac.*, vol. 58, pp. 268–281, 1985.

[33] H. Fastl, "Roughness and temporal masking patterns of sinusoidally amplitude modulated broadband noise," in *Psychophysics and physiology of hearing*, Academic press, 1977.

[34] W. Aures, "Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale," *Acta Acust. united Ac.*, vol. 59, no. 2, pp. 130–141, 1985.

[35] J. C. R. Licklider, "A duplex theory of pitch perception," *J. Acoust. Soc. Am.*, vol. 23, pp. 147–147, 1951.

[36] A. J. Oxenham, "Revisiting place and temporal theories of pitch," *Acoustical science and technology*, vol. 34, no. 6, pp. 388–396, 2013.

[37] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, IEEE, 2014.

[38] L. A. Jeffress, "A place theory of sound localization," *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[39] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, 1999.

[40] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature human behaviour*, vol. 6, no. 1, pp. 111–133, 2022.

[41] L. L. Beranek, *Concert halls and opera houses: music, acoustics, and architecture*, vol. 2. Springer, 2004.

[42] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient, lateral fraction, and apparent source width in concert halls," *J. Acoust. Soc. Am.*, vol. 104, pp. 255–265, 1998.

[43] S.-I. Sato and Y. Ando, "Apparent source width (asw) of complex noises in relation to the interaural cross-correlation function," *J. Temporal Design in Architecture and the Environment*, vol. 2, no. 1, p. 29, 2002.

[44] R. Mason and F. J. Rumsey, "A comparison of objective measurements for predicting selected subjective spatial attributes," in *Audio Eng. Soc. Conv. 112*, 2002.

[45] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound externalization: A review of recent research," *Trends in Hearing*, vol. 24, 2020.

[46] R. Baumgartner and P. Majdak, "Decision making in auditory externalization perception: model predictions for static conditions," *Acta Acustica*, vol. 5, p. 59, 2021.

[47] T. Okano, "Image shift caused by strong lateral reflections, and its relation to inter-aural cross correlation," *J. Acoust. Soc. Am*, vol. 108, no. 5, 2000.

[48] M. Barron, "The subjective effects of first reflections in concert halls—the need for lateral reflections," *J. Sound and Vibration*, vol. 15, pp. 475–494, 1971.

[49] B. C. Moore, "Development and current status of the "Cambridge" loudness models," *Trends in hearing*, vol. 18, 2014.