

AI Stem Separation Tools in the Aural Skills Classroom

FRED HOSKEN

Recent advancements in artificial intelligence and machine learning, particularly AI-powered stem separation tools, offer transformative potential for aural skills education. These tools can isolate individual audio components, such as vocals, bass, drums, and harmonic instruments, from mixed tracks, providing educators and students with novel ways to engage with complex musical textures. One of the key challenges in aural skills training is developing the ability to parse polyphonic textures, a skill known as auditory streaming. AI stem separation tools enable students to focus on individual musical elements, enhancing their ability to isolate and analyze components of multi-voice performances. This paper explores how separators can scaffold auditory streaming development and improve other aspects of aural skills instruction. It also addresses the ethical implications and responsibilities of using such technologies in the classroom. Through case studies, this work demonstrates how AI-driven tools can enrich both music instruction and student learning, while fostering awareness of technology's broader impact on pedagogy.



Introduction

Recent advancements in artificial intelligence have the potential to significantly transform the way we teach music at the collegiate and high-school level. One of the most notable areas of development in artificial intelligence and machine learning is the rapid proliferation of AI-powered stem separation tools, technologies capable of isolating individual audio components (such as vocals, drums, and harmonic instruments) from a mixed musical track. These tools have become increasingly accessible, with many now available at little or no cost and requiring no technical knowledge, drastically lowering the barriers to entry for both educators and students. As a result, we are currently at a unique moment in which the integration of AI into the aural skills classroom in particular offers unprecedented opportunities for innovation in teaching and learning. By capitalizing on these technological advancements, instructors can offer students novel, engaging ways to develop their core aural skills.

One of the most complex challenges in aural skills education is developing students' ability to parse polyphonic textures into their constituent parts, to "hear through" a texture and "get inside" a multi-voice performance. This skill, known as auditory streaming, requires the ability to mentally separate overlapping sound sources in a

musical texture, a process that is often instinctive yet difficult to articulate or train. AI stem separation tools provide a new avenue for addressing this challenge, allowing students to isolate and focus on individual musical elements, facilitating more precise listening and analysis. However, alongside these exciting possibilities comes the responsibility to model the responsible and ethical use of technology in the classroom. As educators, we must ensure that students are not only developing technical skills but also cultivating an awareness of the ethical implications and limitations of AI in music (Airaj 2024).

This paper will explore ways to “scaffold” the development of auditory streaming skills through the use of AI stem separation tools, providing temporary support to students as they learn new concepts or skills through simplification, and gradually reducing that support as students become more proficient. In addition to auditory streaming, I will provide other examples of areas in which AI-driven stem separation can enhance instruction. Through these case studies, I will demonstrate the diverse ways in which AI can support both instruction and learning in music education, while also addressing the broader implications of integrating such technologies into pedagogical practices.

Auditory Stream Analysis

An advanced aural skill we aim to develop in our students is the ability to hear within complex musical textures, enabling them to actively engage with countermelodies, harmony, voice leading, orchestration, and more. Building this aural skill has musical and professional benefits for our students. For example, many of our students go on to work with ensembles, whether as performers or as directors, and the ability to identify with precision where a tuning or phrasing concern lies within the polyphonic texture allows for effective feedback and quick resolution. This is a challenging skill to cultivate, particularly given the “high-voice superiority effect,” which is a systematic perceptual bias towards processing the highest voice in multi-voiced music (Ishida et al. 2024; Marie and Trainor 2012; Marie and Trainor 2014; Trainor et al. 2014). Additionally, the physical characteristics of sound and our auditory anatomy – where higher pitches tend to mask lower ones (Huron 2016, 49–51) – pose additional obstacles. Furthermore, our students’ lived experiences of music, including the music they perform as soloists, the music they listen to daily, and the activities we engage in within the classroom, often heavily emphasize listening to the top voice. As a result of this high-voice superiority bias, we are asking students to work against thousands

of hours of surface-level listening when we try and move their attentional spotlight inside the musical texture.

This challenge of parsing complex auditory textures can be better understood through the lens of *Auditory Scene Analysis* (ASA), a theoretical framework developed by Albert Bregman (1990) to explain how the human brain organizes and interprets complex auditory information. ASA provides insight into how listeners separate and identify sounds from different sources, a process that is essential when navigating polyphonic music. Drawing on and extending principles from Gestalt psychology, Bregman's theory proposes that our auditory system breaks down sounds into distinct "auditory objects" using mechanisms like frequency grouping, temporal coherence, and spatial localization. By applying the principles of ASA, we can better understand the cognitive processes involved in developing the skills needed to navigate dense musical textures and train our students to identify and isolate the components of a complex auditory scene.

More specifically, the element of ASA we are most concerned with is auditory stream segregation or "streaming." This refers to the ability of the auditory system to separate sounds into distinct perceptual streams based on various auditory cues. These cues include differences in frequency, pitch, timbre, amplitude, and spatial location, as well as the rate and abruptness of changes in these properties over time. Relatedly, what we are often working against is the tendency towards "perceptual fusion" of simultaneous components of sound, forming a single perceived sound instead of, say, independent string instruments in a quartet. This is particularly challenging when multiple sounds can be heard as harmonics of the same fundamental. Due to the "harmonicity principle" (Bregman 1990, 232; see also Dewitt and Crowder 1987), the pitches are likely to be fused or vertically grouped and heard as a single sound object. Unfortunately for the task of streaming, a large amount of the core harmonic language of many musical styles encourages this fusion as (putting aside the compromises of equal temperament tuning) major chords can be heard as the fundamental + the fifth harmonic (the major third) + the third harmonic (the perfect fifth) and so they are readily fused into a single object.

In Bregman's book, he discusses streaming in musical contexts, though his focus is more on how to make instruments/voices stand out from the rest of the ensemble by using melodic and rhythmic independence (i.e., playing different pitches and having different onset and offset timings, even by as few as 30-50 milliseconds – Rasch 1979), timbre, spatialization, and register (Bregman 1990, 490ff.). While he addresses the perception of inner voices, Bregman's discussion centers on how composers, specifically

of counterpoint, have employed the principles of ASA to achieve their musical goals involving segregation and fusion. Bregman explains that multi-voice textures involve competition between vertical and horizontal grouping (harmony vs. melody), and that Palestrina-style counterpoint strengthens horizontal bonds by using small melodic steps rather than large skips, as large leaps risk causing a note to fuse with another stream (496). He also suggests using rhythmic independence to facilitate horizontal grouping and auditory streaming of lines.

One factor that has a significant role to play in streaming, and one we can capitalize on in our instruction and performance of streaming, is that of attention. Passive, stimulus-driven/bottom-up processing has been shown to structure and organize sound: “Attention was not required to drive the initial segregation of sounds to streams. That is, stream segregation occurred automatically based on the bottom-up spectrotemporal characteristics of the input” (Sussman 2017, 2991). However, as musicians, most of the time we are not just passively engaging with sound when we are listening, and it has been demonstrated that “attention interacts with the stimulus-driven processes to sharpen the stream segregation process” (Sussman 2017, 2993). Participants in empirical studies (e.g., Sussman and Steinschneider 2009) could separate streams at a smaller frequency separation (i.e., smaller pitch intervals) in active vs. passive listening situations. So, while the preliminary process of stream separation might be passive, we can enhance and refine this streaming through active attentional activities that encourage “event formation” (e.g., hearing a musical phrase or melodic line rather than individual pitches).

Current Ways of Teaching Auditory Streaming in Music Classrooms

As with many of the skills that we develop in the aural skills classroom, it is a challenge to directly instruct and assess the development of auditory streaming. What can instructors do to assist students in locating and tracking inner voices within a complex texture, and how can they know what they are hearing to provide guidance and feedback?

There are few specific resources available that explicitly address the challenges of, or pose solutions to, hearing inner voices. Auralia has a “Chord Imitation” activity where students are tasked with singing back all the pitches of a heard chord, and Teoria has a “Chord and Note” exercise where a named chord is sounded (e.g., E major) followed by one of its pitches (e.g., B) with the task being to identify what note is played. These can be useful at practicing and assessing inner-voice hearing but are

abstract exercises that use synthesized sound, and are removed from real, musical contexts. Jonathan Pitkin created an online tool called Audit in 2018 that plays four-or-more-note chords and then students can use the interface to hear back specific pitches from the chords to assess their own performance at tasks like singing pitches from the chords. This tool could serve as an effective skills scaffold because users can add or remove notes themselves, gradually increasing the challenge, but it, too, is limited in its explicit development of streaming and real-world relevance.

Textbooks and pedagogical books/articles are similarly limited in addressing streaming. Marva Duerksen criticizes Gary Karpinski for only paying “cursory attention” to dictations with more than one part in his book *Aural Skills Acquisition: The Development of Listening, Reading, and Performing Skills in College-Level Musicians* (Duerksen 2009, 398; referring to Karpinski 2000, Chapter 4). Timothy Chenette’s *Foundations of Aural Skills* (2022) frames the task as an extension of hearing bass lines, writing “So there’s not really anything new to teach here; instead, simply practice your attentional focus exercises” (Chapter 9, attention is discussed in Chapter 1). While this is true, and the way that Chenette structures their progression of skill development in the textbook consistently builds on this attentional piece, the exercises and guidance provided assume that students can already “hear it” and there are not any steps offered to lower the barrier of entry to this advanced skill.

One way to assist students in hearing “inside” a piece is to make use of visual information that assists auditory streaming. The integration of visual information has been shown to promote auditory streaming, creating a “bound” multisensory object that enhances the ability, at a neurological level, to parse a complex sound signal into streams (Atilgan et al. 2018; Atilgan and Bizley 2021). In the music classroom, there are a number of ways in which instructors can guide students can engage visual and aural domains simultaneously: For instance, classes can watch an ensemble perform and see the musicians’ hands and bodies move, instructors can mount a camera or phone above a piano or use a MIDI keyboard and project a virtual instrument on the board, or they can employ piano roll notation/MIDI in a projected Digital Audio Workstation window. More traditionally, instructors can engage in active score reading exercises while listening, directing students to read and perhaps even trace with a finger a woodwind part while listening to an orchestral performance or the tenor line of a choral piece.

Instructors can use this multisensory bonding as part of their assessment of students’ streaming skills through error-detection exercises. They can take many of the instructional methods described above and create exercises that have visual/audio

mismatch errors in them to evaluate whether learners are able to successfully stream separate instruments. Most simply, sheet music can be created with plausible but incorrectly notated inner voices, for example writing chord IV but hearing chord ii with an inner voice playing *Re* instead of the written *Do*, and asking students to circle any deviations from the score to assess whether they notice the mismatch. This type of mismatch exercise could also be done with piano roll notation or even, if time and resources allow, by filming an ensemble performing an excerpt twice and creating a mismatch through manipulating certain parameters (rhythmic, orchestration, etc.).

More actively, instructors can physically engage students in the process of streaming and make embodied the shapes of different inner melodic lines by asking them to play multi-voice textures on the keyboard and repeating a phrase multiple times while singing each line within the texture (both supported by the keyboard and unsupported). This is harder to do in a traditional classroom with only one piano but is feasible asynchronously or in a keyboard lab. Alternatively, the embodied experience of inner voices could be felt by singing a cappella as a class and ensuring that everyone performs each part by cycling through the roles.

All of the previously mentioned techniques can be combined with top-down “informed guessing” of the activity of inside voices, using music theory knowledge about the harmonic and melodic context to provide “most likely” options for inner voices (something that Chenette also encourages for developing inner-voice hearing – 2022, Chapter 9). Especially when textures become dense, it is unlikely that one can truly hear each thread, but one can draw on Gestalt ideas such as the principle of good continuation to limit the field of possibilities for where an inner voice may be moving. This is particularly helpful when a line that one was previously able to track becomes buried in a texture. One can use expectations of where the line “should” be to direct their ears and resist the fusion of horizontal streams. This kind of work is important to model for students, showing them times when “experts” struggle to hear every individual line in a texture, acknowledging limits to what can be heard, and showing how to work through these barriers using other skills to make the best representation of what is being heard.

Stem Separation Tools

In recent years there has been a boom in computational tools that enable us to separate full-band recordings into their individual components, producing separate audio tracks for voice, drums, bass, and harmonic instruments. These tools range

from expensive software packages like iZotope RX 11's "Music Rebalance" module to accessible websites like Lalal.ai, which offers free uploads along with paid premium options. Most of the common Digital Audio Workstations (DAWs) now include some kind of stem splitting tool as standard (e.g., Logic Pro X has had a "stem splitter" tool since version 11 was released in 2024). Not that many years ago, these tools required relatively high-level technical know-how to source the code, train and tune the models, and then run them on your desired audio. Now, one can upload audio to a website or right-click the audio file in Logic and it will provide four separate tracks (voice, bass, drums, and harmonic instruments) in a matter of seconds.¹

Stem separation tools rely on machine learning techniques, particularly deep learning, to break down mixed audio recordings into their individual "stems" – the isolated sound files for each instrument including any post-production elements such as compression, reverb, EQ, etc. It does not provide "dry" stems (raw microphone signal). When played together, these stems recreate the original mixed track. These tools are trained on large datasets containing both full ensemble mixes and their corresponding isolated stems, which allow the model to learn the distinctive characteristics and frequency patterns of various sound sources (for more information on how these tools work, see the Appendix: "How Stem Splitters Work").

Case studies

Scaffolding the Development of Auditory Streaming

The motivating question behind this paper has been how we as a community of instructors can help our students develop their abilities to stream different layers in a multi-voice texture while overcoming the high-voice superiority effect. Stem separation tools can help us scaffold the development of this skill by controlling the level of difficulty, bridging the gap between single-line melodies and full ensembles, and can take students from engaging with "extracted music" to "somewhat-real

¹ This field is developing fast and new products are always coming to market, but, at this time, the most flexible, free option that I would recommend to instructors and students would be the MVSEP website (<https://mvsep.com/en>). This site has access to a large, continually expanding and updating number of different separation algorithms that you can choose between from a dropdown menu. It also has a league table where the different algorithms are empirically compared to one another using the "Signal-to-Distortion Ratio." To use this site, you upload your audio file, select the algorithm (or go with the default), and then your audio is queued up for separation on their servers. Depending on the time of day and how many others are using it, this can be very fast, or it can take 15 or more minutes for the queue to clear. Peak times appear to be during US working hours, so I recommend using it at night or first thing in the morning (if you are on the East coast).

music” to “mostly-real music” to “real music” in a more gradual and effective way. When one has access to the stems, one can change the dynamic balance of a full-band mix, increasing or decreasing the volume of layers or even muting them entirely. Or one can present the target audio in isolation and gradually introduce other layers back into the mix to incrementally increase the challenge level. This practice has been possible for a long time, long before AI stem separation tools, by using EQ options on amplifiers to create a high- or low-pass filter (allowing only the treble or bass, respectively, to sound) or, with some recordings, using only the left or right channel as sometimes the instruments are hard-panned to the left or right. For example, because of the limits of early stereo technology like the Ampex 4-channel tube stereo mixer that could only do center/hard left/hard right, many jazz recordings from around the 1960s have the drums and bass on one side and the piano and saxophone on the other, a feature of the recordings that was sometimes used by jazz students to create “music minus one” backing tracks. Stem separation tools simply do this process more precisely and create cleaner end audio that one has more fine-grained control over.

To give an example of how stem separation tools can be used more explicitly in the aural skills classroom to develop streaming: British indie/folk pop singer dodie’s 2019 song “Monster” builds layer by layer across the [song](#). It starts with two bars of brushed snare drum followed by the vocal entry over a solo synth playing an E-E-B (x4) E-E-C (x4) ostinato with a tresillo rhythm. After eight bars, the bass enters with an E pedal and the drums intensify slightly through adding a kick on beat 1. Upon the entry of the bass, the synth line begins to vary: It retains its monophonic identity and the tresillo but is now all arpeggios (I and iv_4^6 in E major). Since there are now more instruments and the vocal part starts to form a more continuous line, hearing through the texture and tracking the synth becomes more challenging.

How can students track the evolution of this melodic/harmonic synth line? An instructor can scaffold this process using stem separation, breaking the task down and providing temporary support to students until they build secure foundations, then gradually and intentionally remove the support until students are able to stream out layers independently.

To start, the song needs to be separated into its stems. Most of the separators mentioned earlier offer voice, bass, drums, and “other” – a catchall category into which harmonic instruments get separated. Many of the online tools, such as Lalal.ai, Moises.ai, and MVSEP, offer premium, paid stem separation services that crack this “other” layer open further into guitar, piano, backing vocals, etc., though these tools are newer to the field and do not have quite as much training, and so sometimes

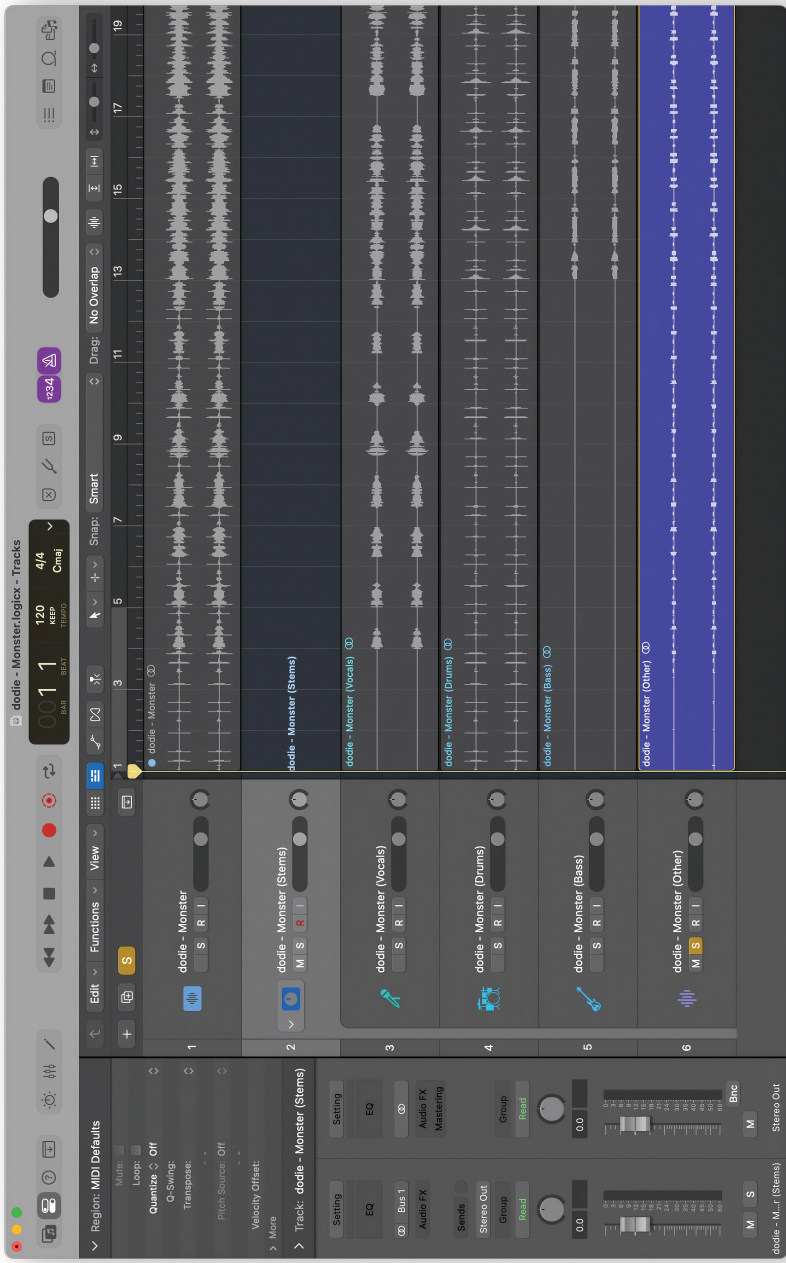
produce odd output that merges multiple layers. As the field progresses, these models will surely become better and free versions will also appear.²

Stems can then be loaded into a DAW and one can alter the relative dynamics of each layer. This DAW could be software on a computer, for example Logic Pro X, or there are online, in-browser DAWs. Moises.ai can serve dual function as it is an online DAW that has a stem separator built in. A starting point might be to “solo” the synth part (the “other” stem) so as to isolate just that layer and place maximum attentional focus on this sound (Example 1). To provide rhythmic and metric structure, we might want to add the drum part in, too. From here, we may wish to add back in the vocal layer, but perhaps use the DAW’s mixer to turn down the volume so it is less likely to steal away the attentional spotlight due to the high-voice superiority effect (Example 2). Lastly, the bass could be added in which, although it only plays an E pedal, adds a new distractor to the task.

Advanced extensions: The chorus of “Monster” (“So maybe I will talk to you”) retains the tonic-subdominant arpeggios but thickens the texture with an electric guitar and faint sustained chords on an organ. Can students still track the arpeggiated synth line through this step-up in texture? How about at the end of the song when the texture thickens considerably for the final chorus with backing vocals singing an independent line, the drums increasing in intensity, and a glitchy low frequency oscillator effect is added to the backing vocals which cuts them in and out multiple times per second? Can students dig through these distractors to still hear the synth line that they have been tracking across multiple levels of difficulty?

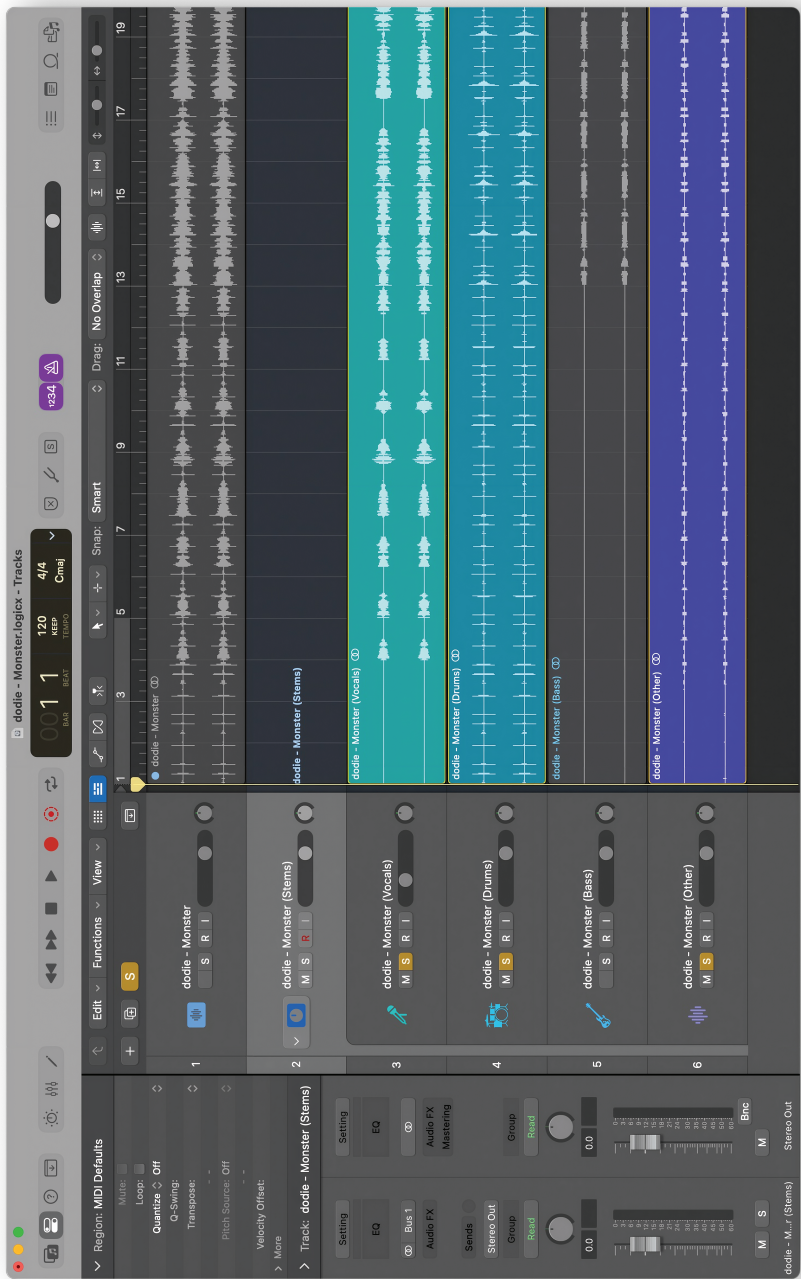
Students can continue this work themselves using online DAWs like Moises.ai, which separates stems and has most typical functions of a DAW. They can review the dodie classwork themselves or import new songs they want to hear inner voices of, self-pacing their continued development of streaming skills and challenging themselves to see how well they can hear past the dominant high voice with minimal instrumental layers removed or quietened.

² As of May 2025, Logic Pro X 11.2 now separates guitar and piano out separately from the “other” instruments.



Example 1.

Screenshot of a Logic Pro X session showing dodie's "Monster" being separated into voice, drums, bass, and "other" with the synth (other) part soloed (note the highlighted "s" under the track name and all other parts greyed out) so only that line is sounded.



Example 2.

Screenshot of a Logic Pro X session showing separated stems for dodie’s “Monster” with the voice at half volume, synth line and drums playing, and the bass omitted.

Creating “Real” Transcription Exercises

In addition to helping with auditory streaming, we can use stem separation tools to create transcription activities that are more “ecologically valid” (i.e., relevant and applicable to real-world settings and situations) by editing recordings. A common, time-consuming challenge with aural skills class preparation is finding appropriate repertoire that introduces, reinforces, or nuances a concept or skill. It is frustrating to find a seemingly perfect piece to use as an example only to discover that it cannot be used because there is one element that is too advanced for the class right now (e.g., a melody that features one chromatic passing tone), because there is a distracting element (e.g., countermelody) that will take focus away from the class’s learning goal, or because a recording does not quite have the clarity needed (due to, e.g., expressive timing). Instructors can solve this repertoire-sourcing challenge with some simple technical knowledge and minimal musical interference by splitting the music into its constituent parts and re-pitching the chromatic passing tone, turning down/erasing the countermelody (as in the dodie example), or re-timing the performance.

For instance, the original audio for Frank Sinatra’s performance of “The Way You Look Tonight” with the Nelson Riddle Orchestra from 1964 can be separated into vocal, bass, drums, and “other” instrument layers using a stem separation tool of choice. From here, one can create multiple audio outputs depending on our pedagogical needs and learning goals. Note that big bands are not what the stem separation tools have been trained on; their training sets are mostly pop and rock, but one still gets remarkably good separation of Sinatra’s voice, the drums, the bass, and then rest of the ensemble (trumpets, saxophones, and trombones) contained within the “other” category. For my classes, I took the individual vocal stem and “quantized” both the rhythmic and pitch elements of Sinatra’s performance. I made his pitches “perfect” (colloquially, I “Autotuned” them) and aligned his rhythms to a quarter-note grid. Currently, only professional-level DAWs have quantizing built in; for example Logic has “Flex Pitch” and “Flex Time” functions. I then presented students with both the quantized solo Sinatra audio and an ensemble audio file that recombines the quantized Sinatra with the original ensemble and used this as the first melodic transcription assignment of Aural Skills I. The melody is almost exclusively stepwise apart from a *Sol-Do* motion at the start and the rhythm is all quarter notes and whole notes with movement on strong beat locations, so it is an approachable melody for early transcribers. I approached this transcription in multiple ways, starting with contour tracking over a grid, then moving into solfège, and finally into staff notation and/or playback. I used this as a first assignment because it connects work in the aural skills

classroom to real, musical performances rather than dry, synthetic exercises, and also because it opens up space to talk about the large gap between music as notated and music as performed, promoting discussion and debate about the purposes of transcriptions and the limits of staff notation. This also provides an early opportunity to talk about technology in the classroom, both regarding the stem separation tools and the manipulation of Sinatra's performance. The original Sinatra performance has a very free approach to both pitch and rhythm and much of the "swagger" of his performance is eradicated through the quantizing process. All of this work enriches what could have been a clinical transcription exercise of a stepwise, quarter-note melody and hopefully serves as somewhat of a "hook" to keep students motivated in the early stages of aural skills classes.

Creating "Music Minus One" Tracks

As a final example of how stem separation tools can be employed in the aural skills classroom, we can create "Music Minus One"-style tracks in which a recording has one of its elements removed. I have used this to create sight singing exercises with more ecological validity, removing the vocal part and effectively making a karaoke version. For example, I have separated out the vocal part from a recording of Louise Reichardt's "Notturmo" from *Zwölf Deutsche und Italiänische Romantische Gesänge* leaving only the piano accompaniment (the "other" stem). This "Lied Minus One" track affords opportunities to practice singing melodies with chromaticism and modulations while having a professional pianist accompanying. Not only does this encourage more musical and expressive performing by the students, it also frees up the instructor to pay attention to the students, rather than having to devote their attention to performing the piano part.³

Responsible Usage of AI Tools

One big concern many have regarding responsible use of these stem splitting tools is that, in cracking open these recordings for use in the classroom, we are putting ourselves and our institutions in legal risk. Intellectual property lawyer Jonathan Coote discusses the copyright implications of using AI tools to separate stems for commercial music production applications and points out that, alongside the issues of licensing of both the composition and the sound recording rights, there are the issues of the "right of paternity" and the "right of integrity" (Coote 2024). These

³ One other potential way of using "Music Minus One" tracks would be to remove the bass from a performance and ask class to improvise a bass line given chord symbols or Roman numerals.

refer, respectively, to the right of the author to be properly identified as the creator of their work and the right to protect the author's work from derogatory treatment, meaning any distortion, modification, or other action that would harm the author's honor or reputation. I have had discussions with copyright lawyers contacted through my university's legal counsel office, and the opinion is that using AI separated stems for research and education are likely covered by the Fair Use Doctrine under U.S. Copyright Law. As such, we can use these materials in our classrooms so long as we abide by the usual Fair Use considerations, and we do not distribute or commercialize these recordings, for example by creating an online repository of examples or as downloadable content alongside a textbook. That being said, I would encourage everyone to talk to their own university's legal counsel since different institutions have different appetites for risk and interpretations of fair use.

As AI technology, including stem separation tools, becomes increasingly integrated into educational and creative contexts, it raises important ethical and philosophical questions. One primary concern is the encroachment of AI into areas traditionally dominated by human creativity and expression. In the music classroom, for instance, instructors must ask whether the use of these tools compromises the authentic, human-centered learning experience they strive to foster. Can the classroom remain a space where students engage deeply with the music – immersing themselves in the process of active listening, analysis, and creative expression – or are we ceding too much of that process to algorithms? While AI tools can undoubtedly enhance learning, it is important to consider whether their pervasive presence detracts from the opportunity to engage with music in a way that is meaningful and grounded in human perception and emotion. Here, I choose to see AI stem splitters, when utilized in the ways similar to those described in the case studies, as augmenting the learning environment, facilitating active and independent learning rather than outsourcing labor and thought to “The Machine” (McGuire et al. 2024, Miller 2020).

Another critical issue is the ethical implications of using AI stem splitters to manipulate musical works. Many of the artists whose music we deconstruct through these technologies have not consented to such computational alterations of their creative output.⁴ A song or composition is often the result of a painstaking and deliberate process,

4 AI stems can be found increasingly often in commercial contexts so some artists, or at least their management and record labels, are consenting to the splitting up of recordings. One example of this is the game *Fortnite*, developed by Epic Games. In the game, players often wait in a virtual lobby before joining an online match. To keep players engaged during this downtime, Epic Games introduced “Jam Tracks” in December 2023, a feature that allows players to step into the shoes of a drummer, guitarist, or singer for famous songs. These tracks are made available for purchase in the in-game store, and

where each element is carefully crafted by multiple people to contribute to the overall artistic vision. When we use AI to isolate individual stems, we risk reducing a complex, cohesive work into its component parts without acknowledging the artistic intent behind it. This “algorithmic shredding” of music can be seen as a form of artistic appropriation, where the integrity of the original work is potentially compromised for the sake of convenience or novelty. As educators and consumers of art, it is our responsibility to approach these tools thoughtfully, considering both the benefits and the ethical responsibilities we hold toward the creators whose work we are engaging with.

To bring some of these techniques into aural skills classrooms without needing to engage with stem separation tools for moral and/or technical reasons, a few alternative solutions can be found with some quick internet searches. One option is to use the Creative Commons’s “Legal Music for Remixing and Sampling” databases.⁵ This points to numerous sites, such as ccMixter and Freesound, where multitrack audio files can be downloaded that have been explicitly provided and legally designated for the purposes of exploring the individual musical layers. For more “real” (and perhaps more familiar) examples, some well-known artists choose to release their original stems with explicit permission given for others to remix, explore, and experiment. This is typically seen in remix contests where artists like Beyoncé, Radiohead, and Charli XCX share their stems and invite fans to vote for the best remixes. Since 1999, Nine Inch Nails has made their stems available, fostering a spirit of creativity, playfulness, and sometimes competition (Anonymous 2024). In even rarer cases, artists use stem-sharing as a form of anti-commercial expression. For instance, Moby released the full stems of his *Innocents* album via BitTorrent, aiming to “bypass the old and arbitrary gate-keepers” and dissolve the boundary between creator and audience (McDonald 2013). Because these are all legitimate, original recording files rather than AI separated stems, they offer more control over layers within the song as you often get audio files for more instruments than simply voice, bass, drums, and “other” and there is much better isolation of files; however, these official stems are rare and so the chances of finding a resource that matches your learning goals are low, which is where stem separators come into their own.

some of the songs use official stems that were directly provided by the record company. For more obscure or older tracks, where the original stems may no longer exist, it would appear that Epic Games has turned to AI stem separation tools to reconstruct individual instrument parts from the full-band mix. I have not been able to talk with anyone from Epic Games to confirm this, but there are occasionally artifacts in the audio that are indicative of stem separation.

5 <https://creativecommons.org/about/arts-culture/arts-culture-resources/legalmusicforremixing/>

Conclusion

The integration of AI-powered stem separation tools into the aural skills classroom represents a significant opportunity to enhance the development of advanced aural abilities in our students. By providing instructors and students with the means to isolate and explore individual components within complex musical textures, these tools facilitate the cultivation of auditory streaming – a critical skill for understanding and engaging with multi-voice music. The case studies presented in this paper illustrate the diverse ways in which stem separation can scaffold the development of streaming as well as support other learning activities such as making performance exercises more musically relevant and fine-tuning dictation exercises that utilize familiar music. Moreover, these tools introduce students to technologies they can continue to explore and utilize in their independent musical lives beyond the classroom. As the world moves further into the AI era, it is essential that pedagogues approach these tools thoughtfully and ethically, ensuring they are used to complement, rather than replace, traditional pedagogical methods. Just as instructors guide students in the responsible use of writing aids like ChatGPT, so too must they model and define AI audio analysis tools as integral elements of a modern, adaptable musical toolbox. By doing so, students can be equipped with both the technical and ethical frameworks to navigate an increasingly technology-driven musical landscape.

Works Cited

- Airaj, Mohammed. 2024. Ethical Artificial Intelligence for Teaching-Learning in Higher Education. *Education and Information Technologies* 29, no. 13: 17145–67. <https://doi.org/10.1007/s10639-024-12545-X>.
- Anonymous. 2024. “Nine Inch Nails – Multitracks.” *nin.wiki*. <https://www.nin.wiki/Multitracks>.
- Atilgan, Huriye, and Jennifer K. Bizley. 2021. “Training Enhances the Ability of Listeners to Exploit Visual Information for Auditory Scene Analysis.” *Cognition* 208: 104529. <https://doi.org/10.1016/j.cognition.2020.104529>.
- Atilgan, Huriye, Stephen M. Town, Katherine C. Wood, Gareth P. Jones, Ross K. Maddox, Adrian K.C. Lee, and Jennifer K. Bizley. 2018. “Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding.” *Neuron* 97, no. 3: 640–55. <https://doi.org/10.1016/j.neuron.2017.12.034>.
- Bregman, Albert. 1990. *Auditory Scene Analysis*. Cambridge, MA: MIT Press.
- Bregman, Albert. 2015. “Progress in Understanding Auditory Scene Analysis.” *Music Perception* 33, no. 1: 12–19. <https://doi.org/10.1525/mp.2015.33.1.12>.
- Calcus, Axelle. 2024. “Development of Auditory Scene Analysis: A Mini-Review.” *Frontiers in Human Neuroscience* 18. <https://doi.org/10.3389/fnhum.2024.1352247>.
- Chenette, Timothy. 2022. *Foundations of Aural Skills*. <https://uen.pressbooks.pub/auralskills/>.
- Coote, Jonathan. 2024. “‘Stem-separating’ AI is Revolutionising the Music Industry.” *Bray & Krais*. <https://www.brayandkrais.com/stem-separating-ai-is-revolutionising-the-music-industry/>.
- Dewitt, Lucinda, and Crowder, Robert. 1987. “Tonal Fusion of Consonant Musical Intervals: The Oomph in Stumpf.” *Perception & Psychophysics* 41, no. 1: 73–84. <https://doi.org/10.3758/BF03208216>.
- Duerksen, Marva. 2009. “Review of Gary S. Karpinski, *Manual for Ear Training and Sight Singing, Anthology for Sight Singing, Student Recordings CD-ROM, Instructor’s Dictation Manual, and Instructor’s CD-ROM*.” *Gamut* 2, no. 1: 391–405.
- Hosken, Fred. Forthcoming. “AI Stem Separation & MIR: Assessing the Validity of Onset Analyses That Use AI-Isolated Audio.” *Empirical Musicology Review*.
- Huron, David. 2016. *Voice Leading: The Science Behind a Musical Art*. The MIT Press. <https://doi.org/10.7551/mitpress/9780262034852.003.0001>.
- Ishida, Kai, and Hiroshi Nittono. 2024. “Different Voice Part Perceptions in Polyphonic and Homophonic Musical Textures.” *Psychology of Music*. <https://doi.org/10.1177/03057356241271027>.
- Karpinski, Gary S. 2000. *Aural Skills Acquisition: The Development of Listening, Reading, and Performing Skills in College-Level Musicians*. New York: Oxford University Press.
- Marie, Céline, and Laurel J. Trainor. 2012. “Development of Simultaneous Pitch Encoding: Infants Show a High Voice Superiority Effect.” *Cerebral Cortex* 23, no. 3: 660–69. <https://doi.org/10.1093/cercor/bhs050>.
- Marie, Céline, and Laurel J. Trainor. 2014. “Early Development of Polyphonic Sound Encoding and the High Voice Superiority Effect.” *Neuropsychologia* 57: 50–8. <https://doi.org/10.1016/j.neuropsychologia.2014.02.023>.

- McDonald, Scott. October 9, 2013. "Moby X BitTorrent: Unlock and Remix Innocents." *BitTorrent*. <https://www.bittorrent.com/blog/2013/10/09/moby-x-bittorrent-unlock-and-remix-innocents/>
- McGuire, Jack, David De Cremer, and Tim Van de Cruys. 2024. "Establishing the Importance of Co-Creation and Self-Efficacy in Creative Collaboration with Artificial Intelligence." *Scientific Reports* 14 (1): 18525. <https://doi.org/10.1038/s41598-024-69423-2>.
- Miller, Arthur I. 2020. *The Artist in the Machine*. London, England: MIT Press.
- Munive Benites, David A., and Philippe Lalitte. 2022. "Employment of Cognitive Science Theories to Improve Aural Training through Digital Tools." In *15th International Conference of Students of Systematic Musicology (SysMus22)*, 50–51. Ghent.
- Rafii, Zafar, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. "MUSDB18-HQ – An Uncompressed Version of MUSDB18." *Zenodo*. <https://doi.org/10.5281/zenodo.3338373>.
- Rasch, Ruth. 1979. "Synchronization in Performed Ensemble Music." *Acustica* 43: 121–31.
- Rouard, Simon, Francisco Massa, and Alexandre Défossez. 2022. "Hybrid Transformers for Music Source Separation." arXiv. <https://doi.org/10.48550/arXiv.2211.08553>.
- Sussman, Elyse S. 2017. "Auditory Scene Analysis: An Attention Perspective." *Journal of Speech, Language, and Hearing Research* 60, no. 10: 2989–3000. <https://doi.org/10.1044/2017-JSLHR-H-17-0041>.
- Sussman, Elyse, and Mitchell Steinschneider. 2009. "Attention Effects on Auditory Scene Analysis in Children." *Neuropsychologia* 47, no. 3: 771–785. <https://doi.org/10.1016/j.neuropsychologia.2008.12.007>.
- Trainor, Laurel J., Céline Marie, Ian C. Bruce, and Gavin M. Bidelman. 2014. "Explaining the High Voice Superiority Effect in Polyphonic Music: Evidence from Cortical Evoked Potentials and Peripheral Auditory Models." *Hearing Research* 308: 60–70. <https://doi.org/10.1016/j.heares.2013.07.014>.

Appendix 1 – How Stem Splitters Work

The process of stem separation typically involves two main stages: encoding and decoding. In the encoding stage, the audio signal is transformed into a high-dimensional representation that captures both time-based (temporal) and frequency-based (spectral) features. Techniques like the Short-Time Fourier Transform are often used to create a spectrogram, which visualizes how the frequency content changes over time. U-Nets, which are particularly effective for tasks requiring localization and context preservation, are commonly applied to process this encoded representation. Their architecture, featuring a contracting path for feature extraction and an expansive path for reconstruction (hence the “U” shape), enables the model to differentiate between sound sources based on their unique features. In the decoding phase, the model then converts the encoded data back into separate audio signals. By applying learned filters, the U-Net reconstructs each stem from the encoded information, effectively isolating individual sound sources. The result is a separated stem that, qualitatively, closely resembles the original element.

The quality of the separated stems varies depending on the content of the original audio file. Since the separators work, in incredibly simplified terms, by slicing bands of the frequency spectrum, if the input audio is an ensemble of instruments that all occupy a similar space in the pitch and timbre domains, for example a string trio or a barbershop quartet, the output of the separator will not be successful with a lot of cross-bleed between elements. We must remember that the stem separation tools are trained on a corpus of pop and rock music, so this is what they are optimized to separate. For instance, one of the top performing tools at the moment, HT Demucs (Rouard et al. 2022), is trained on the musdb18 dataset, which includes 150 full-length music tracks (~10 hours total) across various genres within the broad fields of pop and rock, alongside their isolated stems for drums, bass, vocals, and “other” elements (a catchall stem for any harmonic and mid-frequency instruments; Rafii et al. 2019). Nevertheless, one can have great successes with music from genres outside of these, but results do vary. If, however, one tries to split open a rock band with distinct vocal, harmony, bass, and rhythm layers, the output is often very clear. Some artifacts will inevitably be found from other instruments in the separated stems, which I refer to as “ghosts” as they are otherworldly sounding spectral inclusions, but these are often easily ignored when doing listening tasks or their presence can be reduced by using EQ options. It should be noted that while not prohibitive to aural skills exercises, these artifacts do risk interference with empirical analysis of musical stems, a problem I explore in Hosken (forthcoming).

