Expectation elicits music-evoked chills

Rémi de Fleurian[1,*], Ana Clemente[1,2], Emmanouil Benetos[1], and Marcus T. Pearce[1,3,*]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany
[3]Department of Clinical Medicine, Aarhus University, Denmark
[*]Corresponding authors: remi.defleurian@gmail.com, marcus.pearce@qmul.ac.uk

Abstract

Music-evoked chills (MECs) are physiological responses to pleasurable events in music. Existing research on properties of music that elicit MECs has focused on low-level acoustic features in small samples of music. We created a large dataset of over 1,000 pieces of music timestamped with MECs and used computational methods to predict MEC onsets from both low-level acoustic features and high-level musical expectations. A machine learning classifier was trained to distinguish MEC onsets from non-MEC passages in the same pieces. The results show that MEC onsets are predicted better than chance and corroborate evidence for acoustic elicitors of chills with a much larger dataset. They also produce new empirical evidence that MECs are elicited by expectation, which is a more effective predictor of MEC onsets than acoustic elicitors, and may generalise to pleasurable experience in other domains such as language comprehension or visual perception.

1 Introduction

Music-evoked chills (MECs) are shivers, goosebumps or tingling sensations experienced while listening to music. They are experienced by most individuals and are invariably reported as being intensely pleasurable, often corresponding to moments of peak affective intensity in a musical piece [1]. MECs have attracted scientific interest as an observable physiological response reflecting an unobservable internal state of pleasure while listening to music. MECs have proven to be robust activators of the brain's reward system in neuroimaging experiments [2–4], neuropsychological comparisons [5, 6] and pharmacological interventions [7, 8]. However, the causes of MECs have remained challenging to understand. MECs are thought to arise through the interaction of three factors: the listener, showing particular traits and in a particular state; the music, containing particular eliciting characteristics; and the surrounding context, including social factors and concurrent tasks [1]. The present research identifies eliciting characteristics of music that predict the onset of individual chills.

Three general kinds of elicitor have been identified in previous research: low-level acoustic elicitors (e.g. loudness changes), high-level musical elicitors (e.g., melodic expectation) and emotional elicitors [1]. Research to date has focused on acoustic elicitors, which are relatively straightforward to extract from an audio file, but has been limited to small samples of participants and musical pieces. While expectation has long been considered an important elicitor, this theoretical prediction has yet to be robustly tested. Expectation mechanisms are of broader interest since prediction is thought to be an important general component in neural computation [9], operating in other domains such as language [10, 11] and visual perception [12, 13]. In the present research, we address both issues by creating a large dataset of over 1,000 pieces of music supplied by over 900 individuals containing precise onset times for 1,806 MECs, which we predict with machine learning models that combine acoustic features with computational measures of high-level expectation. Before presenting the results, we review existing research guiding the choice of acoustic and musical elicitors used in the present research.

Acoustic elicitors of MECs are low-level properties of the auditory signal. Early research on MECs identified a relationship with sudden dynamic changes, by analysing music scores for passages reported to elicit MECs by survey participants [14]. This relationship was confirmed empirically in subsequent research [15–22], in which loudness was extracted [or manually inspected from music scores, 19] around the onset of MECs experienced by participants when listening to music in a lab environment, which they reported either retrospectively or continuously by pressing a button or moving a slider. Loudness is by far the best-documented acoustic correlate of MECs, but other relationships have also been identified using similar methods, suggesting that occurrences of MECs tend to correlate with increases in roughness, sensory dissonance or fluctuation strength [16–18, 21, 23], increases in sharpness or brightness [16–18, 20], high spectral centroid and spectral flux [16], high event density [16, 21, 22], or expansion of the frequency range in a high or low register [19, 22]. Using a causal intervention, Bannister [24] experimentally manipulated loudness and brightness in two musical passages that

had elicited MECs in previous research [16]. In one of the musical passages, MECs were experienced more frequently if loudness was increased or if brightness was decreased (in contradiction with previous findings), therefore demonstrating a causal effect of loudness and brightness on MECs, as opposed to the correlational findings discussed above.

Musical elicitors of MECs are high-level properties of musical structure. In a seminal study from 1991, John Sloboda [14] found that musical passages associated with MECs (in 57 pieces supplied by 83 respondents to a questionnaire) included new or unprepared harmonies, sudden textural changes, melodic appoggiaturas, enharmonic changes, specific melodic or harmonic sequences, or prominent events arriving earlier than prepared for, in decreasing order of frequency. These self-reported effects of melodic and harmonic properties on MECs were confirmed in subsequent survey-based and empirical research [15, 16, 19, 25–27], notably through the identification of effects of structural transitions and alterations such as changes in tonality [25]. Rhythmic properties [27–29] and vocals [25, 27] were also found to be involved, although there is a lack of specificity regarding which exact properties were associated with MECs. Additional findings revealed effects of crescendi, build-ups and climaxes [15, 16, 22, 25, 28, 30], as well as textural changes [14, 15, 22, 28], notably through the entrance of new instruments or the interplay between solo and background instruments [15, 16, 19, 25, 26]. It has been argued that research on musical elicitors reflects a hypothesised effect of musical expectation on MECs [2, 14, 31–37], positing that most of the musical elicitors of MECs listed above could be related to violations of expectation. An effect of musical expectation on emotional and aesthetic experience of music has long been hypothesised [38, 39] and has since been confirmed empirically [40–43]. An effect of expectation on MECs, however, remains untested.

Emotional elicitors of MECs also deserve mention, including subjectively perceived valence, emotionality and meaning in music. Research has identified effects of emotional elicitors such as positive valence [44], positive and negative valence [25, 30, 45], perceived emotionality [16, 17, 30, 46] and perceived meaning [25, 26, 47]. However, such characteristics of musical stimuli are difficult to quantify precisely and objectively, especially as continuous features, since they rely on some degree of subjective interpretation and often relate to large sections of music or entire pieces rather than specific, identifiable events. For this reason, emotional elicitors were not considered in the present research, but we acknowledge that they are deemed potent elicitors of MECs and deserve further attention in future research. The Chills in Music (ChiM) dataset, accompanying a systematic review of research on MECs [1], consists of 1,022 musical pieces used in previous empirical research on MECs. These musical pieces were found to have more negative valence scores (according to Spotify API data) than a carefully matched set of tracks [48]. However, the ChiM dataset lacks the precise timing of MEC onsets and consistent information about the exact versions of the pieces of music found to elicit MECs. For these reasons it could not be used in the present research. Mori [29] asked 54 participants to listen to individually self-selected pop songs which induced MECs or tears for them while indicating MECs or tears with button presses. Using multi-class ridge regression and a range of acoustic features, a classification accuracy of 43% was achieved for the onsets of MECs, which was significantly better than chance. Classification performance was driven most strongly by the minor mode immediately preceding MECs (highlighting a possible relationship with expressed valence [48]), higher event density and rhythmic entropy at the onset of MECs, and higher spectral flux and rhythmic entropy after the onset of MECs (replicating hypothesised effects of rhythmic properties [1]).

This existing research on chills has identified acoustic features of music that seem to be related to MECs but has done so for relatively small samples of individuals and musical stimuli. Research has also identified higher-level musical elicitors many of which seem to be related to expectation but the role of expectation has not been robustly validated. The present research aims to address both issues: first, we collect a large dataset time-stamped with onsets of MECs; second, we develop a comprehensive battery of computational measures of low-level acoustic features and high-level expectation hypothesised to elicit chills; third, we use machine learning to compare the impact of acoustic elicitors and expectation-based musical elicitors on MEC classification performance.

2 Results

2.1 Features

Two kinds of feature were computed: low-level acoustic features and high-level expectation-based musical features. The acoustic features included: loudness using a measure of the spectral envelope (Envelope); timbre using measures of roughness, flatness, brightness, spectral centroid, spectral flux, spectral entropy and spectral spread; tonality using measures of key distance and harmonic change (HCDF); and presence of vocals (Vocals). The high-level expectation-based features included melodic entropy and melodic information content. Melodic entropy reflects the prospective uncertainty of a model making probabilistic predictions for the pitch of the next note in a melody based on both a learned model of melodic regularities in a large corpus of Western tonal music

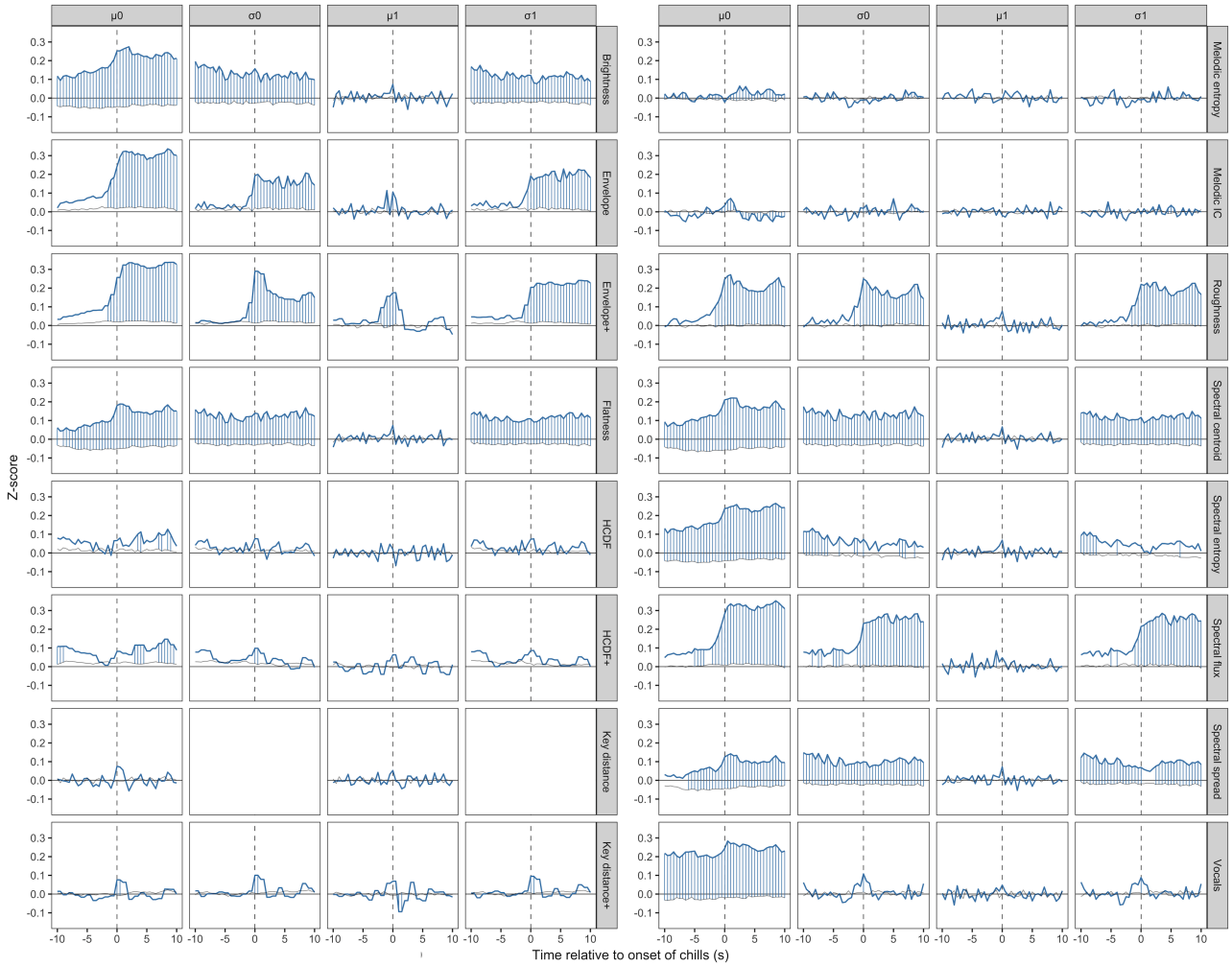

Figure 1: Each row of plots corresponds to a feature (see Methods), with each of the four columns corresponding to the mean (μ) and the variance (σ) of the zeroth (0) and first derivative (1), respectively. The x-axes represent 20-s excerpts centred around MEC onsets ($x = 0$). The y-axes represent normalised feature values (Z-scores) such that the horizontal line at $y = 0$ represents a Z-score of 0 (i.e., the mean feature value). Thick blue lines represent feature values averaged over all excerpts causing MECs. Thin grey lines represent feature values averaged over all control excerpts. Vertical blue lines denote frames for which there were significant differences between the excerpts causing MECs (thick blue line) and control excerpts (thin grey line). Permutation tests were not conducted for key distance variance due to the way these features were computed (see Methods).

and dynamic learning within each piece of music [49, 50]. Information content reflects the unexpectedness or surprisal of the pitch that actually follows. See Methods for details. The impact of each these features on MECs was hypothesised based on previous research (see Introduction).

2.2 Feature differences between MEC and non-MEC excerpts

Figure 1 compares each individual feature in 20-s excerpts centred around the onsets of MECs to 20-s excerpts taken from the same pieces of music but not associated with MECs. Significant differences were obtained by comparing the observed differences with a null distribution computed by permuting the labels 5,000 times (see Methods). The results can be divided into four categories: features that were consistently higher than control excerpts around MEC onsets; features showing a sharp increase around MEC onsets; features showing a significant increase or decrease around an MEC onset followed by a return to the level preceding the MEC; and features that showed no significant differences from control excerpts.

Examining first the averaged features (identified with μ_0), MECs were characterised by elevated brightness, flatness, spectral centroid, spectral entropy, melodic entropy and presence of vocals, as well as sharp increases in envelope, roughness, spectral flux and spectral spread. While several frames showed also significant differences regarding the harmonic change detection function (HCDF) and its associated feature computed with a larger

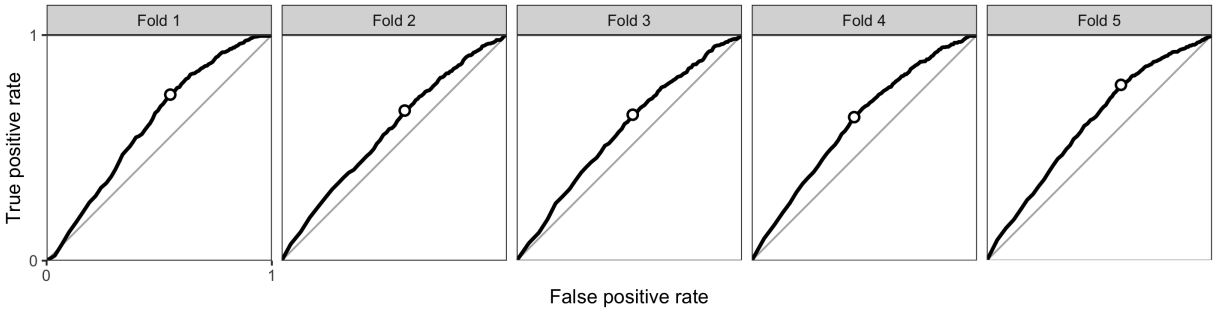

Figure 2: Receiver-operating characteristic (ROC) curves for each cross-validation fold of the trained SVM. Overall AUC for the model was computed by averaging the AUCs for each curve. The threshold returning the highest F_β value is visualised by a circle on each curve and corresponds to the point at which all evaluation metrics were retained before being averaged to return overall evaluation metrics for the model. Chance performance is indicated by the line of equality ($x = y$), with curves above this line indicating that the true positive rate (also known as recall, sensitivity, hit rate or power) exceeds the false positive rate (also known as fall-out or type-I error).

sliding window (HCDF+), the overall differences were less convincing than for other features. Key distance (only when computed with a larger sliding window: Key distance+) and melodic information content significantly increased while melodic entropy significantly decreased around MEC onset followed by a return to their pre-MEC level, suggesting that MECs are elicited by isolated musical events with increased key distance and melodic information content but reduced entropy.

The variance in these features (as identified with σ_0) largely followed similar patterns. The only exception was the presence of vocals, which only contained a single frame with a significant difference between MECs and controls, located at the exact onset of MECs. Note that the presence of vocals was the only binary feature and relatively slow-moving compared to the other features due to how it was preprocessed. Nonetheless, this can be interpreted as vocals being more likely around MEC onsets (as seen with μ_0) and MECs being slightly affected by increased variance in the presence of vocals around their onsets.

Significant effects were much more sparse when looking at the rate of change of each feature (as identified with μ_1). A convincing peak only occurred for envelopes computed with a larger sliding window, providing further evidence that sudden, large changes in loudness are associated with MECs. Variance in these rates of change (as identified with σ_1) almost exactly followed the variance in the original features (as identified with σ_0), consistent with the intuition that when variance in a feature is substantially heightened, this is reflected in the variance of its rate of change.

2.3 Classifying MECs from acoustic and expectation-based musical features

The dimensionality of the feature set (62 features as shown in Table 1) was reduced using PCA with 15 principal components retained (see Methods). These were used to train a linear support vector machine (SVM) to distinguish 500-ms frames within a 5-s segment that also contained a reported onset of chills (positive label) from all other 5-s segments (negative label) (see Methods and Supplementary Information). The SVM was trained and tested, splitting the data track-wise using five-fold cross-validation. The receiver-operating characteristic (ROC) curves showing classification performance for each fold are shown in Figure 2, along with the threshold for the highest F_β values for each curve using $\beta = 2$. In terms of overall performance, the SVM exhibited an AUC of 0.597, F_β of 0.167, and balanced accuracy (the arithmetic mean of the accuracy in each class) of 0.580. See the Supplementary Information for a model comparison.

Finally, the contribution of each feature to the SVM in terms of feature importance (see Methods) is visualised in Figure 3. Melodic entropy and information content were the best predictors of MECs, with melodic entropy reaching a value of 1, meaning it was the most important predictor on all five cross-validation folds. Variance of the first-order difference of melodic entropy, mean spectral flatness, spread, and centroid followed these features. The least contributing features were mean first-order differences of brightness, roughness, spectral entropy, spread and centroid.

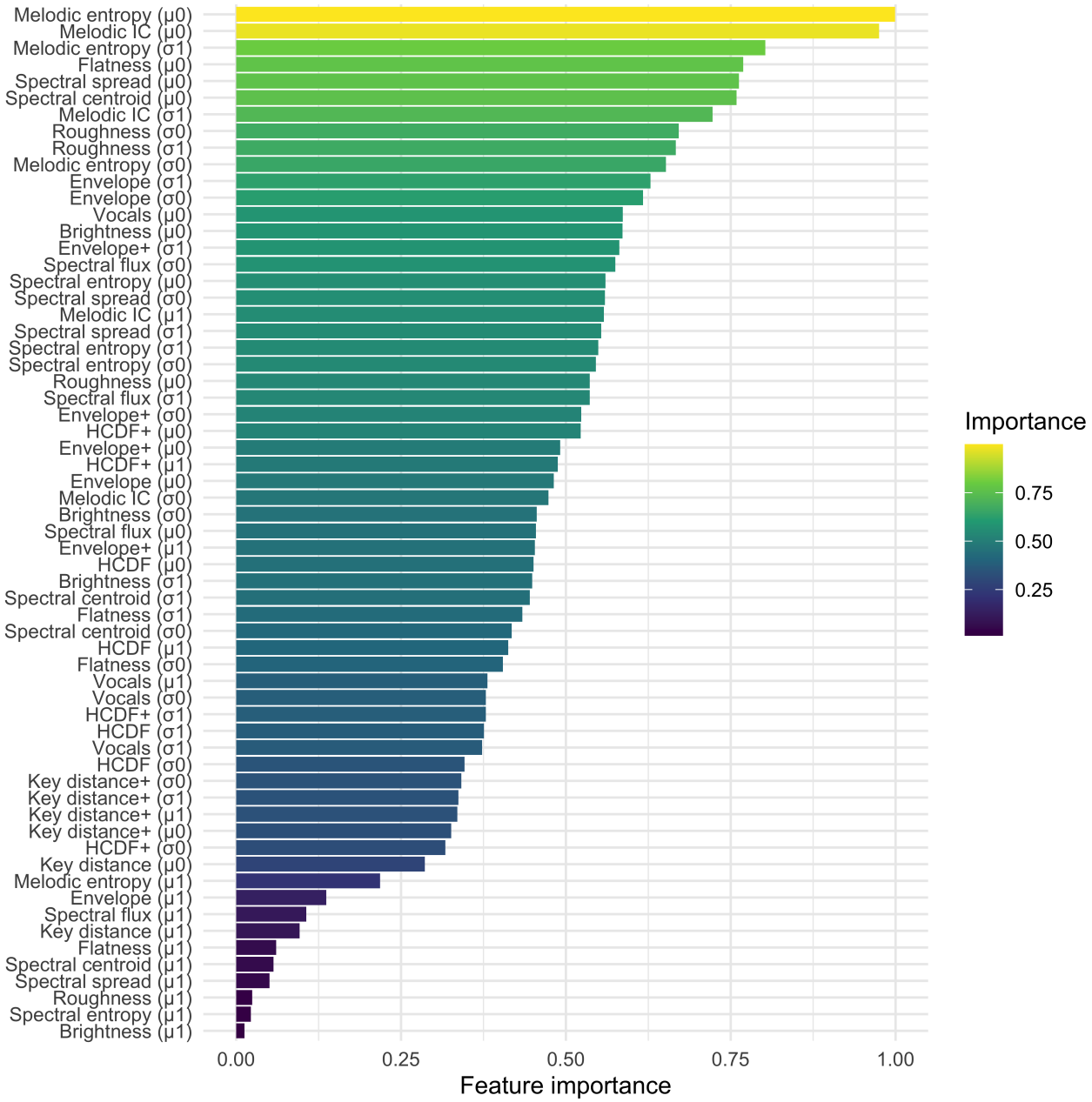

Figure 3: Relative feature importance for the SVM. Feature importance is rescaled linearly such that the feature contributing the most to each cross-validation fold receives a score of 1 (e.g., mean melodic entropy) and the feature contributing the least receives a score of 0, which no feature did for all five folds.

3 Discussion

3.1 Expectation elicits chills

These results provide convincing evidence that the stimulus-based elicitors of MECs include mechanisms related to pitch expectation, newly demonstrated here, alongside acoustic cues, as demonstrated in previous research but corroborated here using machine learning methods on a much larger sample of participants and musical stimuli. In particular, expectation-based measures of pitch surprisal and predictive uncertainty increased classification performance (see Supplementary Information) and were convincingly higher in feature importance than acoustic features such as loudness, timbre (spectral flatness, centroid and spread), roughness and presence of vocals. The permutation testing suggests that MECs are elicited by surprising events (indicated by an increase in information content at MEC onset) in the context of relatively low surprisal (lower information content than baseline surrounding the MEC onset). MECs are also associated with generally higher predictive uncertainty (entropy) than non-MEC passages. This corroborates theoretical proposals dating back more than 150 years

suggesting a prominent role of expectation in aesthetic experience of music [38–40, 42] and more recently in MECs specifically [14, 33, 34, 36, 37].

The permutation tests for individual acoustic elicitors replicated many of the findings from prior research. These results are not exploratory since we only included acoustic elicitors hypothesised based on previous research. In addition, we applied extremely strict Bonferroni correction, reducing the alpha threshold for statistical significance to slightly over 0.001, therefore providing a high degree of confidence in the identified effects. Tying these results back to findings from prior research [1], the permutation test analysis provided a large-scale replication of effects showing MECs as being associated with all the acoustic and musical elicitors approximated with extracted features, including increases in loudness [e.g., 14], crescendi [e.g., 30], increased roughness [e.g., 18], brightness [e.g., 16], event density [e.g., 21], spectral centroid and flux [e.g., 16], expansion of the frequency range [e.g., 19] as well as changes in texture, harmony and tonality [e.g., 14]. The results for harmonic change and key distance were less convincing than for other acoustic features, consistent with previous findings [29], perhaps due to their low-level nature. The results provide novel quantitative evidence that MECs are more likely in the presence of vocals. Other research has found no effect of the content of lyrics on MEC frequency [29], so it seems likely that the effect observed here is related to the acoustic or musical characteristics of the vocal line.

The fact that the SVM classifier performed above chance provides convincing evidence that stimulus-based cues contribute to eliciting MECs. However, while recall was high, precision was very low for all models. Precision refers to the proportion of classified MECs actually annotated as MECs in the dataset, while recall refers to the proportion of MECs in the dataset classified as MECs. In other words, the trained model identified a large proportion of the MECs in the dataset (high recall) but also predicted too many MECs at points where none were recorded in the dataset (low precision), although all the models still performed better than chance as shown by all AUC values exceeding 0.5. It is because of this low precision that we opted to report results based on the highest F_β value returned by each model, since not doing so disproportionately penalised recall by optimising for very small increases in precision. Optimising for recall also made sense conceptually due to the nature of the data. While we can be sure that each MEC recorded represents an experience of chills for at least one person, we cannot infer from the absence of an MEC report that chills would never be experienced by anyone at that point in a particular piece of music. If we were to massively increase the sample size, it seems likely that additional onsets of MECs would have been reported for some of the passages used as controls for model training in the present analysis. This makes the ground truth incomplete, which motivated attempting to predict all the MECs reported in the data by maximising recall (rather than predicting only the MECs present in the data by maximising precision), with the added advantage that an F_β-optimised model should provide more useful information about which features best characterised MEC onsets when looking at feature importance.

The expectation-based measures of melodic entropy and information content were the best predictors of MEC onsets by far, followed by the variance in their rates of change as well as the mean of some spectral features (flatness, spread and centroid). The strong performance of expectation-based measures comes despite them being computed from melodies automatically extracted from the audio tracks—as manual transcription was not possible given the size of the dataset. Automatic melody transcription from audio is a notoriously challenging problem in music informatics [51] and the method used produced much less accurate input data than for the acoustic measures. Therefore, the effects of expectation found here are likely to represent a significant underestimate of the true effects. Furthermore, the inaccuracy of melody extraction precluded examining expectation-based measures for rhythm, an interesting avenue to follow in future research. Finally, future work might also benefit from extracting beat-aligned acoustic features as opposed to frame-based representation in the time domain; this could help integrate low-level acoustic features with high-level expectation-based measures.

Feature importance provides interesting insights when combined with the results of the permutation tests, by allowing comparisons between the magnitude of the difference between MECs and controls in each feature, and the degree to which these differences were predictive of MECs. For instance, the lack of detected effects of first-order differences on MECs in the permutation testing is also apparent in the fact that most of these features were not strong predictors of MECs in the classification analysis. However, in some cases, these comparisons are less easy to interpret. Many features (e.g., sharp increases in loudness) benefit from extensive prior empirical support and strongly displayed the expected behaviour in the permutation tests, but they were not highly ranked in terms of feature importance in the SVM analysis. Conversely, the magnitude of the effects detected in the permutation tests for melodic entropy and information content was much lower than for other features, and yet, these expectation-related features were the best predictors of MECs. This suggests that, while previous correlational evidence for an association between some acoustic features and MECs was replicated, these features were not always strong predictors of MECs. Taking loudness as an example, it could be that MECs are indeed characterised by sharp increases in loudness, but these also occur at other moments, which makes them inaccurate predictors of MECs. In other words, loudness might be a necessary but non-specific elicitor of MECs. Conversely, small differences in expectation between MECs and controls could be due to small

effect sizes or to increases in entropy and information content only being present in a subset of MEC onsets. Nonetheless, these expectation-related elicitors appear to be highly specific to MECs, therefore driving their feature importance in predicting chills onsets.

3.2 Limitations

This study is not without limitations, foremost of which is the generally poor classification performance of the SVM. However, it is worth placing model performance within the context of the task at hand. Automatic detection of MEC onsets is far from trivial, due to the inherent impossibility of collecting exhaustive data and MECs being a subjective, psychophysiological response known to be caused by a wide range of elicitors, some of which would be exceedingly difficult to quantify for modelling purposes. For example, we expect MECs to be influenced by the personality and current affective state of the listener [52] and their context [25, 53, 54]. We also expect MECs to be influenced by emotional elicitors [1] for which expectation is just one of several mechanisms by which music can induce an emotion [34].

Just as importantly, we wanted our models to be hypothesis-driven and interpretable, which meant that we needed to use both interpretable features and interpretable models at the necessary cost of predictive performance. It is highly likely that more powerful approaches (e.g., neural networks, ensemble machine learning methods) or features such as mel-frequency cepstral coefficients [29] would result in better predictive performance. However, such methods would have significantly compromised the interpretable hypothesis-driven nature of the research. Future work will gain from methodological insights gathered in the present work (see Methods and Supplementary Information), such as the use of features computed over large sliding windows and segmented using a 500-ms frame size, or the use of segment-based metrics for evaluation. Exploring the benefits of using larger frame sizes might also lead to improved performance. It may turn out that there is a relatively low ceiling to predictive performance based on acoustic and expectation-based musical elicitors. If so, this would suggest a more important role for emotional elicitors than previously anticipated. Considering the extent of previous findings about the relationships between MECs and various extra-musical factors such as personal meaning or the state of being moved [1], it is possible that only a small proportion of MECs are exclusively caused by acoustic and musical elicitors via the psychological mechanisms of brain stem reflex and musical expectation [34]. Future work should attempt to add emotional elicitors to the analysis of acoustic and musical elicitors as reported here.

Another limitation arises from the reliance of the present study on audio feature extraction, notably in terms of perceptual validity. It is very common in psychology experiments to extract audio features using existing music information retrieval toolboxes—as a rough example, MIRtoolbox [55] has been cited in more than 600 articles including the word *psychology*. However, it is worth noting that many of these features were designed and evaluated for practical tasks, and, therefore, might not necessarily equate to human perceptual representations of sound [56].

3.3 Conclusion

In summary, the present work yields evidence that both low-level acoustic features and high-level expectations influence the occurrence of music-evoked chills. In contrast to previous research, a large sample of participants and pieces of music were collected with precise time stamps for MECs, and machine learning methods were used to investigate whether computational acoustic features and expectation-based measures were predictive of MEC onsets. The results both replicate previous findings for acoustic elicitors of MECs and provide new evidence for a relevant role of high-level musical expectations, unveiling greater surprise and uncertainty for MECs compared to other musical passages and highlighting the prominence of expectation features over acoustic features in classification performance. Expectations are more likely than acoustic features to generalise across domains since prediction plays an important role in other domains such as language comprehension [10, 11] and visual perception [12, 13] and is also fundamental to general mechanisms of neural processing during perception and action [9]. Future research should seek to explore the many remaining gaps in knowledge about the relationship between expectation and MECs. Notably, questions remain about the exact interaction between uncertainty and surprise, the role of harmonic and rhythmic expectation, the differences between schematic and veridical expectation, and the interactions between acoustic and musical elicitors, stylistic preference and familiarity [3, 16, 18, 24, 52], and affect [1]. It also remains to be seen whether the present effects of high-level expectation on MECs generalise to pleasurable experiences in other domains.

4 Methods

4.1 Stimulus selection

4.1.1 Dataset

The *Onsets of Chills in Music (oChiM)* dataset accompanying the present study was collected via an online survey to access a broad sample of individuals and music. Survey responders were asked to report basic demographic information such as age, gender and country of residence. They were then asked to provide up to ten pieces of music that often evoked chills (defined as shivers, goosebumps or a tingling sensation experienced in response to music listening). For each piece of music, participants provided the name of the artist or composer, the title of the piece of music, a link to the piece of music (streamed on YouTube, Spotify, SoundCloud, or a similar platform), and between one and five precise timestamps for the instants at which they frequently experience the MEC onset (i.e., the exact moment at which chills tend to begin). Participants were advised not to include music that could be strongly associated with specific personal memories such as life events (e.g., a wedding, a memorable concert) or periods of time (e.g., last summer, secondary school), or music that was taken from a film soundtrack if they had watched the film in question. These restrictions (taken from a previous study by Salimpoor *et al.* [57]) were intended to maximise the likelihood that the MECs were induced by detectable acoustic and musical elicitors, with the potential to generalise across individual participants, as opposed to individual autobiographical elicitors driven by episodic memory, which are extremely unlikely to generalise across individuals.

Ethics approval was granted by the Queen Mary Ethics of Research Committee (QMREC1752). The survey was hosted on Qualtrics (Qualtrics, Provo, UT) and responses were collected between February 2018 and April 2020. The survey was advertised on a wide range of online platforms, including international academic mailing lists, staff and student mailing lists within the Queen Mary University of London, as well as Twitter[1] and Reddit.[2] Responses were obtained from 2,069 participants aged from 18 to 77 years (M = 23.6 years, SD = 8.8 years) and originating from a wide range of geographical areas (62 % North America, 25% Europe, 8% Asia, 3% Oceania, 1% Africa, 1% South America).

The resulting dataset required extensive manual cleaning comprising five main stages. First, we removed entries from participants who abandoned the survey study before providing any pieces of music, resulting in 1,398 entries being retained. Second, we removed entries that kept the default answers provided in the questionnaire. Third, we processed MEC onsets as follows: time ranges were converted to precise onsets, extraneous qualitative comments about specific musical characteristics leading to MECs were removed, and a few entries in which the onset time provided was greater than the total track duration were removed. Fourth, we cleaned the URLs by removing additional qualitative comments and discarding non-valid URLs. Together, stages 2-4 retained 1,187 pieces of music associated with MECs, corresponding to 1,150 unique pieces with 2,028 MEC onsets. The fifth and final stage involved removing pieces that could not be accessed via the supplied URLs, resulting in 1,019 unique pieces corresponding to 1,806 MEC onsets.

4.1.2 Stimulus preparation

To be suitable for auditory feature extraction, tracks were retrieved as WAV files, downmixed to mono and downsampled to 44.1 kHz when necessary, using the *tuneR* R package [58]. RMS normalisation was applied to all tracks simultaneously, using the *soundgen* R package [59]. This process consisted of rescaling individual audio files so that they had the same peak amplitude, set at 0 dB, and rescaling all files linearly so that the RMS amplitude of each file matched that of the file with the lowest RMS amplitude.

4.2 Feature extraction

Based on the acoustic and musical elicitors reviewed in the Introduction, we extracted a range of features that we believed most closely captured notions of loudness, roughness, brightness, spectral centroid and flux, event density, frequency range, crescendi, tonality, harmony, texture, expectation and presence of vocals. Due to the computationally intensive nature of extracting many features from many tracks, the scripts were run in parallel on multiple Linux-based servers at the Queen Mary University of London. The whole feature extraction process is visualised in Figure 4, along with the source of the features and the elicitors they were intended to approximate.

[1] `https://www.twitter.com`
[2] `https://www.reddit.com`

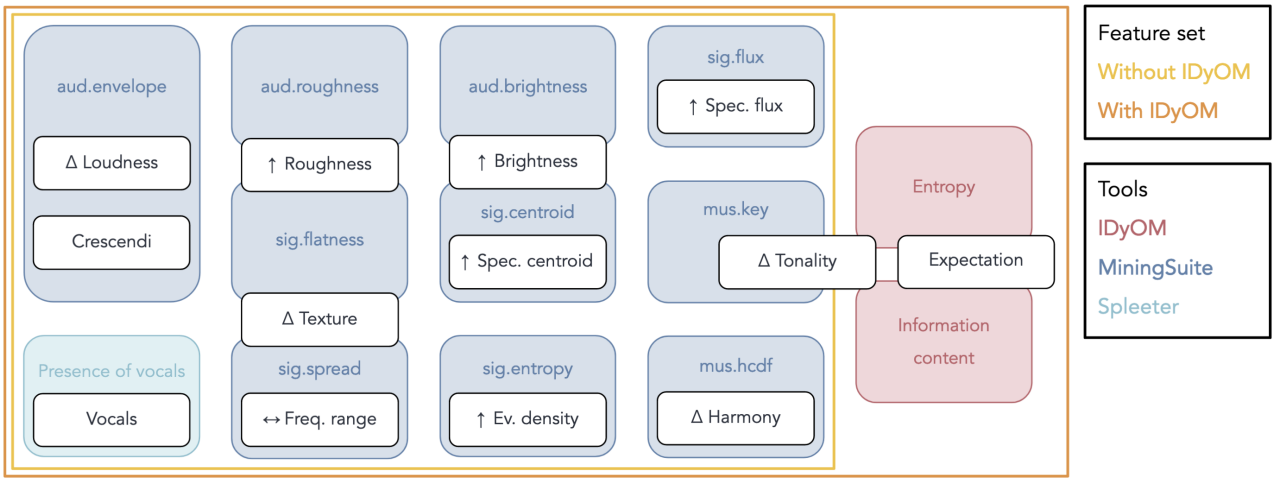

Figure 4: List of features extracted from each track. The features are depicted in boxes with coloured frames indicating the tool used to extract them. The acoustic and musical elicitors of MECs they aim to characterise are shown in white boxes (using the abbreviations *freq.* for "frequency", *spec.* for "spectral", and *ev.* for "event"), along with the hypothesised direction of their respective effects, as identified from prior research and denoted by the following symbols: Δ for changes, ↑ for increases or elevated levels, and ↔ for expansions. The outer boxes represent the feature sets used for model training (see Supplementary Information for details).

4.2.1 MiningSuite

Most features were extracted using *MiningSuite* [60], a MATLAB framework for the analysis of audio and music recordings, which expands on the methods provided by the commonly used *MIRtoolbox* [55]. Default settings for frame size were used as detailed below.

Loudness and crescendi were approximated with the *aud.envelope* function, which consists of a generic envelope extraction method, further processed following a model of human auditory perception [61]. The feature was extracted with a 10-ms frame size, corresponding to a 100-Hz sampling rate.

Spectral frame decomposition was performed using the *sig.spectrum* function, which applies a Fast Fourier Transform to the audio waveforms of each file, using a 50-ms sliding window with a 25-ms overlap. With the same sliding window size and overlap, we approximated roughness with the *aud.roughness* and *sig.flatness* features, respectively estimating sensory dissonance and spectral smoothness; brightness with the *aud.brightness* and *sig.centroid* features, the former capturing the amount of high-frequency energy in the signal and the latter more broadly capturing spectral centroid (the centre of mass of the spectrum—another elicitor of interest); spectral flux with the *sig.flux* feature, calculating the spectral distance between successive frames; event density with the *sig.entropy* feature, computing the relative Shannon entropy of the input [62]; and frequency range with the *sig.spread* feature, capturing spectrum variance.

Tonality was approximated with the *mus.key* feature, which estimates tonal centre positions by choosing the highest key candidate from a key strength curve, itself computed by correlating the chromagram of the signal with known key profiles [63, 64]. This feature used a 1-s sliding window with a 0.5-s overlap. To approximate harmonic change, a six-dimensional tonal centroid was extracted using the *mus.tonalcentroid* function, corresponding to chord projections on the circle of fifths, minor thirds and major thirds, before being processed with the *mus.hcdf* harmonic-change detection function (HCDF), which returns the flux of the tonal centroid, using the default settings of a 743-ms sliding window with a 74.3-ms overlap.

4.2.2 Spleeter

To generate a continuous, binary feature representing the presence of vocals, tracks were first processed using the *Spleeter* source separation Python library [65]. It provides pretrained models to perform source separation of a music track into two, four, or five stems containing separate instruments. For the present study, two-stem separation was conducted, resulting in two separate stems containing vocals and instrumental parts for each track.

We applied an amplitude threshold to the vocals stem. This produced a binary feature reflecting the absence or presence of vocals discarding the small amount of residual noise in the tracks. In practice, amplitude is characterised by a high degree of zero-crossing, which is not suitable for thresholding. Therefore, we computed the loudness of the vocals from their amplitude using the *soundgen* R package, which provides a function allowing

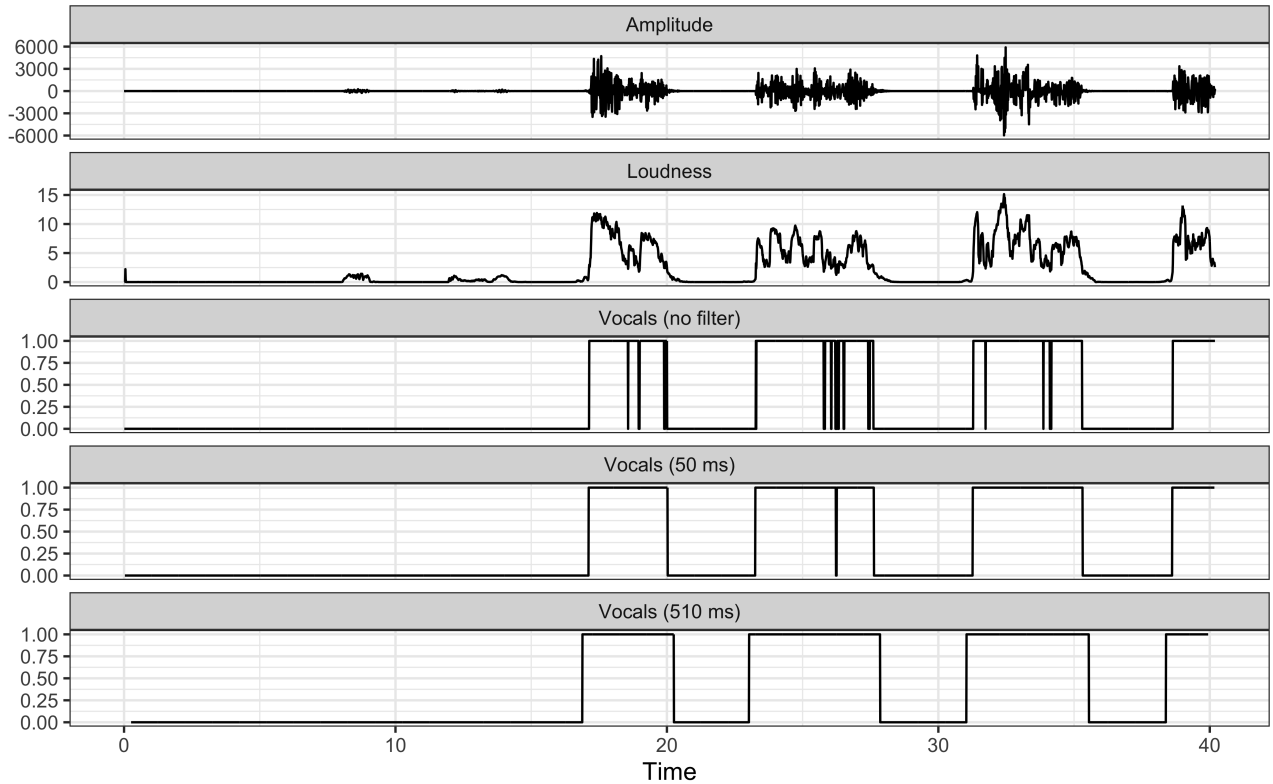

Figure 5: Vocals thresholding process. Data is shown for a short, 40-s excerpt to illustrate the thresholding process used to identify the presence of vocals. From top to bottom row are displayed 1) the amplitude of the waveform of the vocals stem extracted using Spleeter, 2) an estimate of subjective loudness, 3) the results of applying a 2.5-sones loudness threshold with no rolling maximum filter, 4) same as above with a rolling maximum filter with a 50-ms span, and 5) same as above with a 510-ms span. Several combinations of values were tested before the values used in the last plot were retained because they most closely approximated manual annotation in a representative set of examples.

the estimation of subjective loudness in sones (a psychoacoustic unit of perceived loudness) for each 20-ms sliding window with 50% overlap, resulting in a value capturing subjective loudness every 10 ms. To prevent the application of a loudness threshold from returning an overly sensitive vocals-detection feature, we applied a rolling maximum filter with a span of 510 ms for each track before finally applying a loudness threshold categorising vocals as present if above 2.5 sones. This thresholding process is visualised in Figure 5 and resulted in a continuous, binary feature representing the presence of vocals for each 10-ms frame.

Note that both the 510-ms value for the span of the maximum filter and the 2.5-sones value for the loudness threshold were chosen manually after comparing the outputs when altering these parameters. A systematic validation was not possible due to the absence of labelled data in our sample. However, we conducted several manual checks on a representative set of recordings and deemed the feature resulting from these parameters as close as possible to what would have resulted from manual annotation of the tracks. This allowed us not to allocate a disproportionate amount of time to manually extracting a single feature in the wider context of the present analysis. Importantly, while this feature is sensitive to differences in loudness between tracks, such concerns were mitigated by the fact that RMS normalisation was conducted prior to this step in the analysis.

4.2.3 IDyOM

Finally, to compute information about melodic expectation, it was necessary to extract melodies in MIDI format from each audio track. First, the time series of continuous frequency values in Hz was extracted for each melody using the *MELODIA* plugin [66] for *Vamp*,[3] via its associated *vamp* Python library.[4] MELODIA enables the estimation of the fundamental frequency of the pitch of the primary melody in polyphonic tracks, and was therefore particularly well suited for this task.

[3] `https://vamp-plugins.org`
[4] `https://pypi.org/project/vamp`

A simple heuristic provided by the author of MELODIA[5] was reimplemented in R to quantise pitch frequencies into discrete MIDI notes: converting each value in Hz to its closest MIDI note before applying a median filter with a 250-ms span to remove some of the underlying data noise, and finally discarding notes shorter than 100 ms. The resulting MIDI notes were stored in a text file suitable for the next step of the feature extraction process, while note onsets were stored in a reference file separately to convert the extracted features back to time series suitable for model training.

The Information Dynamics of Music (IDyOM) model [49, 50], was used to generate information-theoretic measures of melodic expectation from these sequences of MIDI events. IDyOM generates probabilistic predictions for musical events using variable-order Markov models, which learn from the statistical regularities in symbolic musical sequences such as MIDI representations. Two outputs are relevant here: *entropy*, a measure of uncertainty about which event is predicted to come next given the current context; and *information content*, the negative log probability, a measure of the surprisal of the event. For instance, if at time $t-1$ in a melody, IDyOM is very certain about which note comes next, this will be reflected by low entropy for the note occurring at time t. If the actual next note is indeed very predictable, this will be reflected by low information content, but if it is unpredictable, information content will be high. IDyOM can be configured to learn incrementally from a given piece of music (via its short-term model, capturing dynamic effects on expectation), to generate predictions based on prior learning from a large corpus of music (the long-term model, capturing schematic effects) or both. IDyOM provides a multiple-viewpoint system, which allows sequences to be predicted using a range of different representations of the musical input including, for example, abstract representations of pitch interval and scale degree for predicting the pitch of a note. IDyOM has been the subject of substantial empirical evaluation and has proved capable of simulating expectation, complexity, arousal and pleasure in music perception across a range of experiments [40–42, 50, 67].

In the present study, the *cpint* viewpoint (chromatic pitch interval) was derived from the *cpitch* viewpoint (MIDI pitch number) to assess the entropy and information content of each MIDI event. The model used a combination of the short-term and long-term models, with the long-term model being trained on a set of folk ballads from Nova Scotia, Bach chorale melodies, and German folk songs from the Essen Folk Song Collection, consistent with the geographical origins of most participants in the MEC onset survey [for a description of this corpus, see 49]. The resulting entropy and information content values associated with each MIDI event for each track were linked back to the note onsets, allowing these values to be converted to continuous features synchronised with the tracks.

4.3 Feature preparation

4.3.1 Key distance

Most features required additional processing to capture the hypothesised elicitors laid out in prior research. Tonality, notably, is only thought to affect the occurrence of MECs in occasional cases of changes in tonality. However, the current feature extracted with *mus.key* only captured the tonal centre at a given time, which should have no bearing on the occurrence of MECs. We decided to compute key distance from this feature, following the intuition that more unexpected changes in tonality might be more conducive to experiencing MECs. To do so, we first smoothed the feature by applying majority voting with a 3.5-s span, before assigning a value to tonality changes based on distance in the circle of fifths. For instance, if the tonal centre at time $t-1$ was C, the key distance at time t would be 0 if the tonal centre remained C, 2 if it changed to D, or a maximum of 6 if it changed to F#. The feature was then upsampled to 100 Hz for consistency with the other features, replacing the newly introduced missing values with their nearest existing key distance value.

4.3.2 Interpolation and smoothing

Cubic spline interpolation was applied to all the other auditory features extracted with MiningSuite to match this 100-Hz frame rate, except for envelope, already sampled at 100 Hz. Interpolation was only allowed to fill the gaps between two existing values, to prevent the method from returning wildly unlikely values at the beginning and end of each track. Following this, all features (including envelope) were smoothed using a 50-ms median filter.

4.3.3 First-order difference

As discussed earlier and seen in Figure 4, many MEC elicitors refer to changes, not specific values, in acoustic and musical properties. To allow the analyses in the present study to detect this behaviour, first-order differences

[5] `https://github.com/justinsalamon/audio_to_midi_melodia`

Table 1: List of features included in both sets of analyses

Tool	**Feature category**	**Features**			
		μ_0	σ_0	μ_1	σ_1
MiningSuite	Envelope	✓	✓	✓	✓
	Envelope+	✓	✓	✓	✓
	Roughness	✓	✓	✓	✓
	Flatness	✓	✓	✓	✓
	Brightness	✓	✓	✓	✓
	Spectral centroid	✓	✓	✓	✓
	Spectral flux	✓	✓	✓	✓
	Spectral entropy	✓	✓	✓	✓
	Spectral spread	✓	✓	✓	✓
	Key distance	✓		✓	
	Key distance+	✓	✓	✓	✓
	HCDF	✓	✓	✓	✓
	HCDF+	✓	✓	✓	✓
Spleeter	Vocals	✓	✓	✓	✓
IDyOM	Melodic entropy	✓	✓	✓	✓
	Melodic IC	✓	✓	✓	✓

Note. Four types of features for each feature category. μ_0 = mean of the original values, σ_0 = standard deviation of the original values, μ_1 = mean of the first-order difference, σ_1 = standard deviation of the first-order difference, + = feature computed using a longer, sliding window, IC = information content.

were computed for each feature and included in the analyses such that if a feature had a value of 3 at time $t-1$ and a value of 7 at time t, its first-order difference at time t would be $7-3=4$.

4.3.4 Segmentation

Reducing the quantity of frames being considered for analysis was necessary to speed up computations. Tracks were segmented into larger frames, over which summary statistics were computed to investigate central tendencies and variance for each feature. Segmentation into 200-ms and 500-ms frames was compared with 500-ms performing best (see Supplementary Information) and retained for the main analysis. Mean and standard deviation were computed for each frame, resulting in four dimensions for each original feature: μ_0 and σ_0 (mean and standard deviation of the original values) and μ_1 and σ_1 (mean and standard deviation of the first-order difference). The way key distance was computed prevented us from calculating standard deviations as they would not contain any meaningful information. In the present study, μ_0 can be thought of as the original feature, σ_0 as the variance in that feature, μ_1 as the rate of change in the feature, and σ_1 as the variance in that rate of change.

We predicted that, for envelope, HCDF and key distance, changes on a slower time scale might better capture their hypothesised role as elicitors of MECs. Summary statistics for these features were therefore also computed using a 2-s sliding window with 50% overlap, and upsampled to the target frame size (200-ms or 500-ms) for inclusion in the analyses. The full list of features is shown in Table 1, and the computed features are available in the oChiM dataset accompanying the present study.

4.4 Analysis

4.4.1 Permutation tests

To identify patterns in the behaviour of each feature around the onset of MECs, we ran a series of permutation tests [for a similar approach, see 68]. Permutation tests provide a convenient non-parametric way of constructing a test statistic, comparing the observed data to a null distribution generated by permuting the samples. Here, we examined the impact of each feature on MECs by evaluating how unlikely their values were around the onset of MECs, when compared to other moments within the tracks that were not reported as producing MECs.

To achieve this, we extracted all 20-s excerpts centred around the MEC onsets from each track. We only retained complete excerpts, meaning that MEC onsets within the first or last 10 s of each track were discarded. Rather than choosing individual control excerpts for each excerpt causing MECs, we split every track into sequential 20-s excerpts, discarded every excerpt incomplete or overlapped with any excerpt causing MECs, and retained all remaining excerpts as controls. This process resulted in two unbalanced sets of 20-s excerpts, capturing 10 s before and after each MEC onset for the MEC set and most other moments within the tracks for the control set.

The test statistic was computed for each frame of each feature, consisting of the difference between the average values for excerpts causing MECs and control excerpts. Two-tailed permutation tests were run by randomly permuting the excerpts while keeping the same number of excerpts in each set, using Monte Carlo estimation with 5,000 replications. We only ran the permutation tests on the data with a 500-ms frame size to limit the number of comparisons, and we used Bonferroni correction within each feature to further mitigate the fact that, even with a 500-ms frame size, 41 significance tests would be required for each feature.

4.4.2 Principal component analysis

We trained models to assess whether or not MEC onsets could be predicted using audio-derived acoustic and expectation-based features and, if so, which features were most important in driving such predictions. We expected a high degree of collinearity between the features, with some being exactly identical except for the sliding window over which they were computed. To address this, we ran a principal component analysis (PCA), projecting each data point into a new space, the dimensions of which are called principal components. PCA enables dimensionality reduction when retaining the smallest number of components that account for as much variance in the data as possible.

The results of the PCA are discussed here for simplicity, as they are not crucial to the interpretation of the results reported. We retained 15 principal components, accounting together for 78% of the cumulative proportion of variance explained for the PCA. We did not attempt to interpret how the features were grouped together into specific principal components, due to the difficulty of assigning meaning to such complex combinations of weights without introducing unnecessary subjectivity. However, we stored feature loadings on each principal component and proportion of variance explained by each principal component for later analyses of feature importance.

4.4.3 Support vector machine

Classification analyses were run using the support vector machine (SVM). For classification tasks using an SVM, model training consists of finding the hyperplane that maximises the distance between the two classes and is used as a decision boundary to make predictions. SVMs are widely used for classification tasks and were chosen in the present analysis because they are relatively easy to interpret and have the added benefit of being more forgiving in terms of statistical assumptions than logistic regression——another commonly used classification method. The main drawback is that the SVM does not consider the sequential nature of the data. Therefore, the SVM was compared to a GMM-HMM classifier based on a hidden Markov model (HMM) predicting continuous time series modelled as Gaussian mixtures (a Gaussian mixture model or GMM), which is sensitive to temporal patterns in data; the results showed that the SVM outperforms the GMM-HMM (see Supplementary Information), so we restrict ourselves to the SVM in the main analyses.

SVMs were trained using the *scikit-learn* Python library [69] with a linear kernel, no random feature selection, stopping criterion set at 10^{-5} and by adjusting weights in inverse proportion to class frequencies to account for class imbalance. To further mitigate class imbalance and the imprecision in the time resolution of reported MEC onsets, we assigned positive labels to all frames within 1 s around each MEC onset, corresponding to the resolution with which participants were asked to report MECs in the survey. For instance, with a 500-ms frame size, instead of considering a single frame for an MEC onset at $t = 40$ s, we considered frames at $t = 39$, 39.5, 40, 40.5, and 41 s as positive labels for the classification task. SVMs were trained to distinguish positive (MEC onsets) from negative (control) labels and the regularisation parameter for linear SVMs (using a logarithmic scale ranging from 10^{-24} to 10^{4}, with even-numbered exponents only) was optimised using manual grid search. We used five-fold cross-validation, using for each fold 60% of the tracks as a training set, 20% as a validation set to pick the best-performing model for testing, and 20% as a testing set to return final performance metrics for the learning parameters which performed best across all five folds.

To evaluate classification performance, we computed precision, recall and the F-measure — the harmonic mean of precision and recall — as well as balanced accuracy, true and false positive rates, and the area under the ROC curve (AUC). In a typical classification task, the AUC is calculated by modifying the classification threshold. However, linear SVMs do not return the probabilities associated with the predictions for each frame by default. To obtain these probabilities, we applied Platt scaling [70] on the decision function and computed the AUC with *scikit-learn* by adjusting the classification threshold over the percentiles of the returned probabilities. After observing the results, we also decided to compute F_β, which applies an additional weight β to the F-measure in order to disproportionately favour precision or recall over the other. We picked a value of 2 for β—a standard value to signify that we considered recall twice as important as precision in this classification task to avoid optimising for very small differences in precision.

The results reported in the Supplementary Information show that the best performing model was an SVM with 500-ms frame size and high-level expectation-based features included.

4.4.4 Feature importance

The final step of the analysis was to compute feature importance. To simplify the interpretation of results, feature importance was only extracted from the best-performing model across all combinations of frame size, feature set and model type (see Supplementary Information). Since the best performing model ended up being an SVM, the method described below is specific to feature importance extraction from the parameters of a trained SVM.

Using Platt scaling involves a second layer of cross-validation within each cross-validation fold. We refer to the folds from this second layer as *SVM folds* to differentiate them from the cross-validation folds used in the workflow detailed above. First, feature coefficients were extracted from the parameters of the trained models. Because we were interested in assessing feature importance, as opposed to trying to interpret the directionality of the coefficients (which was more relevant to the permutation test analyses), absolute values were taken, therefore preventing coefficients from cancelling each other out when averaged over SVM folds, if directionally different. These absolute values for the coefficients of each feature were averaged over each SVM fold, resulting in a set of five positive averaged coefficients for each feature——i.e., one averaged coefficient per cross-validation fold.

However, in this case, the features were principal components obtained with PCA. To tie the coefficients back to the original features, we applied weights based on the proportion of variance explained and feature loadings on each principal component, thereby assigning a share of the influence of each principal component on the models to its constituent features based on how much variance was accounted for by the principal component. To do so, we simply took the feature loadings on each principal component, multiplied by the proportion of the variance explained by that principal component and multiplied by the coefficient extracted from the models for that principal component. We then summed the coefficients for each feature across all principal components, resulting, again, in a single coefficient per feature for each of the cross-validation folds (but this time, for each of the original features, as opposed to principal components).

The magnitude of the coefficients differed between cross-validation folds, but as the folds were of equal size and we were interested in feature importance, we rescaled the coefficients linearly for each fold such that 0 corresponded to the least contributing feature for that fold and 1 to the most contributing feature. Finally, we averaged these values across the five cross-validation folds, resulting in a single feature importance value between 0 and 1 for each original feature used for PCA before model training. These values do not exactly correspond to a rank but rather to a continuous spectrum of feature importance, where 0 would correspond to the feature with the lowest feature importance on all five cross-validation folds, and 1 to that with the highest feature importance.

Data availability

The *Onsets of Chills in Music (oChiM)* dataset that supports the findings of this study is available on OSF using doi:10.17605/osf.io/x59fm.

Acknowledgements

This research was supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology [EP/L01632X/1].

References

1. de Fleurian, R. & Pearce, M. T. Chills in music: A systematic review. *Psychological Bulletin* **147,** 890–920 (2021).
2. Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A. & Zatorre, R. J. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14,** 257–262 (2011).
3. Blood, A. J. & Zatorre, R. J. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 11818–11823 (2001).
4. Klepzig, K. *et al.* Brain imaging of chill reactions to pleasant and unpleasant sounds. *Behavioural Brain Research* **380,** 112417 (2020).
5. Satoh, M. *et al.* A case of musical anhedonia due to right putaminal hemorrhage: A disconnection syndrome between the auditory cortex and insula. *Neurocase* **22,** 518–525 (2016).

6. Griffiths, T. D., Warren, J. D., Dean, J. L. & Howard, D. “When the feeling’s gone": A selective loss of musical emotion. *Journal of Neurology, Neurosurgery, and Psychiatry* **72,** 344–345 (2004).

7. Ferreri, L. *et al.* Dopamine modulates the reward experiences elicited by music. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 3793–3798 (2019).

8. Laeng, B. *et al.* ‘Defrosting’ music chills with naltrexone: The role of endogenous opioids for the intensity of musical pleasure. *Consciousness and Cognition* **90,** 103105 (2021).

9. Friston, K. J. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11,** 127–138 (2010).

10. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* **31,** 32–59 (2016).

11. Pickering, M. J. & Clark, A. Getting ahead: forward models and their place in cognitive architecture. *Trends in Cognitive Sciences* **18,** 451–456 (2014).

12. Summerfield, C. & Egner, T. Expectation (and attention) in visual cognition. *Trends in cognitive sciences* **13,** 403–9 (2009).

13. Trapp, S. & Bar, M. Prediction, context, and competition in visual recognition. *Annals of the New York Academy of Sciences* **1339,** 190–198 (2015).

14. Sloboda, J. A. Music structure and emotional response: Some empirical findings. *Psychology of Music* **19,** 110–120 (1991).

15. Auricchio, N. *Natural highs: Timbre and chills in electronic dance music* in *Popular music studies today: Proceedings of the International Association for the Study of Popular Music 2017* (ed Merrill, J.) (Springer, Kassel, Germany, 2017), 11–23.

16. Bannister, S. & Eerola, T. Suppressing the chills: Effects of musical manipulation on the chills response. *Frontiers in Psychology* **9,** 2046 (2018).

17. Beier, E. J., Janata, P., Hulbert, J. C. & Ferreira, F. Do you chill when I chill? A cross-cultural study of strong emotional responses to music. *Psychology of Aesthetics, Creativity, and the Arts* **16,** 74–96 (2022).

18. Grewe, O., Nagel, F., Kopiez, R. & Altenmüller, E. Listening to music as a re-creative process: Physiological, psychological, and psychoacoustical correlates of chills and strong emotions. *Music Perception* **24,** 297–314 (2007).

19. Guhn, M., Hamm, A. & Zentner, M. Physiological and musico-acoustic correlates of the chill response. *Music Perception* **24,** 473–483 (2007).

20. Honda, S. *et al.* Proximal binaural sound can induce subjective frisson. *Frontiers in Psychology* **11,** 316 (2020).

21. Nagel, F., Kopiez, R., Grewe, O. & Altenmüller, E. Psychoacoustical correlates of musically induced chills. *Musicae Scientiae* **12,** 101–113 (2008).

22. Polo, M. J. *Chill responses to post-tonal music: Musical structures and physiological reactions* PhD thesis (University of Florida, 2017).

23. Park, K. S., Hass, C. J., Fawver, B., Lee, H. & Janelle, C. M. Emotional states influence forward gait during music listening based on familiarity with music selections. *Human Movement Science* **66,** 53–62 (2019).

24. Bannister, S. A vigilance explanation of musical chills? Effects of loudness and brightness manipulations. *Music & Science* **3,** 1–17 (2020).

25. Bannister, S. A survey into the experience of musically induced chills: Emotions, situations and music. *Psychology of Music* **48,** 297–314 (2020).

26. Mlejnek, R. *Physically experienced reactions and music: A questionnaire study of musicians and non-musicians* in *Proceedings of the 3rd International Conference on Music and Emotion, ICME3* (eds Luck, G. & Brabant, O.) (University of Jyväskylä, Department of Music, Jyväskylä, Finland, 2013).

27. Schurtz, D. R. *et al.* Exploring the social aspects of goose bumps and their role in awe and envy. *Motivation and Emotion* **36,** 205–217 (2012).

28. Solberg, R. T. & Dibben, N. Peak experiences with electronic dance music: Subjective experiences, physiological responses, and musical characteristics of the break routine. *Music Perception* **36,** 371–389 (2019).

29. Mori, K. Decoding peak emotional responses to music from computational and lyrical features. *Cognition* **222,** 105010 (2022).

30. Panksepp, J. The emotional sources of “chills" induced by music. *Music Perception* **13,** 171–207 (1995).

31. Harrison, L. & Loui, P. Thrills, chills, frissons, and skin orgasms: Toward an integrative model of transcendent psychophysiological experiences in music. *Frontiers in Psychology* **5,** 790 (2014).

32. Huron, D. *Sweet anticipation: Music and the psychology of expectation* (MIT Press, Cambridge, MA, 2006).

33. Huron, D. & Margulis, E. H. in *Handbook of music and emotion: Theory, research, applications* (eds Juslin, P. N. & Sloboda, J. A.) 575–604 (Oxford University Press, Oxford, UK, 2010).

34. Juslin, P. N. & Västfjäll, D. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences* **31,** 559–575 (2008).

35. Mencke, I., Omigie, D., Wald-Fuhrmann, M. & Brattico, E. Atonal music: Can uncertainty lead to pleasure? *Frontiers in Neuroscience* **12,** 979 (2019).

36. Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A. & McIntosh, A. R. Predictions and the brain: How musical sounds become rewarding. *Trends in Cognitive Sciences* **19,** 86–91 (2015).

37. Zatorre, R. J. *From perception to pleasure: The neuroscience of music and why we love it* (Oxford University Press, 2024).

38. Hanslick, E. *Vom Musikalisch-Schönen: Ein Beitrag zur Revision der Ästhetik der Tonkunst* (Rudolph Weigel, Leipzig, Germany, 1854).

39. Meyer, L. B. *Emotion and meaning in music* (Chicago University Press, Chicago, IL, 1956).

40. Cheung, V. K. M. *et al.* Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology* **29,** 4084–4092 (2019).

41. Egermann, H., Pearce, M. T., Wiggins, G. A. & McAdams, S. Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, and Behavioral Neuroscience* **13,** 533–553 (2013).

42. Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A. & Zatorre, R. J. Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience* **39,** 9397–9409 (2019).

43. Steinbeis, N., Koelsch, S. & Sloboda, J. A. The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience* **18,** 1380–1393 (2006).

44. Grewe, O., Katzur, B., Kopiez, R. & Altenmüller, E. Chills in different sensory domains: Frisson elicited by acoustical, visual, tactile and gustatory stimuli. *Psychology of Music* **39,** 220–239 (2011).

45. Mori, K. & Iwanaga, M. Two types of peak emotional responses to music: The psychophysiology of chills and tears. *Scientific Reports* **7,** 46063 (2017).

46. Grewe, O., Kopiez, R. & Altenmüller, E. The chill parameter: Goose bumps and shivers as promising measures in emotion research. *Music Perception* **27,** 61–74 (2009).

47. Goldstein, A. Thrills in response to music and other stimuli. *Physiological Psychology* **8,** 126–129 (1980).

48. de Fleurian, R. & Pearce, M. T. The relationship between valence and chills in music: A corpus analysis. *i-Perception* **12,** 1–11 (2021).

49. Pearce, M. T. *The construction and evaluation of statistical models of melodic structure in music perception and composition* PhD thesis (City University of London, 2005).

50. Pearce, M. T. Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences* **1423,** 378–395 (2018).

51. Salamon, J., Gómez, E., Ellis, D. P. & Richard, G. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine* **31,** 118–134 (2014).

52. Nusbaum, E. C. & Silvia, P. J. Shivers and timbres: Personality and the experience of chills from music. *Social Psychological and Personality Science* **2,** 199–204 (2011).

53. Egermann, H. *et al.* Does music listening in a social context alter experience? A physiological and psychological perspective on emotion. *Musicae Scientiae* **15,** 307–323 (2011).

54. Nusbaum, E. C. *et al.* Listening between the notes: Aesthetic chills in everyday music listening. *Psychology of Aesthetics, Creativity, and the Arts* **8,** 104–109 (2014).

55. Lartillot, O., Toiviainen, P. & Eerola, T. *A Matlab toolbox for music information retrieval* in *Data analysis, machine learning and applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation* (eds Preisach, C., Burkhardt, H., Schmidt-Thieme, L. & Decker, R.) (Springer, Heidelberg, Germany, 2008), 261–268.

56. Aucouturier, J.-J. & Bigand, E. Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems* **41,** 483–497 (2013).

57. Salimpoor, V. N., Benovoy, M., Longo, G., Cooperstock, J. R. & Zatorre, R. J. The rewarding aspects of music listening are related to degree of emotional arousal. *PLOS ONE* **4,** e7487 (2009).

58. Ligges, U., Krey, S., Mersmann, O. & Schnackenberg, S. *tuneR: Analysis of music and speech (R package version 1.3.3.)* `https://CRAN.R-project.org/package=tuneR`.

59. Anikin, A. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods* **51,** 778–792 (2019).

60. Lartillot, O. *MiningSuite: A comprehensive Matlab framework for signal, audio, and music analysis, articulating audio and symbolic approaches* in *SMC 2019 Proceedings of the 16th Sound & Music Computing Conference* (eds Barbancho, I., Tardón, L. J., Peinado, A. & Barbancho, A. M.) (Universidad de Málaga, Málaga, Spain, 2019), 489.

61. Klapuri, A. P., Eronen, A. J. & Astola, J. T. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing* **14,** 342–355 (2006).

62. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 379–423 (1948).

63. Gómez, E. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing* **18,** 294–304 (2006).

64. Krumhansl, C. L. Tonal hierarchies and rare intervals in music cognition. *Music Perception* **7,** 309–324 (1990).

65. Hennequin, R., Khlif, A., Voituret, F. & Moussallam, M. Spleeter: A fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* **5,** 2154 (2020).

66. Salamon, J. & Gómez, E. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing* **20,** 1759–1770 (2012).

67. Clemente, A. *et al.* A set of 200 musical stimuli varying in balance, contour, symmetry, and complexity: Behavioral and computational assessments. *Behavior Research Methods* **52,** 1491–1509 (2020).

68. Grewe, O., Kopiez, R. & Altenmüller, E. Chills as an indicator of individual emotional peaks. *Annals of the New York Academy of Sciences* **1169,** 351–354 (2009).

69. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

70. Platt, J. C. in *Advances in large-margin classifiers* (eds Smola, A. J., Bartlett, P., Schölkopf, B. & Schuurmans, D.) 61–74 (MIT Press, Cambridge, MA, 1999).

Expectation elicits music-evoked chills

Rémi de Fleurian[1,*], Ana Clemente[1,2], Emmanouil Benetos[1], and Marcus T. Pearce[1,3,*]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany
[3]Department of Clinical Medicine, Aarhus University, Denmark
[*]Corresponding authors: remi.defleurian@gmail.com, marcus.pearce@qmul.ac.uk

Supplementary information

Introduction

For the classification, all analyses were run with two segmentation parameters: once for segmentation into 200-ms frames and once for 500-ms frames, because there was no information to determine a priori which frame size would work best to investigate the occurrence of MECs. In addition to these two segmentation parameters, two feature sets were evaluated. The first set did not include the IDyOM features simulating high-level expectation while the second set did. This was done to evaluate the effect of accounting for expectation in the predictive performance of the models. For clarity, we hereafter refer to these differences in model training as differences in *frame size* (200-ms or 500-ms), *feature set* (without or with IDyOM features) and *model type* (described below).

Therefore, four PCAs were conducted with centring and scaling—one for each combination of frame size (200 ms or 500 ms) and feature set (with or without IDyOM). We retained 12 principal components with eigenvalues above one for each PCA using the first feature set (without IDyOM) and 15 principal components for each PCA using the second feature set (with IDyOM), both regardless of frame size. Finally, we compared two kinds of classifier: one based on a hidden Markov model and the other based on a support vector machine, which are covered below.

Hidden Markov models

As opposed to Markov chains, which model the probabilities of sequences of observable states, hidden Markov models (HMMs) model the probabilities of hidden states, which themselves drive the probability distributions of observable events. HMMs are specified by a set of hidden states, the transition probabilities between these states, an initial probability distribution for these states, a set of observations and the observation likelihoods associated with each state, also called emission probabilities—the probability that an observation was generated by a specific state [for excellent introductions to HMMs, see 1, 2].

In its most simple form, for a univariate sequence of observations, an HMM associates each hidden state with a specific emission distribution (e.g., normal distribution with given mean and standard deviation) of the observed values in that sequence. In practice, we are often faced with multivariate sequences of observations, and instead of using multivariate Gaussian emissions, which might be limited in how accurately they can represent observations, Gaussian mixture model emissions (GMMs) are preferred. GMM-HMMs are particularly well suited to modelling auditory events due to their flexibility and sequential nature, and have successfully been applied to speech recognition [see 3] and to segmentation, genre classification, sequence prediction and event detection in music [e.g., 4, 5].

In the present case, the observations correspond to the multidimensional sequence of principal components, of which different configurations were modelled by GMMs, each governed by a different hidden state of the HMM. It was assumed that specific sequences of hidden states gave rise to the occurrence of MECs—as opposed to a single hidden state representing MECs specifically.

HMMs are characterised by three problems [1, 2]: the likelihood problem (determining the likelihood of a specific sequence of observations given the HMM and observations), the decoding problem (discovering the best hidden state sequence given the HMM and observations—not relevant to the present study because we did not attempt to interpret the hidden states themselves), and the learning problem (learning the HMM parameters given the states and observations). The present automatic MEC onset detection task consisted of two steps. First, we trained two separate HMMs (learning problem), one trained on excerpts centred around the MEC onsets, and the other on control excerpts. Second, we presented these trained HMMs with new excerpts to obtain the likelihood of each excerpt according to each HMM (likelihood problem). If the HMM trained on

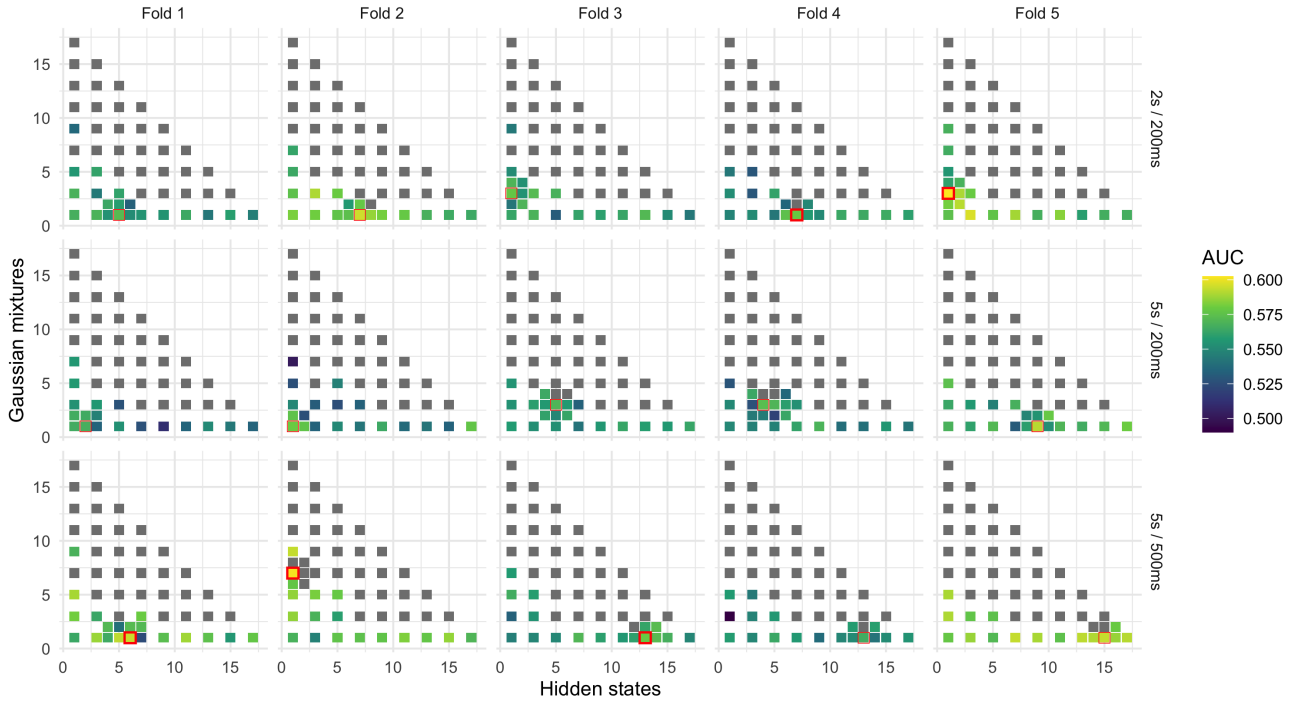

Figure 1: Grid search for GMM-HMM training. In this example, GMM-HMMs were trained on the feature set including IDyOM features. Two GMM-HMMs (MECs and control) were trained for each combination of the number of hidden states (bottom axis), Gaussian mixtures (left axis), cross-validation fold (top axis) and excerpt size and frame size (right axis). Model performance was evaluated as the area under the receiver-operating characteristic curve (AUC). Odd numbers of states and mixtures were first used for training. For each combination of fold, excerpt size and frame size, additional models were trained using the even numbers of states and mixtures closest to the best-performing models (highlighted in red), sometimes resulting in enhanced performance (e.g., Fold 1, 5-s excerpt size, 500-ms frame size). Grey cells signal models which failed to converge. The best-performing GMM-HMMs for each fold are highlighted with a thicker, red border.

excerpts causing MECs returned the highest likelihood, the excerpt was categorised as inducing MECs and vice versa. Implementation details are provided below.

As with the permutation tests, this approach required extracting two sets of excerpts, because HMMs are best trained on short sequences of equal length. The exact same procedure was used, selecting excerpts centred around the MEC onsets, and control excerpts sequentially through the rest of each track. However, instead of 20-s excerpts, we tested two excerpt durations: 2 s (with a 200-ms frame size only), and 5 s (with 200-s or 500-s frame sizes), to speed up computations while ensuring a reasonable amount of frames for each excerpt. *Excerpt size* (2 s or 5 s) thus corresponds to the last aspect of model training we controlled, in addition to frame size, feature set and model type.

Survey participants reported MEC onsets with a 1-s resolution. We consequently decided to augment the training data by also extracting excerpts categorised as causing MECs for each frame within 1 s of the original MEC onset. For instance, with a 500-ms frame size, instead of getting a single excerpt for a MEC onset at $t =$ 40 s, we extracted five excerpts at $t =$ 39, 39.5, 40, 40.5, and 41 s. This also allowed us to reduce the class imbalance, with control excerpts outnumbering excerpts causing MECs by a ratio of 100:1.

Using the *pomegranate* Python library [6], two GMM-HMMs (MECs and control) were trained for each combination of feature set (with or without IDyOM), frame size (200 ms or 500 ms), and excerpt size (2 s or 5 s), using manual grid search by iterating over the number of hidden states and the number of Gaussian mixtures with which the models should be trained. More specifically, to save time on this computationally intensive training process, we trained each model using odd numbers of states and mixtures (ranging from 1 to 17) before testing the even numbers of states and mixtures nearest to the best-performing model. This process is illustrated in Figure 1. We used five-fold cross-validation, using for each fold 60% of the tracks as a training set, 20% as a validation set to pick the best-performing model for testing, and 20% as a testing set to return final performance metrics for the combination of learning parameters that performed best across all five folds. As with feature extraction, model training was run in parallel on university-provided servers.

The process for training an individual GMM-HMM was as follows. All data were concatenated ignoring their sequential nature to identify a cluster for each hidden state using k-means and initialise the parameters of the

Model	IDyOM	Frame size	Excerpt size	Evaluation metrics					
				AUC	F_β	F	P	R	BA
GMM-HMM	✗	200 ms	2 s	0.568	0.156	0.072	0.038	0.692	0.555
			5 s	0.579	0.161	0.075	0.040	0.682	0.564
		500 ms	5 s	0.560	0.159	0.073	0.038	**0.800**	0.561
	✓	200 ms	2 s	0.562	0.157	0.071	0.037	0.788	0.556
			5 s	0.564	0.156	0.071	0.037	0.782	0.552
		500 ms	5 s	0.574	0.164	0.076	0.040	0.755	0.572
SVM	✗	200 ms	-	0.591	0.166	0.077	0.041	0.716	0.578
		500 ms	-	0.592	0.166	0.078	0.041	0.713	0.579
	✓	200 ms	-	0.594	0.166	0.077	0.041	0.724	0.579
		500 ms	-	**0.597**	**0.167**	**0.078**	**0.041**	0.693	**0.580**

Table 1: Evaluation metrics for the SVM and GMM-HMM classifiers. Best performance for each metric (column) is shown in bold. AUC = Area under the receiver-operating characteristic curve, F_β = weighted F-measure, F = F-measure, P = Precision, R = Recall, BA = Balanced accuracy.

corresponding GMM using such a cluster. The model was then initialised with a uniform probability transition matrix before training began using the Baum-Welch algorithm [see 1]. Regularisation was applied by setting a transition and emission pseudocount of 0.1 and an edge and distribution inertia of 0.1 [see 6]. If the model failed to converge, generally due to underflow errors, training was attempted once more. If unsuccessful, training was abandoned and a new model was trained for the next step of the grid search (as in Figure 1).

Full tracks, split into excerpts centred around each consecutive frame (as opposed to excerpts split sequentially for model training), were used for model validation. For each excerpt from each track in the validation set, a log probability value was obtained for each pair of trained GMM-HMMs (one trained on excerpts causing MECs and the other trained on control excerpts), using the forward algorithm [see 1]. If the GMM-HMM trained on excerpts causing MECs returned the highest log probability for the tested excerpt, that excerpt was predicted as an occurrence of MEC, and vice versa. For each fold, this validation process resulted in a univariate time series of binary frame-wise predictions for each track. Performance was assessed by comparing ground truth and predictions in a fixed time grid, splitting each track into consecutive segments and marking these segments as active if they include at least one frame representing an MEC onset, and inactive otherwise. We opted for a 5-s segment length, resulting in a given 5-s segment being categorised as a true positive if it was active in both the ground truth and the predictions.

In a typical classification task, the AUC is calculated by modifying the classification threshold. In the present analysis, however, predictions were not based on a classification threshold but resulted from comparing log probabilities between two GMM-HMMs. To emulate the principle of classification threshold, we collected all frame-wise log-probability differences and extracted their percentiles. Each percentile was used as a proxy for a threshold by generating a new set of predictions based on whether or not the difference between the log probabilities of both GMM-HMMs was higher or lower than that percentile. This allowed us to collect 100 pairs of true positive and false positive rates (one for each percentile), which were used to compute the AUC with the *scikit-learn* Python library [7]. For the validation sets, only AUC was used to select the combination of learning parameters that performed best across all five folds. For each fold, the GMM-HMMs trained using these learning parameters were then used on the testing set to compute the AUC. All other performance metrics were computed for the classification threshold (i.e., the log probability difference threshold) that returned the highest F_β-value. These metrics were then averaged across all five folds to return the final model-performance metrics.

Support vector machines

The SVM classifiers are described in the Methods section of the main text. Predictions were made frame-wise and, as for the GMM-HMM, a 5-s segment length was used for evaluation.

Results

The SVM and GMM-HMM were compared with 200-ms and 500-ms frame sizes, with and without expectation-based features (IDyOM). For performance evaluation, we computed precision, recall and the F-measure——the harmonic mean of precision and recall as well as F_β with a value of 2 for β (see Methods). In addition, we computed balanced accuracy as well as true positive and false positive rates to extract the AUC. Table 1 shows a comparison of the SVMs and GMM-HMMs with and without IDyOM, with 200-ms and 500-ms frame sizes, The best-performing classifierwas an SVM using a 500-ms frame size and including IDyOM. This was used for the main analysis.

References

1. Jurafsky, D. & Martin, J. H. in *Speech and language processing* 1–17 (2021).
2. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77,** 257–286 (1989).
3. Juang, B. H. & Rabiner, L. R. Hidden Markov models for speech recognition. *Technometrics* **33,** 251–272 (1991).
4. Ajmera, J., McCowan, I. & Bourlard, H. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication* **40,** 351–363 (2003).
5. Wang, C., Benetos, E., Meng, X. & Chew, E. *HMM-based glissando detection for recordings of Chinese bamboo flute* in *SMC 2019 Proceedings of the 16th Sound & Music Computing Conference* (eds Barbancho, I., Tardón, L. J., Peinado, A. & Barbancho, A. M.) (Universidad de Málaga, Málaga, Spain, 2019), 545–550.
6. Schreiber, J. pomegranate: Fast and flexible probabilistic modeling in Python. *Journal of Machine Learning Research* **18,** 1–6 (2018).
7. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).