# COMPOSING THE EARS OF THE MACHINE: AUDIO SCORES, SELECTIVE LISTENING, AND CONCATENATIVE AI IMPROVISATION

**Ming Yang**

**Royal Northern College of Music (RNCM)**
nicholasyang2016@gmail.com

## ABSTRACT

This paper proposes an expanded notion of audio This paper proposes an expanded notion of audio score for AI-mediated improvisation, grounded in psychoacoustic theory and practice-based composition. Building on Louis d'Heudière's account of audio scoring with verbal instruc- tions and field recordings, it shifts the focus toward sound without instruction as score: sound that guides performers through selective listening. Drawing on auditory scene analysis and timbre perception, the paper argues that sound can function as a score because listeners extract perceptual dimensions such as brightness, roughness, envelope, and tonal stability and treat them as cues for musical action. It then extends this framework to machine listening, treating corpus-based concatenative systems as computational an- alogues of selective, feature-based listening, in which de- scriptors, weights, and distance models define how the ma- chine "hears" and which aspects of the input acquire pre- scriptive force. A case study of *Overlapping* for alto saxo- phone and live electronics examines how this idea is real- ized through two contrasting corpora, three designed sax- ophone sound types, and three microphones functioning as independent "ears," creating a looping audio score shared by human and algorithmic agents. The paper concludes by discussing the limitations of descriptor-based listening—its partial view of the sound object—and argues for composing with multiple, parallel listening modes as a way of articulating richer audio scores between humans and machines.

## 1. INTRODUCTION

In Louis d'Heudière's thesis, audio scoring is described as a practice in which sound functions as a medium of instruction, enabling performers to respond to recorded or spoken material rather than to symbolic notation [1]. Within his works, both field recordings and verbal instructions are treated as scores that guide performance: he characterises the field recording as a form of "stimulus" [1, p. 33] and allows both the recording and the verbalised instruction to appear in the concert situation as visible and audible agents [1, p. 33]. While d'Heudière explores the coexistence of verbal and sonic instructions, the present study shifts the focus toward sound without instruction as score—that is, toward the capacity of sound itself to establish communication between performers without the mediation of language. I take as a starting point her use of field recording as a guide to performance and extend it by asking what happens if we consider the influence of sound on performers' cognition and attention as a general model of scoring. If field recordings can function as scores by directing human perception, then an analogous principle could apply wherever sound is selectively perceived and acted upon.

On this basis, I propose a further extension of the audio score concept into the domain of AI-mediated improvisation, where machine listening, and corpus-based concatenative systems (CBCS) participate in the same communicative field as human musicians. This move aligns with Pierre Alexandre Tremblay's work on concatenative systems, which demonstrates that musical outcomes are shaped by the mapping of input descriptors to target descriptors and by their relative weights [2]. Loudness, spectral features, envelope, and other dimensions are extracted from the incoming sound and used to select or generate new material. Composition, in this view, resides not in predetermined notes but in the design of listening—in deciding which aspects of sound will be heard and acted upon. Following Tremblay, I understand machine listening as a process that extracts specific perceptual facets—brightness, roughness, gesture—and uses them to guide musical response.

The central question of this paper therefore becomes: how can audio scores operate when one of the listeners is an algorithmic system? In the proposed framework, both human improvisers and machines act through selective perception. Performers choose which elements of a sound to follow; the concatenative system does the same through weighted descriptors. The audio score is not only the sound presented to the algorithm; it is also the sound returned by the algorithm, which in turn influences the subsequent decisions of the performer. The score emerges from the multi-directional interaction between human improvisation and AI improvisation, understood as a continuous exchange of perceptual cues rather than as a one-way transmission of instructions.

## 2. LISTENING AS SELECTION: PSYCHOLACOUSTICS AND THE CONDISIONT OF AUDIO SCORE

### Selective listening during improvisation

As noted before, the fundamental question is: how can a sound, without verbal instruction, become a meaningful instruction for performers or systems? Sound can only function as a score if listeners do not hear all of it equally. Auditory perception is not a neutral recording of acoustic reality but a process of selective attention in which certain features become salient while others recede into the background. Work in auditory scene analysis has shown that the auditory system continuously groups and segregates components of complex sound mixtures into perceptually meaningful streams, based on features such as frequency content, temporal patterning, and spatial cues [3]. Psychoacoustic and timbre research further demonstrates that dimensions such as spectral brightness, roughness, temporal envelope, and tonal stability play different roles in how sounds are recognized, grouped, and remembered [4], [5]. A sharp attack can dominate the identification of an event, a noisy spectrum can attract attention, and a stable pitch can anchor perception, even when these attributes coexist within the same sound.

This selective nature of hearing is what allows sound to operate as instruction. When a performer hears a sound in an improvisational context, they do not respond to its total acoustic content but to those elements that stand out as perceptually or musically significant. A long-sustained tone may invite continuation, a burst of noise may call for contrast, and a bright timbre may suggest further spectral activity. In this way, sound becomes a field of affordances rather than a fixed object, guiding action through perceptual emphasis rather than symbolic encoding.

Psychoacoustic studies of timbre perception provide a formal basis for this idea. Timbre is not a single attribute but a multidimensional perceptual space structured by parameters such as brightness, roughness, attack character, and spectral flux [4]–[6]. Research on psychoacoustic sonification and auditory displays similarly identifies pitch, loudness, brightness, and roughness as separable perceptual dimensions that can be manipulated independently and interpreted reliably by listeners [7]. Listeners weigh these dimensions differently depending on context, task, and attention, resulting in variable judgments of similarity, salience, and grouping. Two sounds that are acoustically complex may be perceived as similar if they share a dominant dimension, or as different if they diverge along a salient axis.

### Psychoacoustic Evidence for Perceptual Dimension Extraction

Having established that an audio score operates through the selective perceptual salience of sound, the question arises: what empirical evidence supports the idea that listeners extract particular dimensions from sound and use them to guide action? Psychoacoustics provides a scientific foundation for understanding how listeners attend to different elements of sound. Decades of research in timbre and auditory perception have identified a set of core perceptual attributes—including loudness, roughness, brightness, tonalness, and temporal shape—those listeners treat as distinct sensations, even when these attributes arise from the same acoustic signal [5], [6]. In a series of experiments that aim to characterise perceptual dimensions of sound for auditory displays, researchers such as Tim Ziemer and colleagues have demonstrated that listeners can reliably extract such qualities from synthesized auditory stimuli and use them to perform cognitive and motor tasks [7], [8], [9].

In Ziemer's work on perceptual dimensions and psychoacoustic auditory displays, the focus is on designing mappings between data and sound such that listeners can extract intended information from specific perceptual features. In one set of studies, participants were presented with sounds that varied systematically along perceptually orthogonal dimensions—for example, changes in spectral centroid (related to brightness), modulation rate (related to perceived roughness), and pitch trajectory—and were asked to use these auditory cues to complete tasks such as judging category differences or navigating toward a target in a two-dimensional space [7], [8], [9]. Trajectory analyses indicate that users are able to analyse the sonification axis-by-axis and to integrate multiple sonified dimensions when approaching a target, showing that listeners can attend to and discriminate several auditory features simultaneously. The success of these experiments demonstrates that listeners do not treat complex sounds holistically or undifferentiated; rather, they hear perceptually salient dimensions and can isolate and act upon these dimensions when they carry meaningful information.

This line of research parallels classic psychoacoustic studies on timbre perception, which show that listeners do not perceive composite sounds as monolithic wholes but as configurations of distinct attributes. Multidimensional scaling of dissimilarity ratings between instrument tones reveals that changes in spectral centroid strongly influence perceived brightness, largely independently of changes in pitch or overall level [10]. Similarly, roughness—the perceptual attribute associated with beating or closely spaced partials—has been shown to contribute to judgments of musical tension and expressive intensity[11]. Recent overviews of timbre research describe the auditory system as organising sounds within a multidimensional space structured by such attributes [6], [12]. Taken together, these findings underscore a central principle of auditory perception: the human auditory system is a multidimensional analyser that parses sound into salient components rather than integrating all of its acoustic features equally.

Importantly for the present study, Ziemer's work and related psychoacoustic experiments show not only those perceptual dimensions exist, but that listeners can use those dimensions to perform tasks that have a procedural or decision-making character. Whether the task is distinguishing categories, matching auditory cues to data dimensions, or navigating toward a spatial target on the basis of perceptual differences, listeners demonstrate that they can extract and selectively attend to perceptual axes embedded

within sound [7], [8], [9]. This is not a passive reception of sound; it is a form of perceptually guided action.

From the perspective of audio-score theory, this psychoacoustic evidence provides a necessary foundation: if listeners can extract and act upon specific perceptual dimensions, then sound can function as a score to the extent that it carries structured, attended information along those dimensions. In the context of improvisation, this means that performers can treat particular aspects of sound as musical cues—to continue, to contrast, or to transform—based on the perceptual weight those dimensions carry for them. The next section examines how this capacity for selective perception underpins interaction in improvisational contexts and lays the groundwork for understanding how machine listening can emulate analogous processes.

### Sound as a Score

The discussion above suggests that sound can function as a score only to the extent that it is selectively heard. In improvisation, this selectivity becomes a core compositional force. Accounts of ensemble practice repeatedly emphasise that listening and attentional focus are central organising principles of musical interaction: performers attend to salient cues in timbre, rhythm, and dynamics, and structure their responses accordingly [13]–[15]. Although they share the same sonic environment, musicians may attend to different aspects of what they hear, leading to different musical responses. This process is grounded not only in individual taste or style, but in the perceptual architecture of hearing itself, as described by auditory scene analysis and timbre research [3], [5].

From this perspective, an audio score is not a fixed sonic object that stands outside performance, but a pattern of selective attention that arises within it. A given sound contains many potential cues; it becomes a score when certain of those cues are treated as instructions. A sustained tone may function as a score for continuity, a noisy attack as a score for rupture, a particular spectral colour as a score for timbral imitation. What the score "is" depends on which dimensions of the sound are foregrounded in listening and how strongly they are weighted in guiding response. The score is thus embedded in the act of listening, rather than inscribed only in notation or recorded media.

This framing also opens a path toward AI-mediated improvisation. If audio scores operate through selective perception, then any system—human or algorithmic—that extracts and weights perceptual dimensions can, in principle, participate in audio scoring. Machine listening models, such as those used in concatenative synthesis, implement their own forms of selective hearing by analysing sound into descriptors and evaluating similarity in a feature space. When such a system is coupled to human improvisers, the sound they produce becomes a score for the machine, and the machine's output, in turn, becomes a score for the performers. Audio scoring then takes the form of a continuous exchange of perceptual cues, in which human and algorithmic listeners guide one another over time.

In this sense, the audio score is no longer a one-way instruction from composer to performer, but a dynamic communicative process that unfolds through reciprocal listening. The composer's role shifts toward designing the conditions under which this reciprocal audio scoring can occur: choosing which aspects of sound will be analysable, how strongly they will be weighted, and how they will be made audible to both human and machine. The following chapter develops this idea by treating concatenative machine listening as a computational analogue of selective, feature-based listening, and by examining how such systems can be composed to act as partners in improvisation rather than as passive processors of sound.

## 3. MACHINE LISTENING AS AUDIO-SCORE INTERPRETATION

### Machine Listening as Perceptual Filtering

The previous section argued that an audio score does not reside in a sound object alone, but in the *perceptual filtering* enacted by listeners: certain dimensions of sound are treated as significant, others recede, and musical action follows those perceptual weights. In this section, I extend that argument from human listeners to machine listeners, asking how similar logics of selection and salience operate within algorithmic systems.

Machine listening systems do not perceive sound holistically. They extract and weight descriptors corresponding to timbral, temporal, and energetic dimensions. In music information retrieval and computer audition, such descriptors are explicitly motivated by psychoacoustic and timbral research and are intended to approximate perceptually relevant aspects of sound [16], [17]. Corpus-based concatenative synthesis, for example, builds on databases of segmented and descriptor-analysed sounds, selecting units according to proximity to a target position in descriptor space. This process constitutes a form of perceptual filtering analogous to selective listening in human improvisation. Rather than treating the audio signal as a continuous physical waveform, machine-listening models transform sound into a set of features—such as MFCCs, spectral centroid, temporal envelope, and energy—that approximate how humans experience qualities like brightness, gesture, and intensity [17]–[20]. What the system "hears" is therefore not sound itself, but a structured representation of perceptual attributes.

The choice of descriptors determines which aspects of the sound are considered meaningful and which are ignored. A system that privileges MFCCs foregrounds timbral colour; one that emphasises envelope descriptors foregrounds gesture and articulation; a model centred on energy responds primarily to dynamics and salience. Work on descriptor sets for timbre analysis has shown that different combinations of features capture different facets of perceived timbre and are more or less effective depending on the task [18], [21]. Machine listening is thus already an interpretive act: it defines in advance what counts as musically or perceptually relevant within the sonic field. In this sense, the architecture of the listening model functions as a hidden score that shapes all subsequent musical decisions.

From the perspective of audio-score theory, this filtering process closely mirrors the way human performers attend selectively to aspects of sound in improvisation. Just as an improviser may follow brightness rather than pitch, or density rather than rhythm, a machine-listening system follows the dimensions encoded in its feature set. In CBCS such as CataRT, navigation through the corpus is driven by distances in descriptor space, making different descriptors and weightings directly audible as different "ways of listening" to the same material [18], [22]. The system's "attention" is distributed according to mathematical weights rather than embodied intuition, yet the structural logic is comparable: both humans and machines respond not to the totality of sound, but to those dimensions that have been marked—by design or by habit—as salient.

### Composing Listening: Machine Audition as Audio Score

If machine listening operates as a form of perceptual filtering, then the role of the composer in AI improvisation shifts from writing sounds to designing how sounds will be heard. In concatenative and corpus-based systems, the composer determines which descriptors are extracted, how these descriptors are weighted, and how similarity is evaluated. These decisions define the system's mode of attention and therefore its musical behaviour. Composition becomes the construction of an auditory perspective rather than the specification of musical events [18].

This shift is clearly articulated in Pierre Alexandre Tremblay's work with interactive concatenative systems such as *Sandbox*. In "Surfing the Waves: Live Audio Mosaicing of an Electric Bass Performance as a Corpus Browsing Interface," Tremblay and Schwarz describe how an electric bass performance is analysed in real time and mapped onto a corpus of samples using CBCS, with "multi-dimensional mapping" and descriptor weighting central to expressive control [2]. The machine does not simply reproduce fragments from a database; it responds to the performer's sound through a descriptor-based representation that maps multidimensional aspects of performance onto a corpus. I argue that this descriptor-based representation fucntions as a perceptual model that privileges ceertain timbral or gestural aspects. Because matching is determined by descriptor selection and weighting, different weighting schemes produce different relationships between live input and corpus response. Tremblay and Schwarz note that both the choice of descriptors and their relative weights are crucial to the perceived responsiveness and expressivity of the system. The composer, in this context, writes the listening strategy that governs the dialogue between human and machine.

Such practices reveal that in AI improvisation the audio score is not located in a fixed sonic object but in the configuration of machine audition. The performer's sound functions as a score only because the system has been prepared to hear it in particular ways. Tremblay's approach demonstrates that altering descriptor weights can transform the same input into radically different musical outcomes, much as two human improvisers might respond differently depending on whether they attend to rhythm, timbre, or dynamics [2]. The compositional act lies in choosing which of these dimensions will carry prescriptive force.

Similar ideas appear in the work of other composers employing machine listening. Aaron Einbond treats real-time audio analysis and corpus-based matching as a means of articulating timbral relationships that could not be specified through notation alone, using descriptor spaces as "timbre spaces" for analysis and composition [23], [24]. Rodrigo Constanzo's *C-C-Combine*, a corpus-based audio mosaicking instrument in Max/MSP, similarly behaves as a responsive partner whose actions are guided by feature-based similarity between live input and pre-analysed corpus grains; Constanzo explicitly highlights the expressive role of feature weighting in shaping the matching process. Across these practices, the machine is conceived not as an automated performer of prewritten material but as a listener whose perceptual orientation has been composed in advance.

Understanding machine listening in this way allows the concept of audio score to be extended beyond human performers. The input sound becomes a prescriptive signal that activates a designed mode of hearing within the system; the corpus provides the potential vocabulary; and the weighting scheme functions as the score that mediates between them. Composition, therefore, consists in shaping the conditions under which sound will influence sound: in short, the composer writes the ears of the machine. This perspective forms the basis for the case study that follows. In my own work with concatenative synthesis, different listening modes were constructed by altering descriptor sets and weights, enabling the same input to generate contrasting musical behaviours. The piece does not begin with a fixed sonic plan but with a designed auditory perspective through which the machine interprets the audio score presented by human performers.

## 4. CASE STUDY

### *OVERLAPPING* (2025-26)—A CONCATENATIVE AUDIO SCORE FOR ALTO SAXPHONE AND ELETRONICS

The preceding sections have proposed that in AI improvisation the audio score is constituted not by a fixed sonic object but by a designed mode of listening. Machine audition—through its descriptors, weights, and distance models—functions as an interpreter that decides which aspects of sound will acquire prescriptive force. The present case study examines how this concept is put into practice in my composition *Overlapping* (2025-26) for alto saxophone and live electronics, in which a machine-listening system built in Max/MSP interacts with the performer through multiple, parallel listening perspectives.

In this piece, audio scoring is realised as a reciprocal process: the saxophonist's sound functions as a score for the machine, and the machine's responses in turn function as a score for the performer. Rather than following a fixed timeline or notated structure, both agents rely on selective

listening. The system analyses the incoming sound into descriptors and generates output according to its current listening mode, while the performer shapes their playing in response to the evolving electronic texture. *Overlapping* thus serves as a practice-based exploration of how listening perception—human and algorithmic—can be organised so that sound itself becomes the medium through which human and AI improvisers guide one another over time.

## Listening Modes and Empirical Observations

As mentioned before, the core assumption of concatenative systems is that similarity in perceptual feature space corresponds to meaningful musical continuity: audio fragments that are "close" in descriptor space produce coherent sequences, and those that are "far" create contrast or collage. In practice, however, this continuity depends not only on the incoming sound but also on the structure of the corpus against which it is matched. To explore this in relation to *Overlapping*, I conducted a series of informal tests using a common input—live alto saxophone—and two contrasting corpora: one built from field recordings, and one built from instrumental sounds.

### Field-recording corpus

The first corpus consisted primarily of field recordings with complex, often diffuse spectral and textural characteristics. When saxophone material was used as input, the system tended to respond most convincingly to noise-based playing: breath noise, key clicks, inharmonic multiphonics, and other extended techniques with broad-band spectra. In these cases, the matching algorithm found neighbours in the field corpus that echoed the gestural envelope and rhythmic articulation of the saxophone while maintaining an environmental or textural character. The result was a concatenated output that, while not pitch-related, followed the shape of the input gesture and embedded it within an environmental texture.

By contrast, clearly pitched saxophone sounds—sustained tones or articulated melodic figures—matched less convincingly in this corpus. The strong periodic structure, clear harmonic content, and discrete pitch events of the saxophone did not find close analogues in the more diffuse, multi-source field recordings. The system still produced output, but the relationship between input and corpus felt weaker: the matches tended to ignore pitch and instead latch onto coarse spectral or energy similarities, which made the connection between performer and electronics less perceptible.

### Instrumental corpus

The second corpus was built from instrumental sounds, including saxophone and other pitched sources, with relatively clear envelopes and harmonic or quasi-harmonic spectra. With this corpus, both noise-based and pitched saxophone materials produced coherent results, but in different ways. Noise gestures matched to noisy or spectrally dense instrumental sounds, again preserving envelope and rhythmic profile. Pitched gestures, however, now found neighbours with comparable spectral centroid trajectories, partial structures, and onset patterns, resulting in concatenated sequences that reflected not only the input's gesture but also its pitch-related structure.

In this instrumental corpus, the system's responses were more clearly "musical" in the sense of perceptual coherence: listeners could more easily hear how the electronics followed or transformed the performer's lines. Noise materials yielded textural and gestural continuity; pitched materials yielded timbral and quasi-melodic continuity.

### Interpretation and implications

These observations suggest that concatenative listening is not neutral: it privileges those acoustic patterns that are both distinct in feature space and well represented in the corpus. When the corpus consists of field recordings, the system is particularly sensitive to broad-band, noisy, and textural input, and tends to disregard fine-grained pitch information. When the corpus is instrumental, the same input material can support both gestural and pitch-related continuity.

From an audio-score perspective, this behaviour aligns with the idea that a score operates through perceptual salience. The field-recording corpus lacks the kind of stable, discrete pitch events that improvisers often use as structural cues, so the machine listener treats the entire field primarily as a texture region. The instrumental corpus, by contrast, contains discrete, articulated sound objects that the system can cluster in feature space in ways that are more directly audible as correspondence with the saxophone's playing.

At the same time, the situation is more complex than a simple field-versus-instrument dichotomy: segmentation, descriptor choice, and corpus diversity all influence the resulting matches. This paper does not aim to resolve these technical variables in detail. Rather, it takes these empirical observations as compositional information: they inform the selection and combination of corpora and listening modes used in *Overlapping*, so that the resulting machine listening supports an unfolding musical interaction rather than merely demonstrating a technique.

## Sound Types and Performer Agency

Because *Overlapping* works with two contrasting corpora—one built from field recordings and one from instrumental sounds—the performer's agency is articulated through a set of three sound types designed to probe how each corpus responds. These types are defined within a two-dimensional space whose x-axis represents temporal character (from static to gestural) and whose y-axis represents spectral character (from noise-based to pitched). Rather than treating the saxophone as a single source, the composition invites the performer to move within this coordinate space (Fig. 1), occupying regions associated with sustained static sounds, rhythmically articulated gestures, and fragmented static textures. Each region is perceived differently by the concatenative listener and therefore functions as a distinct audio score.

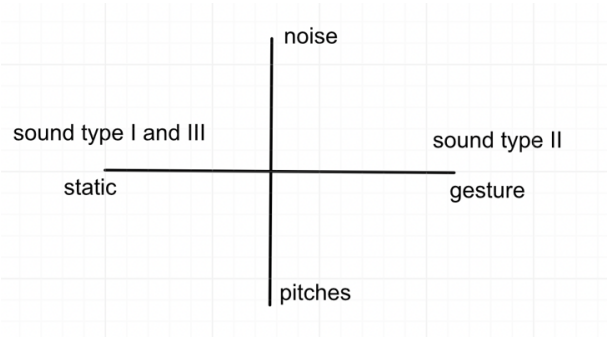


Figure.1. The coordinate for three types of input sound by alto Saxhpone.

**1. Static sustained sounds (Type I)**

The first sound type consists of sustained, relatively static sounds: long pitched tones, airy breath sounds, or stable multiphonics. These sounds produce slowly changing descriptor values over time. When used as input, they tend to elicit outputs that preserve long envelopes and gradual spectral motion. With the instrumental corpus, sustained pitches encourage colour-matching responses, in which the system follows the spectral shape of the saxophone and produces quasi-harmonic continuations. With the field-recording corpus, sustained airy or noisy tones match more readily, and the system embeds the saxophone's sound within broader environmental textures. In both cases, the static temporal profile makes this type a score for continuity, while the choice between pitch and noise steers whether that continuity is timbral/melodic (instrumental corpus, pitched) or textural (field corpus, noisy/airy).

**2. Gestural rhythmic sounds (Type II)**

The second sound type is gestural and rhythmically articulated: tongued figures, slap attacks, key clicks, and rapid sequences that create a strong sense of pulse or kinetic energy (Fig.2). These gestures generate pronounced changes in onset, envelope, and short-term energy, making temporal descriptors highly salient. The concatenator responds to these variations with fragmented or event-like sequences that closely mirror the gesture's contour. With the instrumental corpus, pitched gestures often result in articulated, quasi-melodic mosaics that track both rhythm and rough pitch shape; with the field corpus, noisy or percussive gestures tend to produce bursts of environmental or textural material aligned to the same rhythmic pattern. Type II thus functions as a score for articulation and dynamism, inviting the system to track the morphology of the performer's gesture more than its sustained spectral colour.

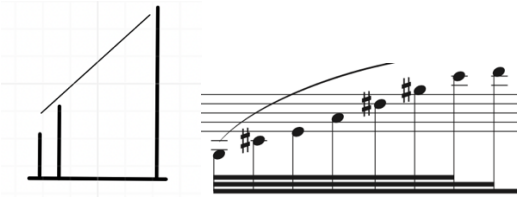

Figure.2. Draft of sound type II.

**3. Fragmented static textures (Type III)**

The third sound type occupies an intermediate position: it is fragmented yet globally static. Here the performer uses short, vivid sound objects—isolated multiphonics, repeated key clicks, small pitch inflections—arranged in a way that does not build toward a clear rhythmic line but rather produces a flickering, pointillistic texture (Fig.3). Temporally, this type is less continuous than Type I but less directional than Type II; spectrally, it can be realised as either pitched or noisy material. For the concatenator, these inputs often produce outputs that echo the local micro-gestures while maintaining an overall sense of textural stasis. In the instrumental corpus, the system tends to pick out clusters of similar instrumental fragments; in the field corpus, it responds with small grains of environmental sound that punctuate the texture. Type III therefore shares with Type I a broadly static temporal quality, but its internal fragmentation makes it a score for textural detail rather than sustained continuity.

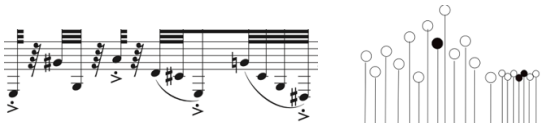

Fig.3. Draft of sound type III.

**Noise and pitch as perceptual control**

Across all three sound types, the performer also controls the spectral character of the input by moving between pitched and noise-based playing. Pitched sounds—conventional tones, microtonal inflections, stable multiphonics—provide periodic spectral organisation that is captured by descriptors such as spectral centroid and MFCC patterns, inviting responses that emphasise timbral resemblance and quasi-melodic continuity, especially in the instrumental corpus. Noise-based sounds—breath noise, inharmonic multiphonics, key clicks—foreground roughness and spectral diffusion, activating regions of the descriptor space that favour textural collage and contrast, particularly in the field-recording corpus.

By combining these three sound types with the pitch/noise distinction, the performer can intentionally steer the electronics. A sustained pitch (Type I) tends to generate elongated, colour-matching responses in the instrumental corpus; a gestural pitch (Type II) produces articulated, quasi-melodic fragments; a sustained noise (Type I realised noisily) encourages dense environmental

textures from the field corpus; and a fragmented noise texture (Type III) triggers pointillistic collage. The saxophonist therefore composes the audio score in real time by choosing not only *what* to play, but *how* it unfolds temporally and spectrally in relation to the two corpora.

### Three Microphones as Three Ears

To open multiple pathways of interaction between performer and system, *Overlapping* employs three microphones, each routed to an independent concatenative process in Max/MSP. The three saxophone sound types described above—static sustained sounds, gestural rhythmic sounds, and fragmented static textures—are not tied to specific microphones in a fixed way; instead, the performer chooses in the moment which sound type to send to which input. In this sense, the microphones function less as technical channels and more as three potential ears for the machine listener. Figure 4 presents an exerpt of the sight-score for the improviser, outlining a possible unfolding of the three sound types in time and space. The vertical axis, read from bottom to top, functions as a timeline, while the three microphones are arranged from left to right to mirror their physical positions on stage. In this way, the diagram maps the performer's movement between microphones onto both temporal progression and spatial orientation during the performance.

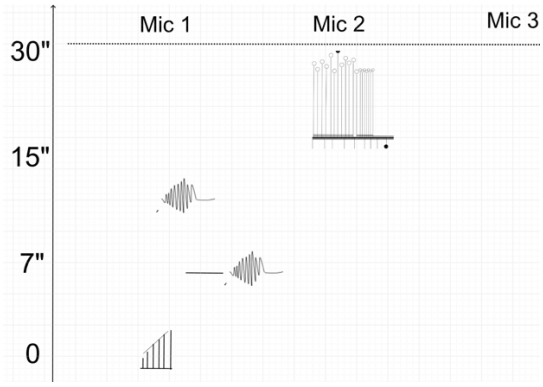


Figure 4. Sight-score for the idea of the improvisation.

Crucially, these ears are not defined by simple labels such as "loudness channel" or "spectral channel." Although the underlying concatenative processes share a similar analytical structure, each microphone–corpus–descriptor combination tends to produce a characteristic response. Over the course of rehearsal and performance, these responses become recognisable to the performer: one microphone might habitually yield smoother, more continuous textures, another might produce fragmented or noisy collage, and a third might oscillate between the two depending on the input. The distinction between the three ears is therefore experiential rather than pre-defined. Each microphone becomes a different audio score because it triggers a different way of listening on the part of the machine.

The compositional design of *Overlapping* seeks to translate the theoretical idea of audio score into a concrete performance situation in which human and machine listeners influence one another. Rather than providing the saxophonist with a traditional notated score, I offer a sight-score of instructions (Figure 4) that describes how sound types are to be produced and how they may be routed to the three microphones. The written material does not prescribe musical results; it prescribes modes of sounding and routing that will function as scores for the machine listeners.

In performance, the improviser chooses among these microphones according to the evolving dialogue with the electronics. A sustained tone might be sent to the ear that tends to produce continuity, while a noisy gesture might be directed to the ear that usually provokes rupture. Through this exploratory process the musician learns how each ear tends to "answer" and adapts their playing accordingly. The machine, in turn, responds to the performer's choices with new sonic material that feeds back into the performer's listening and decision-making. Audio scoring thus becomes a looped process: the saxophone's sound scores the concatenators, the concatenators' output scores the performer, and both adjust their behaviour in response.

This design emphasises that in AI improvisation the score is not fully determined in advance. The microphones provide entry points into the system, but their musical meaning arises only through interaction. Feedback from the concatenators guides the performer toward new sound choices, and these choices in turn redefine what the machine hears. The audio score is therefore an emergent relation among performer, microphones, and machine listening rather than a fixed mapping of descriptors.

## 5. CONCLUSIONS

This paper has proposed a transformation of the concept of audio score from an instruction toward pure sound filtering[1]. Building on practices articulated by Louise d'Heudière, where sound functions as an instruction for performers, I have argued that in AI-mediated improvisation sound can become a score not by representing music, but by actively shaping how music is heard and generated. The audio score, in this sense, is neither a notation nor a fixed object, but a dynamic relationship between sound, listener, and system.

Psychoacoustic research demonstrates that listeners do not perceive sound as an undifferentiated whole but extract salient dimensions such as brightness, roughness, tonalness, and gesture. Improvisation operates through this selective attention: performers choose which aspects of the sonic environment will guide their responses. I have suggested that concatenative machine listening embodies an analogous process. By analysing sound into perceptual descriptors and weighting these descriptors within a feature space, the system performs a form of geometric listening in which similarity, continuity, and collage emerge from computational judgments of perceptual proximity.

[1] Here the pure sound indicates the sound of non-instruction.

The case study of *Overlapping* for alto saxophone and electronics illustrates how this idea can be realised compositionally. Through the use of multiple corpora, three sound types, and three microphones routed to separate concatenative processes, the piece constructs several parallel "ears" that interpret the performer's sound in distinct ways. The written instruction functions only as a metascore; the true score is the sounding interaction between saxophonist and concatenators. In performance, both the acoustic instrument and the electronic responses are audible to the audience, making the audio score perceptible as an evolving dialogue.

At the same time, the case study reveals important limitations of concatenative machine listening as an audio-score interpreter. While the system can respond convincingly to particular aspects of the saxophone sound—such as envelope shape, spectral colour, or energy—it cannot apprehend the sound object as a whole. Each concatenator listens through a restricted set of descriptors, and what lies outside those descriptors remains inaudible to the machine. In practice, no single analytical pathway was sufficient to represent the richness of the instrument; the three-microphone setup emerged from the necessity to distribute listening across multiple, partial perspectives. The audio score interpreted by the system is therefore not the sound itself, but a projection of that sound onto chosen analytical frames.

Recognising this constraint is part of composing with machine listening, but it also becomes a creative resource. Because each object hears differently, the performer can navigate among these partial perspectives, treating them as alternative scores. The dialogue between human and machine thus exposes the constructed nature of listening: neither the saxophonist nor the concatenator possesses a complete understanding of the sound, yet their interaction generates musical form. Composition becomes the design of listening conditions rather than the prescription of musical events. To compose with concatenative systems is to compose the ears of the machine—to decide which dimensions of sound will acquire authority and how these dimensions will influence further sound.

Future work might explore more flexible listening models in which descriptors are dynamically reweighted or combined, allowing the machine to shift its attention as human improvisers do. Hybrid systems that integrate learning-based representations with concatenative methods may also offer richer interpretations of audio scores, and ensembles could engage with multiple machine listeners simultaneously, each offering a different reading of the same performance. More broadly, understanding audio scores as forms of perceptual agency invites a reconsideration of what it means to write music in an age where machines listen. In this context, sound becomes not simply the end of composition, but its beginning.

## 6. REFERENCES


[1] L. P. M. L. d'Heudières, "Scores in Time: Exploring the Use of Sound as a Medium for Notation in Musical Composition," Ph.D. dissertation, Bath Spa Univ., Bath, U.K., 2020.

[2] P. A. Tremblay and D. Schwarz, "Surfing the Waves: Live Audio Mosaicing of an Electric Bass Performance as a Corpus Browsing Interface," in Proc. Int. Conf. New Interfaces for Musical Expression (NIME), Sydney, Australia, 2010, pp. 447–450.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.

[4] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am., vol. 61, no. 5, pp. 1270–1277, May 1977, doi: 10.1121/1.381428.

[5] S. McAdams, "The perceptual representation of timbre," in *Timbre: Acoustics, Perception, and Cognition*, K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds., *Springer Handbook of Auditory Research*, vol. 69. Cham, Switzerland: Springer, 2019, pp. 23–57, doi: 10.1007/978-3-030-14832-4_2.

[6] K. Siedenburg and S. McAdams, "Four distinctions for the auditory 'wastebasket' of timbre," Frontiers in Psychology, vol. 8, Art. no. 1747, 2017, doi: 10.3389/fpsyg.2017.01747.

[7] T. Ziemer, "A psychoacoustic auditory display for navigation," in Proc. Int. Conf. Auditory Display (ICAD), 2018.

[8] T. Ziemer and H. Schultheis, "Three orthogonal dimensions for psychoacoustic sonification," *arXiv preprint* arXiv:1912.00766, 2019.

[9] T. Ziemer, D. Black, and H. Schultheis, "Psychoacoustic sonification design for navigation in surgical interventions," in Proc. Meetings on Acoustics, vol. 30, no. 1. Melville, NY, USA: Acoustical Society of America, 2017.

[10] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?" Acta Acustica united with Acustica, vol. 92, no. 5, pp. 820–825, 2006.

[11] M. M. Farbood, "A parametric, temporal model of musical tension," Music Perception, vol. 29, no. 4, pp. 387–428, 2012.

[12] K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, Eds., Timbre: Acoustics, Perception, and Cognition. Cham, Switzerland: Springer International Publishing, 2019.

[13] P. Oliveros, Deep Listening: A Composer's Sound Practice. New York, NY, USA: iUniverse, 2005.

[14] G. E. Lewis, "Interacting with latter-day musical automata," Contemporary Music Review, vol. 18, no. 3, pp. 99–112, 1999.

[15] I. Monson, Saying Something: Jazz Improvisation and Interaction. Chicago, IL, USA: University of Chicago Press, 2009.

[16] X. Serra, “Musical sound modeling with sinusoids plus noise,” in Musical Signal Processing, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Lisse, The Netherlands: Swets & Zeitlinger, 1997, pp. 91–122.

[17] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, “The Timbre Toolbox: Extracting audio descriptors from musical signals,” J. Acoust. Soc. Am., vol. 130, no. 5, pp. 2902–2916, 2011.

[18] D. Schwarz, “Corpus-based concatenative synthesis,” IEEE Signal Process. Mag., vol. 24, no. 2, pp. 92–104, 2007.

[19] D. Schwarz, “Concatenative sound synthesis: The early years,” *J. New Music Res.*, vol. 35, no. 1, pp. 3–22, 2006.

[20] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in Proc. Int. Symp. Music Inf. Retrieval (ISMIR), 2000.

[21] I. Czedik-Eysenberg and C. Reuter, “Evaluating timbre-related audio descriptors across different libraries and multimodal embeddings,” in *Proc. DAS | DAGA 2025: 51st Annual Meeting on Acoustics*, 2025, pp. 171–174.

[22] A. Einbond, C. Trapani, A. Agostini, D. Ghisi, and D. Schwarz, "Fine-tuned control of concatenative synthesis with CATART using the BACH library for MAX," in Proc. Int. Comput. Music Conf. (ICMC), Athens, Greece, Sep. 2014.

[23] A. Einbond, T. Carpentier, D. Schwarz, and J. Bresson,"Embodying spatial sound synthesis with AI in two compositions for instruments and 3D electronics," Comput. Music J., vol. 46, no. 4, pp. 43–61, 2022, doi: 10.1162/comj_a_00664.

[24] A. Einbond, J. Bresson, D. Schwarz, and T. Carpentier, “Instrumental radiation patterns as models for corpus-based spatial sound synthesis: *Cosmologies* for piano and 3D electronics,” in *Proc. Int. Comput. Music Conf. (ICMC)*, San Francisco, CA, USA, 2021, pp. 148–153.