HARMONIC MIXING BASED ON ROUGHNESS AND PITCH COMMONALITY

Roman Gebhardt
Audio Information Processing,
Technische Universität München
Munich, Germany
roman.gebhardt@tum.de

*Matthew E. P. Davies**
Sound and Music Computing Group,
INESC TEC
Porto, Portugal
mdavies@inesctec.pt

Bernhard Seeber†
Audio Information Processing,
Technische Universität München
Munich, Germany
seeber@tum.de

ABSTRACT

The practice of harmonic mixing is a technique used by DJs for the beat-synchronous and harmonic alignment of two or more pieces of music. In this paper, we present a new harmonic mixing method based on psychoacoustic principles. Unlike existing commercial DJ-mixing software which determine compatible matches between songs via key estimation and harmonic relationships in the circle of fifths, our approach is built around the measurement of musical consonance at the signal level. Given two tracks, we first extract a set of partials using a sinusoidal model and average this information over sixteenth note temporal frames. Then within each frame, we measure the consonance between all combinations of dyads according to psychoacoustic models of roughness and pitch commonality. By scaling the partials of one track over $\pm$ 6 semitones (in 1/8th semitone steps), we can determine the optimal pitch-shift which maximises the consonance of the resulting mix. Results of a listening test show that the most consonant alignments generated by our method were preferred to those suggested by an existing commercial DJ-mixing system.

1. INTRODUCTION

The digital era of DJ-mixing has opened up DJing to a huge range of users, and also enabled new technical possibilities in music creation and remixing. The industry leading DJ-software tools (e.g., Native Instruments Traktor Pro 2[1], djay Pro[2] and Mixed in Key[3]) now offer users of all technical abilities the opportunity to rapidly and easily create DJ mixes out of their personal music collections, or those stored online. Central to these DJ-software tools is the ability to robustly identify tempo and beat locations, which, when combined with high quality audio time-stretching, allow for automatic "beat-matching" (i.e. temporal synchronisation) of music.

* MD is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia within post-doctoral grant SFRH/BPD/88722/2012.

† BS is supported by BMBF 01 GQ 1004B (Bernstein Center for Computational Neuroscience Munich).

[1] http://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-2/

[2] http://www.algoriddim.com/djay-mac

[3] http://www.mixedinkey.com/

In addition to leveraging knowledge of the beat structure, these tools also extract harmonic information – typically in the form of an estimated key. Knowing the key of different pieces of music allows users to engage in so-called "harmonic mixing" where the aim is not only to align music in time, but also in key. Different pieces of music are deemed to be harmonically compatible if their keys exactly match or adhere to well-known relationships within the circle of fifths. When this information is combined with audio pitch-shifting functionality (i.e., the ability to transpose a piece of music by some number of semitones independent of its temporal structure) it provides a powerful means to "force" the harmonic alignment between two pieces of otherwise incompatible music.

While such a combination of robust music understanding and high quality music signal processing techniques is certainly effective within specific musical contexts – in particular for harmonically and temporally stable house music (and other related genres), we believe the key-based matching approach has several important limitations. Putting aside the primary issue that the key estimation itself might be error-prone, the most critical limitation is that a global property such as musical key provides no information regarding the musical composition which gives rise to that key nor how this might affect perceptual harmonic compatibility for listeners when two pieces are mixed. Similarly, music matching based on key alone provides no obvious means for ranking the compatibility between several different pieces of the same key. Likewise, assigning one key for the duration of a piece of music cannot indicate where in time the best possible mixes (or mashups) between different pieces of music might occur. Even with the ability to use pitch-shifting to transpose the musical key, it is important to consider the quantisation effect of only comparing whole semitone shifts. The failure to consider fine-scale tuning could lead to highly dissonant mistuned mixes between songs which still share the same key.

To attempt to address these limitations of key-based harmonic mixing, we propose a new approach based on the analysis of consonance. We base our approach on the well-established psychoacoustic principles of sensory consonance and harmony as defined by Ernst Terhardt [1, 2], where our goal is to discover the optimal, consonance-maximising alignment between two music excerpts. To this end, we first extract a set of frequencies and amplitudes

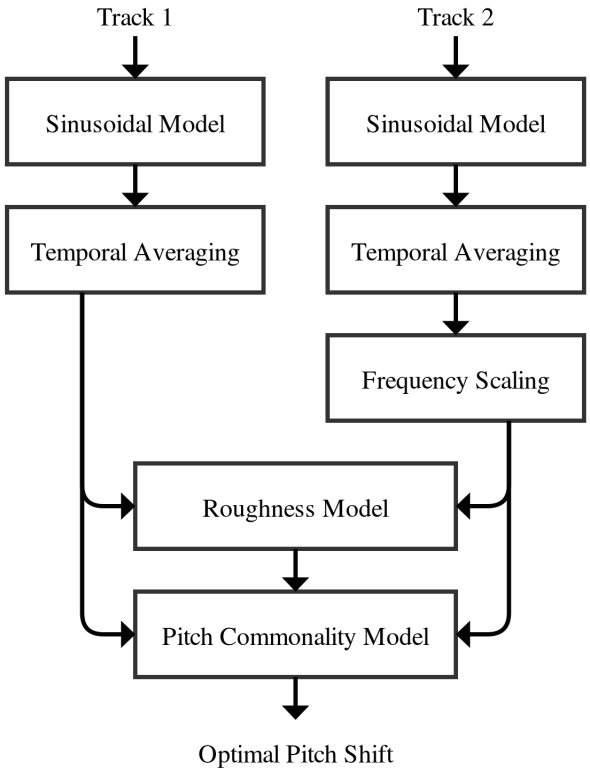

Figure 1: *An overview of the proposed approach for consonance-based mixing.*

using a sinusoidal model and average this information over short temporal frames. We fix the partials of one excerpt, and apply a logarithmic scaling to the partials of the other over a range of one full octave in 1/8th semitone steps. Through an exhaustive search we can identify the frequency shift which maximises the consonance between the two excerpts and then apply the appropriate pitch-shifting factor prior to mixing the two excerpts together. A graphical overview of our approach is given in Figure 1.

Searching across a wide frequency range in small steps allows both for a large number of possible harmonic alignments and the ability to compensate for differences in tuning. In comparison with an existing commercial DJ-mixing system, we demonstrate our approach is able to provide more consonant mixes which are also considered more pleasant by musically trained listeners.

The remainder of this paper is structured as follows. In Section 2 we review existing approaches for the measurement of consonance based on roughness and pitch commonality. In Section 3 we describe our approach for consonance-based music mixing driven by these models. We then address the evaluation of our approach in Section 4 via a listening test. Finally, in Section 5 we present conclusions and areas for future work.

2. CONSONANCE MODELS

In this section, we present the theoretical approaches for the computational estimation of consonance that will form the core of the overall implementation described in Section 3 for estimating the most consonant combination of two tracks. To avoid misunderstandings due to ambiguous terminology, we define consonance by means of Terhardt's psychoacoustic model [1, 2], which is divided into two categories: The first, *sensory consonance* combines *roughness* (and *fluctuations*, standing for slow beatings and therefore equated with roughness throughout), *sharpness* and *tonalness*. The second, *harmony* is mostly built upon Terhardt's virtual pitch theory and inherits *root relationship* and *pitch commonality*. We take these categories as the basis for our approach. To estimate the degree of sensory consonance, we use a modified version of Hutchinson & Knopoff's [3] roughness model. For calculating the pitch commonality of a combination of sonorities, we propose a model that combines Parncutt & Strasburger's [4] pitch categorisation procedure with Hofmann-Engl's [5] virtual pitch model. Both models take a sequence of sinusoids, expressed as frequencies, f_i, and amplitudes, M_i, as input.

2.1. Roughness Model

As stated above, the category of sensory consonance can be divided into three parts: roughness, tonalness and sharpness. While sharpness is closely connected to timbral properties of musical audio, we do not attempt to model or modify this aspect since it can be considered independent of the interaction of two pieces of music, which is the object of our investigation in this paper.

Parncutt & Strasburger [4] discuss the strong relationship between roughness and tonalness as a sufficient reason to only analyse one of the two properties. The fact that roughness has been more extensively explored than tonalness and that most sensory consonance models build exclusively upon it motivates the use of roughness as our sole descriptor for sensory consonance in this work. For each of the partials of a spectrum, the roughness that is evoked by the co-occurrence with other partials is computed, then weighted by the dyads' amplitudes and finally summed for every sinusoid.

The basic structure of this procedure is a modified version of Hutchinson & Knopoff's [6] roughness model for complex sonorities that builds on the roughness curve for pure tone sonorities proposed by Plomp & Levelt [7]. A function that approximates the graph estimated by Plomp & Levelt is proposed by Parncutt [8]:

$$g(y) = \begin{cases} (\exp(1)\frac{y}{0.25}\exp(-\frac{y}{0.25}))^2 & y < 1.2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $g(y)$ is the degree of roughness of a dyad and y the frequency interval between two partials (f_i and f_j) expressed in the critical bandwidth (CBW) of the mean frequency $\bar{f}$, such that:

$$y = \frac{|f_j - f_i|}{\text{CBW}(\bar{f})} \quad (2)$$

and

$$\bar{f} = \frac{f_i + f_j}{2}. \quad (3)$$

Hutchinson & Knopoff's formula for the calculation of the critical bandwidth is often the subject of criticism (see, for example [8, 9]). Parncutt [8] states that better results can be obtained by using Moore & Glasberg's [10] equation for the equivalent rectangular bandwidth (ERB):

$$\text{ERB}(\bar{f}) = 6.23(10^{-3}\bar{f})^2 + 93.39(10^{-3}\bar{f}) + 28.52 \quad (4)$$

and hence we substitute $\text{CBW}(\bar{f})$ with $\text{ERB}(\bar{f})$ in eqn (2). The roughness values $g(y)$ for every dyad are then weighted by the

dyad's amplitudes (M_i and M_j) to obtain a value of the overall roughness D of a complex sonority with N partials:

$$D = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} M_i M_j g_{ij}}{\sum_{i=1}^{N} M_i^2}. \quad (5)$$

2.2. Pitch Commonality Model

As opposed to sensory consonance, which can be applied to any arbitrary sound, the second category of Terhardt's consonance model [1, 2] is largely specified on musical sounds. This is why the incorporation of an aspect based on harmony should be of critical importance in a system that aligns music according to consonance. However, the analysis of audio with a harmonic model of consonance is currently under-explored in the literature. Existing consonance-based tools for music typically focus on roughness alone [11, 12]. Relevant approaches which include harmonic analysis perform note extraction, categorisation in the octave-ranged chromagram and, as a consequence of this, key detection, but the psychoacoustic aspect of harmony is rarely applied. One of our main aims in this work is therefore to use the existing theoretical background to develop a model that estimates the consonance in terms of root relationship and pitch commonality and eventually to combine this with a roughness model.

The fundament of the approach lies in harmonic patterns in the spectrum. The extraction of these patterns is taken from the preprocessing stage of the pitch categorisation procedure of Parncutt & Strasburger's [4] tonalness model.

For a given set of partials, the audibilities of pitch categories in semitone intervals are produced. Since this corresponds directly to the notes of the chromatic scale, the degree of audibility for different pitch categories can be attributed to a chord. Hofmann-Engl's [5] virtual pitch model then will be used to compute the "Hofmann-Engl pitch sets" of these chords which will be compared for their commonality.

2.2.1. Pitch Categorisation

The first step of Parncutt & Strasburger's algorithm is the calculation of the pure-tone height, $H_p(f_i)$, for every frequency peak, f_i, in the spectrum using the analytic formula by Moore & Glasberg [10] that expresses the critical band rate in ERB:

$$H_p(f_i) = H_1 \log_e(\frac{f_i + f_1}{f_i + f_2}) + H_0. \quad (6)$$

As parameters, Moore & Glasberg propose H_1 = 11.17 erb, H_0 = 43.0 erb, f_1 = 312 Hz and f_2 = 14675 Hz. They also estimate the auditory level ΥL of each pure tone with the frequency f_i that is defined as its dB level above the threshold in quiet L_{TH}, which Terhardt [13] formulates as:

$$L_{TH} = 3.64{f_i}^{-0.8} - 6.5\exp\left(-0.6(f_i - 3.3)^2\right) + 10^{-3}{f_i}^4. \quad (7)$$

Then, the partial masking level $ml(f_i, f_j)$ which is the degree of how much every pure-tone in the sonority with the frequency f_i is masked by an adjacent pure-tone with its specific frequency f_j and auditory level $\Upsilon L(f_j)$ is estimated as

$$ml(f_i, f_j) = \Upsilon L(f_j) - k_m |H_p(f_j) - H_p(f_i)| \quad (8)$$

where k_m can take values between 12 and 18 dB (chosen value: 12 dB). The partial masking level is specified in dB. The overall masking level, $ML(f_i)$, of every-pure tone is obtained by adding up its partial masking levels, which are converted first to amplitudes and then, after the addition, back to dB levels:

$$ML(f_i) = \max(0, (20\log_{10} \sum_{P \neq P'} 10^{(ml(f_i, f_j)/20)})). \quad (9)$$

In the case of a pure-tone with frequency f_i that is not masked, $ml(f_i, f_j)$ will take a large negative value. This negative value for $ML(f_i)$ is avoided by use of the the $\max$ operator when comparing the calculated value to zero.

The decision not to analyse pure-tone components in frequency, but in pitch categories is due to the need to extract harmonic patterns. The pitch categories, P, are defined by their centre frequencies in Hz:

$$P(f_i) = 12\log_2(\frac{f_i}{440}) + 57 \quad (10)$$

where the standard pitch of 440 Hz (musical note A_4) is represented by pitch category 57.

Following this procedure for each component, we can now obtain its audible level $AL(P)$ (in dB) by subtracting its overall masking level from its auditory level $\Upsilon L(f)$:

$$AL(P) = \max(0, (\Upsilon L(P) - ML(P))). \quad (11)$$

To incorporate the saturation of each pure-tone with increasing audible level, Parncutt & Strasburger [4] estimate the audibility $A_p(P)$ for each pure-tone component:

$$A_p(P) = 1 - \exp(\frac{-AL(P)}{AL_0}). \quad (12)$$

where they follow Hesse [14] who sets AL_0 = 15.

Once every pure-tone component has been assigned to its corresponding pitch category and its audibility estimated, a template is used to detect partials of harmonic complex tones shifted over the spectrum in a step size of one semitone, i.e., one pitch category. One pattern's element is given by the formula:

$$P_n = P_1 + \lfloor 12\log_2(n) + 0.5 \rfloor \quad (13)$$

where P_1 represents the pitch category of the lowest element (corresponding to the fundamental frequency) and P_n the pitch category of the n^{th} harmonic.

Whenever there is a match between the template and the spectrum for each semitone-shift, a complex-tone audibility $A_c(P_1)$ is assigned to the template's fundamental. To take the lower audibility of higher harmonics into account, they are weighted by their harmonic number, n:

$$A_c(P_1) = \frac{1}{k_T} \left(\sum_n \sqrt{\frac{A_p(P_n)}{n}} \right)^2. \quad (14)$$

Parncutt & Strasburger [4] set the free parameter $k_T = 3$. To estimate the audibility, $A(P)$, of a component which considers both the spectral- and complex-tone audibility of every category, the overall maximum is taken as the general audibility, as Terhardt et al. [13] state that only either a pure or a complex tone can be perceived at once:

$$A(P) = \max(A_p(P), A_c(P)). \quad (15)$$

2.2.2. Pitch-Set Commonality

The resulting set of pitch categories can be understood as a chord with each pitch category's note sounding according to its audibility. With the focus on music, we set a limit of the three notes of the sonority with the highest audibility as the triad – which is seen as the most important chord in Western culture [15]. On this basis we expect it to give a meaningful representation of the harmonic structure.

To compare two chords according to their pitch-commonality, Hofmann-Engl proposes to estimate their similarity by the aid of the pitch-sets that are produced by his virtual pitch model [16]. The obtained triad is first inserted into a table similar to the one Terhardt uses to analyse a chord for its root note (see [2]), with the exception that Hofmann-Engl's table contains one additional subharmonic. The notes are ordered from low to high along with their corresponding different subharmonics. A major difference to Terhardt's model is the introduction of two weights w_1 and w_2 to estimate the strength β_{note} for a specific note to be the root of the chord with $Q = 3$ tones for all 12 notes of an octave:

$$\beta_{\text{note}} = \frac{\sum_{q=1}^{Q} w_{1,\text{note}}\, w_{2,q}}{Q} \tag{16}$$

where the result is a set of 12 strengths of notes, or so-called "Hofmann-Engl pitches" [16]. The fusion weight, $w_{1,\text{note}}$, is based on note similarity and gives the subharmonics more impact in decreasing order. This implies that the unison and the octave have the highest weight, then the fifth, the major third and so on. The maximum value of $w_{1,note}$ is $c = 6$ Hh (Helmholtz, unit set by Hofmann-Engl). The fusion weight is decreased by the variable b, which is $b = 1$ Hh for the fifth, $b = 2$ Hh for the major third, $b = 3$ Hh for the minor seventh, $b = 4$ Hh for the major second and $b = 5$ Hh for the major seventh. All other intervals take the value $b = 6$ and are therefore weighted zero, according to the formula:

$$w_{1,note} = \frac{c^2 - b^2}{c}. \tag{17}$$

The weight according to pitch order, w_2, adds more importance to lower notes, assuming that a lower note is more likely to be perceived as the root of the chord than a higher one and is calculated as:

$$w_{2,q} = \sqrt{\frac{1}{q}} \tag{18}$$

where q represents the position of the note in the chord. For the comparison between two sonorities (e.g. from different tracks), the Pearson correlation $r_{\text{set}_1\text{set}_2}$ is calculated for the pair of Hofmann-Engl pitch sets, as Hofmann-Engl [16] proposes to determine chord similarity and therefore consonance, C, in the sense of harmony as:

$$C = r_{\text{set}_1\text{set}_2}. \tag{19}$$

3. CONSONANCE-BASED MIXING

Based on the models of roughness and pitch commonality presented in the previous section, we now describe our approach for consonance-based mixing between two pieces of music.

3.1. Data Collection and Pre-Processing

We first explain the necessary pre-processing steps which allow the subsequent measurement of consonance between two pieces of music. For the purpose of this paper, which represents our first investigation into consonance-based mixing, we make several simplifications concerning the properties of the musical audio we intend to mix.

Given that our motivation is to compare our approach to key-based matching methods in DJ-mixing software (see Section 4), we currently only consider electronic music (e.g. house music) which is both harmonically stable and typically has a fixed tempo. From a collection of recent electronic music we manually annotated the tempo and beat locations and extracted a set of musical excerpts, each lasting precisely 16 beats (i.e., 4 complete bars).

In order to focus entirely on the issue of harmonic alignment without the need to address temporal alignment, we force the tempo of each excerpt to be exactly 120 beats per minute. For this beat quantisation process, we use the open source pitch-shifting and time-stretching utility, Rubberband[4], to implement any necessary tempo changes. Accordingly, our database of musical excerpts consists of a set of 8 s (i.e., 500 ms per beat) mono .wav files sampled at 44.1 kHz.

To provide an initial set of frequencies and amplitudes, we use a sinusoidal model, namely the "Spectral Modeling Synthesis Tools" Python software package by Serra[5], with which we extract sinusoids using the default window size and hop sizes of 4096 and 256 samples respectively. In order to focus on the harmonic structure present in the musical input, we extract the partials with the highest amplitude under 5 kHz. Through informal experimentation, we set $I = 20$ partials as we found this was able to provide a sufficient harmonic representation for our consonance-based mixing application. However, we intend to explore the effect of this parameter in future work.

For our chosen genre of electronic music, we can assume that the harmonic structure remains largely constant over the duration of each 1/16th note (i.e., 125 ms). Therefore, to strike a balance between temporal resolution and computational complexity, we summarise the frequencies and amplitudes by taking the frame-wise median over the duration of each 1/16th note. Thus, for each excerpt we obtain a set of frequencies and amplitudes, $f_{\gamma,i}$ and $M_{\gamma,i}$, where i indicates the partial number (up to $I = 20$) and γ each 1/16th note frame (up to $\Gamma = 64$).

3.2. Consonance-Based Alignment

For two input musical excerpts, $T^{\mathbf{1}}$ and $T^{\mathbf{2}}$ with corresponding frequencies and amplitudes $f^{\mathbf{1}}_{\gamma,i}$,$M^{\mathbf{1}}_{\gamma,i}$ and $f^{\mathbf{2}}_{\gamma,i}$,$M^{\mathbf{2}}_{\gamma,i}$ respectively, we seek to find the optimal consonance-based alignment between them. To this end, we fix all information regarding $T^{\mathbf{1}}$ and modify $T^{\mathbf{2}}$.

Our approach centres on the calculation of consonance as a function of a frequency shift, s, and is based on the hypothesis that under some frequency shift applied to $T^{\mathbf{2}}$ the consonance between $T^{\mathbf{1}}$ and $T^{\mathbf{2}}$ will be maximised, and this, in turn, will lead to the optimal mix between the two excerpts.

In total we create $S = 97$ shifts which cover the range of $\pm$ 6 semitones in 1/8th semitone steps (i.e., 48 downward and 48 upward shifts around a single "no shift" option). We scale the

[4] https://bitbucket.org/breakfastquay/rubberband

[5] https://github.com/MTG/sms-tools

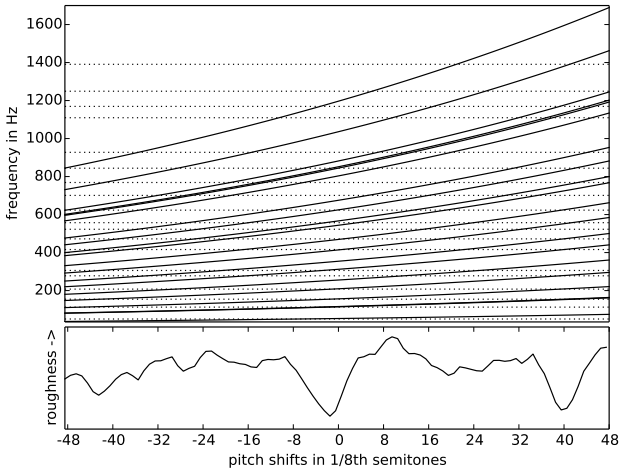

Figure 2: *(upper plot) Frequency scaling applied to the partials of one track (solid lines) compared to the fixed partials of the other (dotted lines) for a single temporal frame. (lower plot) The corresponding roughness as function of frequency scaling over that frame.*

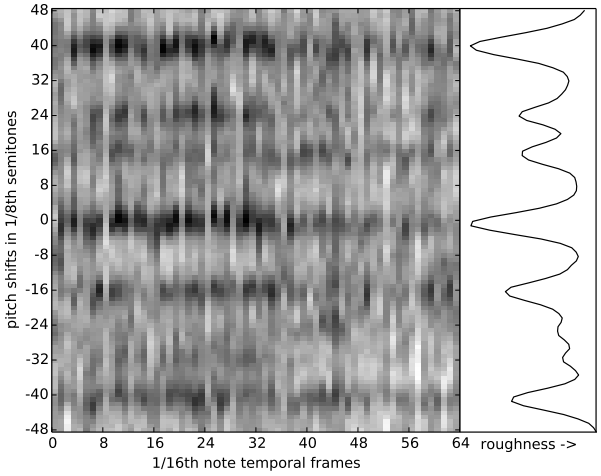

Figure 3: *Visualisation of roughness, $D_\gamma[s]$, over 64 frames for the full range of pitch-shifts. Darker regions indicate lower roughness. The subplot on the right shows the the average roughness curve, $\bar{D}[s]$ as a function of pitch-shift, where the roughness minima point to the left.*

frequencies of the partials $f^{\mathbf{2}}_{\gamma,i}$ as follows:

$$f^{\mathbf{2}}_{\gamma,i}[s] = 2^{\log_2(f^{\mathbf{2}}_{\gamma,i}) + \frac{s-48}{96}} \quad s = 0, \ldots, S-1. \tag{20}$$

For each 1/16th note temporal frame, γ, and per shift, s, we then merge the corresponding frequencies and amplitudes between both tracks (as shown in Figure 2) such that:

$$f_\gamma[s] = \left[f^{\mathbf{1}}_\gamma \;\; f^{\mathbf{2}}_\gamma[s] \right] \tag{21}$$

and

$$M_\gamma[s] = \left[M^{\mathbf{1}}_\gamma \;\; M^{\mathbf{2}}_\gamma[s] \right]. \tag{22}$$

We then calculate the roughness, $D_\gamma[s]$ according to eqn (5) in Section 2.1 with the merged partials and amplitudes as input. Then, to calculate the overall roughness, $\bar{D}[s]$, as a function of frequency shift, s, we average the roughness values $D_\gamma[s]$ across the temporal frames:

$$\bar{D}[s] = \frac{1}{\Gamma} \sum_{\gamma=0}^{\Gamma-1} D_\gamma[s], \tag{23}$$

for which a graphical example is shown in Figure 3.

Having calculated the roughness across all possible frequency shifts, we now turn our focus towards the measurement of pitch commonality as described in Section 2.2. Due both to the high computational demands of the pitch commonality model, and the rounding which occurs to due to the allocation of discrete pitch categories, we do not calculate the harmonic consonance as a function of all possible frequency shifts. Instead we extract all local minima from $\bar{D}[s]$, label these frequency shifts, s^*, and then proceed with this subset. In this way we use the harmonic consonance, C, as a means to filter and rank the set of possible alignments (i.e., minima) arising from the roughness model.

While the calculation of $D_\gamma[s]$ relies on the merged set of frequencies and amplitudes from eqns (21) and (22), the harmonic consonance compares two individually calculated Hoffman-Engl pitch sets. To this end, we calculate eqns. (6) to (16) independently for $f^{\mathbf{1}}_\gamma$ and $f^{\mathbf{2}}_\gamma[s^*]$ to create $\text{set}^{\mathbf{1}}_\gamma$ and $\text{set}^{\mathbf{2}}_\gamma[s^*]$ and hence $C_\gamma[s^*]$ from eqn (19). The overall harmonic consonance $\bar{C}[s^*]$ can then be calculated by averaging across the temporal frames:

$$\bar{C}[s^*] = \frac{1}{\Gamma} \sum_{\gamma=0}^{\Gamma-1} C_\gamma[s^*]. \tag{24}$$

Since no prior method exists for combining the roughness and harmonic consonance we adopt a simple approach to equally weight their contributions to give an overall measure of consonance based on roughness and pitch commonality:

$$\rho[s^*] = \widehat{D}[s^*] + \widehat{C}[s^*] \tag{25}$$

where $\widehat{D}[s^*]$ corresponds to the raw roughness values $\bar{D}[s^*]$ which have been inverted (to reflect sensory consonance as opposed to roughness) and then normalised to the range [0,1], and $\widehat{C}[s^*]$ similarly represents the [0,1] normalised version of $\bar{C}[s^*]$. The overall consonance $\rho[s^*]$ takes values that range from 0 (minimum consonance) to 2 (maximum consonance), as shown in Figure 4. The maximum score of 2 is achieved only if the roughness and harmonic consonance detect the same pitch-shift index as most consonant.

3.3. Post-Processing

The final stage of the consonance-based mixing is to physically implement the mix between tracks $T^{\mathbf{1}}$ and $T^{\mathbf{2}}$ under the consonance-maximising pitch shift, i.e., $\arg\max_{s^*}(\rho[s^*])$. As in Section 3.1, we again use the Rubberband utility to undertake the pitch-shifting on $T^{\mathbf{2}}$. To avoid loudness differences between the two tracks prior to mixing, we normalise each audio excerpt to a reference loudness level using the Replay Gain method [17].

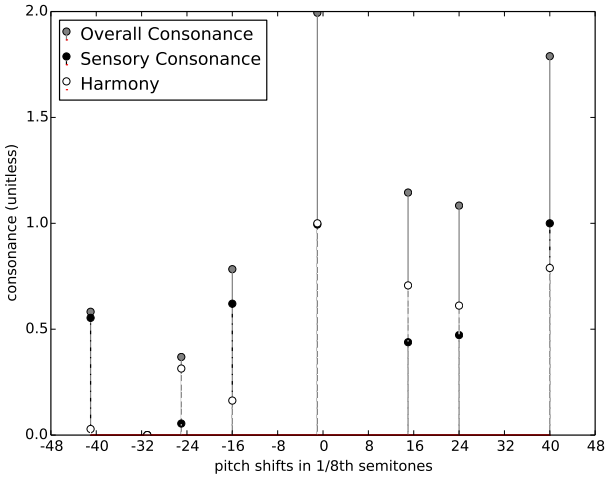

Figure 4: *Values of consonance from the sensory consonance model, $\widehat{D}[s^*]$, the harmonic consonance, $\widehat{C}[s^*]$, and the resulting overall consonance, $\rho[s^*]$. Pitch shift index -1 (i.e., -0.125 semitones) holds the highest consonance value and is the system's choice for the most consonant shift.*

4. EVALUATION

4.1. Listening Test

For the objective evaluation of our consonance-based mixing approach, we conducted a listening test. In this test we asked musically trained participants to rate short mixes created according to different outputs of our system, as well as those derived from the DJ-mixing software Traktor from Native Instruments, for their consonance and pleasantness. In total we created five conditions which are summarised as follows:

- **A** No Shift: we made no attempt to harmonically align the excerpts, instead simply aligned them in time by beat-matching.
- **B** Key Match (Traktor): we ran the key detection algorithm inside Traktor on each excerpt individually to determine the smallest pitch shift required to enable a harmonically compatible match based on the circle of fifths.[6]
- **C** Dissonant: we pitch-shifted according to the highest roughness from the roughness model without considering harmony in terms of pitch commonality.
- **D** Consonant (Sensory): we pitch-shifted according to lowest roughness from the roughness model without considering harmony.
- **E** Consonant (Sensory + Harmony): we pitch-shifted according to the result of the proposed combination of both models of roughness and pitch commonality.

Using a set of 20 excerpts (each 8 s in duration) as described in Section 3.1 we calculated the pitch-shifts required for all possible combinations between excerpts. From this complete set, we extracted a subset of 10 mixes (each made from different source excerpts) for which each of the 5 conditions yielded a unique pitch-shift. In total this gave a set of 50 musical stimuli for use in our experiment. The corresponding pitch-shifts for each mix for each of these stimuli and conditions are shown in Figure 5. Sound examples of some stimuli used in the listening test are available at the following website[7].

In total we recruited 28 participants whose musical training was determined by them being: music students, practicing musicians, or active in DJing. When listening to each mix, the participants were asked to rate two properties: first, how consonant the mixes sounded, and second they were asked to rate pleasantness of the mixes.

Both conditions were rated on a discrete six-point scale using a custom patch developed in Max/MSP. The order of the 50 stimuli was randomised for each participant. After every sound example, the ratings had to be entered before proceeding to the next example. To guarantee familiarity with the experimental procedure and stimuli, a training phase preceded the main experiment. This was also used to ensure all participants understood the concept of consonance and to set the playback volume to a comfortable level.

While the main goal was to assess the ability of our method to measure consonance, the pleasantness question was included to take into account the fact that musical consonance cannot be trivially equated with pleasantness of the sound [18], and furthermore to ensure that the definition of musical consonance was not confused with personal taste.

Regarding our hypotheses on the proposed conditions, we expected condition **C** (Dissonant) to be the least consonant, followed by **A** (No Shift). However, without any harmonic alignment, its behaviour was not predictable. Of the remaining conditions which attempted to find a good harmonic alignment, we expected the following order of consonance: **B** (Traktor) followed by **D** Consonant (Sensory) and finally our proposed combination **E** Consonant (Sensory + Harmony) the most consonant.

While the results of the sensory model have been explored in existing work [19, 11, 20], this experiment is, to the best of our knowledge, the first listener assessment of a combined roughness and harmonic model.

4.2. Results

Inspection of Figure 6, which shows the average ratings per excerpt across all conditions and criteria, reveals a wide range of ratings with some mixes considered very high in terms of consonance and pleasantness, while others were rated very low. In fact, the ratings across the two criteria of consonance and pleasantness were very strongly related with a correlation coefficient of .94. This supports our underlying assumption that a high level of consonance can be seen as a major factor for creating a good sounding mix.

By looking at the difference between different conditions in Figure 6 we can observe that in 8 of 10 cases (mixes), condition **D** (Consonant (Sensory)) - was rated more consonant than condition **A** (No Shift) and condition **C** (Dissonant). Regarding pleasantness, this was the case for every mix. The two mixes that showed the unexpected result of **D** being rated less consonant than **B** were mixes 2 and 3. Both had the lowest average ratings for consonance for all conditions (1.74, respectively 1.89, overall average 2.43). This might suggest that either one or both of the individual input

[6] `http://www.djprince.no/site/camelot_easymix_system.aspx`

[7] `http://telecom.inescporto.pt/~mdavies/dafx15/`

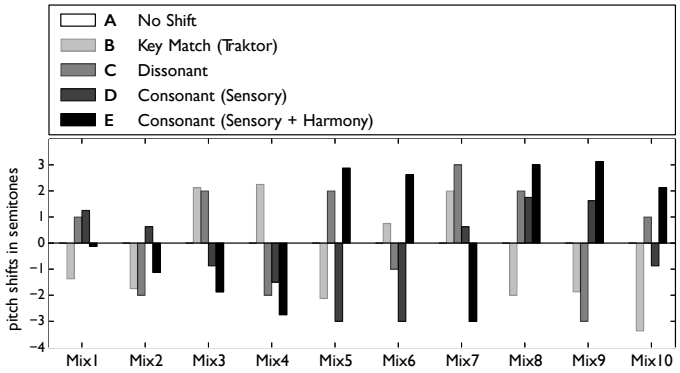

Figure 5: *Comparison of suggested pitch shifts under each condition for the listening experiment, where pitch-shifts are expressed in semitones. Note, the "No Shift" condition is always zero.*

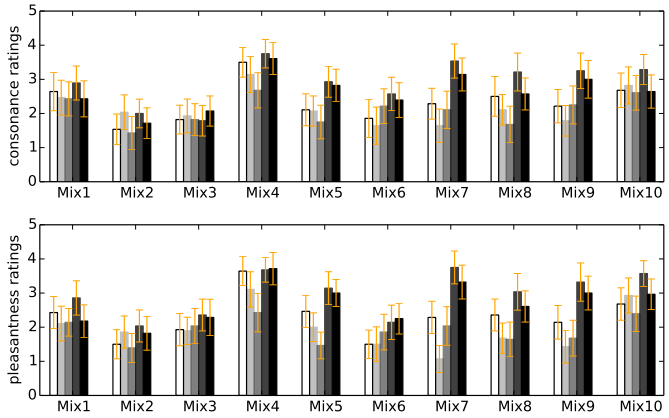

Figure 6: *Average ratings from participants of the listening experiment for consonance (upper plot) and pleasantness (lower plot). The error bars indicate the 95% confidence intervals. The shading to indicate the different conditions is as per Figure 5.*

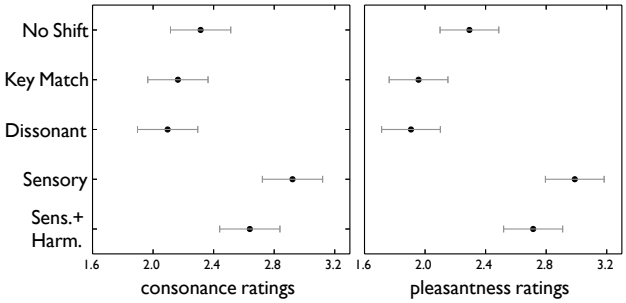

Figure 7: *Summary of multiple comparisons between conditions (with the Bonferroni correction) for consonance and pleasantness ratings. Error bars without overlap indicate statistically significant differences in the mean.*

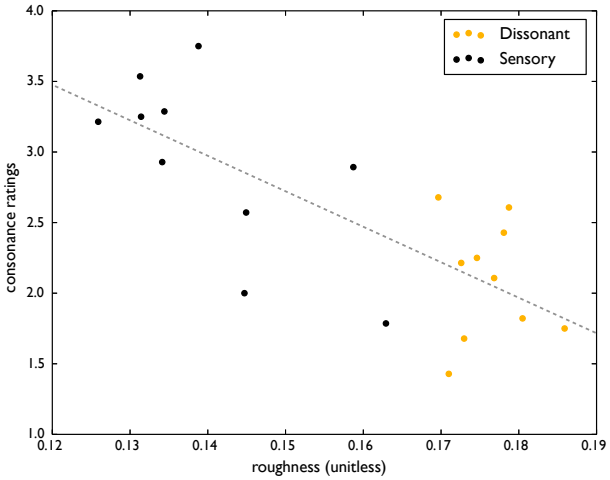

Figure 8: *Scatter plot of roughness versus consonance ratings for the stimuli of conditions* **C** *and* **D**.

tracks (prior to mixing) already contained dissonant sounds and was therefore always understood as dissonant by the participants, no matter what it was mixed with.

Comparing conditions **D** with **E** shows that the addition of the harmonic model, in general, did not improve the consonance or pleasantness ratings. In fact, the harmonic approach (**E**) was only rated more consonant once and more pleasant twice. However, it was still preferred over **A** and **C** for consonance eight times and in terms of pleasantness nine times. Therefore, our simple linear combination of roughness and pitch commonality does not seem obligatory to maximise the consonance. The inclusion of the harmonic model did appear to provide good alternative pitch-shifts, and hence expand the range of "good" harmonic alignments (see Figure 5).

Perhaps the most interesting result found in the listening test was the fact that both developed models (**D** and **E**) were rated more consonant than the mixes from condition **B** (Key Match Traktor) in 8 of 10 cases. These results were even better for pleasantness, where **D** was always preferred over **B**. These observations support the hypothesis that a consonance-based system can produce better harmonic alignments than those using the key detection method of Traktor aligned using the circle of fifths.

In addition to making direct observation of the ratings across conditions and mixes from Figure 6, we also conducted a statistical test to determine if these differences were significant. To this end, we performed a multiple comparison of means test which included the Bonferroni correction to adjust for variance between mixes. The mean ratings per condition with error bars are shown in Figure 7. For both consonance and pleasantness, condition **D** was rated significantly higher than conditions **A**, **B** and **C** ($p < .0001$ comparing **D** to **A** – the highest rated among the three), however there was no significant difference between **D** and **E**. As shown in Figure 7 condition **B** is among the lowest rated and has no significant difference even from **C** which we expected to be rated lowest. A possible explanation for this may be the failure of the key induction algorithm in Traktor to cope with such short music excerpts (each just 8 s in duration). We intend to explore this result and conduct comparisons with other key-based DJ-mixing software systems in future work.

The fact that roughness seems to have a major effect on the

rating of consonance (and hence pleasantness) motivates a closer investigation into their relationship. To this end, the roughness values of all mixes for conditions **C** and **D**, which represent the global extrema of the calculated roughness curves, were compared with their associated consonance ratings, as shown in shown in Figure 8. From the scatter plot we can observe a strong negative correlation (with coefficient of $-.75$) between the two. This relationship further supports the idea that roughness provides a meaningful perceptual scale for harmonic alignment of music signals.

5. CONCLUSIONS

In this paper we have presented a new method for harmonic mixing targeted towards addressing some of the limitations of commercial key-based DJ-mixing systems. Our approach centres on the use of psychoacoustic models of roughness and pitch commonality to identify an optimal harmonic alignment between different pieces of music across a wide range of possible pitch-shifts. Via a listening experiment with musically trained participants we were able to demonstrate that, within the context of the musical stimuli used, mixes based on roughness were considered significantly more consonant than those aligned according to musical key. Furthermore, the inclusion of the harmonic consonance model provided alternative pitch-shifts which were also significantly more pleasant than those of a commercial system.

In terms of future work, we intend to further explore how to weight the contribution of the roughness and harmonic consonance models. We also plan to extend the model to allow it to search across the temporal dimension of music to identify the most consonant temporal alignment between two musical excerpts. To this end, we will investigate more computationally efficient solutions to enable real-time interactive consonance-based music mixing, as well as experimentation with different musical genres.

6. REFERENCES

[1] E. Terhardt, "The concept of musical consonance: A link between music and psychoacoustics," *Music Perception: An Interdisciplinary Journal*, pp. 276–295, 1984.

[2] E. Terhardt, *Akustische Kommunikation (Acoustic Communication)*, Springer, Berlin, 1998, in German.

[3] W. Hutchinson and L. Knopoff, "The significance of the acoustic component of consonance of western triads," *Journal of Musicological Research*, vol. 3, pp. 5–22, 1979.

[4] R. Parncutt and H. Strasburger, "Applying psychoacoustics in composition: "harmonic" progressions of "non-harmonic" sonorities," *Perspectives of New Music*, vol. 32, no. 2, pp. 1–42, 1994.

[5] L. Hoffman-Engl, "Virtual pitch and pitch salience in contemporary composing," in *Proceedings of VI Brazilian Symposium on Computer Music*, Rio de Janeiro, Brazil, 1999.

[6] W. Hutchinson and L. Knopoff, "The acoustic component of western consonance," *Interface*, vol. 7, pp. 1–29, 1978.

[7] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.

[8] R. Parncutt, "Parncutt's implementation of Hutchinson & Knopoff (1978)," Available at http://uni-graz.at/ parncutt/rough1doc.html, accessed May 11, 2015.

[9] D. Huron, "Music 829B: Consonance and Dissonance," Available at http://www.musiccog.ohio-state.edu/Music829B/tonotopic.html, accessed May 11, 2015.

[10] B. Moore and B. Glassberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, 1983.

[11] J. MacCallum and A. Einbond, "Real-time analysis of sensory dissonance," in *Computer Music Modeling and Retrieval. Sense of Sounds*, R. Kronland-Martinet, S. Ystad, and K. Jensen, Eds., vol. 4969 of *Lecture Notes in Computer Science*, pp. 203–211. Springer Berlin Heidelberg, 2008.

[12] P. N. Vassilakis, "SRA: A Web-based Research Tool for Spectral and Roughness Analysis of Sound Signals," in *Proceedings of Sound and Music Computing Conference*, 2007, pp. 319–325.

[13] E. Terhardt, M. Seewan, and G. Stoll, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, pp. 671–678, 1982.

[14] A. Hesse, "Zur Ausgeprägtheit der Tonhöhe gedrosselter Sinustöne (Pitch Strength of Partially Masked Pure Tones)," in *Fortschritte der Akustik*, 1985, pp. 535–538, In German.

[15] W. Apel, *The Harvard Dictionary of Music*, Harvard University Press, Cambridge, 2nd edition, 1970.

[16] L. Hoffman-Engl, "Virtual pitch and the classification of chords in minor and major keys," in *Proceedings of ICMPC10*, Sapporo, Japan, 2008.

[17] D. Robinson, *Perceptual model for assessment of coded audio*, Ph.D. thesis, University of Essex, 2002.

[18] R. Parncutt, *Harmony: A psychoacoustical approach*, Springer, Berlin, 1989.

[19] G. Bernardes, M. E. P. Davies, C. Guedes, and B. Pennycook, "Considering roughness to describe and generate vertical musical structure in content-based algorithmic-assisted audio composition," in *Proceedings of ICMC/SMC*, Athens, Greece, September 2014, pp. 318–324.

[20] B. Hansen, "Modeling sensory dissonance in space: Revelations in sonic sculpture," M.S. thesis, University of California Santa Barbara, 2012.

[21] H. von Helmholtz, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik (On the Sensations of Tone)*, Vieweg, Braunschweig, 1863, in German.