A Hierarchical Harmonic Mixing Method

Chapter · November 2018

DOI: 10.1007/978-3-030-01692-0_11

3 authors:

Gilberto Bernardes
University of Porto
90 PUBLICATIONS **294** CITATIONS
SEE PROFILE

Matthew Davies
Institute for Systems and Computer Engineering, Technology and Science (INESC …
75 PUBLICATIONS **1,383** CITATIONS
SEE PROFILE

Carlos Guedes
New York University Abu Dhabi
60 PUBLICATIONS **759** CITATIONS
SEE PROFILE

A Hierarchical Harmonic Mixing Method

Gilberto Bernardes[1(✉)], Matthew E. P. Davies[1], and Carlos Guedes[1,2]

[1] INESC TEC, Sound and Music Computing Group,
Rua Dr. Roberto Frias, 378, 4200 - 465 Porto, Portugal
{gba,mdavies}@inesctec.pt
[2] New York University Abu Dhabi, PO Box 129188, Saadiyat Island,
Abu Dhabi, United Arab Emirates
carlos.guedes@nyu.edu

Abstract. We present a hierarchical harmonic mixing method for assisting users in the process of music mashup creation. Our main contributions are metrics for computing the harmonic compatibility between musical audio tracks at small- and large-scale structural levels, which combine and reassess existing perceptual relatedness (i.e., chroma vector similarity and key affinity) and dissonance-based approaches. Underpinning our harmonic compatibility metrics are harmonic indicators from the perceptually-motivated Tonal Interval Space, which we adapt to describe musical audio. An interactive visualization shows hierarchical harmonic compatibility viewpoints across all tracks in a large musical audio collection. An evaluation of our harmonic mixing method shows our adaption of the Tonal Interval Space robustly describes harmonic attributes of musical instrument sounds irrespective of timbral differences and demonstrates that the harmonic compatibility metrics comply with the principles embodied in Western tonal harmony to a greater extent than previous approaches.

Keywords: Music mashup · Digital DJ interfaces
Audio content analysis · Music information retrieval

1 Introduction

Mashup creation is a music composition practice strongly linked to the various sub-genres of Electronic Dance Music (EDM) and the role of the DJ [27]. It entails the recombination of existing (pre-recorded) musical audio as a means for creative endeavor [27]. As such, it can be seen as a byproduct of existing mass preservation mechanisms and inscribed within the artistic view of the database as a symbol of postmodern culture [20]. Mashup creation is typically confined to technology-fluent composers, as it requires expertise which extends from the understanding of musical structure to the navigation and retrieval of musical audio from large collections. Both industry and academia have been devoting efforts to enhance the experience of digital tools for mashup creation by streamlining the time-consuming search for compatible musical audio.

M. Aramaki et al. (Eds.): CMMR 2017, LNCS 11265, pp. 151–170, 2018.
https://doi.org/10.1007/978-3-030-01692-0_11

Early research on computational mashup creation, focused on rhythmic-only features, particularly those relevant to the temporal alignment of two or more musical tracks [13]. Recent research on this topic has expanded the range of musical attributes under consideration, notably including harmonic-driven features to identify compatible musical audio, commonly referred to as *harmonic mixing*. We can identify three major harmonic mixing methods: key affinity, chroma vectors similarity, and sensory dissonance minimization.

The affinity between musical keys is a prominent method in commercial applications. It is defined by distances across major and minor keys in a double circular representation, known within the DJ community as the *Camelot Wheel*, shown in Fig. 1. This method favors relative major-minor and intervals of fifth relations across musical keys [21] and enforces some degree of tonal stability and large-scale harmonic coherence of the mix by privileging the use of the same diatonic key pitch set. Chroma vector similarity inspects the cosine distance between chroma vector representations of pitch shifted versions of two given audio tracks as a measure of their compatibility [8,9,19]. Distances are typically computed at the beat level, thus privileging small-scale alignments over large-scale harmonic structure between audio slices with highly similar pitch class content. Sensory dissonance models have been used to search for pitch shifted versions of overlapping musical audio which minimize their combined level of roughness [12]; a motivation well rooted in the Western musical tradition by favoring a less dissonant harmonic lexicon.

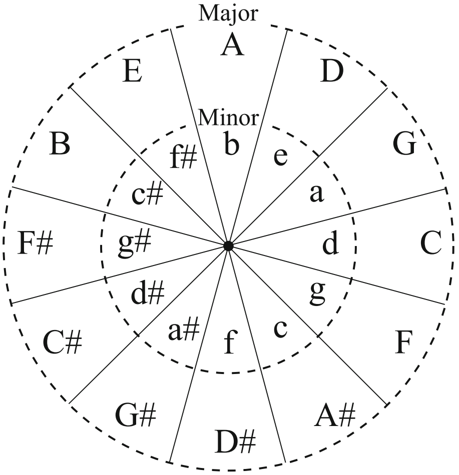

Fig. 1. Key affinity representation based on the two circles of fifths for major and minor keys aligned by relative major-minor key relationships. Enharmonic equivalence is assumed and only sharps (♯) are used.

While existing harmonic compatibility metrics have shown to correlate well with user enjoyment, we argue that they expose promising areas for investigation on harmonic compatibility. First, searching across all possible overlaps between related or in-key (i.e., diatonic) pitch sets result in highly contrasting sonorities with significant levels of enjoyment [3], thus motivating a harmonic compatibility metric below the key level. This metric is also prone to error, namely whenever processing signals with low pitch-to-noise ratio, and despite the perceptual manifestation of the key distances shown in Fig. 1 [17], it denotes temporal phenomena (i.e., key transitions). It remains unclear its validity and usefulness for mixing tracks. Second, while chroma vector distances are effective in capturing highly similar matches between any two given audio tracks, they lack a perceptually-aware basis for comparing pitch configurations [2], and can thus fail to provide an effective ranking between musical audio collections. Third, while psychoacoustic models show enhanced performance over existing approaches, they not only prove to be of limited use when the spectral content of the tracks do not overlap (i.e., when no interaction exists within each critical band), but also violate some perceptual and harmonic principles embodied in Western music, namely at the chordal level, by predicting that an augmented triad is more consonant that a diminished triad [16].

At the design level, existing software for harmonic mixing propose a ranked list of harmonically compatible tracks to a user-defined track [9,21,22]. We believe that this one-to-many mapping is reductive in offering a global view of a music collection and enabling a fluid navigation through an audio collection. Furthermore, it is computationally inefficient, as it recomputes highly intensive audio signal analysis every time a different audio track is selected as target.

In light of these limitations, we propose a new method for computing the small- and large-scale harmonic compatibility between a beat-matched collection of audio tracks, based on indicators from the perceptually-motivated Tonal Interval Space [2] and following the diagram architecture shown in Fig. 2. The proposed method has three aims: (i) to perceptually enhance the manifestation of metrics for harmonic compatibility, (ii) to inspect small- and large-scale structural levels by summarizing existing mashup creation approaches in a single framework, and (iii) to efficiently explore musical audio collections, without the need for intensive computation for each specific target, towards a more fluid user-experience which fosters experimentation.

The remainder of this paper is structured as follows. Section 2 reviews the Tonal Interval Space, which we adapt towards an enhanced representation of the harmonic content of musical audio. Section 3 presents content-driven harmonic analysis of musical audio. Section 4 introduces new metrics for computing the harmonic compatibility between audio tracks. Section 5 details an interactive visualization which exposes the compatibility of a musical audio collection. Section 6 presents an evaluation of the indicators and compatibility metrics which underpin our harmonic mixing method. Finally, Sect. 7 presents conclusions and areas for future work.

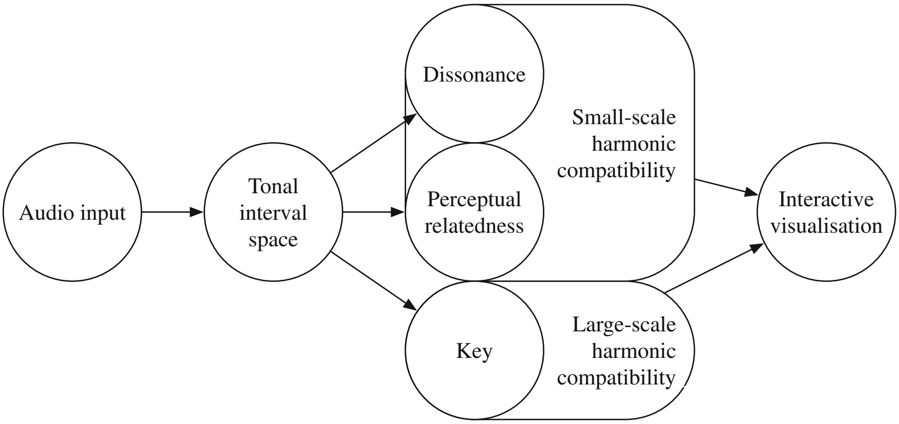
Audio input
Tonal interval space
Dissonance
Perceptual relatedness
Key
Small-scale harmonic compatibility
Large-scale harmonic compatibility
Interactive visualisation

Fig. 2. Diagram of the component modules of our compatibility method for music mixing.

2 Adapting Tonal Interval Vectors for Musical Audio

We represent the harmonic content of musical audio tracks as 12-dimensional Tonal Interval Vectors (TIVs) [2]. This vector space creates an extended representation of tonal pitch in the context of the *Tonnetz* [11], named the Tonal Interval Space, where the most salient pitch levels of tonal Western music—pitch, chord, and key—exist as unique locations. TIVs, $T(k)$, are computed from an audio signal as the weighted Discrete Fourier Transform (DFT) of an L_1 normalized chroma vector, $c(n)$, such that:

$$T(k) = w_a(k) \sum_{n=0}^{N-1} \bar{c}(n) e^{\frac{-j2\pi kn}{N}}, \quad k \in \mathbb{Z}. \tag{1}$$

where $N = 12$ is the dimension of the chroma vector, each of which expresses the energy of the 12 pitch classes, and $w_a(k)$ are weights derived from empirical ratings of dyads consonance used to adjust the contribution of each dimension k (or interpreted musical interval) of the space, which we detail over the next paragraphs. We set k to $1 \leq k \leq 6$ for $T(k)$ since the remaining coefficients are symmetric. $T(k)$ uses $\bar{c}(n)$ which is $c(n)$ normalized by the DC component $T(0) = \sum_{n=0}^{N-1} c(n)$ to allow the representation and comparison of music at different hierarchical levels of tonal pitch [2]. To represent variable-length audio tracks, we accumulate chroma vectors, $c(n)$, resulting from 16384 sample windows analysis at 44.1 kHz sampling rate ($\approx$ 372 ms) with 50% overlap across the track duration.

In [2], we used two complementary sources of empirical data—empirical ratings of dyad consonance shown in Table 1 [14] and the ranking order of triad consonance: {maj, min, sus4, dim, aug} [1,7]—to make the Tonal Interval Space perceptually relevant for symbolic music input representations (i.e., binary chroma vectors). Here, we revisit the task to comply with the timbral components of

Table 1. Composite consonance ratings of dyads consonance [14].

Interval class	m2/M7	M2/m7	m3/M6	M3/m6	P4/P5	TT
Consonance	–1.428	–.582	.594	.386	1.240	–.453

musical audio. Our goal is to find a set of weights, $w_a(k)$, which regulate the importance of the DFT coefficients k in Eq. 1, so that the space conveys a reliable consonance indicator correlated with the aforementioned empirical ratings of dyad [14] and triad consonance [7]. The applied method follows the previously used brute force approach [2], which produced a near optimal result.

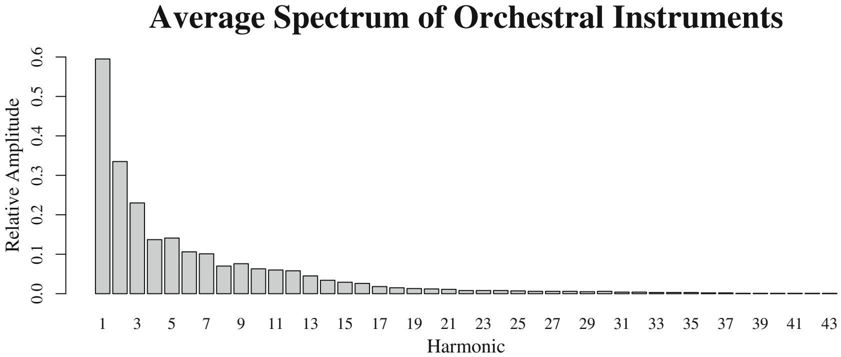

Fig. 3. Average harmonic spectrum of 1338 tones from orchestral instruments [23].

A major problem in defining a set of weights for robustly representing musical audio in the Tonal Interval Space is the variability of timbre across musical instruments and registers. A refined model capable of tracing the idiosyncratic timbral attributes of a particular instrument raises scalability and complexity issues which would defeat the value of the Tonal Interval Space in providing effective and, most importantly, efficient perceptual indicators of tonal pitch. To circumvent these issues, we adopt the 43-partial harmonic spectrum template shown in Fig. 3 to represent the harmonic content of musical audio. The template results from averaging 1338 recorded instrument tones from 23 Western orchestral instruments and can be understood as a time-invariant spectrum of an "average instrument" [23].

To allow a computationally tractable search for weights to represent musical audio in the Tonal Interval Space, we split the task into two steps. In the first step, we find the weights, $w_a(k)$, from all possible 6-element combinations (with repetition and order relevance), of the set $I = \{1, 19\} \in \mathbb{Z} : I = 2I + 1$ (a total of approximately one million combinations), which maintain in the Tonal Interval Space the empirical ranking order of common triads consonance [7]. Following [2], we compute the consonance of musical audio triads in the Tonal Interval Space as the norm of TIVs, $\|T(k)\|$, which we detail in Sect. 3. In the second step, from the resulting set of 111 weight vectors which preserve the ranking order of empirical triads consonance (i.e., a Spearman rank correlation $\rho = 1$), we identify those which have the highest linear correlation to the empirical dyad consonance ratings shown in Table 1.

We repeated the two aforementioned steps to further optimize the two weight vectors ($\{1, 7, 15, 11, 13, 7\}$ and $\{3, 7, 15, 11, 13, 7\}$) with the highest linear correlation ($r = .988$), below our minimal interval using 0.5 increments.

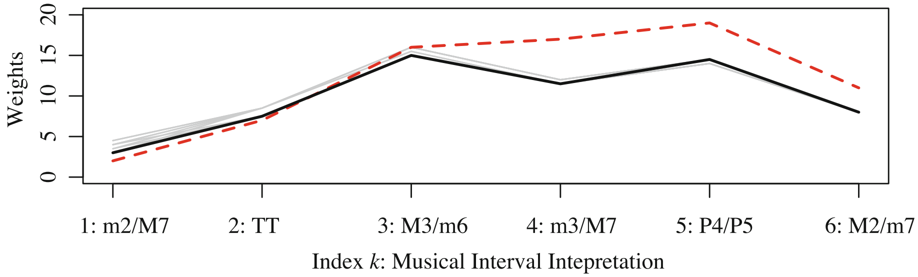

Fig. 4. The set of weights that maximize the linear ($r > 0.995$) and ranking order ($\rho = 1$) correlation of Tonal Interval Space's musical audio consonance indicator with empirical ratings of dyad and triad consonance, respectively. The bold line corresponds to the set of weights $w_a(k)$ used in Eq. 1 and the dashed line to the weights, $w_s(k)$, defined in [2] for a symbolic based Tonal Interval Space.

Figure 4 shows 11 sets of weights, $w_a(k)$, which preserve empirical triads ($\rho = 1$) and dyad ($r > .99$) consonance for musical audio. Given the inherent similarity in shape of the different sets of weights and their almost perfect linear relationship, we do not believe the choice over exactly which set of weights to be critical. Ultimately, we selected the weights with the greatest mutual separation between the triads according to consonance, thus $w_a(k) = \{3, 8, 11.5, 15, 14.5, 7.5\}$.

3 Harmonic Indicators from Musical Audio: Dissonance and Perceptual Relatedness

To provide a mathematical representation of Western tonal harmony perception as distances in the Tonal Interval Space, we distorted a DFT space according to the weights, $w_a(k)$, derived from empirical consonance ratings. In light of this design feature and following previous metrics detailed in [2], we can compute two indicators of musical audio dissonance, D, and perceptual relatedness, R, from the space as distance metrics.

$$D = 1 - \left(\frac{\|T(k)\|}{\|w_a(k)\|} \right), \tag{2}$$

$$D_{ij} = 1 - \left(\frac{\|a_i T_i(k) \ + a_j T_j(k)\|}{a_i + a_j \|w_a(k)\|} \right), \tag{3}$$

and

$$R_{i,j} = \sqrt{\sum_{k=1}^{M} |T_i(k) - T_j(k)|^2}. \tag{4}$$

We adapt the consonance metric presented in [2] to a musical audio dissonance metric, by subtracting the normalized norm of a TIV, $T(k)$, from one. Drawn from the properties of the DFT at the basis of the space, the location of multi-pitch TIVs is equal to the linear combination of its component pitch classes [2]. Thus, we can efficiently compute the dissonance of two combined TIVs, $T_i(k)$ and $T_j(k)$, representing two overlapping audio tracks, i and j, using Eq. 3, where a_i and a_j are the amplitudes of $T_i(k)$ and $T_j(k)$.

Equation 4 computes the perceptual relatedness, $R_{i,j}$, as the Euclidean distance between TIVs. Small values of perceptual relatedness, R, denote voice leading parsimony and controlled transitions of the interval content between neighborhood TIVs, as smaller distances primarily enforce the number of shared tones, and to a lesser degree, the interval relations imposed by the weights, $w_a(k)$.

4 Harmonic Compatibility Metrics

Based on the two dissonance, D, and perceptual relatedness, R, indicators from the Tonal Interval Space presented in Sect. 3, we now propose two metrics that aim at capturing the harmonic compatibility between TIVs to be mixed. Of note is the split between small- and large-scale harmonic compatibility, which roughly correspond to the 'sound object' and 'meso' or 'macro' time scales of music, respectively. In other words, the small-scale denotes the basic units of musical structure, from notes to beats, and the large-scale inspects the structural levels between the phrase and the overall musical piece architecture [24]. In the context of our work, the first aims mostly at finding good harmonic matches between the tracks in a collection, and the second in guaranteeing control over the overall harmonic structure of a mix, i.e., the tonal changes at the key level across its temporal dimension.

4.1 Small-Scale Harmonic Compatibility

The level of small-scale harmonic compatibility is expressed as the combination of two harmonic audio indicators from the Tonal Interval Space detailed in Sect. 3: dissonance, D, and perceptual relatedness, R. The latter indicator finds sonorities which have a strong perceptual affinity and thus range from a perfect match to sonorities with different timbres and similar pitch content, to an array of sonorities with increased levels of perceptual distance. We envisage it as an extension of the chroma vector similarity method used as a measure of harmonic compatibility in prior studies [8,9,19], which offers a refined control over the introduction of new tones as well as its interval relations in the resulting mix between overlapped tracks.

Given the likely increase in dissonance in the mix and following some perceptual evidence from previous research [12], our small-scale harmonic compatibility also privileges the search for less dissonance mixes—a well established principle in the common syntax of Western tonal harmony [1,5,18]. Hence, our small-scale harmonic compatibility metric, H, is then computed as the product of the two indicators, such that:

$$H_{i,j} = \bar{R}_{i,j} \cdot \bar{D}_{ij}, \tag{5}$$

where $\bar{R}$ and $\bar{C}$ are R and C scaled to the range $\{0,1\} \in \mathbb{R}$ to balance the importance of both indicators in the compatibility metric. The main motivation for the simple multiplication of the two variables is rooted in the visualization method we detail in Sect. 5, notably by enforcing a small-scale harmonic compatibility, $H = 0$, when comparing the same track.

4.2 Large-scale Harmonic Compatibility

A derivation of the perceptual relatedness indicator, R, exposes an important property of the Tonal Interval Space: the formation of fuzzy key clusters of diatonic pitch class sets. Neighborhood relations between these clusters in the Tonal Interval Space result in a representation similar to Fig. 1, as a result of common-tone relations between keys. This property is adopted to estimate the global key from musical audio track, which aims to guide users in planning the large-scale harmonic structure of a mix.

We use the method reported in [4] to compute the global key estimate, Q, of an audio track in the Tonal Interval Space, as the minimum Euclidean distance of an audio input TIV, $T(k)$, from the 12 major and 12 minor key TIVs, $T_r(k)$, such that:

$$Q = \text{argmin}_p \sqrt{\sum_{k=1}^{6} |T(k) \cdot \alpha - T_r(k)|^2}, \tag{6}$$

where $T_r(k)$ is derived from a collection of templates (understood here as chroma vectors) representing pitch class distributions for each of the 12 major and 12 minor keys [26]. When $r \leqslant 11$, we adopt the major profile and when $r \geqslant 12$, the minor profile. $\alpha = 0.35$ is a factor which displaces input sample TIVs to balance predictions across modes [4] (Table 2).

The estimated key, Q, ranges between $0 - 11$ for major keys and $12 - 23$ for minor keys, where 0 corresponds to C major, 1 to C# major, and so on through to 23 being B minor.

5 Interactive Visualization

We created a software prototype in Pure Data which implements the proposed hierarchical mixing method, notably the harmonic compatibility metrics. In light

Table 2. Sha'ath's [26] key profiles, p, for the C major and C minor keys.

Key	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
C major	7.239	3.504	3.584	2.845	5.819	4.559	2.448	6.995	3.391	4.556	4.074	4.459
C minor	7.003	3.144	4.359	5.404	3.672	4.089	3.907	6.200	3.634	2.872	5.355	3.832

of the possibility to retrieve compatible harmonic tracks from large musical audio collections, we designed an interactive visualization which aims to (i) provide a global view of the harmonic compatibility across all audio tracks in a collection (i.e., many-to-many relationships); (ii) expose a hierarchical representation over the harmonic compatibility between tracks; and (iii) promote user experimentation and creative endeavorer.

To this end, we pursued an interface design based on crossmodal associations between sound and image, for which a screenshot is shown in Fig. 5. All audio tracks in a collection are represented graphically in a two-dimensional (2-D) space, where regular polygons denote audio tracks and grey circles key centers. Distances among polygons (i.e., audio tracks) indicate small-scale harmonic compatibility, H, and links from circles (i.e., key centers) to polygons the large-scale compatibility, Q, or, in other words, the association to its estimated key.

The computation of 2-D coordinates for each audio track from a square matrix of all pairwise tracks small-scale harmonic compatibility distances, H, is a classical problem, which can be solved by a specific class of algorithms, notably including Multidimensional Scaling (MDS). From m musical audio tracks in a collection, we compute an $H_m \cdot H_m$ harmonic compatibility square matrix, from which an MDS representation extracts two-dimensional coordinates for each sample. The resulting representation attempts to preserve the inter-sample small-scale harmonic compatibility with minimal distortion.

The computation of coordinates for each key center equals the convex combination (or the centroid, in geometric terms) of the 2-D coordinates of its estimated tracks. We adopted this efficient method for the computation of key centers based on the theoretical assumption that the diatonic set of a key is denoted in the Tonal Interval Space by the convex combination of set of diatonic note TIVs [2]. Therefore, assuming that all tracks with the same key estimate represent well its diatonic set, its corresponding key coordinates ought to be represented with minimal distortion.

Two additional rhythmic and spectral musical features of each track are represented by graphical attributes of the polygons (number of sides and color, respectively) to expand the search attributes to fit particular compositional goals. The number of sides, ranging from three to six, expose the note onset density, computed by a threefold approach. First, we extract a spectral flux onset detection function, from a windowed power spectrum representation of the audio signal (2048 analysis windows size at 44.1 kHz sampling rate with 50% overlap), using the `timbreID` [6] library within Pure Data. Second, we identify the peaks from the function above a user-defined threshold, t, whose temporal location we assume to indicate note onset times. Prior to the peak detection stage, we apply a bi-directional low-pass IIR filter, with a cutoff frequency of 5 Hz to avoid spurious detections. Finally, we compute the ratio between the number of onsets and the entire duration of the audio file in seconds and scale the values for a given audio collection to the $\{3, 6\} \in \mathbb{Z}$ range of polygon sides.

The polygons' color, ranging from continuous shades of yellow to red, represents the spectral region a sample occupies in the perceptual perceptually-motivated Bark frequency scale.[1] A threefold strategy is adopted to map these two dimensions. First, we accumulate Bark spectrum B_b, representations computed on short-time windows of 2048 samples size at 44.1 kHz sampling rate with 50% overlap across an audio track, again using the `timbreID` [6] library within Pure Data. Then, we extract the centroid as an indicator of its spectral region, S, using Eq. 7. Finally, we map the spectral region, S, value to the the color scheme. Bark band 1 corresponds to yellow, and bark band 24 to red. Between these values, the colors are linearly mixed.

$$S = \frac{\sum_{i=1}^{19} B_b \cdot b}{\sum_{i=1}^{19} B_b}, \tag{7}$$

where B_b is the energy of the bark band b. The S indicator can range from 1 to 24.

The user can interact with the visualization by clicking on the polygons to trigger their playback, thus promoting an intuitive search for compatible tracks as well as strategies for serendipity and experimentation, rather than a fully automatic method for mashup creation. A demo of this interactive visualization can be found online at: https://sites.google.com/site/tonalintervalspace/mixmash.

6 Evaluation

We undertake a twofold strategy to evaluate our harmonic mixing method. First, we assess the perceptual validity and degree of timbre invariance of the two dissonance, D and perceptual relatedness, R, indicators from the Tonal Interval Space, which underpin our small-scale harmonic compatibility metric. Particular emphasis is given to the implications of the newly proposed weights, $w_a(k)$, for representing musical audio. Second, we examine the level of compliance of the proposed harmonic compatibility and related metrics with Western tonal music principles.

Unless otherwise specified, across evaluation tasks the harmonic spectrum of musical audio is computed as the sum of individual notes spectra using the harmonic template of an average instrument shown in Fig. 3.

[1] The Bark spectrum balances the resolution across the human hearing range in comparison to the typical power spectrum representation, namely increasing the resolution in the low frequency region. It is computed by warping a power spectrum to the 24 critical bands of the human auditory system [28].

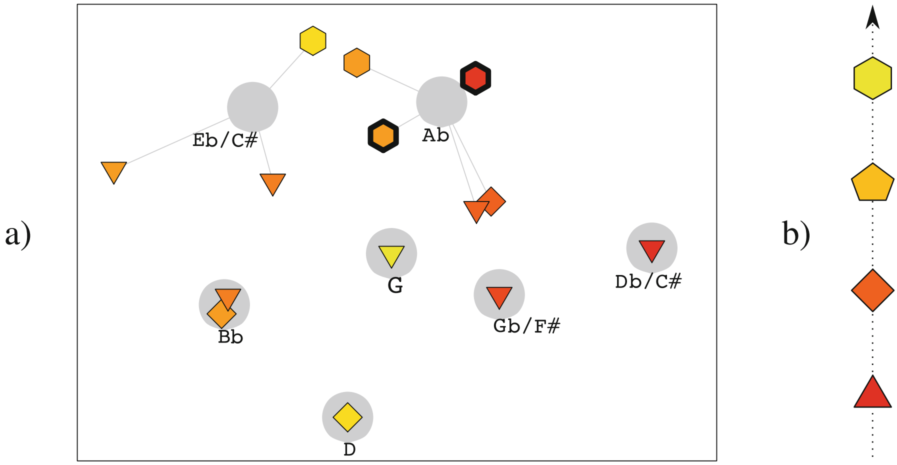

Fig. 5. (a) Interactive visualization of the hierarchical harmonic compatibility between all audio tracks in a collection. Polygons represent audio tracks and circles key centers. Polygon distances indicate small-scale harmonic compatibility and the links from circles the large-scale harmonic compatibility. The graphical attributes of the polygons show onset density (number of sides) and spectral region (color). Polygons with thick outlines indicate the selected files currently playing. (b) The ranking order of low to high onset density and spectral region.

The Spearman rank correlation, ρ, is the metric used to compare most data in our evaluation. It measures the strength and direction of the monotonic relationship between two variables. The motivation to adopt such a metric is due to the importance of the ranking order in proposing harmonic mixes rather than ensuring a linear relationship between the variables (computed, for example, by the Pearson correlation) and the prevalence of ranked data in perceptual studies. The result of the Spearman rank correlation, ρ, is expressed by a single correlation coefficient value in the $\{-1, +1\} \in \mathbb{R}$ range. Positive and negative correlation coefficients express positive and negative relationships between variables, respectively, and a correlation coefficient of $\rho = 0$ indicates that no relationship between the variables exists.

6.1 Harmonic Indicators for Musical Audio

We assess how the weights, $w_a(k)$, implemented as a design feature of the space, to provide a dissonance, D, indicator of musical audio, compare to (i) the sensory dissonance metric by Hutchinson and Knopoff [15] at the basis of Gebhardt et al. [12] mashup creation system and (ii) the previously proposed weights, $w_s(k) = \{2, 11, 17, 16, 19, 7\}$, adjusted for symbolic music representations [2], in measuring triads dissonance—the chordal level at which most mashup creation exists. All theoretical models are additionally compared to perceptual dissonance data [1,7].

Then, we assess how the perceptual relatedness, R, metric and the cosine similarity between chroma vectors, adopted from Davies et al. [9], as a harmonic compatibility metric, compare to perceptual data [25]. The dyad pitch level is used to undertake this comparison as it lays out the basis for distances at all higher hierarchical pitch levels. Furthermore, various corroborating perceptual studies exist for this pitch level, which Schwartz et al. summarize in [25].

Finally, we determine the degree of timbre invariance of both Tonal Interval Space indicators across a wide range of musical instrument timbres. To this end, we extend the two previous tasks beyond the theoretical levels by evaluating the indicators across multiple musical instruments and registers. To constrain the experiment, we limited the pitch sets to triads in the root position with stacked thirds and dyads no larger than one octave.

The pitch sets result from the sum of individual note recordings from acoustic and electronic instruments. IRCAM's Studio OnLine (SOL) database[2] is adopted for the acoustic instruments and the NSynth database [10] for the electronic instruments. From the entire collection of acoustic instruments in the SOL database, we selected four instruments from each family, aiming to cover a wide note range: strings (violin, viola, guitar, and violoncello), woodwinds (flute, B♭ clarinet, alto saxophone, and bassoon), and brass (trumpet, trombone, horn, and bass tuba). From the NSynth database, we selected four electronic and synthetic instruments which commonly feature in EDM: electric keyboard, synth lead, and electric bass, and electric guitar.

The selected acoustic instrument samples are quasi-stationary, i.e., without any extended playing technique, and electronic and synthetic instrument samples are non-stationary, with clear temporal changes at a regular fast pace, as a result of audio effects such as tremolo, vibrato, and filtering.[3] A *mezzo-forte* dynamic was adopted in both cases. Besides the alignment of the samples on detected onsets, no further processing was applied. Due to some discrepancies in the duration of the instrument samples in both databases, we limited their duration to two seconds, thus ensuring the same duration across all pitch sets. Instrument samples are mono WAV files with 44.1 Khz sampling rate and 16 bit depth. We computed the indicators (i.e., the dissonance, D, of triads and the perceptual relatedness, R, of dyads) per instrument as the average value of overlapping 8192 sample windows at 44.1 kHz sample rate with 50% overlap.

6.2 Harmonic Compatibility Metrics

We assess the extent to which the principles embodied in Western tonal harmony, namely the prevalence of common chord sets with reduced dissonance, are promoted by our proposed small-scale harmonic compatibility metric, H, and

[2] We used the version 0.9 of IRCAM's SOL database, retrieved at http://forumnet.ircam.fr/product/orchids-en/ in July, 2017 as the supporting database of the Orchids software.

[3] Please refer to https://sites.google.com/site/tonalintervalspace/mixmash to listen to electronic and synthetic instrument sample examples from the NSynth database.

related approaches, namely chroma similarity [9] and sensory dissonance [12]. To this end, we inspect which pitch class sets are identified as most compatible from a total of 55 triads, which result from overlapping the pitch class `C` or 0 (i.e., index 0 in $c(n)$ from Eq. 1) to all remaining pitch class dyads, in each metric. All dyads resulting from the combination (without repetition and order relevance) of the $\{1-11\} \in \mathbb{Z}$ set are considered.

6.3 Results

Table 3 reports the perceived and computed dissonance level of common musical audio triads. The perceptual data have been corroborated by several experimental studies [1,7]. The values for sensory dissonance are taken from [15] and result from applying the metric to triads that lie within the $\mathtt{C}^4$-$\mathtt{C}^5$ octave. The reported Spearman rank correlations, ρ, and their significance values, p, between the perceptual data and theoretical models show that the Tonal Interval Space is more consistent in ranking the dissonance of common triads than the Hutchinson and Knopoff [15] sensory dissonance model used in related mashup literature [12]. Moreover, we demonstrate that the weights, $w_a(k)$, computed in Sect. 2 are decisive in capturing the dissonance of musical audio triads in the Tonal Interval Space, as the previously proposed weights, $w_s(k)$ for symbolic music inputs [2] fail at providing a ranking of triads consonance from musical audio in line with perceptual data.

Table 3. Ranking of triads dissonance from perceptual data [1,7] and two theoretical models: sensory dissonance [15] and the Tonal Interval Space dissonance, adopting two sets of weights adapted to symbolic representation, $w_s(k)$, and musical audio, $w_a(k)$. The Spearman rank correlations, ρ, and their significance values, p, between the perceptual data and theoretical models are reported.

Triad quality	Perceptual rank [1,7]	Sensory dissonance [15]	Tonal Interval Space (D)	
major	1	1 (.139)	1–2 (.768)	1–2 (.783)
minor	2	2 (.148)	1–2 (.768)	1–2 (.783)
sus4	3	4 (.228)	3 (.769)	3 (.784)
dim	4	5 (.230)	5 (.819)	4 (.805)
aug	5	3 (.149)	4 (.780)	5 (.806)
Correlation ρ		.700	.878	.975
Significance p		.233	.054	<.05

Table 4 reports the perceived and computed dyads relatedness, R. The perceptual data are taken from [25], which summarizes different experimental studies. The chroma similarity was computed as the cosine similarity between chroma vectors following [9]. The reported Spearman rank correlations, ρ, and their significance values, p, between the perceptual data and theoretical models show that the Tonal Interval Space is more consistent in ranking the perceptual relatedness of dyads than the chroma similarity. Moreover, the ranking order of dyad relatedness in the Tonal Interval Space is consistent with tonal harmony principles in the sense it promotes tertian harmony as a result of having fifths and thirds at a closer distance than all remaining intervals. The chroma similarity, adopted as a harmonic compatibility metric in previous computational mashup works [8,9,19], largely agrees with the dyad perceptual ranking, with the notable exception of the minor seconds and major seventh which are closer in this metric space than the major and minor thirds or their complementary minor and major sixth, thus disrupting a preference for tertian harmonies.

Table 4. Ranking of dyad relatedness from perceptual data and theoretical models. The Spearman rank correlations, ρ, and their significance values, p, between the perceptual data and theoretical models are reported.

Dyad	Perceptual rank [25]	Chroma similarity [9]	Tonal Interval Space (R)
Unison (P1)	1	1−2 (0.00)	1−2 (0.00)
Octave (P12)	2	1−2 (0.00)	1−2 (0.00)
Perfect fifth (P5)	3	3−4 (1.11)	3−4 (1.37)
Perfect fourth (P4)	4	3−4 (1.11)	3−4 (1.37)
Major third (M3)	5	7−8 (1.20)	7−8 (1.62)
Major sixth (M6)	6	10−11 (1.26)	5−6 (1.53)
Minor sixth (m6)	7	7−8 (1.20)	7−8 (1.62)
Major third (M3)	8	10−11 (1.26)	5−6 (1.53)
Tritone (TT)	9	9 (1.25)	9 (1.78)
Minor seventh (m7)	10	12−13 (1.27)	10−11 (1.79)
Major second (M2)	11	12−13 (1.27)	10−11 (1.79)
Major seventh (M7)	12	5−6 (1.16)	12−13 (1.95)
Minor second (m2)	13	5−6 (1.16)	12−13 (1.95)
Correlation ρ		.609	.956
Significance p		$< .05$	$< .001$

Table 5 shows the Spearman rank correlations between perceptual data [1, 15, 25] and Tonal Interval Space indicators from musical instrument inputs. We inspected the dissonance, D, of common triads and the perceptual relatedness, R of dyads. With the sole exception of the electric bass, all instruments have a significant Spearman rank correlation between the perceptual data and their computed indicators, thus ensuring a high degree of timbral invariance in computing the harmornic indicators from the Tonal Interval Space.

These results should also be read in light of the Spearman correlation between theoretical musical audio representations and perceptual data shown in Tables 3 and 4. The triad dissonance that results from instrument recordings does not fully mirror the perfect monotonic relationship of the theoretical results. Looking across instrument families it is noticeable the optimal results across all inspected

Table 5. Spearman rank correlation, ρ, of perceptual data for triads consonance and dyads distances and dissonance, D, and perceptual relatedness, R, metrics from the Tonal Interval Space, respectively, across multiple instruments. All results are significant for $p < 0.05$, except for the electric bass dissonance, where $p = 0.233$.

	Intrument	Pitch range (MIDI)	Dissonance (D)	Perceptual relatedness (R)
Strings	Violin	55–100	1	.956
	Viola	48–96	.9	.973
	Guitar	38–83	.8	.896
	Violoncello	36–84	.9	.973
Woodwinds	Flute	59–96	.9	.945
	B♭ clarinet	50–91	1	.956
	Alto saxophone	49–81	.9	.940
	Bassoon	34–75	1	.934
Brass	Trumpet	54–86	1	.951
	Trombone	34–72	1	.951
	Horn	31–77	1	.956
	Bass tuba	30–65	1	.945
Electronic	Electric guitar	36–86	1	.951
	Synth lead	21–108	.9	.934
	Electric keyboard	21–108	.9	.912
	Electric bass	9–96	.7	.900

instruments in the brass family. The remaining families have the approximately same number of instruments which do not fully comply with the perceptual triad ranking. The small sample of observed instruments raises interesting issues which should ultimately be addressed in future adaptations of the Tonal Interval Space. The dyad perceptual relatedness that results from instrument recordings is in line with the theoretical results ($\rho = 956$).

Figure 6 shows the analysis of the electronic instruments, which the "average (acoustic) instrument" spectrum, at the basis of the adaptation of the Tonal Interval Space to musical audio, does not model. In both the triad dissonance and dyad perceptual relatedness plots, an monotonically increasing function is expected, given the order of the x-axis elements according to an ascending perceptual ranking. In the triad dissonance plot, we can observe that the perceptual ranking is not preserved by violating the order of different triads per instrument. The dyad perceptual relatedness plot questions the symmetric property of the space (as a result of the DFT) for complementary intervals (e.g., m2 and M7 or M2 and m7) whose dissonance levels are averaged, thus neglecting any distinction between them.

Figure 7 shows the seven best ranking triads resulting from the overlap of the pitch class `C` and all remaining 55 pitch class dyad combinations for the following five metrics: (i) sensory dissonance [15]; (ii) chroma similarity [9]; (iii) perceptual relatedness, R; (iv) dissonance, D; and (v) small-scale harmonic compatibility,

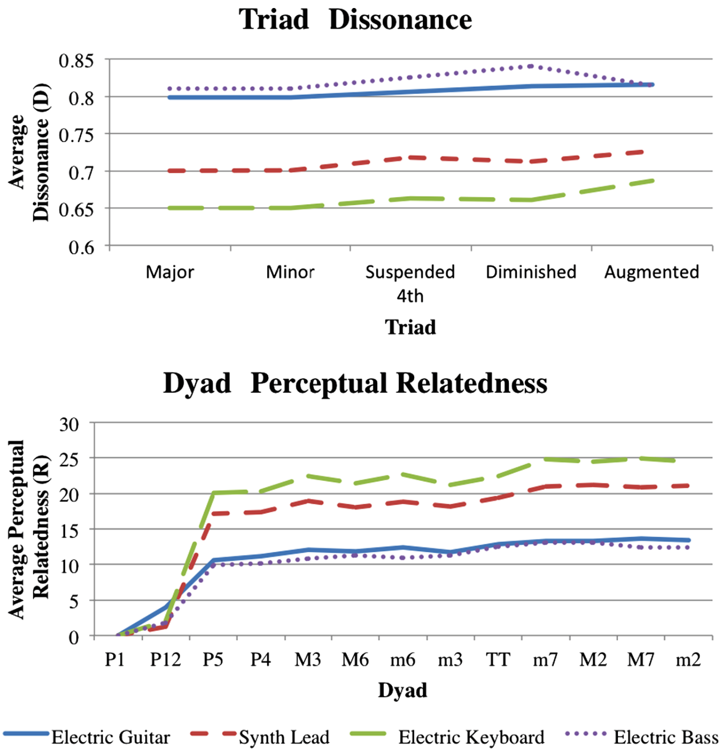

Fig. 6. Computed dyad dissonance and triad perceptual relatedness of electronic and synthetic instrument samples from the Tonal Interval Space. Pitch sets result from summing individual instrument samples. Average values across the inspected instruments' pitch range are reported for both indicators.

H. Sensory dissonance ranking order favors triads in line with the results presented in Table 3. Besides the preference for augmented over suspended 4ths and a few minor triads, to a large extent it conveys the expectancy of the Western tonal syntax. The ranking order of triads in the chroma similarity space and the perceptual relatedness, R, are aligned with the findings shown in Table 4. The chroma similarity favors chords including P4/P5 and m2/M7 intervals. Conversely, the perceptual relatedness, R, favors chords including P5/P4, m3/M6, and M3/m7 intervals. As such, combining sonorities with small R values, results in extended chords with stacked fifths and thirds. However, while the chords resulting from these vertical aggregates are building blocks of Western tonal harmony, the best ranked chords result in multiple seventh chords (with omitted notes), and not, as expected in the ideal case, triads (e.g., major, minor, and diminished).

When combining the perceptual relatedness, R, and the dissonance, D, indicators, i.e., when adopting our small-scale harmonic compatibility, H, metric, we enforce the preference for common major, minor, and suspended fourth triads—the most common building blocks of the Western tonal harmony, by favoring in the former ranking less dissonant triads. Nonetheless, the small-scale harmonic compatibility, H, ranking in Fig. 7 ignores the key level, thus promoting chromaticism (non-diatonic) progressions between neighbor sonorities. Our large-scale harmonic compatibility addresses this issue by providing in our interactive visualization a layer of information which can guide users in selecting (in-key) diatonic mixes.

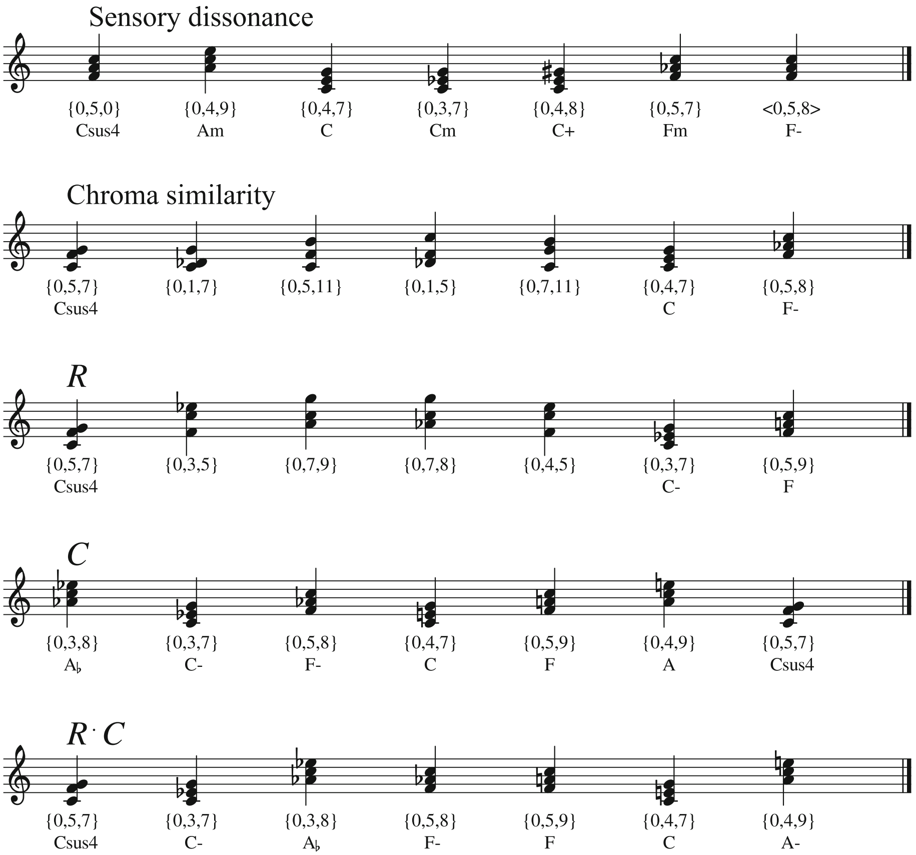

Fig. 7. Ranking order of triads resulting from overlapping the pitch class C and all remaining pitch class dyads as given by computed metrics. Apart from triad sensory dissonance, which is taken from [15], the remaining models were computed using the harmonic spectrum representation of an "average instrument" spectrum shown in Fig. 3. To the pitch class set of the resulting triads, we include the chord label whenever unambiguous and complete triads are formed. Sounding examples of the table contents are available at: https://sites.google.com/site/tonalintervalspace/mixmash.

7 Conclusion and Future Work

In this paper we have presented a hierarchical harmonic mixing method with two underlying metrics that inspect the harmonic compatibility of musical audio tracks at both small- and large-scale structural levels. Small-scale harmonic compatibility results from the combination of dissonance and perceptual relatedness indicators from the Tonal Interval Space, which we adapted to represent musical audio. Our adaptation is largely invariant to timbral differences of instrument sounds, and aims to assist users in finding good local alignments between mixed tracks. Large-scale harmonic compatibility relies on key estimates and aims to assist users in planning the global harmonic structure of a mix. A software prototype in Pure Data presents the metrics to the user in an interactive visualization. Crossmodal associations between sound attributes and geometric elements aim at promoting a global exploration of an audio collection, namely by fostering a fluid strategy to retrieve harmonically compatible tracks.

In future work, we plan to address three important issues raised by the current evaluation. The first is related to the perceptual validity of the harmonic compatibility metrics. Despite the perceptual motivation of its indicators, their combination as a result of a simple multiplicative fashion, remains open. Our evaluation shows that the model leads to an attractive theoretical result, but we speculate that a better perceptually-grounded quantification can be found.

The second issue is related to the relevancy of the proposed method for EDM and the creative flow of the DJ in meaningful world-case application scenarios. One can argue whether the design of the Tonal Interval Space design oriented towards Western tonal music harmony is a prominent dimension in EDM. This ought to be investigated by a broader study which not only takes into account the various dimensions of musical structure at hand, but also the user experience promoted by our interface in the context of EDM practice. In response, we can state that the interface design reflects these concerns as it guides the users through the creative process based on metrics aligned with some perceptual findings rather then dictating or automating the process based on some judgment about the harmonic quality of the music. To this end, we believe that a large degree of creative endeavor can be achieved.

The third aspect under consideration in future work relates to the scalability of the data under analysis. The current interactive interface is inefficient when we scale the musical collection above a certain number of tracks, as it results in dense cluttered visual clusters. Given the goal of inspecting large musical audio collections at the level of the hundreds or thousand tracks, we aim to address strategies to enhance the sparseness of the visualization.

Acknowledgments. This work is supported by national funds through the FCT - Foundation for Science and Technology, I.P., under the project IF/01566/2015.

References

1. Arthurs, Y., Beeston, A.V., Timmers, R.: Perception of isolated chords: Examining frequency of occurrence, instrumental timbre, acoustic descriptors and musical training. Psychol. Music **46**(5), 662–681 (2018). https://doi.org/10.1177/0305735617720834
2. Bernardes, G., Cocharro, D., Caetano, M., Guedes, C., Davies, M.: A multi-level tonal interval space for modelling pitch relatedness and musical consonance. J. New Music Res. **45**(4), 281–294 (2016)
3. Bernardes, G., Cocharro, D., Guedes, C., Davies, M.E.P.: Harmony generation driven by a perceptually motivated tonal interval space. ACM Comput. Entertain. **14**(2), 6 (2016)
4. Bernardes, G., Davies, M., Guedes, C.: Audio key estimation with adaptive mode bias. In: Proceedings of ICASSP, pp. 316–320 (2017)
5. Bidelman, G.M., Krishnan, A.: Brainstem correlates of behavioral and compositional preferences of musical harmony. Neuroreport **22**, 212–216 (2011)
6. Brent, W.: A timbre analysis and classification toolkit for pure data. In: Proceedings of ICMC, pp. 224–229 (2010)
7. Cook, N.: Harmony, Perspective, and Triadic Cognition. Cambridge University Press, Cambridge (2012)
8. Davies, M., Stark, A., Gouyon, F., Goto, M.: Improvasher: a real-time mashup system for live musical input. In: Proceedings of NIME, pp. 541–544 (2014)
9. Davies, M.E.P., Hamel, P., Yoshii, K., Goto, M.: Automashupper: automatic creation of multi-song music mashups. IEEE Trans. ASLP **22**(12), 1726–1737 (2014)
10. Engel, J., et al.: Neural audio synthesis of musical notes with WaveNet autoencoders. In: Proceedings of the 34th International Conference on Machine Learning, pp. 1068–1077 (2017)
11. Euler, L.: Tentamen novae theoriae musicae. Broude (1968/1739)
12. Gebhardt, R., Davies, M., Seeber, B.: Psychoacoustic approaches for harmonic music mixing. Appl. Sci. **6**(5), 123 (2016)
13. Griffin, G., Kim, Y., Turnbull, D.: Beat-sync-mash-coder: A web application for real-time creation of beat-synchronous music mashups. In: Proceedings of ICASSP, pp. 437–440 (2010)
14. Huron, D.: Interval-class content in equally tempered pitch-class sets: common scales exhibit optimum tonal consonance. Music Percept. **11**(3), 289–305 (1994)
15. Hutchinson, W., Knopoff, L.: The acoustic component of western consonance. J. New Music Res. **7**(1), 1–29 (1978)
16. Johnson-Laird, P.N., Kang, O.E., Leong, Y.C.: On musical dissonance. Music Percept. **30**(1), 19–35 (2012)
17. Krumhansl, C.L., Kessler, E.J.: Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys. Psychol. Rev. **89**, 334–368 (1982)
18. Lahdelma, I., Eerola, T.: Mild dissonance preferred over consonance in single chord perception. i-Perception (2016). https://doi.org/10.1177/2041669516655812
19. Lee, C.L., Lin, Y.T., Yao, Z.R., Lee, F.Y., Wu, J.L.: Automatic mashup creation by considering both vertical and horizontal mashabilities. In: Proceedings of ISMIR, pp. 399–405 (2015)
20. Manovich, L.: The Language of New Media. MIT Press, Cambridge (2001)
21. Mixed in Key: Mashup 2 [software]. http://mashup.mixedinkey.com. Accessed 28 Mar 2017

22. Native Instruments: Traktor pro 2 [software]. https://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-2/. Accessed on 1 Sep 2017
23. Plazak, J., Huron, D., Williams, B.: Fixed average spectra of orchestral instrument tones. Empirical Musicol. Rev. **5**(1), 10–17 (2010)
24. Roads, C.: Microsound. MIT Press, Cambridge (2004)
25. Schwartz, D.A., Howe, C., Purves, D.: The statistical structure of human speech sounds predicts musical universals. J. Neurosci. **23**(18), 7160–7168 (2003)
26. Sha'ath, I.: Estimation of key in digital music recordings. Master's thesis, Birkbeck College, University of London (2011)
27. Shiga, J.: Copy-and-persist: the logic of mash-up culture. Crit. Stud. Media Commun. **24**(2), 93–114 (2007)
28. Zwicker, E., Fastl, H.: Psychoacoustics-Facts and Models. Springer, Heidelberg (1990). https://doi.org/10.1007/978-3-540-68888-4