Kirsten Hermes: Enhancing creativity through automatic mixing research: testing spectral clarity predictors in the mix process

Citation:
Hermes. K, (2019). Enhancing creativity through automatic mixing research: testing spectral clarity predictors in the mix process. In Gullö, J.O., Rambarran, S., & Isakoff, K., (Eds.), Proceedings of the 12th Art of Record Production Conference Mono: Stereo: Multi (pp. 155-170). Stockholm: Royal College of Music (KMH) & Art of Record Production.

Abstract

As established in a prior research study, spectral clarity is an important parameter of high-quality mixes. Two predictors for this are the harmonic centroid (a weighted centre mass of energy of a sound spectrum) and spectral inconsistencies related to sharp peaks roughly in the middle of the frequency spectrum (Hermes et al., 2017). The impact of these predictors on the creative process is tested and starting points for further research are established in this paper.

Introduction

Mixing music is a complicated process where several tracks of recorded audio are combined to an overall piece. Difficulties can arise from e.g. time constraints or lack of expertise. Research in automatic mixing seeks to work towards the development of perceptually informed assistive or fully automated mix tools. In a previous PhD research study (Hermes et al.: 2017), two predictors were developed for one important parameter of music mixes, i.e. single sound spectral clarity. As the author is also a creative practitioner (artist and producer), the current paper includes discussion on testing the two spectral clarity predictors in the mix process of an electronica production. The aim here is to assess the usefulness of the predictors and to develop starting points for follow-on research. Another aim is to discuss different research methods that can be employed in furthering the understanding of music mix parameters. The author argues that an interdisciplinary approach, drawing on both scientific and creative knowledge, can yield particularly useful results.

The structure of this paper is as follows. Section 1 introduces the field of automatic mixing, to provide readers that are less familiar with this area of research with an overview of aims and methodologies. Section 2 is a sum-

mary of the author's PhD findings. In section 3, two predictors of single sound spectral clarity are tested in the mix process and findings are related to relevant literature. Section 4 is a discussion with suggestions for further research.

What is automatic mixing?

The democratisation of audio technology and advancements in internet file sharing have resulted in the delocalization of professional recording studios and the decline of traditional record companies (Pras et al.: 2013, pp. 612–626). Almost anyone can create musical outputs and share these online, hence many musical artists are now self-produced (Bell, 2014. pp. 295–312). The large number (982) of music related degree courses offered in the UK indicates that there are many such artists (What Uni: 2014). Not only music artists but also filmmakers and entrepreneurs use media tools which they may not be experts in. Music mixing is equally important in the context of live music. Audience attendance at UK live music events has recently increased by 12% (Ellis-Peterson: 2017) and live sound mix engineers often work under tight time constraints (Biederman and Pattison: 2014).

Mixing music is a complicated process and traditionally requires extensive ear training and an in-depth understanding of specialized tools and techniques. Most mix tools relate to physical parameters of sound, making it difficult for the novice user to understand the connection with perceptual parameters (e.g. compression vs. loudness). When musical artists undertake the entire creative process alone, this can also result in a lack of objective feedback. Recordings taken under less than ideal conditions, such as in "bedroom" studios, can contain unwanted artefacts and spectral problem areas, complicating the mix process further (De Man and Reiss: 2017).

Since all successful mixes seem to have certain qualities in common, it is possible to automate parts of the mix process and to develop powerful, perceptually informed, artificially intelligent (AI) mix tools. The term "automatic mixing" was first used by Dugan (1975) in the context of automatic microphone gain handling for speech. Today, automatic mix tools are commercially available, including Izotope's assistive mix tool *Neutron* and the online mastering platform *Landr*. Existing mix tools range from completely autonomous mixing systems to more assistive, workflow-enhancing tools and perceptually enhanced interfaces (De Man and Reiss: 2017).

Many disciplines, including signal processing, music cognition, machine learning and human computer interaction contribute to automatic mixing research (Scott: 2014). Different approaches exist to solving this complex problem. The most common is knowledge engineering (De Man and Reiss: 2013), where informally known rules for creating high quality mixes are implemented in technology. These mixing rules are derived from prosumer

mixing guides and the expertise of mix engineers. The second approach is grounded theory, which was first presented by Glaser and Strauss (1967) in the context of social research. The authors propose that conclusions that are grounded in data can be more reliable than conclusions based on existing theories. Hence, grounded theory is the discovery of theory from data systematically obtained through research. De Man and Reiss (2103) relate the grounded theory approach to the field of automatic mixing. Here, basic knowledge about high quality mixes is acquired first and subsequently transferred to an intelligent system. In this approach, psychoacoustic studies are undertaken to define mix attributes, and perceptual audio evaluation (i.e. listener-based experimentation) is employed to determine listener preference for mix approaches (Bech and Zacharov: 2006). The grounded theory approach can be slow and resource intensive. De Man and Reiss (2013) argue that therefore, it is too limited to constitute a sufficient knowledge base for the implementation of an overall system. Knowledge engineering is a less formalized approach (Scott: 2014) and many commonly accepted rules in mixing do not hold true in formalized studies, for example the notion that most elements should be high pass filtered above their fundamental frequency (Pestana and Reiss: 2014a). Hence, both approaches have advantages and disadvantages.

De Man and Reiss (2017) provide a useful overview over existing studies that seek to automate parts of the mix process. Initial studies contributed to the development of tools for the automatic adjustment of e.g. stereo panning (Gonzales and Reiss: 2010). Further studies have focussed on automating mix parameters such as level (e.g. Wilson and Fazenda: 2016a), reverb (e.g. Benito and Reiss: 2017), panning (e.g. Pestana and Reiss: 2014), EQ (spectral equalization, e.g. Hafezi and Reiss: 2015, pp. 312-323) and compression (e.g. Ma et al., 2014, pp. 412-426).

Some studies do not directly develop prototype automatic mix tools but instead help further the understanding of important perceptual parameters that can feed into the development of such tools (grounded theory). For example, Fenton and Wakefield measure perceived punch and clarity in produced music (Fenton and Wakefield: 2012). Pestana et al. (2013) investigate average spectra of commercially recorded pop songs. Wilson and Fazenda (2016b) investigate the perception of audio quality in productions of popular music. Research in spatial quality perception can also be useful in the field of automatic mixing (e.g. Conetta et al.: 2015, pp. 847–860). The author took a similar approach in her PhD, as summarized in section 2.

Motivation and summary of PhD research findings

The author's motivation for pursuing a PhD in the field of automatic mixing was to investigate whether the quality of music mixes could be measured

objectively. There can be disagreement as to what constitutes quality in any creative product. Generally agreed quality parameters can help guide this discussion, which can be especially useful in an educational context. The author is also fascinated with the way in which scientific research and creativity can enhance each other. In this case, findings in psychoacoustics and auditory perception can be used to explain preferences in recorded music.

During her PhD, the author and her PhD supervisors worked towards measuring and modelling the perceived quality of music mixes, taking a grounded theory approach. Key findings summarized in the current section are all based on the author's PhD thesis (Hermes et al., 2017). First, relevant high-level descriptive mix quality criteria were established through a search of scientific and creative literature. These are "clarity and separation", "balance", "impact and interest" and "freedom from technical faults", alongside context-specific parameters. Clarity and separation is the extent to which individual components can be heard in a mix. Balance is an even distribution of energy in the spatial and frequency domains. Three sub-categories of balance are horizontal or stereo balance, depth and tonal balance. Horizontal or stereo balance is the extent to which sound energy is distributed symmetrically and evenly between the left and right channels (within any given frequency range). Depth is a sense of perspective in a mix, where sound sources can be placed at various distances from the listener and inside a fictional, reverberant space of a certain size and shape. Tonal balance is the extent to which sound energy is distributed evenly across the frequency spectrum. Impact and interest is the extent to which the mix grabs the listener's attention. Freedom from technical faults is the absence of e.g. unwanted recording artefacts or clipping. Lastly, context specific characteristics are the extent to which the mix fits current trends, fashions and norms, complements artistic purpose and supports the musical content. The latter category relates to mix quality parameters that are difficult to generalize, whereas the other categories can be measured automatically. An overview of all parameters is shown in figure 1.

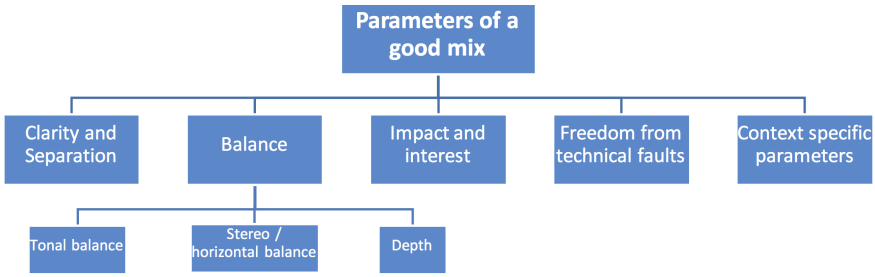

Fig. 1: The parameters of high quality music mixes.

Clarity and separation were deemed particularly important for music mix quality and were therefore investigated further. As established in a literature review, clarity and separation in general depend on spectral, spatial and intensity factors and temporal changes in these factors and this is also likely to be the case in music mixes. Spectral factors play a particularly important role across all areas of literature consulted, i.e. timbral clarity, clarity in concert halls, masking, loudness, auditory scene analysis and speech intelligibility. Hence, the impact of spectral equalization (EQ) on spectral clarity was investigated in a series of listener-based experiments. The focus here was on changes in the spectral clarity of single, isolated sounds to keep the complexity low. Single sound spectral clarity is the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard. It was established that the clarity of naturally-occurring sounds can be increased when low-Q EQ is applied to boost the less-audible higher frequency regions. If EQ exaggerates or introduces timbrally unpleasant spectral inconsistencies, then these can mask or distract from other sonic components and lead to a clarity reduction (Hermes et al., 2017).

Based on these findings, two predictors of relative changes in single sound spectral clarity were established. These are the harmonic centroid (HC, a weighted centre mass of energy of a sound spectrum, Hermes et al., 2016) and mid-range spectral peakiness (measuring spectral inconsistencies related to sharp peaks roughly in the middle of the frequency spectrum). The HC is a weighted mean of a sound's spectrum, indicating the harmonic at which the centre mass of energy is situated. It is defined in Equation 1.

$$HC = \frac{\sum_{k=0}^{K-1} f(k)X(k)}{F \cdot \sum_{k=0}^{K-1} X(k)} \qquad \text{(Equation 1)}$$

$X(k)$ is the magnitude of frequency bin number $f(k)$ is the centre frequency (Hz) of k, K is the number of bins output from a discrete Fourier transform of the sound, and F is the sound's median fundamental frequency. F is defined as the pitch directly in the middle between the highest and lowest note played (Hermes et al. 2017). The HC needs to be raised above around 1.5 harmonics before clarity increases. Mid-range spectral peakiness is calculated by measuring the height of sharp peaks in the middle of stimulus long-term average spectra (LTAS). A computational model was developed that fits a curve to the LTAS, such that potentially unpleasant-sounding peaks lie above it, while the remaining frequency areas lie below it (Hermes et al., 2017). The number of data points above the curve are to estimate relative changes in mid-range spectral peakiness. The computational model is used in

section 3. More information about the model, including a MATLAB download link can be found in Hermes et al. (2017).

As mentioned earlier, the above findings were mainly tested for isolated sounds in order to keep the complexity low. Sounds in mixes were only briefly investigated, in the context of one additional listening test (Hermes et al., 2017). It was concluded that the predictors are still somewhat useful for measuring spectral clarity changes for individual sounds in mixes but the presence of the backing track (rest of the mix) means that complex masking and fusion phenomena need to be considered additionally. Following the analysis of listening test data, it appeared that the more that peaks in the target protrude through the backing track, the clearer the target is perceived (peak audibility). Hence, interestingly, the contribution of spectral peaks to relative changes in clarity appears to depend on the context: when EQ is used to increase peaks on the target sound itself, clarity is reduced. However, if the audibility of peaks is increased by cutting this area in the surrounding backing track, clarity is increased. Therefore, spectral peaks on target sounds appear to contribute to clarity in a complex way and further research needs to be carried out to investigate this. Spectral peaks will be the focus of the next section.

Testing spectral clarity predictors in the mix process

In the previous section, PhD findings on the parameters of high quality mixes and the spectral clarity of sounds were summarized. The aim of the current section is to make informed suggestions for further research by applying the findings to the creative process. As an electronic artist (*Nyokee*), the author has been writing, producing, performing, mixing and mastering original songs for approximately ten years. Since the entire creative process is undertaken by one person, external, objective feedback on the mix process can be useful. Such feedback could be provided through automatic mix tools. Therefore, as an initial step towards furthering the understanding of spectral clarity in mixes, an autoethnographic study is carried out where the above findings are used to mix a track. Like all research methods, autoethnography has strengths and weaknesses.

Ellis et al. (2011) present some of the criticism that autoethnography has received as a research method. It is occasionally described as insufficiently rigorous, theoretical, and analytical, as conclusions may be based on biased data. However, the authors point out that these criticisms erroneously position art and science at odds with each other. Similarly, Dwyer investigates the question as to whether qualitative researchers should be members of the population they are studying. She comes to the conclusion that the dichotomy of "insider versus outsider" is misplaced here and that we should instead explore the complexity and richness of "the space between entrenched per-

spectives". As mentioned in section 2, many disciplines are involved in the field of automatic mixing. Therefore, the use of autoethnography may help develop a more holistic understanding of spectral clarity and help guide follow-on research. The remainder of this section is structured as follows. In section 3.1, the predictors are tested in the mix process and findings are presented. In section 3.2, the role of peaks in the natural character of a sound is discussed. In section 3.3. the potential impact of phase issues is presented and in section 3.4 the influence of masking and auditory scene analysis phenomena on clarity is discussed. Lastly, section 3.5. argues that a more holistic understanding of mix quality may be necessary.

Testing the spectral clarity predictors in the mix process — observations

The previously established predictors of single sound spectral clarity were applied to the mix process of a vocal in an electronica production ("Serendipity") in order to assess whether they may be able to help improve lead sound clarity in this type of production. In particular, the contribution of spectral peaks to clarity is investigated further. Feedback was also informally gathered from a group of additional audio professionals with no previous knowledge of the predictors. The finished track, "Serendipity" can be auditioned online (https://soundcloud.com/kirsten-hermes/serendipity). Conclusions in this paper are based on one mix only, and, as mentioned above, the author is both the scientist and subject. While this can be seen as a limitation of the study, the aim here is not to develop a universal model for sound clarity in mixes but rather to explore whether EQ-related clarity changes may be related to harmonic centroid changes and mid-range spectral peakiness for sounds in mixes. For a more holistic understanding of vocal clarity, a larger collection of mixes will need to be investigated. Vocal clarity in the current mix is also more formally tested in a publication under review (Hermes: 2018). Here, ten participants compare versions of the mix in terms of clarity in a custom GUI. The participants are experienced in critical listening and in verbalising sensations of timbre and have no previous knowledge of the predictors.

Having completed an arrangement of synthesizers and electronic sounds, the author recorded her vocal into *Apple Logic Pro X* in an acoustically untreated home studio, using an *SE2200* microphone, an *SSL* channel strip and an *Focusrite Saffire Pro 24* audio interface. The recording lacked clarity and had unpleasant spectral peaks which made it a useful starting point for testing the predictors. Three vocal mixes were created for the first (30s) verse, as follows. First, EQ, compression, deEssing and reverb were applied without explicitly consulting the predictors (version 1). Previous knowledge of the predictors still influenced the EQ process: low frequencies were cut and high frequencies were boosted in order to raise the HC and it was attempted

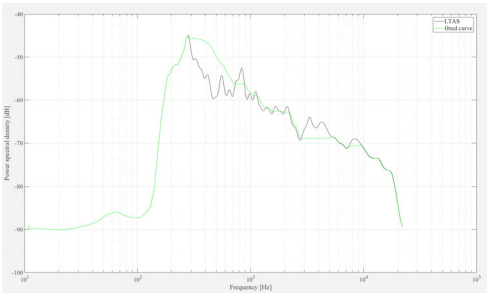

Fig. 2: the curve for the mid-range spectral peakiness metric (green) is fitted to the version 1 LTAS (black). Potentially unpleasant-sounding peaks in the LTAS lie above the curve, while the remaining frequency areas lie below it (Hermes et al., 2017).

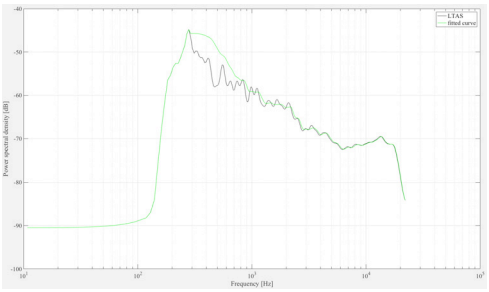

Fig. 3: the curve for the mid-range spectral peakiness metric (green) is fitted to the version 2 LTAS (black). Peakiness is greatly reduced in comparison on version 1 and an additional HF boost leads to a HC increase of 6 harmonics.additional HF boost leads to a HC increase of 6 harmonics.

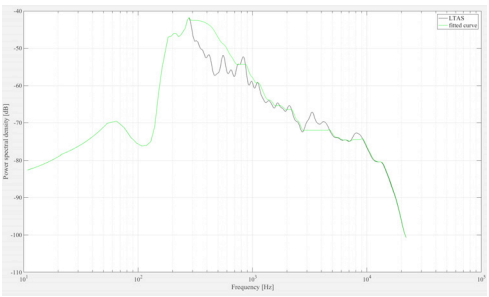

Fig. 4: the curve for the mid-range spectral peakiness metric (green) is fitted to the default version. LTAS (black). Mid-range spectral peakiness is greater than in version 1 but lower than in version 2. The HC is lower than in both versions 1 and 2 due to increased area around the fundamental frequency.

to reduce some of the unwanted spectral peaks. Since the increased audibility of spectral peaks of target sounds in mixes appears to contribute positively to clarity (section 2), it was attempted to remove energy near peaks in the vocal from surrounding instruments with EQ. However, since there were many small adjacent peaks, this was not feasible in practice.

A second version of the mix was created, guided by the clarity predictors (version 2). The HC of this version was increased by another 6 harmonics. Using the computational model of mid-range spectral peakiness introduced in section 2 (Hermes et al, 2017), it was attempted to eliminate peakiness entirely. This process required a large number of additional fine EQ adjustments. The resulting version was spectrally much flatter and free from mid-range spectral peakiness. The LTAS and mid-range spectral peakiness for versions 1 and 2 are shown in figures 2 and 3. A third version was also created where all EQ was removed entirely for comparison (default version). In this version, the HC was lower than in version 1 by 6 harmonics. Mid-range spectral peakiness was greater than in version 1 but lower than in version 2. For comparison, the LTAS and peakiness for this version is shown in fig. 4.

Since the author's previous knowledge of the predictors had

influenced the EQ of the first version, the difference between versions 1 and 2 was small. The author felt that version 2 sounded somewhat smoother and clearer than version 1. Both versions sounded much clearer than the default version, which appears to indicate that the HC is a useful clarity predictor in this case: the HC was considerably lower in the default version than in versions 1 and 2. In the following, each version is described in greater detail.

When played in isolation, the default version sounded fairly natural but had excessive energy in the 300Hz area. This gave it a 'dull', 'muffled' and 'muddy' timbre. There also seemed to be some unwanted noise and distortion as a result of the recording process, leading to a 'fuzzy' quality. In the mix, the default version sounded particularly unclear and not separate enough from other sounds.

Version 1 sounded much clearer than the default version, both in isolation and in the mix. The increased high frequency to low frequency balance made it appear more 'present' and 'thin'. Version 1 did sound somewhat less natural than the default version when played in isolation but clarity and separation was much improved in the mix. However, by boosting high frequencies and cutting low frequencies, an unpleasant peak in the high frequency are was made more obvious. The author struggled to identify the exact location of this peak. This lead to a degree of 'harshness' and the overall timbre did not appear to be tonally balanced. Additional boosts and cuts seemed to either increase this harshness or make the timbre 'duller' like in the default version.

The use of the clarity predictors in version 2 allowed the author to locate the aforementioned peak in the 3—5khz area. It was possible to flatten the peak without affecting other areas, as the model had correctly identified it as contributing to unpleasant mid-range spectral peakiness. Hence, the peak can be seen in the LTAS for version 1 (figure 3) but not in the LTAS for version 2 (figure 4), where it had been removed. Additional, smaller peaks could also be treated. In the author's opinion, the resulting sound was smoother, clearer, and more present than in the other two versions with reduced sharpness. The noise and distortion in the default version had been altered to a pleasant 'sizzle'. Despite this improvement, version 2 sounded less natural and more processed than version 1. In the mix, the difference between versions 1 and 2 was small.

Informal discussions with other audio professionals (students, lecturing staff and sound technicians at the University of Westminster with no previous knowledge of the predictors) revealed some disagreement. While the general consensus was that both equalized versions were much clearer than the unequalized version, some listeners felt that version 1 was clearer than version 2, others the other way around. While some listeners did also describe version 2 as smoother, they also commented on the fact that it sounded overprocessed and unnatural. It seems that some listeners perceive naturalness as an important aspect of clarity, while others do not (including the

author). It appears that the HC is a suitable clarity predictor, since the HC in versions 1 and 2 was notably higher than in the default version. Peakiness still seems relevant, albeit to a lesser degree, since the version containing no peaks (version 2) was not perceived as clearer by all audio professionals consulted. It is concluded that the HC would be useful in an overall clarity model. The contribution of spectral peaks to naturalness, and therefore spectral clarity, however, needs to be investigated further. In the following sections, starting points for this investigation are presented.

When is a spectral peak part of natural character of sound?

It is possible that some spectral peaks are perceived to be part of the natural character of a sound while others are considered unpleasant. If naturalness is important for clarity, it is possible that an entirely flat spectrum resulting from fine EQ adjustments can reduce clarity much in the same way as a peaky spectrum can in other cases. Therefore, it is possible that some spectral peaks appear to increase clarity while others reduce it. However, what constitutes a natural timbre may be difficult to measure, as discussed in the following part of the paper.

The impact of spectral peaks on clarity may also, to some degree, depend on the instrument. While some instruments tend have fewer peaks (e.g. cello), others feature natural, strong resonances, such as the Erhu (Chinese violin) (Hermes et al., 2017). It is possible that recordings of acoustic instruments and voices sound clear when their spectra resemble their natural, unrecorded spectra. However, spectra can only be measured by recording sound and the recording process always introduces spectral distortions.

Spectral clarity may be even more difficult to establish for newly created timbres such as electronic synths. Zagorski-Thomas (2007) relates musical elements to "physical manifestations of emotions, gesture and being in space", describing music recordings as sonic metaphors for physiologically and culturally determined gestures and morphologies. Further, Zagorski-Thomas (2017) states that multi-track and electronic music can be interpreted as something impossible and yet understandable. Similarly, Théberge (1997) links adjectives used to describe low-level mix parameters to bodily sensations. It is possible that listeners might still agree on the clarity of newly synthesized sounds, since there might be a shared understanding of what constitutes a 'believable' timbre. In the author's experience, combinations of contrasting timbres may result in greater separation in electronic music mixes. The author tends to prefer 'edgy', 'hard' timbres for her own productions, some of which have strong peaks (e.g. chiptune and 8-bit timbres). For that reason, it would be useful to establish whether there are spectral areas that always contribute to clarity in the same way for these sounds.

Phase issues

The LTAS of sounds can be useful for measuring timbral attributes and is used in a multitude of research studies (a literature review on clarity can be found in Hermes et al., 2017). However, this measurement method ignores the relative phases of spectral components and short-term spectral fluctuations. Bregman (2007) states that sounds with the same frequency content but differing phases can sound different, which can also influence separation in sound mixtures. Laitinen et al. (2013) confirm that humans can perceive differences in the phase spectrum of otherwise identical sounds and that the phase spectrum affects the perceived timbre, especially in sounds with lower fundamental frequencies. Toulson (2008) argues that therefore, it can be difficult to fix spectral problem areas in the mix. He points out that an overuse or incorrect implementation of EQ can be detrimental to the sound quality due to the resulting phase distortion. Several listeners confirmed that the second version of the "Serendipity" vocal sounded over-processed, which may have been due to distortions in the phase spectrum resulting from the many fine EQ adjustments.

The recorded vocal take was 30 seconds long, hence the predictors would have ignored any large spectral fluctuations throughout. It could be argued that in order to keep the complexity low, it is useful to fully understand in which way features of steady-state spectra contribute to clarity before considering temporal factors and phase spectra. However, a fully functional clarity model would most likely be more accurate if these parameters were also considered.

Sounds in isolation and in the mix

As indicated in section 2, the interaction of the target sound (in this case the vocal) with the backing track (rest of the mix) also needs to be considered, due to masking and fusion phenomena. Parts of the target sound spectrum may become masked by the backing track. According to the American National Standards Institute (ANSI/ASA: 2013), masking is the process by which the threshold of audibility for one sound is raised by the presence of another, masking sound. Instruments occupying similar frequency regions in music mixes are likely to mask each other. Partial masking can reduce the loudness of the target in the mix (Ma et al.: 2014) and is therefore likely to lead to a reduction in clarity. Pestana and Reiss (2014) point out that in music mixes, EQ should be applied to ensure that no element masks any of the frequency content of lead sounds. Overall, it is apparent that masking can reduce clarity but it is not clear whether there may be any frequency areas that particularly should be unmasked and how this relates to the audibility of spectral peaks in the target sound. As mentioned in section 2, it is possible that natural peaks in the target sound should be unmasked. Phenomena such

as upwards masking, where lower frequencies in the masker can mask higher frequencies in the target and temporal phenomena like forward and backward masking add to the complexity (Moore: 2012).

The audibility of target sounds can also be compromised when not enough separation exists from the backing track, even when the target sound is unmasked. Fusion and separation phenomena are assessed in the field of auditory scene analysis. Auditory scene analysis (ASA) is the process of forming mental representations of individual sound sources from the summed waveforms that reach the ears. The ASA process consists of the following two conceptual stages (Bregman: 2007). First, the auditory system divides the input into its constituent atomic units, i.e. packages of acoustic evidence (*segmentation*). Following segmentation, any packages that appear to have arisen from the same source are either *grouped* (to form a stream for a given source) or *segregated* (to form separate streams for different sources). Elements that fall in the same auditory stream are perceived as stemming from the same sound source. Grouping and segregation are related to the perception of separation in music mixes. The factors that determine whether sounds are fused or separated are complex but spectral similarity between target and backing track is particularly important.

Woszczyk and Bregman (2005, pp. 13–25) state that the ear is more easily able to follow a sound in a mix if it has a unique timbre, leading to greater separation from other sounds. They state that unique timbres usually have obvious features that the listener can track over time. Strong spectral peaks may constitute an obvious feature, increasing clarity in mixes. Bregman (1990) provides the example of classically trained singers who can enlarge their pharynx cavity and produce a strong resonance in the mid-frequency area, an area that is usually not occupied by a lot of other instruments in the orchestra. This resonance allows e.g. opera singers to be heard over their accompaniment, even without the availability of amplification.

A potentially similar technique appears to be prevalent in pop singing, i.e. singing with ‘twang’. ‘Twang’ is a vocal timbre produced through increased subglottal pressure, leading to increased energy in the first two formants, decreased energy in formants 3 and 5 and an overall higher sound pressure level (Sundberg and Thalén: 2010). It is possible that ‘twang’ leads to a spectral peak around the first two formants that can increase audibility and separation of vocals in mixes. If this peak were to be removed with EQ, clarity may decrease as a result. Therefore, spectral flatness in lead sounds such as vocals may decrease spectral clarity in mixes. It would be interesting to establish whether there are specific, generalizable frequency areas that contribute to the clarity of lead sounds. Izhaki (2008, p.251) suggests that the spectral area for vocal clarity area lies around 2kHz – 9kHz. An increase in energy in this area is likely to correlate with an increase in the HC.

Holistic perception of clarity in music mixes

It is possible that target clarity in mixes is not an isolated concept but ties in with the holistic perception of the overall mix. The clarity of each sound in the mix may bias the perception of the clarity of each other sound. As established in section 2, overall tonal balance is another important parameter of music mixes. Pestana et al. (2013) explain that spectra of professionally produced commercial recordings show consistent trends, which can roughly be described as a linearly decaying distribution of around 5 dB per octave between 100Hz and 4000Hz, becoming gradually steeper with higher frequencies, and a severe low-cut around 60Hz. It is possible that overall tonal balance influences the perception of clarity of each sound therein.

Similarly, so far, only spectral parameters of clarity have been investigated. For a complete model of clarity in music mixes, spatial and intensity related factors should also be considered, alongside temporal changes in all these factors.

Discussion and suggestions for further research

As established in the last section, spectral clarity is likely to be a multifaceted, complex concept. The two predictors of single sound spectral clarity still appear to be important, in particular the HC. Mid-range spectral peakiness can help measure the influence of resonances on the clarity of isolated sounds. In music mixes, the influence of peaks on clarity appears to be more complex. For a better overall spectral clarity model, further research needs to be carried out. First, it would need to be established which spectral peaks constitute a part of the natural character of a sound. This is especially difficult for newly synthesized timbres, where no natural reference exists. Naturalness is likely to influence clarity; however, it should be established how strongly these two attributes correlate. Second, it would be useful to assess how spectral parameters that are not measured by the LTAS contribute to spectral clarity. Phase spectra appear to be particularly important in this context, since phase distortions, as introduced by EQ, can be detrimental to sound quality. Third, the impact of the complex relationship between the target and backing track spectra on clarity needs to be understood further. Not just masking phenomena, but also fusion and separation between target and backing track need to be taken into consideration. It would therefore be useful to reassess spectral target clarity in mixes using computational auditory scene analysis (CASA) models. The clarity of sounds appears to depend on fusion phenomena in a complex way. Fourth, it would be useful to assess the impact of non-spectral parameters on clarity, such as spatial, intensity related and temporal factors.

Lastly, in order to measure mix quality successfully, it would be necessary to measure all high-level parameters of mixes, that is, not only clarity and separation, but also balance, impact and interest and the freedom from technical faults. Some context-specific parameters could be measured through comparison to a reference, e.g. mixes of a similar fashion or style. Since the perception of each of these mix parameters may not be isolated, their influence on each other should be assessed.

When spectral clarity is more fully understood, the implications of the findings on other research areas should also be explored. For instance, if specific resonances do increase clarity in singing, it would be useful if singers could integrate this knowledge into their training. Similarly, sound synthesis tools could be based on more perceptually informed parameters, such as a spectral clarity control. Since all research methods presented in this paper (grounded theory, knowledge engineering and autoethnography) have limitations, ultimately, an interdisciplinary approach may yield the most useful results. Therefore, in the author's opinion, scientists and creatives should continue to collaborate in furthering the understanding of music mix quality, since this is likely to lead to a rich, holistic understanding of the subject.

The aim of the current study was to test whether the HC and mid-range spectral peakiness may be able predict clarity for sounds in mixes and to develop starting points for follow-on research. To conclude, the HC appears to be a strong predictor of spectral clarity in music mixes. Mid-range spectral peakiness also seems useful but should be supported by a metric for naturalness. Additionally, metrics of phase and masking, as well as a CASA model should be included in an overall spectral clarity model. Follow-on research should not only consider spectral clarity but also consider temporal, spatial and intensity related factors. An interdisciplinary approach is likely to produce useful results.

References

ANSI/ASA S1.1 (2013) 'American National Standards. Acoustical Terminology'. In: *S1. American National Standards Institute*, New York, USA.

Bech, S., Zacharov, N. (2006) *Perceptual audio evaluation: theory, method and application*. *Hoboken, USA:* John Wiley & Sons.

Bell, A. P. (2014), 'Trial-by-fire: A case study of the musician–engineer hybrid role in the home studio'. In: *Journal of Music, Technology and Education*, 7,1, pp. 295–312.

Benito, A. L. and Reiss, J. D. (2017) 'Intelligent multitrack reverberation based on hinge-loss markov random fields'. In: *Audio Engineering Society International Conference (Semantic Audio)*, Erlangen, Germany, June 21st.

Biederman, R. and Pattison, P. (2014) *Basic Live Sound Reinforcement: A Practical Guide for Starting Live Audio*. Abingdon: Focal Press.

Bregman, A.S. (1990), *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: Bradford Books, MIT Press.

Bregman, A.S. (2007) 'Auditory scene analysis.' In: A.I. Basbaum, A. Koneko, G.M. Shepherd and G.Westheimer (eds.) *The Senses: A Comprehensive Reference*, *Vol. 3, Audition*, P. Dallos & D. Oertel (Volume Eds.) San Diego: Academic Press, pp. 861–870.

Conetta, R., Brookes, T., Rumsey, F., Zielinski, S., Dewhirst, M., Jackson, P., Bech, S., Meares, D. and George, S., (2015) 'Spatial Audio Quality Perception (Part 2): A Linear Regression Model' In: *Journal of the Audio Engineering Society*, 62, 12, pp. 847–860.

De Man, B. and Reiss, J. D. (2017) '*Ten Years of Automatic Mixing'*. In: *Proceedings of the 3rd Workshop on Intelligent Music Production*, Salford, UK, September 15th.

De Man, B. and Reiss, J.D. (2013) 'A Semantic Approach To Autonomous Mixing.' In: 2013 *Art of Record Production Conference*, Québec, Canada, July 12th–14th.

Dugan, D. (1975) 'Automatic microphone mixing'. In: *Audio Engineering Society 52nd Convention,* New York, USA, June 23rd.

Dwyer, S. C. and Buckle, J. L. (2009) 'The Space Between: On Being an Insider-Outsider in Qualitative Research'. In: *International Journal of Qualitative Methods* [Online] March 1st. Available at: https://doi.org/10.1177/160940690900800105 (Accessed February 2018).

Ellis-Peterson, H. (2017) 'UK music industry gets boost from 12% rise in audiences at live events.' In: The Guardian. [Online] July 10th. Available at: https://www.theguardian.com/music/2017/jul/10/uk-music-industry-gets-boost-from-12-rise-in-audiences-at-live-events (Accessed February 2018).

Ellis, C., Adams, T. E. and Bochner, A. P. (2011) 'Autoethnography: An Overview'. In: *Forum: Qualitative Social Research*, 21, 1 [Online] January 10th. Available at: http://www.qualitative-research.net/index.php/fqs/article/view/1589/3095 (Accessed February 2018).

Fenton, S., Wakefield, J. (2012) 'Objective Profiling of Perceived Punch and Clarity in Produced Music.' In: *Audio Engineering Society 132nd Convention*. Budapest, Hungary, April 26th-29th.

Glaser, B. G., Strauss, A. L. (1967) *The discovery of grounded theory: strategies for qualitative research. Chicago:* Aldine.

Gonzalez, E. P. and Reiss, J. D., (2010) 'A real-time semi- autonomous audio panning system for music mixing,' In: *EURASIP Journal on Advances in Signal Processing.*, 2010: 436895.

Hafezi, S. and Reiss, J. D. (2015) 'Autonomous multitrack equalization based on masking reduction'. In: *Journal of the Audio Engineering Society*, 63, 5, pp. 312-323.

Hermes (2018, under review) 'What makes a clear vocal mix?'. In: Toulson, R., Paterson, J., Hodgson, J. and Hepworth-Sawyer, R. (eds.) *Innovation In Music: performance, production, technology and business.* London: Routledge.

Hermes, K., Brookes, T., Hummersone, C. (2017) 'Towards Measuring Music Mix Quality: the factors contributing to the spectral clarity of single sounds.' *PhD thesis*, University of Surrey, Guildford, UK.

Izhaki, R. (2008) *Mixing Audio: Concepts, Practices and Tools, Second Edition*. Oxford: Focal Press, Oxford, UK, p. 251.

Laitinen, M. V., Disch, S. and Pulkki, V. (2013) 'Sensitivity of Human Hearing to Changes in Phase Spectrum'. In: *Journal of the Audio Engineering Society,* 61, 11, pp. 860–877.

Ma, Z., De Man, B., Pestana, P. D. L., Black, D. A. A., Reiss, J. D. (2015) 'Intelligent multitrack dynamic range compression.' *Journal of the Audio Engineering*, 63, 6, pp. 412-426.

Ma, Z., Reiss, J. D. and Black, D.A.A. (2014), 'Partial Loudness in Multitrack Mixing.' In: *Audio Engineering Society 53rd International Conference: Semantic Audio*. London, UK, January 26–29.

Moore, B.C.J., 2012, *An introduction to the psychology of hearing*. *Bingley*: Emerald.

Pestana, P. D. and Reiss, J. D. (2014) 'A cross-adaptive dynamic spectral panning technique'. In: *17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, Sep 1st-5th.

Pestana, P.D., Ma, Z., Reiss, J.D., Barbosa, A., Black, D.A.A., 2013, 'Spectral Characteristics of Popular Commercial Recordings 1950-2010.' In: *Audio Engineering Society 135th Convention.* New York, USA, October 17th–20th.

Pestana, P.D., Reiss, J.D. (2014), 'Intelligent audio production strategies informed by best practices.' In: *Audio Engineering Society 53rd International Conference*, London, UK, Jan 26th–29th.

Pras, A., Guastavino, C. and Lavoie, M. (2013) 'The impact of technological advances on recording studio practices'. In: *Journal of the Association for Information Science and Technology*, 64, 3, pp. 612–626.

Scott, J. (2014), 'Automated Multi-Track Mixing and Analysis of Instrument Mixtures.' In: *22nd ACM International Conference on Multimedia*. Orlando, Florida, November 3rd–7th.

Sundberg, J., Thalén, M. (2010) 'What is "Twang"?' In: *Journal of Voice*. 24, 6, pp. 654–660

Théberge, P. (1997) *Any Sound You Can Imagine: Making Music/Consuming Technology*. *Hanover*: Wesleyan University Press.

Toulson, R. (2008) 'Can We Fix It? – The Consequences Of 'Fixing It In The Mix' With Common Equalisation Techniques Are Scientifically Evaluated.' In: *Journal of the Art of Record Production* [Online]. November. Available at: http://arpjournal.com/can-we-fix-it-–-the-consequences-of-'fixing-it-in-the-mix'-with-common-equalisation-techniques-are-scientifically-evaluated/ (Accessed: February 2018).

What Uni (2017) 'Music Degrees.' What Uni?, London, UK [Online]. Available at: https://www.whatuni.com/degree-courses/search?subject=music (Accessed February 2018).

Wilson, A. and Fazenda, B. (2016a) 'An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge'. In: *2nd Workshop on Intelligent Music Production*, London, UK, Sep 13th.

Wilson, A. and Fazenda, M. (2016b) 'Perception of Audio Quality in Productions of Popular Music'. In: *Journal of the Audio Engineering Society*, 64, 1/2, pp. 23–34.

Woszczyk, W. and Bregman, A.S. (2005) 'Creating mixtures: The application of auditory scene analysis (ASA) to audio recording.' In: K. Greenebaum and R. Barzel (Eds.) *Audio Anecdotes III: Tools, tips and techniques for digital audio*. *Natick:* Peters, pp. 13– 25.

Zagorski-Thomas, S. (2007) 'The Musicology of Record Production.' In: *Twentieth-Century Music*, 4, 2, pp. 189–207.

Zagorski-Thomas, S. (2017) 'Looney Tunes: Sonic Cartoons and Semantic Audio'. In: *The 12th Art of Record Production Conference. Mono: Stereo: Multi*. Stockholm, Sweden, December 1st–3rd.